LETTER Communicated by Justin Dauwels

Reverse-Engineering Neural Networks to Characterize Their Cost Functions

Takuya Isomura
takuya.isomura@riken.jp
Brain Intelligence Theory Unit, RIKEN Center for Brain Science, Wako, Saitama 351-0198, Japan

Karl Friston
k.friston@ucl.ac.uk
Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, WC1N 3AR, U.K.

This letter considers a class of biologically plausible cost functions for neural networks, where the same cost function is minimized 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 partially observed Markov decision processes (POMDP), we show that neural activity and plasticity perform Bayesian inference and learning, respectively, by maximizing model evidence. Using mathematical and numerical analyses, we establish the formal equivalence between neural network cost functions and variational free energy under some prior beliefs about latent states that generate inputs. These prior beliefs are determined by particular constants (e.g., thresholds) that define the cost function. This means that the Bayes optimal encoding of latent or hidden states is achieved when the network's implicit priors match the process that generates its inputs. This equivalence is potentially important because it suggests that any hyperparameter of a neural network can itself be optimized—by minimization with respect to variational free energy. Furthermore, it enables one to characterize a neural network formally, in terms of its prior beliefs.

1 Introduction

Cost functions are ubiquitous in scientific fields that entail optimization—including physics, chemistry, biology, engineering, and machine learning. Furthermore, any optimization 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 optimization process (Marr, 1969; Albus, 1971; Schultz, Dayan, & Montague, 1997; Sutton & Barto, 1998; Linsker, 1988; Brown, Yamada, & Sejnowski, 2001). These

Neural Computation **32, 1–37** (2020)
https://doi.org/10.1162/neco_a_01315

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 characterize the neural network's behavior in terms of fixed points, basin of attraction and structural stability—based 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 (Bishop, 2006; Dayan & Abbott, 2001). Generally the evaluation of surprise is intractable (especially for neural networks) as it entails a logarithm of an intractable marginal (i.e., integral) However, this evaluation can be converted into an optimization problem by inducing a variational bound on surprise. In machine learning, this is known as an evidence lower bound (ELBO; Blei, Kucukelbir, & McAuliffe, 2017), while the same quantity is known as variational free energy in statistical physics and theoretical neurobiology.

Variational free energy minimization is a candidate principle that governs neuronal activity and synaptic plasticity (Friston, Kilner, & Harrison, 2006; Friston, 2010). Here, surprise reflects the improbability of sensory inputs given a model of how those inputs were caused. In turn, minimizing variational free energy, as a proxy for surprise, corresponds to inferring the (unobservable) causes of (observable) consequences. To the extent that biological systems minimize 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, Zoccolan, & Rust, 2012) and consequently predict those inputs (Rao & Ballard, 1999; Friston, 2005). This is generally referred to as perceptual inference based on an internal generative model about the external world (Dayan, Hinton, Neal, & Zemel, 1995; George & Hawkins, 2009; Bastos et al., 2012).

Variational free energy minimization provides a unified mathematical formulation of these inference and learning processes in terms of self-organizing 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, Mattout, & Kilner, 2011). The ensuing free-energy principle suggests that active inference and learning are mediated by changes in neural activity, synaptic strengths, and the behavior of an organism to minimize variational free energy as a proxy for surprise. Crucially, variational free energy and model evidence rest on 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 minimize variational free energy (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2016; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017; Friston, Parr, & de Vries, 2017; Friston, Lin et al., 2017). This minimization reproduces various interesting dynamics and behaviors of real neuronal networks and biological organisms. However, it remains to be established whether variational free energy minimization is an apt explanation for any given neural network, as opposed to the optimization 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 minimizing 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. However, although the complete class theorem guarantees the existence of a generative model, it does not specify its form In what follows, we show that a ubiquitous class of neural networks implements approximates Bayesian inference under a generic discrete state space model with a known form.

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 follow a gradient descent on the cost function (assumption 1). For simplicity, we consider single-layer feedforward neural networks comprising firing-rate neuron models—receiving sensory inputs weighted by synaptic strengths—whose firing intensity is determined by the sigmoid activation function (assumption 2). We focus on blind source separation (BSS), namely the problem of separating sensory inputs into multiple hidden sources or causes (Belouchrani, Abed-Meraim, Cardoso, & Moulines, 1997; Cichocki, Zdunek, Phan, & Amari, 2009; Comon & Jutten, 2010), which provides the minimum setup for modeling causal inference. A famous example of BSS is the cocktail party effect: the ability of a partygoer to disambiguate an individual's voice from the noise of a crowd (Brown et al., 2001; Mesgarani & Chang, 2012). Previously, we observed BSS performed by in vitro neural networks (Isomura, Kotani, & Jimbo, 2015) and reproduced this self-supervised process using an MDP and variational free energy minimization (Isomura & Friston, 2018). These works suggest that variational free energy minimization offers a plausible account of the empirical behavior of in vitro networks.

In this work, we ask whether variational free energy minimization can account for the normative behavior of a canonical neural network that minimizes 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. The gradient descent on the ensuing cost function naturally leads to Hebbian plasticity (Hebb, 1949; Bliss & Lømo, 1973; Malenka & Bear, 2004) with an activity-dependent homeostatic term. We show that these cost functions are formally homologous to variational free energy under an MDP. Crucially, this means the hyperparameters (i.e., any variables or constants) of the neural network can be associated with prior beliefs of the generative model. In principle, this allows one to optimize the neural network hyperparameters (e.g., thresholds and learning rates), given some priors over the causes (i.e., latent states) of inputs to the neural network. Furthermore, estimating hyperparameters from the dynamics of (in silico or in vitro) neural networks allows one to quantify the network's implicit prior beliefs. In this letter, we focus on the mathematical foundations for applications to in vitro and in vivo neuronal networks in subsequent work.

2 Methods

In this section, we formulate the same computation in terms of variational Bayesian inference and neural networks to demonstrate their correspondence. We first derive the form of a variational free energy cost function under a specific generative model, a Markov decision process.[1] We present the derivations carefully, with a focus on the form of the ensuing Bayesian belief updating. The functional form of this update will reemerge later, when reverse engineering the cost functions implicit in neural networks. These correspondences are depicted in Figure 1 and Table 1. This section starts with a description of Markov decision processes as a general kind of generative model and then considers the minimization of variational free energy under these models.

2.1 Generative Models. Under an MDP model (see Figure 1A), a minimal BSS setup (in a discrete space) reduces to the likelihood mapping from N_s hidden sources or states $s_t \equiv \left(s_t^{(1)}, \ldots, s_t^{(N_s)}\right)^T$ to N_o observations $o_t \equiv \left(o_t^{(1)}, \ldots, o_t^{(N_o)}\right)^T$. Each source and observation takes a value of one (ON state)

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

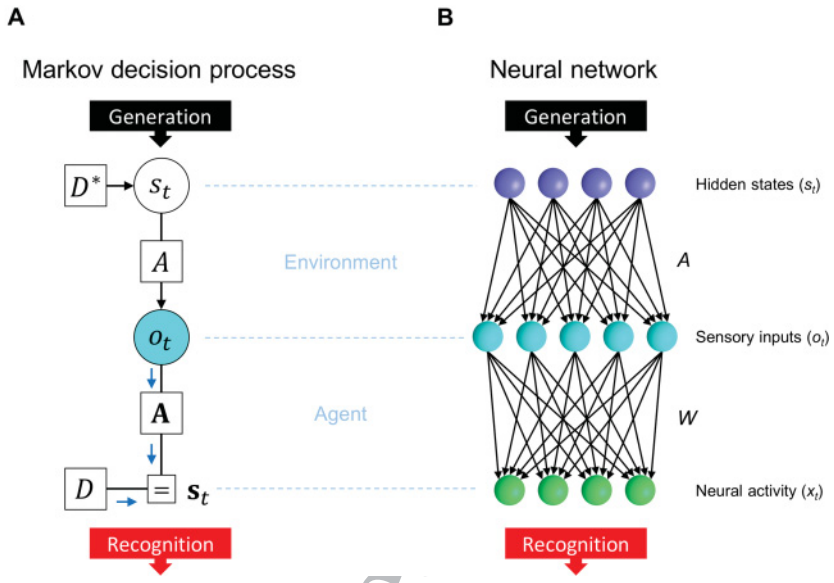

Figure 1: Comparison between an MDP scheme and a neural network. (A) MDP scheme expressed as a Forney factor graph (Forney, 2001; Dauwels, 2007) based on the formulation in Friston, Parr et al., 2017). In this BSS setup, the prior D determines hidden states s_t, while s_t determines observation o_t through 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 singlelayer feedforward network with a sigmoid activation function. The network receives sensory inputs $o_t = (o_t^{(1)}, \ldots, o_t^{(N_o)})^T$ that are generated from hidden states $s_t = (s_t^{(1)}, \ldots, s_t^{(N_s)})^T$ and outputs neural activities $x_t = (x_{t1}, \ldots, x_{tN_x})^T$. Here, x_{tj} should encode the posterior expectation about a binary state $s_t^{(j)}$. In an analogy with the cocktail party effect, s_t and o_t correspond to individual speakers and auditory inputs, respectively.

or zero (OFF state) at each time step, that is, $s_t^{(j)} o_t^{(i)} \in \{1, 0\}$. Throughout this letter, j denotes the jth hidden state, while i denotes the ith observation. The probability of $s_t^{(j)}$ follows a categorical distribution $P(s_t^{(j)}) = \mathrm{Cat}(D^{(j)})$, where $D^{(j)} \equiv (D_1^{(j)}, D_0^{(j)}) \in \mathbb{R}^2$ with $D_1^{(j)} + D_0^{(j)} = 1$ (see Figure 1A, top).

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, $P(o_t^{(i)} | s_t, A^{(i)}) = \mathrm{Cat}(A^{(i)})$ (see Figure 1A, middle). Here, each element of the tensor $A^{(i)} \in \mathbb{R}^{2 \times 2^{N_s}}$ parameterizes the probability that

Table 1: Correspondence of Variables and Functions.

Neural Network Formation		Variational Bayes Formation
Neural activity	$x_{tj} \Longleftrightarrow \mathbf{s}_{t1}^{(j)}$	State posterior
Sensory inputs	$o_t \Longleftrightarrow o_t$	Observations
Synaptic strengths	$W_{j1} \Longleftrightarrow \mathrm{sig}^{-1}\left(\mathbf{A}_{11}^{(\cdot,j)}\right)$	
	$\hat{W}_{j1} \equiv \mathrm{sig}(W_{j1}) \Longleftrightarrow \mathbf{A}_{11}^{(\cdot,j)}$	Parameter posterior
Perturbation term	$\phi_{j1} \Longleftrightarrow \ln D_1^{(j)}$	State prior
Threshold	$h_{j1} \Longleftrightarrow \ln\left(\vec{1} - \mathbf{A}_{11}^{(\cdot,j)}\right) \cdot \vec{1} + \ln D_1^{(j)}$	
Initial synaptic strengths	$\lambda_{j1} \odot \hat{W}_{j1}^{init} \Longleftrightarrow a_{11}^{(\cdot,j)}$	Parameter prior

$P\left(o_t^{(i)} = k | s_t = \vec{l}\right)$, 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 zero and one. Furthermore, learning the likelihood mapping leads to biologically plausible update rules, which have the form of associative or Hebbian plasticity (see below and Friston et al., 2016, for details).

We use $\tilde{o} \equiv (o_1, \ldots, o_t)$ and $\tilde{s} \equiv (s_1, \ldots, s_t)$ 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

$$\begin{aligned} P\left(\tilde{o}, \tilde{s}, A\right) &= P\left(A\right) \prod_{\tau=1}^{t} P\left(o_\tau | s_\tau, A\right) P\left(s_\tau\right) \\ &= \prod_{i=1}^{N_o} P(A^{(i)}) \cdot \prod_{\tau=1}^{t} \left\{ \prod_{i=1}^{N_o} P\left(o_\tau^{(i)} | s_\tau, A^{(i)}\right) \prod_{j=1}^{N_s} P\left(s_\tau^{(j)}\right) \right\}. \end{aligned} \tag{2.1}$$

Throughout this letter, t denotes the current time, and τ denotes an arbitrary time from the past to the present, $1 \leq \tau \leq t$.

2.2 Minimization of Variational Free Energy. In this MDP scheme, the aim is to minimize surprise by minimizing variational free energy as a

proxy, that is, 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,

$$Q\left(\tilde{s}, A\right) = Q\left(A\right) Q\left(\tilde{s}\right) = \prod_{i=1}^{N_o} Q(A^{(i)}) \cdot \prod_{\tau=1}^{t} \prod_{j=1}^{N_s} Q\left(s_\tau^{(j)}\right), \tag{2.2}$$

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 $A^{(i)}$ factorizes into the product of the likelihood mappings from the jth hidden state to the ith observation: $A_{k,:}^{(i)} \approx A_{k,:}^{(i,1)} \otimes \cdots \otimes A_{k}^{(i,N_s)}$ (where $\otimes$ denotes the outer product and $A^{(i,j)} \in \mathbb{R}^{2\times 2}$). This (mean-field) approximation simplifies the computation of state posteriors and serves to specify a particular form of Bayesian model which corresponds to a class of canonical neural networks (see below).

In what follows, a case variable in bold 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 zero and one. 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

$$\begin{aligned} \ln \mathbf{A}^{(i,j)} &\equiv \mathrm{E}_{Q\left(A^{(i,j)}\right)}\left[\ln A^{(i,j)}\right] = \psi\left(\mathbf{a}_{\cdot l}^{(i,j)}\right) - \psi\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) \\ &= \ln \mathbf{a}_{\cdot l}^{(i,j)} - \ln\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)}\right) + O\left(\left(\mathbf{a}_{\cdot l}^{(i,j)}\right)^{-1}\right), \end{aligned} \tag{2.3}$$

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

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

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

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

$$+\underbrace{\sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\{(\mathbf{a}^{(i,j)}-a^{(i,j)})\cdot\ln\mathbf{A}^{(i,j)}-\ln\mathcal{B}(\mathbf{a}^{(i,j)})\}}_{\text{parameter complexity}}, \quad (2.4)$$

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 on the order of $\ln t$—and is thus negligible when t is large (see section A.1 for details). This means that we can ignore parameter complexity when the scheme has experienced a sufficient number of outcomes. We drop the parameter complexity in subsequent sections. In the remainder of this section, we show how the minimization of variational free energy transforms (i.e., updates) priors into posteriors when the parameter complexity is evaluated explicitly.

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

$$\mathbf{s}_t^{(j)}=\sigma\left(\sum_{i=1}^{N_o}\ln\mathbf{A}^{(i,j)}\cdot o_t^{(i)}+\ln D^{(j)}\right)=\sigma\left(\ln\mathbf{A}^{(\cdot,j)}\cdot o_t+\ln D^{(j)}\right), \quad (2.5)$$

where $\sigma(\cdot)$ is the softmax function (see Figure 1A, bottom). 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

$$\mathbf{s}_{t1}^{(j)} = \frac{\exp\left(\ln \mathbf{A}_{\cdot 1}^{(\cdot,j)} \cdot o_t + \ln D_1^{(j)}\right)}{\exp\left(\ln \mathbf{A}_{\cdot 1}^{(\cdot,j)} \cdot o_t + \ln D_1^{(j)}\right) + \exp\left(\ln \mathbf{A}_{\cdot 0}^{(\cdot,j)} \cdot o_t + \ln D_0^{(j)}\right)}$$
$$= \mathrm{sig}\left(\ln \mathbf{A}_{\cdot 1}^{(\cdot,j)} \cdot o_t - \ln \mathbf{A}_{\cdot 0}^{(\cdot,j)} \cdot o_t + \ln D_1^{(j)} - \ln D_0^{(j)}\right) \tag{2.6}$$

using the sigmoid function $\mathrm{sig}(z) \equiv 1/(1+\exp(-z))$. Thus, the posterior expectation of $s_t^{(j)}$ taking a value zero (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 and only when the prior beliefs match the genuine prior distribution: $D_1^{(j)} = D_0^{(j)} = 0.5$ in this BSS setup.

This concludes our treatment of inference about hidden states under this minimal scheme. Note that the updates in equation 2.5 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, Movshon, & Dean, 1983; Newsome, Britten, & Movshon, 1989), while the arguments of the sigmoid (softmax) function can be associated with neuronal depolarization, rendering the softmax function a voltage-firing rate activation function. (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, and preplay activity 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

$$\mathbf{a}^{(i,j)} = a^{(i,j)} + \sum_{\tau=1}^{t} o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)} = a^{(i,j)} + t\overline{o_t^{(i)} \otimes \mathbf{s}_t^{(j)}}, \tag{2.7}$$

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

$$\begin{cases} \mathbf{A}_{11}^{(i,j)} = \dfrac{\mathbf{a}_{11}^{(i,j)}}{\mathbf{a}_{11}^{(i,j)} + \mathbf{a}_{01}^{(i,j)}} = \dfrac{t\overline{o_t^{(i)}\mathbf{s}_{t1}^{(j)}} + a_{11}^{(i,j)}}{t\overline{\mathbf{s}_{t1}^{(j)}} + a_{11}^{(i,j)} + a_{01}^{(i,j)}} = \dfrac{\overline{o_t^{(i)}\mathbf{s}_{t1}^{(j)}}}{\overline{\mathbf{s}_{t1}^{(j)}}} + O\left(\dfrac{1}{t}\right) \\ \mathbf{A}_{10}^{(i,j)} = \dfrac{\mathbf{a}_{10}^{(i,j)}}{\mathbf{a}_{10}^{(i,j)} + \mathbf{a}_{00}^{(i,j)}} = \dfrac{t\overline{o_t^{(i)}\mathbf{s}_{t0}^{(j)}} + a_{10}^{(i,j)}}{t\overline{\mathbf{s}_{t0}^{(j)}} + a_{10}^{(i,j)} + a_{00}^{(i,j)}} = \dfrac{\overline{o_t^{(i)}\mathbf{s}_{t0}^{(j)}}}{\overline{\mathbf{s}_{t0}^{(j)}}} + O\left(\dfrac{1}{t}\right) \end{cases}, \tag{2.8}$$

where $\overline{o_t^{(i)}\mathbf{s}_{t1}^{(j)}} = \frac{1}{t}\sum_{\tau=1}^{t} o_\tau^{(i)}\mathbf{s}_{\tau 1}^{(j)}$, $\overline{\mathbf{s}_{t1}^{(j)}} = \frac{1}{t}\sum_{\tau=1}^{t}\mathbf{s}_{\tau 1}^{(j)}$, $\overline{o_t^{(i)}\mathbf{s}_{t0}^{(j)}} = \frac{1}{t}\sum_{\tau=1}^{t} o_\tau^{(i)}\mathbf{s}_{\tau 0}^{(j)}$, and $\overline{\mathbf{s}_{t0}^{(j)}} = \frac{1}{t}\sum_{\tau=1}^{t}\mathbf{s}_{\tau 0}^{(j)}$. 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 on the order of one and is thus negligible when t is large. The matrix $\mathbf{A}^{(i,j)}$ expresses 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 normalized 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 2.7. 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 (see Figure 1B). The generation of observations o_t is exactly the same as in the MDP model introduced in section 2.1 (see Figure 1B, top to middle). We assume that the jth neuron's activity x_{tj} (see Figure 1B, bottom) is given by

$$\dot{x}_{tj} \propto -\underbrace{f'(x_{tj})}_{\text{leakage}} + \underbrace{W_{j1}o_t - W_{j0}o_t}_{\text{synaptic input}} + \underbrace{h_{j1} - j_{j0}}_{\text{threshold}}. \tag{2.9}$$

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) so that the fixed point of x_{tj} (i.e., the state of x_{tj} that gives $\dot{x}_{tj} = 0$) is given in the form of the sigmoid function:

$$\begin{aligned} x_{tj} &= \mathrm{sig}\left(W_{j1}o_t - W_{j0}o_t + h_{j1} - h_{j0}\right) \\ &= \frac{\exp\left(W_{j1}o_t + h_{j1}\right)}{\exp\left(W_{j1}o_t + h_{j1}\right) + \exp\left(W_{j0}o_t + h_{j0}\right)}. \end{aligned} \tag{2.10}$$

Equations 2.9 and 2.10 are a mathematical expression of assumption 2. Further, we consider a class of synaptic plasticity rules that comprise Hebbian

plasticity with an activity-dependent homeostatic term as follows:

$$\begin{cases} \Delta \mathrm{W_{j1}}\left(\mathrm{t}\right) \equiv \mathrm{W_{j1}}\left(\mathrm{t}+1\right) - \mathrm{W_{j1}}\left(\mathrm{t}\right) \propto \mathit{Hebb}_1\left(\mathrm{x_{tj}}, \mathrm{o_t}, \mathrm{W_{j1}}\right) + \mathit{Home}_1\left(\mathrm{x_{tj}}, \mathrm{W_{j1}}\right) \\ \Delta W_{j0}\left(t\right) \equiv W_{j0}\left(t+1\right) - W_{j0}\left(t\right) \propto \mathit{Hebb}_0\left(x_{tj}, o_t, W_{j0}\right) + \mathit{Home}_0\left(x_{tj}, W_{j0}\right) \end{cases}, \tag{2.11}$$

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. Equation 2.11 can be read as an ansatz: we will see below that a synaptic update rule with the functional form of equation 2.11 emerges as a natural consequence of assumption 1.

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 (see equation 2.6) and the activity in the neural network (see equation 2.10; see also Table 1). 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, $\vec{1} = (1, \ldots, 1) \in \mathbb{R}^{N_o}$ is a vector of ones. In particular, the optimal posterior of a hidden state taking a value of one (see equation 2.6) 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 $f'\left(x_{tj}\right)$ is the inverse of the sigmoid function (see equation 2.13). Under this condition the fixed point or solution for x_{tk} (see equation 2.10) compares inputs from ON and OFF pathways, and thus x_{tj} straightforwardly encodes the posterior of the jth 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 2.11 ensure that the neural activity and synaptic strengths asymptotically encode Bayes optimal posterior beliefs about hidden states $\left(x_{tj} \to \mathbf{s}_{t1}^{(j)}\right)$ and parameters $\left(W_{j1} \to \mathrm{sig}^{-1}\left(\mathbf{A}_{11}^{(\cdot,j)}\right)\right)$, respectively? To this end, 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 jth neuron's activity (see equation 2.9) is determined by the gradient of cost function L_j: $\dot{x}_{tj} \propto -\partial L_j / \partial x_{tj}$. By integrating the right-hand side of equation 2.9, we obtain a class of cost functions as

$$\begin{aligned} L_j &= \sum_{\tau=1}^{t} \left(f\left(x_{\tau j}\right) - x_{\tau j} W_{j1} o_\tau - \left(1 - x_{\tau j}\right) W_{j0} o_\tau - x_{\tau j} h_{j1} - \left(1 - x_{\tau j}\right) h_{j0} \right) + O\left(1\right) \\ &= \sum_{\tau=1}^{t} \left(f\left(x_{\tau j}\right) - \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left(\begin{pmatrix} W_{j1} \\ W_{j0} \end{pmatrix} o_\tau + \begin{pmatrix} h_{j1} \\ h_{j0} \end{pmatrix} \right) \right) + O(1), \end{aligned} \tag{2.12}$$

where the $O(1)$ term, which depends on W_{j1} and W_{j0}, is of a lower order than the other terms (as they are $\mathcal{O}\left(t\right)$) and is thus negligible when t is large (See section A.2 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 $p\left(A\right)$). The cost function of the entire network is defined by $L \equiv \sum_{j=1}^{N_x} L_j$. When $f'\left(x_{\tau j}\right)$ is the inverse of the sigmoid function, we have

$$f\left(x_{\tau j}\right) = x_{\tau j} \ln x_{\tau j} + \left(1 - x_{\tau j}\right) \ln\left(1 - x_{\tau j}\right) \tag{2.13}$$

up to a constant term (ensure $f'\left(x_{\tau j}\right) = \mathrm{sig}^{-1}\left(x_{\tau j}\right)$). We further assume that the synaptic weight update rule is given as the gradient descent on the same cost function L_j (see assumption 1). Thus, the synaptic plasticity is derived as follows:

$$\begin{cases} \dot{W}_{j1} \propto -\dfrac{1}{t} \dfrac{\partial L_j}{\partial W_{j1}} = \overline{x_{tj} o_t^T} + \overline{x_{tj}} h'_{j1} \\ \dot{W}_{j0} \propto -\dfrac{1}{t} \dfrac{\partial L_j}{\partial W_{j0}} = \overline{\left(1 - x_{tj}\right) o_t^T} + \overline{1 - x_{tj}} h'_{j0} \end{cases}, \tag{2.14}$$

where $\overline{x_{tj} o_t^T} \equiv \frac{1}{t} \sum_{\tau=1}^{t} x_{\tau j} o_\tau^T$, $\overline{x_{tj}} \equiv \frac{1}{t} \sum_{\tau=1}^{t} x_{\tau j}$, $\overline{\left(1 - x_{tj}\right) o_t^T} \equiv \frac{1}{t} \sum_{\tau=1}^{t} (1 - x_{\tau j}) o_\tau^T$, $\overline{1 - x_{tj}} \equiv \frac{1}{t} \sum_{\tau=1}^{t} \left(1 - x_{\tau j}\right)$, $h'_{j1} \equiv \partial h_{j1} / \partial W_{j1}$, 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, Girolami, Bell, & Sejnowski, 2000; Kusmierz, Isomura, & Toyoizumi, 2017). 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

$$\begin{cases} W_{j1} = h_1'^{-1}\left(-\dfrac{\overline{x_{tj}o_t^T}}{\overline{x_{tj}}}\right) \\ W_{j0} = h_0'^{-1}\left(-\dfrac{\overline{(1-x_{tj})\,o_t^T}}{\overline{1-x_{tj}}}\right) \end{cases}. \tag{2.15}$$

Crucially, these synaptic strength updates are a subclass of the general synaptic plasticity rule in equation 2.11 (see also section A.3 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. The updates of neural activity and synaptic strengths—via gradient descent on the cost function—enable us to associate neural and synaptic dynamics with optimization. Although the steepest descent method gives the simplest implementation, other gradient descent schemes, such as adaptive moment estimation (Adam; Kingma & Ba, 2015), can be considered, while retaining the local learning property.

2.5 Comparison with Variational Free Energy. Here, we establish a formal relationship between the cost function L and variational free energy. We define $\hat{W}_{j1} \equiv \mathrm{sig}(W_{j1})$ and $\hat{W}_{j0} \equiv \mathrm{sig}\left(W_{j0}\right)$ as the sigmoid functions of synaptic strengths. We consider the case in which neural activity is expressed as a sigmoid function and thus equation 2.13 holds. As $W_{j1} = \ln \hat{W}_{j1} - \ln\left(\vec{1} - \hat{W}_{j1}\right)$, equation 2.12 becomes

$$\begin{aligned} L = \sum_{j=1}^{N_x}\sum_{\tau=1}^{t}\begin{pmatrix} x_{\tau j} \\ 1-x_{\tau j}\end{pmatrix}^T \Bigg\{ \begin{pmatrix} \ln x_{\tau j} \\ \ln\left(1-x_{\tau j}\right)\end{pmatrix} - \begin{pmatrix} \ln \hat{W}_{j1} & \ln\left(\vec{1}-\hat{W}_{j1}\right) \\ \ln \hat{W}_{j0} & \ln\left(\vec{1}-\hat{W}_{j0}\right)\end{pmatrix} \\ \times \begin{pmatrix} o_\tau \\ \vec{1}-o_\tau\end{pmatrix} - \begin{pmatrix} h_{j1} \\ h_{j0}\end{pmatrix} + \begin{pmatrix} \ln\left(\vec{1}-\hat{W}_{j1}\right) \\ \ln\left(\vec{1}-\hat{W}_{j0}\right)\end{pmatrix}\vec{1}\Bigg\} + O(1), \end{aligned} \tag{2.16}$$

where $\vec{1} = (1, \ldots, 1) \in \mathbb{R}^{N_o}$. One can immediately see a formal correspondence between this cost function and variational free energy (see equation 2.4). That is, when we assume that $\left(x_{tj}, 1-x_{tj}\right)^T = \mathbf{s}_t^{(j)}$, $\hat{W}_{j1} = \mathbf{A}_{11}^{(\cdot,j)}$, and $\hat{W}_{j0} = \mathbf{A}_{10}^{(\cdot,j)}$, equation 2.16 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 2.4).

Specifically, when the thresholds satisfy $h_{j1} - \ln\left(\vec{1} - \hat{W}_{j1}\right) \cdot \vec{1} = \ln D_1^{(j)}$ and $h_{j0} - \ln\left(\vec{1} - \hat{W}_{j0}\right) \cdot \vec{1} = \ln D_0^{(j)}$, equation 2.16 becomes equivalent to equation 2.4 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 $\phi_{j1} \equiv h_{j1} - \ln\left(\vec{1} - \hat{W}_{j1}\right) \cdot \vec{1}$ and $\phi_{j0} \equiv h_{j0} - \ln\left(\vec{1} - \hat{W}_{j0}\right) \cdot \vec{1}$ as functions of W_{j1} and W_{j0}, respectively, and can express the cost function as

$$L = \sum_{j=1}^{N_x} \sum_{\tau=1}^{t} \begin{pmatrix} x_{\tau j} \\ 1 - x_{\tau j} \end{pmatrix}^T \left\{ \begin{pmatrix} \ln x_{\tau j} \\ \ln\left(1 - x_{\tau j}\right) \end{pmatrix} - \begin{pmatrix} \ln \hat{W}_{j1} & \ln\left(\vec{1} - \hat{W}_{j1}\right) \\ \ln \hat{W}_{j0} & \ln\left(\vec{1} - \hat{W}_{j0}\right) \end{pmatrix} \times \begin{pmatrix} o_\tau \\ \vec{1} - o_\tau \end{pmatrix} - \begin{pmatrix} \phi_{j1} \\ \phi_{j0} \end{pmatrix} \right\} + O(1). \tag{2.17}$$

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

$$\begin{cases} \phi_{j1} = \ln D_1^{(j)} \\ \phi_{j0} = \ln D_0^{(j)} \end{cases} \tag{2.18}$$

and thus equation 2.17 is formally equivalent to the accuracy and state complexity terms of variational free energy.

This means that when the prior belief about states $\left(D^{(j)}\right)$ is a function of the parameter posteriors $\left(\mathbf{A}^{(\cdot, j)}\right)$, the general cost function under consideration can be expressed in the form of variational free energy, up to the $O(\ln t)$ term. A generic cost function L is suboptimal from the perspective of Bayesian inference unless ϕ_{j1} and ϕ_{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 ϕ_{j1} and ϕ_{j0} as linear functions of W_{j1} and W_{j0}, respectively, given by

$$\begin{cases} \phi_{j1} = \alpha_{j1} + W_{j1}\beta_{j1} \\ \phi_{j0} = \alpha_{j0} + W_{j0}\beta_{j0} \end{cases}, \tag{2.19}$$

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

$$\begin{cases} W_{j1} = \mathrm{sig}^{-1}\left(\dfrac{\overline{x_{tj}o_t^T}}{\overline{x_{tj}}} + \beta_{j1}^T\right) \\ W_{j0} = \mathrm{sig}^{-1}\left(\dfrac{\overline{(1-x_{tj})\,o_t^T}}{\overline{1-x_{tj}}} + \beta_{j0}^T\right) \end{cases}. \tag{2.20}$$

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

$$\begin{cases} \Delta \mathrm{W_{j1}} = \underbrace{\dfrac{\{\hat{W}_{j1} \odot (\vec{1} - \hat{W}_{j1})\}^{\odot -1}}{\overline{x_{tj}}}}_{\text{adaptive learning rate}} \odot \left\{ \underbrace{x_{(t+1)j}o_{t+1}^T}_{\text{Hebbian plasticity}} - \underbrace{(\hat{W}_{j1} - \beta_{j1}^T)x_{(t+1)j}}_{\text{homeostatic plasticity}} \right\} \\ \Delta W_{j0} = \underbrace{\dfrac{\{\hat{W}_{j0} \odot (\vec{1} - \hat{W}_{j0})\}^{\odot -1}}{\overline{1-x_{tj}}}}_{\text{adaptive learning rate}} \\ \qquad \odot \left\{ \underbrace{(1-x_{(t+1)j})o_{t+1}^T}_{\text{anti-Hebbian plasticity}} - \underbrace{(\hat{W}_{j0} - \beta_{j0}^T)(1-x_{(t+1)j})}_{\text{homeostatic plasticity}} \right\} \end{cases}, \tag{2.21}$$

where $\odot$ denotes the elementwise (Hadamard) product and $\left\{\hat{W}_{j1} \odot \left(\vec{1} - \hat{W}_{j1}\right)\right\}^{\odot -1}$ denotes the element-wise inverse of $\hat{W}_{j1} \odot \left(\vec{1} - \hat{W}_{j1}\right)$. This synaptic plasticity rule is a subclass of the general synaptic plasticity rule in equation 2.11.

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 minimizing 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 2.17):

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

where $\hat{W}_{j1} = \mathrm{sig}(W_{j1})$ and $\hat{W}_{j0} = \mathrm{sig}\left(W_{j0}\right)$ hold and ϕ_{j1} and ϕ_{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 (assumption 2). Additional terms denoted by ϕ_{j1} and ϕ_{j0} express the state prior, indicating that a generic cost function L is variational free energy under a suboptimal prior belief about hidden states: $\ln P\left(s_t^{(j)}\right) = \ln D^{(j)} = \phi_j$, where $\phi_j \equiv \left(\phi_{j1}, \phi_{j0}\right)$. This prior alters the landscape of the cost function in a suboptimal 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_{j1} and ϕ_{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 2.17 becomes

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

$$\times \begin{pmatrix} o_\tau \\ \vec{1} - o_\tau \end{pmatrix} - \begin{pmatrix} \alpha_{j1} + W_{j1}\beta_{j1} \\ \alpha_{j0} + W_{j0}\beta_{j0} \end{pmatrix} \Bigg\} \Bigg] + O(1), \tag{3.1}$$

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 section 4 for details).

The cost function of the neural networks considered is characterized only by ϕ_j. Thus, after fixing ϕ_j by fixing constraints $(\alpha_{j1}, \alpha_{j0})$ and (β_{j1}, β_{j0}), 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 (see section A2).

The fixed point of synaptic strengths that give the minimum of L is given analytically as equation 2.20, expressing that (β_{j1}, β_{j0}) deviates the center of the nonlinear mapping—from Hebbian products to synaptic strengths—from the optimal position (shown in equation 2.8). As shown in equation 2.14, 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 2.14 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 $t + 1$ (see equation 2.21).

Hence, based on assumptions 1 and 2 (irrespective of assumption 3), we find that the cost function approximates variational free energy. Table 1 summarizes this correspondence. Under this condition, neural activity encodes the posterior expectation about hidden states, $x_{\tau j} = \mathbf{s}_{\tau 1}^{(j)} = Q(s_\tau^{(j)} = 1)$, and synaptic strengths encode the posterior expectation of the parameters, $\hat{W}_{j1} = \mathrm{sig}(W_{j1}) = \mathbf{A}_{11}^{(\cdot,j)}$ and $\hat{W}_{j0} = \mathrm{sig}(W_{j0}) = \mathbf{A}_{10}^{(\cdot,j)}$. In addition, based on assumption 3, the threshold is characterized by constants $\{\alpha_{j1}, \alpha_{j0}, \beta_{j1}, \beta_{j0}\}$. From a Bayesian perspective, these constants can be viewed as prior beliefs, $\ln P(s_t^{(j)}) = \ln D^{(j)} = (\alpha_{j1} + W_{j1}\beta_{j1}, \alpha_{j0} + W_{j0}\beta_{j0})$. When and only when $(\alpha_{j1}, \alpha_{j0}) = (-\ln 2, -\ln 2)$ and $(\beta_{j1}, \beta_{j0}) = (\vec{0}, \vec{0})$, 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 suboptimal) 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 3.1. We considered a BSS comprising two hidden sources (or states) and 32 observations (or sensory inputs), formulated as an MDP. The two hidden sources show 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

$$\begin{cases} P\left(o_t^{(i)} = 1 | s_t, A^{(i)}\right) = A_{1\cdot}^{(i)} = \left(0, \frac{3}{4}, \frac{1}{4}, 1\right) & \text{for } 1 \le i \le 16 \\ P\left(o_t^{(i)} = 1 | s_t, A^{(i)}\right) = A_{1\cdot}^{(i)} = \left(0, \frac{1}{4}, \frac{3}{4}, 1\right) & \text{for } 17 \le i \le 32 \end{cases}. \tag{3.2}$$

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 remaining elements were given by $A_{0\cdot}^{(i)} = \vec{1} - A_{1}^{(i)}$. The state priors were varied between zero and one in keeping with $D_1^{(j)} + D_0^{(j)} = 1$. Synaptic strengths were initialized as values close to zero. The simulations preceded over $T = 10^4$ time steps. The simulations and analyses were conducted using Matlab. (The scripts are available at https://github.com/takuyaisomura/reverse_engineering.) 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 feedforward neural network model, and variational Bayesian formulations.

First, as in Isomura and Friston (2018), we demonstrated that a network with a cost function with optimized constants $\left(\left(\alpha_{j1}, \alpha_{j0}\right) = \left(-\ln 2, -\ln 2\right)\right.$ and $\left.\left(\beta_{j1}, \beta_{j0}\right) = \left(\vec{0}, \vec{0}\right)\right)$ can perform BSS successfully (see Figure 2). The responses of neuron 1 came to recognize source 1 after training, indicating that neuron 1 learned to encode source 1 (see Figure 2A). Meanwhile, neuron 2 learned to infer source 2 (see Figure 2B). During training, synaptic plasticity followed gradient descent on the cost function (see Figures 2C and 2D). This demonstrates that minimization of the cost function, with optimal constants, is equivalent to variational free energy minimization and hence is sufficient to emulate BSS. This process establishes a concise representation of the hidden causes and allows maximizing information retained in the neural network (Linsker, 1988; Isomura, 2018).

Next, we quantified the dependency of BSS performance on the form of the cost function, by varying the above-mentioned constants (Figure 3). We varied $\left(\alpha_{j1}, \alpha_{j0}\right)$ in a range of $0.05 \le \exp\left(\alpha_{j1}\right) \le 0.95$, while maintaining $\exp\left(\alpha_{j1}\right) + \exp\left(\alpha_{j0}\right) = 1$ and found that changing $\left(\alpha_{j1}, \alpha_{j0}\right)$ from

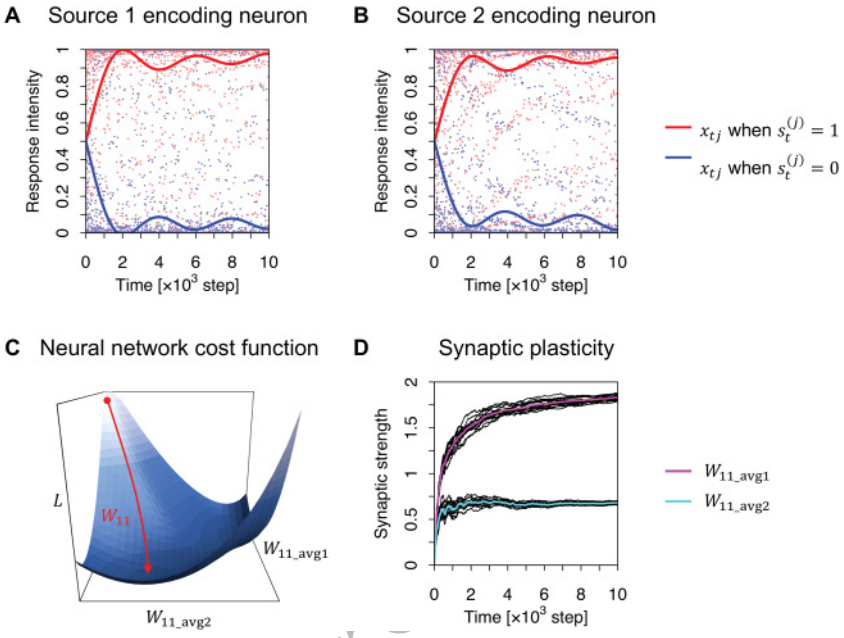

Figure 2: Emergence of response selectivity for a source. (A) Evolution of neuron 1's responses that learn to encode source 1, in the sense that the response is high when source 1 takes a value of one (red dots), and it is low when source 1 takes a value of zero (blue dots). Lines correspond to smoothed 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. (C) Neural network cost function L. It is computed based on equation 3.1 and plotted against the averaged synaptic strengths, where W_{11_avg1} (z-axis) is the average of 1 to 16 elements of W_{11}, while W_{11_avg2} (x-axis) is the average of 17 to 32 elements of W_{11}. The red line depicts a trajectory of averaged synaptic strengths. (D) Trajectory of synaptic strengths. Black lines show elements of W_{11}, and magenta and cyan lines indicate W_{11_avg1} and W_{11_avg2}, respectively.

$(-\ln 2, -\ln 2)$ led to a 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 the response of neuron 1 is nearly zero. In the case of suboptimal α, these correlations fall to around 0.5, indicating that the response of neuron 1 encodes a mixture of sources 1 and 2 (Figure 3A). Moreover, a failure of BSS can be induced when the elements of β take values far from zero (see Figure 3B). When the elements of β are generated from a zeromean gaussian distribution,

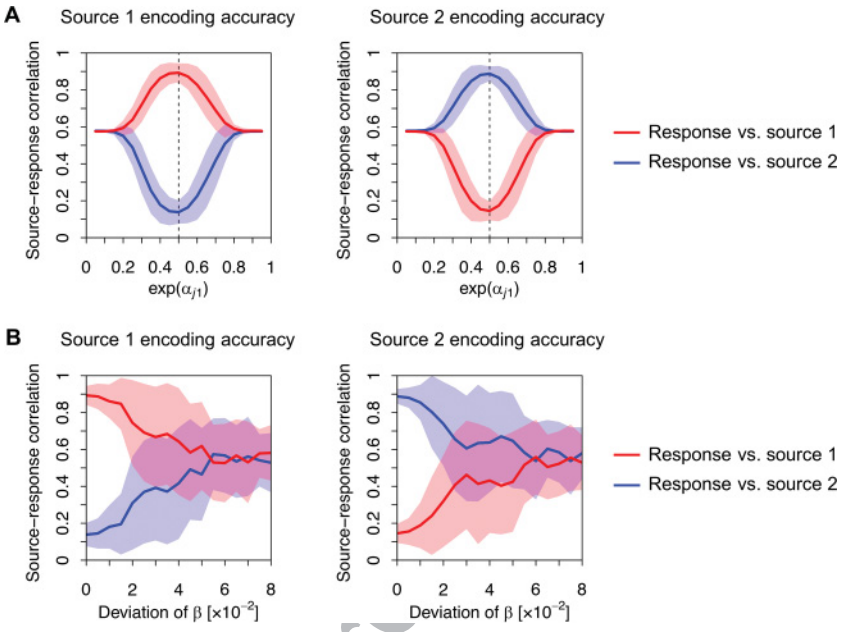

Figure 3: Dependence of source encoding accuracy on constants. Left panels show the magnitudes of the correlations between sources and responses of a neuron expected to encode source 1: $|\mathrm{corr}(s_t^{(1)}, x_{t1})|$ and $|\mathrm{corr}(s_t^{(2)}, x_{t1})|$. The right panels show the magnitudes of the correlations between sources and responses of a neuron expected to encode source 2: $|\mathrm{corr}(s_t^{(1)}, x_{t2})|$ and $|\mathrm{corr}(s_t^{(2)}, x_{t2})|$. (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 $\exp(\alpha_{j1})$. (B) Dependence on constant β when α is fixed as $(\alpha_{j1}, \alpha_{j0}) = (-\ln 2, -\ln 2)$. 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. Lines and shaded areas indicate the mean and standard deviation of the source-response correlation, evaluated with 50 different sequences.

the accuracy of BSS—measured using the correlation between sources and responses—decreases as the standard deviation increases.

Our numerical analysis, under assumptions 1 to 3, 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 formally homologous to 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 ϕ_j provides degrees of freedom in the shape of the neural network cost functions under consideration that determine the purpose or function of the neural network. Among various ϕ_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 neural networks (of the sort considered in this letter: single-layer feedforward networks that minimize their cost function) as performing approximate Bayesian inference under priors that may or may not be optimal. This result is as predicted by the complete class theorem (Brown, 1981; Wald, 1947) as it implies that any response of a neural network is Bayes optimal under some prior beliefs (and cost function). Therefore, in principle, under the theorem, 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 characterizing a neural network model—and indeed a real neuronal network—in terms of its implicit prior beliefs.

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

From a practical perspective for optimizing 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 optimize the constants by updating the priors themselves such that they minimize the variational free energy for BSS. Under the Dirichlet form for the priors, the implicit threshold constants of the objective function can then be optimized using the following updates:

$$\begin{aligned} \phi_j &= \ln D^{(j)} = \psi\left(\mathbf{d}^{(j)}\right) - \psi\left(\mathbf{d}_1^{(j)} + \mathbf{d}_0^{(j)}\right) \\ \mathbf{d}^{(j)} &= d^{(j)} + \sum_{\tau=1}^{t} \mathbf{s}_\tau^{(j)}. \end{aligned} \tag{3.3}$$

(See 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.

3.4 Reverse-Engineering Implicit Prior Beliefs. Another situation important from a neuroscience perspective is when belief updating in a neural network is slow in relation to experimental observations. In this case, the implicit prior beliefs can be viewed as being fixed over a short period of time. This is likely when such a firing threshold is determined by a homeostatic plasticity over longer timescales (Turrigiano & Nelson, 2004).

The considerations in the previous section speak to the possibility of using empirically observed neuronal responses to infer implicit prior beliefs. The synaptic weights (W_{j1}, W_{j0}) can be estimated statistically from response data, through equation 2.20. By plotting their 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_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 (β_{j1}, β_{j0}) fails as a Bayes optimal encoder for BSS. Since actual neuronal networks can perform BSS (Isomura et al., 2015; Isomura & Friston, 2018), one would envisage that the implicit cost function will exhibit a near-optimal ϕ_j.

In particular, when ϕ_j can be viewed as a constant $\left(\text{i.e., } \phi_j = (\alpha_{j1}, \alpha_{j0})^T\right)$ during a period of experimental observation, the characterization of thresholds is fairly straightforward: using empirically observed neuronal responses, through variational free energy minimization, under the constraint of $e^{\phi_{j1}} + e^{\phi_{j0}} = 1$, the estimator of ϕ_j is obtained as follows:

$$\phi_j = \ln\left(\frac{1}{t}\sum_{\tau=1}^{t}\begin{pmatrix} x_{tj} \\ 1 - x_{tj} \end{pmatrix}\right). \tag{3.4}$$

Crucially, equation 3.4 is exactly the same as equation 3.3 up to a negligible term $d^{(j)}$. However, equation 3.3 represents the adaptation of a neural network, while equation 3.4 expresses Bayesian inference about ϕ_j based on empirical data. We applied equation 3.4 to sequences of neural activity generated from the synthetic neural networks used in the simulations reported in Figures 2 and 3, and confirmed that the estimator was a good approximation to the true ϕ_j (see Figure 4A). This characterization enables the identification of implicit prior beliefs (D) and consequently, the reconstruction of the neural network cost function, that is, variational free energy; see Figure 4B). Furthermore, because a canonical neural network is supposed to use the same prior beliefs for the same sort of tasks, one can

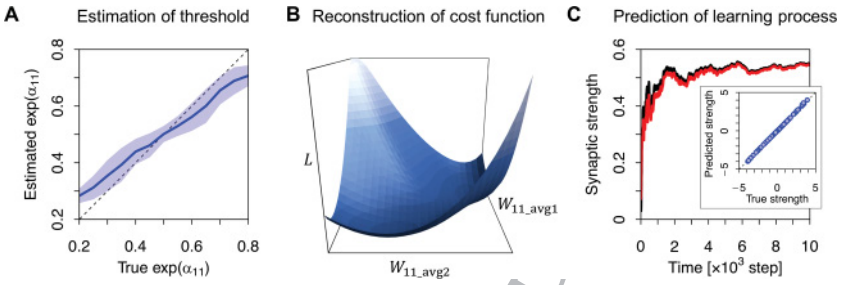

Figure 4: Estimation of prior beliefs enables the prediction of subsequent learning. (A) Estimation of constants (α_{11}) that characterize thresholds (i.e., prior beliefs) based on sequences of neural activity. Lines and shaded areas indicate the mean and standard deviation. (B) Reconstruction of cost function L using the obtained threshold estimator through equation 3.1. The estimated L well approximates the true L shown in Figure 2C. (C) Prediction of learning process with new sensory input data generated through different likelihood mapping characterized by a random matrix A. Supplying the sensory (i.e., input) data to the ensuing cost function provides a synaptic trajectory (red line), which predicts the true trajectory (black line) in the absence of observed neural responses. Inset panel depicts a comparison between elements of the true and predicted W_{11} at $t = 10^4$.

use the reconstructed cost functions to predict subsequent inference and learning without observing neural activity (see Figure 4C). These results highlight the utility of reverse-engineering neural networks to predict their activity, plasticity, and assimilation of input data.

4 Discussion

In this work, we investigated a class of biologically plausible cost functions for neural networks. A single-layer feedforward 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 minimize 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 terms. We have shown that the dynamics of a single-layer feedforward neural network, which minimizes 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 characterize 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. In short, this equivalence provides an insight that any neural and synaptic dynamics (in the class considered) have functional meaning and any neural network variables and constants can be formally associated with quantities in the variational Bayesian formation, implying that Bayesian inference is universal characterisation of canonical neural networks.

According to the complete class theorem (Brown, 1981; Wald, 1947), any dynamics that minimizes a cost function can be viewed as performing Bayesian inference under some prior beliefs (Wald, 1947; Brown, 1981). This implies that any neural network whose activity and plasticity minimize the same cost function can be cast as performing Bayesian inference. Moreover, when a system has reached a (possibly nonequilibrium) steady state, the conditional expectation of internal states of an autonomous system can be shown to parameterize a posterior belief over the hidden states of the external milieu (Friston, 2013, 2019; Parr, Da Costa, & Friston, 2020). Again, this suggests that any (nonequilibrium) steady state can be interpreted as realising some elemental Bayesian inference.

Having said this, we note that the implicit generative model that underwrites any (e.g., neural network) cost function is a more delicate problem—one that we have addressed in this work. In other words, it is a mathematical truism that certain systems can always be interpreted as minimizing a variational free energy under some prior beliefs (i.e., generative model). However, this does not mean it is possible to identify the generative model by simply looking at systemic dynamics. To do this, one has to commit to a particular form of the model, so that the sufficient statistics of posterior beliefs are well defined. We have focused on discrete latent variable models that can be regarded as special (reduced) cases of partially observable Markov decision processes (POMDP).

Note that 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, such as hierarchical predictive coding models (Friston, 2008; Whittington & Bogacz, 2017; Ahmadi & Tani, 2019). 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). We have described the generative model in terms of an MDP; however, we ignored state transitions. This means the generative model in this letter reduces to a simple latent variable model, with categorical states and outcomes. We have considered 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 on control variables, we have an MDP, and when states are only partially observed, we

have a partially observed MDP (POMDP). To deal with these general cases, extensions of the current framework are required, which we hope to consider in future work, perhaps with recurrent neural networks.

Our theory implies that Hebbian plasticity is a corollary (or realization) of cost function minimization. In particular, Hebbian plasticity with a homeostatic term emerges naturally from a gradient descent on the neural network cost function defined via the integral of neural activity. In other words, the integral of synaptic inputs $W_{j1}o_t$ in equation 2.9 yields $x_{tj}W_{j1}o_t$, and its derivative yields a Hebbian product $x_{tj}o_t^T$ in equation 2.14. This relationship indicates that this form of synaptic plasticity is natural for canonical neural networks. In contrast, a naive Hebbian plasticity (without a homeostatic term) fails to perform BSS because it updates synapses with false prior beliefs (see Figure 3). It is well known that a modification of Hebbian plasticity is necessary to realize BSS (Földiák, 1990; Linsker, 1997; Isomura & Toyoizumi, 2016), speaking to the importance of selecting the right priors for BSS.

The proposed equivalence between neural networks and Bayesian inference may offer insights into designing neural network architectures and synaptic plasticity rules to perform a given task—by selecting the right kind of prior beliefs—while retaining their biological plausibility. An interesting extension of the proposed framework is an application to spiking neural networks. Earlier work has highlighted relationships between spiking neural networks and statistical inference (Bourdoukan, Barrett, Deneve, & Machens, 2012; Isomura, Sakai, Kotani, & Jimbo, 2016). The current approach might be in a position to formally link spiking neuron models and spike-timing dependent plasticity (Markram et al., 1997; Bi & Poo, 1998; Froemke & Dan, 2002; Feldman, 2012) with variational Bayesian inference.

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 (Markram et al., 2004; Isaacson & Scanziani, 2011) and neuromodulators (Pawlak, Wickens, Kirkwood, & Kerr, 2010; Frémaux & Gerstner, 2016) in biological neuronal networks. 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 suboptimal manner—in relation to the process that generates inputs. Meanwhile, (β_{j1}, β_{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 (compare equation 2.8 and equation 2.20). 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 minimizing 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 ϕ_j in the cost functions L. In particular, if ϕ_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 future work, we plan to apply this scheme to empirical data and examine the biological plausibility of variational free energy minimization.

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.

Minimization of the cost function can render the neural network Bayes optimal in a Bayesian sense, including the choice of the prior, as described in the previous section. The dependence between the likelihood function and the state prior vanishes when the network uses an optimal threshold to perform inference—if the true generative process does not involve dependence between the likelihood and the state prior. In other words, the dependence arises from a suboptimal choice of the prior. Indeed, any free parameters or constraints in a neural network can be optimized by minimizing variational free energy. This is because only variational free energy with the optimal priors—that match the true generative process of the external world—can provide the global minimum among a class of neural network cost functions under consideration. This is an interesting observation because it suggests that the global minimum of the class of cost functions—that determine neural network dynamics—is characterized by and only

by statistical properties of the external world. This implies that the recapitulation of external dynamics is an inherent feature of canonical neural systems.

Finally, the free energy principle and complete class theorem imply that any brain function can be formulated in terms of variational Bayesian inference. Our reverse engineering may enable the identification of neuronal substrates or process models underlying brain functions by identifying the implicit generative model from empirical data. Unlike conventional connectomics (based on functional connectivity), reverse engineering furnishes a computational architecture (e.g., neural network), which encompasses neural activity, synaptic plasticity, and behavior. This may be especially useful for identifying neuronal mechanisms that underlie neurological or psychiatric disorders—by associating pathophysiology with false prior beliefs that may be responsible for things like hallucinations and delusions Fletcher & Frith, 2009; Friston , Stephan, Montague, & Dolan, 2014).

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 optimized 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 modeled using neural networks of the class that we have considered. In other words, if one can fit neuronal responses—using a neural network model parameterized 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 modeling of distributed neuronal responses from noninvasive data (Daunizeau, David, & Stephan, 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).

Appendix: Supplementary Methods

A.1 Order of the Parameter Complexity. The order of the parameter complexity term

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

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 ε. This means that all the elements of $\mathbf{a}_{\cdot l}^{(i,j)}$ are in the order of t. The first term of equation A) 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 2.3, $\mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln \mathbf{A}_{\cdot l}^{(i,j)} = \mathbf{a}_{\cdot l}^{(i,j)} \cdot \left(\ln \mathbf{a}_{\cdot l}^{(i,j)} - \ln \left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)} \right) + \mathcal{O}\left(\left(\mathbf{a}_{\cdot l}^{(i,j)} \right)^{-1} \right) \right) = \mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln \left(\mathbf{A}_{\cdot l}^{(i,j)} \right) + \mathcal{O}(1)$. Meanwhile, the second term of equation A.1 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,

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

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

$$\begin{aligned} \ln \Gamma\left(\mathbf{a}_{1l}^{(i,j)} \right) &= -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{1l}^{(i,j)} \left(\ln \mathbf{a}_{1l}^{(i,j)} - 1 \right) \\ &\quad + \ln \left(1 + \mathcal{O}\left(\left(\mathbf{a}_{\cdot l}^{(i,j)} \right)^{-1} \right) \right) \\ &= \mathbf{a}_{1l}^{(i,j)} \ln \mathbf{a}_{1l}^{(i,j)} - \mathbf{a}_{1l}^{(i,j)} + \mathcal{O}(\ln t) . \end{aligned} \tag{A.3}$$

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

$$\begin{aligned} \ln \mathcal{B}\left(\mathbf{a}_{\cdot l}^{(i,j)} \right) &= \mathbf{a}_{1l}^{(i,j)} \ln \mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)} \ln \mathbf{a}_{0l}^{(i,j)} - \left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)} \right) \\ &\quad \times \ln \left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{0l}^{(i,j)} \right) + \mathcal{O}(\ln t) \\ &= \mathbf{a}_{\cdot l}^{(i,j)} \cdot \ln \left(\mathbf{A}_{\cdot l}^{(i,j)} \right) + \mathcal{O}(\ln t) . \end{aligned} \tag{A.4}$$

Hence, equation 5.1 becomes

$$\mathcal{D}_A = \sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\sum_{l\in\{1,0\}}\left\{\mathbf{a}_{\cdot l}^{(i,j)}\cdot\ln\left(\mathbf{A}_{\cdot l}^{(i,j)}\right)+\mathcal{O}(1)-\left(\mathbf{a}_{\cdot l}^{(i,j)}\cdot\ln\left(\mathbf{A}_{\cdot l}^{(i,j)}\right)\right.\right.$$
$$\left.\left.+\ \mathcal{O}(\ln t)\right)\right\}=\mathcal{O}(\ln t)\,. \tag{A.5}$$

Therefore, we obtain

$$F\left(\tilde{o},Q\left(\tilde{s}\right),Q\left(A\right)\right)=\sum_{j=1}^{N_s}\sum_{\tau=1}^{t}\mathbf{s}_\tau^{(j)}\cdot\left\{\ln\mathbf{s}_\tau^{(j)}-\sum_{i=1}^{N_o}\ln\mathbf{A}^{(i,j)}\cdot o_\tau^{(i)}-\ln D^{(j)}\right\}$$
$$+\,O(\ln t)\,. \tag{A.6}$$

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 $O(\ln t)$ term in equation A.6 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

$$\delta F=\sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\sum_{\tau=1}^{t}\mathbf{s}_\tau^{(j)}\cdot\left\{-\delta\ln\mathbf{A}_{1\cdot}^{(i,j)}o_\tau^{(i)}-\delta\ln\left(\vec{1}-\mathbf{A}_{1\cdot}^{(i,j)}\right)\left(1-o_\tau^{(i)}\right)\right\}$$
$$=t\sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\left\{-\left(\delta\mathbf{A}_{1\cdot}^{(i,j)}\odot\left(\mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot-1}\right)\cdot\overline{o_t^{(i)}\otimes\mathbf{s}_t^{(j)}}\right.$$
$$\left.+\left(\delta\mathbf{A}_{1\cdot}^{(i,j)}\odot\left(\vec{1}-\mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot-1}\right)\cdot\overline{\left(1-o_t^{(i)}\right)\mathbf{s}_t^{(j)}}\right\}$$
$$=t\sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\left(\delta\mathbf{A}_{1\cdot}^{(i,j)}\odot\left(\mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot-1}\odot\left(\vec{1}-\mathbf{A}_{1\cdot}^{(i,j)}\right)^{\odot-1}\right)$$
$$\cdot\left(\mathbf{A}_{1\cdot}^{(i,j)}\odot\bar{\mathbf{s}}_t^{(j)}-\overline{o_t^{(i)}\mathbf{s}_t^{(j)}}\right) \tag{A.7}$$

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

$$\mathbf{A}_{1\cdot}^{(i,j)}=\overline{o_t^{(i)}\mathbf{s}_t^{(j)}}\odot\left(\overline{\mathbf{s}_t^{(j)}}\right)^{\odot-1}+\mathcal{O}\left(\frac{\ln t}{t}\right). \tag{A.8}$$

Therefore, we obtain the same result as equation 2.8 up to the order of $\ln t/t$.

A.2 Correspondence between Parameter Prior Distribution and Initial Synaptic Strengths. In general, optimizing 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 minimization 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. We have demonstrated that a class of cost functions—for a single-layer feedforward 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 ϕ_j, which characterize 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 do 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 2.12 to make the dependence on the initial synaptic strengths explicit. Specifically, we modify equation 2.12 as

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

$$+\left(\ln\left(\vec{1}-\hat{W}_{j1}\right),\ln\left(\vec{1}-\hat{W}_{j0}\right)\right)\left(\lambda_{j1},\lambda_{j0}\right)^{T}, \tag{A.9}$$

where $\hat{W}_{j1}^{init} \equiv \mathrm{sig}(W_{j1}^{init})$ and $\hat{W}_{j0}^{init} \equiv \mathrm{sig}(W_{j0}^{init})$ 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 A9 expresses the integral of $\hat{W}_{j1}$ and $\hat{W}_{j0}$ (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 A9 is zero), the derivative of L_j is given by $\partial L_j/\partial W_{j1} = \lambda_{j1} \odot \hat{W}_{j1}^{init} - \lambda_{j1} \odot \hat{W}_{j1}$, 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 2.12 to equation 2.17, we compute equation A9) as

$$\begin{aligned}
L = &\sum_{j=1}^{N_x}\sum_{\tau=1}^{t}\begin{pmatrix} x_{\tau j} \\ 1-x_{\tau j}\end{pmatrix}^{T}\left\{\begin{pmatrix} \ln x_{\tau j} \\ \ln\left(1-x_{\tau j}\right)\end{pmatrix} - \begin{pmatrix} \ln\hat{W}_{j1} & \ln\left(\vec{1}-\hat{W}_{j1}\right) \\ \ln\hat{W}_{j0} & \ln\left(\vec{1}-\hat{W}_{j0}\right)\end{pmatrix}\right. \\
&\left.\times\begin{pmatrix} o_\tau \\ \vec{1}-o_\tau\end{pmatrix} - \begin{pmatrix}\phi_{j1}\\ \phi_{j0}\end{pmatrix}\right\} \\
&+\sum_{j=1}^{N_x}\left\{\left(\ln\hat{W}_{j1},\ln\left(\vec{1}-\hat{W}_{j1}\right)\right)\left(\lambda_{j1}\odot\hat{W}_{j1}^{init},\lambda_{j1}\odot\left(\vec{1}-\hat{W}_{j1}^{init}\right)\right)^{T}\right. \\
&\left.+\left(\ln\hat{W}_{j0},\ln\left(\vec{1}-\hat{W}_{j0}\right)\right)\left(\lambda_{j0}\odot\hat{W}_{j0}^{init},\lambda_{j0}\odot\left(\vec{1}-\hat{W}_{j0}^{init}\right)\right)^{T}\right\}.
\end{aligned} \tag{A.10}$$

Note that we used $W_{j1} = \ln\hat{W}_{j1} - \ln\left(\vec{1}-\hat{W}_{j1}\right)$. Crucially, analogous to the correspondence between $\hat{W}_{j1}$ and the Dirichlet parameters of the parameter posterior $\mathbf{a}_{11}^{(\cdot,j)}$, $\lambda_{j1}\odot\hat{W}_{j1}^{init}$ 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 A.10 and the expectation of the log parameter prior in equation 2.4:

$$\begin{aligned}
\mathrm{E}_{Q(A)}\left[\ln P\left(A\right)\right] &= \sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\ln\mathbf{A}^{(i,j)}\cdot a^{(i,j)} \\
&= \sum_{i=1}^{N_o}\sum_{j=1}^{N_s}\left\{\ln\mathbf{A}_{\cdot 1}^{(i,j)}\cdot a_{\cdot 1}^{(i,j)} + \ln\mathbf{A}_{\cdot 0}^{(i,j)}\cdot a_{\cdot 0}^{(i,j)}\right\}.
\end{aligned} \tag{A.11}$$

Furthermore, the synaptic update rules are derived from equation A.10 as

$$\begin{cases} \dot{W}_{j1} \propto -\frac{1}{t}\frac{\partial L}{\partial W_{j1}} = \overline{x_{tj} o_t^T} - \overline{x_{tj}}\hat{W}_{j1} + \overline{x_{tj}}\phi'_{j1} + \frac{1}{t}\left(\lambda_{j1} \odot \hat{W}_{j1}^{init} - \lambda_{j1} \odot \hat{W}_{j1}\right) \\ \dot{W}_{j0} \propto -\frac{1}{t}\frac{\partial L}{\partial W_{j0}} = \overline{\left(1 - x_{tj}\right) o_t^T} - \overline{1 - x_{tj}}\hat{W}_{j0} + \overline{1 - x_{tj}}\phi'_{j0} \\ \qquad\quad + \frac{1}{t}\left(\lambda_{j0} \odot \hat{W}_{j0}^{init} - \lambda_{j0} \odot \hat{W}_{j0}\right) \end{cases} \tag{A.12}$$

The fixed point of equation A.12 is provided as

$$\begin{cases} W_{j1} = \mathrm{sig}^{-1}\left(\left(t\overline{x_{tj}}\vec{1} + \lambda_{j1}\right)^{\odot -1} \odot \left(t\overline{x_{tj} o_t^T} + t\overline{x_{tj}}\phi'_{j1} + \lambda_{j1} \odot \hat{W}_{j1}^{init}\right)\right) \\ W_{j0} = \mathrm{sig}^{-1}\left(\left(t\overline{1 - x_{tj}}\vec{1} + \lambda_{j0}\right)^{\odot -1}\right. \\ \qquad\quad \left. \odot \left(t\overline{\left(1 - x_{tj}\right) o_t^T} + t\overline{1 - x_{tj}}\phi'_{j0} + \lambda_{j0} \odot \hat{W}_{j0}^{init}\right)\right) \end{cases} \tag{A.13}$$

Note that the synaptic strengths at $t = 0$ are computed as $W_{j1} = \mathrm{sig}^{-1}\left(\left(\lambda_{j1}\right)^{\odot -1} \odot \left(\lambda_{j1} \odot \hat{W}_{j1}^{init}\right)\right) = W_{j1}^{init}$. Again, one can see the formal correspondence between the final values of the synaptic strengths given by equation A.13 in the neural network formation and the parameter posterior given by equation 2.8 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 $\overline{x_{tj} o_t^T}$ and $\lambda_{j1} \odot \hat{W}_{j1}^{init}$, respectively. Thus, equation 2.8 corresponds to equation A.13. 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

$$a^{(\cdot,j)} \equiv \begin{pmatrix} a_{11}^{(\cdot,j)} & a_{10}^{(\cdot,j)} \\ a_{01}^{(\cdot,j)} & a_{00}^{(\cdot,j)} \end{pmatrix} = \begin{pmatrix} \lambda_{j1} \odot \hat{W}_{j1}^{init} & \lambda_{j0} \odot \hat{W}_{j0}^{init} \\ \lambda_{j1} \odot \left(\vec{1} - \hat{W}_{j1}^{init}\right) & \lambda_{j0} \odot \left(\vec{1} - \hat{W}_{j0}^{init}\right) \end{pmatrix}. \tag{A.14}$$

In summary, one can establish the formal correspondence between neural network and variational Bayesian formations in terms of the cost functions (see equation 2.4 versus equation A.1), priors (see equations 2.18 and A.14), and posteriors (see equations 2.8 versus. equation 5.13). This means that a neural network successively transforms priors $P\left(s_t\right) P\left(A\right)$ into posteriors $Q\left(s_t\right) Q\left(A\right)$, as parameterized with neural activity, and initial and final synaptic strengths (and thresholds). Crucially, when increasing the number of observations, this process is asymptotically equivalent to that of variational Bayesian inference under a specific likelihood function.

A.3 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 2.15, $h_1'(W_{j1})$ satisfies both

$$h_1'\left(W_{j1} + \Delta W_{j1}\right) - h_1'(W_{j1}) = h_1''(W_{j1}) \odot \Delta W_{j1} + O\left(\left|\Delta W_{j1}\right|^2\right) \quad \text{(A.15)}$$

and

$$\begin{aligned} &h_1'\left(W_{j1} + \Delta W_{j1}\right) - h_1'(W_{j1}) \\ &= -\frac{x_{(t+1)j} o_{t+1}^T + t\overline{x_{tj} o_t^T}}{x_{(t+1)j} + t\overline{x_{tj}}} + \frac{\overline{x_{tj} o_t^T}}{\overline{x_{tj}}} \\ &\approx -\frac{x_{(t+1)j} o_{t+1}^T}{t\overline{x_{tj}}} + \frac{\overline{x_{tj} o_t^T}}{t\overline{x_{tj}}^2} x_{(t+1)j} = -\frac{1}{t\overline{x_{tj}}}\left(x_{(t+1)j} o_{t+1}^T - x_{(t+1)j} h_1'(W_{j1})\right). \end{aligned} \quad \text{(A.16)}$$

Thus, we find

$$\Delta W_{j1} = -\underbrace{\frac{h_1''(W_{j1})^{\odot -1}}{t\overline{x_{tj}}}}_{\text{adaptive learning rate}} \odot \left(\underbrace{x_{(t+1)j} O_{t+1}^T}_{\text{Hebbian term}} - \underbrace{x_{(t+1)j} h_1'(W_{j1})}_{\text{homeostatic term}} \right). \quad \text{(A.17)}$$

Similarly,

$$\Delta W_{j0} = -\underbrace{\frac{h_0''(W_{j0})^{\odot -1}}{t\overline{1 - x_{tj}}}}_{\text{adaptive learning rate}} \odot \left(\underbrace{(1 - x_{(t+1)j}) O_{t+1}^T}_{\text{anti-Hebbian term}} - \underbrace{(1 - x_{(t+1)j}) h_0'(W_{j0})}_{\text{homeostatic term}} \right). \quad \text{(A.18)}$$

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

Data Availability

All relevant data are within the letter. Matlab scripts are available at https://github.com/takuyaisomura/reverse_engineering.

Acknowledgments

T.I. is funded by the RIKEN Center for Brain Science. K.J.F. is funded by a Wellcome Principal Research Fellowship (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

Ahmadi, A., & Tani, J. (2019). A novel predictive-coding-inspired variational RNN model for online prediction and recognition. *Neural Comput.*, *31*, 2025–2074.

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.

Bi, G. Q., & Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.*, *18*, 10464–10472.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, *112*, 859–877.

Bliss, T. V., & Lømo, T. (1973). Longasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J. Physiol. 232*, 331–356.

Bourdoukan, R., Barrett, D., Deneve, S., & Machens, C. K. (2012). Learning optimal spike-based representations. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *25* (pp. 2285–2293) Red Hook, NY: Curran.

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*. Hoboken, NY: Wiley.

Comon, P., & Jutten, C. (2010). *Handbook of blind source separation: Independent component analysis and applications*. Orlando, FL: 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. In *Proceedings of the International Symposiun on Information Theory*. Piscataway, NJ: IEEE.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems.* Cambridge, MA: MIT Press.

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.

Feldman, D. E. (2012). The spike-timing dependence of plasticity. *Neuron*, *75*, 556–571.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neuros.*, *10*, 48–58.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.*, *64*, 165–170.

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. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.*, *4*, e1000211.

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

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface*, *10*, 20130475.

Friston, K. (2019). *A free energy principle for a particular physics*. arXiv:1906.10184.

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

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Comput.*, *29*, 2633–2683.

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

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

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *Lancet Psychiatry*, *1*, 148–158.

Froemke, R. C., & Dan, Y. (2002). Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature*, *416*, 433–438.

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

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

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

Isaacson, J. S., & Scanziani, M. (2011). How inhibition shapes cortical activity. *Neuron*, *72*, 231–243.

Isomura, T. (2018). A measure of information available for inference. *Entropy*, *20*, 512.

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

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., Sakai, K., Kotani, K., & Jimbo, Y. (2016). Linking neuromodulated spike-timing dependent plasticity with the free-energy principle. *Neural Comput.*, *28*, 1859–1888.

Isomura, T., & Toyoizumi, T. (2016). A local learning rule for independent component analysis. *Sci. Rep.*, *6*, 28073.

Kingma, D. P., & Ba, J. (2015). ADAM: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*. ICLR-15.

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.

Linsker, R. (1997). A local learning rule that enables information maximization for arbitrary input distributions. *Neural Comput.*, *9*, 1661–1665.

Malenka, R. C., & Bear, M. F. (2004). LTP and LTD: An embarrassment of riches. *Neuron*, *44*, 5–21.

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, *275*, 213–215.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., & Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nat. Rev. Neurosci.*, *5*, 793–807.

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

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*, 233–236.

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

Parr, T., Da Costa, L., & Friston, K. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Phil. Trans. R. Soc. A*, *378*, 20190159.

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.

Rolfe, J. T. (2016). *Discrete variational autoencoders*. arXiv:1609.02200.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

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

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

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.

Turrigiano, G. G., & Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nat Rev. Neurosci.*, *5*, 97–107.

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

Whittington, J. C., & Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Comput.*, *29*, 1229–1262.

Received December 20, 2019; accepted June 22, 2020.