Synaptic pruning facilitates online Bayesian model selection

Ukyo T. Tazawa[a,b*], Takuya Isomura[b*]

[a] Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo, 113-0033, Japan

[b] Brain Intelligence Theory Unit, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351- 0198, Japan

* Corresponding authors. E-mails: ukyo.tazawa@a.riken.jp; takuya.isomura@riken.jp

Abstract

Identifying appropriate structures for generative or world models is essential for both biological organisms and machines. This work shows that synaptic pruning facilitates efficient statistical structure learning. We extend previously established canonical neural networks to derive a synaptic pruning scheme that is formally equivalent to an online Bayesian model selection. The proposed scheme, termed Bayesian synaptic model pruning (BSyMP), utilizes connectivity parameters to switch between the presence (ON) and absence (OFF) of synaptic connections. Mathematical analyses reveal that these parameters converge to zero for uninformative connections, thus providing reliable and efficient model reduction. This enables the identification of a plausible structure for the environmental model, particularly when the environment is

characterized by sparse likelihood and transition matrices. Through causal inference and rule learning simulations, we demonstrate that BSyMP achieves model reduction more efficiently than the conventional Bayesian model reduction scheme. These findings indicate that synaptic pruning could be a neuronal substrate underlying structure learning and generalizability in the brain.

Keywords: Bayesian model selection; Synaptic pruning; Free-energy principle; Structure learning; Bayesian model reduction

1. Introduction

Characterizing brain intelligence is a significant challenge in the field of biology. Biological organisms recognize the external milieu by inferring underlying structures. Despite the importance of structure learning—the process of identifying an appropriate structure to represent the environment (Braun et al., 2010; Lynn & Bassett, 2020; Tenenbaum et al., 2011)—its neuronal mechanisms in the brain are not yet fully understood (Tervo et al., 2016). The structure of neural networks is shaped by the pruning of synaptic connections during development (Sakai, 2020), and several computational studies have explored the advantages of synaptic pruning for memory and learning (Chechik et al., 1998; Scholl et al., 2021; Thomas et al., 2015). However, analyzing the relationship between biologically grounded synaptic pruning and information-theoretically derived structure learning is still required.

The free-energy principle has been proposed in theoretical neuroscience to account for various brain functions, including perception, learning, and action, in terms of minimizing variational free energy (Friston, 2010; Friston et al., 2006). It models the brain as an agent that performs variational Bayesian inference of external states and parameters by minimizing the variational free energy. These inferences rest upon a generative or world model that describes how latent variables in the external milieu generate sensory inputs. Recent works have showed that the dynamics of canonical neural networks that have a certain biological plausibility can be conceptualized as variational Bayesian inference under implicit generative models characterized by network structures (Isomura et al., 2022; Isomura & Friston, 2020). This lends the explainability of Bayesian inference to investigate the functional properties of biologically plausible neural networks. Additionally, Bayesian inference can be employed in selecting model structures (Friston et al., 2019; Friston et al., 2017; Jafarian et al., 2019). However, the neural substrate of Bayesian model selection is yet to be established.

Herein, we show that biologically plausible synaptic pruning enables structure learning in a Bayes-optimal manner by extending the previously established equivalence between neural networks and Bayesian inference. We reveal that incorporating synaptic pruning into a canonical neural network is equivalent to online Bayesian model selection, a method we term Bayesian synaptic model pruning (BSyMP). BSyMP adopts a set of connectivity parameters that enables switching between the presence (ON) and absence (OFF) of the model parameters. These connectivity parameters, in addition to states and other parameters, are updated by minimizing the variational free energy that is equivalent to a cost function for canonical neural networks. Thereby, BSyMP optimizes model structures via variational Bayesian inference of the hyper- or

connectivity parameters of the generative model. Mathematical analyses show that BSyMP is guaranteed to prune parameters unnecessary for propagating hidden state information, resulting in efficient model reduction.

We compared the performance of BSyMP with that of the previously proposed Bayesian model reduction (BMR). BMR reduces the number of parameters in a learned full model by minimizing the variational free energy in a post hoc manner (Friston et al., 2019). Owing to online and continuous (i.e., gradual) pruning, BSyMP outperformed BMR in numerical simulations of causal inference and rule learning. BSyMP could deliver successful results even when the full model and BMR failed. Finally, we discuss possible neuronal substrates that realize synaptic pruning.

2. Results

2.1. Derivation of BSyMP from synaptic pruning

Herein, we consider a recurrent network of N_x neurons that receives N_o-dimensional binary observations (sensory inputs). The activity of the neurons follows that of the previously defined canonical neural network model (Isomura et al., 2022). However, we extend the model by adding connectivity parameters (C_W and C_K) that express the presence (ON) or absence (OFF) of synaptic connections, which is formulated as follows:

$$\dot{x}(\tau) \propto -\mathrm{sig}^{-1}\big(x(\tau)\big) + (C_W \odot W)o(\tau) + (C_K \odot K)x(\tau - \Delta t) + h, \tag{1}$$

where $x(\tau) \coloneqq \left(x_1(\tau), \dots, x_{N_x}(\tau)\right)^{\mathrm{T}}$ denotes the firing rates of the neurons; $o(\tau) \coloneqq \left(o_1(\tau), \dots, o_{N_o}(\tau)\right)^{\mathrm{T}}$ denotes binary observations; $\mathrm{sig}^{-1}(x(\tau))$ denotes the inverse function of the sigmoid function of $x(\tau)$ that models the leakage current; $W \in \mathbb{R}^{N_x \times N_o}$ and $K \in \mathbb{R}^{N_x \times N_x}$ are the feedforward and recurrent synaptic strength matrices, respectively; the symbol $\odot$ denotes the element-wise (Hadamard) product; $\Delta t > 0$ is the delay in the recurrent input; and h represents the adaptive firing thresholds that depend on the synaptic strengths. The additional parameter matrices $C_W \in [0,1]^{N_x \times N_o}$ and $C_K \in [0,1]^{N_x \times N_x}$ express the probability of the presence of synaptic connections, providing a major difference from the previous work. As shown in Eq. (1), when a certain element in C_W takes the value of 0, the signal from the corresponding synapse does not affect the neural activity. With this property, switching the values of C_W and C_K models the synaptic pruning process.

A previous work established a scheme for reverse-engineering cost function L for canonical neural networks from neural activity equations and showed the formal equivalence between L and the variational free energy F under a partially observable Markov decision process (POMDP) model (Isomura et al., 2022). Following the scheme, we reconstruct the cost function L from Eq. (1) (Fig. 1, left) and identify the variational free energy F that formally corresponds to L (Fig. 1, right).

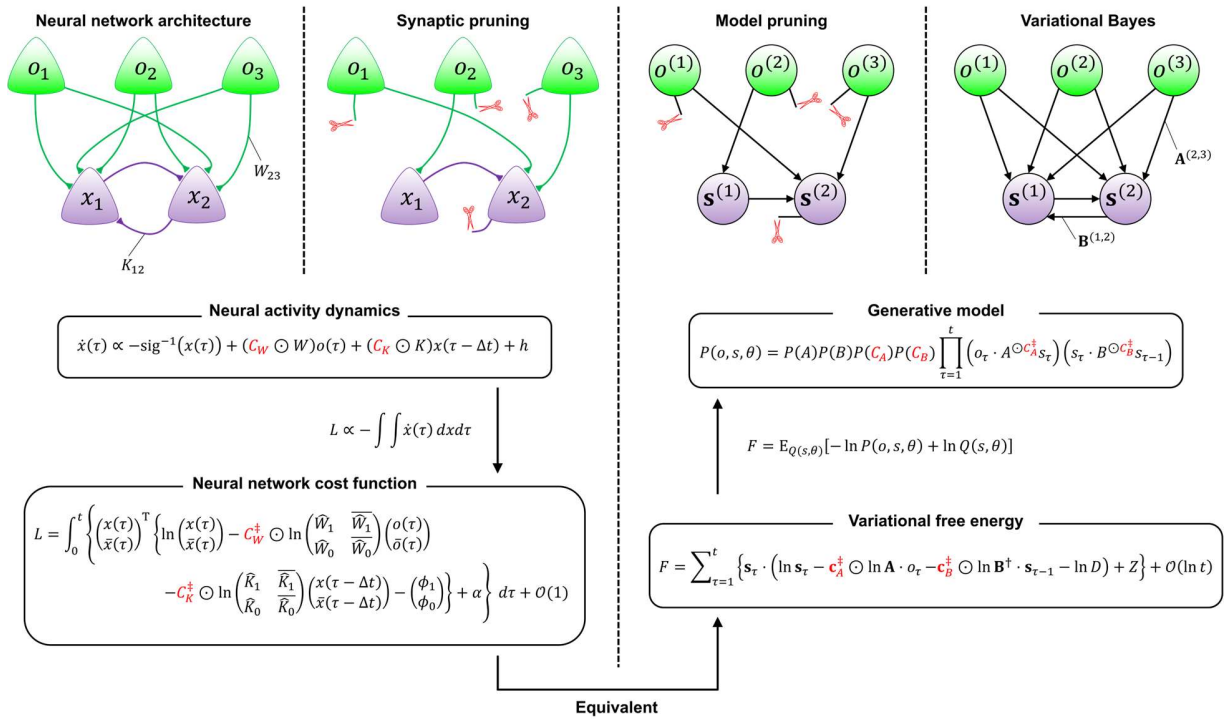

Fig. 1. Equivalence between synaptic pruning in neural networks and Bayesian pruning of connections in a generative model. Top left: A recurrent neural network receives sensory inputs. The occurrence of synaptic pruning is mediated by the connectivity parameters C_W and C_K. The neural activity dynamics are given by Eq. (1). Bottom left: The neural network cost function L is derived by integrating Eq. (1) with respect to x and time. The reconstructed L involves a free parameter α (i.e., integration constant). Bottom right: The variational free energy F equivalent to L is identified by extending the previously established equivalence between F and L. All variables are introduced in the main text and Table 1. Top right: The generative model $P(o, s, \theta)$ is reverse engineered from F. This enables the identification of the normalization factor Z that is involved in F. The details of the notations and derivations are provided in Section 4.1.

Subsequently, a biologically plausible generative model under which the neural network in Eq. (1) operates can be reverse engineered, as follows:

$$P(o_{1:t}, s_{1:t}, \theta) = P(A)P(B)P(C_A)P(C_B) \prod_{\tau=1}^{t} P(o_\tau | s_\tau, A, C_A) P(s_\tau | s_{\tau-1}, B, C_B)\,. \qquad (2)$$

Here and throughout, $\theta \coloneqq \{A, B, C_A, C_B\}$ denotes a set of model parameters.

Equation (2) presents a POMDP model that generates N_o observations $o_\tau \coloneqq \left(o_\tau^{(1)}, \dots, o_\tau^{(N_o)}\right)^{\mathrm{T}}$ from N_s hidden states $s_\tau \coloneqq \left(s_\tau^{(1)}, \dots, s_\tau^{(N_s)}\right)^{\mathrm{T}}$ through the likelihood mapping A and its connectivity C_A in the form of a categorical distribution (also called a generalized Bernoulli distribution), $P(o_\tau | s_\tau, A, C_A) = \mathrm{Cat}\left(A^{\odot C_A^{\ddagger}} s_\tau\right)$ (Bishop, 2006). Each state and observation has a binary value. The hidden states s_τ depend on the previous states $s_{\tau-1}$, following the state transition matrix B and its connectivity C_B, $P(s_\tau | s_{\tau-1}, B, C_B) = \mathrm{Cat}\left(B^{\odot C_B^{\ddagger}} s_{\tau-1}\right)$. The prior belief about initial states is described by the block vector D, $P(s_1) = \mathrm{Cat}(D)$, and throughout this work, D is a constant for simplicity. We consider the case in which A and B are factorized into products of 2×2 matrices $A^{(i,j)}$ and $B^{(i,j)}$. The new parameters $C_A \in \{1,0\}^{N_o \times N_s}$ and $C_B \in \{1,0\}^{N_s \times N_s}$ express the connectivity structure of graphical models. The double dagger ‡ on C denotes the block matrix of C, $C^{\ddagger} \coloneqq \begin{pmatrix} C & C \\ C & C \end{pmatrix}$, and the superscript $\odot$ on A and B indicates the element-wise power. Owing to the factorial property of A, $C_{A_{ij}} = 0$ means the reduction (i.e., elimination or pruning) of $A^{(i,j)}$ from the generative model, and the same applies to B (see Section 4.1 for further details). If $C_{A_{ij}} = 1$ holds for all connectivities, Eq. (2) is reduced to a conventional full model without model reduction. The prior distributions of A, B and C_A, C_B follow the Dirichlet distributions $P(A) = \mathrm{Dir}(a)$, $P(B) = \mathrm{Dir}(b)$ and categorical distributions $P(C_A) = \mathrm{Cat}(c_A)$, $P(C_B) = \mathrm{Cat}(c_B)$, respectively (Bishop, 2006). A simplified notation for the factorial structure was adopted here; please refer to previous works (Isomura et al., 2022; Isomura & Friston, 2020) for the detailed definition of a generative model that explicitly expresses the factorial structure.

Variational Bayesian inference is the process of optimizing an approximate posterior belief $\mathrm{Q}(s_{1:t}, \theta)$ based on a sequence of observations by minimizing the variational free energy as a tractable proxy for surprise minimization. Based on the mean-field approximation, the posterior distribution is expressed as follows:

$$Q(s_{1:t}, \theta) = Q(A)Q(B)Q(C_A)Q(C_B)\prod_{\tau=1}^{t} Q(s_\tau). \tag{3}$$

The marginal posteriors about A, B, C_A, C_B, s_τ follow the Dirichlet distributions $Q(A) = \mathrm{Dir}(\mathbf{a})$, $Q(B) = \mathrm{Dir}(\mathbf{b})$ and categorical distributions $Q(C_A) = \mathrm{Cat}(\mathbf{c}_A)$, $Q(C_B) = \mathrm{Cat}(\mathbf{c}_B)$, $Q(s_\tau) = \mathrm{Cat}(\mathbf{s}_\tau)$, respectively. Here and throughout, the bold case variables (i.e., $\mathbf{c}_A$, $\mathbf{c}_B$, $\mathbf{s}_\tau$, $\mathbf{A}$, $\mathbf{B}$) indicate the posterior expectations, while $\mathbf{a}$ and $\mathbf{b}$ denote the concentration parameters (Table 1). The analytical expression for the variational free energy under the abovementioned generative model and posterior beliefs is given as follows:

$$F\big(o_{1:t}, Q(s_{1:t}, \theta)\big) = \sum_{\tau=1}^{t}\big\{\mathbf{s}_\tau \cdot \big(\ln \mathbf{s}_\tau - \mathbf{c}_A^{\ddagger} \odot \ln \mathbf{A} \cdot o_\tau - \mathbf{c}_B^{\ddagger} \odot \ln \mathbf{B}\, \mathbf{s}_{\tau-1}\big) + Z_{C_A} + Z_{C_B}\big\} + \mathcal{D}_{\mathrm{KL}}[Q(\theta)||P(\theta)], \tag{4}$$

where Z_{C_A} and Z_{C_B} denote the normalization factors associated with $\mathbf{c}_A$ and $\mathbf{c}_B$, respectively; $\mathcal{D}_{\mathrm{KL}}[Q(\theta)||P(\theta)]$ denotes the complexity scored by the Kullback–Leibler divergence between $Q(\theta)$ and $P(\theta)$. Minimization of the variational free energy furnishes posterior beliefs. By solving the fixed point of the implicit gradient descent, $\partial F/\partial \mathbf{c}_A = O$, the update rule for $\mathbf{c}_{A_{ij}}$ is formulated as follows:

$$\mathbf{c}_{A_{ij}} = \mathrm{sig}\left(\mathrm{sig}^{-1}\left(c_{A_{ij}}\right) + \sum_{\tau=1}^{t}\left(\ln \mathbf{A}^{(i,j)} \cdot \left(o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}\right) + \ln 2\right)\right). \tag{5}$$

A similar equation holds for $\mathbf{c}_B$. The derivation details and explicit forms of the other posterior beliefs (i.e., $\mathbf{s}_\tau$, $\mathbf{a}$, and $\mathbf{b}$) are provided in Section 4.2.

Equation (5) can be viewed as the dynamics of synaptic pruning in neural network formation. The probability of pruning is controlled by the similarities between the current synaptic strength ($\ln \mathbf{A}^{(i,j)}$) and its change, as calculated by the Hebbian rule ($o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}$) (Hebb, 1949). When these values are considerably different, their product becomes a large negative value over time, leading to the pruning of the corresponding connection. In addition, mathematical analyses show that an element of the connectivity parameter $\mathbf{c}_{A_{ij}}$ decreases and converges to 0 when the corresponding element of observation $o_\tau^{(i)}$ is randomly determined and independent of all elements of s_τ, and the likelihood mapping posterior $\mathbf{A}^{(i,j)}$ and the mean state posterior $\mathrm{E}_{P(o_\tau)}\left[\mathbf{s}_\tau^{(j)}\right]$ are sufficiently flat (refer to Section 4.3 for further details). This means that BSyMP is guaranteed to prune unnecessary connections in the generative model as long as the prior of the likelihood mapping and the mean state posterior are unbiased.

In short, the connectivity parameters C_A and C_B provide an implementation of the Bayesian model selection that enables switching between the presence (ON) and absence (OFF) of each element of matrices A and B. As shown in Fig. 1, this model reduction is formally equivalent to synaptic pruning in neural network formation, which is mediated by C_W and C_K in Eq. (1). Table 1 summarizes the correspondence between variables in neural networks and those in variational Bayes.

Table 1. Glossary of variables

	Variables in neural networks	Variables in Bayesian inference	
Sensory inputs	$o(\tau) \coloneqq \left(o_1(\tau), \dots, o_{N_o}(\tau)\right)^{\mathrm{T}} \in \{0,1\}^{2\times N_o}$	$o_\tau \coloneqq \left(o_\tau^{(1)}, \dots, o_\tau^{(N_o)}\right)^{\mathrm{T}} \in \{0,1\}^{2\times N_o}$	Observations
Neural activity	$x(\tau) \coloneqq \left(x_1(\tau), \dots, x_{N_x}(\tau)\right)^{\mathrm{T}} \in [0,1]^{N_x}$	$\mathbf{s}_\tau \coloneqq \left(\mathbf{s}_\tau^{(1)}, \dots, \mathbf{s}_\tau^{(N_s)}\right)^{T} \in [0,1]^{2\times N_s}$	State posterior
Feedforward synaptic strengths	$W \in \mathbb{R}^{N_x\times N_o}$	$\mathbf{a} \in (\mathbb{R}^{2\times 2})^{N_o\times N_s}$	Likelihood mapping posterior
Recurrent synaptic strengths	$K \in \mathbb{R}^{N_x\times N_x}$	$\mathbf{b} \in (\mathbb{R}^{2\times 2})^{N_s\times N_s}$	State transition posterior
Survival probability of feedforward synapses	$C_W \in [0,1]^{N_x\times N_o}$	$\mathbf{c}_A \in [0,1]^{N_o\times N_s}$	Connectivity posterior for A
Survival probability of recurrent synapses	$C_K \in [0,1]^{N_x\times N_x}$	$\mathbf{c}_B \in [0,1]^{N_s\times N_s}$	Connectivity posterior for B
Adaptive firing thresholds	$h \in \mathbb{R}^{N_x}$	$D \in [0,1]^{2\times N_s}$	State prior (constant)

Note: The variables in bold for the Dirichlet distributions (i.e., $\mathbf{a}$ and $\mathbf{b}$) indicate the concentration parameters, and those for categorical distributions (i.e., $\mathbf{c}_A$, $\mathbf{c}_B$, $\mathbf{s}_\tau$) indicate the posterior expectations of the corresponding variables in italics.

2.2. Numerical simulation 1: Blind source separation

The performance of BSyMP was evaluated using a blind source separation (BSS) task, in which the performance was expected to improve with appropriate model reduction. This BSS task was analogous to a visual shape recognition task in which sources 1 and 2 displayed a low-contrast circle and rectangle, respectively, on a 32 × 32 pixel screen with a certain fidelity following a likelihood mapping (Fig. 2a). Each pixel on the display took either 0 or 1 stochastically, where some parts of the shapes overlapped. Thus, the agent needed to disentangle the input signals to perceive the hidden states. The field of view of the agent comprised 1024-dimensional observations, and the states of the two sources (ON/OFF) changed randomly. This setup was effectively expressed as a generative model without matrix *B*; thus, the agents optimized only $\mathbf{s}_\tau$, $\mathbf{a}$, and $\mathbf{c}_A$.

Three agent types (full model, BMR, and BSyMP) were trained with identical input sequences. The agents optimized $\mathbf{s}_\tau$ at every time point; conversely, they optimized $\mathbf{a}$ and $\mathbf{c}_A$ at every 10 time points, which we refer to as a training session. One simulation comprised 1000 training sessions and a subsequent test session. In every test session, the agents inferred $\mathbf{s}_\tau$ for 1000 time-point-length test data that were newly generated from the same likelihood mapping used in the training sessions. We prepared three likelihood mappings, generating inputs that differed in the size of the circle and rectangle and the overlap between them. The simulation was repeated 100 times for each mapping (i.e., 300 simulations in total).

The agents with BSyMP successfully self-organized to infer the hidden sources as a result of successful learning of posterior beliefs about the likelihood mapping (Figs. 2b and 2c). The result

was underwritten by an appropriate model reduction, as depicted by the efficient reduction of the redundant connections (Fig. 2d). When the source and observation were independent of one another, BSyMP reliably reduced the connectivity between them to 0 (Fig. 2d; blue lines). This result was consistent with the mathematical analysis described in the previous subsection and Section 4.3. Figure 2d further shows that the connectivity was retained if the source and the observation were not independent (red lines), indicating that no over-pruning occurred in this task. Owing to these desirable properties of the connectivity parameters, the variational free energy of the BSyMP agent was smaller than that of the full model agent (Fig. 2e), and the inference and learning performance of the BSyMP agent in this BSS task significantly exceeded those of the full model agent (Figs. 2f-h). For ease of discussion, we labeled a simulation as successful if the state inference error in the test session was less than 20%, and as a failure otherwise (Fig. 2f; vertical and horizontal dashed lines). Under this criterion, the success rate of the full model agent was 45.0% (135/300), whereas that of the BSyMP agent was 92.3% (277/300). These results demonstrate that BSyMP achieved appropriate model reduction and significant effects on the performance of the BSS task.

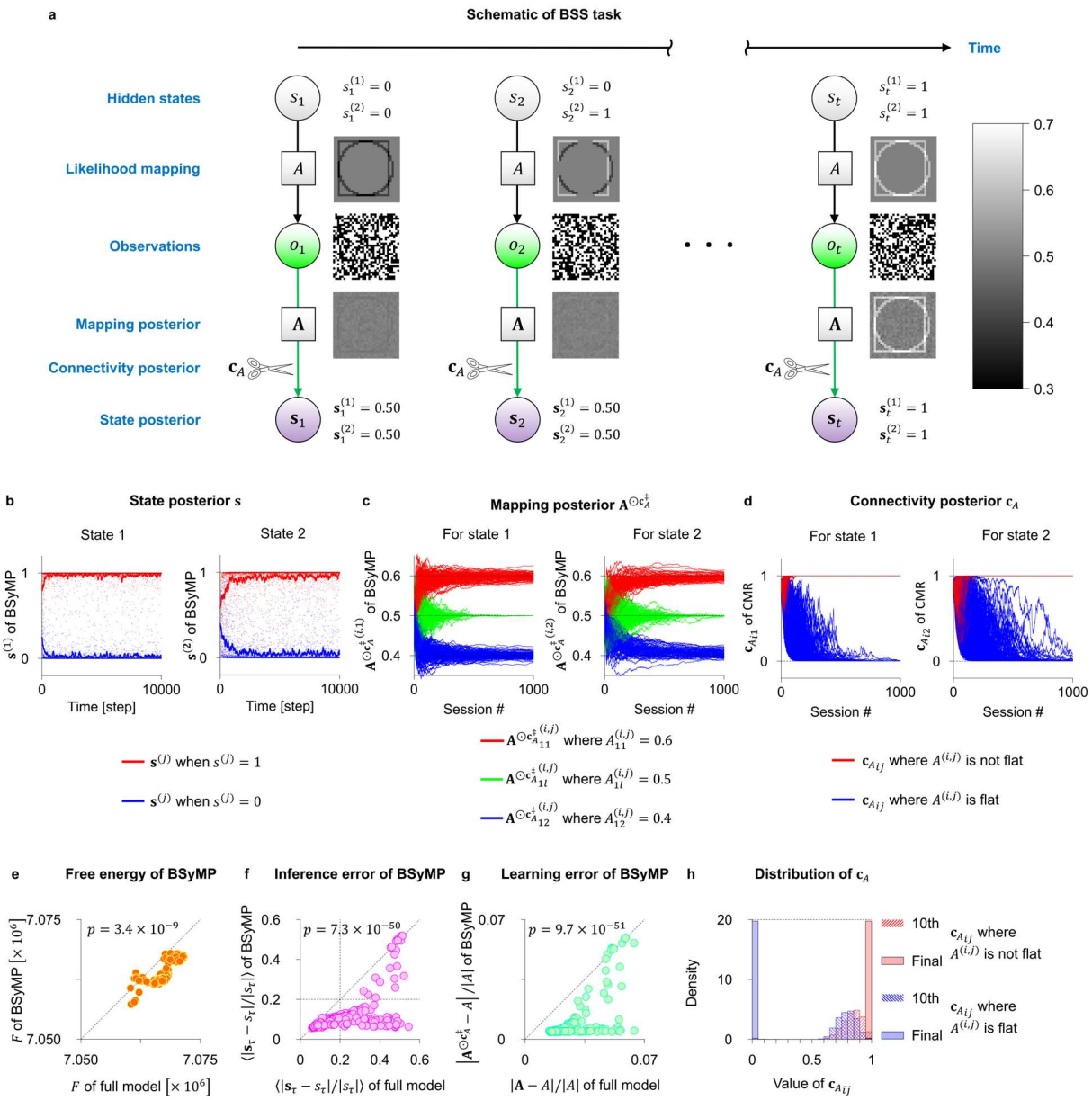

Fig. 2. Performance of BSyMP agent on BSS task. (**a**) A schematic of the task, illustrating a generative process with one likelihood mapping out of three. Two hidden sources s_τ randomly take the value of 1 (ON) or 0 (OFF). Binary observations o_τ are stochastically generated from s_τ through the likelihood mapping A, $P(o_\tau|s_\tau, A) = \mathrm{Cat}(As_\tau)$. The three vertically aligned images depict As_τ, o_τ, and $\mathbf{A}s_\tau$ in order from top to bottom. (**b**) The posterior beliefs about two states during a single simulation. The red and blue dots indicate $\mathbf{s}_\tau^{(j)}$ when the true hidden state $s_\tau^{(j)}$ is 1 or 0, respectively. The red and blue trajectories indicate the moving averages of the red and blue dots with a window length of 100 time points. (**c**) The posterior beliefs about the likelihood mapping during a single simulation. The elements in the posterior beliefs are grouped and colored according to the value of the corresponding elements in the true mapping, that is, 0.6 (red), 0.4 (blue), or 0.5 (green). (**d**) The posterior beliefs about the connectivity during a single

simulation. The elements in the posterior beliefs are grouped and colored according to whether the corresponding submatrix in the true mapping is flat (blue; to be reduced) or not (red; to be retained). (**e**) A comparison of the free energy between BSyMP and the full model. In (**e**), (**f**), and (**g**), each point represents data obtained in a single simulation. Throughout this work, the *p* value was obtained by the two-sided Wilcoxon signed-rank test for the pair. (**f**) A comparison of the state inference error between BSyMP and the full model. The error was evaluated by the mean of the Euclidean norm ratio $|\mathbf{s}_\tau - s_\tau|/|s_\tau|$ in the test session. (**g**) A comparison of the parameter learning error between BSyMP and the full model. The error was evaluated by the Frobenius norm ratio $\left|\mathbf{A}^{\odot \mathbf{c}_A^{\ddagger}} - A\right|/|A|$ for BSyMP and $|\mathbf{A} - A|/|A|$ for the full model at the end of the training sessions. (**h**) The distribution of the posterior beliefs about the connectivity. All the data at the 10th and final training sessions of 300 simulations were counted in 0.05 intervals. Please refer to Section 4.4 for further details.

Subsequently, BSyMP was compared with a standard Bayesian model reduction method, referred to as BMR. BMR compares the free energy of the learned full model with that of the reduced version of the model at a small computational cost. Thus, BMR is a post hoc model reduction method that is based on an already acquired model.

Although the BMR agent decreased the variational free energy more than the BSyMP agent, the performance of the BMR agent was close to that of the full model and did not exceed that of the BSyMP agent (Figs. 3a-c). The success rate of BMR was 46.7% (140/300), which was only 1.7% higher than that of the full model, whereas the rate of BSyMP was twice as high as that of the full model. The inconsistency between the free energy and the performance was arguably because BMR decreased the complexity term $\mathcal{D}_{\mathrm{KL}}[Q(A)||P(A)]$ in the free energy by heuristically altering priors (Friston et al., 2017), but this free energy reduction had little contribution to the performance in this setting. By contrast, BSyMP could prune synapses by evaluating the likelihood through its construction. Consistent with the performance, the accuracy of model reduction by BMR is worse than BSyMP (Fig. 3d).

We attribute the underperformance of BMR to the post hoc manner of its model reduction. This hypothesis was supported by the comparison of parameter trajectories. Figure 3e illustrates the typical outcomes of three approaches in a simulation. Owing to stochasticity in the generative processes and priors, the agent with the full model often (in 55% of the simulations) failed to learn the likelihood mapping (Fig. 3e, left panel). BMR could reduce the redundant elements but could not change the learning process because it operates in a post hoc manner (Fig. 3e, center panel). By contrast, as BSyMP adopts online reduction during learning, it could learn the likelihood mapping based on the state inference that was improved by the model reduction, resulting in its reliable success (Fig. 3e, right panel). In summary, we demonstrated that BSyMP successfully eliminated connections independent of the sources and improved the performance of the BSS task more significantly than related methods.

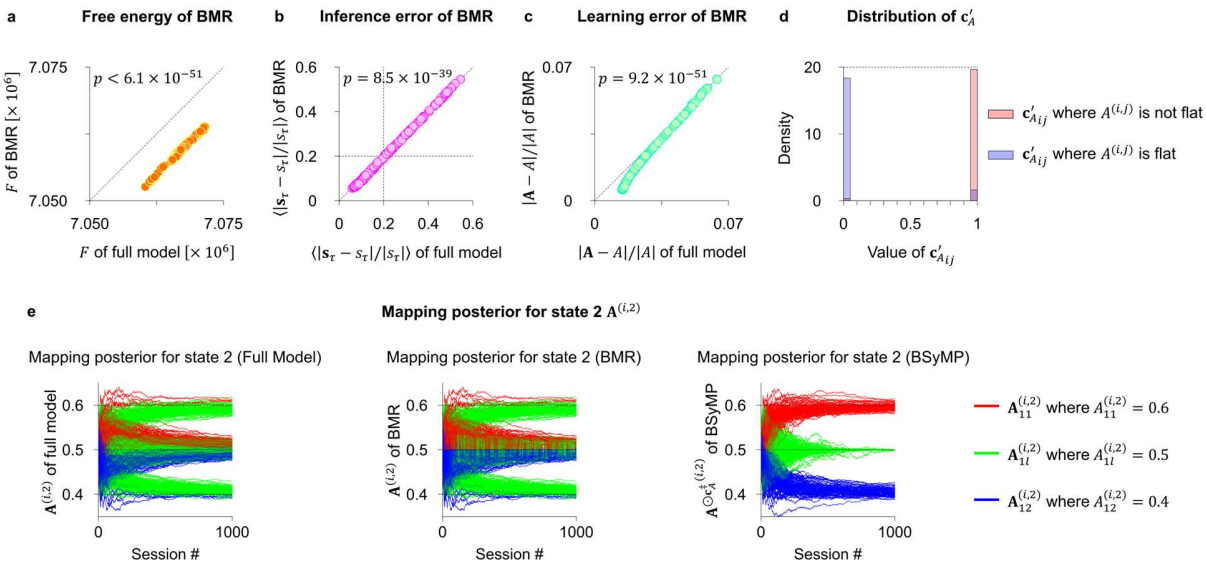

Fig. 3. Comparison of BSyMP with BMR on BSS task. (**a**) A comparison of the free energy between BMR and the full model. (**b**) A comparison of the state inference error between BMR and the full model. (**c**) A comparison of the parameter learning error between BMR and the full model. (**d**) The distribution of the BMR's counterpart, $\mathbf{c}'_A$, to the connectivity parameter in BSyMP, $\mathbf{c}_A$, at the end of the training sessions. Here, $\mathbf{c}'_A$ has the same number of elements as $\mathbf{c}_A$, and each element in $\mathbf{c}'_A$ takes a binary value, indicating whether the corresponding submatrix of the mapping posterior is reduced (0) or not (1). (**e**) The posterior beliefs about the likelihood mapping regarding state 2 during a single simulation obtained by three methods. Note that the time course of the posterior beliefs of a BMR agent was calculated by performing BMR at every session.

2.3. Numerical simulation 2: Rule learning

The performance of BSyMP was further evaluated using a rule-learning task in which the agents learned the transition rule behind the visual inputs in addition to BSS. The agents received observations comprising 32 × 96 pixels, which is analogous to three screens. Four sources on each screen stochastically displayed a circle, rectangle, x, and plus, respectively (Fig. 4a). In other words, the agents received 3072 binary observations that depend on the states of the 12 sources (ON/OFF). The states changed according to the transition rule described by matrix B. The agents had to learn the rule to predict the next state of the sources.

We trained the three agent types in the same manner as in the BSS task. The agents optimized $\mathbf{s}_\tau$ at every time point and optimized $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}_A$, and $\mathbf{c}_B$ at every 100 time points during the training sessions. In this case, we provided the agents with accurate and strong prior beliefs about the likelihood mapping *A* to focus on the learning in matrix *B*, thereby ensuring accurate state inference after obtaining the observations. One simulation comprised 100 training sessions and a subsequent 1000 time-point-length test session in which the prediction error for the upcoming states was evaluated. We prepared three transition rules, and the simulation was repeated 100 times for each rule (i.e., 300 simulations in total).

The BSyMP agents scored a smaller variational free energy than the full model agent and successfully predicted the upcoming hidden states by learning the posterior beliefs about the transition rules (Figs. 4b-d). The success was underwritten by the appropriate pruning of redundant recurrent connections between the state posteriors (Fig. 4e). The prediction and learning performance of the BSyMP agent in the rule-learning task significantly exceeded those

295 of the full model agent (Figs. 4c and d) owing to the desirable properties of the connectivity
296 parameters. For ease of discussion, we labeled a simulation as successful if the state prediction
297 error in the test session was less than 10%, and as a failure otherwise (Fig. 4c). Under this criterion,
298 the success rate of the full model agent was 9.0% (27/300), whereas that of the BSyMP agent
299 was 82.3% (247/300). These results show that BSyMP achieved appropriate model reduction and
300 significant effects on the performance of the rule-learning task.

301 Subsequently, BSyMP was compared with BMR. As in the BSS task, although BMR decreased
302 the variational free energy more than BSyMP owing to the decrease in complexity (Fig. 4f), and
303 the performance of the BMR agent exceeded that of the full model agent (Figs. 4g and h), the
304 success rate of the BMR agent, 58.7% (176/300), was less than that of BSyMP. This performance
305 difference could not be attributed to the difference in the manner of model reduction (i.e., online
306 or post hoc) because the model reduction had no significant effect on learning matrix B in this
307 rule-learning task, owing to accurate state inference, unlike in the BSS task. Nevertheless, pruning
308 unnecessary connections is more challenging in this task than in the BSS task because some
309 sources can be correlated with others despite the absence of direct connections. For example, in
310 the rule shown in Fig. 4a, the state of the circle on screen 1 at time τ is positively correlated with
311 the state of the x on screen 1 at time $\tau + 1$. The blue bars in Figs. 4e and 4i show that BSyMP was
312 better at pruning these "false positive connections" than BMR, which can explain the
313 outperformance of BSyMP in rule-learning tasks. In summary, we demonstrated that BSyMP
314 successfully eliminated unnecessary connections between the sources and improved the rule-
315 learning performance to a greater extent than the previously developed method.

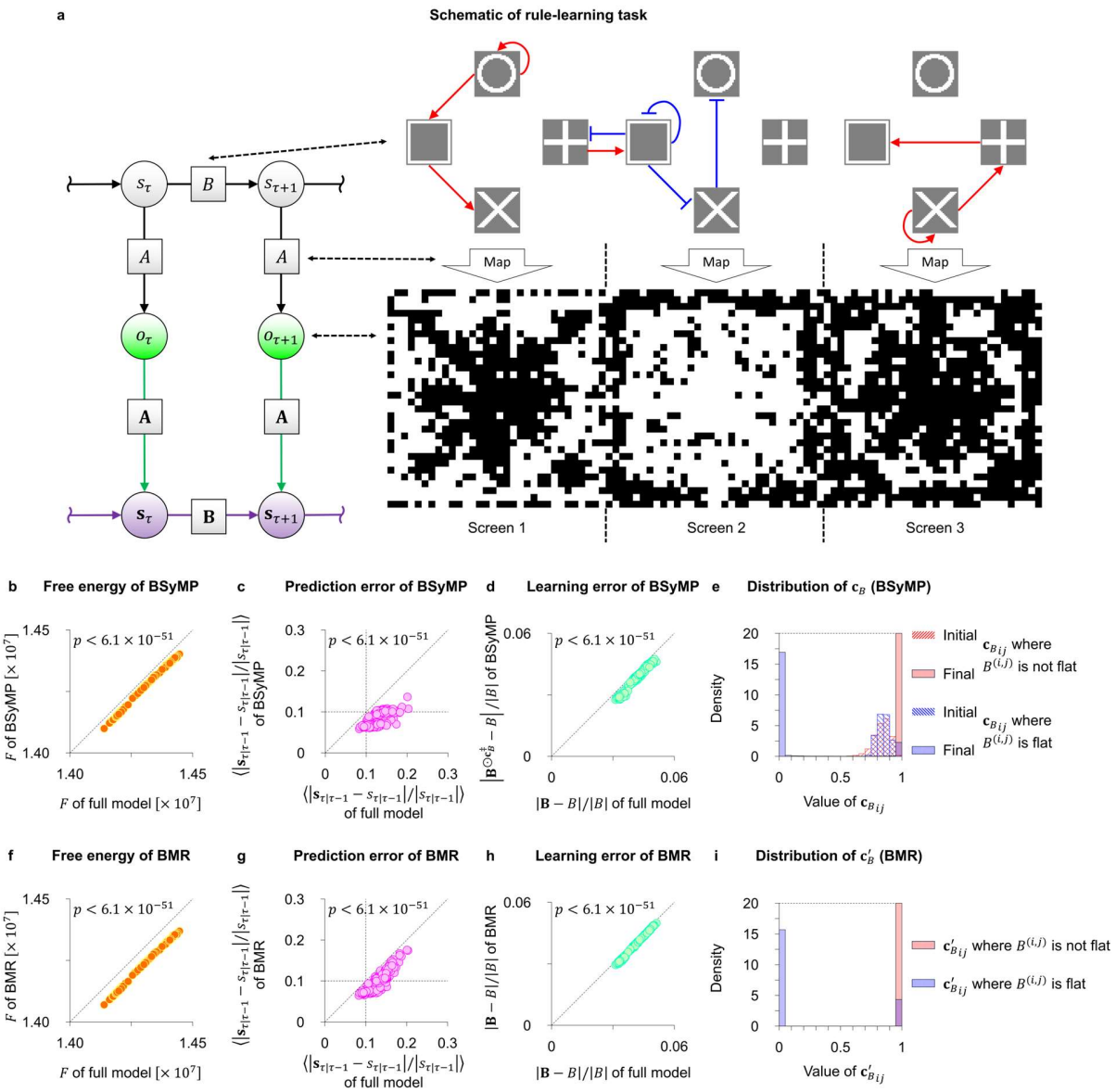

Fig. 4. Performance of BSyMP agent on rule-learning task. (**a**) A schematic of the task. The factor graph on the left side illustrates a generative process. Images on the right side depict one of three transition rules and one example of observations at a moment. The states of the sources stochastically change depending on the states at the previous time point. The red arrows from one source $s^{(j)}$ to another source $s^{(i)}$ indicate the existence of the positive connection from $s_{\tau}^{(i)}$ to $s_{\tau-1}^{(j)}$, whereas the blue T-shaped lines indicate the negative connection. Each source displays a shape in one of three 32 × 32-pixel screens and has no effect on the other 32 × 64 pixels. The agent optimizes the posterior beliefs about s_{τ}, A, B, C_A, and C_B. (**b**) A comparison of the free energy between BSyMP and the full model. (**c**) A comparison of the state prediction error

between BSyMP and the full model. The error was evaluated by the mean of the normalized Euclidean distance between the actual state and agent's prediction $\left|\mathbf{s}_{\tau|\tau-1} - s_{\tau|\tau-1}\right| / \left|s_{\tau|\tau-1}\right|$ in the test session. Please refer to Section 4.5 for the definitions of $\mathbf{s}_{\tau|\tau-1}$ and $s_{\tau|\tau-1}$. (**d**) A comparison of the tule learning error between BSyMP and the full model. The error was evaluated by the Frobenius norm ratio $\left|\mathbf{B}^{\odot \mathbf{c}_B^{\ddagger}} - B\right| / |B|$ for BSyMP and $|\mathbf{B} - B| / |B|$ for the full model at the end of the training sessions. (**e**) The distribution of the posterior beliefs about the connectivity parameter. (**f**) A comparison of the free energy between BMR and the full model. (**g**) A comparison of the state prediction error between BMR and the full model. (**h**) A comparison of the rule-learning error between BMR and the full model. (**i**) The distribution of the BMR's counterpart, $\mathbf{c}_B'$, to the connectivity parameter in BSyMP, $\mathbf{c}_B$, at the end of the training sessions.

3. Discussion

We derived the biologically plausible synaptic pruning rule, BSyMP (Eq. (5)), from the cost function minimization based on the equivalence between the dynamics of the canonical neural networks and variational Bayesian inference. Equation (5) states that the probability of pruning depends on the similarities between the current synaptic strength ($\ln \mathbf{A}^{(i,j)}$) and its change calculated by the Hebbian rule ($o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}$) during the latest session. The Hebbian plasticity in the brain can be achieved by long-term potentiation (LTP) and depression (LTD), and its molecular mechanisms are well studied (Malenka & Bear, 2004). The Hebbian term in Eq. (5) can be encoded in the molecules in the postsynaptic neuron (Nicoll, 2017; Scannevin & Huganir, 2000), while the values of synaptic strengths lie in the density of AMPA receptors in the synaptic plasma membrane. Thus, it is possible that the molecular mechanisms in the postsynaptic neuron compute the correlation between $\ln \mathbf{A}^{(i,j)}$ and $o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}$ to determine which synapses are pruned. Experimentally observed similarities between LTP, LTD, and synaptic pruning—in their physiological properties and molecular mechanisms—further support the biological plausibility

of our model, highlighting its potential to guide the experimental targets to shed light on the mechanisms underlying synaptic connectivity control (Helias et al., 2008; Le Bé & Markram, 2006; Lüscher et al., 2000).

One can contrast BSyMP with earlier pruning methods. Pruning connections with small weight is the most basic and widely used rule to prune artificial neural networks (Hagiwara, 1994; Han et al., 2015, 2016; Hoefler et al., 2021). Our analysis shows that BMR constitutes a family of weight-base pruning rules (please refer to Appendix B for details). In addition to weight-based methods, pruning methods based on the activity of neurons exist (Scholl et al., 2021; Sietsma & Dow, 1991; Sun et al., 2015). BSyMP utilizes temporal consistency to organize the network structure by comparing synaptic weight (i.e., the history of neural activity in the past) with neural activity in the latest session. This allows the neural network to spontaneously detect changes in the statistical features of neural activity. This would be useful for reorganizing the network structure flexibly to adapt to environmental changes, enabling BSyMP to outstand in the activity-based pruning rules. This might be a mechanistic explanation of the dynamics of the network structure in the brain (Ziv & Brenner, 2018). Moreover, most pruning methods, including BMR, prune synapses in a binary manner, and their processes involve some heuristics. BSyMP, by contrast, computes the probability of the presence of synapses and prunes synapses in a continuous (i.e., differentiable) manner, thereby enabling stochastic gradient descent. This rule is formally derived from variational free energy minimization, achieving more accurate pruning than the binary methods. This is a great advantage over heuristic methods. Appendix A delves into this contrast analytically.

Biological organisms utilize structural learning to identify the structure of their environment (Braun et al., 2010; Lynn & Bassett, 2020; Tenenbaum et al., 2011). Structure learning facilitates generalization by transferring the structures obtained in the past to newly encountered environments (Collins & Frank, 2013). Our work suggests that synaptic pruning plays a vital role in structure learning in the brain. This would be useful for designing next-generation artificial intelligence with high data efficiency and robustness and for understanding the mechanisms underlying the generalizability of the brain and its attenuation caused by psychiatric disorders (Kasai et al., 2021).

Conclusion

In this work, we derived a reliable online Bayesian model selection scheme, BSyMP, that can be implemented by canonical neural networks equipped with synaptic pruning. Mathematical analyses revealed that BSyMP is guaranteed to prune unnecessary connections in the generative model. Numerical simulations of BSS and rule-learning tasks demonstrated the advantage of BSyMP over the conventional BMR scheme. These findings suggest that synaptic pruning can be a possible neuronal substrate underlying structure learning and generalizability in the brain.

4. Methods

4.1. Derivation of generative model from neural network

The generative model of BSyMP originates from the canonical neural network described in Eq. (1). We derive the cost function L for canonical neural networks from Eq. (1) and identify the variational free energy F under a certain POMDP generative model that formally corresponds to L. For this purpose, we introduce excitatory and inhibitory synapses for feedforward circuits W_1 and W_0 that satisfy $W = W_1 - W_0$ and those for recurrent circuits K_1 and K_0 that satisfy $K = K_1 - K_0$. The corresponding adaptive firing thresholds are denoted as h_1 and h_0, which are functions of $W_1, K_1, W_0, K_0, C_W, C_K$ and satisfy $h = h_1 - h_0$. This treatment separates the positive (variables with subscript 1) and negative (variables with subscript 0) effects on neural activity. As the gradient descent on L yields Eq. (1), L can be derived by integrating Eq. (1) with respect to neural activity x and time τ, as follows:

$$L = \int_0^t \left\{ \begin{pmatrix} x(\tau) \\ \bar{x}(\tau) \end{pmatrix}^T \left\{ \ln \begin{pmatrix} x(\tau) \\ \bar{x}(\tau) \end{pmatrix} - \left\{ C_W^{\ddagger} \odot \ln \begin{pmatrix} \widehat{W}_1 & \overline{\widehat{W}_1} \\ \widehat{W}_0 & \overline{\widehat{W}_0} \end{pmatrix} \right\} \begin{pmatrix} o(\tau) \\ \bar{o}(\tau) \end{pmatrix} \right.\right.$$

$$\left.\left. - \left\{ C_K^{\ddagger} \odot \ln \begin{pmatrix} \widehat{K}_1 & \overline{\widehat{K}_1} \\ \widehat{K}_0 & \overline{\widehat{K}_0} \end{pmatrix} \right\} \begin{pmatrix} x(\tau - \Delta\tau) \\ \bar{x}(\tau - \Delta\tau) \end{pmatrix} - \begin{pmatrix} \phi_1 \\ \phi_0 \end{pmatrix} \right\} + \alpha \right\} d\tau + \mathcal{O}(1), \tag{6}$$

where $\hat{X} \coloneqq \mathrm{sig}(X)$ denotes the sigmoid function, $\bar{X}$ is defined as $\bar{X} \coloneqq 1 - X$ for any variable X, and α denotes the integration constant that is independent of x, W, and K but can depend on C_W and C_K.

The previous work established the equivalence between the cost function of neural networks without connectivity parameters C_W and C_K, denoted as L_{woC}, and the variational free energy under the POMDP model F_{woC} (Isomura et al., 2022). When all elements of C_W and C_K are equal

to 1, Eq. (6) is reduced to L_{woC} and the corresponding variational free energy is expressed as follows:

$$F_{woC} = \sum_{\tau=1}^{t} \{\mathbf{s}_\tau \cdot (\ln \mathbf{s}_\tau - \ln \mathbf{A} \cdot o_\tau - \ln \mathbf{B}^\dagger \cdot \mathbf{s}_{\tau-1} - \ln D + \ln Z)\} + \mathcal{O}(\ln t), \tag{7}$$

where $\mathbf{B}^\dagger \coloneqq \mathbf{B}^{\mathrm{T}} \mathrm{diag}(D)^{-1}$ denotes the inverse transition matrix and D is the prior belief about the hidden states, in which $\ln \mathbf{B}\, \mathbf{s}_{\tau-1} = \ln \mathbf{B}^\dagger \cdot \mathbf{s}_{\tau-1} + \ln D$ holds. For simplicity, we consider D as a constant. The partition function Z is a constant that is independent of $\mathbf{s}$, $\mathbf{A}$, and $\mathbf{B}$. Although Z was omitted in the previous work because its derivatives with respect to $\mathbf{s}$, $\mathbf{A}$, and $\mathbf{B}$ are 0, it is involved in the generative model as a normalization factor. This Z is a counterpart to α in L. With this equivalence, the variational free energy F that is equivalent to cost function L in Eq. (6) takes the following form:

$$\begin{aligned} F &= \sum_{\tau=1}^{t} \left\{\mathbf{s}_\tau \cdot \left(\ln \mathbf{s}_\tau - \mathbf{c}_A^\ddagger \odot \ln \mathbf{A} \cdot o_\tau - \mathbf{c}_B^\ddagger \odot \ln \mathbf{B}^\dagger \cdot \mathbf{s}_{\tau-1} - \ln D + \ln Z\right)\right\} + \mathcal{O}(\ln t) \\ &= \sum_{\tau=1}^{t} \left\{\mathbf{s}_\tau \cdot \left(\ln \mathbf{s}_\tau - \ln \mathbf{A}^{\odot \mathbf{c}_A^\ddagger} \cdot o_\tau - \ln \mathbf{B}^{\odot \mathbf{c}_B^\ddagger}\, \mathbf{s}_{\tau-1} + \ln Z\right)\right\} + \mathcal{O}(\ln t). \end{aligned} \tag{8}$$

The new parameters $\mathbf{c}_A$ and $\mathbf{c}_B$ are the counterparts to C_W and C_K in Eq. (6), respectively. At this point, the POMDP generative model underlying Eq. (8) can be reverse engineered by adding the roles of $\mathbf{c}_A$ and $\mathbf{c}_B$ to the conventional POMDP, as per Eq. (2) in Section 2:

$$P(o_{1:t}, s_{1:t}, \theta) = P(A)P(B)P(C_A)P(C_B) \prod_{\tau=1}^{t} P(o_\tau | s_\tau, A, C_A) P(s_\tau | s_{\tau-1}, B, C_B)\,.$$

In the following subsection, we derive the explicit form of the free energy of BSyMP and confirm that Z in Eq. (8) is independent of $\mathbf{s}$, $\mathbf{A}$, and $\mathbf{B}$, which is necessary for making this Bayesian formulation equivalent to canonical neural networks with synaptic pruning.

The explicit form of this generative model is expressed as follows:

$$
\begin{aligned}
P(o_{1:t}, s_{1:t}, \theta) &= P(A)P(B)P(C_A)P(C_B)\prod_{\tau=1}^{t} P(o_\tau|s_\tau, A, C_A)P(s_\tau|s_{\tau-1}, B, C_B) \\
&= \prod_{i=1}^{N_o}\prod_{j=1}^{N_s} P\left(A^{(i,j)}\right)P\left(C_{A_{ij}}\right)\cdot\prod_{i=1}^{N_s}\prod_{j=1}^{N_s} P\left(B^{(i,j)}\right)P\left(C_{B_{ij}}\right)\cdot \\
&\cdot\prod_{\tau=1}^{t}\left\{\prod_{i=1}^{N_o}\frac{\prod_{j=1}^{N_s} P\left(o_\tau^{(i)}\middle|s_\tau^{(j)}, A^{(i,j)}, C_{A_{ij}}\right)}{Z_{C_A}^{(i)}}\cdot\prod_{i=1}^{N_s}\frac{\prod_{j=1}^{N_s} P\left(s_\tau^{(i)}\middle|s_{\tau-1}^{(j)}, B^{(i,j)}, C_{B_{ij}}\right)}{Z_{C_B}^{(i)}}\right\}.
\end{aligned}
\tag{9}
$$

The prior beliefs about $A^{(i,j)}$, $B^{(i,j)}$ and $C_{A_{ij}}$, $C_{B_{ij}}$ follow the Dirichlet distributions $P\left(A^{(i,j)}\right) = \mathrm{Dir}\left(a^{(i,j)}\right)$, $P\left(B^{(i,j)}\right) = \mathrm{Dir}\left(b^{(i,j)}\right)$ and categorical distributions $P\left(C_{A_{ij}}\right) = \mathrm{Cat}\left(c_{A_{ij}}\right)$, $P\left(C_{B_{ij}}\right) = \mathrm{Cat}\left(c_{B_{ij}}\right)$, respectively. The initial states follow a categorical distribution: $P(s_1) = \mathrm{Cat}(D)$. As C_A takes the value of 1 or 0, the following equation holds:

$$
A^{(i,j)\odot C_{A_{ij}}^{\ddagger}} = \begin{cases} A^{(i,j)} & \left(C_{A_{ij}} = 1\right) \\ \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} & \left(C_{A_{ij}} = 0\right) \end{cases},
\tag{10}
$$

which means that $C_{A_{ij}} = 0$ reduces $A^{(i,j)}$ from the model owing to the factorial property of the model. The same applies to B. As mentioned in Section 2, $P\left(o_\tau^{(i)}\middle|s_\tau^{(j)}, A^{(i,j)}, C_{A_{ij}}\right)$ and $P\left(s_\tau^{(i)}\middle|s_{\tau-1}^{(j)}, B^{(i,j)}, C_{B_{ij}}\right)$ are given by categorical distributions $P\left(o_\tau^{(i)}\middle|s_\tau^{(j)}, A^{(i,j)}, C_{A_{ij}}\right) = \mathrm{Cat}\left(\left(A^{(i,j)\odot C_{A_{ij}}^{\ddagger}}\right)s_\tau^{(j)}\right)$ and $P\left(s_\tau^{(i)}\middle|s_{\tau-1}^{(j)}, B^{(i,j)}, C_{B_{ij}}\right) = \mathrm{Cat}\left(\left(B^{(i,j)\odot C_{B_{ij}}^{\ddagger}}\right)s_{\tau-1}^{(j)}\right)$, respectively.

The partition functions $Z_{C_A}^{(i)}$ and $Z_{C_B}^{(i)}$ serve as normalization factors to counterbalance the deviations of $\sum_{o_\tau^{(i)}=1,0} \prod_{j=1}^{N_s} P\left(o_\tau^{(i)} \middle| s_\tau^{(j)}, A^{(i,j)}, C_{A_{ij}}\right)$ and $\sum_{o_\tau^{(i)}=1,0} \prod_{j=1}^{N_s} P\left(s_\tau^{(i)} \middle| s_{\tau-1}^{(j)}, B^{(i,j)}, C_{B_{ij}}\right)$ from 1, which are expressed as

$$Z_{C_A}^{(i)} = \sum_{k=1}^{2} \prod_{j=1}^{N_s} \left(A_{k1}^{(i,j)^{C_{A_{ij}}}} \quad A_{k2}^{(i,j)^{C_{A_{ij}}}} \right) s_\tau^{(j)}, \qquad Z_{C_B}^{(i)} = \sum_{k=1}^{2} \prod_{j=1}^{N_s} \left(B_{k1}^{(i,j)^{C_{B_{ij}}}} \quad B_{k2}^{(i,j)^{C_{B_{ij}}}} \right) s_{\tau-1}^{(j)}. \quad (11)$$

4.2. Variational free energy and belief update rule

In this subsection, we describe the explicit forms of posterior beliefs and variational free energy and derive belief update rules that minimize variational free energy. The explicit form of Eq. (3) (adopting the factorial structure) can be expressed as

$$Q(s_{1:t}, \theta) = \prod_{i=1}^{N_o} \prod_{j=1}^{N_s} \left\{ Q\left(A^{(i,j)}\right) Q\left(C_{A_{ij}}\right) \right\} \cdot \prod_{i=1}^{N_s} \prod_{j=1}^{N_s} \left\{ Q\left(B^{(i,j)}\right) Q\left(C_{B_{ij}}\right) \right\} \cdot \prod_{\tau=1}^{t} \prod_{j=1}^{N_s} Q\left(s_\tau^{(j)}\right). \quad (12)$$

The marginal posterior beliefs about $A^{(i,j)}$, $B^{(i,j)}$, $C_{A_{ij}}$, $C_{B_{ij}}$, $s_\tau^{(j)}$ follow the Dirichlet distributions $Q\left(A^{(i,j)}\right) = \mathrm{Dir}\left(\mathbf{a}^{(i,j)}\right)$, $Q\left(B^{(i,j)}\right) = \mathrm{Dir}\left(\mathbf{b}^{(i,j)}\right)$ and categorical distributions $Q\left(C_{A_{ij}}\right) = \mathrm{Cat}\left(\left(\mathbf{c}_{A_{ij}}, \overline{\mathbf{c}_{A_{ij}}}\right)^{\mathrm{T}}\right)$, $Q\left(C_{B_{ij}}\right) = \mathrm{Cat}\left(\left(\mathbf{c}_{B_{ij}}, \overline{\mathbf{c}_{B_{ij}}}\right)^{\mathrm{T}}\right)$, $Q\left(s_\tau^{(j)}\right) = \mathrm{Cat}\left(\mathbf{s}_\tau^{(j)}\right)$, respectively.

Under the generative model defined in Eq. (9), the variational free energy is expressed as follows:

$$
\begin{aligned}
&F\big(o_{1:t}, Q(s_{1:t}, \theta)\big) \\
&\qquad = \sum_{\tau=1}^{t} \mathrm{E}_Q[-\ln P(o_\tau|s_\tau, A, C_A) - \ln P(s_\tau|s_{\tau-1}, B, C_B) + \ln Q(s_\tau)] \\
&\qquad + \mathcal{D}_{\mathrm{KL}}[Q(\theta)||P(\theta)] \\
&= \sum_{\tau=1}^{t} \big\{\mathbf{s}_\tau \cdot \big(\ln \mathbf{s}_\tau - \mathbf{c}_A^{\ddagger} \odot \ln \mathbf{A} \cdot o_\tau - \mathbf{c}_B^{\ddagger} \ln \mathbf{B}\, \mathbf{s}_{\tau-1}\big) + E_Q\big[\ln Z_{C_A}\big] + E_Q\big[\ln Z_{C_B}\big]\big\} \\
&\qquad + (\mathbf{a} - a) \cdot \ln \mathbf{A} - \ln \mathcal{B}(\mathbf{a}) + (\mathbf{b} - b) \cdot \ln \mathbf{B} - \ln \mathcal{B}(\mathbf{b}) \\
&\qquad + \left(\frac{\mathbf{c_A}}{\overline{\mathbf{c}}_A}\right) \cdot \left(\ln\left(\frac{\mathbf{c_A}}{\overline{\mathbf{c}}_A}\right) - \ln\left(\frac{C_A}{\overline{C}_A}\right)\right) + \left(\frac{\mathbf{c}_B}{\overline{\mathbf{c}}_B}\right) \cdot \left(\ln\left(\frac{\mathbf{c}_B}{\overline{\mathbf{c}}_B}\right) - \ln\left(\frac{C_B}{\overline{C}_B}\right)\right).
\end{aligned}
\tag{13}
$$

Note that factorial structures (i.e., summations over N_s and N_o) are implicitly involved for simplification. In the above, $\mathcal{D}_{\mathrm{KL}}[Q(\theta)||P(\theta)]$ denotes the complexity scored by the Kullback–Leibler divergence between $Q(\theta)$ and $P(\theta)$, $\mathcal{B}\left(\mathbf{a}_{\cdot l}^{(i,j)}\right) \coloneqq \Gamma\left(\mathbf{a}_{1l}^{(i,j)}\right)\Gamma\left(\mathbf{a}_{2l}^{(i,j)}\right)/\Gamma\left(\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{2l}^{(i,j)}\right)$ denotes the beta function characterized by the gamma function $\Gamma(\cdot)$, $\overline{X}$ denotes $1 - X$, and $\ln \mathbf{A}^{(i,j)} \coloneqq \mathrm{E}_{Q(A^{(i,j)})}\big[\ln A^{(i,j)}\big]$ denotes the posterior expectation of $\ln A^{(i,j)}$. Using a first-order approximation on $A^{(i,j)}$ and $B^{(i,j)}$, $\mathrm{E}_Q[\ln Z]$ can be calculated as follows:

$$
\mathrm{E}_Q\big[\ln Z_{C_A}\big] \approx \ln 2 \cdot \sum_{i=1}^{N_o}\left(1 - \sum_{j=1}^{N_s} \mathbf{c}_{A_{ij}}\right), \qquad \mathrm{E}_Q\big[\ln Z_{C_B}\big] \approx \ln 2 \cdot \sum_{i=1}^{N_s}\left(1 - \sum_{j=1}^{N_s} \mathbf{c}_{B_{ij}}\right). \tag{14}
$$

The detailed derivations of this approximation are provided in Appendix B. The term $\mathrm{E}_Q\big[\ln Z_{C_A}\big] + \mathrm{E}_Q\big[\ln Z_{C_B}\big]$ is independent of $\mathbf{s}$, $\mathbf{A}$, and $\mathbf{B}$ and corresponds to $\ln Z$ in Eq. (8), which also corresponds to α in Eq. (6) that is independent of x, W, and K. Thus, this relationship confirms the formal equivalence between the variational free energy under BSyMP (Eq. (8)) and the cost function for canonical neural networks with synaptic pruning (Eq. (6)).

Variational free energy minimization yields posterior beliefs about the hidden states and parameters. By solving the fixed point of the implicit gradient descent $\partial F/\partial \mathbf{s}_\tau = 0$ and $\partial F/\partial \boldsymbol{\theta} = O$, the update rules are derived as follows:

$$\mathbf{s}_\tau^{(n)} = \sigma\left(\sum_{i=1}^{N_o} \mathbf{c}_{A_{in}} \ln \mathbf{A}^{(i,n)} \cdot o_\tau^{(i)} + \sum_{j=1}^{N_s} \mathbf{c}_{B_{nj}} \ln \mathbf{B}^{(n,j)}\, \mathbf{s}_{\tau-1}^{(j)}\right), \tag{15}$$

$$\mathbf{a}^{(i,j)} = a^{(i,j)} + \mathbf{c}_{A_{ij}} \sum_{\tau=1}^{t} o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}, \tag{16}$$

$$\mathbf{b}^{(i,j)} = b^{(i,j)} + \mathbf{c}_{B_{ij}} \sum_{\tau=2}^{t} \mathbf{s}_\tau^{(i)} \otimes \mathbf{s}_{\tau-1}^{(j)}, \tag{17}$$

$$\mathbf{c}_{A_{ij}} = \mathrm{sig}\left(\mathrm{sig}^{-1}\left(c_{A_{ij}}\right) + \sum_{\tau=1}^{t} \left(\ln \mathbf{A}^{(i,j)} \cdot \left(o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}\right) + \ln 2\right)\right), \tag{18}$$

$$\mathbf{c}_{B_{ij}} = \mathrm{sig}\left(\mathrm{sig}^{-1}\left(c_{B_{ij}}\right) + \sum_{\tau=2}^{t} \left(\ln \mathbf{B}^{(i,j)} \cdot \left(\mathbf{s}_\tau^{(i)} \otimes \mathbf{s}_{\tau-1}^{(j)}\right) + \ln 2\right)\right). \tag{19}$$

Equations (18) and (19) are equivalent to the equations for the dynamics of C_W and C_K in the neural network formation that are derived from the fixed point of the gradient descent on L, $\partial L/\partial C_W = O$ and $\partial L/\partial C_K = O$. That is, Eqs. (18) and (19) are plausible synaptic pruning rules for the canonical neural network.

4.3. Properties of connectivity parameters

In this subsection, we show that $\mathbf{c}_{A_{ij}}$ converges to 0 when $o_\tau^{(i)}$ is randomly determined and independent of all elements of s_τ, and the likelihood mapping posterior $\mathbf{A}^{(i,j)}$ and the mean state posterior $\mathrm{E}_{P(o_\tau)}\left[\mathbf{s}_\tau^{(j)}\right]$ are sufficiently flat. The direction of the change in $\mathbf{c}_{A_{ij}}$ is determined by the sign of the second term on the right-hand side of Eq. (18). We adopt the following approximation for analytical tractability: $\ln \mathbf{A}^{(i,j)} \coloneqq \mathrm{E}_{Q(A^{(i,j)})}\left[\ln A^{(i,j)}\right] \approx \ln \mathrm{E}_{Q(A^{(i,j)})}\left[A^{(i,j)}\right] = \ln \begin{pmatrix} 0.5+\Delta_{ij1} & 0.5+\Delta_{ij2} \\ 0.5-\Delta_{ij} & 0.5-\Delta_{ij} \end{pmatrix}$. This approximation holds when $\mathbf{a}^{(i,j)}$ is sufficiently large, owing to the properties of the digamma functions (Isomura & Friston, 2020). Subsequently, the Taylor expansion of $\ln \mathbf{A}^{(i,j)}$ at $\frac{1}{2}$ yields the following equation:

$$\ln \mathbf{A}^{(i,j)} = -\ln 2 + \mathrm{Odd}\begin{pmatrix} \Delta_{ij1} & \Delta_{ij2} \\ -\Delta_{ij1} & -\Delta_{ij2} \end{pmatrix} + \mathrm{Even}\begin{pmatrix} \Delta_{ij1} & \Delta_{ij2} \\ \Delta_{ij1} & \Delta_{ij} \end{pmatrix}, \tag{20}$$

where $\mathrm{Odd}(x)$ and $\mathrm{Even}(x)$ are odd and even function components of $\ln(0.5+x)$. This equation allows the following transformation:

$$\ln \mathbf{A}^{(i,j)} \cdot \left(o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}\right) + \ln 2 = \sum_{n=1}^{\infty} \left\{\mathrm{Odd}(\Delta_{ij})\left(o_{\tau 1}^{(i)} - o_{\tau 2}^{(i)}\right) + \mathrm{Even}(\Delta_{ij})\right\} \cdot \mathbf{s}_\tau^{(j)}, \tag{21}$$

where Δ_{ij} and $o_{\tau l}^{(i)}$ $(l = 1,2)$ are given by $\Delta_{ij} \coloneqq \left(\Delta_{ij1}, \Delta_{ij}\right)^{\mathrm{T}}$ and $o_\tau^{(i)} = \left(o_{\tau 1}^{(i)}, o_{\tau 2}^{(i)}\right)^{\mathrm{T}}$. Because we consider the case where $o_\tau^{(i)}$ is independent of all elements of s_τ, $o_\tau^{(i)}$ is independent of all the other observation. Here and after, $o_\tau^{(i')} = o_\tau \backslash o_\tau^{(i)}$ denotes the set of all the observations other than *i*th observation $o_\tau^{(i)}$. The expectation of $\ln \mathbf{A}^{(i,j)} \cdot \left(o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}\right) + \ln 2$ over observations is as follows:

$$\begin{aligned} &\mathrm{E}_{P(o_\tau)}\left[\ln \mathbf{A}^{(i,j)} \cdot \left(o_\tau^{(i)} \otimes \mathbf{s}_\tau^{(j)}\right) + \ln 2\right] \\ &= \sum_{n=1}^{\infty} \mathrm{Odd}(\Delta_{ij}) \cdot \mathrm{E}_{P\left(o_\tau^{(i)}\right)}\left[\left(o_{\tau 1}^{(i)} - o_{\tau 2}^{(i)}\right)\mathbf{s}'^{(j)}_\tau\left(o_\tau^{(i)}\right)\right] + \mathrm{Even}(\Delta_{ij}) \cdot \mathrm{E}_{P\left(o_\tau^{(i)}\right)}\left[\mathbf{s}'^{(j)}_\tau\left(o_\tau^{(i)}\right)\right]. \end{aligned} \tag{22}$$

where $\mathbf{s}'^{(j)}_{\tau}\left(o^{(i)}_{\tau}\right) \coloneqq \mathrm{E}_{P\left(o^{(i')}_{\tau}\right)}\left[\mathbf{s}^{(j)}_{\tau}\right]$ denotes the expected state posterior inferred without observing $o^{(i)}_{\tau}$. This is a random variable that depends on $o^{(i)}_{\tau}$. Because we consider the case in which $o^{(i)}_{\tau}$ changes randomly, i.e., $p\left(o^{(i)}_{\tau 1} = 1 \middle| s_{\tau}\right) = p\left(o^{(i)}_{\tau 2} = 1 \middle| s_{\tau}\right) = 1/2$, the expectation of $\ln \mathbf{A}^{(i,j)} \cdot \left(o^{(i)}_{\tau} \otimes \mathbf{s}^{(j)}_{\tau}\right) + \ln 2$ over observations is calculated as follows:

$$\mathrm{E}_{P(o_{\tau})}\left[\ln \mathbf{A}^{(i,j)} \cdot \left(o^{(i)}_{\tau} \otimes \mathbf{s}^{(j)}_{\tau}\right) + \ln 2\right] = \frac{1}{2}\mathbf{s}'^{(j)}_{\tau}\left(\begin{pmatrix}1\\0\end{pmatrix}\right) \cdot \ln 2\mathbf{A}^{(i,j)}_{1\cdot} + \frac{1}{2}\mathbf{s}'^{(j)}_{\tau}\left(\begin{pmatrix}0\\1\end{pmatrix}\right) \cdot \ln 2\mathbf{A}^{(i,j)}_{2\cdot}. \quad (23)$$

Furthermore, in the case where $\left|\Delta_{ij}\right| \ll 1$ holds, $\ln \mathbf{A}^{(i,j)} \approx -\ln 2 + 2\begin{pmatrix}\Delta_{ij1} & \Delta_{ij2}\\ -\Delta_{ij1} & -\Delta_{ij2}\end{pmatrix} + \mathcal{O}\left(\Delta_{ij}^{\odot 2}\right)$ holds. With this approximation, the Taylor expansion of $\mathbf{s}'^{(j)}_{\tau}\left(o^{(i)}_{\tau}\right)$ provides the following approximation:

$$\mathbf{s}'^{(j)}_{\tau}\left(o^{(i)}_{\tau}\right) \approx S + 2\mathbf{c}_{A_{ij}}\left(o^{(i)}_{\tau 1} - o^{(i)}_{\tau 2}\right) S \odot (1 - S) \odot \Delta_{ij} + \mathcal{O}\left(\Delta_{ij}^{\odot 2}\right), \quad (24)$$

where S is defined by $S \coloneqq \mathrm{E}_{P\left(o^{(i')}_{\tau}\right)}\left[\sigma\left(\sum_{i' \neq i} \mathbf{c}_{A_{i'j}} \ln \mathbf{A}^{(i',j)} \cdot o^{(i')}_{\tau}\right)\right]$. Besides, $\mathrm{E}_{P(o_{\tau})}\left[\mathbf{s}^{(j)}_{\tau}\right] = \mathrm{E}_{P\left(o^{(i)}_{\tau}\right)}\left[\mathbf{s}'^{(j)}_{\tau}\left(o^{(i)}_{\tau}\right)\right] = S + \mathcal{O}\left(\Delta_{ij}^{\odot 2}\right)$ holds in this case. Subsequently, when $\mathrm{E}_{P(o_{\tau})}\left[\mathbf{s}^{(j)}_{\tau}\right] = \left(\frac{1}{2} + \epsilon, \frac{1}{2} - \epsilon\right)^{\mathrm{T}}$ with $|\epsilon| \ll 1$ holds, the expectation of interest is approximated as follows:

$$\mathrm{E}_{P(o_{\tau})}\left[\ln \mathbf{A}^{(i,j)} \cdot \left(o^{(i)}_{\tau} \otimes \mathbf{s}^{(j)}_{\tau}\right) + \ln 2\right] \approx \left({\Delta_{ij1}}^2 + {\Delta_{ij2}}^2\right)\left(\mathbf{c}_{A_{ij}} - 1 - 2\epsilon \cdot \frac{{\Delta_{ij1}}^2 - {\Delta_{ij2}}^2}{{\Delta_{ij1}}^2 + {\Delta_{ij2}}^2}\right). \quad (25)$$

This means that the expectation above is negative when $\mathbf{c}_{A_{ij}} < 1 + 2\epsilon \cdot \frac{{\Delta_{ij1}}^2 - {\Delta_{ij2}}^2}{{\Delta_{ij1}}^2 + {\Delta_{ij2}}^2}$ holds, that is, when $\mathbf{c}_{A_{ij}}$ is not sufficiently close to 1. In addition, $\sum_{\tau=1}^{t}\left(\ln \mathbf{A}^{(i,j)} \cdot \left(o^{(i)}_{\tau} \otimes \mathbf{s}^{(j)}_{\tau}\right) + \ln 2\right)$ has a $\left|\Delta_{ij}\right|\sqrt{t}$-order standard deviation according to the central limit theorem. Taken together, $\sum_{\tau=1}^{t}\left(\ln \mathbf{A}^{(i,j)} \cdot \left(o^{(i)}_{\tau} \otimes \mathbf{s}^{(j)}_{\tau}\right) + \ln 2\right)$ is a $\left|\Delta_{ij}\right|^2 t$-order negative value with a $\left|\Delta_{ij}\right|\sqrt{t}$-order standard deviation, which decreases monotonically with time when t is sufficiently large. Therefore, the direction of change in $\mathbf{c}_{A_{ij}}$ is negative when $o^{(i)}_{\tau}$ randomly determined and independent of all elements of s_{τ}, and $\mathbf{A}^{(i,j)}$ and $E_{P(o_{\tau})}\left[\mathbf{s}^{(j)}_{\tau}\right]$ are sufficiently flat. Hence, $\mathbf{c}_{A_{ij}}$ converges to 0 after sufficient sessions.

4.4. Numerical simulation 1: Blind source separation

The agents received 1024 (i.e., 32 × 32 pixels) observations that were generated using the generative model defined in Eq. (2). The model comprised two states, $s^{(1)}$ and $s^{(2)}$, which generated the observations. Both states randomly took values of 0 or 1 with a probability of 0.5 each (as depicted by the flat matrix B with all elements of 0.5). When $s^{(1)}$ or $s^{(2)}$ was ON, the shape of a circle or rectangle was displayed on the screen as sensory input for the agent (Fig. 2a). This was implemented using the following matrix A:

$$A^{(i,1)} = \begin{cases} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} & (i \notin \text{the points on a circle}) \\ \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix} & (i \in \text{the points on a circle}) \end{cases}, \tag{26}$$

$$A^{(i,2)} = \begin{cases} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} & (i \notin \text{the points on a rectangle}) \\ \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix} & (i \in \text{the points on a rectangle}) \end{cases}. \tag{27}$$

The agent optimized $\mathbf{s}$, $\mathbf{a}$, and $\mathbf{c_A}$ (i.e., $\mathbf{b}$ and $\mathbf{c_B}$ were omitted from this simulation). The prior beliefs about A and C_A characterized the inference of the agents. The elements of prior belief $a^{(i,j)}$ with the i-index denoting a point on the shape is expressed as follows: $a^{(i,j)} = \begin{pmatrix} 25.5 & 24.5 \\ 24.5 & 25.5 \end{pmatrix}$. The other elements $a^{(i,j)}$ were randomly sampled from a uniform distribution with the center of 25 and interval of 1. All elements of prior belief about C_A were sampled from a uniform distribution with the center of 0.9 and interval of 0.1. Further details are provided in the source code.

The full model and BSyMP agents optimized their posterior beliefs: they optimized $\mathbf{s}_\tau$ at every time point and optimized $\mathbf{a}$ and $\mathbf{c_A}$ at every 10 time points. Following optimization over 10,000

time points, BMR was applied to the learned full model (Friston et al., 2019). BMR computed the change in free energy $\Delta F^{(i,j)}$ when each $a^{(i,j)}$ was varied. A prior for reduction $\tilde{a}^{(i,j)}$ was set to a uniform matrix with large positive values. The element of the posterior $\mathbf{a}^{(i,j)}$ was reduced if the change in the approximate free energy $\Delta F^{(i,j)} \coloneqq \ln \mathcal{B}\left(\tilde{a}^{(i,j)}\right) - \ln \mathcal{B}\left(\tilde{\mathbf{a}}^{(i,j)}\right) - \ln \mathcal{B}\left(a^{(i,j)}\right) + \ln \mathcal{B}\left(\mathbf{a}^{(i,j)}\right)$ was negative and retained otherwise. Further details regarding BMR can be found in Friston et al., 2019.

After performing BMR, the variational free energy was calculated by following Eq. (4) at the end of the training sessions, however, the complexity of the connectivity parameters was excluded for a consistent evaluation among the three methods. Subsequently, the normalized mapping posteriors for all agents were calculated. The normalized mapping posterior $\mathbf{A}$ for the full model and BMR and $\mathbf{A}^{\odot \mathbf{c}_A^{\ddagger}}$ for BSyMP are expressed as follows:

$$\mathbf{A}_{kl}^{(i,j)} = \frac{\mathbf{a}_{kl}^{(i,j)}}{\mathbf{a}_{1l}^{(i,j)} + \mathbf{a}_{2l}^{(i,j)}}, \qquad \mathbf{A}^{\odot \mathbf{c}_A^{\ddagger}}{}_{kl}^{(i,j)} = \frac{\left(\mathbf{a}_{kl}^{(i,j)}\right)^{\mathbf{c}_{A_{ij}}}}{\left(\mathbf{a}_{1l}^{(i,j)}\right)^{\mathbf{c}_{A_{ij}}} + \left(\mathbf{a}_{2l}^{(i,j)}\right)^{\mathbf{c}_{A_{ij}}}}. \tag{28}$$

Finally, 1000 time-point-length data points were created using the same generative process to test the performance of the agents.

4.5. Numerical simulation 2: Rule learning

The agents received 3072 (i.e., 32 × 32 × 3 pixels) observations generated using the generative model defined in Eq. (2). The model comprised 12 states, each of which exhibited a certain shape

on one of the three screens. Matrix A was set similarly to that in Simulation 1, but the fidelity of the observations was determined as follows:

$$A^{(i,j)} = \begin{cases} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} & (i \notin \text{the points on a shape}) \\ \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} & (i \in \text{the points on a shape}) \end{cases}. \tag{29}$$

The state change followed the transition rule characterized by matrix B. All 2 × 2 submatrices were symmetric. Therefore, $B_{11}^{(i,j)}$ was sufficient to characterize $B^{(i,j)}$, as follows:

$$B^{(i,j)} = \begin{pmatrix} B_{11}^{(i,j)} & 1 - B_{11}^{(i,j)} \\ 1 - B_{11}^{(i,j)} & B_{11}^{(i,j)} \end{pmatrix}. \tag{30}$$

The initial states were determined randomly, as depicted by a flat block vector D with all elements of 0.5.

The agent employed the prior beliefs about A, B and C_A, C_B. The elements of prior belief $a^{(i,j)}$ were set as follows:

$$a^{(i,j)} = \begin{cases} \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} \cdot 10^5 & (i \notin \text{the points on a shape}) \\ \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \cdot 10^5 & (i \in \text{the points on a shape}) \end{cases}. \tag{31}$$

All elements of the prior belief $b^{(i,j)}$ were sampled from a uniform distribution with the mean of 50 and interval of 2. All elements of the prior beliefs c_A and c_B were sampled from a uniform distribution with the mean of 0.9 and interval of 0.1. Further details are provided in the source code.

The full model and BSyMP agents optimized their posterior beliefs: they optimized $\mathbf{s}_\tau$ at every time point and optimized $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}_A$, and $\mathbf{c}_B$ at every 100 time points. Following optimization over 10,000 time points, BMR was applied to the learned full model in the same manner as in Simulation 1.

After performing BMR, the variational free energy and the normalized transition rule posteriors $\mathbf{B}$ and $\mathbf{B}^{\odot \mathbf{c}_B^{\ddagger}}$ were calculated in the same manner as in Simulation 1. Finally, 1000 time-point-length data points were created using the same generative process to test the prediction performance of the agents. The predicted state of the *i*-th source was defined as $\mathbf{s}_{\tau|\tau-1}^{(i)} \coloneqq \sigma\left(\sum_{j=1}^{N_s} \mathbf{c}_{B\,ij} \ln \mathbf{B}^{(i,j)}\, \mathbf{s}_{\tau-1}^{(j)}\right)$ for BSyMP and $\mathbf{s}_{\tau|\tau-1}^{(i)} \coloneqq \sigma\left(\sum_{j=1}^{N_s} \ln \mathbf{B}^{(i,j)}\, \mathbf{s}_{\tau-1}^{(j)}\right)$ for the full model and BMR. The actual probability of the *i*-th source state was defined as follows: $s_{\tau|\tau-1}^{(i)} = \sigma\left(\sum_{j=1}^{N_s} \ln B^{(i,j)} s_{\tau-1}^{(j)}\right)$.

Acknowledgment

T.I. is funded by the Japan Society for the Promotion of Science (JSPS) KAKENHI (Ref: JP23H04973) and the Japan Science and Technology Agency (JST) CREST (Ref: JPMJCR22P1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CRediT authorship contribution statement

Ukyo T. Tazawa: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

Takuya Isomura: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no competing interests or personal relationships that could have appeared to influence the work reported in this paper.

References

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

Braun, D. A., Mehring, C., & Wolpert, D. M. (2010). Structure learning in action. *Behavioural Brain Research*, *206*(2), 157–165. https://doi.org/10.1016/j.bbr.2009.08.031

Chechik, G., Meilijson, I., & Ruppin, E. (1998). Synaptic pruning in development: a computational account. *Neural Computation*, *10*(7), 1759–1777. https://doi.org/10.1162/089976698300017124

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active Inference, curiosity and insight. *Neural Computation*, *29*(10), 2633–2683. https://doi.org/10.1162/neco_a_00999

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Friston, K., Parr, T., & Zeidman, P. (2019). Bayesian model reduction. arXiv preprint arXiv:1805.07092. https://doi.org/10.48550/arXiv.1805.07092

Hagiwara, M. (1994). A simple and effective method for removal of hidden units and weights. *Neurocomputing*, *6*(2), 207–218. https://doi.org/10.1016/0925-2312(94)90055-8

Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv preprint arXiv:1510.00149. https://doi.org/10.48550/arXiv.1510.00149

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. arXiv preprint arXiv:1506.02626. https://doi.org/10.48550/arXiv.1506.02626

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

Helias, M., Rotter, S., Gewaltig, M.-O., & Diesmann, M. (2008). Structural plasticity controlled by calcium based correlation detection. *Frontiers in Computational Neuroscience*, *2*. https://doi.org/10.3389/neuro.10.007.2008

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks. arXiv preprint arXiv: 2102.00554. https://doi.org/10.48550/arXiv.2102.00554

Isomura, T., & Friston, K. (2020). Reverse-engineering neural networks to characterize their cost functions. *Neural Computation*, *32*(11), 2085–2121. https://doi.org/10.1162/neco_a_01315

Isomura, T., Shimazaki, H., & Friston, K. J. (2022). Canonical neural networks perform active inference. *Communications Biology*, *5*(1), Article 55. https://doi.org/10.1038/s42003-021-02994-2

Jafarian, A., Zeidman, P., Litvak, V., & Friston, K. (2019). Structure learning in coupled dynamical systems and dynamic causal modelling. *Philosophical Transactions of the Royal Society A*, *377*(2160), 20190048. https://doi.org/10.1098/rsta.2019.0048

Kasai, H., Ziv, N. E., Okazaki, H., Yagishita, S., & Toyoizumi, T. (2021). Spine dynamics in the brain, mental disorders and artificial neural networks. *Nature Reviews Neuroscience*, *22*(7), 407–422. https://doi.org/10.1038/s41583-021-00467-3

Le Bé, J.-V., & Markram, H. (2006). Spontaneous and evoked synaptic rewiring in the neonatal neocortex. *Proceedings of the National Academy of Sciences*, *103*(35), 13214–13219. https://doi.org/10.1073/pnas.0604691103

Lüscher, C., Nicoll, R. A., Malenka, R. C., & Muller, D. (2000). Synaptic plasticity and dynamic modulation of the postsynaptic membrane. *Nature Neuroscience*, *3*(6), 545–550. https://doi.org/10.1038/75714

Lynn, C. W., & Bassett, D. S. (2020). How humans learn and represent networks. *Proceedings of the National Academy of Sciences*, *117*(47), 29407–29415. https://doi.org/10.1073/pnas.1912328117

Malenka, R. C., & Bear, M. F. (2004). LTP and LTD: An embarrassment of riches. *Neuron*, *44*(1), 5–21. https://doi.org/10.1016/j.neuron.2004.09.012

Nicoll, R. A. (2017). A brief history of long-term potentiation. *Neuron*, *93*(2), 281–290. https://doi.org/10.1016/j.neuron.2016.12.015

Sakai, J. (2020). How synaptic pruning shapes neural wiring during development and, possibly, in disease. *Proceedings of the National Academy of Sciences*, *117*(28), 16096–16099. https://doi.org/10.1073/pnas.2010281117

Scannevin, R. H., & Huganir, R. L. (2000). Postsynaptic organisation and regulation of excitatory synapses. *Nature Reviews Neuroscience*, *1*(2), 133–141. https://doi.org/10.1038/35039075

Scholl, C., Rule, M. E., & Hennig, M. H. (2021). The information theory of developmental pruning: optimizing global network architectures using local synaptic rules. *PLOS Computational Biology*, *17*(10), e1009458-. https://doi.org/10.1371/journal.pcbi.1009458

Sietsma, J., & Dow, R. J. F. (1991). Creating artificial neural networks that generalize. *Neural Networks*, *4*(1), 67–79. https://doi.org/10.1016/0893-6080(91)90033-2

Sun, Y., Wang, X., & Tang, X. (2015). Sparsifying neural network connections for face recognition. arXiv preprint arXiv: 1512.01891. https://doi.org/10.48550/arXiv.1512.01891

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Current Opinion in Neurobiology*, *37*, 99–105. https://doi.org/10.1016/j.conb.2016.01.014

Thomas, B. T., Blalock, D. W., & Levy, W. B. (2015). Adaptive synaptogenesis constructs neural codes that benefit discrimination. *PLOS Computational Biology*, *11*(7), e1004299-. https://doi.org/10.1371/journal.pcbi.1004299

Ziv, N. E., & Brenner, N. (2018). Synaptic tenacity or lack thereof: spontaneous remodeling of synapses. *Trends in Neurosciences*, *41*(2), 89–99. https://doi.org/10.1016/j.tins.2017.12.003