Chapter 11
Neural Engine Hypothesis

Hideaki Shimazaki

11.1 Introduction

Humans and animals change sensitivity to sensory stimulus either adaptively to the stimulus conditions or following a behavioral context even if the stimulus does not change. A potential neurophysiological basis underlying these observations is gain modulation that changes responsiveness of neurons to stimulus; an example is contrast gain-control found in retina (Sakmann and Creutzfeldt 1969) and primary visual cortex under anesthesia (Ohzawa et al. 1985; Laughlin 1989), or in higher visual area caused by attention (Reynolds et al. 2000; Martínez-Trujillo and Treue 2002). Theoretical considerations suggested the gain modulation as a nonlinear operation that integrates information from different origins, offering ubiquitous computation performed in neural systems (see Salinas and Sejnowski (2001), Carandini and Heeger (2012) for reviews). Regulation of the level of background synaptic inputs (Chance et al. 2002; Burkitt et al. 2003), shunting inhibition (Doiron et al. 2001; Prescott and De Koninck 2003; Mitchell and Silver 2003), and synaptic depression (Abbott et al. 1997; Rothman et al. 2009) among others have been suggested as potential biophysical mechanisms of the gain modulation (see Silver (2010) for a review). While such modulation of the informative neural activity is a hallmark of computation performed internally in an organism, a principled view to quantify the internal computation has not been proposed yet.

Neurons convey information about the stimulus in their activity patterns. To describe probabilities of a combinatorially large number of activity patterns of the neurons with a smaller number of activity features, the maximum entropy principle has been successfully used (Schneidman et al. 2006; Shlens et al. 2006). This

H. Shimazaki (✉)
Kyoto University, Kyoto, Japan and Honda Research Institute Japan, Saitama, Japan
e-mail: h.shimazaki@i.kyoto-u.ac.jp

Z. Chen, S.V. Sarma (eds.), *Dynamic Neuroscience*,
https://doi.org/10.1007/978-3-319-71976-4_11

principle constructs the least structured probability distribution given the small set of specified constraints on the distribution, known as a maximum entropy model. It explains probabilities of activity patterns as a result of nonlinear operation on the specified features using a softmax function. Moreover, the model belongs to an exponential family distribution, or a Gibbs distribution. Equivalence of inference under the maximum entropy principle with aspects of the statistical mechanics and thermodynamics was explicated through the work by Jaynes (1957). Recently thermodynamic quantities were used to assess criticality of neural activity (Tkačik et al. 2014, 2015). However, analysis of neural populations under this framework only recently started to include "dynamics" of a neural population (Shimazaki et al. 2009, 2012; Shimazaki 2013; Kass et al. 2011; Kelly and Kass 2012; Granot-Atedgi et al. 2013; Nasser et al. 2013; Donner et al. 2017), and has not yet reached maturity to include computation performed internally in an organism.

Based on a neural population model obtained under the maximum entropy principle, this study investigates neural dynamics during which gain of neural response to a stimulus is modulated with a delay by an internal mechanism to enhance the stimulus information. The delayed gain modulation is observed at different stages of visual pathways (McAdams and Maunsell 1999; Reynolds et al. 2000; Lee et al. 2003). For example, effect of contrast gain-control by attention on response of V4 neurons to high contrast stimulus appears 200–300 ms after the stimulus presentation, but is absent during 100–200 ms time period during which the neural response is returning to a spontaneous rate (Reynolds et al. 2000). This process is expected for dynamics of neurons subject to a feedback gain-modulation mechanism, e.g., via recurrent networks (Salinas and Abbott 1996; Spratling and Johnson 2004; Sutherland et al. 2009). Similar modulation of the late activity component of neurons is discussed as underpinnings of working memory (Supèr et al. 2001), sensory perception (Cauller and Kulics 1991; Sachidhanandam et al. 2013; Manita et al. 2015), and reward value (Schultz 2016). We demonstrate that our hypothetical neural dynamics with delayed gain-modulation forms an information-theoretic cycle that generates entropy ascribed to the stimulus-related activity using entropy supplied by the internal gain-modulation mechanism. The process works analogously to a heat engine that produces work from heat supplied by reservoirs. We hypothesize that neurons in the brain act in this manner when it actively modulates the incoming sensory information to enhance perceptual capacity.

This chapter is organized as follows. In Sect. 11.2, we construct a maximum entropy model of a neural population by constraining two types of activities, one that is directly regulated by stimulus and the other that represents background activity of neurons, termed "internal activity." We point out that modulation of the internal activity realizes gain-modulation of stimulus response. In Sect. 11.3, we explain the conservation of entropy, equation of state for the neural population, and information on stimulus. In Sect. 11.4, we construct cycles of neural dynamics that model stimulus-evoked activity during which the stimulus information is enhanced by the internal gain-modulation mechanism. We define entropic efficiency of gain-modulation performed to retain the stimulus information. An ideal cycle introduced in this section achieves the highest efficiency. The chapter ends with discussion in which the state-space model of the neural population is argued as a potential approach to test the hypothesis. Thermodynamic formulation and derivations of free energies for a neural population are summarized in Appendix.

11.2 A Simple Model of Gain Modulation by a Maximum Entropy Model

11.2.1 Maximum Entropy Model of Spontaneous Neural Activity

We start by modeling spontaneous activity of N spiking neurons. We represent a state of the i-th neuron by a binary variable $x_i = (0, 1)$ $(i = 1 \cdots N)$. Here silence of the neuron is represented by "0" whereas activity, or a spike, of the neuron is denoted by "1." The simultaneous activity of the N neurons is represented by a vector of the binary variables, $\mathbf{x} = (x_1, \ldots, x_N)$. The joint probability mass function, $p(\mathbf{x})$, describes the probability of generating the pattern $\mathbf{x}$. There are 2^N different patterns. We characterize the combinatorial neural activity with a smaller number of characteristic features $F_i(\mathbf{x})$ $(i = 1, \ldots, d$, where $d < 2^N)$, based on the maximum entropy principle. Here $F_i(\mathbf{x})$ is the i-th feature that combines the activity of individual neurons. For example, these features can be the first and second order interactions, $F_i(\mathbf{x}) = x_i$ for $i = 1, \ldots, N$, and $F_{N+(N-i/2)(i-1)+j-i}(\mathbf{x}) = x_i x_j$ for $i < j$. The maximum entropy principle constructs the least structured probability distribution while expected values of these features are specified (Jaynes 1957). By representing expectation by $p(\mathbf{x})$ using a bracket $\langle \cdot \rangle$, these constraints are written as $\langle F_i(\mathbf{x}) \rangle = c_i$ $(i = 1, \ldots, d)$, where c_i is the specified constant.

Maximization of a function subject to the equality constraints is formulated by the method of Lagrange multipliers that alternatively maximizes the following Lagrange function

$$\mathscr{L}[p] = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) - a \sum_{\mathbf{x}} p(\mathbf{x}) - \sum_{i} b_i \left\{ \sum_{\mathbf{x}} p(\mathbf{x}) F_i(\mathbf{x}) - c_i \right\} , \tag{11.1}$$

where a and b_i $(i = 1, \ldots, d)$ are the Lagrange multipliers. The Lagrange function is a functional of the probability mass function. By finding a zero point of its variational derivative, we obtain

$$p(\mathbf{x}) \sim \exp\left(-\sum_{i} b_i F_i(\mathbf{x}) \right) . \tag{11.2}$$

The Lagrange parameters b_i are obtained by simultaneously solving $\frac{\partial \mathscr{L}}{\partial b_i} = \langle F_i(\mathbf{x}) \rangle - c_i = 0$ for $i = 1, \ldots, d$. Many gradient algorithms and approximation methods have been developed to search the parameters. Activities of retinal ganglion cells (Schneidman et al. 2006; Shlens et al. 2006; Tkačik et al. 2014, 2015), hippocampal (Shimazaki et al. 2015), and cortical neurons (Tang et al. 2008; Yu et al. 2008; Shimazaki et al. 2012) were successfully characterized using Eq. (11.2). In the following, we use a vector notation $\mathbf{b}_0 = (b_1, \ldots, b_d)^\top$ and $\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), \ldots, F_d(\mathbf{x}))^\top$. Here $\mathscr{H}_0 \equiv \mathbf{b}_0^\top \mathbf{F}(\mathbf{x})$ is a Hamiltonian of the spontaneously active neurons. In statistical mechanics, Eq. (11.2) is identified as the Boltzmann distribution with a unit thermodynamic *beta*. If the features contain only up to the second order interactions, the model is equivalent to the Ising or spin-glass model for ferromagnetism.

11.2.2 Maximum Entropy Model of Evoked Neural Activity

In this subsection, we model evoked activity of neurons caused by changes in extrinsic stimulus conditions. We define a feature of stimulus-related activity as $X(\mathbf{x}) = \mathbf{b}_1^\top \mathbf{F}(\mathbf{x})$, where elements of $\mathbf{b}_1$ dictate response properties of each feature in $\mathbf{F}(\mathbf{x})$ to a stimulus. For simplicity, we represent the stimulus-related activity by this single feature, and consider that the evoked activity is characterized by the two summarized features, $\mathscr{H}_0(\mathbf{x})$ and $X(\mathbf{x})$. To model it, we constrain expectation of the internal and stimulus features using U and X, respectively. Here we assume that $\mathbf{F}(\mathbf{x})$, $\mathbf{b}_0$, and $\mathbf{b}_1$ are known and fixed. For example, this would model responses of visual neurons when we change contrast of a stimulus while fixing the rest of the stimulus properties. The maximum entropy distribution subject to these constraints is again given by the method of Lagrange multipliers. The Lagrange function is given as

$$\begin{aligned}\mathscr{L}[p] = &- \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) \\ &- a \sum_{\mathbf{x}} p(\mathbf{x}) - \beta \left\{ \sum_{\mathbf{x}} p(\mathbf{x}) \mathscr{H}_0(\mathbf{x}) - U \right\} + \alpha \left\{ \sum_{\mathbf{x}} p(\mathbf{x}) X(\mathbf{x}) - X \right\}.\end{aligned} \tag{11.3}$$

Here a, β, and α are the Lagrange parameters. By maximizing the functional $\mathscr{L}$ with respect to p, we obtain the following maximum entropy model,

$$p(\mathbf{x}) = \exp[-\beta \mathscr{H}_0(\mathbf{x}) + \alpha X(\mathbf{x}) - \psi(\beta, \alpha)], \tag{11.4}$$

where $\psi(\beta, \alpha) (= 1 + a)$ is a logarithm of a normalization term. It is computed as

$$\psi(\beta, \alpha) = \log \sum_{\mathbf{x}} e^{-\beta \mathscr{H}_0(\mathbf{x}) + \alpha X(\mathbf{x})}. \tag{11.5}$$

We call $\psi(\beta, \alpha)$ a log-partition function. The Lagrange multipliers, β and α, are adjusted such that $\langle \mathscr{H}_0(\mathbf{x}) \rangle = U$ and $\langle X(\mathbf{x}) \rangle = X$. Equation (11.4) is a softmax function (generalization of a logistic function to multinomial outputs) that returns the population output from a linear sum of the features weighted by $-\beta$ and α. With this view, we may alternatively regard β or α as an input parameter that controls U and X. Hereafter we simply call U internal activity, and X stimulus-related activity. Similarly, we call β an internal component, and α a stimulus component. We consider that the stimulus component α can be controlled by changing extrinsic stimulus conditions that an experimenter can manipulate. The stimulus component is written as $\alpha(s)$ if it is a function of a scalar stimulus condition s, such as stimulus contrast for visual neurons. In contrast, the internal component β is not directly controllable by the stimulus conditions. The spontaneous activity is modeled at $\beta = 1$ and $\alpha = 0$.

11.2.3 Gain Modulation by Internal Activity

We give a simple example of the maximum entropy model to show how the internal activity modulates the stimulus-related activity. Figure 11.1a illustrates an exemplary model composed of 5 neurons. With these particular model parameters (see figure caption), the stimulus component α controls activity rates of the first three neurons and their correlations. The internal component β controls background activity rates of all neurons. In our settings, decreasing β increases the baseline activity level of all neurons. Figure 11.1b displays activity rates of the individual neurons ($\langle x_i \rangle$ for $i = 1, \ldots, 5$) as a function of the stimulus component α with a fixed internal component β. Increasing α under these conditions activates the first three neurons without changing the activity rates of Neuron 4 and 5.[1] Furthermore, the response functions of the three neurons shift toward left when the background activity rates of all neurons is increased by *decreasing* the internal component β (Fig. 11.1b dashed lines). Thus Neuron 1–3 increase sensitivity to stimulus component α. This type of modulation is called input-gain control. For example, if α is a logarithmic function of contrast s of visual stimulation presented to an animal while recording visual neurons ($\alpha(s) = \log s$), increasing the modulation (decreasing β) makes neurons respond to multiplicatively smaller stimulus contrast. This models the contrast gain-control observed in visual pathways (Sakmann and Creutzfeldt 1969; Ohzawa et al. 1985; Reynolds et al. 2000; Martínez-Trujillo and Treue 2002). Other types of nonlinearity in the input-output relation can be constructed, depending on the nonlinearity in $\alpha(s)$.

[1] The activity rates of Neuron 4, 5 do not depend on α because $\mathbf{b}_0$ does not contain interactions that relate Neuron 1–3 with Neuron 4, 5. If there are non-zero interactions between any pair from Neuron 1–3 and Neuron 4, 5 in $\mathbf{b}_0$, the activity rates of Neuron 4, 5 increase with the increased rates of Neuron 1–3.

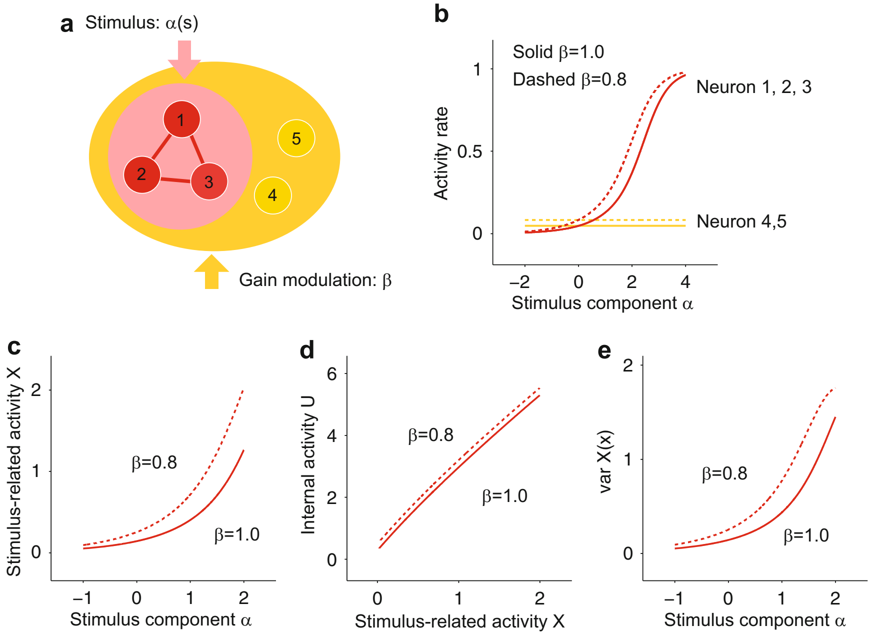

Fig. 11.1 A simple model of gain modulation by a maximum entropy model of 5 neurons. (**a**) An illustration of neurons that are activated by a stimulus (neurons in a pink area) and controlled by an internal mechanism (neurons in a yellow area). The model is constrained by features containing up to the second order statistics: $\mathbf{F}(\mathbf{x}) = (x_1, \ldots, x_5,\ x_1x_2, x_1x_3, x_2x_3, \ldots, x_4x_5)^\top$, where the first 5 elements are parameters for the individual activities x_i $(i = 1, \ldots, 5)$ and the rest of the elements is the joint activities of two neurons x_ix_j $(i < j)$. We assume that the stimulus-related activity is characterized by $\mathbf{b}_1 = (1, 1, 1, 0, 0,\ 0.3, 0.3, 0.3, 0, \ldots, 0)$. The first 3 elements are parameters for individual activity of the first three neurons x_i $(i = 1, 2, 3)$. The value 0.3 is assigned to the joint activities of the first three neurons, namely the features specified by x_1x_2, x_1x_3, and x_2x_3. The internal activity is characterized by $\mathbf{b}_0 = (2, 2, 2, 2, 2, 0, \ldots, 0)$, which regulates activity rates of individual neurons but does not change their interactions. (**b**) The activity rates of neurons as a function of the stimulus component α at fixed internal components, $\beta = 1.0$ (solid line) and $\beta = 0.8$ (dashed line). (**c**) The stimulus component X as a function of α at different internal components. (**d**) The relation between the stimulus-related activity X and internal activity U. (**e**) The Fisher information about the stimulus component α

Figure 11.1c displays a relation of the stimulus component α with the stimulus-related activity X at different internal component β. Similarly to the activity rates (Fig. 11.1b), the stimulus-related activity X is augmented if the internal component β is decreased. This nonlinear interaction between α and β is caused by the neurons that belong to both stimulus-related and internal activities. In this example, the stimulus component α also increases the internal activity U (Fig. 11.1d) because of increased activity rates of the shared neurons 1, 2, 3. Finally, Fig. 11.1e displays the variance of stimulus feature $X(\mathbf{x})$ as a function of α. It quantifies the information about the stimulus component α, which we will discuss in the next section.

11.3 The Conservation of Entropy, Equation of State, and Stimulus Information for a Neural Population

11.3.1 Conservation of Entropy for Neural Dynamics

The probability mass function, Eq. (11.4), belongs to the exponential family distribution. The Lagrange parameters are called natural or canonical parameters. The activity patterns of neurons are modeled as a linear combination of the two features $\mathscr{H}_0(\mathbf{x})$ and $X(\mathbf{x})$ using the canonical parameters $(-\beta, \alpha)$ in the exponent. Expectation of the features are called the expectation parameters U and X. Either natural or expectation parameters are sufficient to specify the probability distribution. We review dual structure of the two representations (Amari and Nagaoka 2000), and show that the relation provides the conservation law of entropy.

Negative entropy of the neural population is computed as

$$\begin{aligned} -S &= \langle \log p(\mathbf{x}) \rangle \\ &= -\beta \langle \mathscr{H}_0(\mathbf{x}) \rangle + \alpha \langle X(\mathbf{x}) \rangle - \psi(\beta, \alpha) \\ &= -U\beta + X\alpha - \psi(\beta, \alpha). \end{aligned} \tag{11.6}$$

Since the log-partition function of Eq. (11.4) is a cumulant generating function, U and X are related to the derivatives of $\psi(\beta, \alpha)$ as

$$\frac{\partial \psi(\beta, \alpha)}{\partial \beta} = -\langle \mathscr{H}_0(\mathbf{x}) \rangle = -U, \tag{11.7}$$

$$\frac{\partial \psi(\beta, \alpha)}{\partial \alpha} = \langle X(\mathbf{x}) \rangle = X. \tag{11.8}$$

Equations (11.6)–(11.8) form a Legendre transformation from $\psi(\beta, \alpha)$ to $-S(U, X)$. The inverse Legendre transformation is constructed using Eq. (11.6) as well: $\psi(\beta, \alpha) = -\beta U + \alpha X - (-S(U, X))$. Thus dually to Eqs. (11.7) and (11.8), the natural parameters are obtained as derivatives of the entropy with respect to the expectation parameters,

$$\left(\frac{\partial S}{\partial U} \right)_X = \beta, \tag{11.9}$$

$$\left(\frac{\partial S}{\partial X} \right)_U = -\alpha. \tag{11.10}$$

The natural parameters represent sensitivities of the entropy to the independent variables U and X. From these results, the total derivative of $S(U, X)$ is written as

$$\begin{aligned} dS &= \left(\frac{\partial S}{\partial U} \right)_X dU + \left(\frac{\partial S}{\partial X} \right)_U dX \\ &= \beta dU - \alpha dX. \end{aligned} \tag{11.11}$$

This explains a change of neurons' entropy by changes in the internal and stimulus-related activities. We denote an entropy change caused by the internal activity as $dS^{\text{int}} \equiv \beta dU$, and an entropy change caused by the extrinsic stimulus as $dS^{\text{ext}} \equiv \alpha dX$, respectively. Then Eq. (11.11) is written as

$$dS = dS^{\text{int}} - dS^{\text{ext}}. \tag{11.12}$$

We remark that dS is an infinitesimal difference of entropies at two close states, and its integral does not depend on a specific transition between the two states. In contrast, dS^{int} and dS^{ext} represent production of entropy separately by the internal and stimulus-related activities, and their integrals depend on the specific paths. Equation (11.12) constitutes the conservation of entropy for neural dynamics. We stress that although it is the first law of thermodynamics, the neurons considered here interact with an environment differently from conventional thermodynamic systems.[2] While internal energy of the conventional systems is indirectly controlled via work and heat, we consider that the internal activity of neurons is controlled directly by the organism's internal mechanism. Thus we use dS^{int} and dS^{ext}, rather than the work and heat, as quantities that neurons exchange with an environment.

11.3.2 Equation of State for a Neural Population

Equation (11.8) is an equation of the state for a neural population, which we rewrite here as

$$X(\beta, \alpha) = \frac{\partial \psi(\beta, \alpha)}{\partial \alpha}. \tag{11.13}$$

Through the log-partition function ψ, this equation relates state variables, β, α, and X, similarly to, e.g., the classical ideal gas law that relates temperature, pressure, and volume. Figure 11.1c displayed the equation of state. We note that ψ is related to the Gibbs free energy (see Appendix). Furthermore, without loss of generality, we can assume that the hamiltonian of the silent state is zero: $\mathscr{H}_0(\mathbf{0}) = X(\mathbf{0}) = 0$, where $\mathbf{x} = \mathbf{0}$ denotes the simultaneous silence of all neurons. We then obtain $p(\mathbf{0}) = e^{-\psi}$, namely

$$-\psi(\beta, \alpha) = \log p(\mathbf{0}). \tag{11.14}$$

[2] We obtain $dU = TdS - fdX$, using $\beta \equiv 1/T$ and $\alpha \equiv \beta f$ in Eq. (11.11). In this form, the expectation parameter U is a function of (S, X). According to the conventions of thermodynamics, we may call U internal energy, T temperature of the system, and f force applied to neurons by a stimulus. It is possible to describe the evoked activity of a neural population using these standard terms of thermodynamics. However, this introduces the concepts of work and heat, which may not be relevant quantities for neurons to exchange with environment.

Thus $-\psi(\beta, \alpha)$ is a logarithm of the simultaneous silence probability.[3] Since $d(\log p(\mathbf{0})) = dp(\mathbf{0})/p(\mathbf{0})$, $-d\psi$ gives a fractional increase of the simultaneous silence probability of the neurons. Accordingly, Eq. (11.13) states that the stimulus-related activity X equals to the fractional decrease of the simultaneous silence probability by a small change of α, given β.

The opposite representation of the equation of state, α as a function of X given β, is obtained as follows. From Eq. (11.13), we have $d\psi = X d\alpha$ given that β is fixed. Let ψ_0 and X_0 be ψ and X at $\alpha = 0$. Then, if the internal component β is fixed, the stimulus component α at X is given by

$$\alpha(\beta, X) = \int_{\psi_0}^{\psi} \left(\frac{1}{X}\right)_{\beta} d\psi' = \int_{X_0}^{X} \left(\frac{1}{X'} \frac{\partial \psi}{\partial X'}\right)_{\beta} dX'. \tag{11.15}$$

Here $\left(\frac{\partial \psi}{\partial X}\right)_{\beta}$ is a fractional decrease of the simultaneous silence probability when X shifts to $X + dX$ while β is fixed.

11.3.3 Information About Stimulus

The Fisher information $J(\alpha)$ provides the accuracy of estimating a small change in the stimulus component α by an optimal decoder. More specifically, the inverse of the Fisher information provides a lower bound of variance of an unbiased estimator for α from a sample. For the exponential family distribution, it is given as the second order derivative of the log-partition function with respect to α, which is also the variance of stimulus feature $X(\mathbf{x})$:

$$\begin{aligned} J(\alpha) \equiv \left\langle \left(\frac{\partial \log p(\mathbf{x})}{\partial \alpha} \right)^2 \right\rangle &= \frac{\partial^2 \psi(\beta, \alpha)}{\partial \alpha^2} \\ &= \frac{\partial X}{\partial \alpha} = \langle X(\mathbf{x})^2 \rangle - \langle X(\mathbf{x}) \rangle^2. \end{aligned} \tag{11.16}$$

The first equality in the second line of Eq. (11.16) is obtained using the first order derivative of ψ, namely the equation of state (Eq. (11.13)). The second equality in Eq. (11.16) represents the fluctuation-dissipation relation of the stimulus feature. The equalities show that the Fisher information can be computed in three different manners given that the internal component β is fixed: (1) the second derivative of

[3] Importantly, $-\psi$ is a logarithm of the simultaneous silence probability predicted by the model, Eq. (11.4). The observed probability of the simultaneous silence could be different from the prediction if the model is inaccurate. For example, an Ising model may be inaccurate, and it was shown that neural higher-order interactions may significantly contribute to increasing the silence probability (Ohiorhenuan et al. 2010; Shimazaki et al. 2015).

ψ with respect to α using the simultaneous silence probability, (2) the derivative of X with respect to α using the equation of state, or (3) the variance of the stimulus feature.

The Fisher information computed at two fixed internal components was shown in Fig. 11.1e. The stimulus component α becomes relatively dominant in characterizing the neural activity if the internal component β decreases. This results in the larger Fisher information $J(\alpha)$ for the smaller internal component β at given α. If the stimulus condition s controls the stimulus component as $\alpha(s)$, and it is not related to β, the information about s is given as $\frac{\partial \alpha(s)}{\partial s} J(\alpha) \frac{\partial \alpha(s)}{\partial s}$.

11.4 Information-Theoretic Cycles by a Neural Population

We now introduce neural dynamics that models dynamical gain-modulation performed by an internal mechanism while neurons are processing stimulus. Since there are neurons that belong to both stimulus-related and internal activities, the internal mechanism changes not only the internal activity but also the stimulus-related activity, which realizes the modulation. From an information-theoretic point of view, this process converts entropy generated by the internal mechanism to entropy associated with stimulus-related activity after one cycle of the neural response is completed. To explain this in detail, we first provide an intuitive example of delayed gain-modulation using a dynamical model, and then provide an ideal cycle that efficiently enhances stimulus information. Using the latter model, we explain why the process works similarly to a heat engine, and show how to quantify efficiency of the gain-modulation performed by the internal mechanism.

11.4.1 An Example of Delayed Gain-Modulation

We first consider a simple dynamical model of delayed gain-modulation. We use the feature vector, $\mathbf{b}_0$ and $\mathbf{b}_1$ based on those used in Fig. 11.1. In this model, neurons are activated by a stimulus input, which subsequently increases modulation by an internal mechanism (Fig. 11.2a). Such a process can be modeled through dynamics of the controlling parameters given by

$$\tau_\alpha \dot{\alpha}(t) = -\alpha(t) + s\, e^{-t/\tau_\alpha} \tag{11.17}$$

$$\tau_\beta \dot{\beta}(t) = -\beta(t) + \beta_0 - \gamma \alpha(t) \tag{11.18}$$

for $t \geq 0$. Here s is intensity of an input stimulus. Neurons are initially at a spontaneous state: $\alpha(0) = 0$ and $\beta(0) = \beta_0 = 1$. The top panel of Fig. 11.2b displays the dynamics of $\alpha(t)$ and $\beta(t)$. The population activity is sampled from

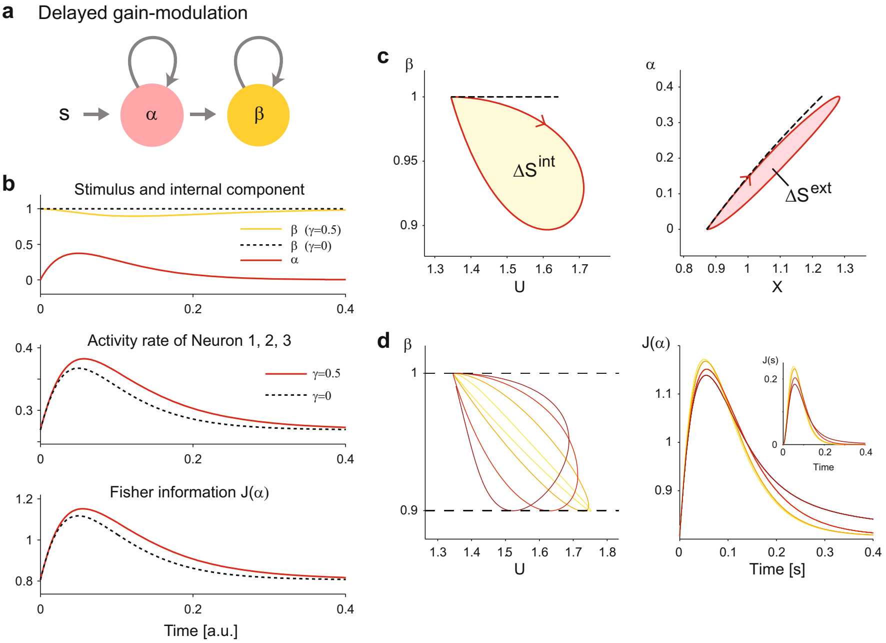

Fig. 11.2 The delayed gain-modulation by internal activity. The parameters of the maximum entropy model ($N = 5$) follow those in Fig. 11.1. (**a**) An illustration of delayed gain-modulation described in Eqs. (11.17) and (11.18). The stimulus increases the stimulus component α that activates Neuron 1, 2, and 3. Subsequently, the internal component β is increased, which increases the background activity of all 5 neurons. We assume a slower time constant for the gain-modulation than the stimulus activation ($\tau_\beta = 0.1$ and $\tau_\alpha = 0.05$). (**b**) *Top:* Dynamics of the stimulus and internal components (solid lines, $\gamma = 0.5$). The internal component β without the delayed gain-modulation ($\gamma = 0$) is shown by a dashed black line. *Middle:* Activity rates [a.u.] of Neuron 1–3 with (solid red) and without (dashed black) the delayed gain-modulation. *Bottom:* The Fisher information about stimulus component α (Eq. (11.16)). (**c**) The X-α (left) and U-β (right) phase diagrams. A red solid cycle represents dynamics when the delayed gain-modulation is applied ($\gamma = 0.5$). The dashed line is a trajectory when the delayed gain-modulation is not applied to the population ($\gamma = 0$). (**d**) *Left:* The U-β phase diagrams of neural dynamics with different combinations of τ_β and γ that achieve the same level of the maximum modulation (the minimum value of $\beta = 0.9$). *Right:* The Fisher information about the stimulus component α for different cycles. The color code is the same as in the left panel. The inset shows the Fisher information about the stimulus intensity s (Eq. (11.19))

the maximum entropy model with these dynamical parameters. Here we consider a continuous-time representation of the maximum entropy model[4] (Kass et al. 2011;

[4] Under the assumption that rates of synchronous spike events scale with $\mathcal{O}(\Delta^k)$, where Δ is a bin size of discretization and k is the number of synchronous neurons. It was proved (Kass et al. 2011) that it is possible to construct a continuous-time limit ($\Delta \to 0$) of the maximum entropy model that takes the synchronous events into account. Here we follow their result to consider the continuous-time representation.

Kelly and Kass 2012). The activity rates of neurons are increased by the delayed gain-modulation (solid lines in Fig. 11.2b, middle panel) from those obtained without the modulation ($\gamma = 0$; dashed lines). Accordingly, the information about the stimulus component α contained in the population activity as quantified by the Fisher information (Eq. (11.16)) increases and lasts longer by the delayed gain-modulation (Fig. 11.2b, bottom panel). Note that in this example, the information about the stimulus strength s is carried in both $\beta(t)$ and $\alpha(t)$ as time passes. The result obtained from the Fisher information about s using both $\beta(t)$ and $\alpha(t)$ is qualitatively the same as the result of the Fisher information about α (not shown).[5]

The U-β phase diagram (Fig. 11.2c, left panel) shows that dynamics without the gain-modulation is represented as a line because β is constant. In contrast, dynamics with the gain-modulation forms a cycle because weaker and then stronger modulation (larger and then smaller β) is applied to neurons when the internal activity U increases and then decreases, respectively. Similarly, the dynamics forms a cycle in the X-α plane (Fig. 11.2c, right panel) if the stimulus activity X is augmented by the delayed gain-modulation. By applying the conservation law for entropy (Eq. (11.12)) to the cycle, we obtain

$$0 = \oint \beta dU - \oint \alpha dX. \tag{11.20}$$

Here $\oint \beta dU \equiv \Delta S^{\text{int}}$ is entropy produced by the internal activity during the cycle due to the delayed gain-modulation, and $\oint \alpha dX \equiv \Delta S^{\text{ext}}$ is entropy produced by the activity related to extrinsic stimulus conditions. These are the areas within the circles in the phase diagrams. Equation (11.20) states that the two cycles have the same area ($\Delta S^{\text{int}} = \Delta S^{\text{ext}}$).

The left panel in Fig. 11.2d displays the U-β phase diagram for dynamics with given maximum strength of modulation (the minimum value of β). Among these cycles, larger cycles retain the information about the stimulus component α for a longer time period (Fig. 11.2d, right panel). The same conclusion is made from the Fisher information about s (Fig. 11.2d, an inset in right panel). The larger cycles were made because the modulation was only weakly applied to neurons when the internal activity U increased, then the strong modulation was applied when U decreased. Such modulation is considered to be efficient because it allows neurons to retain the stimulus information for a longer time period by using the slow time-scale of β without excessively increasing activity rates of neurons at its initial rise. In the

[5] When α and β are both dependent on the stimulus, the Fisher information about s is given as

$$J(s) = \frac{\partial \boldsymbol{\theta}(s)^{\top}}{\partial s} \mathbf{J} \frac{\partial \boldsymbol{\theta}(s)}{\partial s}, \tag{11.19}$$

where $\boldsymbol{\theta}(s) \equiv [-\beta, \alpha]^{\top}$ and $\mathbf{J}$ is a Fisher information matrix given by Eq. (11.24), which will be discussed in the later section. We computed Eq. (11.19) using analytical solutions of the dynamical equations given as $\alpha(t) = \frac{st}{\tau_\alpha} e^{-t/\tau_\alpha}$ and $\beta(t) = 1 - \frac{s\gamma}{\tau_\beta - \tau_\alpha} \left\{ \frac{\tau_\alpha \tau_\beta}{\tau_\beta - \tau_\alpha} (e^{-t/\tau_\beta} - e^{-t/\tau_\alpha}) - te^{-t/\tau_\alpha} \right\}$.

next section, we introduce the largest cycle that maximizes the entropy produced by the gain-modulation when the maximum strength of the modulation is given. Using this cycle, we explain how the cycle works analogously to a heat engine, and define efficiency of the cycle to retain the stimulus information.

11.4.2 The Efficient Cycle by a Neural Population

The largest cycle is made if the modulation is not applied when the internal activity U increases, then applied when U decreases. Figure 11.3 displays a cycle of hypothetical neural dynamics that maximizes the entropy production when the ranges of the internal component and activity are given. The model parameters follow those in Fig. 11.1. This cycle is composed of four steps. The process starts at the state A at which neurons exhibit spontaneous activity ($\beta = \beta_H = 1, \alpha = 0$). Figure 11.3a displays a sample response of the neural population to a stimulus change. Figure 11.3b and c display the X-α and U-β phase diagrams of the cycle. Heat capacity of the neural population and the Fisher information about α are shown in Fig. 11.3d. Details of the cycle steps are now described as follows.

A→B **Increased stimulus response** The stimulus-related activity X is increased by increasing the stimulus component α while the internal component is fixed at $\beta = \beta_H$. In this process the internal activity U also increases.

B→C **Internal computation** An internal mechanism decreases the internal component β while keeping the internal activity ($dU = 0$). In this process the stimulus-related activity X decreases. The process ends at $\beta = \beta_L$.

C→D **Decreased stimulus response** The stimulus-related activity X is decreased by decreasing the stimulus component α while the internal component is fixed at $\beta = \beta_L$. In this process the internal activity U also decreases.

D→A **Internal computation** An internal mechanism increases the internal component β while keeping the internal activity ($dU = 0$). In this process the stimulus-related activity X increases. The process ends at $\beta \equiv \beta_H$.

The processes B→C and D→A represent additional computation performed by an internal neural mechanism on the neurons' stimulus information processing. It is applied after the initial increase of stimulus-related activity during A→B, therefore manifests delayed modulation. Without these processes, the neural dynamics is represented as a line in the phase diagrams. The Fisher information about α also increases during the process between C and D (Fig. 11.3d, right panel). We reiterate that the Fisher information quantifies the accuracy of estimating a small change in α by an optimal decoder. Thus operating along the path between C and D is more advantageous than the path between A and B for downstream neurons if their goal is to detect a change in the stimulus-related activity of the upstream neurons that is not explained by the internal activity.

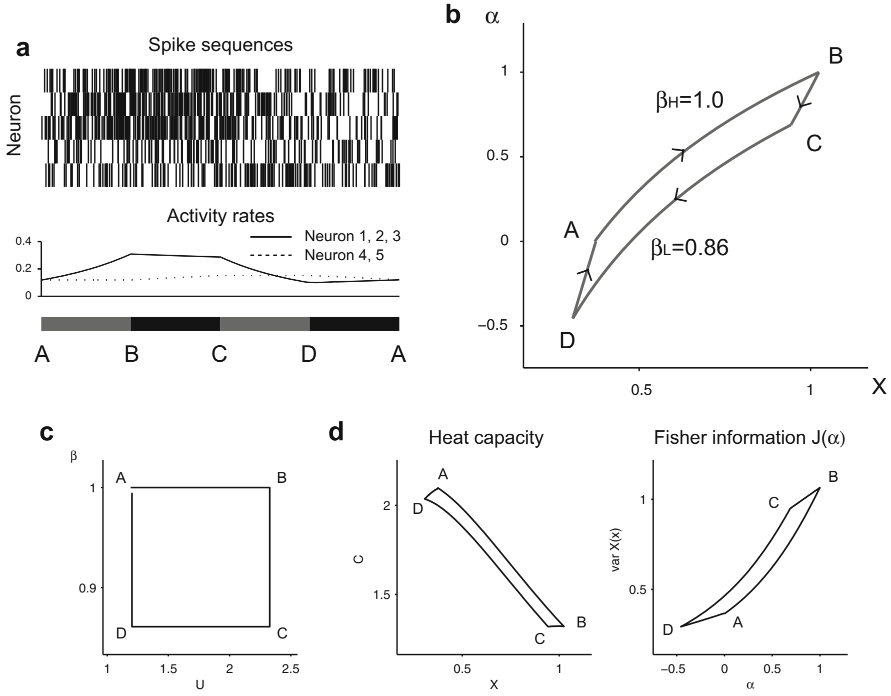

Fig. 11.3 The efficient circle by a neural population ($N = 5$). The parameters of the maximum entropy model follow those in Fig. 11.1. The cycle starts from the state A at which $\beta = \beta_H = 1$ and $\alpha = 0$. See the main text for details of the steps. The efficiency of this cycle is 0.14. (**a**) *Top:* Spike raster plots during the cycle. *Middle:* Activity rates of neurons. *Bottom:* The cycle steps. (**b**) The X-α phase diagram. (**c**) The U-β phase diagram. (**d**) *Left:* X v.s. heat capacity. The heat capacity is defined as $C = \langle h^2 \rangle - \langle h \rangle^2$, where $h = -\log p(\mathbf{x})$ is information content. *Right:* Fisher information about the stimulus component α

11.4.3 Interpretation as an Information-Theoretic Cycle

We start our analysis on the cycle by examining how much entropy is generated by the internal and stimulus-related activities at each step. First, we denote by $\Delta S^{\text{int}}_{\text{AB}}$ and $\Delta S^{\text{int}}_{\text{CD}}$ the entropy changes caused by the internal activity during the process A→B and C→D, respectively. Since the internal component β is fixed at β_H during the process A→B, we obtain $\Delta S^{\text{int}}_{\text{AB}} = \beta_H \Delta U$, where ΔU is a change of the internal activity (see Fig. 11.3c). This change in the internal activity is positive ($\Delta U > 0$). Since the internal activity does not change during B→C and D→A, a change of the internal activity during C→D is given by $-\Delta U$ (Note that the internal activity is a state variable). We obtain $\Delta S^{\text{int}}_{\text{CD}} = -\beta_L \Delta U$ for the process during C→D. The total entropy change caused by the internal activity during the cycle is given as $\Delta S^{\text{int}}_{\text{AB}} + \Delta S^{\text{int}}_{\text{CD}} = (\beta_H - \beta_L)\Delta U$, which is positive because $\beta_H > \beta_L$ and $\Delta U > 0$. Thus the internal activity contributes to increasing the entropy of neurons during the

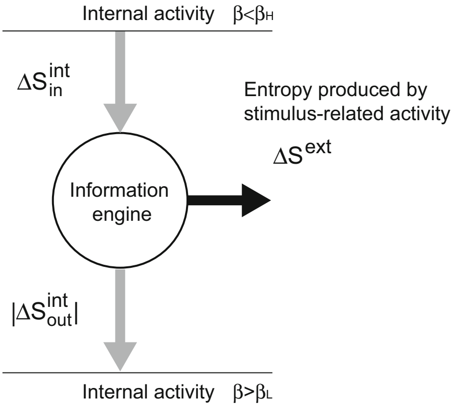

Fig. 11.4 An information-theoretic cycle by a neural population

cycle. Second, we denote by ΔS^{ext} the total entropy change caused by the stimulus-related activity during the cycle. According to the conservation law (Eq. (11.12)) applied to this cycle, we obtain

$$0 = \Delta S^{\text{int}}_{\text{AB}} + \Delta S^{\text{int}}_{\text{CD}} - \Delta S^{\text{ext}}. \tag{11.21}$$

Note that the sign of $\Delta S^{\text{ext}} = \Delta S^{\text{int}}_{\text{AB}} + \Delta S^{\text{int}}_{\text{CD}}$ is positive. Hence the stimulus-related activity contributes to decreasing the entropy of neurons during the cycle.

This cycle belongs to the following cycle that is analogous to a heat engine (Fig. 11.4). In this paragraph, we temporarily use *receive entropy* and *emit entropy* to express the positive and negative path-dependent entropy changes caused by the internal or stimulus-related activity in order to facilitate comparison with a heat engine.[6] In this cycle, neurons receive *entropy* as internal activity from an environment ($\Delta S^{\text{int}}_{\text{in}} > 0$) and emit *entropy* to the environment ($\Delta S^{\text{int}}_{\text{out}} < 0$). The received *entropy* as the internal activity is larger than the emitted *entropy* ($\Delta S^{\text{int}}_{\text{in}} + \Delta S^{\text{int}}_{\text{out}} > 0$). The surplus *entropy* is emitted to the environment in the form of the stimulus-related activity ($-\Delta S^{\text{ext}} < 0$). Thus we may regard the cycle as the process that produces stimulus-related entropy using entropy supplied by the internal mechanism. We hereafter denote this cycle as an information-theoretic cycle, or engine. The cycle in Fig. 11.2 is also regarded as an information-theoretic cycle by separating the process at which the internal activity is maximized. The conservation law prohibits a perpetual information-theoretic cycle that can indefinitely produce the stimulus-related entropy without entropy production by the internal mechanism.[7]

[6] Here we use *entropy* synonymously with heat in thermodynamics to facilitate the comparison with a heat engine. However this is not an accurate description because the entropy is a state variable.

[7] This is synonymous with the statement that the first law prohibits a perpetual motion machine of the first kind, a machine that can work indefinitely without receiving heat.

11.4.4 Efficiency of a Cycle

As we discussed for the example dynamics in Fig. 11.2, we may consider that the modulation is efficient if it helps neurons to retain stimulus information without excessively increasing the internal and stimulus-related activities during the initial response. Such a process was achieved when gain-modulation was only weakly applied to neurons when the internal activity U increased, then strong gain modulation was applied when U decreased. We can formally assess this type of efficiency by defining entropic efficiency, similarly to thermal efficiency of a heat engine. It is given by a ratio of the entropy change caused by the stimulus-related activity as opposed to the entropy change gained by the internal activity as:

$$\eta \equiv \frac{\Delta S^{\text{ext}}}{\Delta S^{\text{int}}_{\text{in}}} = 1 - \frac{|\Delta S^{\text{int}}_{\text{out}}|}{\Delta S^{\text{int}}_{\text{in}}}. \tag{11.22}$$

For the proposed information-theoretic cycle in Fig. 11.3, it is computed as

$$\eta_e = 1 - \frac{|\Delta S^{\text{int}}_{\text{CD}}|}{\Delta S^{\text{int}}_{\text{AB}}} = 1 - \frac{\beta_L}{\beta_H}, \tag{11.23}$$

which is a function of the internal components, β_H and β_L. This cycle is the most efficient in terms of the entropic efficiency defined by Eq. (11.22) when the highest and lowest internal components and activities are given. The square cycle in the U-β phase diagram (Fig. 11.3c) already suggests this claim, and we can formally prove this by comparing the information-theoretic cycle with an arbitrary cycle $\mathscr{C}$ whose internal component β satisfies $\beta_L \leq \beta \leq \beta_H$.[8] Thus the proposed cycle bounds efficiency of the additional computation made by the delayed gain-modulation mechanism. Here we now call the proposed cycle in Fig. 11.3, the ideal information-theoretic cycle. Note that this cycle is similar to, but different from the Carnot cycle (Carnot 1824) that can be realized by replacing the processes B→C and D→A with adiabatic processes. The Carnot cycle achieves the highest *thermal* efficiency.

[8] Let us consider the efficiency η achieved by an arbitrary cycle $\mathscr{C}$ during which the internal component β satisfies $\beta_L \leq \beta \leq \beta_H$. Let the minimum and maximum internal activity in the cycle be $U_{\min}$ and $U_{\max}$. We decompose $\mathscr{C}$ into the path $\mathscr{C}_1$ from $U_{\min}$ to $U_{\max}$ and the path $\mathscr{C}_2$ from $U_{\max}$ to $U_{\min}$ during which the internal component is given as $\beta_1(U)$ and $\beta_2(U)$, respectively. Because the cycle acts as an engine, we expect $\beta_1(U) > \beta_2(U)$. The entropy changes produced by the internal activity during the path C_i $(i = 1, 2)$ is computed as $\Delta S^{\text{int}}_{\mathscr{C}_1} = \int_{U_{\min}}^{U_{\max}} \beta_1(U)\, dU \leq \beta_H \int_{U_{\min}}^{U_{\max}} dU = \beta_H(U_{\max} - U_{\min})$ and $|\Delta S^{\text{int}}_{\mathscr{C}_2}| = |\int_{U_{\max}}^{U_{\min}} \beta_2(U)\, dU| \geq |\beta_L \int_{U_{\max}}^{U_{\min}} dU| = \beta_L(U_{\max} - U_{\min})$. Hence we obtain $|\Delta S^{\text{int}}_{\mathscr{C}_2}|/\Delta S^{\text{int}}_{\mathscr{C}_1} \geq \beta_L/\beta_H$, or $\eta \leq \eta_e$.

11.4.5 Geometric Interpretation

Finally, we introduce geometric interpretation of the cycle, and consider conditions that realize the information-theoretic cycle. Let us denote the internal and stimulus components as $\boldsymbol{\theta} = [-\beta, \alpha]^{\top}$. In addition, we represent the expected internal and stimulus features by $\boldsymbol{\eta} = [U, X]^{\top}$. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ form dually flat affine coordinates, and are called θ and η-coordinates in information geometry (Amari and Nagaoka 2000).

A small change in $\boldsymbol{\theta}$ is related to a change in $\boldsymbol{\eta}$ as $d\boldsymbol{\eta} = \mathbf{J} d\boldsymbol{\theta}$. Here $\mathbf{J}$ is the Fisher information matrix with respect to $\boldsymbol{\theta}$. It is given as

$$\mathbf{J} = \begin{bmatrix} \langle \mathbf{b}_0, \mathbf{b}_0 \rangle & \langle \mathbf{b}_0, \mathbf{b}_1 \rangle \\ \langle \mathbf{b}_1, \mathbf{b}_0 \rangle & \langle \mathbf{b}_1, \mathbf{b}_1 \rangle \end{bmatrix}, \tag{11.24}$$

where $\langle \mathbf{b}_i, \mathbf{b}_j \rangle \equiv \mathbf{b}_i^{\top} \mathbf{G} \mathbf{b}_j$ $(i, j = 0, 1)$ is an inner product of the vectors $\mathbf{b}_i$ and $\mathbf{b}_j$ with a metric given by $\mathbf{G} = \langle \mathbf{F}(\mathbf{x}) \mathbf{F}(\mathbf{x})^{\top} \rangle - \langle \mathbf{F}(\mathbf{x}) \rangle \langle \mathbf{F}(\mathbf{x}) \rangle^{\top}$. Note that $\langle \mathbf{b}_0, \mathbf{b}_0 \rangle$ is equivalent to Eq. (11.16). In general, in order to make a change of the internal component β influence the stimulus-related activity X, therefore controls stimulus information, one requires $\langle \mathbf{b}_0, \mathbf{b}_1 \rangle \neq 0$ because $dX = -\langle \mathbf{b}_1, \mathbf{b}_0 \rangle d\beta + \langle \mathbf{b}_1, \mathbf{b}_1 \rangle d\alpha$ from $d\boldsymbol{\eta} = \mathbf{J} d\boldsymbol{\theta}$. This condition indicates that the modulation by an internal mechanism is achieved through the activity features shared by the two components. Accordingly, this condition is violated if neurons participate in the stimulus-related activity and neurons subject to the internal modulation do not overlap (namely if neurons that appear in the features corresponding to non-zero elements of $\mathbf{b}_0$ are separable from those of $\mathbf{b}_1$).

For the ideal information-theoretic cycle, we indicate the parameters at A, B, C, and D using a subscript of $\boldsymbol{\theta}$ or $\boldsymbol{\eta}$. For example, the parameters at A are $\boldsymbol{\theta}_A$ and $\boldsymbol{\eta}_A$. The first process A→B of the ideal information-theoretic cycle is a straight line (geodesic) between $\boldsymbol{\theta}_{\mathrm{A}}$ and $\boldsymbol{\theta}_{\mathrm{B}}$ in the curved space of θ-coordinates. It is called e-geodesic. In addition, the internal component β is fixed while the stimulus component decreases, therefore the e-geodesic is a vertical line in the θ-coordinates. The second process B→C is the shortest line between $\boldsymbol{\eta}_{\mathrm{B}}$ and $\boldsymbol{\eta}_{\mathrm{C}}$ in the curved space of η-coordinates. The path is called an m-geodesic. In addition, the internal activity U is fixed while the stimulus-related activity decreases, therefore the m-geodesic is a vertical line in the η-coordinates. Similarly, the process C→D is an e-geodesic, and the process D→A is an m-geodesic.

The change in the internal component β during the processes along m-geodesic manifested the internal computation in the ideal information-theoretic cycle. The small change in $\boldsymbol{\eta}$ is related to the change in $\boldsymbol{\theta}$ by $d\boldsymbol{\theta} = \mathbf{J}^{-1} d\boldsymbol{\eta}$. Since the m-geodesic processes B→C and D→A are characterized by $d\boldsymbol{\eta} = [0, dX]^{\top}$, the small change in θ-coordinates is given as

$$d\boldsymbol{\theta} = \begin{bmatrix} -\langle \mathbf{b}_0, \mathbf{b}_1 \rangle \\ \langle \mathbf{b}_0, \mathbf{b}_0 \rangle \end{bmatrix} |\mathbf{J}|^{-1} dX, \tag{11.25}$$

Conversely, the internal mechanism needs to change the internal and stimulus component according to the above gradient in order to accomplish the most efficient cycle. Thus the internal mechanism need to access the stimulus component α in order to realize the ideal information-theoretic cycle. Again, if $\langle \mathbf{b}_0, \mathbf{b}_1 \rangle = 0$, the internal component β is not allowed to change, which however means that the entire process does not form a cycle. Therefore we impose $\langle \mathbf{b}_0, \mathbf{b}_1 \rangle \neq 0$.

11.5 Discussion

In this study, we provided hypothetical neural dynamics that efficiently encodes stimulus information with the aid of delayed gain-modulation by an internal mechanism, and demonstrated that the dynamics forms an information-theoretic cycle that acts similarly to a heat engine. This view provided us to quantify the efficiency of the gain-modulation in retaining the stimulus information. The ideal information-theoretic cycle introduced here bounded the entropic efficiency.

As an extension of a logistic activation function of a single neuron to multinomial outputs, the maximum entropy model explains probabilities of activity patterns by a softmax function of the features, therefore allows nonlinear interaction of the inputs (here β and α) in producing the stimulus-related activity X (Fig. 11.1). This interaction was caused by shared activity features in $\mathbf{b}_1$ and $\mathbf{b}_0$. The gain modulation more effectively changes the stimulus-related activity if the features of the stimulus-related and internal activities resemble (i.e., $\langle \mathbf{b}_1, \mathbf{b}_0 \rangle$ is close to 1), which may have implications in similarity between evoked and spontaneous activities (Kenet et al. 2003) that can be acquired during development (Berkes et al. 2011).

The model's statistical structure common to thermodynamics (the Legendre transformation; see Appendix) allowed us to construct the first law for neural dynamics (Eq. (11.12)), the equation of state (Eq. (11.13)), fluctuation-dissipation relation (Eq. (11.16)), and neural dynamics similar to a thermodynamic cycle (Figs. 11.2 and 11.3) although we emphasized the differences from conventional thermodynamics in terms of the controllable quantities. The dynamics forms a cycle if the gain modulation is applied after the initial increase of the stimulus-related activity. This scenario is expected when the stimulus response is modulated by a feedback mechanism of recurrent networks (Salinas and Abbott 1996; Spratling and Johnson 2004; Sutherland et al. 2009), and is associated with short-term memory of the stimulus (Salinas and Abbott 1996; Salinas and Sejnowski 2001; Supèr et al. 2001). Consistently with the idea of efficient stimulus-encoding by a cycle, effect of attentional modulation on neural response typically appears several hundred milliseconds after stimulus onset (later than the onset of the stimulus response) (Motter 1993; Luck et al. 1997; McAdams and Maunsell 1999; Seidemann and Newsome 1999; Reynolds et al. 2000; Ghose and Maunsell 2002) although the temporal profile can be altered by task design (Luck et al. 1997; Ghose and Maunsell 2002). Further, the modulation of late activity components is ubiquitously observed in different neural systems (Cauller and Kulics 1991; Supèr et al. 2001; Sachidhanandam et al. 2013; Manita et al. 2015; Schultz 2016).

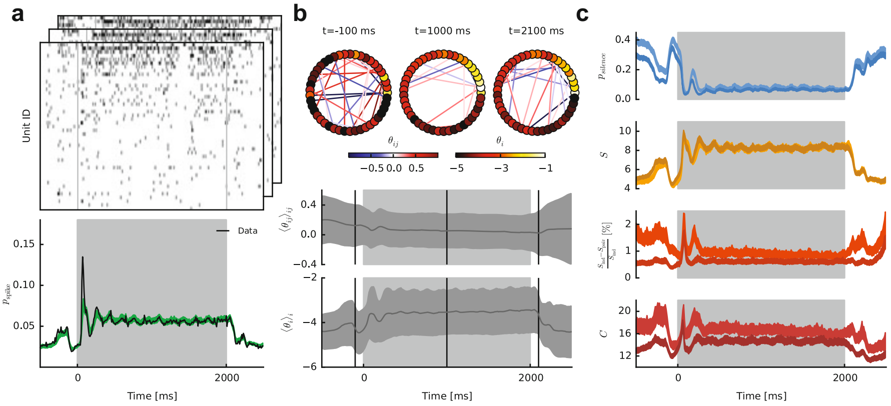

Fig. 11.5 The state-space method for estimating time-varying Ising model for monkey V4 data. (**a**) *Top:* Simultaneously recorded spiking data from 45 neurons while grating stimulus is presented to a monkey. *Bottom:* spiking probability (black, data; green, model fit). Gray area indicates the period of stimulus presentation. (**b**) *Top:* Time-varying parameters of an Ising model (i.e., individual and pairwise interaction parameters) are estimated by fitting the state-space model using an expectation-maximization (EM) algorithm. *Bottom:* the means and standard deviations of the Ising parameters. (**c**) Estimated dynamics of thermodynamic quantities (from top to bottom: silence probability, entropy, fractional entropy for correlations, heat capacity). The figure is modified from (Donner et al. 2017)

To test the hypothesis that neurons act as an information-theoretic engine using empirical data, the internal and stimulus feature need to be specified. Since even spontaneous neural activity is known to exhibit ongoing dynamics (Kenet et al. 2003), estimation of these features is nontrivial. The optimal sequential Bayesian algorithms have been proposed to smoothly estimate the parameters of the neural population model when they vary in time (Shimazaki et al. 2009, 2012; Shimazaki 2013; Donner et al. 2017), based on the paradigm developed by Brown and colleagues (Brown et al. 1998; Smith and Brown 2003) for joint estimation of the state-space and parameter estimation for point process observations. With the recent advances in applying various approximation methods to this model, it was demonstrated that the method is applicable to simultaneously analyzing a large number of neurons, and trace dynamics of thermodynamic quantities of the network such as the free energy, entropy, and heat capacity (Donner et al. 2017) (see Fig. 11.5). Hence this and similar approaches can be used to select dominant features of spontaneous and evoked activities, and then to estimate the time-varying internal and stimulus-related components. Efficiency of the cycles computed from the data can be used to test the hypothesis that the neurons are working as an information-theoretic engine. Further, by including multiple stimulus features in the model, the theory is expected to make quantitative predictions on competitive mechanisms of selective attention (Moran and Desimone 1985; Motter 1993; Luck et al. 1997; Reynolds et al. 1999). The conservation law of entropy imposes competition among the stimuli given a limited entropic resource generated by the internal mechanism.

The current theory assumes a quasi-static process for a neural response as we use an equilibrium model of the neural population at each point of time. For this to be a good approximation of neural dynamics, network activity caused by stimulus presentation may need to change more slowly than the time-scale of individual neurons under the examination, which may be expected as several tens of milliseconds for cortical neurons based on synaptic and membrane time constants and axonal delays. Otherwise, the theory needs to be extended to account for non-equilibrium processes by considering causal relations of past population activity on a current state of the population. It is possible to include the history effect on the population activity in the model (Shimazaki et al. 2012) or by using non-equilibrium models such as a kinetic Ising model. It will be an important challenge to consider a thermodynamic paradigm for a neural population including the second law for such non-equilibrium processes based on the recent advances in the field, where the second law of thermodynamics was generalized for a causal system with feedback (Sagawa and Ueda 2010, 2012; Ito and Sagawa 2013, 2015).

In summary, a neural population that works as an information-theoretic engine produces entropy ascribed to stimulus-related activity out of entropy supplied by an internal mechanism. This process is expected to appear during stimulus response of neurons subject to feedback gain-modulation. It is thus hoped that quantitative assessment of the neural dynamics as an information-theoretic engine contributes to understanding neural computation performed internally in an organism.

Acknowledgements This chapter is an extended edition of the manuscript submitted to the arXiv (Shimazaki H., Neurons as an information-theoretic engine. arXiv:1512.07855, 2015). The author thanks J. Gaudreault, C. Donner, D. Hirashima, S. Koyama, S. Amari, and S. Ito for critically reading the original manuscript.

Appendix: Free Energies of Neurons

In this appendix, we introduce thermodynamic formulation and free energies of a neural population. Let us first discuss the relation of state variables and free energies that appear in our analysis of the neural population with those found in conventional thermodynamics. Assume that the small change in internal activity of neurons has the following linear relations to entropy S, expected feature X, and the number of neurons N:

$$dU = TdS + fdX + \mu dN. \tag{11.26}$$

Equation (11.26) is the first law of thermodynamics, and the parameters are temperature T, force f, and chemical potential μ. The first law describes the internal activity as a function of (S, X, N). In thermodynamics, the Helmholtz free energy $F = U - TS$, Gibbs free energy $G = F - fX$, or enthalpy $H = U - fX$ is introduced to change the independent variables to (T, X, N), (T, f, N), and (S, f, N), respectively.

These free energies are useful to analyze isothermal or other processes in which only one of the independent variables is changed. For example, the Helmholtz free energy can be used to compute the work done by force f under the isothermal condition. However, the concepts of the force and work may not be directly relevant to information-theoretic analysis of a neural population. Here we introduce the free energies that are more consistent with the framework based on entropy changes.

The first law is alternatively written as

$$dS = \beta dU - \alpha dX - \gamma dN, \tag{11.27}$$

Here we used $\beta = 1/T$, $\alpha = f/T$, and $\gamma = \mu/T$. This first law describes a small entropy change as a function of (U, X, N). The parameters are defined as

$$\beta(U,X,N) = \left(\frac{\partial S}{\partial U}\right)_{X,N}, \tag{11.28}$$

$$\alpha(U,X,N) = -\left(\frac{\partial S}{\partial X}\right)_{N,U}, \tag{11.29}$$

$$\gamma(U,X,N) = -\left(\frac{\partial S}{\partial N}\right)_{U,X}. \tag{11.30}$$

We change the independent variable U to β. For this goal, here we define the *scaled* Helmholtz free energy $\mathscr{F}$ as

$$\mathscr{F} = S - \beta U. \tag{11.31}$$

Note that $\mathscr{F} = -\beta F$. It is a function that changes the independent variables from (S, X, N) to (β, X, N). This can be confirmed from the total derivative of $\mathscr{F}$: $d\mathscr{F} = dS - d(\beta U) = -Ud\beta - \alpha dX - \gamma dN$. From this equation, we have

$$U(\beta,X,N) = -\left(\frac{\partial \mathscr{F}}{\partial \beta}\right)_{X,N}, \tag{11.32}$$

$$\alpha(\beta,X,N) = -\left(\frac{\partial \mathscr{F}}{\partial X}\right)_{N,\beta}, \tag{11.33}$$

$$\gamma(\beta,X,N) = -\left(\frac{\partial \mathscr{F}}{\partial N}\right)_{\beta,X}. \tag{11.34}$$

The entropy change caused by the stimulus-related activity when X changes from X_1 to X_2 is given by the area under the curve of $\alpha(\beta, X, N)$ in the X-α phase plane. From Eq. (11.33), if the process satisfies $d\beta = dN = 0$, the entropy change is computed as reduction of the scaled Helmholtz free energy as

$$\Delta S^{\text{ext}} = \int_{X_1}^{X_2} \alpha(\beta, X, N)\, dX = \mathscr{F}(\beta, X_2, N) - \mathscr{F}(\beta, X_1, N). \tag{11.35}$$

Further change of the independent variables from (β, X, N) to (β, α, N) is done by introducing the *scaled* Gibbs free energy:

$$\mathscr{G} = \mathscr{F} + \alpha X = S - \beta U + \alpha X. \tag{11.36}$$

Note that $\mathscr{G} = -\beta G$. The independent variables of the Gibbs free energy are (β, α, N) since $d\mathscr{G} = d\mathscr{F} + (d\alpha X + X d\alpha) = -U d\beta + X d\alpha - \gamma dN$. From this equation, we find

$$\left(\frac{\partial \mathscr{G}}{\partial \beta}\right)_{\alpha,N} = -U(\beta, \alpha, N), \tag{11.37}$$

$$\left(\frac{\partial \mathscr{G}}{\partial \alpha}\right)_{\beta,N} = X(\beta, \alpha, N). \tag{11.38}$$

Note that the definition of the Gibbs free energy by Eq. (11.36) is obtained from Eq. (11.6) if we identify $\mathscr{G} = \psi$. Accordingly, Eqs. (11.37) and (11.38) coincide with Eqs. (11.7) and (11.8).

The Legendre transformation that changes the state variable N to μ is given by

$$\mathscr{G} + \gamma N = S - \beta U + \alpha X + \gamma N. \tag{11.39}$$

Since $d(\mathscr{G} + \mu N) = d\mathscr{G} + (d\gamma N + \gamma dN) = -U d\beta + X d\alpha + N d\gamma$, the natural independent variables is now (β, α, γ). From the extensive property of S, X, and N, we have the Gibbs-Duhem relation,

$$-U d\beta + X d\alpha + N d\gamma = 0. \tag{11.40}$$

Thus this free energy is identical to zero, and we obtain $\mathscr{G} = -\gamma N$.

References

Abbott, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science, 275*(5297), 220–224.

Amari, S.-I., & Nagaoka, H. (2000). *Methods of information geometry*. Providence: The American Mathematical Society.

Berkes, P., Orbán, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science, 331*(6013), 83–87.

Brown, E. N., Frank, L. M., Tang, D., Quirk, M. C., & Wilson, M. A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience, 18*(18), 7411–7425.

Burkitt, A. N., Meffin, H., & Grayden, D. B. (2003). Study of neuronal gain in a conductance-based leaky integrate-and-fire neuron model with balanced excitatory and inhibitory synaptic input. *Biological Cybernetics, 89*(2), 119–125.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Review Neuroscience, 13*(1), 51–62.

Carnot, S. (1824). *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*, Bachelier, Paris.

Cauller, L. J., & Kulics, A. T. (1991). The neural basis of the behaviorally relevant N1 component of the somatosensory-evoked potential in SI cortex of awake monkeys: Evidence that backward cortical projections signal conscious touch sensation. *Experimental Brain Research, 84*(3), 607–619.

Chance, F. S., Abbott, L. F., & Reyes, A. D. (2002). Gain modulation from background synaptic input. *Neuron, 35*(4), 773–782.

Doiron, B., Longtin, A., Berman, N., & Maler, L. (2001). Subtractive and divisive inhibition: Effect of voltage-dependent inhibitory conductances and noise. *Neural Computation, 13*(1), 227–248.

Donner, C., Obermayer, K., & Shimazaki, H. (2017). Approximate inference for time-varying interactions and macroscopic dynamics of neural populations. *PLoS Computational Biology, 13*(1), e1005309.

Ghose, G. M., & Maunsell, J. H. R. (2002). Attentional modulation in visual cortex depends on task timing. *Nature, 419*(6907), 616–620.

Granot-Atedgi, E., Tkačik, G., Segev, R., & Schneidman, E. (2013). Stimulus-dependent maximum entropy models of neural population codes. *PLoS Computational Biology, 9*(3), e1002922.

Ito, S., & Sagawa, T. (2013). Information thermodynamics on causal networks. *Physics Review Letter, 111*(18), 180603.

Ito, S., & Sagawa, T. (2015). Maxwell's demon in biochemical signal transduction with feedback loop. *Nature Communication, 6*, Article number: 7498.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review, 106*(4), 620–630.

Kass, R. E., Kelly, R. C., & Loh, W.-L. (2011). Assessment of synchrony in multiple neural spike trains using loglinear point process models. *Annals of Applied Statistics, 5*, 1262–1292.

Kelly, R. C., & Kass, R. E. (2012). A framework for evaluating pairwise and multiway synchrony among stimulus-driven neurons. *Neural Computation, 24*(8), 2007–2032.

Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., & Arieli, A. (2003). Spontaneously emerging cortical representations of visual attributes. *Nature, 425*(6961), 954–956.

Laughlin, S. B. (1989). The role of sensory adaptation in the retina. *Journal of Experimental Biology, 146*, 39–62.

Lee, B. B., Dacey, D. M., Smith, V. C., & Pokorny, J. (2003). Dynamics of sensitivity regulation in primate outer retina: The horizontal cell network. *Journal of Vision, 3*(7), 513–526.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology, 77*(1), 24–42.

Manita, S., Suzuki, T., Homma, C., Matsumoto, T., Odagawa, M., Yamada, K., et al. (2015). A top-down cortical circuit for accurate sensory perception. *Neuron, 86*(5), 1304–1316.

Martínez-Trujillo, J., & Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron, 35*(2), 365–370.

McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience, 19*(1), 431–441.

Mitchell, S. J., & Silver, R. A. (2003). Shunting inhibition modulates neuronal gain during synaptic excitation. *Neuron, 38*(3), 433–445.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*(4715), 782–784.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology, 70*(3), 909–919.

Nasser, H., Marre, O., & Cessac, B. (2013). Spatio-temporal spike train analysis for large scale networks using the maximum entropy principle and monte carlo method. *Journal of Statistical Mechanics, 2013*(03), P03006.

Ohiorhenuan, I. E., Mechler, F., Purpura, K. P., Schmid, A. M., Hu, Q., & Victor, J. D. (2010). Sparse coding and high-order correlations in fine-scale cortical networks. *Nature, 466*(7306), 617–621.

Ohzawa, I., Sclar, G., & Freeman, R. D. (1985). Contrast gain control in the cat's visual system. *Journal Neurophysiology, 54*(3), 651–667.

Prescott, S. A., & De Koninck, Y. (2003). Gain control of firing rate by shunting inhibition: roles of synaptic noise and dendritic saturation. *Proceedings of National Academy of Science USA, 100*(4), 2076–2081.

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience, 19*(5), 1736–1753.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron, 26*(3), 703–714.

Rothman, J. S., Cathala, L., Steuber, V., & Silver, R. A. (2009). Synaptic depression enables neuronal gain control. *Nature, 457*(7232), 1015–1018.

Sachidhanandam, S., Sreenivasan, V., Kyriakatos, A., Kremer, Y., & Petersen, C. C. (2013). Membrane potential correlates of sensory perception in mouse barrel cortex. *Nature Neuroscience, 16*(11), 1671–1677.

Sagawa, T., & Ueda, M. (2010). Generalized Jarzynski equality under nonequilibrium feedback control. *Physics Review Letter, 104*(9), 090602.

Sagawa, T., & Ueda, M. (2012). Fluctuation theorem with information exchange: Role of correlations in stochastic thermodynamics. *Physics Review Letter, 109*(18), 180602.

Sakmann, B., & Creutzfeldt, O. D. (1969). Scotopic and mesopic light adaptation in the cat's retina. *Pflügers Archiv: European Journal of Physiology, 313*(2), 168–185.

Salinas, E., & Abbott, L. F. (1996). A model of multiplicative neural responses in parietal cortex. *Proceedings of National Academy of Sciences USA, 93*(21), 11956–11961.

Salinas, E., & Sejnowski, T. J. (2001). Gain modulation in the central nervous system: Where behavior, neurophysiology, and computation meet. *Neuroscientist, 7*(5), 430–440.

Schneidman, E., Berry, M. J., Segev, R., & Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature, 440*(7087), 1007–1012.

Schultz, W. (2016). Dopamine reward prediction-error signalling: A two-component response. *Nature Review Neuroscience, 17*(3), 183–195.

Seidemann, E., & Newsome, W. T. (1999). Effect of spatial attention on the responses of area MT neurons. *Journal of Neurophysiology, 81*(4), 1783–1794.

Shimazaki, H. (2013). Single-trial estimation of stimulus and spike-history effects on time-varying ensemble spiking activity of multiple neurons: a simulation study. *Journal of Physics: Conference Series, 473*, 012009.

Shimazaki, H., Amari, S.-I., Brown, E. N., & Grün, S. (2009). State-space analysis on time-varying correlations in parallel spike sequences. In *Proceedings of IEEE ICASSP*, pp. 3501–3504.

Shimazaki, H., Amari, S.-i., Brown, E. N., & Grün, S. (2012). State-space analysis of time-varying higher-order spike correlation for multiple neural spike train data. *PLoS Computational Biology, 8*(3), e1002385.

Shimazaki, H., Sadeghi, K., Ishikawa, T., Ikegaya, Y., & Toyoizumi, T. (2015). Simultaneous silence organizes structured higher-order interactions in neural populations. *Scientific Reports, 5*, 9821.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., et al. (2006). The structure of multi-neuron firing patterns in primate retina. *Journal of Neuroscience, 26*(32), 8254–8266.

Silver, R. A. (2010). Neuronal arithmetic. *Nature Review Neuroscience, 11*(7), 474–489.

Smith, A. C., & Brown, E. N. (2003). Estimating a state-space model from point process observations. *Neural Computation, 15*(5), 965–991.

Spratling, M. W., & Johnson, M. H. (2004). A feedback model of visual attention. *Journal of Cognitive Neuroscience, 16*(2), 219–237.

Supèr, H., Spekreijse, H., & Lamme, V. A. (2001). A neural correlate of working memory in the monkey primary visual cortex. *Science, 293*(5527), 120–124.

Sutherland, C., Doiron, B., & Longtin, A. (2009). Feedback-induced gain control in stochastic spiking networks. *Biological Cybernetics, 100*(6), 475–489.

Tang, A., Jackson, D., Hobbs, J., Chen, W., Smith, J. L., Patel, H., et al. (2008). A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *Journal of Neuroscience, 28*(2), 505–518.

Tkačik, G., Marre, O., Amodei, D., Schneidman, E., Bialek, W., & Berry, M. J. (2014). Searching for collective behavior in a large network of sensory neurons. *PLoS Computational Biology, 10*(1), e1003408.

Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S. E., Berry, M. J., et al. (2015). Thermodynamics and signatures of criticality in a network of neurons. *Proceedings of National Academy of Sciences USA, 112*(37), 11508–11513.

Yu, S., Huang, D., Singer, W., & Nikolic, D. (2008). A small world of neuronal synchrony. *Cerebral Cortex, 18*(12), 2891–2901.