frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 21 November 2013
doi: 10.3389/fncom.2013.00165

Information maximization principle explains the emergence of complex cell-like neurons

Takuma Tanaka* and Kiyohiko Nakamura

Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan

Edited by:
Tomoki Fukai, RIKEN Brain Science Institute, Japan

Reviewed by:
Ko Sakai, University of Tsukuba, Japan
James McFarland, University of Maryland, USA

****Correspondence:***
Takuma Tanaka, Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259-G3-46, Nagatsuta-cho, Midori-ku, Yokohama, 224-8502, Japan
e-mail: tanaka.takuma@gmail.com

We propose models and a method to qualitatively explain the receptive field properties of complex cells in the primary visual cortex. We apply a learning method based on the information maximization principle in a feedforward network, which comprises an input layer of image patches, simple cell-like first-output-layer neurons, and second-output-layer neurons (Model 1). The information maximization results in the emergence of the complex cell-like receptive field properties in the second-output-layer neurons. After learning, second-output-layer neurons receive connection weights having the same size from two first-output-layer neurons with sign-inverted receptive fields. The second-output-layer neurons replicate the phase invariance and iso-orientation suppression. Furthermore, on the basis of these results, we examine a simplified model showing the emergence of complex cell-like receptive fields (Model 2). We show that after learning, the output neurons of this model exhibit iso-orientation suppression, cross-orientation facilitation, and end stopping, which are similar to those found in complex cells. These properties of model neurons suggest that complex cells in the primary visual cortex become selective to features composed of edges to increase the variability of the output.

Keywords: information maximization principle, complex cell, primary visual cortex, extraclassical receptive field, computational model

1. INTRODUCTION

A fundamental question that is often raised in neuroscience is how to determine the principle that underlies neural information coding in the brain. In terms of sensory information processing, this question can be answered by explaining how sensory neurons acquire their selectivity to inputs. The primary visual cortex (V1) is an ideal subject for this type of investigation because experimental results regarding the receptive field properties, single-cell electrophysiology, and topographic selectivity map are accumulated in V1. These experimental results allow us to screen the proposed principles of neural information coding by comparing the behavior of the model on the basis of each principle with the receptive field properties of neurons in V1. This screening provides us with a way to deal with the general principle of neural information coding in the cerebral cortex. Several principles having similar mathematical structures, such as the information maximization principle (Linsker, 1988; Bell and Sejnowski, 1997) and sparse coding hypothesis (Barlow, 1959; Olshausen and Field, 1996), were proposed to explain the receptive field properties of simple cells in V1. The statistical independence of the output neurons is the most essential assumption of these models. In the framework of independent component analysis (ICA), the output neurons acquire selectivity to stimuli such that the outputs of these neurons are uncorrelated and as statistically independent as possible. The statistical independence is closely related to the information maximization principle and sparse coding. If the output neurons are statistically dependent, the amount of information conveyed by these neurons reduces because many of them share the same information. Thus, maximizing the amount of information conveyed by the output neurons gives a result similar to that obtained when increasing the statistical independence of the output neurons. If the activity of neurons is independent and uncorrelated, the number of neurons firing simultaneously decreases, and therefore, the neuronal activity becomes sparse.

Since publication of the groundbreaking work by Hubel and Wiesel (1959), it is widely accepted that simple cells have wavelet-like receptive fields, which respond when a wavelet or an edge is positioned appropriately. The ICA models have revealed that ICA of natural images generates output units with simple cell-like receptive field properties. These results suggest that the assumption of the statistical independence of output neurons is a promising principle of neural information coding. However, the ability of this principle to explain the receptive field properties of complex cells has not been addressed. In contrast to simple cells, the response of complex cells is predominantly determined by the orientation of gratings and edges, and these cells are less sensitive to the positions of edges (Hubel and Wiesel, 1962). Thus, it is assumed that complex cells respond to abstract orientation and that the shift-invariant representation of the stimuli in the visual field is accomplished by complex cells. However, recent studies reveal that complex cells are not simple orientation detectors or shift-invariance detectors, and they exhibit surround suppression and facilitation by the gratings outside their classical receptive fields (Jones et al., 2001, 2002). Superimposing gratings perpendicular to the preferred orientation in the receptive field of

complex cells suppresses the firing of these cells (Bonds, 1989). These results suggest that the outputs of complex cells are not a simple pooling of simple cell inputs with similar orientation preferences. Using learning principles that can explain the emergence of these complex properties, we will be able to approach the general principle of neural information coding. Conversely, a general principle may shed light on what characterizes the features detected by complex cells. A number of theoretical studies were reported to explain the properties of complex cells (Földiák, 1991; Hyvärinen and Hoyer, 2001; Berkes and Wiskott, 2005; Karklin and Lewicki, 2005, 2009; Shan et al., 2007). Extraclassical receptive fields and complicated receptive field properties were replicated by these studies. However, these models assume that the complex cell-like units receive inputs whose magnitude does not depend on the polarity of the input image pixels (black and white). This type of input facilitates the emergence of the shift-invariant complex cell-like units; however, this assumption should be justified by a general principle.

In this paper, we show that this assumption is justified by using the information maximization principle. In addition, we show that the information maximization principle explains the receptive field properties of complex cells. The differential entropy of the output was used as the measure of information transmitted from the input to the output in the previous ICA models (Bell and Sejnowski, 1995; Shriki et al., 2001). In the Methods section, we introduce mutual information, entropy, and the models we propose in this paper. In the first part of the Results section, we use a three-layer feedforward network comprising an input layer of natural image patches, simple cell-like first-output-layer neurons, and second-output-layer neurons that receive inputs from the first-output-layer neurons (Model 1). Our simulation results obtained using a learning rule that maximizes information transmission show that second-output-layer neurons that receive inputs from first-output-layer neurons with rectifying nonlinearity become shift invariant after the learning process. Some second-output-layer neurons exhibit surround suppression similar to that reported for complex cells. A theoretical calculation based on the fact that the edges detected by simple cells are almost statistically independent proves that the phase insensitivity of complex cells maximizes the output entropy of V1. Model 1 contains realistic first-output-layer neurons with non-negative-definite firing rates. However, Model 1 is computationally expensive so we develop a simplified model of complex cells and more precisely examine the properties of model neurons. In Model 2, the output neurons receive the absolute value of the simple cell-like neurons as inputs. This simplification reduces the computational load and accelerates the convergence. In the Discussion section, we compare our proposed models with previous models and discuss the biological implications of our models.

2. METHODS

2.1. INFORMATION MAXIMIZATION PRINCIPLE

Information transmitted from input $\mathbf{x}$ to output $\mathbf{y}$ is measured by the mutual information

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x},\mathbf{y}),$$

where $H(\mathbf{x})$, $H(\mathbf{y})$, and $H(\mathbf{x},\mathbf{y})$ are the entropies of the input, the output, and combination of the input and the output. Bell and Sejnowski (1995) showed that the amount of information transmitted from the input layer to the output layer equals the output entropy plus a constant dependent only on the probability distribution of the input. The mutual information can be given by

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}),$$

where $H(\mathbf{y}|\mathbf{x})$ is the entropy of the noise. In the noiseless system, the maximization of the mutual information of input and output can be achieved by the maximization of output entropy. Similar to Bell and Sejnowski (1995), we ignore the intrinsic noise of neurons although it is an important factor contributing to the variability of neuronal firings. We can ignore the effect of the intrinsic noise because qualitatively similar results are obtained by ICA models when we add a small noise to the input and the output of first-output-layer neurons. Thus, by maximizing the output entropy, we expect to have an information-efficient representation of the input in the output layer.

Entropy is a measure of the variability of a random variable, and is defined by

$$H(\mathbf{x}) = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}, \quad (1)$$

where $p(\mathbf{x})$ is the probability density function of the random variable $\mathbf{x}$. Taking the logarithm to base 2, we have the entropy measured in bits. However, in the following sections, we take logarithms to base e because this simplifies the analytical treatment. The entropy of the univariate normal distribution $p(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, where μ is the mean and σ is the standard deviation, is $\log\sqrt{2\pi e\sigma^2}$; the entropy is a monotonically increasing function of the standard deviation. The uniform distribution

$$p(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

has the largest entropy, $\log(b-a)$, among the probability distributions whose domain is between a and b (Cover and Thomas, 2006). In other words, the uniform distribution is the probability distribution with the highest variability. This means that, under the assumption of rate coding, the entropy of the output of a neuron is large if the maximal firing rate is large and if the firing rate over time is uniformly distributed. The entropy of the n-variate normal distribution

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mu)\right) \quad (3)$$

is $\log\sqrt{(2\pi e)^n \det \mathbf{\Sigma}}$, where μ is the mean vector and $\mathbf{\Sigma}$ is the covariance matrix. The normal distributions with the mean

vectors

$$\mu_1 = \mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{4}$$

and the covariance matrices

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{5}$$

and

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 \end{pmatrix} \tag{6}$$

have the same marginal distributions, i.e., x_1 and x_2 obey the standard normal distribution in both cases. Thus, the entropies of the marginalized x_1 and x_2 assume the same value for both distributions. However, the joint entropy of the joint distribution of x_1 and x_2 with $\mathbf{\Sigma}_1$, $\log_2(2\pi e) \approx 4.09$ bits, is greater than that of the joint distribution with $\mathbf{\Sigma}_2$, $\log_2\sqrt{(2\pi e)^2/2} \approx 3.59$ bits, because the variables x_1 and x_2 are independent in the former and correlated in the latter. Entropy is maximized if the random variables are independent of each other. We can see that the joint entropy of the output of the neurons takes a large value if these neurons fire vigorously and independently. Entropy can therefore be used as a measure of the output variability of a population of neurons. The entropy of the output of neurons in the sensory areas is large if their population firing pattern varies with the input, and it is small if the firing pattern is less affected by the change of input. The output of sensory neurons with large output entropy can be easily utilized by higher sensory areas because different inputs are well separated in the space of the output firing patterns. For this reason, entropy is used as an objective function to train the models of sensory information processing. Another reason for using entropy as an objective function is that entropy is a measure of sensitivity. If the input vector $\mathbf{x}$ is transformed to the output vector $\mathbf{y}$ having the same number of elements, the entropy of the output is given by

$$H(\mathbf{y}) = \int p(\mathbf{x}) \log\det \mathbf{J} d\mathbf{x} + H(\mathbf{x}), \tag{7}$$

where $J_{ij} = \frac{\partial y_i}{\partial x_j}$ is the Jacobian matrix of the transform. Because the (i, j) entry of the Jacobian matrix is the sensitivity of output i to the change of input j, the maximization of the output entropy $H(\mathbf{y})$ can be regarded as the maximization of the sensitivity of output neurons. It is desirable that sensory neurons are maximally sensitive to the change of the stimuli. Note that $\det \mathbf{J}$ cannot always take an arbitrary large value when we use bounded functions, such as in Equation 13, as the activation function of output units.

2.2. MODEL 1

We assume that the system has an N-dimensional input vector $\mathbf{x}$ and output comprising a $2N$-dimensional first-output-layer neuron vector $\mathbf{y} = [\mathbf{y}^+, \mathbf{y}^-]$ and an N-dimensional second-output-layer neuron vector $\mathbf{z}$ (**Figure 1A**). The output of the first-output-layer neurons is a deterministic function of input $\mathbf{x}$ and is described by

$$y_i^+ = R(u_i) \tag{8}$$

and

$$y_i^- = R(-u_i), \tag{9}$$

where

$$u_i = f(a_i) \tag{10}$$

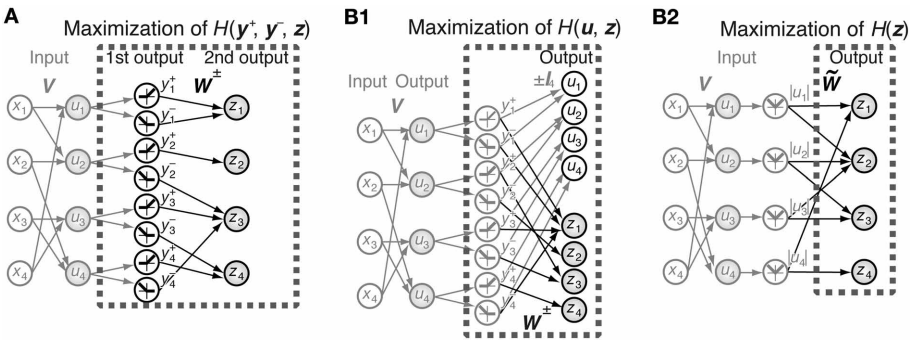

FIGURE 1 | Schematic representation of the models. (A) Joint entropy of the first-output-layer $\mathbf{y}^+$ and $\mathbf{y}^-$ and second-output-layer $\mathbf{z}$ is maximized in Model 1. $\mathbf{V}$ is the connection weight matrix between $N = 4$ input neurons and $2N = 8$ first-output-layer neurons, and $\mathbf{W}$ is the connection weight matrix between the first-output-layer neurons and $N = 4$ second-output-layer neurons. The first-output-layer neurons y_i^+ take on values greater than zero if $u_i > 0$, and zero otherwise, whereas y_i^- take on values greater than zero if $u_i < 0$, and zero otherwise. Black and gray arrows indicate modifiable and fixed connections, respectively. Gray filled circles and open circles are units with and without a nonlinear activation function $f(x)$, respectively. **(B1)** Entropy of the outputs $\mathbf{u}$ and $\mathbf{z}$ is maximized in Model 2. Output z_i is a function of the linear superposition of inputs y_i^+ and y_i^- by the connection weight matrices $\mathbf{W}^+$ and $\mathbf{W}^-$, respectively. **(B2)** Model 2 is equivalent to the maximization of the entropy of the output $\mathbf{z}$, which is a function of the linear superposition of input $|u_i|$ by the connection weight matrix $\tilde{\mathbf{W}}$.

and

$$a_i = \sum_{1 \le j \le N} V_{ij} x_j. \tag{11}$$

Here,

$$R(x) = \begin{cases} x & x \ge 0 \\ 0 & x < 0 \end{cases} \tag{12}$$

is the ramp function, and

$$f(x) = 2 \arctan \tanh \frac{x}{2} \tag{13}$$

is the activation function used in this paper. This activation function gives the same algorithm as the ICA algorithm of "tanh nonlinearity" (Hyvärinen et al., 2001) because $f''(x)/f'(x) = -\tanh x$. This activation function corresponds to the assumption that the independent components follow a sparse distribution $p(x) = 1/(\pi \cosh x)$ in terms of maximal likelihood. Different activation functions corresponding to dense distributions such as that satisfying $f''(x)/f'(x) = -x^3$ cannot predict the receptive field properties of simple cells. Any decomposition matrix of natural images obtained using ICA algorithms can be used as matrix $\mathbf{V} = (V_{ij})$. In all simulations of this paper, we use the decomposition matrix obtained by the method described in Section 2.4. The second-output-layer output $\mathbf{z}$ is a function of the first-output-layer output $\mathbf{y}$:

$$z_i = f\left(-h_i + \sum_{1 \le j \le N} W_{ij}^{+}\left(y_j^{+} - \bar{y}_j^{+}\right) + \sum_{1 \le j \le N} W_{ij}^{-}\left(y_j^{-} - \bar{y}_j^{-}\right)\right), \tag{14}$$

where W_{ij}^{+} and W_{ij}^{-} are connection weights, h_i is the offset of the second-output-layer neuron i, $\bar{y}_j^{+}$ is the average of y_j^{+}, and $\bar{y}_j^{-}$ is the average of y_j^{-}. The terms $-\bar{y}_j^{+}$ and $-\bar{y}_j^{-}$ are introduced to accelerate the convergence. These terms do not affect the values of the parameters after convergence because the increase of $\bar{y}_j^{\pm}$ to $\bar{y}_j^{\pm} + \Delta\bar{y}_j^{\pm}$ is offset by the change of h_i to $h_i - \sum_{1 \le j \le N} W_{ij}^{\pm} \Delta\bar{y}_j^{\pm}$.

To enable readers to replicate our results without the need to follow the details of the derivation of the algorithm, here we summarize the simulation process and describe how to evaluate the entropy of the output and derivation of the learning rule in the next section. We use a batch learning process to accelerate the simulation. The weights W_{ij}^{+}, W_{ij}^{-}, and thresholds h_i are updated once every 100 steps using

$$\mathbf{W}^{\pm} \leftarrow \mathbf{W}^{\pm} + \epsilon \sum_{1 \le t \le 100} \Delta\mathbf{W}^{\pm}(t) \tag{15}$$

and

$$\mathbf{h} \leftarrow \mathbf{h} + \epsilon \sum_{1 \le t \le 100} \Delta\mathbf{h}(t), \tag{16}$$

where $\Delta W_{ij}^{\pm}(t)$ and $\Delta h_i(t)$ are defined by Equations 29 and 32, respectively, and ϵ is the learning rate. The update was performed 1.63×10^6 times with $\epsilon = 10^{-4}$ and then 3×10^4 times with $\epsilon = 10^{-5}$.

2.3. ENTROPY OF OUTPUT AND DERIVATION OF THE LEARNING RULE OF MODEL 1

Here we describe the objective function, the entropy of the output to be maximized, and the derivation of the learning algorithm. First, we define the joint entropy of the first- and second-output-layer neurons, and then we calculate the derivative of the joint entropy with respect to the connection weights. In our model, the N-dimensional input vector gives rise to a $2N$-dimensional first-output-layer vector and an N-dimensional second-output-layer vector. The representation is overcomplete because there are more output components than input components. Shriki et al. (2001) considered the maximization of the entropy of the overcomplete representation of the input. Although their model does not contain a multilayer structure and rectifying nonlinearity, the same idea can be applied to Model 1. The probability density of $\mathbf{y}$ and $\mathbf{z}$ is related to the input distribution and is given by the relation

$$P(\mathbf{y}, \mathbf{z}) \propto \frac{P(\mathbf{x})}{\sqrt{\det(\boldsymbol{\chi}^T \boldsymbol{\chi})}}, \tag{17}$$

where the susceptibility matrix $\boldsymbol{\chi} \in \mathbb{R}^{3N \times N}$ is defined as

$$\boldsymbol{\chi} = \begin{pmatrix} \mathbf{Y}^{+} \\ \mathbf{Y}^{-} \\ \mathbf{Z} \end{pmatrix}. \tag{18}$$

Here we define

$$\begin{aligned} Y_{ij}^{\pm} &= \frac{\partial y_i^{\pm}}{\partial x_j} \\ &= \pm s(\pm u_i) f'(a_i) V_{ij}, \end{aligned} \tag{19}$$

$$\begin{aligned} Z_{ij} &= \frac{\partial z_i}{\partial x_j} \\ &= f'(b_i) \sum_{1 \le k \le N} \left(W_{ik}^{+} s(u_k) - W_{ik}^{-} s(-u_k)\right) f'(a_k) V_{kj}, \end{aligned} \tag{20}$$

$$b_i = -h_i + \sum_{1 \le j \le N} W_{ij}^{+}\left(y_j^{+} - \bar{y}_j^{+}\right) + \sum_{1 \le j \le N} W_{ij}^{-}\left(y_j^{-} - \bar{y}_j^{-}\right), \tag{21}$$

and

$$s(x) = \begin{cases} 1 & x > 0 \\ \frac{1}{\sqrt{2}} & x = 0 \\ 0 & x < 0 \end{cases}, \tag{22}$$

because $\frac{\mathrm{d}}{\mathrm{d}x} R(f(x)) = s(x) f'(x)$ and $f'(-x) = f'(x)$. We define $s(0) = 1/\sqrt{2}$ to simplify the equation. Here the dependence of $\mathbf{Y}$ and $\mathbf{Z}$ on the time step t is not explicitly shown. Thus, we

maximize the entropy of the output

$$\begin{aligned} H(\mathbf{y},\mathbf{z}) &= -\int\int P(\mathbf{y},\mathbf{z})\log P(\mathbf{y},\mathbf{z})\,\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z} \\ &= -\int P(\mathbf{x})\log\left(\frac{P(\mathbf{x})}{\sqrt{\det(\boldsymbol{\chi}^T\boldsymbol{\chi})}}\right)\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\mathbb{E}\left[\log\det(\boldsymbol{\chi}^T\boldsymbol{\chi})\right] + H(\mathbf{x}), \end{aligned} \tag{23}$$

where $\mathbb{E}[\cdot]$ indicates averaging over the input distribution and we used $P(\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{z} = P(\mathbf{x})\mathrm{d}\mathbf{x}$ (Shriki et al., 2001).

The first term of Equation 23 is given by the average of

$$\frac{1}{2}\log\det\left(\boldsymbol{\chi}^T\boldsymbol{\chi}\right) = \frac{1}{2}\log\det\left(\mathbf{Y}^{+T}\mathbf{Y}^{+} + \mathbf{Y}^{-T}\mathbf{Y}^{-} + \mathbf{Z}^T\mathbf{Z}\right). \tag{24}$$

Here note that

$$\begin{aligned} \left(\mathbf{Y}^{+T}\mathbf{Y}^{+} + \mathbf{Y}^{-T}\mathbf{Y}^{-}\right)_{ij} &= \sum_{1\le k\le N} V_{ki}\left(s(u_k)^2 f'(a_k)^2 \right. \\ &\quad \left. + s(-u_k)^2 f'(a_k)^2\right) V_{kj} \\ &= \left(\mathbf{V}^T\,\mathrm{diag}(f'(a_k)^2)\mathbf{V}\right)_{ij}, \end{aligned} \tag{25}$$

where $\mathrm{diag}(d_k)$ is a diagonal matrix whose diagonal entries are $d_1,\ d_2,\ \ldots,\ d_N$. We also note that

$$\left(\mathbf{Z}^T\mathbf{Z}\right)_{ij} = \left(\mathbf{V}^T\,\mathrm{diag}(f'(a_k))\mathbf{C}^T\mathbf{C}\,\mathrm{diag}(f'(a_k))\mathbf{V}\right)_{ij}, \tag{26}$$

where

$$C_{ij} = f'(b_i)\left[W^{+}_{ij}s(u_j) - W^{-}_{ij}s(-u_j)\right]. \tag{27}$$

From these relations we obtain

$$\begin{aligned} H(\mathbf{y},\mathbf{z}) &= \frac{1}{2}\log\det\left(\mathbf{V}^T\,\mathrm{diag}(f'(a_i)^2)\mathbf{V} + \mathbf{V}^T\,\mathrm{diag}(f'(a_i))\right. \\ &\quad \left.\times\,\mathbf{C}^T\mathbf{C}\,\mathrm{diag}(f'(a_i))\mathbf{V}\right) + H(\mathbf{x}) \\ &= \frac{1}{2}\log\det\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right) + \sum_{1\le k\le N}\log f'(a_k) + \log\det\mathbf{V} \\ &\quad + H(\mathbf{x}). \end{aligned} \tag{28}$$

The second and third terms are the objective function used in the ICA model to generate the connectivity of the first-output-layer neurons (see Section 2.4). The fourth term is the entropy of the input. Differentiating the first term with respect to $W^{\pm}_{ij}$, we obtain

$$\begin{aligned} \Delta W^{\pm}_{ij}(t) &= \frac{\partial}{\partial W^{\pm}_{ij}}\frac{1}{2}\log\det\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right) \\ &= \frac{1}{2}\sum_{1\le l,\,m\le N}\left[\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}\right]_{ml}\frac{\partial}{\partial W^{\pm}_{ij}}\left(\mathbf{C}^T\mathbf{C}\right)_{lm} \\ &= \sum_{1\le l,\,m\le N}\left[\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}\right]_{ml}\sum_k \frac{\partial C_{kl}}{\partial W^{\pm}_{ij}}C_{km} \\ &= \sum_{1\le l,\,m\le N}\left[\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}\right]_{ml}\left(\pm\delta_{jl}f'(b_i)s(\pm u_l)\right. \\ &\quad \left. + C_{il}\frac{f''(b_i)}{f'(b_i)}(y^{\pm}_j - \bar{y}^{\pm}_j)\right)C_{im} \\ &= \pm\left[\mathbf{C}\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}\right]_{ij}f'(b_i)s(\pm u_j) \\ &\quad + \left[\mathbf{C}\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\right]_{ii}\frac{f''(b_i)}{f'(b_i)}\left(y^{\pm}_j - \bar{y}^{\pm}_j\right), \end{aligned} \tag{29}$$

where

$$f'(x) = \frac{1}{\cosh x} \tag{30}$$

and

$$\frac{f''(x)}{f'(x)} = -\tanh x. \tag{31}$$

Differentiating the first term of Equation 28 with respect to h_i, we obtain

$$\begin{aligned} \Delta h_i(t) &= \sum_{1\le l,\,m\le N}\left[\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}\right]_{ml}\sum_{1\le k\le N}\frac{\partial C_{kl}}{\partial h_i}C_{km} \\ &= -\sum_{1\le l,\,m\le N}\left[\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}\right]_{ml}\sum_{1\le k\le N}\delta_{ik}C_{kl}\frac{f''(b_k)}{f'(b_k)}C_{km} \\ &= -\left[\mathbf{C}\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\right]_{ii}\frac{f''(b_i)}{f'(b_i)}. \end{aligned} \tag{32}$$

Maximization of the entropy of the second output layer, $H(\mathbf{z})$, can be achieved by replacing $\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}$ in the above learning rules with $\left(\mathbf{C}^T\mathbf{C}\right)^{-1}$. Results similar to those presented in this paper are obtained by this modified learning rule, which is equivalent to Model 2 without the constraint $W^{+}_{ij} = W^{-}_{ij}$.

2.4. NEWTON METHOD

We used the Newton method for the ICA model proposed in Amari et al. (1997) and Palmer et al. (2008) to obtain the connection matrix from the input units to first-output-layer neurons and perform the simulation of Model 2. In general, the entropy of real data in ICA models is not a convex function and has multiple local optima. Thus, the global optimum cannot be found

using the Newton method in most cases. However, it accelerates the convergence to one of the local optima. The following is the summary of the learning algorithm by Palmer et al. (2008). The entropy of the output is given by the expectation of

$$\log\det\mathbf{V}+\sum_{1\le i\le N}\log f'(y_i), \tag{33}$$

where

$$y_i=\sum_{1\le j\le N}V_{ij}x_j. \tag{34}$$

According to their results, the optimal direction $\Delta V_{ij}(t)$ at step t is given by

$$\Delta\mathbf{V}(t)=\mathbf{B}(t)\mathbf{V}, \tag{35}$$

where

$$B_{ij}=\frac{1+\frac{f''(y_i)}{f'(y_i)}y_i}{1+\eta_i}\delta_{ij}+\frac{\kappa_j\sigma_i^2\frac{f''(y_i)}{f'(y_i)}y_j-y_i\frac{f''(y_j)}{f'(y_j)}}{\kappa_i\kappa_j\sigma_i^2\sigma_j^2-1}\left(1-\delta_{ij}\right), \tag{36}$$

$$\kappa_i=\mathbb{E}\left[\frac{f''(y_i)^2-f'''(y_i)f'(y_i)}{f'(y_i)^2}\right], \tag{37}$$

$$\sigma_i^2=\mathbb{E}\left[y_i^2\right], \tag{38}$$

and

$$\eta_i=\mathbb{E}\left[\frac{f''(y_i)^2-f'''(y_i)f'(y_i)}{f'(y_i)^2}y_i^2\right], \tag{39}$$

where

$$\frac{f''(x)^2-f'''(x)f'(x)}{f'(x)^2}=\frac{1}{\cosh^2 x}. \tag{40}$$

Here the dependence on t is not explicitly shown. For the online algorithm, we updated κ_i, σ_i^2, and η_i using

$$\kappa_i(t)=\kappa_i\,(t\,-\,1)+\frac{1}{\tau}\left(\frac{1}{\cosh^2 y_i(t)}-\kappa_i(t\,-\,1)\right), \tag{41}$$

$$\sigma_i^2(t)=\sigma_i^2\,(t\,-\,1)+\frac{1}{\tau}\left(y_i(t)^2-\sigma_i^2\,(t\,-\,1)\right), \tag{42}$$

and

$$\eta_i(t)=\eta_i(t-1)+\frac{1}{\tau}\left(\frac{y_i(t)^2}{\cosh^2 y_i(t)}-\eta_i(t-1)\right) \tag{43}$$

at each time step. The integration time constant τ was set to be $10,000$ steps in all simulations. The weights V_{ij} were updated once every 100 steps using

$$\mathbf{V}\leftarrow\mathbf{V}+\epsilon\sum_{1\le t\le 100}\Delta\mathbf{V}(t), \tag{44}$$

where ϵ is the learning rate.

The decomposition matrix $\mathbf{V}$ was obtained using this method. We used randomly selected 20×20 pixels image patches from the images and converted the pixels in these image patches to 400-dimensional real-valued inputs, $\mathbf{x}$. The mean of the pixels of the image was subtracted from each image. The input images were not prewhitened (Olshausen and Field, 1997) because the input images without the prewhitening process yielded clearer results than the prewhitened ones. The receptive field properties of the second-output-layer neurons obtained using prewhitened images were qualitatively similar to those obtained using non-prewhitened images. The update of the matrix $\mathbf{V}$ using Equation 44 was performed 10^5 times with $\epsilon=10^{-5}$, 2.9×10^6 times with $\epsilon=10^{-4}$, and 9.7×10^6 times with $\epsilon=10^{-5}$. The iteration of the learning process of $\mathbf{V}$ must be sufficiently long as insufficient optimization disrupts the receptive field properties of output neurons.

2.5. MODEL 2

In Model 1, we decomposed N pixels into $2N$ first-output-layer values and N second-output-layer values. In Model 2, we decomposed $2N$ input values from the rectified linear simple cell-like elements into N sign-dependent simple cell-like neurons $\mathbf{u}$ and N complex cell-like neurons $\mathbf{z}$ (**Figure 1B1**). In other words, we fixed the output of the first half of the neurons to u_i, which are the nonlinear transformations of independent components of the images, and vary the connection weights to the second half of the neurons to obtain complex cell-like properties. $2N$ inputs are given by y_i^+ and y_i^- and are defined by Equations 8 and 9, respectively. The outputs $\mathbf{u}$ and $\mathbf{z}$ are functions of $\mathbf{y}=\left[\mathbf{y}^+,\mathbf{y}^-\right]$, and are defined as

$$u_i=y_i^+-y_i^- \tag{45}$$

and

$$z_i=f\left(c_i\right), \tag{46}$$

where

$$c_i=\sum_{1\le j\le N}W_{ij}^+\left(y_j^+-\bar{y}_j^+\right)+\sum_{1\le j\le N}W_{ij}^-\left(y_j^--\bar{y}_j^-\right). \tag{47}$$

We do not introduce the offset parameter h_i in this model because h_i takes on values close to zero in the simulation of Model 1. Here we define

$$\mathbf{W}=\begin{pmatrix}\mathbf{I}_N & -\mathbf{I}_N\\ \mathbf{W}^+ & \mathbf{W}^-\end{pmatrix}, \tag{48}$$

where $\mathbf{I}_N$ is an N-dimensional identity matrix. We fix the connection weights of the first half of the neurons, $\mathbf{u}$, to simple cell-like receptive fields, and update $\mathbf{W}^+$ and $\mathbf{W}^-$ using an ICA algorithm. Because the covariance matrix of independent components is a diagonal matrix (Hyvärinen et al., 2001), $\mathbf{W}^+$ and $\mathbf{W}^-$

must be appropriately set to decorrelate the outputs. Below, we set $\mathbf{W}^+ = \mathbf{W}^- = \tilde{\mathbf{W}}$, because assuming $\bar{y}_i^+ = \bar{y}_i^-$, $\mathbb{E}\left[y_i^+ y_j^+\right] = \mathbb{E}\left[y_i^- y_j^-\right]$, and $\mathbb{E}\left[y_i^+ y_j^-\right] = \mathbb{E}\left[y_i^- y_j^+\right]$, all of which approximately hold for independent components of natural scenes, the covariances $\mathbb{E}\left[u_i c_j\right] - \mathbb{E}\left[u_i\right]\mathbb{E}\left[c_j\right]$ vanish only if $\mathbf{W}^+ = \mathbf{W}^-$.

Then, the output z_i is given by

$$z_i = f\left(\sum_{1 \le j \le N} \tilde{W}_{ij}\left(|u_j| - \overline{|u_i|}\right)\right), \tag{49}$$

where $\overline{|u_i|}$ is the average of $|u_i|$, which equals $\bar{y}_j^+ + \bar{y}_j^-$. If input $y_i^\pm$ is greater than 0, $y_i^\mp$ is equal to 0; the probability density $p\left(y_i^+, y_i^-\right)$ is not a function with finite values, and therefore, ICA algorithms cannot be applied. However, assuming that $p\left(y_i^+, y_i^-\right)$ is a continuous function, the entropy of the output of this system is given by the expectation of

$$\begin{aligned} &\log\det\mathbf{W} + \sum_{1 \le i \le N} \log f'(c_i) + H(\mathbf{y}) \\ &= \log\det\tilde{\mathbf{W}} + \sum_{1 \le i \le N} \log f'(c_i) + N\log 2 + H(\mathbf{y}), \end{aligned} \tag{50}$$

where the last two terms do not depend on $\tilde{\mathbf{W}}$. Hence, the maximization of the first two terms is sufficient for the maximization of the output. This is equivalent to the simulation of the ICA model with the N-dimensional input $|u_i| - \overline{|u_i|}$ (**Figure 1B2**). Note that ICA can be applied to it because the probability density of $p(|u_i|)$ takes finite values, and that the first two terms of Equation 50 equal the first term of Equation 28 if $\left(\mathbf{I} + \mathbf{C}^T\mathbf{C}\right)^{-1}$ is replaced by $\left(\mathbf{C}^T\mathbf{C}\right)^{-1}$ and $W_{ij}^+ = W_{ij}^- = \tilde{W}_{ij}$. We perform the Newton method for N-dimensional inputs and outputs, which is much faster than the gradient descent of Model 1. The update of the matrix $\tilde{\mathbf{W}}$ obtained using Equation 44 was performed 1×10^5 times with $\epsilon = 10^{-6}$ and then 2.99×10^7 times with $\epsilon = 10^{-5}$.

2.6. CHARACTERIZATION OF MODEL NEURONS

We fit the connection weights to the first-output-layer neurons from the pixel at (i, j) with the Gabor function

$$A\exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\cos\left(kx - \phi\right) + B, \tag{51}$$

where

$$\begin{aligned} x &= (i - x_0)\cos\theta + \left(j - y_0\right)\sin\theta, \\ y &= -(i - x_0)\sin\theta + \left(j - y_0\right)\cos\theta \end{aligned}$$

with parameters A, B, x_0, y_0, σ_x, σ_y, θ, ϕ, and k by using the gradient descent method (**Figure 2**). The sum of the square of the difference between the fitted function and the connection weights is less than 10% of the sum of the square of the connection weights for 398 out of 400 first-output-layer neurons.

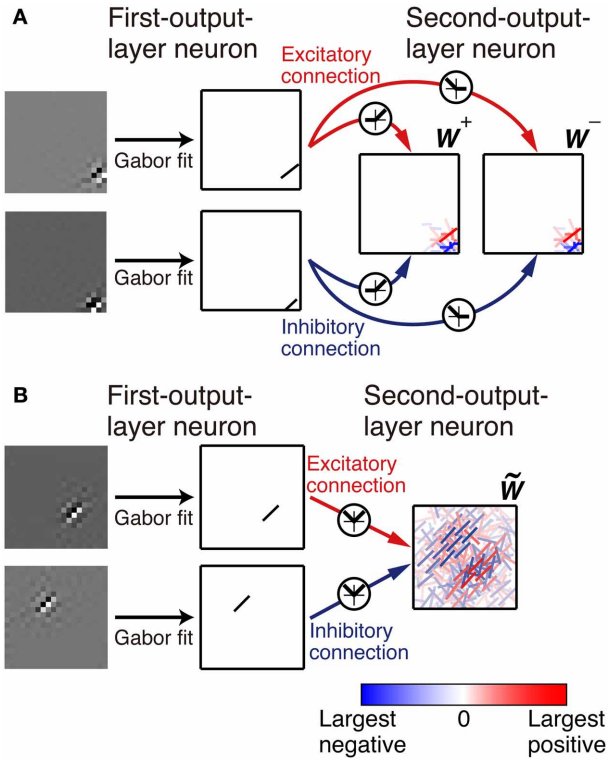

FIGURE 2 | Representation of the connection weights. The connection weights from input image patches to the simple cell-like first-output-layer neurons are fitted by a Gabor function. The connection weights from the first-output layer neurons to second-output layer neurons are represented by bars in the second-output-layer neurons. The color and opacity of each bar indicate the sign (red indicates positive, i.e., the excitatory connections, and blue indicates negative, i.e., the inhibitory connections) and magnitude of the connection weight from the first-output-layer neuron, respectively. **(A)**. In Model 1, the left box represents the connection weights from the first-output-layer neurons y_1^+ through y_N^+, and the right box represents the connection weights from y_1^- through y_N^-. **(B)**. In Model 2, connection weights $\tilde{W}_{ij}$ are represented by bars in a box.

The phase-dependent (F1) to phase-invariant (F0) component ratio (F1/F0 ratio) has been used to characterize simple and complex cells (Shapley and Lennie, 1985; Skottun et al., 1991). Simple cells are identified by the F1/F0 ratio greater than 1, and complex cells are identified by an F1/F0 ratio less than 1. We calculate the F1/F0 ratio of model neuron i by using

$$\frac{8}{\pi}\frac{\sqrt{\left(\sum_\phi R\left(z_i(\phi)\right)\sin\phi\right)^2 + \left(\sum_\phi R\left(z_i(\phi)\right)\cos\phi\right)^2}}{\sum_\phi R\left(z_i(\phi)\right)}, \tag{52}$$

where ϕ is the phase of the grating and $z_i(\phi)$ is the output of neuron i in response to the grating of the phase ϕ. Under this definition, the F1/F0 ratio equals 2 for the simple cell-like case, $z_i(\phi) = R(\sin\phi)$, and equals 0 for the perfectly phase-invariant case, $z_i(\phi) = 1$. We first choose the optimal grating for each neuron and obtain $z_i(\phi)$ by varying the phase of

the optimal grating. The optimal grating is chosen from the gratings with various radii (2, 3, 4, 5, and 6 pixels), center positions (x, $y = 1, 2, \ldots, 20$), orientations ($0°, 20°, \ldots, 340°$), spatial frequencies ($60°$/pixel, $75°$/pixel, $\ldots$, $120°$/pixel), and phases ($0°, 20°, \ldots, 340°$). The center of the optimal grating is used as the center of the gratings when examining the receptive field properties of model neurons (**Figures 4**, **6**).

Neurons tend to adapt to static stimuli and decrease their firing rates. Moving gratings are frequently used to evoke a large response in experiments. Because the response of this model does not depend on the previous stimuli, we use single stationary stimuli as inputs. This does not necessarily mean that the results in this paper correspond to the experimental results obtained using stationary stimuli. We compare the responses of model neurons with experimental results obtained using moving gratings.

3. RESULTS

3.1. MODEL 1: INFORMATION MAXIMIZATION IN A THREE-LAYER FEEDFORWARD NETWORK

Model 1 is a multilayer feedforward network (**Figure 1A**). This network contains N input units (x_i), $2N$ first-output-layer neurons (y_i^+ and y_i^-), and N second-output-layer neurons (z_i). We used randomly selected 20 pixels × 20 pixels image patches from natural photographs distributed by Prof. Bruno Olshausen on his homepage (Olshausen and Field, 1997) and converted the pixels in these image patches to $N = 400$ real-valued inputs. The input units correspond to the relay neurons in the lateral geniculate nucleus, and first- and second-output-layer neurons correspond to simple and complex cells in V1, respectively. The intensity of a pixel in the input images is represented by a real value in the present paper. In the preprocessing of images, we subtracted the mean pixel intensity from each image; the mean pixel intensity of an image patch is not necessarily zero. First, we performed the learning of the first-output-layer neurons, followed by the learning of the second-output-layer neurons. For first-output-layer neurons to acquire simple cell-like receptive field properties, we set the connection weights from the input units to first-output-layer neurons to the decomposition matrix $\mathbf{V} = (V_{ij})$ obtained from the linear ICA of natural image patches. A standard linear ICA algorithm decomposes the N-dimensional input vector $\mathbf{x}$ into an N-dimensional independent component vector whose elements can take both signs (Amari et al., 1996). Because the firing rate of a simple cell is not less than zero, we made two first-output-layer neurons y_i^+ and y_i^- from a nonlinear transformation of independent component i; if $u_i \geq 0$, we set $y_i^+ = u_i$ and $y_i^- = 0$, and if $u_i < 0$, we set $y_i^+ = 0$ and $y_i^- = -u_i$, where

$$u_i = f\left(\sum_{1 \leq j \leq N} V_{ij} x_j\right) \tag{53}$$

is the nonlinear transformation of independent component i by the sigmoidal activation function $f(x)$. The first-output-layer neuron y_i^+ is selective to the sign-inversion of the image to which the first-output-layer neuron y_i^- is selective, and vice versa. The output of the second-output-layer neuron i is defined by

$$z_i = f\left(-h_i + \sum_{1 \leq j \leq N} W_{ij}^+ y_j^+ + \sum_{1 \leq j \leq N} W_{ij}^- y_j^-\right), \tag{54}$$

where W_{ij}^+ and W_{ij}^- are the connection weights from the first-output-layer neurons y_j^+ and y_j^-, respectively, and h_i is the threshold. We updated the connection weight matrices $\mathbf{W}^+$ and $\mathbf{W}^-$, and maximized the entropy of the outputs $\mathbf{y}^+$, $\mathbf{y}^-$, and $\mathbf{z}$, $H(\mathbf{y}^+, \mathbf{y}^-, \mathbf{z})$.

First, we examine the connection weights of the model neurons. Note that we do not impose the constraint $W_{ij}^+ = W_{ij}^-$, which makes second-output-layer neurons invariant to the sign inversion of input images. However, after the learning process, the connection weights W_{ij}^+ and W_{ij}^- result in similar values (**Figure 3A**, $p < 0.01$, permutation test of Spearman's rank correlation coefficient). To visualize connection weights from first-output-layer neurons to second-output-layer neurons, we have to represent each first-output-layer neuron compactly. We used a Gabor function to fit the connection weights to the first-output-layer neurons, as shown in **Figure 2A**, and represent the fitted Gabor function with a bar. Thus, the first-output-layer neurons are indicated by bars that represent the optimal orientation and the spatial location of the fitted Gabor function. In the boxes on the right of **Figure 2A**, we plotted the bars corresponding to the first-output-layer neurons. The left one shows the weights W_{ij}^+ of a second-output-layer neuron, and the right one shows the weights W_{ij}^- of the same neuron. The approximate relation $W_{ij}^+ \approx W_{ij}^-$ holds in this neuron. Other examples are shown in **Figure 3B**. The strongest excitatory and inhibitory connections of each neuron are represented by red (RGB[100%, 0%, 0%]) and blue (RGB[0%, 0%, 100%]), respectively. The other connections are represented by a paler red (RGB[100%, 100(1 − r)%, 100(1 − r)%]) and a paler blue (RGB[100(1 − r)%, 100(1 − r)%, 100%]), where r is the ratio of the strength of the connection to the strongest excitatory or inhibitory connection. These figures indicate that the second-output-layer neurons tend to receive an input having the same sign from pairs of first-output-layer neurons with sign-inverted receptive fields and neurons with similar orientation preferences in a small region.

Figure 4 shows the response properties of second-output-layer neurons. The image patches presented to the network as the input are shown on the top of the panels. The output of the neuron in **Figure 2A** in response to the phase-shifted gratings is shown in **Figure 4A**. This figure shows that the output is positive for all phases, i.e., this neuron is less sensitive to the phase of the gratings. Insensitivity to the phase of the gratings is a feature of complex cells. Previous experiments have characterized simple and complex cells by measuring the relative modulation or phase-dependent (F1) to phase-invariant (F0) component ratio (F1/F0 ratio) in their responses to the optimal gratings (Shapley and Lennie, 1985; Skottun et al., 1991).

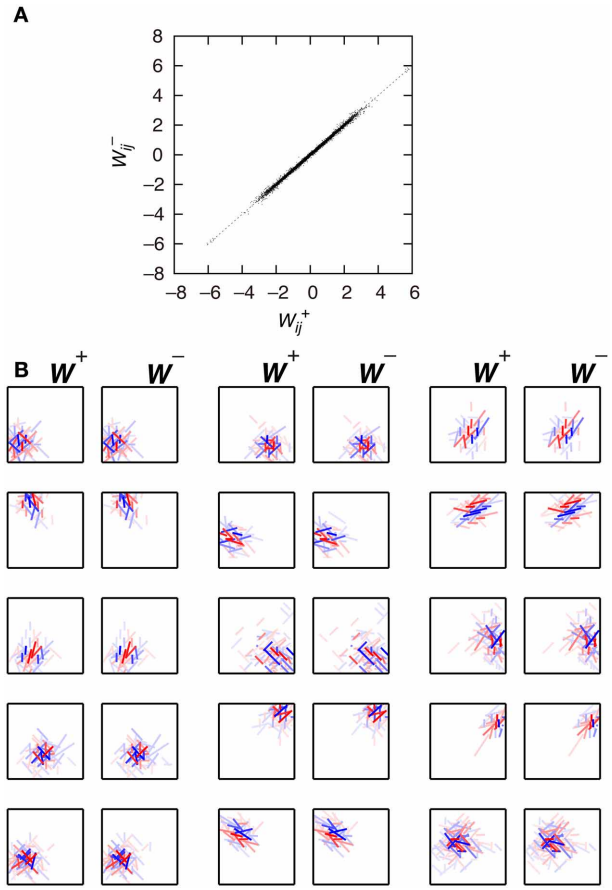

FIGURE 3 | Connection weights from the first-output-layer neurons to the second-output-layer neurons in Model 1. (A) Each dot corresponds to the connection weights from two first-output-layer neurons y_j^+ and y_j^- with sign-inverted receptive fields to the second-output-layer neuron z_i. **(B)** Connection weights from the first-output-layer neurons to 15 second-output-layer neurons.

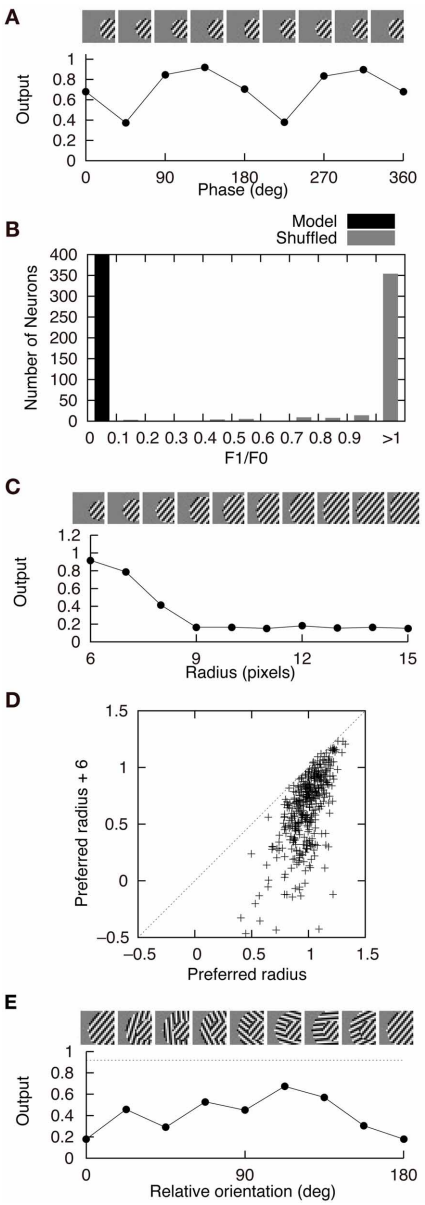

FIGURE 4 | Properties of second-output-layer neurons in Model 1. (A) Modulation of the output of the second-output-layer neuron whose connection weights are shown in **Figure 2A** by the phase of the gratings presented on the top. **(B)** Histogram of the F1/F0 ratio of the model neurons after the learning process (black) and the model neurons whose connection weights are randomly shuffled (gray). **(C)** Change in the output of the same neuron by the radius of the gratings. **(D)** Distribution of the output of 400 second-output-layer neurons in response to the optimal gratings with the preferred radius and to the same gratings whose radius is enlarged by 6 pixels. **(E)** Modulation of the output of the same neuron by the relative orientation of the gratings of the outer annulus to the inner patch. The dashed line indicates the output of the neuron in response to the stimulus with the grating of the inner patch only.

Neurons with the F1/F0 ratio greater than 1 are identified as simple cells, whereas those with the F1/F0 ratio less than 1 are identified as complex cells. The F1/F0 ratio of the neuron in **Figure 2A** is 0.02, which suggests that this cell should be classified as a complex cell. This phase insensitivity is a result of the convergence of the connections having the same sign from two first-output-layer neurons with sign-inverted receptive fields. If a second-output-layer neuron receives connection weights having the same size from each pair of first-output-layer neurons with sign-inverted receptive fields, the sign inversion of the input does not change the output of this second-output-layer neuron. In addition, the convergence of connections having the same sign from the first-output-layer neurons with similar orientation preferences facilitates the phase insensitivity of the second-output-layer neuron. The black bars in **Figure 4B** represent the histogram of the F1/F0 ratio of second-output-layer neurons. This histogram shows that almost all of them are classified as complex cells. In contrast, most of the model neurons with randomly shuffled connection weights exhibit F1/F0 ratios greater than 1 (**Figure 4B**, gray bars). Thus, the F1/F0 ratio close to zero is not a result of the multilayer structure because the second-output-layer neurons in the network with random first- to second-output-layer connections do not exhibit the F1/F0 ratio

close to zero. On the contrary, this is a result of the information maximization in the multilayer network and the resultant approximate relation $W_{ij}^{+} \approx W_{ij}^{-}$. These results suggest that the phase insensitivity of complex cells originates from an efficient encoding of the visual input. However, F1/F0 ratios of most model neurons are much smaller than experimentally obtained values. It is reported that a substantial proportion of complex cells have F1/F0 ratios greater than 0.5 (Skottun et al., 1991). This discrepancy may suggest that the simple and complex cells in V1 are not as clearly segregated as the first- and second-output-layer neurons in Model 1.

Stimuli presented in the silent receptive field surrounds can modulate the response of the cells in V1 to the stimuli presented in the classical receptive field (Jones et al., 2002). In most of the cells, the suppression is greatest when the orientation of the gratings in the silent surrounds is the same as the optimal orientation of the classical receptive field. The model neuron shown in **Figure 2A** is suppressed when the radius of the grating is increased (**Figure 4C**). Inhibitory connections play an important role in this suppression. Strong inhibitory connections originate from the first-output-layer neurons with similar orientation preferences as the first-output-layer neurons with strong excitatory connections. In this neuron, excitatory and inhibitory areas are separated from each other (**Figure 2A**). The optimal grating (the leftmost grating of **Figure 4C**) does not cover the inhibitory area. To systematically examine the suppression of the second-output-layer neurons, we first choose the optimal disk of gratings with the preferred radius for each neuron, and then enlarge them by 6 pixels. **Figure 4D** shows that enlarging the radius of the optimal gratings decreases the output of most of the second-output-layer neurons. The activity of a second-output-layer neuron is large if the presented grating is restricted to the area of the receptive fields of first-output-layer neurons with excitatory connections, whereas the activity diminishes if the presented grating covers the receptive fields of first-output-layer neurons with inhibitory connections. Jones et al. (2002) reported that iso-orientation suppression was found in 94% of the V1 cells. Similarly, almost all of the second-output-layer neurons (396/400) in Model 1 exhibit the surround suppression. **Figure 4E** shows the response of the same neuron as in **Figure 4C** to the surrounding annulus for various orientations in the presence of the inner patch of grating at its preferred orientation. When the orientation of the grating of the outer annulus deviates from the preferred orientation, the degree of suppression diminishes. This type of response, which is classified as iso-orientation suppression (Jones et al., 2002), is a result of the convergence of the excitatory and inhibitory connections from first-output-layer neurons with similar orientation preferences. The first-output-layer neurons with inhibitory connections correspond to the silent surroundings outside the classical receptive field of complex cells. Because the surrounding gratings perpendicular to the center grating suppress the response, this model neuron is classified into the group called "mixed general suppression and orientation alignment suppression" in Jones et al. (2002). The responses to the stimulus with the surrounding grating perpendicular to the center grating are greater than that with the surrounding grating parallel to the center in 376 out of 400 second-output-layer neurons ($p < 0.01$, binomial test). However, the responses to the stimulus with the surrounding grating perpendicular to the center grating are greater than that to the center stimulus alone in only 175 second-output-layer neurons. Jones et al. (2002) reported that 63% of neurons in V1 exhibited cross-orientation facilitation. Model 1 fails to predict the result.

These results suggest that the second-output-layer neurons acquire phase insensitivity and complex cell-like receptive field properties. However, the simulation of Model 1 is computationally expensive, and the learning is very slow to converge. Therefore, we examine the receptive field properties of the neurons in a simplified model.

3.2. MODEL 2: INFORMATION MAXIMIZATION IN A TWO-LAYER NETWORK WITH SIGN-INVARIANT INPUT

Connection weights W_{ij}^{+} and W_{ij}^{-} take on close values in the three-layer network of Model 1 after learning. To examine why this relation holds, we constructed another model network. In Model 2, we maximize the entropy of outputs $\mathbf{u}$ and $\mathbf{z}$ (**Figure 1B1**). Here u_i is identical to the nonlinear transformation of independent component i and corresponds to a first-output-layer neuron in Model 1. Output $\mathbf{z}$ corresponds to second-output-layer neurons in Model 1, and is connected from $\mathbf{y}^{+}$ and $\mathbf{y}^{-}$ with connection matrices $\mathbf{W}^{+}$ and $\mathbf{W}^{-}$, respectively. We assume that the probability density functions of the outputs of first-output-layer neurons are even functions, and that the signs of these outputs are not correlated. This assumption is supported by the observation that less than 1% of the 2×2 contingency tables of the signs of two first-output-layer neurons contain entries that are greater than 0.26 or less than 0.24. Using this assumption, a simple calculation proves that the entropy of the output is maximized if $W_{ij}^{+} = W_{ij}^{-}$ (see Methods). Assuming that $W_{ij}^{+} = W_{ij}^{-} = \tilde{W}_{ij}$, Equation 54 leads to

$$z_i = f\left(-h_i + \sum_{1 \le j \le N} \tilde{W}_{ij}|u_j|\right) \tag{55}$$

(**Figure 1B2**). In this maximization, we use an ICA algorithm with N input units, $|u_i|$, and N output units, z_i. Model 2 reduces the computational load and accelerates the convergence of the learning.

Figure 5 shows the connection weights of 100 output neurons after learning. Because we assume that $W_{ij}^{+} = W_{ij}^{-}$ in this model, the connection weights to a second-output-layer neuron can be represented by a single box (**Figure 2B**). The inputs to the second output-layer neurons have sparse distributions, whose kurtosis is greater than 3, with the exception of only one model neuron. None of the outputs of first-output-layer neurons, u_i, have kurtosis greater than 3. The receptive fields for more than one third of the output neurons are similar to that of the output neuron shown in **Figure 2A**. Some of them have excitatory and inhibitory connections from input neurons with similar orientation preferences, e.g., the third neuron from the right in the second row.

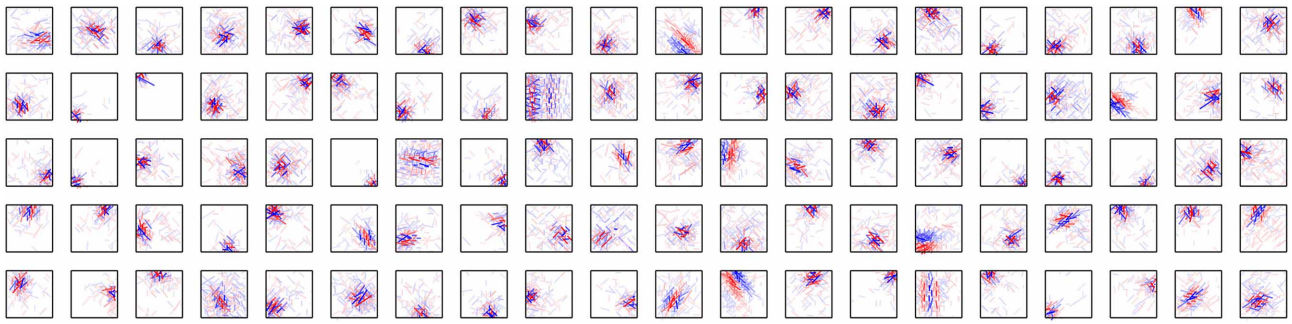

FIGURE 5 | Connection weights from the input neurons to 100 output neurons in Model 2.

The connection weights from the input neurons with the same orientation selectivity in distant areas of the image patches tend to have inverted signs. **Figure 6A** shows that this type of neuron is suppressed if a larger grating is presented, in a way similar to that shown in **Figure 4C**. **Figure 6B** shows that enlarging the radius of the optimal gratings decreases the output of most neurons. 393 out of 400 neurons of Model 2 also exhibit surround suppression. The neuron in **Figure 6C** is also suppressed by enlarging the radius of the gratings. However, the orientation preference and spatial alignment of input neurons sending strong inputs to the neuron in **Figure 6C** are different from those in **Figure 6A**. The neuron in **Figure 6A** is suppressed by parallel bars; in contrast, the neuron in **Figure 6C** is suppressed by a long bar. Thus, the neuron in **Figure 6C** is an end-stopped neuron (Hubel and Wiesel, 1968), whereas the neuron in **Figure 6A** is not. Some neurons have much more complex receptive fields and different response properties. **Figure 6D** shows that the activity of a second-output-layer neuron is suppressed if a grating perpendicular to the preferred orientation of this neuron is superimposed on the center of the receptive field of the neuron. This type of suppression was reported in complex cells (Bonds, 1989). This neuron has excitatory and inhibitory connections from input neurons with orientation selectivity perpendicular to the preferred orientation. In this type of neuron, the connections from the input neurons with orientation selectivity perpendicular to each other in the same area of the image patches tend to have inverted signs. This type of suppression is found in 264 out of 400 second-output-layer neurons ($p < 0.01$, binomial test). **Figure 6E** shows an example of the response classified as cross-orientation facilitation (Jones et al., 2002). This neuron is suppressed if the radius of the optimal grating is enlarged in a manner similar to that shown in **Figures 4C**, **6A**. When the orientation of the grating of the outer annulus deviates from that of the inner patch, the response exceeds the response to the optimal grating only (shown by the dashed line). In this type of neuron, the connections from the input neurons with orientation selectivity perpendicular to each other in distant areas of the image patches tend to have the same sign. This type of receptive field configuration facilitates the response of the neuron when the orientation of the outer annulus is perpendicular to the preferred orientation of the neuron. This type of facilitation is found in 369 out of 400 second-output-layer neurons ($p < 0.01$, binomial test). The responses to the stimulus with the surrounding grating perpendicular to the center grating are greater than that to the center stimulus alone in 149 second-output-layer neurons only. Model 2 also fails to predict the experimental result of cross-orientation facilitation (Jones et al., 2002). In some other neurons, the orientation preference is unclear (**Figure 5**). However, these neurons receive strong excitatory and inhibitory connections from input neurons whose preferred positions are restricted to a small area. Thus, these neurons respond selectively to edges in these areas.

Ringach et al. (2002) reported that the circular variance defined by

$$1 - \frac{\left|\sum_\theta R(z_i(\theta)) \exp(2i\theta)\right|}{\sum_\theta R(z_i(\theta))}, \tag{56}$$

where θ is the orientation of the grating and $z_i(\theta)$ is the output of neuron i in response to the grating, tends to be greater than 0.5 for complex cells. **Figure 6F** shows that the circular variance of output neurons of Model 2 tends to be greater than 0.5, which is consistent with experimental results of complex cells.

4. DISCUSSION

Our models differ from previous models in several ways. First, to generate sign-insensitive complex cells, Model 1 does not require inputs to be insensitive to the signs of pixels. The model in Hyvärinen and Hoyer (2001) assumes that complex cells receive the square of the output of simple cells. Complex cells in the model by Berkes and Wiskott (2005) are insensitive to the sign of image patches because their output is given by the degree two polynomials of pixel intensities. In the model by Shan et al. (2007), the output of simple cells is transformed to the absolute value and subjected to a nonlinear transformation. The models by Karklin and Lewicki (2005, 2009) also ignore the sign of the output of simple cells because the output of complex cells in these models depends only on the variance of the simple cells. In contrast, Model 1 can generate sign-insensitive complex cells without assuming sign-insensitive inputs to complex cells. The information maximization principle gives a possible explanation as to why sign-insensitive complex cells arise from sign-sensitive simple cell inputs. The results of Model 1 justify the assumption of Model 2 that the output neurons receive the absolute values of the output of simple cell-like neurons as inputs. Second, our models can

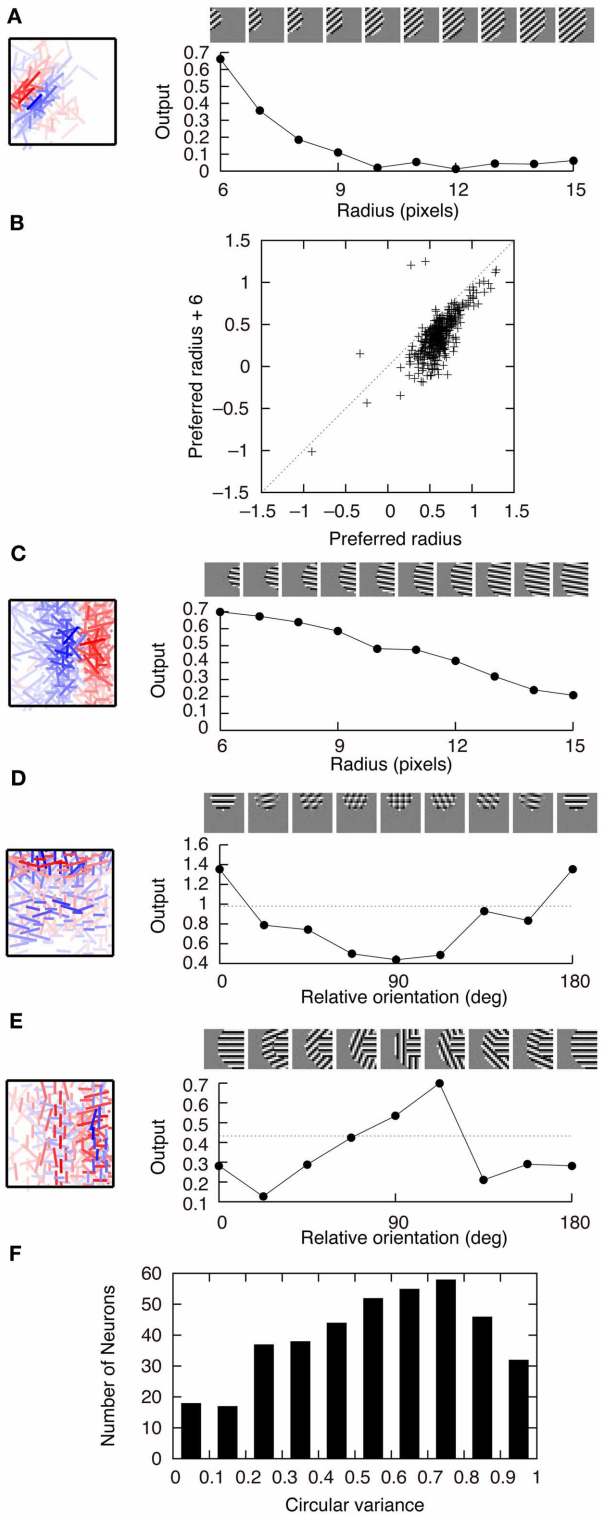

FIGURE 6 | Properties of output neurons in Model 2. (A) Change in the output of a neuron by the radius of the gratings. **(B)** Distribution of the output of 400 neurons in response to the optimal gratings with the preferred radius and to the same gratings whose radius is enlarged by 6 pixels. **(C)** End stopping exhibited by a neuron changing its output depending on the radius of the gratings. **(D)** Modification of the output of a neuron by the relative orientation of the gratings superimposed on the optimal grating. The dashed line indicates the output of the neuron in response to the optimal grating only. **(E)** Modulation of the output of a neuron by the orientation of the gratings of the outer annulus. The dashed line indicates the output of the neuron in response to the stimulus with the grating of the inner patch only. **(F)** Histogram of circular variance of orientation tuning of output neurons.

predict the receptive fields exhibiting surround suppression and facilitation. The complex cell-like receptive fields produced by the models of Földiák (1991) and Hyvärinen and Hoyer (2001) are devoid of these properties. The models of Földiák (1991) and Berkes and Wiskott (2005) also differ from our models in that their models require time-varying sequences of image patches as inputs. Although these models give results similar to our models, none of these models are based on the information maximization principle, but they are based on other optimization criteria such as the parameter fitting of the probability distributions and the minimization of the temporal change of the response.

In Model 2, we transform the i-th independent component, a_i, to $|u_i| = |f(a_i)|$. Shan et al. (2007) used a nonlinear function that transforms the absolute value of the independent components into a standard normal distribution. In Model 2, $|u_i|$ is almost uniformly distributed rather than normally distributed because the entropy of $f(a_i)$ is maximized when $f(a_i)$ is uniformly distributed in the range of the bounded function $f(x)$. Our results show that a uniformly distributed input can form complex cell-like receptive fields. The model proposed by Karklin and Lewicki (2005) also resembles Model 2. Each output unit of their model detects a specific set of covariances among input variables. For example, an output becomes large when $|x_1|$ and $|x_2|$ are large, and another output becomes large when $|x_1|$ is large and $|x_2|$ is small. The linear superposition of $|u_i|$ plays a similar role in Model 2. A large $|u_i|$ indicates that the absolute value of the i-th simple cell-like input is large, and a small $|u_i|$ indicates that the absolute value of the i-th simple cell-like input is small. Their model requires the sparseness of the output distribution, which is also required in Model 2, because the linear ICA algorithms can be derived by assuming the sparseness of the source distribution.

The nonlinearity of V1 complex cells was studied by using the models with the pooling of simple cell-like units (Sakai and Tanaka, 2000; Martinez and Alonso, 2001). We attempted to examine the second-order nonlinearity of the second-output-layer neurons. However, reverse correlation of these neurons did not exhibit second-order Wiener-like kernels that are similar to those observed for complex cells (Szulborski and Palmer, 1990). This is presumably because the widths of on- and off-regions of most first-output-layer neurons are as narrow as 1 pixel.

From our study, we speculate that complex features can be detected by combining the inputs from a large number of simpler elements. By increasing the number of output layers in our

models, we will have higher-order neurons that are selective to more complex features. The information maximization of a multilayer network with higher-order neurons would be capable of explaining the complicated selectivity of neurons in higher visual areas.

However, there are some limitations to our models, such as the fact that our models use feedforward networks. Although this structure simplifies the models and facilitates the derivation of the learning rules, cortical networks have feedback and recurrent structures as well as feedforward structures. It is known that the feedback from higher to lower sensory areas plays an essential role in sensory cortices. Bardy et al. (2006) reported that the inactivation of the feedback from the higher visual areas affected the selectivity of the neurons in V1. They found that this inactivation changed the responses of a substantial proportion of neurons classified as complex cells to simple cell-like responses, indicating that the feedback from higher visual areas modifies the receptive fields of complex cells. The response of complex cells appears to be formed by both a feedforward mechanism and a feedback and recurrent mechanism. Simple cells, which are assumed to provide inputs to complex cells in this paper, receive recurrent connections so that they exhibit surround suppression (Burr et al., 1981; Walker et al., 2000). Ringach et al. (2002) showed that simple cells with odd-symmetric receptive fields in V1 are greater in number than those predicted from ICA models. This may be the reason for which reverse correlation of the second-output-layer neurons did not exhibit Wiener-like kernels that are similar to those observed for complex cells. Although the receptive fields of some second-output-layer neurons in our models are irregular and different from model neurons with typical complex cell-like receptive fields, the introduction of higher-order neurons and recurrent connections (Tanaka et al., 2009) could cause the first-output-layer neurons to exhibit properties that are more similar to simple cells; in turn, it may increase the number of model neurons with complex cell-like receptive fields.

ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas "Mesoscopic Neurocircuitry" (No. 23115512) of The Ministry of Education, Science, Sports and Culture of Japan.

REFERENCES

Amari, S., Chen, T.-P., and Cichocki, A. (1997). Stability analysis of learning algorithms for blind source separation. *Neural Networks* 10, 1345–1351. doi: 10.1016/S0893-6080(97)00039-7

Amari, S., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. *Adv. Neural Inf. Process. Syst.* 8, 757–763.

Bardy, C., Huang, J., Wang, C., FitzGibbon, T., and Dreher, B. (2006). 'Simplification' of responses of complex cells in cat striate cortex: suppressive surrounds and 'feedback' inactivation. *J. Physiol.* 574, 731–750. doi: 10.1113/jphysiol.2006.110320

Barlow, H. B. (1959). "Sensory mechanisms, the reduction of redundancy, and intelligence," in *The Mechanisation of Thought Processes*, London: Her Majesty's Stationery Office. 535–539.

Bell, A., and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Bell, A., and Sejnowski, T. (1997). The "independent components" of natural scenes are edge filters. *Vision Res.* 37, 3327–3338. doi: 10.1016/S0042-6989(97)00121-1

Berkes, P., and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.* 5, 579–602. doi: 10.1167/5.6.9

Bonds, A. B. (1989). Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neurosci.* 2, 41–55. doi: 10.1017/S0952523800004314

Burr, D., Morrone, C., and Maffei, L. (1981). Intra-cortical inhibition prevents simple cells from responding to textured visual patterns. *Exp. Brain Res.* 43, 455–458.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory, 2nd Edn.* Hoboken, NJ: John Wiley and Sons, Inc.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.* 3, 194–200. doi: 10.1162/neco.1991.3.2.194

Hubel, D. H., and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol. (Lond.)* 148, 574–591.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Hubel, D. H., and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195, 215–243.

Hyvärinen, A., and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res.* 41, 2413–2423. doi: 10.1016/S0042-6989(01)00114-6

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis.* (New York, NY: John Wiley and Sons, Inc.) doi: 10.1002/0471221317

Jones, H., Grieve, K., Wang, W., and Sillito, A. (2001). Surround suppression in primate v1. *J. Neurophysiol.* 86, 2011–2028. doi: 10.1152/jn.00403.2001

Jones, H. E., Wang, W., and Sillito, A. M. (2002). Spatial organization and magnitude of orientation contrast interactions in primate V1. *J. Neurophysiol.* 88, 2796–2808.

Karklin, Y., and Lewicki, M. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Comput.* 17, 397–423. doi: 10.1162/0899766053011474

Karklin, Y., and Lewicki, M. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* 457, 83–86. doi: 10.1038/nature07481

Linsker, R. (1988). Self-organization in a perceptual network. *Computer* 21, 105–117. doi: 10.1109/2.36

Martinez, L. M., and Alonso, J.-M. (2001). Construction of complex receptive fields in cat primary visual cortex. *Neuron* 32, 515–525. doi: 10.1016/S0896-6273(01)00489-5

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Olshausen, B. A., and Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* 37, 3311–3325. doi: 10.1016/S0042-6989(97)00169-7

Palmer, J., Makeig, S., Delgado, K., and Rao, B. (2008). "Newton method for the ICA mixture model," in *Proceedings of the 33rd IEEE International Conference on Acoustics and Signal Processing (ICASSP 2008)*, (Las Vegas), 1805–1808. doi: 10.1109/ICASSP.2008.4517982

Ringach, D. L., Shapley, R. M., and Hawken, M. J. (2002). Orientation selectivity in macaque v1: diversity and laminar dependence. *J. Neurosci.* 22, 5639–5651.

Sakai, K., and Tanaka, S. (2000). Spatial pooling in the second-order spatial structure of cortical complex cells. *Vision Res.* 40, 855–871. doi: 10.1016/S0042-6989(99)00230-8

Shan, H., Zhang, L., and Cottrell, G. (2007). Recursive ICA. *Adv. Neural Inf. Process. Syst.* 19, 1273–1280.

Shapley, R., and Lennie, P. (1985). Spatial frequency analysis in the visual system. *Ann. Rev. Neurosci.* 8, 547–581. doi: 10.1146/annurev.ne.08.030185.002555

Shriki, O., Sompolinsky, H., and Lee, D. (2001). An information maximization approach to overcomplete and recurrent representations. *Adv. Neural Inf. Process. Syst.* 13, 612–618.

Skottun, B., De Valois, R., Grosof, D., Movshon, J., Albrecht, D., and Bonds, A. (1991). Classifying simple and complex cells on the basis of response modulation. *Vision Res.* 31, 1079–1086. doi: 10.1016/0042-6989(91)90033-2

Szulborski, R. G., and Palmer, L. A. (1990). The two-dimensional spatial structure of nonlinear subunits in the receptive fields of complex cells. *Vision Res.* 30, 249–254. doi: 10.1016/0042-6989(90)90040-R

Tanaka, T., Kaneko, T., and Aoyagi, T. (2009). Recurrent infomax generates cell assemblies, neuronal avalanches, and simple cell-like selectivity. *Neural Comput.* 21, 1038–1067. doi: 10.1162/neco.2008.03-08-727

Walker, G., Ohzawa, I., and Freeman, R. (2000). Suppression outside the classical cortical receptive field. *Visual Neurosci.* 17, 369–379. doi: 10.1017/S0952523800173055

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 July 2013; accepted: 26 October 2013; published online: 21 November 2013.

*Citation: Tanaka T and Nakamura K (2013) Information maximization principle explains the emergence of complex cell-like neurons. Front. Comput. Neurosci. **7**:165. doi: 10.3389/fncom.2013.00165*

This article was submitted to the journal Frontiers in Computational Neuroscience.