Replicating receptive fields of simple and complex cells in primary visual cortex in a neuronal network model with temporal and population sparseness and reliability

Takuma Tanaka[1], **Toshio Aoyagi**[2, 3], **Takeshi Kaneko**[4]

[1]Department of Morphological Brain Science, Graduate School of Medicine, Kyoto University, Japan.*

[2]Department of Applied Analysis and Complex Dynamical Systems, Graduate School of Informatics, Kyoto University, Japan.

[3]JST, CREST.

[4]Department of Morphological Brain Science, Graduate School of Medicine, Kyoto University, Japan.

Keywords: Sparse coding, Primary visual cortex, Learning

Abstract

We propose a new principle for replicating receptive field properties of neurons in the primary visual cortex. We derive a learning rule for a feedforward network, which maintains a low firing rate for the output neurons (resulting in temporal sparseness) and which allows only a small subset of the neurons in the network to fire at any given time (resulting in population sparseness). Our learning rule also sets the firing rates of the output neurons at each time step to near-maximum or near-minimum levels, resulting in

*Present address: Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Japan.

neuronal reliability. The learning rule is simple enough to be written in almost spatially and temporally local forms. After the learning stage is performed using input image patches of natural scenes, output neurons in the model network are found to exhibit simple-cell-like receptive field properties. When the output of these simple-cell-like neurons are input to another model layer using the same learning rule, the second-layer output neurons after learning become less sensitive to the phase of gratings than the simple-cell-like input neurons. In particular, some of the second-layer output neurons become completely phase invariant, owing to the convergence of the connections from first-layer neurons with similar orientation selectivity to second-layer neurons in the model network. We examine the parameter dependencies of the receptive field properties of the model neurons after learning and discuss their biological implications. We also show that the localized learning rule is consistent with experimental results concerning neuronal plasticity and can replicate the receptive fields of simple and complex cells.

1 Introduction

Neuronal networks and their learning rules have evolved to maximize metabolic and information efficiencies in coding stimuli by reducing redundancy in the coded signals (Barlow, 2001). Neuronal networks can reduce redundancy in response to natural stimuli that exhibit statistical regularities, thereby reducing the metabolic cost required for spike generation. For example, neurons in the visual cortex do not respond to constant stimuli with constant firing rates. Because luminance in the visual field changes infrequently, these neurons can accommodate and reduce their firing rates for constant stimuli and increase their firing rates only when the stimuli change. This accommodation allows the neurons to reduce the number of spikes without reducing or losing information about the stimuli. This principle of redundancy reduction explains phenomena such as light and dark adaptation, lateral inhibition, coding of motion, and the orientation selectivity of simple cells, and the accommodation of sensory discharges to constant stimuli.

Barlow (2001) proposed sparse coding as a way for neurons to represent stimuli with reduced firing rates. Studies by Attwell and Laughlin (2001) and Lennie (2003) estimated the energy consumed by neurons in the brain, and according to Lennie's estimate, the average spike rate is as low as 0.16 spikes/s/neuron. Sensory input is thus represented by this small number of firings in the cerebral cortex, with each neuron discharging spikes at a low rate and only a small number of neurons in a large population active at any given time.

The most successful models based on the idea of sparse coding are those with simple cells in the primary visual cortex (V1). Olshausen and Field (1996) showed that when natural images are represented with minimal neuronal activity, receptive field properties similar to those found in the simple cells of V1 emerge. Sparse coding appears to be closely related to independent component analysis (ICA). For example, Bell and Sejnowski (1997) showed that the decomposition of natural scenes into independent sparse representations replicates the receptive fields of simple cells. In addition to these statistical and information-theoretical models, Falconbridge et al. (2006) showed that the biologically based neuronal network model proposed by Földiák (1990) can replicate the receptive fields of simple cells.

The most successful model of complex cells to date is the model proposed by Karklin and Lewicki (2009), which is based on probability estimations of natural scenes and which assumes sparse neuronal activity. This model replicates the emergence of complex-cell-like receptive fields and of receptive fields for spirals that have been observed in neurons in higher visual areas (Gallant et al., 1996). However, the model has overly complicated dynamics, and the learning rule, in which the output of each complex cell depends on the output of other complex cells in the network, is not biologically plausible because of this non-local property.

In this paper, we propose a new model that replicates the receptive fields of not only simple cells but also complex cells. The remainder of this paper is organized as follows. In Section 2, we define temporal and population sparseness, which constitute two measures of the sparseness of neuronal firings in a network, and describe how to increase firing sparseness in a feedforward network. In Section 2, we first identify reliability as an additional key property of functional neuronal networks, which must be imposed on our model because random and input-independent firing patterns can be perfectly sparse. Then we derive a learning rule that improves both sparseness and reliability in a feedforward network. In Section 3, we show that our proposed model replicates the receptive fields of simple cells in V1 and examine how the neurons' receptive field properties depend on the values of the model's parameters. We then show that if the simple-cell-like output of a layer is fed to another layer that has the same learning rule, the model neurons in the second layer become less phase sensitive, exhibiting one of the properties of complex cells in V1. We also derive a spatially and temporally local learning rule and offer a physiological interpretation of the local learning rule. In Section 4, we compare our proposed model with previous models. We also discuss the limitations of our proposed model along with some open questions, describing how to extend our proposed model to overcome these limitations.

2 Models

2.1 Concepts

Sparse coding includes two concepts: temporal sparseness and population sparseness.

Temporal sparseness Willmore and Tolhurst (2001) defined temporal sparseness as lifetime sparseness. A neuron's activity is temporally sparse if its average firing rate over time is low. Figure 1A shows an example of a temporally sparse firing pattern in which each neuron fires only once, whereas Fig. 1D shows the example in which the firing rate of the neurons at the top is much more frequent and is therefore not temporally sparse. From a biological viewpoint, because decreasing the average firing rate of a neuron reduces the energy required for generating spikes, temporally sparse code is advantageous for survival.

Furthermore, temporally sparse codes represent external stimuli in a biologically plausible manner, for example, the idea of temporal sparseness underlies the most widely used ICA algorithms (Bell and Sejnowski, 1995; Olshausen and Field, 1996; Bell and Sejnowski, 1997; Karklin and Lewicki, 2005, 2009). In these algorithms, the probability distribution of each component is assumed to have high kurtosis, and thus, the outputs rarely have large values. The temporally sparse coding used in these algorithm can extract statistically independent edge-like components from natural scenes, components for which simple cells in V1 are also selective. Thus, temporally sparse code is an efficient and biologically plausible code for representing stimuli.

It is important to note that the definition of temporal sparseness in previous models and in the present model is independent of the temporal succession of input patterns that are fed to the model network. What matters in temporal sparseness is the average firing rate, and a temporally sparse firing pattern is not necessarily a firing sequence with low temporal correlation. Neurons in our model will fire frequently during a short time interval when they are given a series of inputs for which they are selective.

Population sparseness In population sparseness of Willmore and Tolhurst (2001) and the sparse-dispersed encoding of Field (1987), only a small subset of the coding population is active for each stimulus, and different small subsets of the population will be activated by different stimuli (Willmore and Tolhurst, 2001). In population-sparse code, because only a small subset of neurons in a network fire at each time step, there is minimal redundancy, which is desirable from a biological viewpoint. Figure 1B shows an example of population-sparse firings, with only one or two neurons firing at each time step. In contrast, every neuron in Fig. 1C fires at $t = 2$, and so this pattern is not population sparse.

Note that temporally sparse firing patterns are not necessarily population sparse. For example, although all the neurons in Fig. 1C have temporally sparse firing rates, the overall pattern is not population sparse because of the synchronous burst of firing. Similarly, Fig. 1D shows that population-sparse firing patterns are not necessarily temporally sparse.

Reliability An efficient representation of external stimuli in a neuronal network requires not only temporal and population sparseness but also reliable output in response to stimuli. A sensory neuron should respond to some specific stimuli with a high firing rate and to other stimuli with a low firing rate. In other words, the neuron must reliably respond to inputs. Figure 1E shows a reliable neuron that fires with high probability for some stimuli and with low probability for the others, while Fig. 1F shows an unreliable neuron that fires with an intermediate probability in response to all stimuli.

2.2 Model Network and Learning Algorithm

In this section, we describe our proposed model. Our neuronal network is a feedforward network in which the output $y_i(t)$ of neuron i at time t depends only on the input vector $[x_j(t)]$ at time t, and thus independent of earlier inputs and the states of other output neurons at time t. The output $y_i(t)$ of neuron i corresponds to the firing rate of the neuron in response to the stimulus $[x_j(t)]$ presented at time t. We assume that the firing rate of each output neuron ranges between 0 and 1. N output neurons receive input from M input neurons according to the weight matrix (W_{ij}). The output of neuron i at time t is calculated as

$$y_i(t) = \frac{1}{1 + \exp\Big(-\sum_{1 \le j \le M} W_{ij} x_j(t) + h_i(t)\Big)}, \tag{1}$$

where $x_j(t)$ is the firing rate of input neuron j at time t and $h_i(t)$ is the firing threshold for neuron i at time t.

All simulations except those shown in Figs. 2B and 8B were performed for $10\,000$ blocks, each containing $T = 10\,000$ time steps. The simulation in Fig. 2B was performed for $30\,000$ blocks because the connections between input and model neurons developed slowly in the simulation. The simulation in Fig. 8B was performed for 3000 blocks to reduce simulation time. Learning was implemented by updating the weight matrix (W_{ij}). We used batch learning process to accelerate the simulation. Although batch learning does not seem to be biologically plausible, the online learning described later in this paper produces results that are similar to the formation of the receptive fields of simple cells (Fig. 8). Our method is both simple and biologically plausible.

We imposed temporal sparseness using equation 2 to update the threshold $h_i(t)$ for each output neuron i at each time step t with the mean firing rate of neuron $\bar{p}$,

$$h_i(t+1) = h_i(t) + \epsilon(y_i(t) - \bar{p}). \tag{2}$$

This threshold update corresponds to the homeostatic plasticity observed by Desai et al. (1999). If neuron i fires too frequently, the threshold h_i rises; conversely, if neuron i fires too rarely, the threshold drops. In all simulations presented in this paper, we set the output neurons' mean firing rates to values less than 0.05 and the parameter ϵ to 0.01. These values produced a stable emergence of the receptive field properties similar to those found in the neurons in V1.

To maximize population sparseness, we derived a function that measures population sparseness. As shown in Fig. 1B, firings in a neuronal network are more population sparse when they are uncorrelated, while firings are less population sparse when they are correlated (i.e., if most neurons in the network fire simultaneously). Consequently, a population's firing rate at each time step is either very high and or very low if it is less population sparse. As an example, all the neurons in Fig. 1C fire at certain time steps, whereas they are all quiescent at other time steps. We can therefore measure population sparseness as the sign-inverted summation of the firing correlation among the model neurons,

$$S = -\mathbb{E}\left[\sum_{\substack{1\le i\le N\\ 1\le j\le N\\ i\ne j}} y_i(t)y_j(t)\right],$$

where $\mathbb{E}[\cdot]$ is the block temporal average over the inputs, that is, $\mathbb{E}[f(t)] = \frac{1}{T}\sum_{1\le t\le T} f(t)$. Maximizing S anti-correlates the activity of the neurons, thereby making the firing patterns population sparse. Although functions other than S could be used to measure population sparseness, the simplicity of S allowed us to effectively maximize population sparseness by using a simple learning rule.

Similarly, we define a function to measure reliability so that we could maximize reliability. A reliable model neuron should respond to some specific stimuli with a high firing rate and to other stimuli with a low firing rate. We define the reliability function R_i for neuron i as

$$R_i = \mathbb{E}\left[y_i(t)^2\right],$$

where different stimuli are given to the network at each time step in order to obtain the block temporal average. Because $f(x) = x^2$ is a convex function, this function is maximized when the firing rate $y_i(t)$ reaches its maximum 1 in $\bar{p}T$ out of the T time steps and 0 in the other time steps. In other words, the reliability function R_i is maximized if neuron i exclusively responds to $\bar{p}T$ out of T stimuli.

Combining the functions that measure population sparseness and reliability, we obtain the objective function given as

$$\begin{aligned} F &= \alpha \sum_{1 \le i \le N} R_i + \beta S \\ &= \alpha \mathbb{E}\left[\sum_{1\le i\le N} y_i(t)^2\right] - \beta \mathbb{E}\left[\sum_{\substack{1\le i\le N \\ 1\le j\le N \\ i\ne j}} y_i(t) y_j(t)\right]. \end{aligned} \tag{3}$$

This objective function is maximized through changes in the connection weight matrix (W_{ij}) from input to output neurons. A small ratio α/β makes the neurons in the network more population sparse and less reliable, whereas a large ratio α/β makes the neurons in the network less population sparse and more reliable. Parameters α and β should be tuned appropriately for each learning problem. Because the first and second terms are the summations of N and $N(N-1)$ terms, respectively, β should be scaled proportionally to $1/(N-1) \approx 1/N$ in order to balance population sparseness and reliability. We present the value of $\beta' = N\beta$ instead of β in the following specification of simulation parameters. For most of our simulations, the value of the parameter α is 1, and the value of β' is 1 in all simulations except for the simulation shown in Fig. 2E.

To adjust the connection weight matrix, we differentiate F with respect to z_i (W_{ij} or h_i) as follows:

$$\frac{\partial F}{\partial z_i} = \mathbb{E}\left[\frac{\partial y_i(t)}{\partial z_i}\left(2\alpha y_i(t) - 2\beta \sum_{\substack{1\le k\le N \\ k\ne i}} y_k(t)\right)\right]. \tag{4}$$

Substituting

$$\frac{\partial y_i(t)}{\partial W_{ij}} = y_i(t)[1 - y_i(t)]x_j$$

and

$$\frac{\partial y_i(t)}{\partial h_i} = -y_i(t)[1 - y_i(t)]$$

into this equation, we obtained the following gradients of the objective function with respect to W_{ij} and h_i:

$$\frac{\partial F}{\partial W_{ij}} = \mathbb{E}\left[y_i(t)[1-y_i(t)]x_j \left(2\alpha y_i(t) - 2\beta \sum_{\substack{1\le k\le N \\ k\neq i}} y_k(t) \right) \right], \tag{5}$$

$$\frac{\partial F}{\partial h_i} = \mathbb{E}\left[-y_i(t)[1-y_i(t)] \left(2\alpha y_i(t) - 2\beta \sum_{\substack{1\le k\le N \\ k\neq i}} y_j(t) \right) \right]. \tag{6}$$

Because the threshold h_i is adjusted by equation 2 to set the mean firing rate $\mathbb{E}[y_i(t)]$ to $\bar{p}$, h_i is dependent on W_{ij}. In other words, h_i is an implicit function of W_{ij} defined by

$$\mathbb{E}[y_i(t,\, \{W_{ij}\},\, h_i)] = \bar{p}.$$

Thus, the gradient of F with respect to W_{ij} is given by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}W_{ij}} F(t,\, \{W_{ij}\},\, \{h_i\}) &= \frac{\partial F}{\partial W_{ij}} + \frac{\partial F}{\partial h_i}\frac{\partial h_i}{\partial W_{ij}} \\ &= \frac{\partial F}{\partial W_{ij}} - \frac{\partial F}{\partial h_i}\frac{\partial}{\partial W_{ij}}\mathbb{E}[y_i(t)] \Big/ \frac{\partial}{\partial h_i}\mathbb{E}[y_i(t)] \\ &\equiv \Delta W_{ij}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} \frac{\partial}{\partial h_i}\mathbb{E}[y_i(t)] &= -\mathbb{E}\left[y_i(t)[1-y_i(t)]\right], \\ \frac{\partial}{\partial W_{ij}}\mathbb{E}[y_i(t)] &= \mathbb{E}\left[y_i(t)[1-y_i(t)]x_j\right]. \end{aligned}$$

The learning rule for the connection weight matrix W_{ij} is given by

$$W_{ij} \leftarrow W_{ij} + \eta \Delta W_{ij},$$

where η, the learning rate, was set to 1000 in all simulations in this paper. In our simulations, the connection weight matrix W_{ij} was updated at the end of each block, and the average of f in each block was used as the expectation $\mathbb{E}[f]$ in the above equations. We did not impose any bounds on the value of W_{ij}. To ensure that h_i converged, we performed the calculation for $10\,000$ time steps, updating h_i with equation 2 before starting the next block. Before starting the first block, we performed a simulation consisting of $500\,000$ steps in order to ensure the convergence of $\mathbb{E}[y_i(t)]$ to $\bar{p}$.

We preprocessed natural image scenes using the method described in Olshausen and Field (1997) in order to generate the firing patterns of the input neurons. We used randomly selected 16×16 image patches from the preprocessed images and converted the pixels in these image patches to 256 real-valued inputs. This input corresponds to the output of the neurons in the lateral geniculate nucleus (LGN), because the receptive fields of these neurons have a spatial profile similar to the inverse Fourier transform of the whitening filter (Olshausen, 2003).

3 Results

3.1 Simple-cell-like receptive field in the first layer

We first examined the receptive field properties of neurons in the first layer of the feed-forward network after learning, using image patches of size 16×16 from the natural image scenes as input. This network layer had 256 input neurons, corresponding to the pixels in the input image patches. At each time step, a new image was used as input. There were also $N = 256$ output neurons in the network, and we used the parameter values $\bar{p} = 0.01$, $\alpha = 1$, $\beta' = 1$, $\eta = 1000$, and $\epsilon = 0.01$. At the beginning of the simulation, W_{ij} was drawn from a uniform distribution on $[-0.5, 0.5]$ and h_i was set to 0.

The connection weights after learning show that most of the model neurons have receptive fields that are selective to edge-like stimuli (Fig. 2A1). For example, the leftmost connection weight in the top row of Fig. 2A1 shows that this neuron responds to a vertical edge. This edge detector-like receptive field is characteristic of the simple cells in V1 (Hubel and Wiesel, 1962).

We imposed temporal and population sparseness and reliability on the model neurons during the optimization process. Figure 2A2 shows the sparseness and reliability after learning by depicting the firing activity of the first 15 neurons in Fig. 2A1 in response to the image patches. The connection weights of these neurons, which determine the receptive field properties, are shown at the top of the figure. We presented the image patches shown on the left of Fig. 2A2 to these neurons. The black squares indicate that the firing rate of the neuron in the top row in response to the image patch to the left was close to 1. Gray and white squares indicate that the firing rate was low and close to 0, respectively. This figure shows that few neurons fired in response to a given image patch (population sparseness) and that each neuron fired in response to few image patches (temporal sparseness). The fact that most squares are black or white and few are gray means that the neurons were highly reliable. Thus, the optimization algorithm described in Section 2 successfully improved the sparseness and reliability of the model neurons. These results suggest that given the natural scene image patches as input, the model network successfully replicated the emergence of simple-cell-like receptive field properties for its neurons (Hubel and Wiesel, 1962). However, the receptive field properties of the neurons varied after learning, depending on the values of the parameters $\bar{p}$, α, and β'.

To examine this dependency of the output neurons' post-learning receptive field properties, Figs. 2B-E show the connection weights from the input to the output neurons that resulted from simulations with different values of $\bar{p}$, α, and β'. Figures 2B and 2C show the receptive field properties of the output neurons after learning with $\bar{p} = 0.002$ and $\bar{p} = 0.05$, respectively. The value of α in these simulations was set to 1, which is the same value used in the simulation shown in Fig. 2A1.

Figure 2C shows that with $\bar{p} = 0.05$, most neurons acquired a receptive field property with high spatial frequencies, while the others acquired a receptive field property with low spatial frequencies, unlike the receptive field properties of the neurons with $\bar{p} = 0.01$ (Fig. 2A1). In contrast, the receptive field properties of the neurons in the simulation with $\bar{p} = 0.002$ (Fig. 2B) are not very different from those of the neurons shown in Fig. 2A1.

We interpret these results as follows. When $\bar{p}$ is set to a large value, some neurons become selective to stimuli with low spatial frequencies. To make the output population sparse, these neurons prevent other neurons from becoming selective to stimuli with low spatial frequencies, and thereby making other neurons selective to stimuli with high spatial frequencies. Conversely, when $\bar{p}$ is small, competition among the output neurons is less severe.

Figures 2D and 2E show the receptive field properties of the output neurons after learning with $\alpha = 0$ and $\beta' = 0$, respectively. The average firing rate $\bar{p}$ in these simulations was set to 0.01, which is the same value used in the simulation shown in Fig. 2A1.

The ratio of α to β' determines the balance between population sparseness and reliability. Smaller α/β' ratios make the resultant network less reliable and more population sparse. We note that the neurons are temporally sparse and population sparse when they fire completely randomly and independently. Therefore, many neurons lost their selectivity, making the output of neurons population sparse in the simulation with $\alpha = 0$ (Fig. 2D). This figure shows that reliability is essential for neurons to exhibit input selectivity. Networks with large α/β' ratios are less population sparse and more reliable. A small β' allows neurons in the network to respond to similar stimuli (Fig. 2E) because the simultaneous firing of neurons is not prohibited in this case. Thus, the parameters $\bar{p}$, α, and β' affect the receptive fields of neurons after learning, with different effects.

3.2 Complex-cell-like receptive field in the second layer

In this section, we examine the receptive field properties of the second-layer model neurons to which the output of the neurons with simple-cell-like receptive fields is given as input. As shown in Fig. 3A, we first trained 1024 neurons with simple-cell-like receptive fields by performing a simulation with parameter values $\bar{p} = 0.01$, $\alpha = 1$, $\beta' = 1$, $\eta = 1000$, and $\epsilon = 0.01$, which are the same values that were used in Fig. 2A1. After completing the first layer's training, we trained the second layer using the output of these simple-cell-like model neurons as input with parameter values $N = 1024$, $\bar{p} = 0.04$, $\alpha = 1$, $\beta' = 1$, $\eta = 1000$, and $\epsilon = 0.01$ (Fig. 3A). At the beginning of the second-layer network simulation, we set the weights W_{ij} between the first and second layers with $W_{ii} = 1$ for all neurons i and $W_{ij} = 0$ $(i \neq j)$ for all other connections; this is done in order to improve the model neurons' selectivity after learning. The value of the objective function for the model initialized with $W_{ii} = 1$ and $W_{ij} = 0$ $(i \neq j)$ is larger than that for the model initialized with random W_{ij} values. Drawing W_{ij} from a uniform distribution produces a considerable number of non-selective neurons, and the large value of the objective function suggests that this is a better initial condition than random connections.

After 10 000 learning blocks, we examined the connections from the first-layer neurons to the second-layer neurons. To visualize these connections, we use a Gabor function to fit the connection weights of the first-layer neurons, as shown in Fig. 3B. In this figure, the first-layer neurons are represented by bars, and their orientation and spatial positions in the boxes reflect the optimal fit provided by the Gabor function.

In the boxes on the right of Fig. 3B, we plotted the bars corresponding to the first-layer neurons. The color of each bar indicates the signs (red indicates excitatory connections and blue indicates inhibitory connections) and magnitudes of the weights of the connections from the first-layer neuron. Figure 4 shows the connection strengths from the first-layer neurons to the first 32 second-layer neurons, again using red and blue to indicate excitatory and inhibitory connections, respectively.

Most second-layer neurons received strong excitatory connections from several first-layer neurons with similar orientation selectivity. Inhibitory inputs were much weaker than excitatory inputs. The orientation selectivity of inhibitory inputs to some second-layer neurons was parallel to that of excitatory inputs. For example, the leftmost neuron in the first row in Fig. 4 receives excitatory input from the first-layer neurons that are selective to vertical edges in the left side of image patches and inhibitory input from the neurons that are selective to vertical edges in the middle left of the image patches. Second-layer neurons tended to receive excitatory and inhibitory inputs from first-layer neurons with similar orientation selectivity.

To investigate the receptive fields of the neurons in the second layer after learning, we examined the model neurons' firing rates in response to grating stimuli. Previous experiments have successfully identified and characterized simple and complex cells by measuring the phase-dependent (F1) to phase-invariant (F0) component ratio (F1/F0 ratio) in their responses to optimal gratings (Skottun et al., 1991). By varying the phase of the gratings that cover the receptive fields of the neurons, researchers have been able to identify neurons with the F1/F0 ratio greater than 1 as simple cells and neurons with the F1/F0 ratio less than 1 as complex cells.

Accordingly, we presented the model neurons with gratings such as those shown in Fig. 5. We varied the orientation, spatial frequency, and phase of the gratings, and selected the best orientation and frequency for each neuron as follows. To quantify the phase sensitivity of the neurons, we first formed sets of gratings that have the same spatial frequency and orientation but with different phases: $0°,\ 10°,\ 20°,\ \ldots,\ \mathrm{and}\ 350°$. Then, we counted the number of gratings in each set for which a neuron's firing rate was greater than 0.5. We refer to this number, which ranges between 0 and 36, as the response number, and we define the set of gratings with the largest response number as the optimal set of gratings.

Our method differs from the method used in previous experiments to quantify the phase invariance of neurons. Researchers in previous experiments first selected the grating for which a neuron fired most strongly and then varied the phase of this grating to determine the neuron's phase invariance. This method does not work well in our model. Because our model maximized the reliability of the model neurons, they responded to most gratings with near-maximal ($y_i \approx 1$) and near-minimal ($y_i \approx 0$) firing rates, as shown in Fig. 5A. Thus, there were a large number of quasi-optimal gratings, and we could not choose one as the best stimulus. Therefore, we chose the optimal set of gratings with different phases for each model neuron, rather than an optimal grating with a fixed phase.

Similarly, we could not use the F1/F0 ratio to quantify phase invariance. The neurons in our model responded to most of the phase-shifted optimal gratings with near-maximal and near-minimal firing rates (Fig. 5A), and we observed a step-like profile rather than the sine curve-like profile that was observed in experiments (Bardy et al., 2006), Thus, we used the response number to quantify the phase sensitivity of the first-layer and second-layer neurons. In Figs. 5C1 and 5C2, we show the receptive fields of the second-layer neurons. The gratings shown around the boxes are the phase-shifted gratings to which neurons responded with a firing rate greater than 0.5. Gratings covering the entire 16×16 pixel area are shown in the inside circle, and half-sized gratings are shown in the outside circle. The second-layer neurons shown in Figs. 5C1 and 5C2 responded to a larger number of phase-shifted gratings than the first-layer neuron shown in Fig. 5B. In particular, the neuron in Fig. 5C2 was selective to oblique gratings and was completely phase-invariant. This suggests that second-layer neurons show more complex-cell-like properties than the first-layer neurons.

To quantitatively verify this observation, we compared the response numbers for each set of gratings. For example, the second-layer neuron shown in Fig. 5C2 responded to all the 36 phase-shifted optimal gratings, while the first-layer neurons shown in Fig. 5B responded to only 11 of the 36 phase-shifted gratings. Figure 6A shows a histogram of the response numbers for optimal gratings that caused firings at a rate greater than 0.5 in the first-layer and second-layer neurons. This histogram shows that no neurons in the first layer responded to more than 18 of 36 phase-shifted gratings, while in contrast, most of the neurons in the second layer responded to more than 18 of the 36 gratings. As shown in the histograms, neurons in the second layer were less sensitive to the phases of the gratings and some were even completely phase invariant, whereas neurons in the first layer were much more phase sensitive. The first-layer and second-layer neurons in our model seem to respectively correspond to simple and complex cells in V1.

Complex cells are reported to have classical receptive fields and silent surroundings. In most cases, stimuli presented in the silent surrounding suppress the response of the cell to the stimuli presented in the classical receptive field (Bardy et al., 2006). In a high proportion of cells, the suppression is greatest when the orientation of the gratings in the silent surrounding is the same as the best orientation for the classical receptive field. Reducing the area of the gratings sometimes increases the response of complex cells. This is consistent with the observation that second-layer neurons in our simulation tended to have excitatory and inhibitory inputs with similar orientation selectivity (Fig. 4). However, it is not clear from Fig. 4 to what extent these inhibitory inputs affect the model neurons' output. We therefore examined the effect of the silent surrounding of the second-layer neurons in the model. Using gratings whose length is half that of the image patches, we counted the response numbers of the neurons for the best gratings. The exterior gratings in Figs. 5C1 and 5C2 are the phase-shifted gratings to which neurons responded with a firing rate greater than 0.5. The response numbers of the second-layer neurons of Fig. 5C1 increased when we presented half-ranged gratings. Figure 6B shows a histogram of the response numbers in response to full-ranged and half-ranged gratings. Smaller gratings tend to make second-layer neurons more phase invariant. This shift of phase invariance is accounted for by the fact that smaller gratings are free from suppression by inhibitory inputs whose orientation selectivity is the same as the excitatory inputs. The receptive field, composed of an excitatory component and an inhibitory surrounding, is a result of our learning rule that makes neuronal firings sparse and reliable.

To examine the parameter dependency of the receptive fields after learning, we varied the values of the second-layer network parameter $\bar{p}$. The results are summarized in Fig. 7. By changing the value of $\bar{p}$ to 0.02, which determines the temporal sparseness of output neurons, we found that the second-layer neurons became more phase sensitive after learning, as shown in Fig. 7B. In both cases, nearly half of the neurons responded to more than 18 of 36 phase-shifted gratings, whereas the remaining neurons were as phase sensitive as the first-layer neurons.

Setting $\bar{p}$ to 0.01, we found that after learning, most second-layer neurons responded to fewer than 18 of the 36 phase-shifted gratings, as shown in Fig. 7C. That is, the second-layer neurons with $\bar{p} = 0.01$ failed to become phase invariant. Thus, the phase sensitivity of the second-layer neurons was highly dependent on the model's parameter values. Second-layer neurons with higher $\bar{p}$ values than the first-layer neurons tended to become less phase sensitive and thereby more complex-cell-like, because the second-layer neurons with higher $\bar{p}$ values must receive inputs from a larger number of the first-layer neurons than the second-layer neurons with lower $\bar{p}$ values.

3.3 Local learning rule

The learning rule for our model is neither temporally nor spatially local; however, the learning rule can be rewritten in a form that is almost spatially and temporally local. In equation 7, $\frac{\partial}{\partial h_i}\mathbb{E}[y_i(t)]$ and $\frac{\partial}{\partial W_{ij}}\mathbb{E}[y_i(t)]$ are spatially local but not temporally local. However, if we introduce neuron-specific variables $a_i(t)$ and synapse-specific variables $b_{ij}(t)$ such that

$$a_i(t+1) - a_i(t) = \left[-y_i(t)\left(1 - y_i(t)\right) - a_i(t)\right]/\tau$$

and

$$b_{ij}(t+1) - b_{ij}(t) = \left[y_i(t)\left(1 - y_i(t)\right)x_j - b_{ij}(t)\right]/\tau,$$

respectively, where τ is a sufficiently large time constant, then equation 7 can be approximated by the temporally local equation

$$\Delta W_{ij} = \frac{\partial F}{\partial W_{ij}} - \frac{\partial F}{\partial h_i}\frac{b_{ij}(t)}{a_i(t)}. \tag{8}$$

We obtain a spatially and temporally local learning rule if $\frac{\partial F}{\partial W_{ij}}$ and $\frac{\partial F}{\partial h_i}$ are approximated by spatially and temporally local equations. The first term on the right-hand side of equation 8 can be estimated by

$$c_{ij}(t) = y_i(t)\left(1 - y_i(t)\right)x_j(t)\left\{2\alpha y_i(t) - 2\beta[n(t) - y_i(t)]\right\}, \tag{9}$$

where $n(t) = \sum_{1 \le i \le N} y_i(t)$. This equation corresponds to equation 5 and requires only the population firing rate n and spatially local information. Similarly, $\frac{\partial F}{\partial h_i}$ can be estimated from the population firing rate n and spatially local information by

$$d_i(t) = -y_i(t)\,(1 - y_i(t))\,\{2\alpha y_i(t) - 2\beta[n(t) - y_i(t)]\}\,, \tag{10}$$

which corresponds to equation 6. Thus, each neuron requires only the population firing rate, its input, and its firing rate at each time step to update the connection weights as

$$\begin{aligned} \Delta W_{ij} &= c_{ij}(t) - d_i(t)\frac{b_{ij}(t)}{a_i(t)} \\ &= y_i(t)\,(1 - y_i(t))\,\{2\alpha y_i(t) - 2\beta[n(t) - y_i(t)]\}\left(x_j(t) + \frac{b_{ij}(t)}{a_i(t)}\right). \end{aligned} \tag{11}$$

In this form, the learning rule is almost spatially and temporally local. Spatially, each neuron's learning process requires only the population firing rate rather than the firing rate of each of the other neurons. The population firing rate can be provided for the synapse from neuron j to i by local interneurons. The information required for updating the synaptic strength can be stored in the neuron-specific value $a_i(t)$ and the synapse-specific value $b_{ij}(t)$, both of which are updated at each time step.

Equation 11 can be interpreted as follows. We ignore the factor $y_i(t)[1 - y_i(t)]$ because this factor is always positive. Assuming $\beta' \approx \alpha$, that is, $\beta \approx \alpha/(N-1)$, the factor $2\alpha y_i(t) - 2\beta[n(t) - y_i(t)]$ changes its sign depending on whether the firing rate of neuron i is greater than the average firing rate of the other neurons at time t. This factor introduces competition among the neurons. The factor $b_{ij}(t)/a_i(t)$ is a weighted average of x_j over time. This weighted average is dominated by the time steps at which $y_i(t) \approx 0.5$. Thus, $b_{ij}(t)/a_i(t)$ is close to the typical value of x_j in the image patches to which neuron i is intermediately selective. In other words, the neuron tends to fire if $x_j(t)$ is larger than $b_{ij}(t)/a_i(t)$. This means that the factor $x_j(t)+b_{ij}(t)/a_i(t)$ is positive if $x_j(t)$ falls within the range of the values of x_j in the image patches to which neuron i is selective, otherwise it is negative. The product of these factors, i.e., the synaptic update, is positive if the selectivity of neuron i to the input at t is unique among the population of neurons and $x_i(t)$ is sufficiently large or the selectivity to the input is not unique and $x_i(t)$ is small.

The present learning rule is a modified Hebbian learning rule. The BCM rule (Bienenstock et al., 1982), a well-known model of Hebbian learning, uses the product of input j and the difference between the activity of neuron i and a threshold. In our model, the activity of other neurons in the population determines the threshold, whereas in the BCM rule, the threshold is independent of the activity of other neurons. Similarly, our learning rule uses the difference between input j and the typical level of input j for the image patches to which neuron i is selective, rather than the input j itself. In other words, stronger input is required for increasing the strength of synapses that strongly influence the neuron's firing than synapses that does not influence the neuron's firing. This is consistent with experimental observations that large spines tend to be resistant to long-term potentiation (Matsuzaki et al., 2004).

The output neurons in our localized model exhibited simple-cell-like receptive field properties after learning from natural scenes (Fig. 8A), which are very similar to the receptive fields formed by the original model. In Fig. 8A, the value of parameter τ was set to 1000 and the those of other parameters were set to the same values used in Fig. 2A1. The receptive field of neuron i in Fig. 8A is similar to that of neuron i in Fig. 2A1 because we used the same initial weight W_{ij}. In a similar way, the connections from the first-layer neurons to the second-layer neurons after the localized learning (Fig. 8B) are similar to those shown in Fig. 4. We did not use this localized learning rule in the simulations in the previous sections because this localized form of the rule requires a much longer simulation time than the learning rule presented in Section 2.

In addition to the localized model, we examined a model network with stochastic model neurons that have a binary output value of 1 with the probability $y_i(t)$ and an output value of 0 with the probability $1 - y_i(t)$. Results obtained with the stochastic model were very similar to those discussed in the present paper.

4 Discussion

In this paper, we derived a learning rule that maximizes the temporal and population sparseness and neuronal firing reliability in a feedforward network. Using image patches from natural scenes as input, we found that after learning, the neurons in the network exhibited simple-cell-like receptive field properties. We then examined the effect of our model's parameter values on the simple-cell-like receptive fields of the model neurons. Using the output from these simple-cell-like neurons as input to a second-layer network, the neurons in the second-layer network tended to acquire greater phase invariance than the simple-cell-like neurons, that is, replicated complex-cell-like receptive fields. These results indicate that our proposed model successfully replicates the receptive fields of simple and complex cells in V1.

Previous models based on sparse coding have succeeded in explaining the receptive fields of simple cells. Földiák (1990) proposed a very simple model in which anti-Hebbian plasticity among the output neurons forms sparse representations of the input after the learning phase. Subsequently, Falconbridge et al. (2006) showed a similar model that was capable of extracting Gabor function-like components of natural scenes. The emergence of simple-cell-like receptive fields in our proposed model is consistent with these earlier results; however, our model's explanations of the receptive field properties of complex cells differ greatly from that of previous models. Previous models of complex cells utilize temporal correlation of the natural scenes or the local connectivity to train the complex cells. For example, the neuronal network models proposed in Földiák (1991) and Berkes and Wiskott (2005) used temporal sequences of gradually evolving images (i.e., gradually changing sequences of images) to develop complex-cell-like shift invariance. In another previous model of neurons in V1 (Hyvarinen and Hoyer, 2001), complex cells have fixed connections from local simple cells. The connection weights from the input layer to the simple cells are updated such that the firing of complex cells becomes sparse. This model assumes that simple cells connect to nearby complex cells, and the learning rule produces the local simple cell selectivity to edges with similar orientations and spatial frequencies.

Our model and the model proposed by Karklin and Lewicki (2009) are in sharp contrast to these previous models by virtue of neither requiring a temporal sequence of correlated images nor assuming local connectivity. In addition, the dynamics of the neurons in our model is simpler than those in previous models such as Földiák (1990), Falconbridge et al. (2006), and Karklin and Lewicki (2009), where ordinary differential equations must be solved to determine the firing states of the output neurons. This means that the neurons in these models interact with each other to decide whether to fire in response to a given stimulus. In contrast, the firing state of each output neuron in our model can be determined by a single equation (equation 1) that is independent of the firing states of other neurons.

The learning rule in our proposed model can be written in a form that is nearly spatially and temporally local form. The localized learning rule can replicate the results of the model described in Section 2. Our localized model explains the emergence of complex-cell-like receptive field properties without requiring a non-local learning rule. This makes our model more biologically plausible than the previous models. In addition, the localized learning rule supports an interpretation that is consistent with experimental results concerning the plasticity of pyramidal neurons. Although our model's neurons must be provided with information about the population firing rate (even in the localized form), this information can be transmitted to excitatory neurons from local inhibitory interneurons. This correlates with the observation that some subtypes of inhibitory neurons in the cortex are locally connected to almost every pyramidal cell (Fino and Yuste, 2011). Földiák (1990) and Falconbridge et al. (2006) also assumed interaction among local cells through inhibitory interneurons. These models assume a plasticity rule that strengthens inhibitory synapses between two simple cells if the simultaneous firing of the two cells occurs too frequently. This plasticity rule makes it difficult for simple cells to fire simultaneously and thereby makes the firing population sparse. Although neurons in the same layer do not interact in our model, $n(t) - y_i(t)$ in equation 9 plays a similar role. This term weakens input synapses that are activated when a large number of neurons fire. Thus, an effective mutual inhibition is introduced by this term without assuming the dynamics described by ordinary differential equations in Földiák (1990) and Falconbridge et al. (2006). The simplicity of our dynamics and the spatially and temporally local forms of our learning rule suggest that our model can be implemented by biological neurons.

This simplicity allows us to draw some predictions based on our model. First, changing the firing activity of neurons in the critical period can affect the development of the receptive fields of neighboring neurons whose activity is not changed. This is determined by the factor $2\alpha p_i(t) - 2\beta[n(t) - y_i(t)] = 2\alpha p_i(t) + 2\beta y_i(t) - 2\beta n(t)$ in equation 11. The direction of the plastic change of a neuron can be inverted by activating other neurons. If our speculation that the third term, $-2\beta n(t)$, is mediated by local inhibitory interneurons is correct, then modifying the strength of GABAergic synapses disrupts the selectivity of simple and complex cells after the critical period. More specifically, neurons will tend to have large receptive fields selective to low frequency edges if the GABAergic synapses are weakened. Many of our model's neurons tended to exhibit similar selectivity in this case (Fig. 2E). Second, changes in the excitability of neurons can shift the neurons' frequency tuning. Most of our model's neurons became selective to edges with higher spatial frequencies when we increased $\bar{p}$ (Fig. 2C). The phase sensitivity of complex cells can also be affected (Figs. 7A and 7B).

Nonetheless, there are some limitations to our proposed model. First, the homeostatic plasticity described by equation 2 may be non-biological. Recent research has revealed that there are two major ways to achieve homeostasis in neuronal activities; through intrinsic excitability and through the efficacy of individual synapses (Pozo and Goda, 2010). Equation 2 corresponds to the regulation of intrinsic excitability, because increasing and decreasing the threshold change the firing rate of a neuron in response to a given input without changing the synaptic strength. Increases and decreases of the firing threshold are not bounded in our model, even though plastic changes in biological neurons may be rather limited. The other way to achieve homeostasis, by scaling the efficacy of individual synapses, might be a more biologically based explanation of the stability of the average firing rate.

Second, our model's results depend on the value of parameter $\bar{p}$. In our model, the $\bar{p}$ value for the second-layer neurons must be greater than that for the first-layer neurons in order to replicate the selectivity of complex cells. This parameter setting is justified by the fact that the average firing rate of complex cells in response to dot stimuli and sine-wave gratings is two to three times greater than the firing rate of simple cells (Skottun et al., 1988). However, because the average firing rate depends on the stimuli used in the experiments, more realistic settings for $\bar{p}$ should be investigated in future works.

Third, our network model uses a feedforward network because this structure simplifies the model and facilitates the derivation of the learning rule. In contrast, the neuronal networks in the brain also have feedback and recurrent structures apart from feedforward structures, and it is known that feedback from higher to lower sensory areas plays an important role in sensory information processing. Bardy et al. (2006) reported that the inactivation of feedback from the posterotemporal visual cortex affected the selectivity of the neurons in V1. They found that this inactivation changed the responses of substantial proportions of neurons classified as complex cells in V1 to simple-cell-like responses, indicating that the feedback from higher sensory areas modifies and determines the receptive fields of complex cells to some extent. Thus, the receptive fields of the complex cells seem to be formed not only by a feedforward mechanism but also by a feedback or recurrent mechanism. Introducing higher-order neurons and providing feedback to the second-order neurons from these higher-order neurons would improve the receptive fields of the second-layer neurons in our model network.

Acknowledgments

This work was supported by Grants-in-Aid from the Ministry of Education, Science, Sports, and Culture (MEXT) of Japan: Grant numbers 23115512, 19GS0208, 21120002, 23115511, 21650083, 23115509, and 22300113. This work was also supported by Global COE Program "Center for Frontier Medicine", MEXT, Japan.

References

Attwell, D. & Laughlin, S. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21(10), 1133–1145.

Bardy, C., Huang, J., Wang, C., FitzGibbon, T., & Dreher, B. (2006). 'Simplification' of responses of complex cells in cat striate cortex: suppressive surrounds and 'feedback' inactivation. *J. Physiol.*, 574(3), 731–750.

Barlow, H. (2001). The exploitation of regularities in the environment by the brain. *Behav. Brain Sci.*, 24, 602–607.

Bell, A. & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.*, 7(6), 1129–1159.

Bell, A. & Sejnowski, T. (1997). The "independent components" of natural scenes are edge filters. *Vis. Res.*, 37(23), 3327–3338.

Berkes, P. & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.*, 5(6), 579–602.

Bienenstock, E., Cooper, L., & Munro, P. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2(1), 32–48.

Desai, N. S., Rutherford, L. C., & Turrigiano, G. G. (1999). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat. Neurosci.*, 2, 515–520.

Falconbridge, M., Stamps, R., & Badcock, D. (2006). A simple Hebbian/anti-Hebbian network learns the sparse, independent components of natural images. *Neural Comput.*, 18(2), 415–429.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4, 2379–2394.

Fino, E. & Yuste, R. (2011). Dense inhibitory connectivity in neocortex. *Neuron*, 69, 1188–1203.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.*, 64(2), 165–170.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Comput.*, 3, 194–200.

Gallant, J., Connor, C., Rakshit, S., Lewis, J., & Van Essen, D. (1996). Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J. Neurophysiol.*, 76(4), 2718–2739.

Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160(1), 106–154.

Hyvarinen, A. & 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.

Karklin, Y. & Lewicki, M. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Comput.*, 17(2), 397–423.

Karklin, Y. & Lewicki, M. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225), 83–86.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, 13(6), 493–497.

Matsuzaki, M., Honkura, N., Ellis-Davies, G., & Kasai, H. (2004). Structural basis of long-term potentiation in single dendritic spines. *Nature*, 429(6993), 761–766.

Olshausen, B. (2003). Principles of image representation in visual cortex. In Chalupa, L. & Werner, J., editors, *The visual neurosciences*, pages 1603–1615. Cambridge: MIT Press.

Olshausen, B. A. & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583), 607–609.

Olshausen, B. A. & Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.*, 37(23), 3311–3325.

Pozo, K. & Goda, Y. (2010). Unraveling mechanisms of homeostatic synaptic plasticity. *Neuron*, 66, 337–351.

Skottun, B., De Valois, R., Grosof, D., Movshon, J., Albrecht, D., & Bonds, A. (1991). Classifying simple and complex cells on the basis of response modulation. *Vis. Res.*, 31(7-8), 1078–1086.

Skottun, B., Grosof, D., & De Valois, R. (1988). Responses of simple and complex cells to random dot patterns: a quantitative comparison. *J. Neurophysiol.*, 59(6), 1719–1735.

Willmore, B. & Tolhurst, D. J. (2001). Characterizing the sparseness of neural codes. *Network*, 12, 255–270.

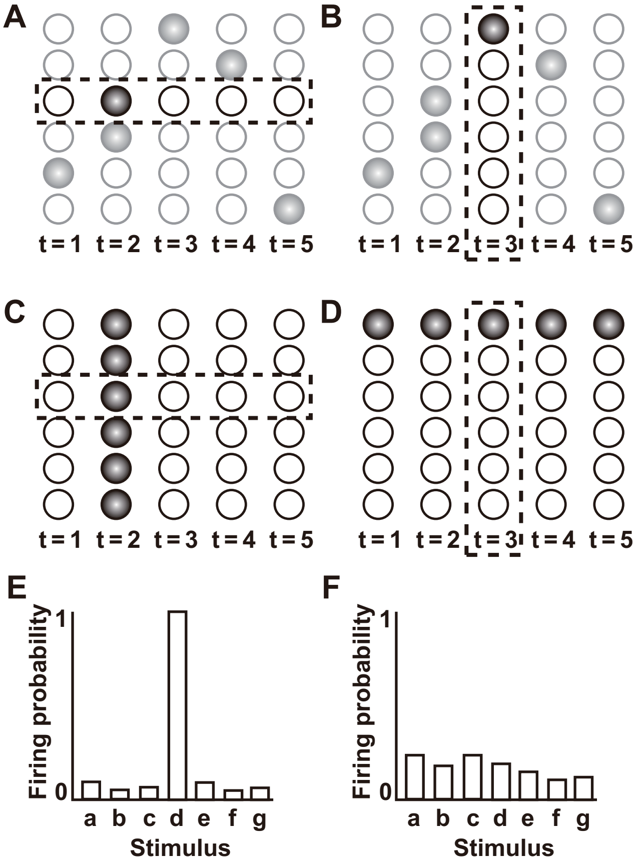

Figure 1: Temporal and population sparseness and reliability. (A) The third neuron from the top fires only once in five time steps; this firing of this neuron is therefore temporally sparse. (B) At $t = 3$, only one neuron fires; this firing is therefore population sparse at $t = 3$. (C,D) Temporal sparseness does not necessarily imply population sparseness and vice versa. (E) A reliable neuron responds to specific stimuli with a high firing rate and does not respond to other inputs; this neuron is selective to stimulus d. (F) An unreliable neuron fires with intermediate probabilities to all the stimuli.

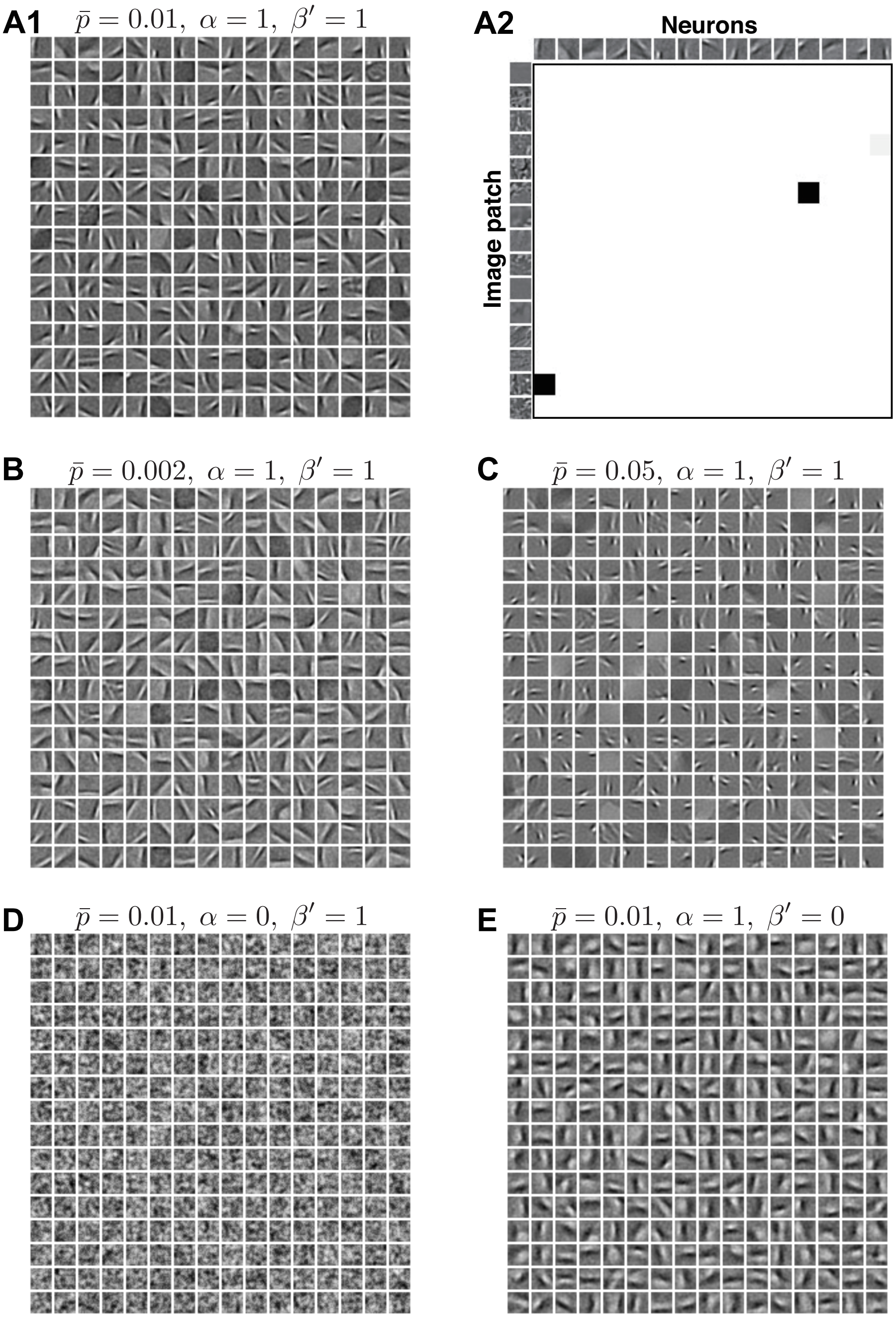
A1 $\bar{p} = 0.01,\ \alpha = 1,\ \beta' = 1$
A2
Neurons
Image patch
B $\bar{p} = 0.002,\ \alpha = 1,\ \beta' = 1$
C $\bar{p} = 0.05,\ \alpha = 1,\ \beta' = 1$
D $\bar{p} = 0.01,\ \alpha = 0,\ \beta' = 1$
E $\bar{p} = 0.01,\ \alpha = 1,\ \beta' = 0$

Figure 2: Connection weights reveal the simple-cell-like receptive field properties of the neurons after learning. (A1,B,C,D,E) The following parameter values are used: (A1) $\bar{p} = 0.01$, $\alpha = 1$, and $\beta' = 1$; (B) $\bar{p} = 0.002$, $\alpha = 1$, and $\beta' = 1$; (C) $\bar{p} = 0.05$, $\alpha = 1$, and $\beta' = 1$; (D) $\bar{p} = 0.01$, $\alpha = 0$, and $\beta' = 1$; and (E) $\bar{p} = 0.01$, $\alpha = 1$, and $\beta' = 0$. (A2) The activity of the first 15 neurons of the model shown in A1 in response to natural scenes. Black boxes indicate that the neuron with the connection weight shown on the top row responded to the image patch shown in the leftmost column.

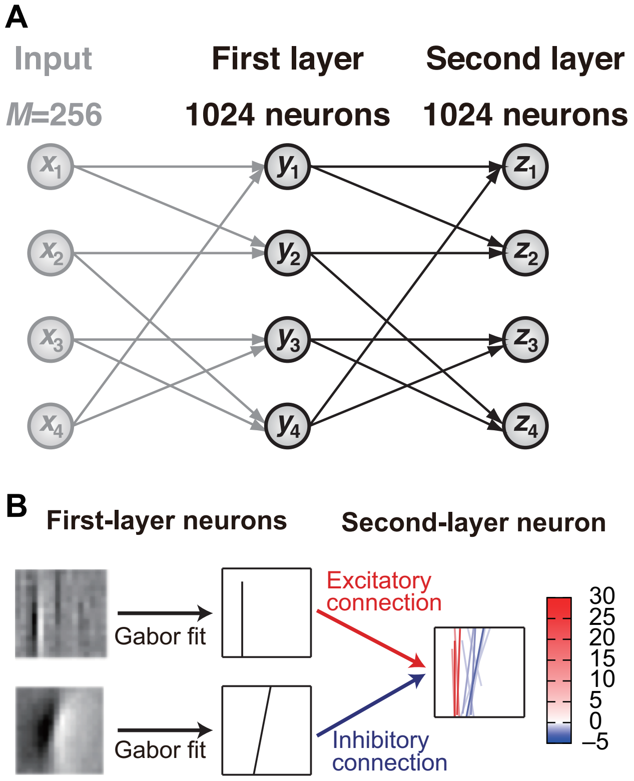

Figure 3: A two-layer feedforward network and the visualization of its connections. (A) The outputs from 1024 first-layer neurons with simple-cell-like receptive fields are used as the inputs to 1024 second-layer neurons. (B) The connection weights from the simple-cell-like first-layer neurons are fitted by a Gabor function and are represented by bars in the second-layer neurons; red indicates excitatory connections and blue indicates inhibitory connections.

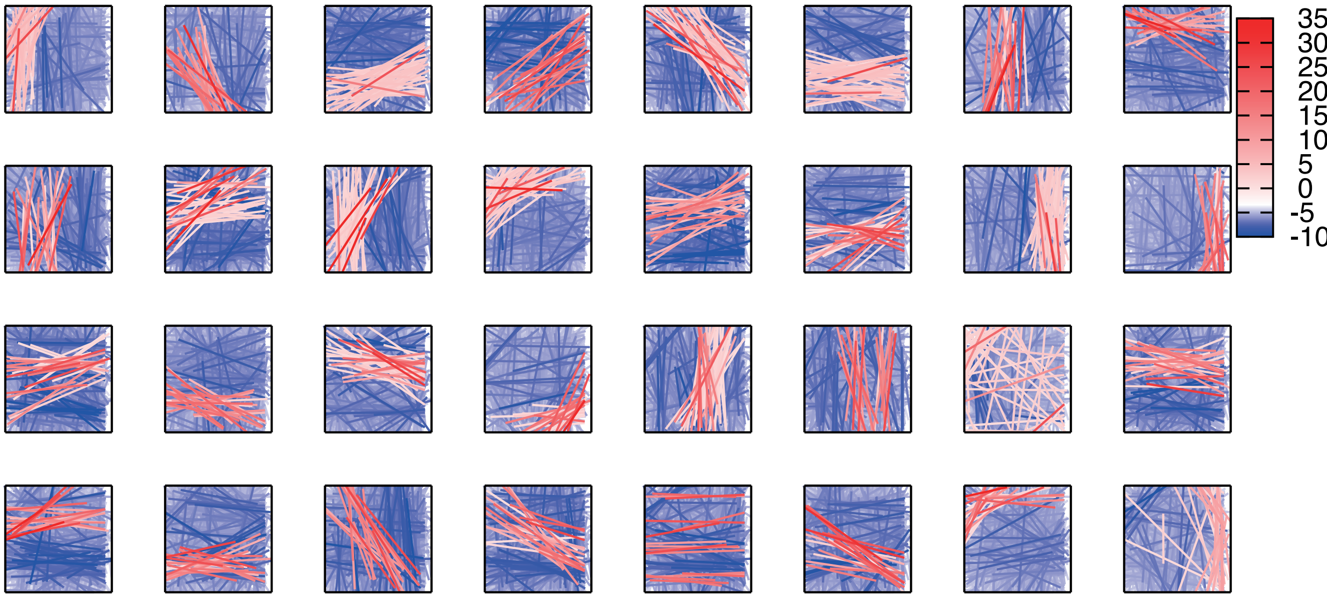

Figure 4: Excitatory (red) and inhibitory (blue) inputs from first-layer neurons to second-layer neurons.

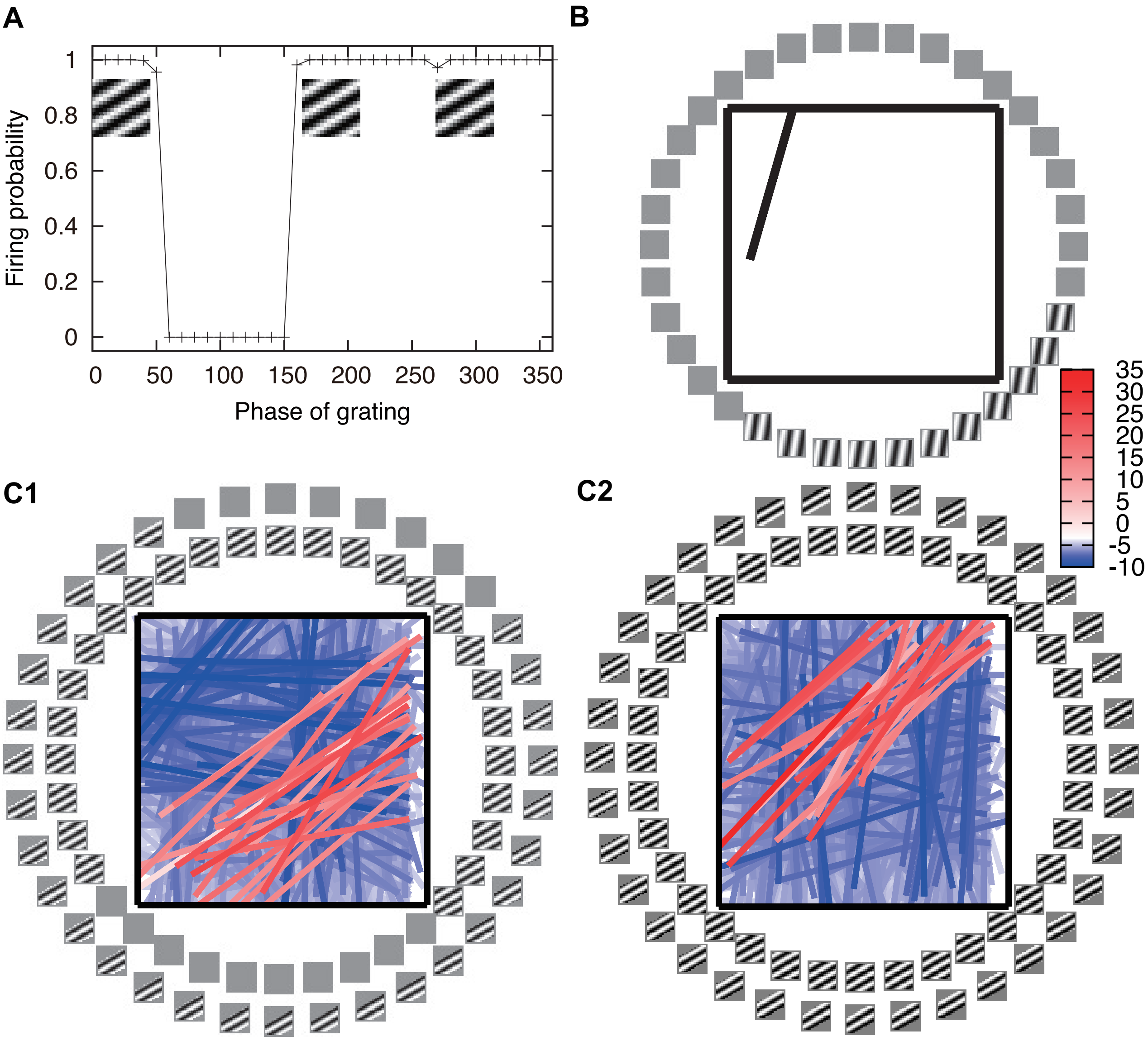

Figure 5: Receptive fields of first-layer and second-layer neurons in response to phase-shifted gratings. (A) Firing rate of the second-layer neuron shown in C1 in response to the phase-shifted gratings. (B) Receptive field of the first-layer neuron shown as the leftmost neuron in the first row of Fig. 2. (C) Excitatory (red) and inhibitory (blue) inputs from first-layer neurons to second-layer neurons is shown in boxes. For B, C1, and C2, the gratings that cause firings with a rate greater than 0.5 after learning are shown around the boxes. Gratings covering the entire area of 16×16 image patches are shown on the inside and half-ranged gratings are shown on the outside.

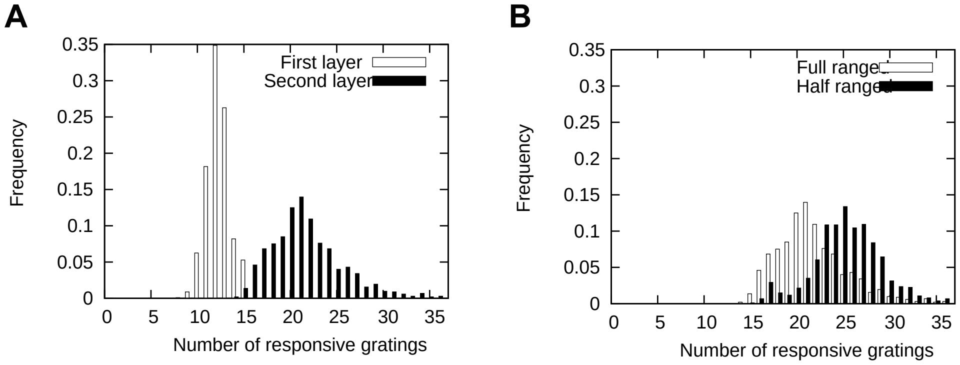

Figure 6: The response of second-layer neurons to gratings. (A) A histogram of the number of phases (out of a total of 36 phases) of gratings that caused model neurons to fire. No first-layer neuron fired in response to more than half of the phases. Most second-layer neurons with $\bar{p} = 0.04$ fired in response to more than half of the phases. (B) Second-layer neurons were more phase invariant in response to smaller gratings than in response to larger gratings.

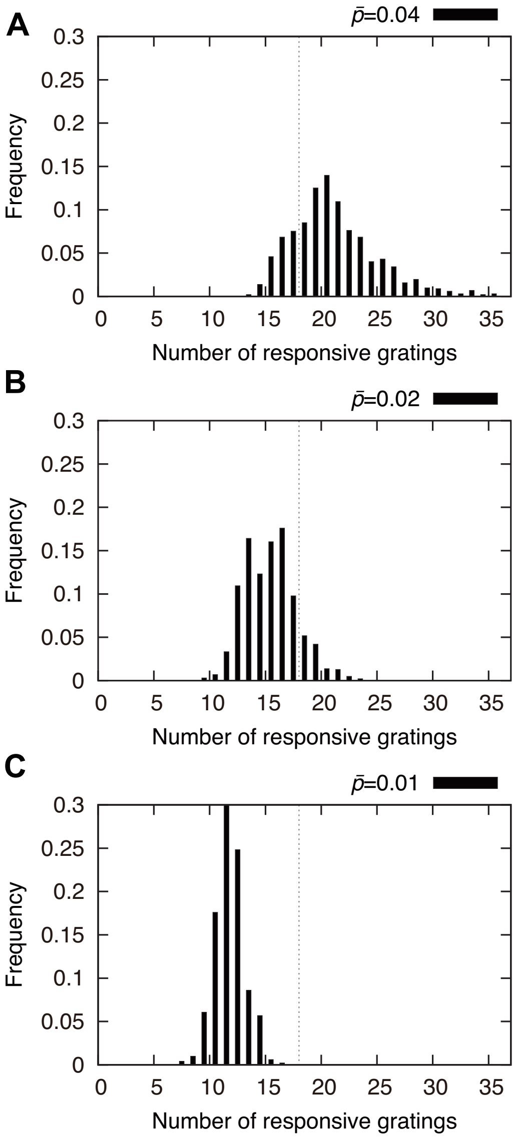

Figure 7: Histograms of the number of phases of gratings that caused output neurons with different parameter values to fire; the parameter differences were (A) $\bar{p} = 0.04$; (B) $\bar{p} = 0.02$; and (C) $\bar{p} = 0.01$. The parameter value affects the receptive field properties of the second-layer neurons.

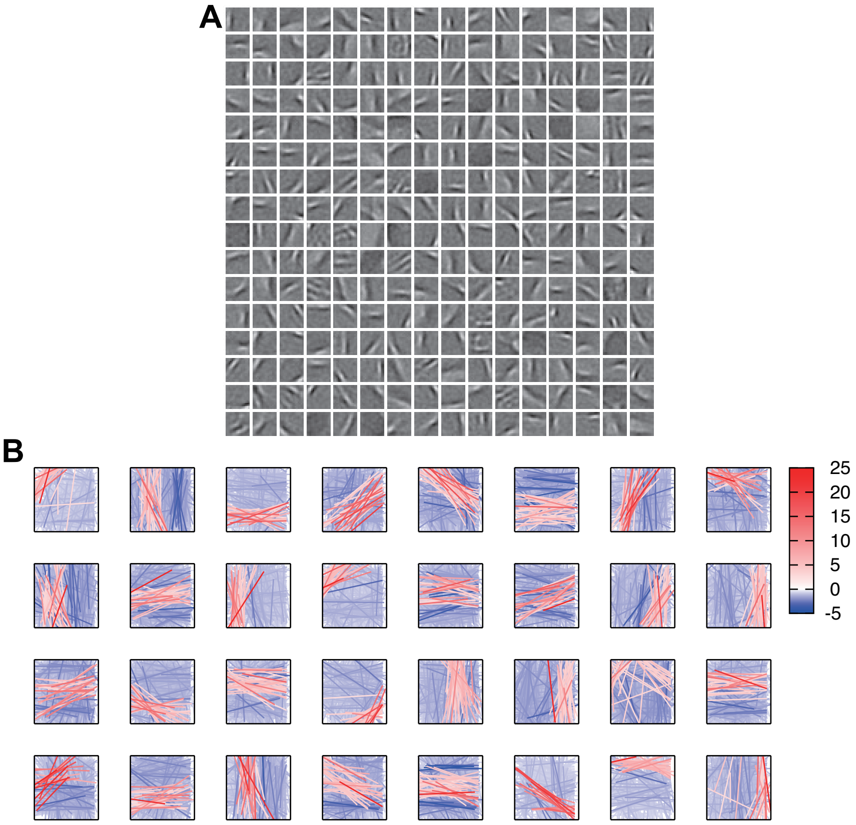

Figure 8: Online and almost local learning rule. (A) Simple-cell-like connection weights of output neurons in the first layer after learning. (B) Complex-cell-like connection weights of second-layer neurons after learning.