Rao, R. P. N. and Sejnowski, T. J., Predictive coding, cortical feedback, and spike-timing dependent plasticity, In: R. P. N. Rao, B. A. Olshausen and M. S. Lewicki (Eds.), Probabilistic models of the brain: Perception and neural function, MIT Press, Cambridge, MA 297-315 (2002).

16 Predictive Coding, Cortical Feedback, and Spike-Timing Dependent Plasticity

Rajesh P. N. Rao and Terrence J. Sejnowski

Introduction

One of the most prominent but least understood neuroanatomical features of the cerebral cortex is feedback. Neurons within a cortical area generally receive massive excitatory feedback from other neurons in the same cortical area. Some of these neurons, especially those in the superficial layers, send feedforward axons to higher cortical areas while others neurons, particularly those in the deeper layers, send feedback axons to lower cortical areas. What is the functional significance of these local and long-range feedback connections?

In this chapter, we explore the following two hypotheses: (a) feedback connections from a higher to a lower cortical area carry predictions of expected neural activity in the lower area, while the feedforward connections carry the differences between the predictions and the actual neural activity; and (b) recurrent feedback connections between neurons within a cortical area are used to learn, store, and predict temporal sequences of input neural activity. Together, these two types of feedback connections help instantiate a hierarchical spatiotemporal generative model of cortical inputs.

The idea that feedback connections may instantiate a hierarchical generative model of sensory inputs has been proposed previously in the context of the Helmholtz machine [14, 15]. However, feedback connections in the Helmholtz machine were used only during training and played no role in perception, which involved a single feedforward pass through the hierarchical network. On the other hand, the possibility of feedback connections carrying expectations of lower level activity and feedforward connections carrying error signals was first studied by MacKay in the context of his epistemological automata [24]. More recently, similar ideas have been suggested by

Pece [34] and Mumford [31] as a model for corticothalamic and cortical networks. The idea of using lateral or recurrent feedback connections for storing temporal dynamics has received much attention in the neural networks community [21, 19, 36] and in models of the hippocampus [28, 1]. However, in the case of cortical models, recurrent connections have been used mainly to amplify weak thalamic inputs in models of orientation [7, 42] and direction selectivity [17, 44, 29]. Recent results on synaptic plasticity of recurrent cortical connections indicate a dependence on the temporal order of pre- and postsynaptic spikes: synapses that are activated slightly before the postsynaptic cell fires are strengthened whereas those that are activated slightly after are weakened [26]. In this chapter, we explore the hypothesis that such a synaptic learning rule allows local recurrent feedback connections to be used for encoding and predicting temporal sequences. Together with corticocortical feedback, these local feedback connections could allow the implementation of spatiotemporal generative models in recurrent cortical circuits.

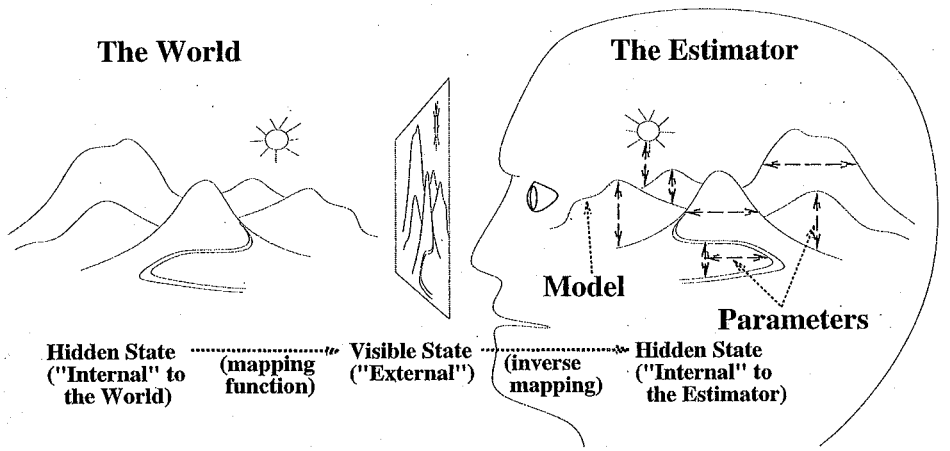
Spatiotemporal Generative Models

Figure 16.1A depicts the problem faced by an organism perceiving the external world. The organism does not have access to the hidden states of the world that are causing its sensory experiences. Instead, it must solve the "inverse" problem of *estimating* these hidden state parameters using only the sensory measurements obtained from its various sensing devices in order to correctly interpret and understand the external world [35]. Note that with respect to the cortex, the definition of an "external world" need not be restricted to sensory modalities such as vision or audition. The cortex may learn and use internal models of "extra-cortical" systems such as the various musculo-skeletal systems responsible for executing body movements [47].

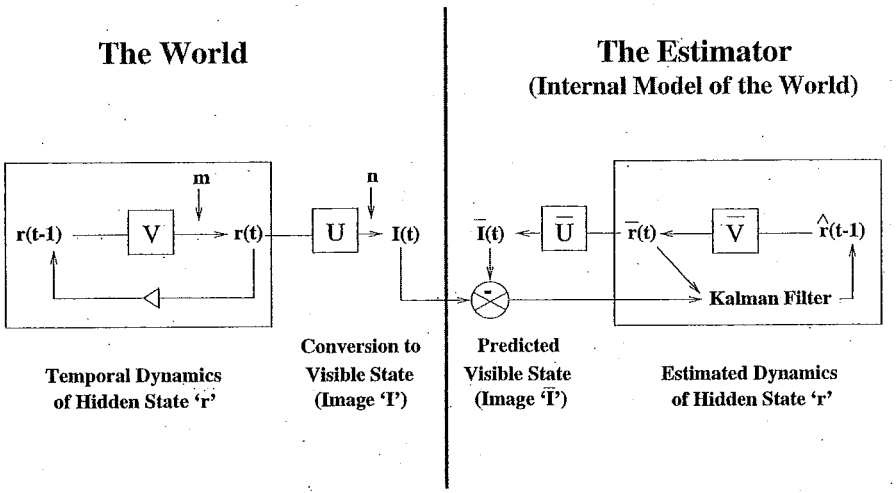
Perhaps the simplest mathematical form one can ascribe to an internal model is to assume a *linear generative model* for the process underlying the generation of sensory inputs. In particular, at any time instant t, the state of the given input generating process is assumed to be characterized by a k-element *hidden state vector* $\mathbf{r}(t)$. Although not directly accessible, this state vector is assumed to generate a measurable and observable output $\mathbf{I}(t)$ (for example, an image of n pixels) according to:

$$\mathbf{I}(t) = U\mathbf{r}(t) + \mathbf{n}(t) \tag{16.1}$$

where U is a (usually unknown) generative (or measurement) matrix that relates the $k \times 1$ state vector $\mathbf{r}(t)$ to the $n \times 1$ observable output vector $\mathbf{I}(t)$, and $\mathbf{n}(t)$ is a Gaussian stochastic noise process with mean zero and a covariance matrix given by



A



B

Figure 16.1. Internal Models and the Problem of Optimal Estimation of Hidden State (A) The problem faced by an organism relying on an internal model of its environment (from [33]). The underlying goal is to optimally estimate, at each time instant, the hidden state of the environment given only the sensory measurements **I**. (B) depicts a single-level Kalman filter solution to the estimation problem. The internal model is encoded jointly by the state transition matrix $\overline{V}$ and the generative matrix $\overline{U}$, and the filter uses this internal model to compute optimal estimates $\hat{\mathbf{r}}$ of the current state $\mathbf{r}$ of the environment.

$\Sigma = E[\mathbf{n}\mathbf{n}^T]$ (E denotes the expectation operator and T denotes transpose). Note that this is a sufficient description of $\mathbf{n}$ since a Gaussian distribution is completely specified by its mean and covariance.

In addition to specifying how the hidden state of the observed process generates a spatial image, we also need to specify how the state itself changes with time t. We assume that the transition from the state $\mathbf{r}(t-1)$ at time instant $t-1$ to the state $\mathbf{r}(t)$ at the next time instant can be modeled as:

$$\mathbf{r}(t) = V\mathbf{r}(t-1) + \mathbf{m}(t-1) \tag{16.2}$$

where V is a (usually unknown) *state transition (or prediction) matrix* and $\mathbf{m}$ is a Gaussian noise process with mean $\overline{\mathbf{m}}(t)$ and covariance $\Pi = E[(\mathbf{m} - \overline{\mathbf{m}})(\mathbf{m} - \overline{\mathbf{m}})^T]$. In other words, the matrix V is used to characterize the dynamic behavior of the observed system over the course of time. Any difference between the actual state $\mathbf{r}(t)$ and the prediction from the previous time step $V\mathbf{r}(t-1)$ is modeled as the stochastic noise vector $\mathbf{m}(t-1)$.

Optimization Functions

The parameters $\mathbf{r}$, U, and V in the spatiotemporal generative model above can be estimated and learned directly from input data if we can define an appropriate optimization function with respect to $\mathbf{r}$, U, and V. For the present purposes, assume that we know the true values of U and V, and we therefore wish to find, at each time instant, an optimal estimate $\widehat{\mathbf{r}}(t)$ of the current state $\mathbf{r}(t)$ of the observed process using only the measurable inputs $\mathbf{I}(t)$.

Suppose that we have already computed a prediction $\overline{\mathbf{r}}$ of the current state $\mathbf{r}$ based on prior data. In particular, let $\overline{\mathbf{r}}(t)$ be the mean of the current state vector *before* measurement of the input data $\mathbf{I}$ at the current time instant t. The corresponding covariance matrix is given by $E[(\mathbf{r} - \overline{\mathbf{r}})(\mathbf{r} - \overline{\mathbf{r}})^T] = M$. A common optimization function whose minimization yields an estimate for $\mathbf{r}$ is the *least-squares criterion*:

$$J_1 = \sum_{i=1}^{n} \left(\mathbf{I}^i - U^i\mathbf{r}\right)^2 + \sum_{i=1}^{k} (\mathbf{r}^i - \overline{\mathbf{r}}^i)^2 = (\mathbf{I} - U\mathbf{r})^T(\mathbf{I} - U\mathbf{r}) + (\mathbf{r} - \overline{\mathbf{r}})^T(\mathbf{r} - \overline{\mathbf{r}}) \tag{16.3}$$

where the superscript i denotes the ith element or row of the superscripted vector or matrix. For example, in the case where $\mathbf{I}$ represents an image, the value for $\mathbf{r}$ that minimizes this quadratic function is the value that (1) yields the smallest sum of pixel-wise differences (squared residual errors) between the image $\mathbf{I}$ and its reconstruction $U\mathbf{r}$ obtained using the matrix U, and (2) is also as close as possible to the prediction $\overline{\mathbf{r}}$ computed from prior data.

The quadratic optimization function above is a special case of the more general *weighted least-squares criterion* [10, 35]:

$$J = (\mathbf{I} - U\mathbf{r})^T \Sigma^{-1} (\mathbf{I} - U\mathbf{r}) + (\mathbf{r} - \bar{\mathbf{r}})^T M^{-1} (\mathbf{r} - \bar{\mathbf{r}}) \tag{16.4}$$

The weighted least-squares criterion becomes meaningful when interpreted in terms of the stochastic model described in the previous section. Recall that the measurement equation 16.1 was characterized in terms of a Gaussian with mean zero and covariance Σ. Also, as given in the previous paragraph, $\mathbf{r}$ follows a Gaussian distribution with mean $\bar{\mathbf{r}}$ and covariance M. Thus, it can be shown that J is simply the sum of the negative log of the (Gaussian) probability of generating the data $\mathbf{I}$ given the state $\mathbf{r}$, and the negative log of the (Gaussian) prior probability of the state $\mathbf{r}$:

$$J = (-\log P(\mathbf{I}|\mathbf{r})) + (-\log P(\mathbf{r})) \tag{16.5}$$

The first term in the above equation follows from the fact that $P(\mathbf{I}|\mathbf{r}) = P(\mathbf{I}, \mathbf{r})/P(\mathbf{r}) = P(\mathbf{n}, \mathbf{r})/P(\mathbf{r}) = P(\mathbf{n})$, assuming $P(\mathbf{n}, \mathbf{r}) = P(\mathbf{n})P(\mathbf{r})$. Now, note that the *posterior* probability of the state given the the input data is given by (using Bayes theorem):

$$P(\mathbf{r}|\mathbf{I}) = P(\mathbf{I}|\mathbf{r})P(\mathbf{r})/P(\mathbf{I}) \tag{16.6}$$

By taking the negative log of both sides (and ignoring the term due to $P(\mathbf{I})$ since it is a fixed quantity), we can conclude that minimizing J is exactly the same as maximizing the posterior probability of the state $\mathbf{r}$ given the input data $\mathbf{I}$.

Predictive Coding

The optimization function J formulated in the previous section can be minimized to find the optimal value $\hat{\mathbf{r}}$ of the state $\mathbf{r}$ by setting $\frac{\partial J}{\partial \hat{\mathbf{r}}} = 0$:

$$-U^T \Sigma^{-1} (\mathbf{I} - U\hat{\mathbf{r}}) + M^{-1} (\hat{\mathbf{r}} - \bar{\mathbf{r}}) = 0 \tag{16.7}$$

which yields:

$$(U^T \Sigma^{-1} U + M^{-1}) \hat{\mathbf{r}} = M^{-1} \bar{\mathbf{r}} + U^T \Sigma^{-1} \mathbf{I} \tag{16.8}$$

Using the substitution $N(t) = (U^T\Sigma^{-1}U + M^{-1})^{-1}$ and rearranging the terms in the above equation, we obtain the following predictive coding equation (also known as the *Kalman filter* in optimal control theory [10]):

$$\hat{\mathbf{r}}(t) = \bar{\mathbf{r}}(t) + N(t)U^T\Sigma(t)^{-1}(\mathbf{I}(t) - U\bar{\mathbf{r}}(t)) \tag{16.9}$$

This equation is of the form:

$$\text{New Estimate} = \text{Old Estimate} + \text{Gain} \times \text{Sensory Residual Error} \tag{16.10}$$

The gain matrix $K(t) = N(t)U^T\Sigma(t)^{-1}$ in Equation 16.9 determines the weight given to the sensory residual in correcting the old estimate $\bar{\mathbf{r}}$. Note that this gain can be interpreted as a form of "signal-to-noise" ratio: it is determined by the covariances Σ and M, and therefore effectively trades off the prior estimate $\bar{\mathbf{r}}$ against the sensory input $\mathbf{I}$ according to the *uncertainties* in these two sources. The Kalman filter estimate $\hat{\mathbf{r}}$ is in fact the *mean* of the Gaussian distribution of the state $\mathbf{r}$ *after* measurement of $\mathbf{I}$ [10]. The matrix N, which performs a form of divisive normalization, can likewise be shown to be the corresponding *covariance* matrix.

Recall that $\bar{\mathbf{r}}$ and M were the mean and covariance *before* measurement of $\mathbf{I}$. We can now specify how these quantities can be updated over time:

$$\bar{\mathbf{r}}(t) = V\hat{\mathbf{r}}(t-1) + \overline{\mathbf{m}}(t-1) \tag{16.11}$$

$$M(t) = VN(t-1)V^T + \Pi(t-1) \tag{16.12}$$

The above equations propagate the estimates of the mean and covariance ($\hat{\mathbf{r}}$ and N respectively) forward in time to generate the predictions $\bar{\mathbf{r}}$ and M for the next time instant. Figure 16.2A summarizes the essential components of the predictive coding model (see also Figure 16.1B).

Predictive Coding and Cortical Feedback

The cerebral cortex is usually characterized as a 6-layered structure, where layer 4 is typically the "input" layer and layer 5 is typically the "output" layer. Neurons in layers 2/3 generally project to layer 4 of "higher" cortical areas while the deeper layers, including layer 6, project back to the "lower" area (see Figure 16.2B). Further details and area-specific variations of these rules can be found in the review article by Van Essen [46].

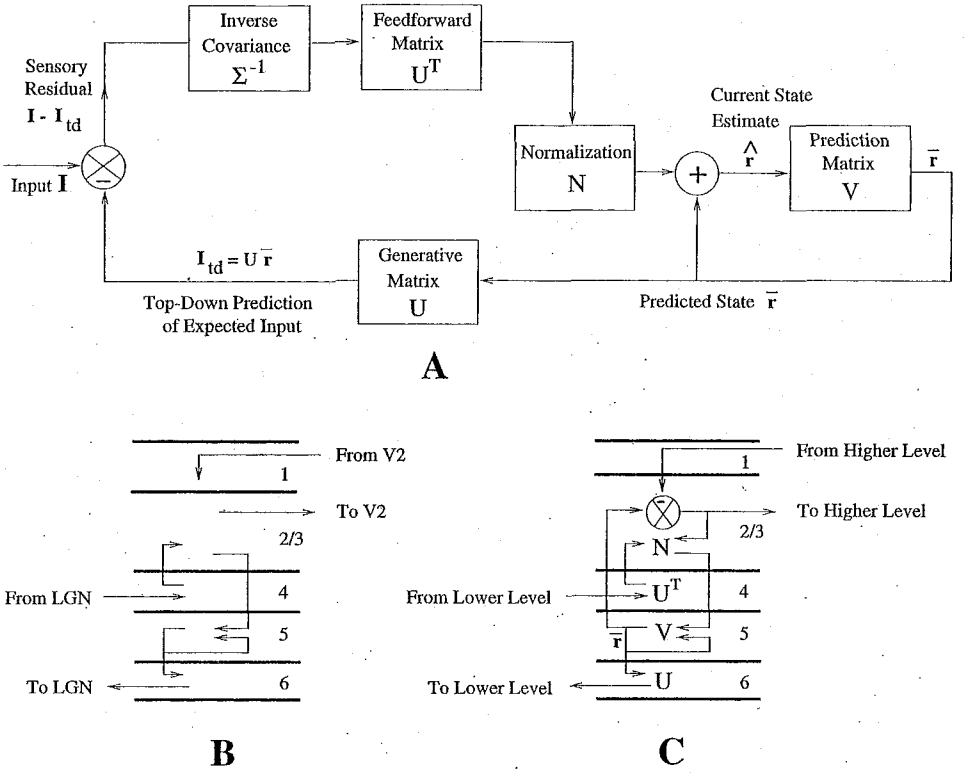


Figure 16.2. The Predictive Coding Model. (A) Schematic diagram of the predictive coding model. (B) The pattern of interlaminar connectivity in primary visual cortex (after [9]). (C) A possible mapping of the components of the predictive coding model onto cortical circuitry.

In the predictive coding model, one needs to predict one step into the future using Equation 16.11, obtain the next sensory input $\mathbf{I}(t)$, and then correct the prediction $\bar{\mathbf{r}}(t)$ using the sensory residual error $(\mathbf{I}(t) - U\bar{\mathbf{r}}(t))$ and the gain $K(t) = N(t)U^T\Sigma^{-1}$. This yields the corrected estimate $\hat{\mathbf{r}}(t)$, which is then used to make the next prediction $\bar{\mathbf{r}}(t+1)$.

This suggests the following mapping between the predictive coding model and cortical anatomy. Feedback connections from a higher cortical area to a lower area may carry the prediction $U\bar{\mathbf{r}}(t)$ to the lower area, while the feedforward connections may carry the prediction error $(\mathbf{I}(t) - U\bar{\mathbf{r}}(t))$. Here, $\mathbf{I}(t)$ is the input signal at the lowest level (for example, the lateral geniculate nucleus (LGN)). The deeper layer neurons, for example those in layer 6, are assumed to implement the feedback weights U while the connections to input layer 4 implement the synaptic weights U^T. Neurons in the "output" layer 5 maintain the current estimate $\hat{\mathbf{r}}(t)$ and the recurrent intracortical connections between neurons in layer 5 are assumed to implement the synaptic weights V. This suggested mapping is depicted in Figure 16.2C. Note that this

mapping is at best extremely coarse and neglects several important issues, such as how the covariance and gain matrices are implemented.

The model described above can be extended to the hierarchical case, where each level in the hierarchy receives not only bottom-up error signals (as described above) but also top-down errors from higher levels (see [36, 37] for more details). Such a model provides a statistical explanation for nonclassical receptive field effects involving orientation contrast exhibited by neurons in layer 2/3 [37]. In these experiments, an oriented stimulus, such as a grating, evokes a strong response from a cortical cell but this response is suppressed when the surrounding region is filled with a stimulus of identical orientation. The neural response is strongest when the orientation of the central stimulus is orthogonal to the stimulus orientation of the surrounding region. In the predictive coding model, neurons in layers 2/3 carry error signals. Thus, assuming that the synaptic weights of the network have been developed based on natural image statistics, the error (and hence the neural response in layers 2/3) is smallest when the surrounding context can predict the central stimulus; the response is largest when the central stimulus cannot be predicted from the surrounding context, resulting in a large error signal. Such an explanation is consistent with Barlow's redundancy reduction hypothesis [3, 4] and Mumford's Pattern Theoretic approach [32]. It differs from the explanation suggested by Wainwright, Schwartz, and Simoncelli based on divisive normalization (see their chapter in this book), although the goal in both approaches is redundancy reduction.

Spike-Timing Dependent Plasticity and Predictive Sequence Learning

The preceding section sketched a possible mapping between cortical anatomy and an algorithm for predictive coding. An important question then is whether there exists neurophysiological evidence supporting such a mapping. In this section, we focus specifically on the hypothesis, put forth in the previous section, that recurrent intracortical connections between neurons in layer 5 implement the synaptic weights V that are used in the predictive coding model to encode temporal sequences of the state vector $\mathbf{r}(t)$.

Recent experimental results suggest that recurrent excitatory connections between cortical neurons are modified according to a spike-timing dependent Hebbian learning rule: synapses that are activated slightly before the cell fires are strengthened whereas those that are activated slightly after are weakened [26] (see also [22, 48, 8, 1, 20, 41, 43]). Such a time-sensitive learning rule is especially well-suited for learning temporal sequences [1, 28, 38].

To investigate how such a timing-dependent learning rule could allow predictive

learning of sequences, we used a two-compartment model of a cortical neuron consisting of a dendrite and a soma-axon compartment. The compartmental model was based on a previous study that demonstrated the ability of such a model to reproduce a range of cortical response properties [25]. To study synaptic plasticity in this model, excitatory postsynaptic potentials (EPSPs) were elicited at different time delays with respect to postsynaptic spiking by presynaptic activation of a single excitatory synapse located on the dendrite. Synaptic currents were calculated using a kinetic model of synaptic transmission [18] with model parameters fitted to whole-cell recorded AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole proprionic acid) currents (see Methods for more details). Other inputs representing background activity were modeled as sub-threshold excitatory and inhibitory Poisson processes with a mean firing rate of 3 Hz. Synaptic plasticity was simulated by incrementing or decrementing the value for maximal synaptic conductance by an amount proportional to the temporal-difference in the postsynaptic membrane potential at time instants $t + \Delta t$ and t for presynaptic activation at time t [38]. The delay parameter Δt was set to 5 ms for these simulations; similar results were obtained for other values in the 5–15 ms range.

Figure 16.3A shows the results of pairings in which the postsynaptic spike was triggered 5 ms after and 5 ms before the onset of the EPSP respectively. While the peak EPSP amplitude was increased 58.5% in the former case, it was decreased 49.4% in the latter case, qualitatively similar to experimental observations [26]. The critical window for synaptic modifications in the model was examined by varying the time interval between presynaptic stimulation and postsynaptic spiking (with $\Delta t = 5$ ms). As shown in Figure 16.3B, changes in synaptic efficacy exhibited a highly asymmetric dependence on spike timing similar to physiological data [8]. Potentiation was observed for EPSPs that occurred between 1 and 12 ms before the postsynaptic spike, with maximal potentiation at 6 ms. Maximal depression was observed for EPSPs occurring 6 ms after the peak of the postsynaptic spike and this depression gradually decreased, approaching zero for delays greater than 10 ms. As in rat neocortical neurons [26], *Xenopus* tectal neurons [48], and cultured hippocampal neurons [8], a narrow transition zone (roughly 3 ms in the model) separated the potentiation and depression windows.

To see how a network of model neurons can learn to predict sequences using the learning mechanism described above, consider the simplest case of two excitatory neurons N1 and N2 connected to each other, receiving inputs from two separate input neurons I1 and I2 (Figure 16.4A). Suppose input neuron I1 fires before input neuron I2, causing neuron N1 to fire (Figure 16.4B). The spike from N1 results in a sub-threshold EPSP in N2 due to the synapse S2. If input arrives from I2 any time between 1 and 12 ms after this EPSP and the temporal summation of these two EPSPs causes N2 to fire, the synapse S2 will be strengthened. The synapse S1, on the other hand, will be weakened because the EPSP due to N2 arrives a few milliseconds after N1 has fired. Thus, on a subsequent trial, when input I1 causes neuron N1 to fire,

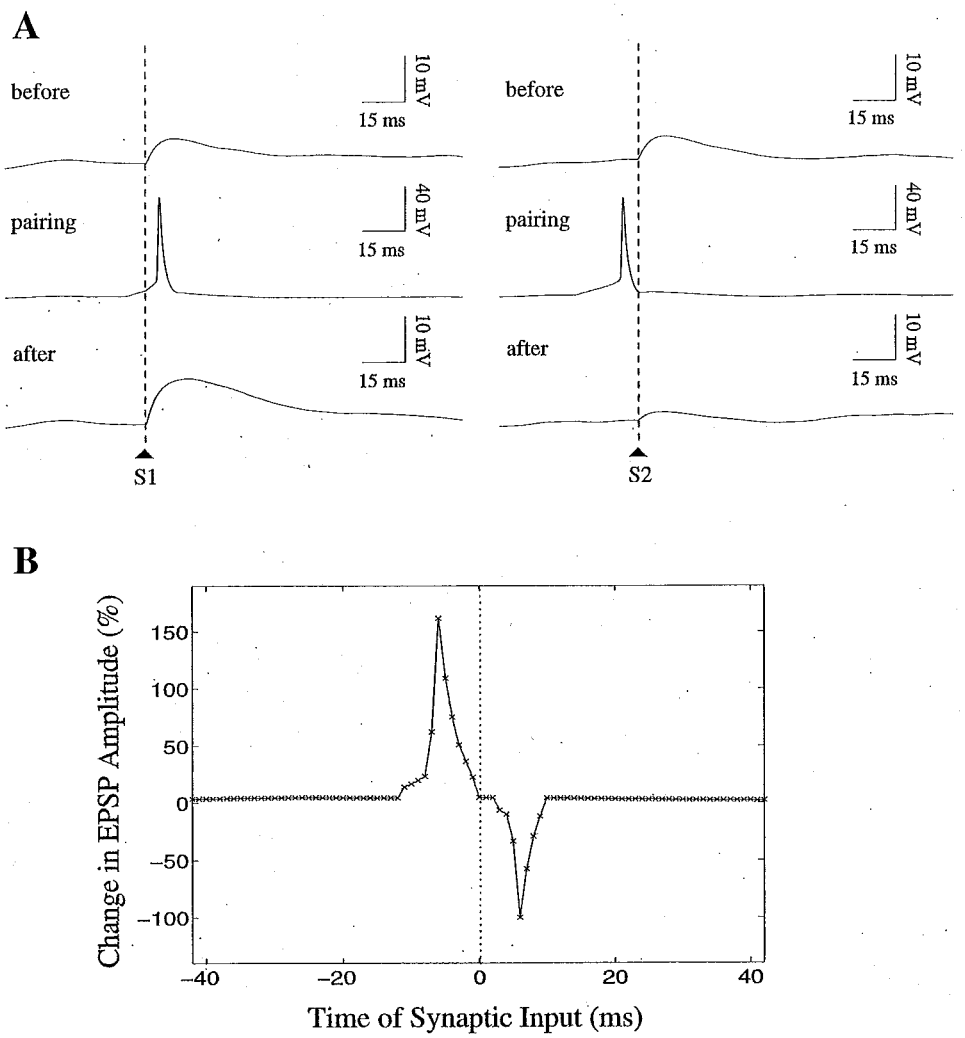


Figure 16.3. Synaptic Plasticity in a Model Neocortical Neuron. (from [39]) **(A)** (Left Panel) The response at the top (labeled "before") is the EPSP invoked in the model neuron due to a presynaptic spike (S1) at an excitatory synapse. Pairing this presynaptic spike with postsynaptic spiking after a 5 ms delay ("pairing") induces long-term potentiation as revealed by an enhancement in the peak of the EPSP evoked by presynaptic simulation alone ("after"). (Right Panel) If presynaptic stimulation (S2) occurs 5 ms after postsynaptic firing, the synapse is weakened resulting in a corresponding decrease in peak EPSP amplitude. **(B)** Critical window for synaptic plasticity obtained by varying the delay between presynaptic stimulation and postsynaptic spiking (negative delays refer to cases where presynaptic stimulation occurred before the postsynaptic spike).

it in turn causes N2 to fire several milliseconds *before* input I2 occurs due to the potentiation of the recurrent synapse S2 in previous trial(s) (Figure 16.4C). Input neuron I2 can thus be inhibited by the predictive feedback from N2 just before the occurrence of imminent input activity (marked by an asterisk in Figure 16.4C). This

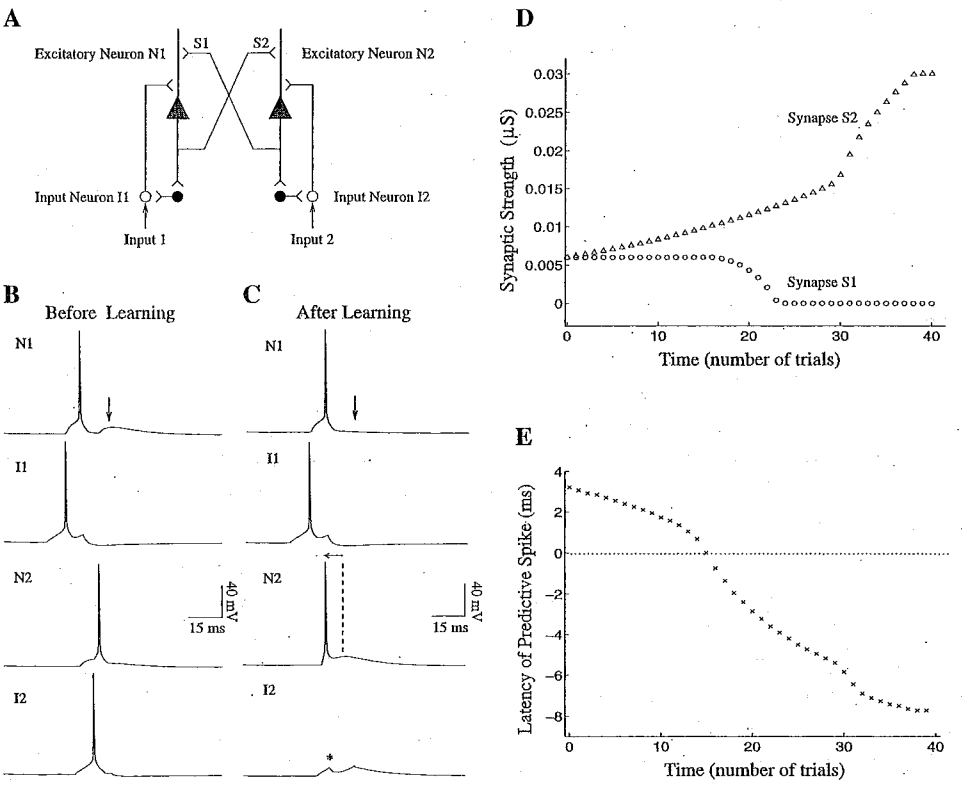


Figure 16.4. Learning to Predict using Spike-Timing Dependent Hebbian Plasticity. (from [39]) (**A**) A simple network of two model neurons N1 and N2 recurrently connected via excitatory synapses S1 and S2. Sensory inputs are relayed to the two model neurons by input neurons I1 and I2. Feedback from N1 and N2 inhibit the input neurons via inhibitory interneurons (darkened circles). (**B**) Activity in the network elicited by the input sequence I1 followed by I2. Notice that N2 fires after its input neuron I2 has fired. (**C**) Activity in the network elicited by the same input sequence after 40 trials of learning. Notice that due to the strengthening of synapse S2, neuron N2 now fires several milliseconds before the time of expected input from I2 (dashed line), allowing it to inhibit I2 (asterisk). On the other hand, synapse S1 has been weakened, thereby preventing re-excitation of N1 (downward arrows show the corresponding decrease in EPSP). (**D**) Potentiation and depression of synapses S1 and S2 respectively during the course of learning. Synaptic strength was defined as maximal synaptic conductance in the kinetic model of synaptic transmission (see Methods). (**E**) Latency of the predictive spike in neuron N2 during the course of learning measured with respect to the time of input spike in I2 (dotted line). Note that the latency is initially positive (N2 fires after I2) but later becomes negative, reaching a value of up to 7.7 ms before input I2 as a consequence of learning.

inhibition prevents input I2 from further exciting N2. Similarly, a positive feedback loop between neurons N1 and N2 is avoided because the synapse S1 was weakened in previous trial(s) (see arrows in Figures 16.4B and 16.4C). Figure 16.4D depicts the process of potentiation and depression of the two synapses as a function of the number of exposures to the I1-I2 input sequence. The decrease in latency of the

predictive spike elicited in N2 with respect to the timing of input I2 is shown in Figure 16.4E. Notice that before learning, the spike occurs 3.2 ms after the occurrence of the input whereas after learning, it occurs 7.7 ms before the input. This simple example helps to illustrate how subsets of neurons may learn to selectively trigger other subsets of neurons in anticipation of future inputs while maintaining stability in the recurrent network.

Comparisons to Awake Monkey Visual Cortex Data

To facilitate comparison with published neurophysiological data, we have focused specifically on the problem of predicting moving visual stimuli. We used a network of recurrently connected excitatory neurons (as shown in Figure 16.5A) receiving retinotopic sensory input consisting of moving pulses of excitation (8 ms pulse of excitation at each neuron) in the rightward and leftward directions. The task of the network was to predict the sensory input by learning appropriate recurrent connections such that a given neuron in the network can fire a few milliseconds before the arrival of its input pulse of excitation. The network was comprised of two parallel chains of neurons with mutual inhibition (dark arrows) between corresponding pairs of neurons along the two chains. The network was initialized such that within a chain, a given excitatory neuron received both excitation and inhibition from its predecessors and successors. This is shown in Figure 16.5B for a neuron labeled '0'. Inhibition at a given neuron was mediated by an inhibitory interneuron (dark circle) which received excitatory connections from neighboring excitatory neurons (Figure 16.5B, lower panel). The interneuron received the same input pulse of excitation as the nearest excitatory neuron. Maximum conductances for all synapses were initialized to small positive values (dotted lines in Figure 16.5C) with a slight asymmetry in the recurrent excitatory connections for breaking symmetry between the two chains. The initial asymmetry elicited a single spike slightly earlier for neurons in one chain than neurons in the other chain for a given motion direction, allowing activity in the other chain to be inhibited.

To evaluate the consequences of synaptic plasticity, the network of neurons was exposed alternately to leftward and rightward moving stimuli for a total of 100 trials. The excitatory connections (labeled 'EXC' in Figure 16.5B) were modified according to the spike-timing dependent Hebbian learning rule in Figure 16.3B while the excitatory connections onto the inhibitory interneuron (labeled 'INH') were modified according to an asymmetric anti-Hebbian learning rule that reversed the polarity of the rule in Figure 16.3B [6].

The synaptic conductances learned by two neurons (marked N1 and N2 in Figure 16.5A) located at corresponding positions in the two chains after 100 trails of

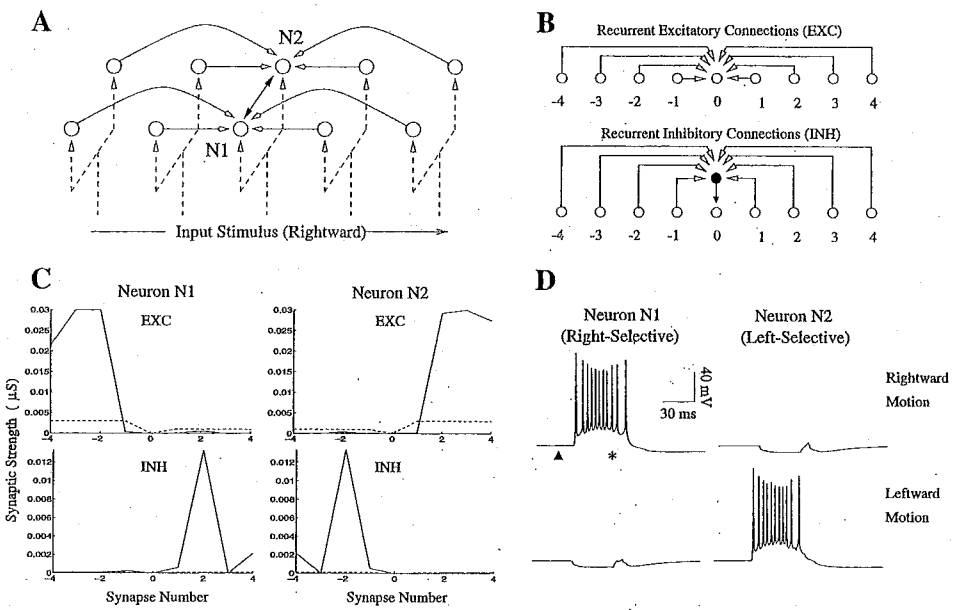


Figure 16.5. Emergence of Direction Selectivity in the Model. **(A)** A model network consisting of two chains of recurrently connected neurons receiving retinotopic inputs. A given neuron receives recurrent excitation and recurrent inhibition (white-headed arrows) as well as inhibition (dark-headed arrows) from its counterpart in the other chain. **(B)** Recurrent connections to a given neuron (labeled '0') arise from 4 preceding and 4 succeeding neurons in its chain. Inhibition at a given neuron is mediated via a GABAergic interneuron (darkened circle). **(C)** Synaptic strength of recurrent excitatory (EXC) and inhibitory (INH) connections to neurons N1 and N2 before (dotted lines) and after learning (solid lines). Synapses were adapted during 100 trials of exposure to alternating leftward and rightward moving stimuli. **(D)** Responses of neurons N1 and N2 to rightward and leftward moving stimuli. As a result of learning, neuron N1 has become selective for rightward motion (as have other neurons in the same chain) while neuron N2 has become selective for leftward motion. In the preferred direction, each neuron starts firing several milliseconds before the actual input arrives at its soma (marked by an asterisk) due to recurrent excitation from preceding neurons. The dark triangle represents the start of input stimulation in the network.

exposure to the moving stimuli are shown in Figure 16.5C (solid line). As expected from the learned asymmetric pattern of connectivity, neuron N1 was found to be selective for rightward motion while neuron N2 was selective for leftward motion (Figure 16.5D). Moreover, when stimulus motion is in the preferred direction, each neuron starts firing a few milliseconds before the time of arrival of the input stimulus at its soma (marked by an asterisk) due to recurrent excitation from preceding neurons. Conversely, motion in the non-preferred direction triggers recurrent inhibition from preceding neurons as well as inhibition from the active neuron in the corresponding position in the other chain.

Similar to complex cells in primary visual cortex, model neurons are direction selective throughout their receptive field because at each retinotopic location, the

corresponding neuron in the chain receives the same pattern of asymmetric excitation and inhibition from its neighbors as any other neuron in the chain. Thus, for a given neuron, motion in any local region of the chain will elicit direction selective responses due to recurrent connections from that part of the chain. This is consistent with previous modeling studies [11] suggesting that recurrent connections may be responsible for the spatial-phase invariance of complex cell responses. Assuming a 200 μm separation between excitatory model neurons in each chain and utilizing known values for the cortical magnification factor in monkey striate cortex [45], one can estimate the preferred stimulus velocity of model neurons to be 3.1°/s in the fovea and 27.9°/s in the periphery (at an eccentricity of 8°). Both of these values fall within the range of monkey striate cortical velocity preferences (1°/s to 32 °/s) [46, 23].

The model predicts that the neuroanatomical connections for a direction selective neuron should exhibit a pattern of asymmetrical excitation and inhibition similar to Figure 16.5C. A recent study of direction selective cells in awake monkey V1 found excitation on the preferred side of the receptive field and inhibition on the null side consistent with the pattern of connections learned by the model [23]. For comparison with this experimental data, spontaneous background activity in the model was generated by incorporating Poisson-distributed random excitatory and inhibitory alpha synapses on the dendrite of each model neuron. Post stimulus time histograms (PSTHs) and space-time response plots were obtained by flashing optimally oriented bar stimuli at random positions in the cell's activating region. As shown in Figure 16.6, there is good qualitative agreement between the response plot for a direction-selective complex cell and that for the model. Both space-time plots show a progressive shortening of response onset time and an increase in response transiency going in the preferred direction: in the model, this is due to recurrent excitation from progressively closer cells on the preferred side. Firing is reduced to below background rates 40-60 ms after stimulus onset in the upper part of the plots: in the model, this is due to recurrent inhibition from cells on the null side. The response transiency and shortening of response time course appears as a slant in the space-time maps, which can be related to the neuron's velocity sensitivity (see [23] for more details).

Conclusions

This chapter reviewed the hypothesis that (i) feedback connections between cortical areas instantiate probabilistic generative models of cortical inputs, and (ii) recurrent feedback connections within a cortical area encode the temporal dynamics associated with these generative models. We formalized this hypothesis in terms of a predictive coding framework and suggested a possible implementation of the predictive coding

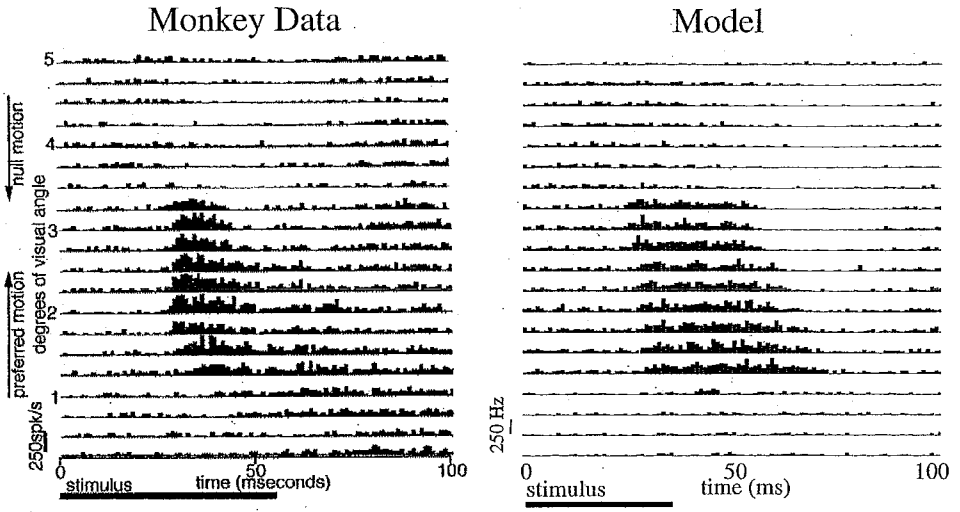


Figure 16.6. Comparison of Monkey and Model Space-Time Response Plots. (Left) Sequence of PSTHs obtained by flashing optimally oriented bars at 20 positions across the 5°-wide receptive field (RF) of a complex cell in alert monkey V1 (from [23]). The cell's preferred direction is from the part of the RF represented at the bottom towards the top. Flash duration = 56 ms; inter-stimulus delay = 100 ms; 75 stimulus presentations. (Right) PSTHs obtained from a model neuron after stimulating the chain of neurons at 20 positions to the left and right side of the given neuron. Lower PSTHs represent stimulations on the preferred side while upper PSTHs represent stimulations on the null side.

model within the laminar structure of the cortex. At the biophysical level, we showed that recent results on spike-timing dependent plasticity in recurrent cortical synapses are consistent with our suggested roles for cortical feedback. Data from model simulations were shown to be similar to electrophysiological data from awake monkey visual cortex.

An important direction for future research is exploring hierarchical models of spatiotemporal predictive coding based on spike-timing dependent sequence learning at multiple levels. A related direction of research is elucidating the role of spike timing in predictive coding. The chapter by Ballard, Zhang, and Rao in this book investigates the hypothesis that cortical communication may occur via synchronous volleys of spikes. The spike-timing dependent learning rule appears to be especially well-suited for learning synchrony [20, 1], but the question of whether the same learning rule allows the formation of multi-synaptic chains of synchronously firing neurons remains to be ascertained.

The predictive coding model is closely related to models based on sparse coding (see the chapters by Olshausen and Lewicki) and to competitive/divisive normalization models (see the chapters by Piepenbrock and Wainwright, Schwartz, and Simoncelli). These models share the goal of redundancy reduction but attempt to achieve

this goal via different means (for example, by using sparse prior distributions on the state vector $\mathbf{r}$ or by dividing it with a normalization term). The model described in this chapter additionally includes a separate component in its generative model for temporal dynamics, which allows prediction in time as well as space. The idea of sequence learning and prediction in the cortex and the hippocampus has been explored in several previous studies [1, 28, 30, 36, 5, 13, 27]. Our biophysical simulations suggest a possible implementation of such predictive sequence learning models in cortical circuitry. Given the general uniformity in the structure of the neocortex across different areas [12, 40, 16] as well as the universality of the problem of learning temporal sequences in both sensory and motor domains, the hypothesis of predictive coding and sequence learning may help provide a unified probabilistic framework for investigating cortical information processing.

Acknowledgments

This work was supported by the Alfred P. Sloan Foundation and Howard Hughes Medical Institute. We thank Margaret Livingstone, Dmitri Chklovskii, David Eagleman, and Christian Wehrhahn for discussions and comments.

References

[1] L. F. Abbott and K. I. Blum, "Functional significance of long-term potentiation for sequence learning and prediction," *Cereb. Cortex* **6**, 406-416 (1996).

[2] L. F. Abbott and S. Song, "Temporally asymmetric Hebbian learning, spike timing and neural response variability," in *Advances in Neural Info. Proc. Systems 11*, M. S. Kearns, S. A. Solla and D. A. Cohn, Eds. (MIT Press, Cambridge, MA, 1999), pp. 69–75.

[3] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. Cambridge, MA: MIT Press, 1961.

[4] H. B. Barlow, "What is the computational goal of the neocortex?" in C. Koch and J. L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 1–22. Cambridge, MA: MIT Press, 1994.

[5] H. Barlow, "Cerebral predictions," *Perception* **27**, 885-888 (1998).

[6] C. C. Bell, V. Z. Han, Y. Sugawara, and K. Grant, "Synaptic plasticity in a cerebellum-like structure depends on temporal order," *Nature* **387**, 278-281 (1997).

[7] R. Ben-Yishai, R. L. Bar-Or, and H. Sompolinsky, "Theory of orientation tuning in visual cortex," *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3844-3848 (1995).

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

[9] J. Bolz, C. D. Gilbert, and T. N. Wiesel, "Pharmacological analysis of cortical circuitry," *Trends in Neurosciences*, 12(8):292–296, 1989.

[10] A. E. Bryson and Y.-C. Ho. *Applied Optimal Control.* New York: John Wiley and Sons, 1975.

[11] F. S. Chance, S. B. Nelson, and L. F. Abbott, "Complex cells as cortically amplified simple cells," *Nature Neuroscience* **2**, 277-282 (1999).

[12] O. D. Creutzfeldt, "Generality of the functional structure of the neocortex," *Naturwissenschaften* **64**, 507-517 (1977).

[13] J. G. Daugman and C. J. Downing, "Demodulation, predictive coding, and spatial vision," *J. Opt. Soc. Am. A* **12**, 641-660 (1995).

[14] P. Dayan, G.E. Hinton, R.M. Neal, and R.S. Zemel, "The Helmholtz machine," *Neural Computation*, 7:889–904 (1995).

[15] P. Dayan and G. E. Hinton, "Varieties of Helmholtz machine," *Neural Networks* **9**(8), 1385-1403 (1996).

[16] R. J. Douglas, K. A. C. Martin, and D. Whitteridge, "A canonical microcircuit for neocortex," *Neural Computation* **1**, 480-488 (1989).

[17] R. J. Douglas, C. Koch, M. Mahowald, K. A. Martin, and H. H. Suarez, "Recurrent excitation in neocortical circuits," *Science* **269**, 981-985 (1995).

[18] A. Destexhe, Z. F. Mainen, and T. J. Sejnowski, "Kinetic models of synaptic transmission," in *Methods in Neuronal Modeling*, C. Koch and I. Segev, Eds. (MIT Press, Cambridge, MA, 1998).

[19] J. L. Elman, "Finding structure in time," *Cognitive Science* **14**, 179-211 (1990).

[20] W. Gerstner, R. Kempter, J. L. van Hemmen, and H. Wagner, "A neuronal learning rule for sub-millisecond temporal coding," *Nature* **383**, 76-81 (1996).

[21] M. I. Jordan, "Attractor dynamics and parallelism in a connectionist sequential machine," in *Proceedings of the Annual Conf. of the Cog. Sci. Soc.*, pp. 531-546 (1986).

[22] W. B. Levy and O. Steward, "Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus," *Neuroscience* **8**, 791-797 (1983).

[23] M.S. Livingstone, "Mechanisms of direction selectivity in macaque V1," *Neuron*, 20:509–526 (1998).

[24] D. M. MacKay, "The epistemological problem for automata," in *Automata Studies*, pages 235–251. Princeton, NJ: Princeton University Press, 1956.

[25] Z. F. Mainen and T. J. Sejnowski, "Influence of dendritic structure on firing pattern in model neocortical neurons," *Nature* **382**, 363-366 (1996).

[26] H. Markram, J. Lubke, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs," *Science* **275**, 213-215 (1997).

[27] M. R. Mehta and M. Wilson, "From hippocampus to V1: Effect of LTP on spatiotemporal dynamics of receptive fields," in *Computational Neuroscience, Trends in Research 1999*, J. Bower, Ed. (Elsevier Press, Amsterdam, 2000).

[28] A. A. Minai and W. B. Levy, "Sequence learning in a single trial," in *Proceedings of the 1993 INNS World Congress on Neural Networks* II, (Erlbaum, NJ, 1993), pp. 505-508.

[29] P. Mineiro and D. Zipser, "Analysis of direction selectivity arising from recurrent cortical interactions," *Neural Comput.* **10**, 353-371 (1998).

[30] P. R. Montague and T. J. Sejnowski, "The predictive brain: Temporal coincidence and temporal order in synaptic learning mechanisms," *Learning and Memory* **1**, 1-33 (1994).

[31] D. Mumford, "On the computational architecture of the neocortex. II. The role of cortico-cortical loops," *Biological Cybernetics*, 66:241–251 (1992).

[32] D. Mumford, "Neuronal architectures for pattern-theoretic problems," in C. Koch and J. L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 125–152. Cambridge, MA: MIT Press, 1994.

[33] R. C. O'Reilly. *The LEABRA model of neural interactions and learning in the neocortex*. PhD thesis, Department of Psychology, Carnegie Mellon University, 1996.

[34] A. E. C. Pece, "Redundancy reduction of a Gabor representation: a possible computational role for feedback from primary visual cortex to lateral geniculate nucleus," in I. Aleksander and J. Taylor, editors, *Artificial Neural Networks 2*, pages 865–868. Amsterdam: Elsevier Science, 1992.

[35] R. P. N. Rao, "An optimal estimation approach to visual perception and learning," *Vision Research* **39**, 1963-1989 (1999).

[36] R. P. N. Rao and D. H. Ballard, "Dynamic model of visual recognition predicts neural response properties in the visual cortex," *Neural Computation* **9**, 721-763 (1997).

[37] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects," *Nature Neuroscience* **2**, 79-87 (1999).

[38] R. P. N. Rao and T. J. Sejnowski, "Predictive sequence learning in recurrent neocortical circuits", in *Advances in Neural Information Processing Systems 12*, S. A. Solla and T. K. Leen and K.-R. Müller, Eds. (MIT Press, Cambridge, MA, 2000), pp. 164-170.

[39] R. P. N. Rao and T. J. Sejnowski, "Spike-timing-dependent Hebbian plasticity

as Temporal Difference learning," *Neural Computation* **13**, 2221-2237, (2001).

[40] T. J. Sejnowski, "Open questions about computation in cerebral cortex," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 2*, J. L. McClelland *et al.*, editors (MIT Press, Cambridge, MA, 1986), pp. 372-389.

[41] T. J. Sejnowski, "The book of Hebb," *Neuron* **24**(4), 773-776 (1999).

[42] D. C. Somers, S. B. Nelson, and M. Sur, "An emergent model of orientation selectivity in cat visual cortical simple cells," *J. Neurosci.* **15**, 5448-5465 (1995).

[43] S. Song, K. D. Miller, and L. F. Abbott, "Competitive Hebbian learning through spike-timing dependent synaptic plasticity," *Nature Neuroscience* **3**, 919–926 (2000).

[44] H. Suarez, C. Koch, and R. Douglas, "Modeling direction selectivity of simple cells in striate visual cortex with the framework of the canonical microcircuit," *J. Neurosci.* **15**, 6700-6719 (1995).

[45] R. B. Tootell, E. Switkes, M. S. Silverman, and S. L. Hamilton, "Functional anatomy of macaque striate cortex. II. Retinotopic organization," *J. Neurosci.* **8**, 1531-1568 (1988).

[46] D.C. Van Essen, "Functional organization of primate visual cortex," in A. Peters and E.G. Jones, editors, *Cerebral Cortex*, volume 3, pages 259–329. New York, NY: Plenum, 1985.

[47] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An internal model for sensorimotor integration," *Science*, 269:1880–1882 (1995).

[48] L. I. Zhang, H. W. Tao, C. E. Holt, W. A. Harris, and M. M. Poo, "A critical window for cooperation and competition among developing retinotectal synapses," *Nature* **395**, 37-44 (1998).