10.1098/rsta.2003.1190

THE ROYAL SOCIETY

Self-organizing neural systems based on predictive learning

By Rajesh P. N. Rao[1] and Terrence J. Sejnowski[2,3]

[1]*Department of Computer Science and Engineering, University of Washington, Box 352350, Seattle, WA 98195-2350, USA* (rao@cs.washington.edu)
[2]*Computational Neurobiology Laboratory, Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA*
[3]*Department of Biology, University of California at San Diego, La Jolla, CA 92037, USA* (terry@salk.edu)

Published online 6 May 2003

The ability to predict future events based on the past is an important attribute of organisms that engage in adaptive behaviour. One prominent computational method for learning to predict is called temporal-difference (TD) learning. It is so named because it uses the difference between successive predictions to learn to predict correctly. TD learning is well suited to modelling the biological phenomenon of conditioning, wherein an organism learns to predict a reward even though the reward may occur later in time. We review a model for conditioning in bees based on TD learning. The model illustrates how the TD-learning algorithm allows an organism to learn an appropriate sequence of actions leading up to a reward, based solely on reinforcement signals. The second part of the paper describes how TD learning can be used at the cellular level to model the recently discovered phenomenon of spike-timing-dependent plasticity. Using a biophysical model of a neocortical neuron, we demonstrate that the shape of the spike-timing-dependent learning windows found in biology can be interpreted as a form of TD learning occurring at the cellular level. We conclude by showing that such spike-based TD-learning mechanisms can produce direction selectivity in visual-motion-sensitive cells and can endow recurrent neocortical circuits with the powerful ability to predict their inputs at the millisecond time-scale.

Keywords: neuroscience; cerebral cortex; conditioning; synaptic plasticity; visual perception; prediction

1. Introduction

Learning and predicting temporal sequences from experience underlies much of adaptive behaviour in both animals and machines. The smell of freshly baked bread may bring to mind the image of a loaf; the unexpected ring of a doorbell may prompt thoughts of a salesperson at the door; the disappearance of a car behind a slow moving bus elicits an expectation of the car's reappearance after an appropriate delay; the initial notes from an oft-repeated Beatles song prompts a recall of the entire song. These examples illustrate the ubiquitous nature of prediction in behaviour. Our ability to predict depends crucially on the statistical regularities that characterize the

One contribution of 18 to a Theme 'Self-organization: the quest for the origin and evolution of structure'.

natural world (Atick & Redlich 1992; Barlow 1961; Bell & Sejnowski 1997; Dong & Atick 1995; Eckert & Buchsbaum 1993; MacKay 1956; Olshausen & Field 1996; Rao 1999; Rao & Ballard 1997, 1999; Schwartz & Simoncelli 2001). Indeed, prediction would be impossible in a world that is statistically random.

The role of prediction in behavioural learning was investigated in early psychological experiments by Pavlov and others (see Rescorla (1988) for a review). In the famous Pavlovian conditioning experiments, a dog learned to salivate when a bell was rung, after a training session in which an appetizing food stimulus was presented right after the bell. The dog thus learned to predict a food reward (the unconditioned stimulus) based on a hitherto unrelated auditory stimulus (the conditioned stimulus). Several major areas in the brain have been implicated in the learning of rewards and punishments, such as the dopaminergic system, the amygdala, and the cerebellum. At a more general level, it has been suggested that one of the dominant functions of the neocortex is prediction and sequence learning (Barlow 1998; MacKay 1956; Rao 1999; Rao & Ballard 1997, 1999).

A major challenge from a computational point of view is to devise algorithms for prediction and sequence learning that rely solely on interactions with the environment. Several approaches have been suggested, especially in control theory and engineering, such as Kalman filtering, hidden Markov models, and dynamic Bayesian networks (see Ghahramani (2001) for a review). A popular algorithm for learning to predict is temporal-difference (TD) learning (Sutton 1988). TD learning was proposed by Sutton as an 'on-line' algorithm for reinforcement-based learning, wherein an agent is given a scalar reward typically after the completion of a sequence of actions that lead to a desired goal state. The TD-learning algorithm has been enormously influential in the machine learning community, with a wide variety of applications, having even produced a world-class backgammon playing program (Tesauro 1989). We review the basic TD-learning model in § 2.

TD learning has been used to model the phenomenon of conditioning wherein an animal learns to predict a reward based on past stimuli. Sutton & Barto (1990) studied a TD-learning model of classical conditioning. Montague *et al*. (1995) have applied TD learning to the problem of reinforcement learning in foraging bees. There is also evidence for physiological signals in the primate brain that resemble the prediction error seen in TD learning (Schultz *et al*. 1997). We review some of these results in §§ 2 and 3.

The idea of learning to predict based on the temporal difference of successive predictions can also be applied to learning at the cellular level (Dayan 2002; Rao & Sejnowski 2000, 2001). In § 4, we link TD learning to spike-timing-dependent synaptic plasticity (Bi & Poo 1998; Gerstner *et al*. 1996; Levy & Steward 1983; Markram *et al*. 1997; Sejnowski 1999; Zhang *et al*. 1998) and review simulation results. We show that spike-based TD learning causes neurons to become direction selective when exposed to moving visual stimuli. Our results suggest that spike-based TD learning is a powerful mechanism for prediction and sequence learning in recurrent neocortical circuits.

2. Temporal-difference learning

TD learning is a popular computational algorithm for learning to predict inputs (Montague & Sejnowski 1994; Sutton 1988). Learning takes place based on whether the difference between two temporally separated predictions is positive or negative.

This minimizes the errors in prediction by ensuring that the prediction generated after adapting the parameters (for example, the synapses of a neuron) is closer to the desired value than before.

The simplest example of a TD-learning rule arises in the problem of predicting a scalar quantity z using a neuron with synaptic weights $w(1), \ldots, w(k)$ (represented as a vector $\boldsymbol{w}$). The neuron receives as presynaptic input the sequence of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$. The output of the neuron at time t is assumed to be given by $P_t = \sum_i w(i)x_t(i)$. The goal is to learn a set of synaptic weights such that the prediction P_t is as close as possible to the target z. According to the temporal-difference (TD(0)) learning rule (Sutton 1988), the weights at time $t+1$ are given by

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \lambda(P_{t+1} - P_t)\boldsymbol{x}_t, \tag{2.1}$$

where λ is a learning rate or gain parameter and the last value of P is set to the target value, i.e. $P_{m+1} = z$. Note that learning is governed by the temporal difference in the outputs at time instants $t+1$ and t in conjunction with the input $\boldsymbol{x}_t$ at time t.

To understand the rationale behind the simple TD-learning rule, consider the case where all the weights are initially zero, which yields a prediction $P_t = 0$ for all t. However, in the last time-step $t = m$, there is a non-zero prediction error $(P_{m+1} - P_m) = (z - 0) = z$. Given that the prediction error is z at the last time-step, the weights are changed by an amount equal to $\lambda z \boldsymbol{x}_t$. Thus, in the next trial, the prediction P_m will be closer to z than before, and after several trials, will tend to converge to z. The striking feature of the TD algorithm is that, because P_m acts as a training signal for P_{m-1}, which in turn acts as a training signal for P_{m-2} and so on, information about the target z is propagated backwards in time such that the predictions P_t at all previous time-steps are corrected over many trials and will eventually converge to the target z, even though the target only occurs at the end of the trial.

One way of to interpret z is to view it as the reward delivered to an animal at the end of a trial. We can generalize this idea by assuming that a reward r_t is delivered at each time-step t, where r_t could potentially be zero. As Sutton & Barto (1990) originally suggested, the phenomenon of conditioning in animals can be modelled as the prediction of the sum of future rewards in a trial, starting from the current time-step t: $\sum_{i>t}^{m} r_i$. In other words, we want $P_t = \sum_i w(i)x_t(i)$ to approximate $\sum_{i>t}^{m} r_i$. Note that, ideally,

$$P_t = \sum_{i>t}^{m} r_i = r_{t+1} + \sum_{i>t+1}^{m} r_i = r_{t+1} + P_{t+1}. \tag{2.2}$$

Therefore, the error in prediction is given by

$$\delta_t = r_{t+1} + P_{t+1} - P_t \tag{2.3}$$

and the weights can be updated as follows to minimize the prediction error:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \lambda(r_{t+1} + P_{t+1} - P_t)\boldsymbol{x}_t. \tag{2.4}$$

This equation implements the standard TD-learning rule (also known as TD(0)) (Sutton 1988; Sutton & Barto 1998). Note that it depends on both the immediate reward r_{t+1} and the temporal difference between the predictions at time $t+1$ and t.

Considerable theory exists to show that the rule and its variants converge to the correct values under appropriate circumstances (see Sutton & Barto 1998).

Beginning with Sutton & Barto's early work on TD learning as a model for classical conditioning, a number of researchers have used TD learning to explain both behavioural and neural data. One important application of TD learning has been in interpreting the transient activity of cells in the dopamine system of primates: the activity of many of these cells (for example, in the ventral tegmental area) is strikingly similar to the temporal-difference error δ_t that would be expected during the course of learning to predict rewards in a particular task (Schultz *et al*. 1995, 1997). Another demonstration of the utility of the TD-learning algorithm has been in modelling foraging behaviour in bees. Results from this study are reviewed in the next section.

3. TD-learning model of conditioning in bees

In addition to the sensory and motor systems that guide the behaviour of vertebrates and invertebrates, all species also have a set of small nuclei that project axons throughout the brain and release neurotransmitters such as dopamine, norepinephrine, and acetylcholine (Morrison & Magistretti 1983). The activity in some of these systems may report on expectation of future reward (Cole & Robbins 1992; Schultz *et al*. 1995, 1997; Wise 1982). For example, honeybees can be conditioned to a sensory stimulus such as colour, shape or smell of a flower when paired with application of sucrose to the antennae or proboscis. An identified neuron, VUMmx1, projects widely throughout the entire bee brain, becomes active in response to sucrose, and its firing can substitute for the unconditioned odour stimulus in classical conditioning experiments. A simple model based on TD learning can explain many properties of bee foraging (Montague *et al*. 1994, 1995).

Real and co-workers (Real 1991; Real *et al*. 1990) performed a series of experiments on bumble bees foraging on artificial flowers whose colours, blue and yellow, predicted the delivery of nectar. They examined how bees respond to the mean and variability of this delivery in a foraging version of a stochastic two-armed-bandit problem (Berry & Fristedt 1985). All the blue flowers contained 2 µl of nectar, $\frac{1}{3}$ of the yellow flowers contained 6 µl, and the remaining $\frac{2}{3}$ of the yellow flowers contained no nectar at all. In practice, 85% of the bees' visits were to the constant-yield blue flowers despite the equivalent mean return from the more variable yellow flowers. When the contingencies for reward were reversed, the bees switched their preference for flower colour within one to three visits to flowers. Real and co-workers further demonstrated that the bees could be induced to visit the variable and constant flowers with equal frequency if the mean reward from the variable flower type was made sufficiently high.

This experimental finding shows that bumble bees, like honeybees, can learn to associate colour with reward. Further, colour and odour learning in honeybees has approximately the same time course as the shift in preference described above for the bumble bees (Gould 1987). It also indicates that under the conditions of a foraging task, bees prefer less variable rewards and compute the reward availability in the short term. This is a behavioural strategy used by a variety of animals under similar conditions for reward (Krebs *et al*. 1978; Real 1991; Real *et al*. 1990), suggesting a common set of constraints in the underlying neural substrate.

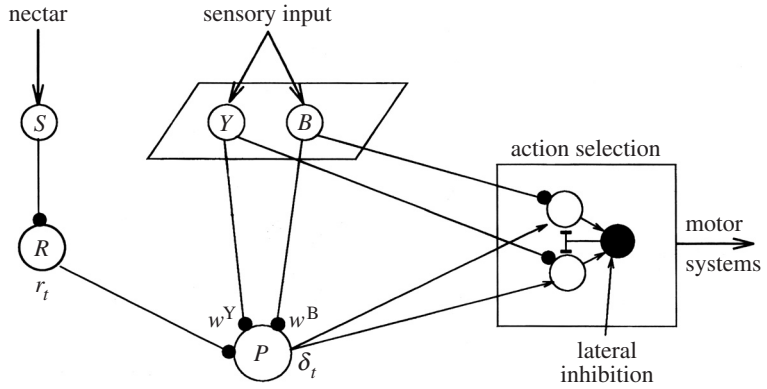

Figure 1. Neural architecture of the bee-foraging model. During bee foraging (Real 1991), sensory input drives the units B and Y representing blue and yellow flowers. These units project to a reinforcement neuron P through a set of variable weights (filled circles w^{B} and w^{Y}) and to an action selection system. Unit S provides input to R and fires while the bee sips the nectar. R projects its output r_t through a fixed weight to P. The variable weights onto P implement predictions about future reward r_t (see text) and P's output is sensitive to temporal changes in its input. The output projections of P, δ_t (lines with arrows), influence learning and also the selection of actions such as steering in flight and landing, as in equation (3.2) (see text). Modulated lateral inhibition (dark circle) in the action selection layer symbolizes this. Before encountering a flower and its nectar, the output of P will reflect the temporal difference only between the sensory inputs B and Y. During an encounter with a flower and nectar, the prediction error δ_t is determined by the output of B or Y and R, and learning occurs at connections w^{B} and w^{Y}. These strengths are modified according to the correlation between presynaptic activity and the prediction error δ_t produced by neuron P as in equation (3.1) (see text). Learning is restricted to visits to flowers. (Adapted from Montague *et al.* (1994).)

Figure 1 shows a diagram of the model architecture, which is based on the anatomy and physiological properties of VUMmx1. Sensory input drives the units 'B' and 'Y' representing blue and yellow flowers. These neurons (outputs x_t^{B} and x_t^{Y}, respectively, at time t) project through excitatory connection weights both to a diffusely projecting neuron P (weights w^{B} and w^{Y}) and to other processing stages which control the selection of actions such as steering in flight and landing. P receives additional input r_t through unchangeable weights. In the absence of nectar ($r_t = 0$), the net input to P becomes $P_t \equiv \boldsymbol{w}_t \cdot \boldsymbol{x}_t = w_t^{\mathrm{B}} x_t^{\mathrm{B}} + w_t^{\mathrm{Y}} x_t^{\mathrm{Y}}$.

Assume that the firing rate of P is sensitive only to changes in its input over time and habituates to constant or slowly varying input. Under this assumption, the error in prediction is given by δ_t in equation (2.3), and the weights can be updated according to the TD-learning rule in equation (2.4). This permits the weights onto P to act as predictions of the expected reward consequent on landing on a flower.

When the bee actually lands on a flower and samples the nectar, R influences the output of P through its fixed connection (figure 1). Suppose that just prior to sampling the nectar the bee switched to viewing a blue flower, for example. Then, since $r_{t-1} = 0$, δ_t would be $r_t - x_{t-1}^{\mathrm{B}} w_{t-1}^{\mathrm{B}}$. In this way, the term $x_{t-1}^{\mathrm{B}} w_{t-1}^{\mathrm{B}}$ is a prediction of the value of r_t and the difference $r_t - x_{t-1}^{\mathrm{B}} w_{t-1}^{\mathrm{B}}$ is the error in that prediction. Adjusting the weight w_t^{B} according to the TD rule in equation (2.4) allows the weight w_t^{B}, through P's outputs, to report to the rest of the brain the amount of reinforcement r_t expected from blue flowers when they are sensed.

As the model bee flies between flowers, reinforcement from nectar is not present ($r_t = 0$) and δ_t is proportional to $P_t - P_{t-1}$. w^{B} and w^{Y} can again be used as predictions but through modulation of action choice. For example, suppose the learning process in equation (2.4) sets w^{Y} less than w^{B}. In flight, switching from viewing yellow flowers to viewing blue flowers causes δ_t to be positive and biases the activity in any action selection units driven by outgoing connections from B. This makes the bee more likely than chance to land on or steer towards blue flowers.

The biological assumptions of this neural architecture are explicit:

(i) the diffusely projecting neuron changes its firing according to the temporal difference in its inputs;

(ii) the output of P is used to adjust its weights upon landing; and

(iii) the output otherwise biases the selection of actions by modulating the activity of its target neurons.

For the particular case of the bee, both the learning rule described in equation (2.4) and the biasing of action selection described above can be further simplified. Significant learning about a particular flower colour only occurs in the 1–2 s just prior to an encounter (Menzel & Erber 1978). This is tantamount to restricting weight changes to each encounter with the reinforcer, which allows only the sensory input just preceding the delivery or non-delivery of r_t to drive synaptic plasticity. We therefore make the learning rule punctate, updating the weights on a flower by flower basis. During each encounter with the reinforcer in the environment, P produces a prediction error $\delta_t = r_t - P_{t-1}$, where r_t is the actual reward at time t, and the last flower colour seen by the bee at time t, say blue, causes a prediction $P_{t-1} = w^{\mathrm{B}}_{t-1} x^{\mathrm{B}}_{t-1}$ of future reward r_t to be made through the weight w^{B}_{t-1} and the input activity x^{B}_{t-1}. The weights are then updated using the TD-learning rule,

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} + \lambda \delta_t \boldsymbol{x}_{t-1}, \tag{3.1}$$

where λ is the learning rate.

We model the temporal biasing of actions such as steering and landing with a probabilistic algorithm that uses the same weights onto P to choose which flower is actually visited on each trial. At each flower visit, the predictions are used directly to choose an action, according to

$$Q(Y) = \frac{\exp(\mu(w^{\mathrm{Y}} x^{\mathrm{Y}}))}{\exp(\mu(w^{\mathrm{B}} x^{\mathrm{B}})) + \exp(\mu(w^{\mathrm{Y}} x^{\mathrm{Y}}))}, \tag{3.2}$$

where $Q(Y)$ is the probability of choosing a yellow flower. Values of $\mu > 0$ amplify the difference between the two predictions, so that larger values of μ make it more likely that the larger prediction will result in choice toward the associated flower colour. In the limit as $\mu \to \infty$ this approaches a winner-take-all rule.

To apply the model to the foraging experiment, it is necessary to specify how the amount of nectar in a particular flower gets reported to P. We assume that the reinforcement neuron R delivers its signal r_t as a saturating function of nectar volume (figure 2*a*). Harder & Real (1987) suggest just this sort of decelerating function of nectar volume and justify it on biomechanical grounds. Figure 2*b* shows the behaviour

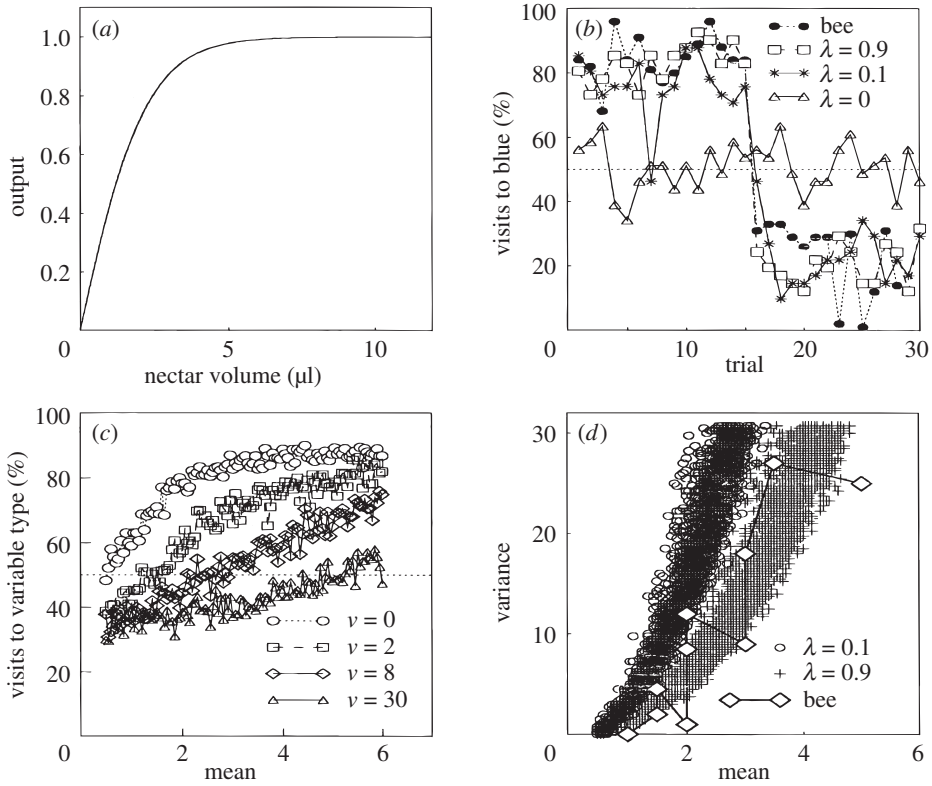

Figure 2. Simulations of bee foraging behaviour using TD learning. (*a*) Reinforcement neuron output as a function of nectar volume for a fixed concentration of nectar (Real 1991; Real *et al*. 1990). (*b*) Proportion of visits to blue flowers. Each trial represents approximately 40 flower visits averaged over five real bees and exactly 40 flower visits for a single model bee. Trials 1–15 for the real and model bees had blue flowers as the constant type; the remaining trials had yellow flowers as constant. At the beginning of each trial, w^{Y} and w^{B} were set to 0.5, which is consistent with evidence that information from past foraging bouts is not used (Menzel & Erber 1978). The real bees were more variable than the model bees: sources of stochasticity such as the two-dimensional feeding ground were not represented. The real bees also had a slight preference for blue flowers (Menzel *et al*. 1974). Note the slow drop for $\lambda = 0.1$ when the flowers are switched. (*c*) Method of selecting indifference points. The indifference point is taken as the first mean for a given variance ($\boldsymbol{v}$ in the legend) for which a stochastic trial demonstrates the indifference. This method of calculation tends to bias the indifference points to the left. (*d*) Indifference plot for model and real bees. Each point represents the (mean, variance) pair for which the bee sampled each flower type equally. The circles are for $\lambda = 0.1$ and the pluses are for $\lambda = 0.9$. (Adapted from Montague *et al*. (1994).)

of model bees compared with that of real bees (Real 1991). Further details are presented in the figure legend.

The behaviour of the model matched the observed data for $\lambda = 0.9$, suggesting that the real bee uses information over a small time window for controlling its foraging (Real 1991). At this value of λ, the average proportion of visits to blue was 85% for the real bees and 83% for the model bees. The constant and variable flower types were switched at trial 15 and both bees switched flower preference in one to

three subsequent visits. The average proportion of visits to blue changed to 23% and 20%, respectively, for the real and model bee. Part of the reason for the real bees' apparent preference for blue may come from inherent biases. Honey bees, for instance, are known to learn about shorter wavelengths more quickly than others (Menzel *et al*. 1974). In our model, the learning rate λ is a measure of the length of time over which an observation exerts an influence on flower selection rather than being a measure of the bee's time horizon in terms of the mean rate of energy intake (Real 1991; Real *et al*. 1990).

Real bees can be induced to forage equally on the constant and variable flower types if the mean reward from the variable type is made sufficiently large (figure 2*c*, *d*). For a given variance, the mean reward was increased until the bees appeared to be indifferent between their choice of flowers. In this experiment, the constant flower type contained 0.5 µl of nectar. The data for the real bee are shown as points connected by a solid line in order to make clear the envelope of the real data. The indifference points for $\lambda = 0.1$ (circles) and $\lambda = 0.9$ (pluses) also demonstrate that a higher value of λ is again better at reproducing the bee's behaviour. The model captured both the functional relationship and the spread of the real data.

This model was implemented and tested in several ways. First, a virtual bee was simulated foraging in a virtual field of coloured flowers. In these simulations, the field of view of the bee was updated according to the decision rule above (equation (3.2)), so that the bee eventually 'landed' on a virtual flower and the trial was repeated (Montague *et al*. 1995). In a second test, an actual robot bee was constructed and placed in the centre of a circular field. The robot bee had a camera that detected coloured paper on the walls of the enclosure and moved toward the wall using the above decision rule (P. Yates, P. R. Montague, P. Dayan & T. J. Sejnowski, unpublished results). In each of these tests, the statistics of flower visits qualitatively confirmed the results shown in figure 2, despite the differences in the dynamics of the model bees in the two circumstances. This is an important test since the complicated contingencies of the real world, such as the slip in the wheels of the robot and random influences that are not taken into account in the idealized simulations shown here, did not affect the regularities in the overall behaviour that emerged from the use of the TD-learning rule.

A similar model has been used to model the primate dopamine pathways that also may be involved in the prediction of future reward (Montague *et al*. 1996; Schultz *et al*. 1997). In this case, the neurons are located in the ventral tegmental area and project diffusely through the basal ganglia and the cerebral cortex, particularly to the prefrontal cortex, which is involved in planning actions. Thus, there is a remarkable evolutionary convergence of reward prediction systems in animals as diverse as bees and primates.

4. TD learning at the cellular level: spike-timing-dependent plasticity

A recently discovered phenomenon in spiking neurons appears to share some of the characteristics of TD learning. Known as spike-timing-dependent synaptic plasticity or temporally asymmetric Hebbian learning, the phenomenon captures the influence of relative timing between input and output spikes in a neuron. Specifically, an input synapse to a given neuron that is activated slightly before the neuron fires is strengthened, whereas a synapse that is activated slightly after is weakened. The

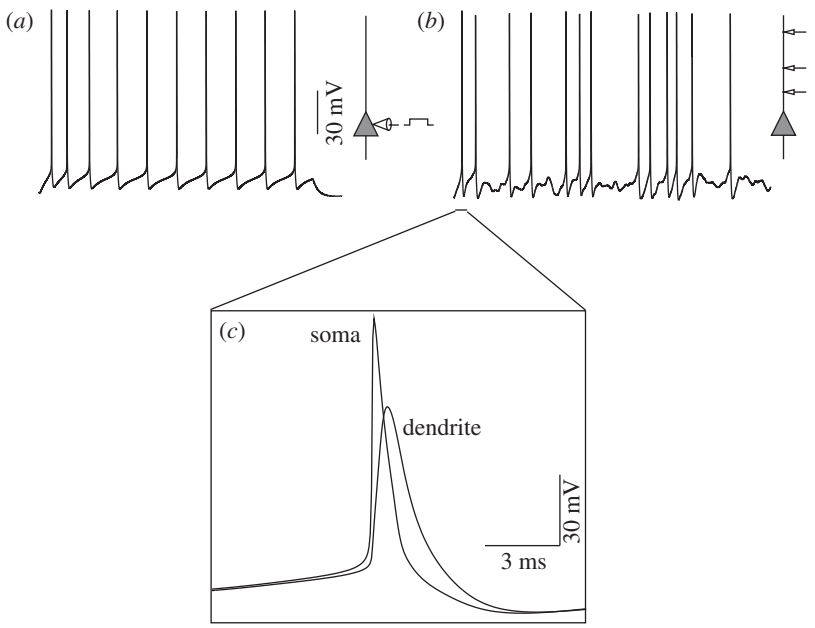

Figure 3. Model neuron response properties. (*a*) Response of a model neuron to a 70 pA current pulse injection into the soma for 900 ms. (*b*) Response of the same model neuron to Poisson distributed excitatory and inhibitory synaptic inputs at random locations on the dendrite. (*c*) Example of a back-propagating action potential in the dendrite of the model neuron as compared with the corresponding action potential in the soma (enlarged from the initial portion of the trace in (*b*)). (From Rao & Sejnowski (2001).)

window of plasticity typically ranges from +40 to −40 ms. Such a form of synaptic plasticity has been observed in recurrent cortical synapses (Markram *et al*. 1997), in the hippocampus (Bi & Poo 1998; Levy & Steward 1983), in the tectum (Zhang *et al*. 1998), and in layer II/II of rat somatosensory cortex (Feldman 2000).

In order to ascertain whether spike-timing-dependent plasticity in cortical neurons can be interpreted as a form of TD learning, we used a two-compartment model of a cortical neuron consisting of a dendrite and a soma-axon compartment (figure 3). 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 (Mainen & Sejnowski 1996). Four voltage-dependent currents and one calcium-dependent current were simulated, as in Mainen & Sejnowski (1996): fast Na^+, I_{Na}; fast K^+, I_{Kv}; slow non-inactivating K^+, I_{Km}; high voltage-activated Ca^{2+}, I_{Ca} and calcium-dependent K^+ current, I_{KCa}. The following active conductance densities were used in the soma-axon compartment (in $pS\,\mu m^{-2}$): $\bar{g}_{Na} = 40\,000$ and $\bar{g}_{Kv} = 1400$. For the dendritic compartment, we used the following values: $\bar{g}_{Na} = 20$, $\bar{g}_{Ca} = 0.2$, $\bar{g}_{Km} = 0.1$, and $\bar{g}_{KCa} = 3$, with leak conductance $33.3\,\mu S\,cm^{-2}$ and specific membrane resistance $30\,k\Omega\,cm^{-2}$. The presence of voltage-activated sodium channels in the dendrite allowed back propagation of action potentials from the soma into the dendrite as shown in figure 3*c*.

Conventional Hodgkin–Huxley-type kinetics were used for all currents (integration time-step, 25 μs; temperature, 37 °C). Ionic currents I were calculated using the

ohmic equation

$$I = \bar{g} A^x B (V - E), \tag{4.1}$$

where $\bar{g}$ is the maximal ionic conductance density, A and B are activation and inactivation variables, respectively (x denotes the order of kinetics; see Mainen & Sejnowski 1996), and E is the reversal potential for the given ion species ($E_K = -90\,\mathrm{mV}$, $E_{\mathrm{Na}} = 60\,\mathrm{mV}$, $E_{\mathrm{Ca}} = 140\,\mathrm{mV}$, $E_{\mathrm{leak}} = -70\,\mathrm{mV}$). For all compartments, the specific membrane capacitance was $0.75\,\mu\mathrm{F\,cm}^{-2}$. Two key parameters governing the response properties of the model neuron are (Mainen & Sejnowski 1996) the ratio of axo-somatic area to dendritic membrane area (ρ) and the coupling resistance between the two compartments (κ). For the present simulations, we used the values $\rho = 150$ (with an area of $100\,\mu\mathrm{m}^2$ for the soma-axon compartment) and a coupling resistance of $\kappa = 8\,\mathrm{M}\Omega$. Poisson-distributed synaptic inputs to the dendrite (see figure 3*b*) were simulated using alpha-function-shaped (Koch 1999) current pulse injections (time constant 5 ms) at Poisson intervals with a mean presynaptic firing frequency of 3 Hz.

To study plasticity, 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 with model parameters fitted to whole-cell recorded AMPA (a-amino 3-hydroxy 5-methylisoxazole 4-proprionic acid) currents (see Destexhe *et al*. (1998) for more details). 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. The delay parameter Δt was set to 10 ms to yield results consistent with previous physiological experiments (Bi & Poo 1998; Markram *et al*. 1997). Presynaptic input to the model neuron was paired with postsynaptic spiking by injecting a depolarizing current pulse (10 ms, 200 pA) into the soma. Changes in synaptic efficacy were monitored by applying a test stimulus before and after pairing, and recording the EPSP evoked by the test stimulus.

Figure 4 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 by 58.5% in the former case, it was decreased by 49.4% in the latter case, qualitatively similar to experimental observations (Bi & Poo 1998; Markram *et al*. 1997). The critical window for synaptic modifications in the model depends on the parameter Δt as well as the shape of the back-propagating action potential (AP). This window of plasticity was examined by varying the time-interval between presynaptic stimulation and postsynaptic spiking (with $\Delta t = 10\,\mathrm{ms}$). As shown in figure 4*c*, changes in synaptic efficacy exhibited a highly asymmetric dependence on spike timing similar to physiological data (Markram *et al*. 1997). 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 (Markram *et al*. 1997), *Xenopus* tectal neurons (Zhang *et al*. 1998), and cultured hippocampal neurons (Bi & Poo 1998), a narrow transition zone (roughly 3 ms in the model) separated the potentiation and depression windows. The stability of this spike-based TD rule is analysed in Rao & Sejnowski (2001). It is shown that the stability of the TD-learning rule for spike-timing-dependent synaptic

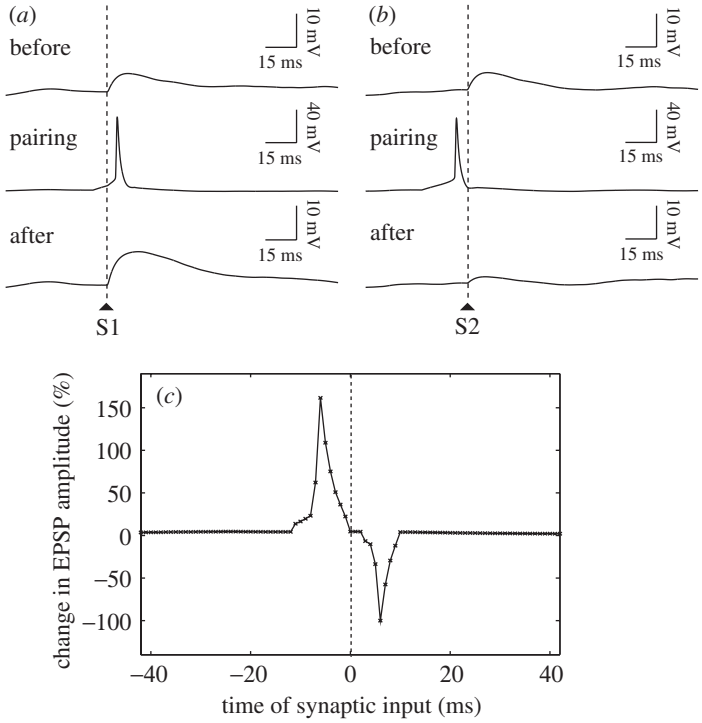

Figure 4. Synaptic plasticity in a model neocortical neuron. (*a*) EPSP in the model neuron evoked by a presynaptic spike (S1) at an excitatory synapse ('before'). Pairing this presynaptic spike with postsynaptic spiking after a 5 ms delay ('pairing') induces long-term potentiation ('after'). (*b*) If presynaptic stimulation (S2) occurs 5 ms after postsynaptic firing, the synapse is weakened, resulting in a corresponding decrease in peak EPSP amplitude. (*c*) Temporally asymmetric window of synaptic plasticity obtained by varying the delay between pre- and postsynaptic spiking (negative delays refer to presynaptic before postsynaptic spiking). (From Rao & Sejnowski (2001).)

plasticity depends crucially on whether the temporal window parameter Δt is comparable in magnitude to the width of the back-propagating AP at the location of the synapse. An upper bound on the maximal synaptic conductance may be required to ensure stability in general (Abbott & Song 1999; Gerstner *et al*. 1996; Song *et al*. 2000). Such a saturation constraint is partly supported by experimental data (Bi & Poo 1998). An alternative approach that might be worth considering is to explore a learning rule that uses a continuous form of the TD error, where, for example, an average of postsynaptic activity is subtracted from the current activity (Doya 2000; Montague & Sejnowski 1994). Such a rule may offer better stability properties than the discrete TD rule that we have used, although other parameters, such as the window over which average activity is computed, may still need to be carefully chosen.

(*a*) *Biophysical mechanisms for spike-based TD learning*

An interesting question is whether a biophysical basis can be found for the TD-learning model described above. Neurophysiological and imaging studies suggest a

role for dendritic Ca^{2+} signals in the induction of spike-timing-dependent long-term potentiation (LTP) and long-term depression (LTD) in hippocampal and cortical neurons (Koester & Sakmann 1998; Magee & Johnston 1997; Paulsen & Sejnowski 2000). In particular, when an EPSP preceded a postsynaptic action potential, the Ca^{2+} transient in dendritic spines, where most excitatory synaptic connections occur, was observed to be larger than the sum of the Ca^{2+} signals generated by the EPSP or AP alone, causing LTP; on the other hand, when the EPSP occurred after the AP, the Ca^{2+} transient was found to be a sublinear sum of the signals generated by the EPSP or AP alone, resulting in LTD (Koester & Sakmann 1998; Linden 1999; Paulsen & Sejnowski 2000). Possible sources contributing to the spinous Ca^{2+} transients include Ca^{2+} ions entering through NMDA (N-methyl-D-aspartate) receptors (Bliss & Collingridge 1993; Koester & Sakmann 1998), voltage-gated Ca^{2+} channels in the dendrites (Schiller *et al*. 1998), and calcium release from intracellular stores (Emptage 1999).

How do the above experimental observations support the TD model? In a recent study (Franks *et al*. 1999), a Monte Carlo simulation program MCell (Stiles *et al*. 2000) was used to model the Ca^{2+} dynamics in dendritic spines following pre- and postsynaptic activity, and to track the binding of Ca^{2+} to endogenous proteins. The influx of Ca^{2+} into a spine is governed by the rapid depolarization pulse caused by the back-propagating AP. The width of the back-propagating AP is much smaller than the time course of glutamate binding to the NMDA receptor. As a result, the dynamics of Ca^{2+} influx and binding to calcium-binding proteins such as calmodulin depends highly nonlinearly on the relative timing of presynaptic activation (with release of glutamate) and postsynaptic depolarization (due to the back-propagating AP). In particular, due to its kinetics, the binding protein calmodulin could serve as a differentiator of intracellular calcium concentration, causing synapses to either potentiate or depress depending on the spatio-temporal profile of the dendritic Ca^{2+} signal (Franks *et al*. 1999). As a consequence of these biophysical mechanisms, the change in synaptic strength depends, to a first approximation, on the time derivative of the postsynaptic activity, as postulated by the TD model.

(*b*) *Learning to predict using spike-based TD learning*

Our results suggest that spike-timing-dependent plasticity in neocortical synapses can be interpreted as a form of TD learning for prediction. To see how a network of model neurons can learn to predict sequences using such a learning mechanism, consider the simple case of two excitatory neurons, N1 and N2, connected to each other, receiving inputs from two separate input neurons, I1 and I2 (figure 5*a*). Model neuron parameters were the same as those used in § 4. Suppose input neuron I1 fires before input neuron I2, causing neuron N1 to fire (figure 5*b*). The spike from N1 results in a sub-threshold EPSP in N2 due to the synapse S2. If input arrives from I2 between 1 ms and 12 ms after this EPSP and if the temporal summation of these two EPSPs causes N2 to fire, 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.

After several exposures to the I1–I2 training sequence, when I1 causes neuron N1 to fire, N1 in turn causes N2 to fire several milliseconds *before* input I2 occurs due to the potentiation of the recurrent synapse S2 in previous trials (figure 5*c*). Input neuron I2

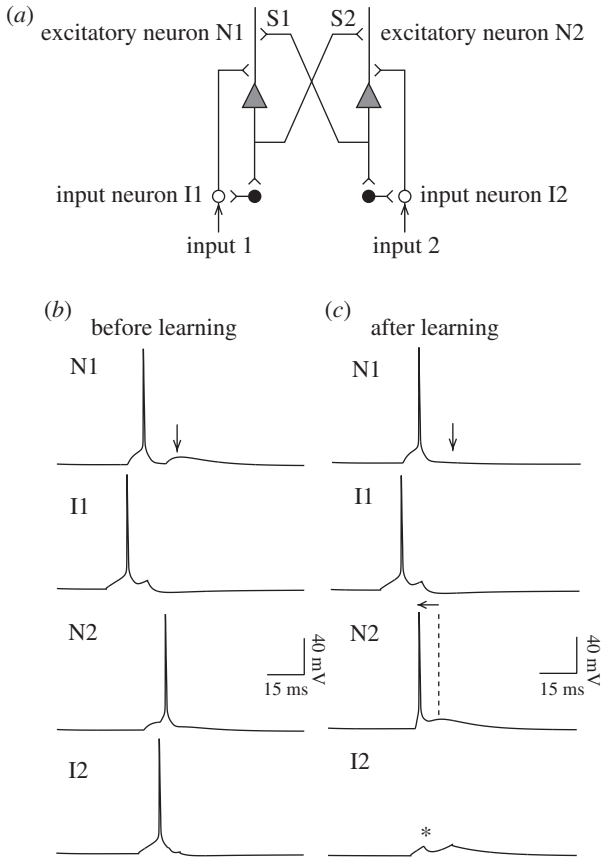

Figure 5. Learning to predict using spike-based TD learning. (*a*) Network of two model neurons N1 and N2 recurrently connected via excitatory synapses S1 and S2, with input neurons I1 and I2. N1 and N2 inhibit the input neurons via inhibitory interneurons (filled circles). (*b*) Network activity elicited by the sequence I1 followed by I2. (*c*) Network activity for the same sequence after 40 trials of learning. Due to strengthening of recurrent synapse S2, recurrent excitation from N1 now causes N2 to fire several ms before the expected arrival of input I2 (dashed line), allowing it to inhibit I2 (asterisk). Synapse S1 has been weakened, preventing re-excitation of N1 (downward arrows show decrease in EPSP). (From Rao & Sejnowski (2001).)

can thus be inhibited by the predictive feedback from N2 just before the occurrence of imminent input activity (marked by an asterisk in figure 5*c*). This inhibition prevents input I2 from further exciting N2, thereby implementing a negative feedback-based predictive coding circuit (Rao & Ballard 1999). Similarly, a positive feedback loop between neurons N1 and N2 is avoided because the synapse S1 was weakened in previous trials (see arrows in figure 5*b*, *c*, top row). Figure 6*a* 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 6*b*. 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. Although the postsynaptic spike continues to occur shortly after the activation of synapse S2, this synapse is

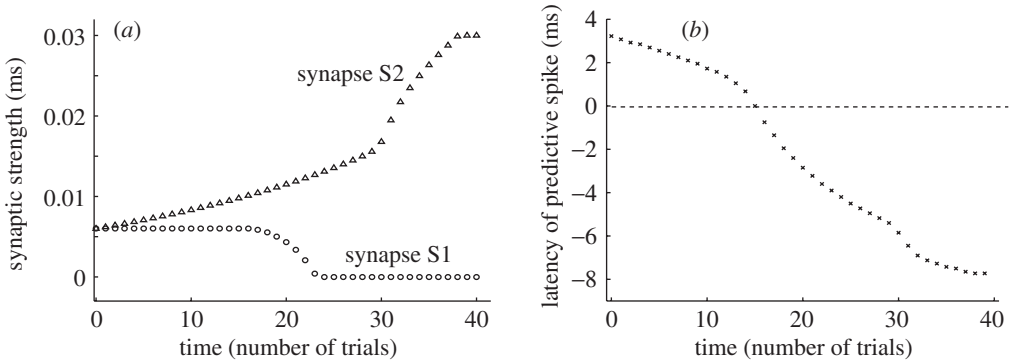

Figure 6. Synaptic strength and latency reduction due to learning. (*a*) Potentiation and depression of synapses S1 and S2, respectively, during the course of learning. Synaptic strength was defined as maximal synaptic conductance in the kinetic model of synaptic transmission (Destexhe *et al*. 1998). (*b*) Latency of predictive spike in N2 during the course of learning measured with respect to the time of input spike in I2 (dotted line). (From Rao & Sejnowski (2001).)

prevented from assuming larger values due to a saturation constraint of 0.03 μS on the maximal synaptic conductance (see above for a discussion of this constraint).

(*c*) *Prediction, visual motion detection and direction selectivity*

In related work (Rao & Sejnowski 2000), we have shown how spike-based TD learning can explain the development of direction selectivity in recurrent cortical networks, yielding receptive field properties similar to those observed in awake monkey V1. We first simulated a simple motion detection circuit consisting of a single chain of nine recurrently connected excitatory cortical neurons (figure 7*a*). Each neuron in the chain initially received symmetric excitatory and inhibitory inputs of the same magnitude (maximal synaptic conductance 0.003 μS) from its preceding and successor neurons (figure 7*b*, 'before learning'). Excitatory and inhibitory synaptic currents were calculated using kinetic models of synaptic transmission based on properties of AMPA and $GABA_A$ (γ-aminobutyric acid A) receptors as determined from whole-cell recordings (see Destexhe *et al*. 1998). Neurons in the network were exposed to 100 trials of retinotopic sensory input consisting of moving pulses of excitation in the rightward direction (5 ms pulse of excitation at each neuron). These inputs, which approximate the depolarization caused by retinotopic inputs from the LGN (lateral geniculate nucleus), were sufficient to elicit a spike from each neuron.

The effects of spike-timing-dependent learning on the excitatory and inhibitory synaptic connections in the network are shown in figure 7*b* ('after learning'). There is a profound asymmetry in the developed pattern of excitatory connections from the preceding and successor neurons to neuron 0 in figure 7*b*. The synaptic conductances of excitatory connections from the left-side have been strengthened, while the ones from the right-side have been weakened. This result can be explained as follows: due to the rightward motion of the input stimulus, neurons on the left side fire (on average) a few milliseconds before neuron 0, while neurons on the right side fire (on average) a few milliseconds after neuron 0; as a result, the synaptic strength of connections from the left side are increased, while the synaptic strength for connections from the right side are decreased, as prescribed by the spike-timing-dependent

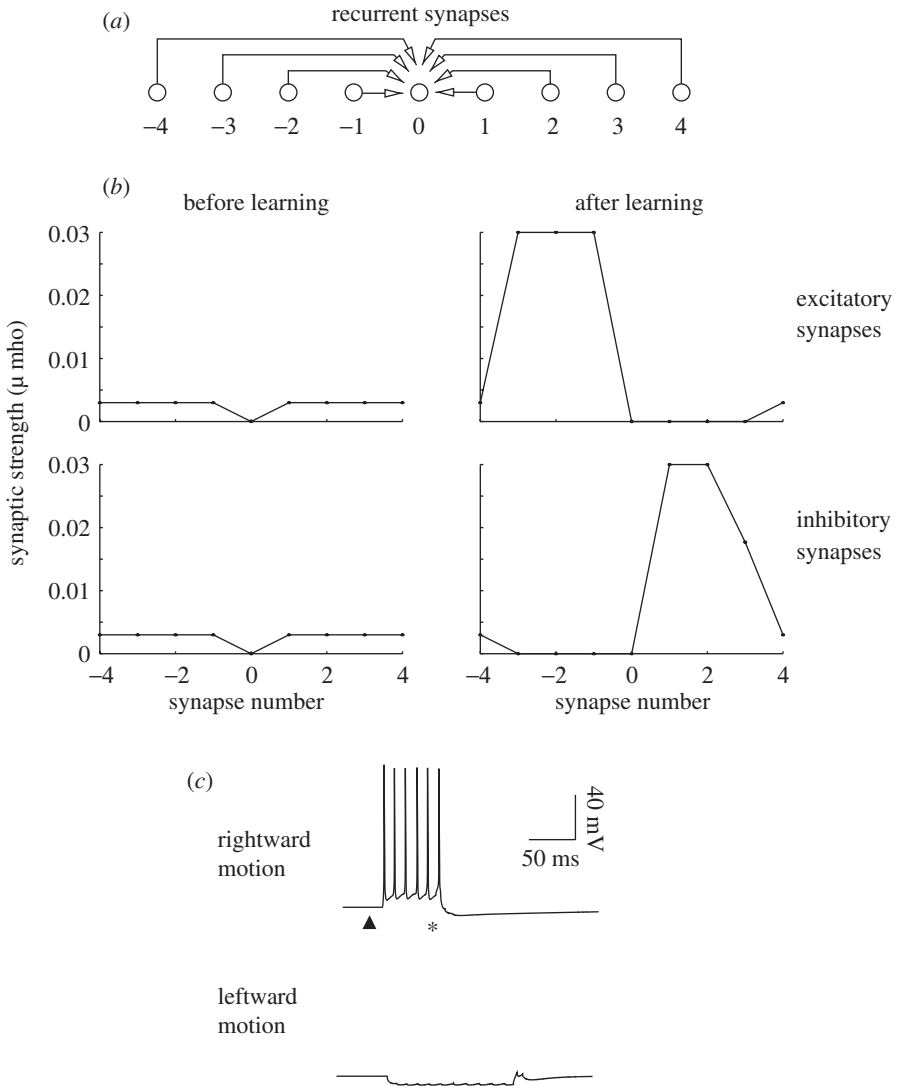

Figure 7. Emergence of direction selectivity in the spike-based TD-learning model. (*a*) Schematic depiction of recurrent connections to a given neuron (labelled '0') from four preceding and four successor neurons in its chain. (*b*) Synaptic strength of recurrent excitatory and inhibitory connections to neuron 0 before and after learning. Note the symmetry in connections before learning and the asymmetry in connections after spike-timing-dependent learning. Synapses were adapted during 100 trials of exposure to rightward moving stimuli. (*c*) Direction-selective response of neuron 0 to rightward moving stimuli after learning. Due to recurrent excitation from preceding neurons, the neuron starts firing a few milliseconds before the expected arrival time of its input (marked by an asterisk). The black triangle represents the time at which the input stimulus begins its rightward motion.

learning window in figure 4*c*. The opposite pattern of connectivity develops for the inhibitory connections because these were modified according to an asymmetric anti-Hebbian learning rule that reversed the polarity of the rule in figure 4*c*. Such a rule

is consistent with spike-timing-dependent anti-Hebbian plasticity observed in some classes of inhibitory interneurons (Bell *et al*. 1997). Alternatively, one could keep the level of inhibition constant (for example, at 0.015 μS) and obtain qualitatively similar results because a decrease in the strength of the corresponding excitatory connections, as shown in figure 7*b*, would again tilt the balance in favour of inhibition on the right side of neuron 0.

The responses of neuron 0 to rightward and leftward moving stimuli are shown in figure 7*c*. As expected from the learned pattern of connections, the neuron responds vigorously to rightward motion but not to leftward motion. Similar responses selective for rightward motion were exhibited by the other neurons of which the network is composed. More interestingly, each neuron fires 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. Such predictive neural activity is characteristic of temporally asymmetric learning rules (see, for example, Abbott & Blum 1996; Rao & Sejnowski 2001). In contrast, motion in the non-preferred direction triggered recurrent inhibition and little or no response from the model neurons.

(i) *Detecting multiple directions of motion*

To investigate the question of how selectivity for different directions of motion may emerge simultaneously, we simulated a network comprising two parallel chains of neurons (see figure 8*a*), each containing 55 neurons, with mutual inhibition (black arrows) between corresponding pairs of neurons along the two chains. As in the previous simulation, a given excitatory neuron received both excitation and inhibition from its predecessors and successors, as shown in figure 8*b* for a neuron labelled '0'. Inhibition at a given neuron was mediated by an inhibitory interneuron (black circle) which received excitatory connections from neighbouring excitatory neurons (figure 8*b*, 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 8*c*). To break the symmetry between the two chains, one may: (i) select small randomly chosen values for the synaptic conductances in the two chains, or (ii) provide a slight bias in the recurrent excitatory connections, so that neurons in one chain may fire slightly earlier than neurons in the other chain for a given motion direction. Both alternatives succeed in breaking symmetry during learning. We report here the results for alternative (ii), which is supported by experimental evidence indicating the presence of a small amount of initial direction selectivity in cat visual cortical neurons before eye opening (Movshon & Sluyters 1981).

To evaluate the consequences of spike-based TD learning in the two-chain network, model neurons were exposed alternately to leftward- and rightward-moving stimuli for a total of 100 trials. The excitatory connections (labelled 'EXC' in figure 8*b*) were modified according to the TD-learning rule in figure 4*c*, while the excitatory connections onto the inhibitory interneuron (labelled 'INH') were modified according to the asymmetric anti-Hebbian learning rule, as in the previous simulation. The synaptic conductances learned by two neurons (marked N1 and N2 in figure 8*a*) located at corresponding positions in the two chains after 100 trials of exposure to the moving stimuli are shown in figure 8*c* (solid line). The excitatory and inhibitory connections to neuron N1 exhibit a marked asymmetry, with excitation originating

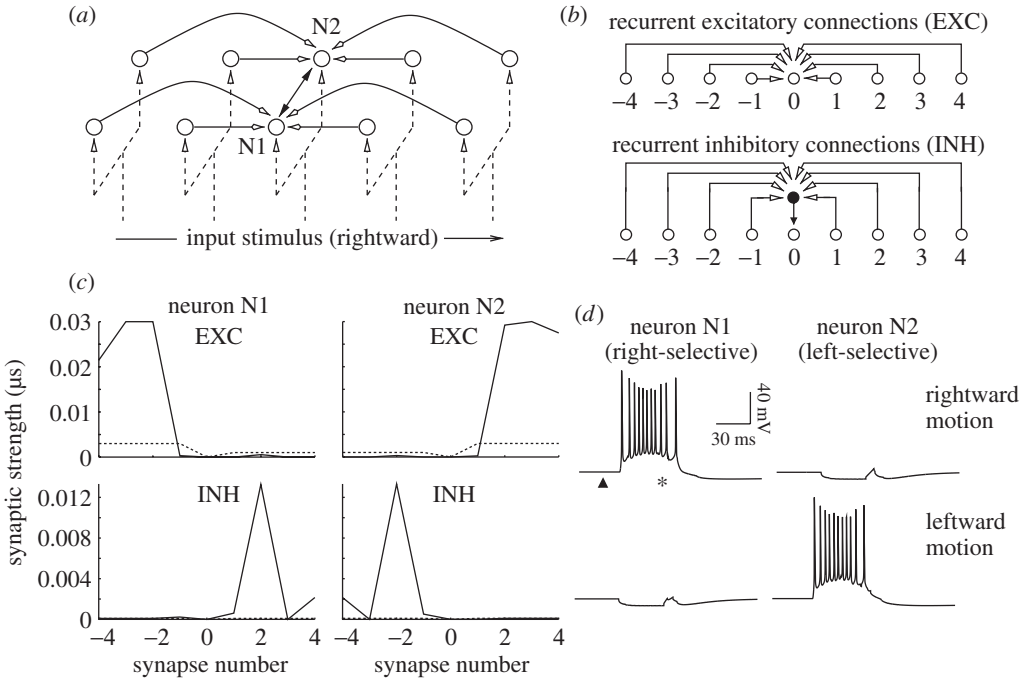

Figure 8. Detecting multiple directions of motion. (*a*) A model network consisting of two chains of recurrently connected neurons receiving retinotopic inputs. A given neuron receives recurrent excitation and recurrent inhibition (white-headed arrows) as well as inhibition (black-headed arrows) from its counterpart in the other chain. (*b*) Recurrent connections to a given neuron (labelled '0') arise from four preceding and four succeeding neurons in its chain. Inhibition at a given neuron is mediated via a GABAergic interneuron (black 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. After 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 input arrives at its soma (marked by an asterisk) due to recurrent excitation from preceding neurons. The black triangle represents the start of input stimulation in the network.

from neurons on the left and inhibition from neurons on the right. Neuron N2 exhibits the opposite pattern of connectivity.

As expected from the learned pattern of connectivity, neuron N1 was found to be selective for rightward motion, while neuron N2 was selective for leftward motion (figure 8*d*). 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. Thus, the learned pattern of connectivity allows the direction selective neurons comprising the network to conjointly code for and predict the moving input stimulus in each possible direction of motion.

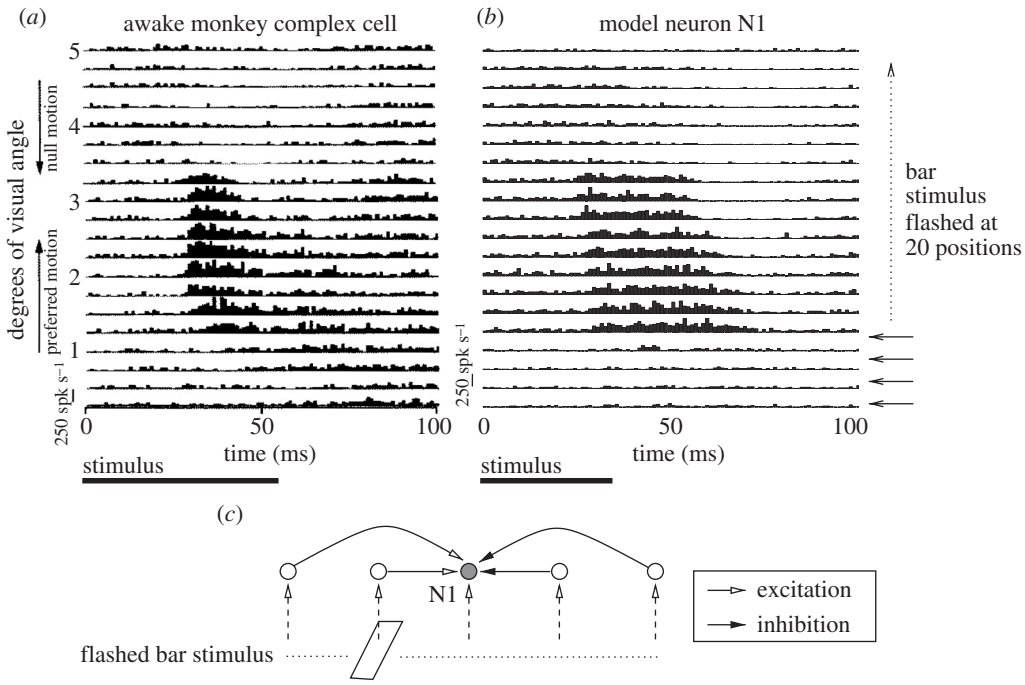

Figure 9. Comparison of monkey and model space-time response plots to single flashed bars. (*a*) Sequence of PSTHs obtained by flashing optimally oriented bars at 20 positions across the 5°-wide receptive field (RF) of a complex cell in alert monkey V1 (from Livingstone 1998). The cell's preferred direction 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. (*b*) 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. The lower PSTHs represent stimulations on the preferred side, while upper PSTHs represent stimulations on the null side. (*c*) Interpretation of the space-time plots in the model. Bars flashed on the left (preferred) side of the recorded cell (shaded) cause progressively greater excitation as the stimulation site approaches the recorded cell's location. Bars flashed to the right of the cell cause inhibition due to the predominantly inhibitory connections that develop on the right (null) side during learning.

(ii) *Comparison with awake-monkey complex-cell responses: first-order analysis*

Like complex cells in the primary visual cortex, model neurons were found to be direction selective throughout their receptive field. This phase-invariant direction selectivity is a consequence of the fact that at each retinotopic location, the corresponding neuron in the chain receives the same pattern of asymmetric excitation and inhibition from its neighbours 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 modelling studies (Chance *et al.* 1999), suggesting that recurrent connections may be responsible for the spatial-phase invariance of complex-cell responses.

The model predicts that the neuro-anatomical connections for a direction selective neuron should exhibit a pattern of asymmetrical excitation and inhibition similar to figure 8*c*. A recent study of complex 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 (Livingstone 1998). In this study, optimally oriented bars were flashed at random positions in a cell's receptive field, and a reverse correlation map was calculated from a record of eye position, spike occurrence and stimulus position. Figure 9*a* depicts an eye-position corrected reverse correlation map for a complex cell, with time on the x-axis and stimulus position on the y-axis: each row of the map is the post-stimulus time histogram of spikes elicited for a bar flashed at that spatial position. The map thus depicts the firing rate of the cell as a function of the retinal position of the stimulus and time after stimulus onset.

For comparison with these 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. As shown in figure 9*a*, *b*, there is good qualitative agreement between the space-time response plot for the 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, but unlike space-time maps in simple cells, this slant cannot be used to predict the neuron's velocity preference (see Livingstone (1998) for more details). However, assuming a 200 μm separation between excitatory model neurons in each chain and using known values for the cortical magnification factor in monkey striate cortex (Tootell *et al*. 1988), one can estimate the preferred stimulus velocity of model neurons to be in the range of $3.1^\circ\ s^{-1}$ in the fovea and $27.9^\circ\ s^{-1}$ in the periphery (at an eccentricity of 8°), which is within the range of monkey V1 velocity preferences ($1–32^\circ\ s^{-1}$) (Livingstone 1998; Van Essen 1985).

(iii) *Comparison with awake-monkey complex-cell responses: second-order analysis*

Complex cells are known to exhibit higher-order interactions between two successively presented stimuli. For example, the response to two bars presented sequentially at two different positions is generally not a linear function of the responses to the bars presented individually. In the case of the model network, we would expect the asymmetry in synaptic connections to give rise to nonlinear facilitation if the two bars are flashed along the preferred direction relative to each other and a reduction in response for bars flashed in the opposite direction (see figure 10*a*).

To study such two-bar interactions in complex cells in awake monkey V1, single bars of optimal orientation were flashed within a direction-selective cell's receptive field at a series of locations along the dimension perpendicular to stimulus orientation (these experiments were conducted in Margaret Livingstone's laboratory at Harvard Medical School). A continuous record was kept of eye position (at 250 Hz), spike occurrence (1 ms resolution), and stimulus position. A reverse correlation analysis was performed, after correcting for eye position, to produce two-bar interaction maps as shown in figure 10*b*. These maps show how the response to one stimulus is influenced by a preceding stimulus, as a function of the two stimulus locations. Thus, for each of the plots shown, the y-axis represents the spatial position of bar 1,

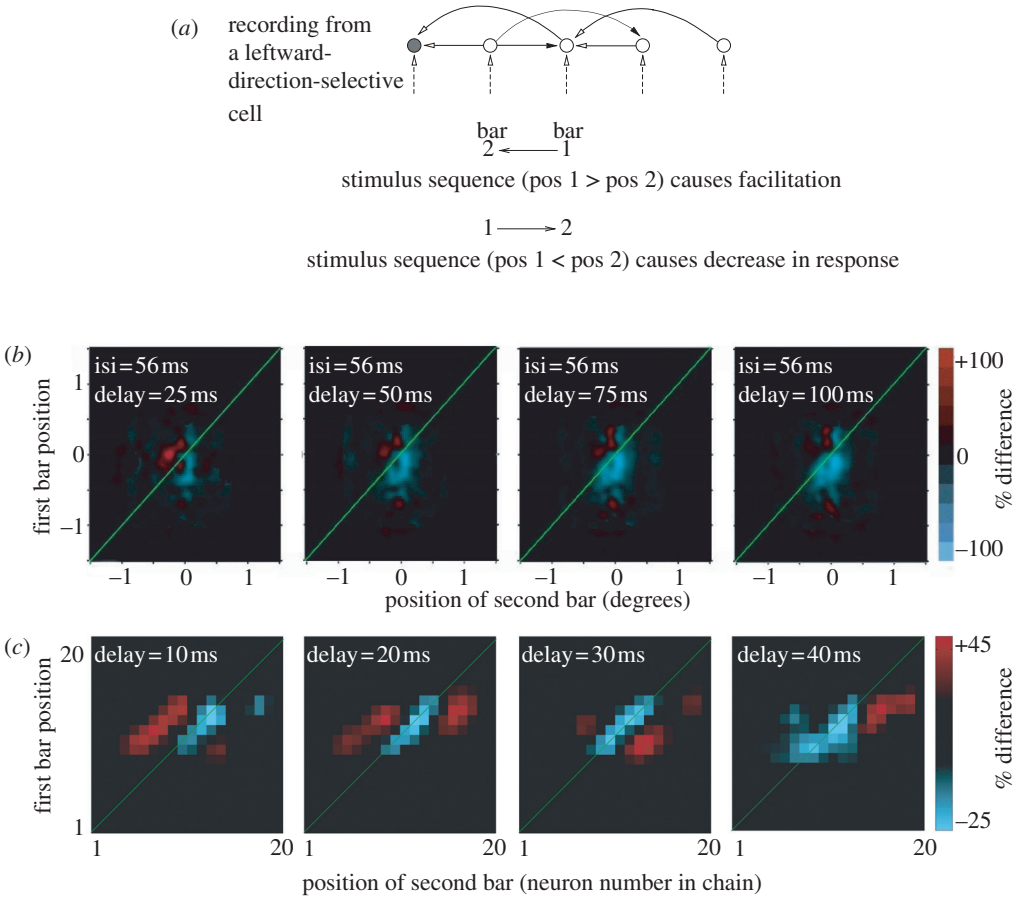

Figure 10. Two-bar interactions. (*a*) Model predictions for sequential presentation of two optimally oriented bars. Black arrowheads represent inhibitory connections. Facilitation is predicted when spatial position of bar 1 is greater than that of bar 2 (relative motion in the preferred direction); a reduction in response is expected when position of bar 2 is higher than that of bar 1 due to recruitment of inhibition. (*b*) Sequential two-bar interaction maps for a direction selective complex cell in awake monkey V1. The maps show the percent difference in response after subtracting the average responses to the individual bars from the cell's two-bar response. (*c*) Two-bar interaction maps for the model network. Note the slightly different scale bar for model data, compared with the experimental data. Both model and experimental data show qualitatively similar sequential interactions consistent with the expectations in (*a*), namely, facilitation (red) for spatial points above the diagonal (bar 1 position higher than bar 2 position) and a reduction in response (blue) for points close to and below the diagonal.

while the x-axis represents the position of bar 2, which was flashed after an inter-stimulus-interval (ISI) of 56 ms after bar 1. The four plots represent the evolution of the cell's response to the two-bar sequence at delays of 25 ms, 50 ms, 75 ms, and 100 ms, respectively, after the onset of bar 2. The average responses for the individual bars were subtracted from each of the plots to show any facilitation or reduction in responses due to sequential interactions.

Facilitation (red) can be observed above the diagonal line for all four plots in figure 10*b*. Locations above the diagonal represent cases where the two-bar sequence is flashed in the preferred direction of the cell (see figure 10*a*). A reduction in the cell's response (blue) occurs at longer delays, predominantly at positions below the diagonal. This is consistent with the model predictions sketched in figure 10*a*. A repetition of the two-bar experiment in the model yielded interaction plots that were qualitatively similar to the physiological data (figure 10*c*). The main differences are in the time-scale and magnitude of facilitation/reduction in the responses, both of which could be fine-tuned, if necessary, by adjusting model parameters such as the maximal allowed synaptic conductance, synaptic delays and the number of neurons used in the simulated network.

5. Discussion

Several theories of prediction and sequence learning in the brain have been proposed, based on statistical and information theoretic ideas (Abbott & Blum 1996; Barlow 1998; Daugman & Downing 1995; Dayan & Hinton 1996; Minai & Levy 1993; Montague & Sejnowski 1994; Rao & Ballard 1999). The bee-foraging model illustrates the utility of the TD model in understanding how animals learn to predict at the behavioural level. Our biophysical simulations suggest a possible implementation of TD-like models in cortical circuitry. Given the universality of the problem of encoding and generating temporal sequences in both sensory and motor domains, the hypothesis of TD-based sequence learning in recurrent neocortical circuits may help provide a unifying principle for studying prediction and learning at both the behavioural and the cellular levels.

Other researchers have suggested temporally asymmetric Hebbian learning as a possible mechanism for sequence learning in the hippocampus (Abbott & Blum 1996; Minai & Levy 1993) and as an explanation for the asymmetric expansion of hippocampal place fields during route learning (Mehta *et al*. 1997). Some of these models used relatively long temporal windows of synaptic plasticity, of the order of several hundreds of milliseconds (Abbott & Blum 1996), while others used temporal windows in the submillisecond range for coincidence detection (Gerstner *et al*. 1996). Sequence learning in our spike-based TD model is based on a window of plasticity that spans *ca.* ± 20 ms, which is roughly consistent with recent physiological observations (Markram *et al*. 1997; see also Abbott & Song 1999; Mehta & Wilson 2000; Roberts 1999; Song *et al*. 2000; Westerman *et al*. 1999).

(*a*) *Predictions of the spike-based TD-learning model*

The spike-timing-dependent TD-learning model makes several predictions that could potentially be tested in future experiments. First, it is known that the shape and size of back-propagating APs at different locations in a cortical dendrite depends on the distance of the dendritic location from the soma. For example, as back-propagating APs progress from the soma to the distal parts of a dendrite, they tend to become broader than more proximal parts of the dendrite. This is shown in figure 11 for a reconstructed layer-5 neocortical neuron (Douglas *et al*. 1991) from cat visual cortex with ionic channels based on those used for this neuron in the study by Mainen & Sejnowski (1996).

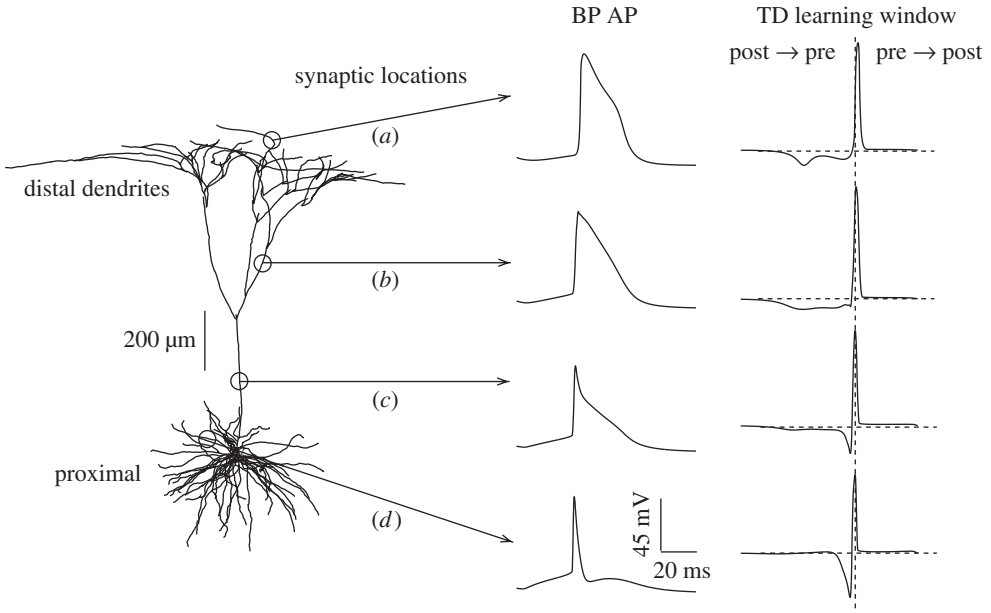

Figure 11. Dependence of learning window on synaptic location. Size and shape of a back-propagating AP at different dendritic locations in a compartmental model of a reconstructed layer-5 neocortical neuron. The corresponding TD-learning windows for putative synapses at these dendritic locations is shown on the right. Note the gradual broadening of the learning window in time as one progresses from proximal to distal synapses. (From Rao & Sejnowski (2001).)

Since synaptic plasticity in our model depends on the temporal difference in post-synaptic activity, the model predicts that synapses situated at different locations on a dendrite should exhibit different temporally asymmetric windows of plasticity. This is illustrated in figure 11 for the reconstructed model neuron. Learning windows were calculated by applying the temporal-difference operator to the back-propagating APs with $\Delta t = 2$ ms. The model predicts that the window of plasticity for distal synapses should be broader than for proximal synapses. Having broader windows would allow distal synapses to encode longer time-scale correlations between pre- and post-synaptic activity. Proximal synapses would encode correlations at shorter time-scales due to sharper learning windows. Thus, by distributing its synapses throughout its dendritic tree, a cortical neuron could in principle capture a wide range of temporal correlations between its inputs and its output. This would in turn allow a network of cortical neurons to accurately predict sequences and reduce possible ambiguities such as aliasing between learned sequences by tracking the sequence at multiple time-scales (Rao 1999; Rao & Ballard 1997).

The temporal-difference model also predicts asymmetries in size and shape between the LTP and LTD windows. For example, in figure 11, the LTD window for the two apical synapses (labelled (*a*) and (*b*)) is much broader and shallower than the corresponding LTP window. Such an asymmetry between LTP and LTD has recently been reported for synapses in rat primary somatosensory cortex (S1) (Feldman 2000). In particular, the range of time delays between pre- and postsynaptic spiking that induces LTD was found to be much longer than the range of delays that induces

LTP, generating learning windows similar to the top two windows in figure 11. One computational consequence of such a learning window is that synapses that elicit subthreshold EPSPs in a manner uncorrelated with postsynaptic spiking will, over time, become depressed. In rat primary somatosensory cortex, plucking one whisker but sparing its neighbour causes neuronal responses to the deprived whisker in layer II/III to become rapidly depressed. The asymmetry in the LTP/LTD-learning windows provides an explanation for this phenomenon: spontaneously spiking inputs from plucked whiskers are uncorrelated with postsynaptic spiking and, therefore, synapses receiving these inputs will become depressed (Feldman 2000). Such a mechanism may contribute to experience-dependent depression of responses and related changes in the receptive field properties of neurons in other cortical areas as well.

(*b*) *Future work*

The precise biophysical mechanisms underlying spike-timing-dependent TD-learning remain unclear. However, as discussed in § 4 *a*, calcium fluctuations in dendritic spines are known to be strongly dependent on the timing between pre- and postsynaptic spikes. Such calcium transients may cause, via calcium-mediated signaling cascades, asymmetric synaptic modifications that are dependent, to a first approximation, on the temporal derivative of postsynaptic activity. An interesting topic worthy of further investigation is therefore the development of more realistic implementations of TD learning based on, for instance, the temporal derivative of postsynaptic *calcium activity*, rather than the temporal difference in postsynaptic membrane potential as modelled here.

An alternate approach to analysing spike-timing-dependent learning rules is to decompose an observed asymmetric learning window into a TD component plus noise, and to analyse the noise component. However, the advantage of our approach is that one can predict the shape of plasticity windows at different dendritic locations based on an estimate of postsynaptic activity, as described in § 5 *a*. Conversely, given a particular learning window, one can use the model to explain the temporal asymmetry of the window as a function of the neuron's postsynaptic activity profile.

6. Conclusions

In this article, we have shown that two types of learning that have been considered quite different in character—classical conditioning and unsupervised learning in cortical neurons—may be reflections of the same underlying learning algorithm operating on different time-scales. The key to understanding this similarity is that they both make predictions about future states of the world. In the case of classical conditioning, the prediction is of future rewards. In the case of the cortex, the prediction is of the next location of a moving object. In both cases, temporal order is an important clue to causality. This type of learning rule was foreshadowed in Hebb's influential book *The organization of behaviour*, which contains the following statement.

> When an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.
>
> (Hebb 1949, p. 62)

Thus, Hebb had already anticipated the spike-timing-dependent version of synaptic plasticity that has recently been discovered in the cortex, and also realized that the principle of causality embodied in this learning rule could be exploited, in a manner later formalized by TD learning, to help self-organize complex systems in the brain (Sejnowski 1999).

This research was supported by the National Science Foundation, the Alfred P. Sloan foundation and the Howard Hughes Medical Institute. We are grateful to Margaret Livingstone for her collaboration, and for providing the data in figure 10*b*. We also thank Peter Dayan, David Eagleman and Christian Wehrhahn for their comments and suggestions.

References

Abbott, L. F. & Blum, K. I. 1996 Functional significance of long-term potentiation for sequence learning and prediction. *Cereb. Cortex* **6**, 406–416.

Abbott, L. F. & Song, S. 1999 Temporally asymmetric Hebbian learning, spike timing and neural response variability. In *Advances in neural information processing systems* vol. 11, pp. 69–75. Cambridge, MA: MIT Press.

Atick, J. J. & Redlich, A. N. 1992 What does the retina know about natural scenes? *Neural Comput.* **4**, 196–210.

Barlow, H. B. 1961 Possible principles underlying the transformation of sensory messages. In *Sensory communication* (ed. W. A. Rosenblith), pp. 217–234. Cambridge, MA: MIT Press.

Barlow, H. B. 1998 Cerebral predictions. *Perception* **27**, 885–888.

Bell, A. J. & Sejnowski, T. J. 1997 The 'independent components' of natural scenes are edge filters. *Vision Res.* **37**, 3327–3338.

Bell, C., Bodznick, D., Montgomery, J. & Bastian, J. 1997 The generation and subtraction of sensory expectations within cerebellum-like structures. *Brain Behav. Evol.* **50**, 17–31.

Berry, D. A. & Fristedt, B. 1985 *Bandit problems: sequential allocation of experiments*. London: Chapman and Hall.

Bi, G. Q. & Poo, M. M. 1998 Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18**, 10 464–10 472.

Bliss, T. V. & Collingridge, G. L. 1993 A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361**, 31–39.

Chance, F. S., Nelson, S. B. & Abbott, L. F. 1999 Complex cells as cortically amplified simple cells. *Nature Neurosci.* **2**, 277–282.

Cole, B. J. & Robbins, T. W. 1992 Forebrain norepinephrine: role in controlled information processing in the rat. *Neuropsychopharmacology* **7**, 129–142.

Daugman, J. G. & Downing, C. J. 1995 Demodulation, predictive coding, and spatial vision. *J. Opt. Soc. Am.* A **12**, 641–660.

Dayan, P. 2002 Matters temporal. *Trends Cogn. Sci.* **6**, 105–106.

Dayan, P. & Hinton, G. 1996 Varieties of Helmholtz machine. *Neural Netw.* **9**, 1385–1403.

Destexhe, A., Mainen, Z. & Sejnowski, T. 1998 Kinetic models of synaptic transmission. In *Methods in neuronal modeling* (ed. C. Koch & I. Segev). Cambridge, MA: MIT Press.

Dong, D. W. & Atick, J. J. 1995 Statistics of natural time-varying images. *Network Computat. Neural Syst.* **6**, 345–358.

Douglas, R. J., Martin, K. A. C. & Whitteridge, D. 1991 An intracellular analysis of the visual responses of neurons in cat visual cortex. *J. Physiol.* **440**, 659–696.

Doya, K. 2000 Reinforcement learning in continuous time and space. *Neural Comput.* **12**, 219–245.

Eckert, M. P. & Buchsbaum, G. 1993 Efficient encoding of natural time varying images in the early visual system. *Phil. Trans. R. Soc. Lond.* B **339**, 385–395.

Emptage, N. J. 1999 Calcium on the up: supralinear calcium signaling in central neurons. *Neuron* **24**, 495–497.

Feldman, D. 2000 Timing-based LTP and LTD at vertical inputs to layer II/III pyramidal cells in rat barrel cortex. *Neuron* **27**, 45–56.

Franks, K. M., Bartol, T. M., Egelman, D. M., Poo, M. M. & Sejnowski, T. J. 1999 Simulated dendritic influx of calcium ions through voltage- and ligand-gated channels using MCell. *Abstr. Soc. Neurosci.* **25**, 1989.

Gerstner, W., Kempter, R., Hemmen, J. L. & van Wagner, H. 1996 A neuronal learning rule for sub-millisecond temporal coding. *Nature* **383**, 76–81.

Ghahramani, Z. 2001 An introduction to hidden Markov models and Bayesian networks. *Int. J. Patt. Recog. Art. Intel.* **15**, 9–42.

Gould, J. L. 1987 In *Foraging behavior* (ed. A. C. Kamil, J. R. Krebs & H. R. Pulliam). New York: Plenum.

Harder, L. D. & Real, L. A. 1987 Why are bumble-bees risk averse? *Ecology* **68**, 1104–1108.

Hebb, D. O. 1949 *The organization of behaviour: a neuropsychological theory.* Wiley.

Koch, C. 1999 *Biophysics of computation: information processing in single neurons.* Oxford University Press.

Koester, H. J. & Sakmann, B. 1998 Calcium dynamics in single spines during coincident pre- and postsynaptic activity depend on relative timing of back-propagating action potentials and subthreshold excitatory postsynaptic potentials. *Proc. Natl Acad. Sci. USA* **95**, 9596–9601.

Krebs, J. R., Kacelnik, A. & Taylor, P. 1978 *Nature* **275**, 27.

Levy, W. & Steward, O. 1983 Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience* **8**, 791–797.

Linden, D. J. 1999 The return of the spike: postsynaptic action potentials and the induction of LTP and LTD. *Neuron* **22**, 661–666.

Livingstone, M. 1998 Mechanisms of direction selectivity in macaque V1. *Neuron* **20**, 509–526.

MacKay, D. M. 1956 The epistemological problem for automata. In *Automata studies* pp. 235–251. Princeton, NJ: Princeton University Press.

Magee, J. C. & Johnston, D. 1997 A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. *Science* **275**, 209–213.

Mainen, Z. & Sejnowski, T. 1996 Influence of dendritic structure on firing pattern in model neocortical neurons. *Nature* **382**, 363–366.

Markram, H., Lubke, J., Frotscher, M. & Sakmann, B. 1997 Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215.

Mehta, M. R. & Wilson, M. 2000 From hippocampus to V1: effect of LTP on spatiotemporal dynamics of receptive fields. In *Computational neuroscience, trends in research 1999* (ed. J. Bower). Amsterdam: Elsevier Press.

Mehta, M. R., Barnes, C. A. & McNaughton, B. L. 1997 Experience-dependent, asymmetric expansion of hippocampal place fields. *Proc. Natl Acad. Sci. USA* **94**, 8918–8921.

Menzel, R. & Erber, J. 1978 Learning and memory in bees. *Scient. Am.* **239**, 80–87.

Menzel, R., Erber, J. & Masuhr, J. 1974 *Experimental analysis of insect behavior*, vol. 195. Springer.

Minai, A. A. & Levy, W. 1993 Sequence learning in a single trial. In *Proc. 1993 INNS World Congr. Neural Networks II*, pp. 505–508. New Jersey: Erlbaum.

Montague, P. R. & Sejnowski, T. J. 1994 The predictive brain: temporal coincidence and temporal order in synaptic learning mechanisms. *Learn. Mem.* **1**, 1–33.

Montague, P. R., Dayan, P. & Sejnowski, T. J. 1994 Foraging in an uncertain environment using predictive Hebbian learning. In *Advances in neural information processing systems*, vol. 6, pp. 598–605. San Mateo, CA: Morgan Kaufmann.

Montague, P. R., Dayan, P., Person, C. & Sejnowski, T. J. 1995 Bee foraging in uncertain environments using predictive Hebbian learning. *Nature* **377**, 725–728.

Montague, P. R., Dayan, P. & Sejnowski, T. J. 1996 A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947.

Morrison, J. H. & Magistretti, P. J. 1983 Monoamines and peptides in cerebral cortex—contrasting principles of cortex organization. *Trends Neurosci.* **6**, 146–151.

Movshon, J. A. & Sluyters, R. C. V. 1981 Visual neural development. *A. Rev. Psychol.* **32**, 477–522.

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

Paulsen, O. & Sejnowski, T. J. 2000 Natural patterns of activity and long-term synaptic plasticity. *Curr. Opin. Neurobiol.* **10**, 172–179.

Rao, R. P. N. 1999 An optimal estimation approach to visual perception and learning. *Vision Res.* **39**, 1963–1989.

Rao, R. P. N. & Ballard, D. H. 1997 Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comput.* **9**, 721–763.

Rao, R. P. N. & Ballard, D. H. 1999 Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nature Neurosci.* **2**, 79–87.

Rao, R. P. N. & Sejnowski, T. J. 2000 Predictive sequence learning in recurrent neocortical circuits. In *Advances in neural information processing systems*, vol. 12, pp. 164–170. Cambridge, MA: MIT Press.

Rao, R. P. N. & Sejnowski, T. J. 2001 Spike-timing dependent Hebbian plasticity as temporal difference learning. *Neural Comput.* **13**, 2221–2237.

Real, L. A. 1991 Animal choice behavior and the evolution of cognitive architecture. *Science* **253**, 980–986.

Real, L. A., Ellner, S. & Harder, L. D. 1990 Short-term energy maximization and risk-aversion in bumblebees: a reply to Possingham *et al*. *Ecology* **71**, 1625–1628.

Rescorla, R. A. 1988 Behavioral studies of Pavlovian conditioning. *A. Rev. Neurosci.* **11**, 329–352.

Roberts, P. D. 1999 Computational consequences of temporally asymmetric learning rules. I. Differential Hebbian learning. *J. Comput. Neurosci.* **7**, 235–246.

Schiller, J., Schiller, Y. & Clapham, D. E. 1998 NMDA receptors amplify calcium influx into dendritic spines during associative pre- and postsynaptic activation. *Nature Neurosci.* **1**, 114–118.

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R. & Dickinson, A. 1995 Reward-related signals carried by dopamine neurons. In *Models of information processing in the basal ganglia* (ed. J. C. Houk, J. L. Davis & D. G. Beiser), pp. 233–248. Cambridge, MA: MIT Press.

Schultz, W., Dayan, P. & Montague, P. R. 1997 A neural substrate of prediction and reward. *Science* **275**, 1593–1598.

Schwartz, O. & Simoncelli, E. P. 2001 Natural signal statistics and sensory gain control. *Nature Neurosci.* **4**, 819–825.

Sejnowski, T. J. 1999 The book of Hebb. *Neuron* **24**, 773–776.

Song, S., Miller, K. D. & Abbott, L. F. 2000 Competitive Hebbian learning through spike-timing dependent synaptic plasticity. *Nature Neurosci.* **3**, 919–926.

Stiles, J. S., Bartol, T. M., Salpeter, M. M., Salpeter, E. E. & Sejnowski, T. J. 2000 Synaptic variability: new insights from reconstructions and Monte Carlo simulations with MCell. In *Synapses* (ed. M. Cowan & K. Davies). Baltimore: Johns Hopkins University Press.

Sutton, R. S. 1988 Learning to predict by the method of temporal differences. *Mach. Learn.* **3**, 9–44.

Sutton, R. S. & Barto, A. 1990 Time-derivative models of Pavlovian reinforcement. In *Learning and computational neuroscience: foundations of adaptive networks* (ed. M. Gabriel & J. W. Moore). Cambridge, MA: MIT Press.

Sutton, R. S. & Barto, A. G. 1998 *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Tesauro, G. 1989 Neurogammon wins Computer Olympiad. *Neural Comput.* **1**, 321–323.

Tootell, R. B., Switkes, E., Silverman, M. S. & Hamilton, S. L. 1988 Functional anatomy of macaque striate cortex. II. Retinotopic organization. *J. Neurosci.* **8**, 1531–1568.

Van Essen, D. 1985 Functional organization of primate visual cortex. In *Cerebral cortex* (ed. A. Peters & E. Jones), vol. 3, pp. 259–329. New York: Plenum.

Westerman, W. C., Northmore, D. P. M. & Elias, J. G. 1999 Antidromic spikes drive Hebbian learning in an artificial dendritic tree. *Analog Integr. Circuits Signal Process.* **18**, 141–152.

Wise, R. A. 1982 Neuroleptics and operant behavior: the anhedonia hypothesis. *Behav. Brain Sci.* **5**, 39–53.

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A. & Poo, M. M. 1998 A critical window for cooperation and competition among developing retinotectal synapses. *Nature* **395**, 37–44.