J Neurophysiol 98: 3648–3665, 2007.
First published October 10, 2007; doi:10.1152/jn.00364.2007.

Reinforcement Learning With Modulated Spike Timing–Dependent Synaptic Plasticity

Michael A. Farries[1] and Adrienne L. Fairhall[2]

[1]*Department of Biology, University of Texas at San Antonio, San Antonio, Texas; and* [2]*Department of Physiology and Biophysics, University of Washington, Seattle, Washington*

Submitted 2 April 2007; accepted in final form 9 October 2007

Farries MA, Fairhall AL. Reinforcement learning with modulated spike timing–dependent synaptic plasticity. *J Neurophysiol* 98: 3648–3665, 2007. First published October 10, 2007; doi:10.1152/jn.00364.2007. Spike timing–dependent synaptic plasticity (STDP) has emerged as the preferred framework linking patterns of pre- and postsynaptic activity to changes in synaptic strength. Although synaptic plasticity is widely believed to be a major component of learning, it is unclear how STDP itself could serve as a mechanism for general purpose learning. On the other hand, algorithms for reinforcement learning work on a wide variety of problems, but lack an experimentally established neural implementation. Here, we combine these paradigms in a novel model in which a modified version of STDP achieves reinforcement learning. We build this model in stages, identifying a minimal set of conditions needed to make it work. Using a performance-modulated modification of STDP in a two-layer feedforward network, we can train output neurons to generate arbitrarily selected spike trains or population responses. Furthermore, a given network can learn distinct responses to several different input patterns. We also describe in detail how this model might be implemented biologically. Thus our model offers a novel and biologically plausible implementation of reinforcement learning that is capable of training a neural population to produce a very wide range of possible mappings between synaptic input and spiking output.

INTRODUCTION

Synaptic plasticity is widely believed to be at least a component of the neurobiological changes underlying learning, but it is still far from clear exactly how the forms of synaptic plasticity studied in vitro contribute to learning and memory. An early problem was that many protocols used to induce synaptic plasticity in vitro, such as tetanic stimulation (Andersen et al. 1977), were difficult to translate into precise plasticity rules. This left the modeler with a great deal of freedom in formulating plasticity rules that were "consistent" with experimental data, leaving considerable doubt as to which rules might accurately represent processes occurring in vivo. Over the last few years, new protocols for inducing synaptic plasticity in vitro have been devised that more closely emulate processes that might occur in the intact nervous system. Spike timing–dependent plasticity (STDP) is a prominent example of such a protocol. In STDP, synaptic changes are induced by repeatedly pairing presynaptic and postsynaptic action potentials (APs) with precisely controlled timing. At glutamatergic synapses in the isocortex and hippocampus, postsynaptic APs arriving after the onset of presynaptically evoked excitatory postsynaptic potentials (EPSPs) induce long-term potentiation (LTP) of that synapse (Fig. 1*A*), whereas APs arriving before EPSPs induce long-term depression (LTD) (Bi and Poo 1998; Debanne et al. 1998; Feldman 2000; Froemke and Dan 2002; Markram et al. 1997). Although much remains to be discovered about how arbitrary activity patterns change synaptic strength, STDP can be relatively directly translated into a precise plasticity rule suitable for computer modeling.

STDP-based plasticity rules have already been used in models describing certain kinds of learning, including predictive learning (Abbott and Blum 1996; Blum and Abbott 1996; Rao and Sejnowski 2001; Roberts 1999), learning to respond to correlated inputs (Gerstner et al. 1996; Gütig et al. 2003; Song and Abbott 2001; Song et al. 2000; van Rossum et al. 2000), stabilization of postsynaptic firing rate (Kempter et al. 1999, 2001; Tegnér and Kepecs 2002), enhancement of synchronous firing (Suri and Sejnowski 2002), and coordinate transformations (Davison and Frégnac 2006). These forms of learning and self-organization, while interesting in their own right, cover only a small fraction of the kinds of adaptive changes that presumably must occur within the nervous system. More specifically, there may be occasions in which a neural population must learn semiarbitrary mappings between spatiotemporal patterns of input and evoked patterns of output. Vocal learning in songbirds is probably example of this kind of task, where a motor nucleus must translate patterned synaptic input from a premotor nucleus into activity patterns that reproduce the tutor song. In the DISCUSSION, we explain how our model could serve as a model of song learning and how it could provide a starting point for modeling basal ganglia–dependent learning in cortical networks.

The general problem outlined above is not addressed by most models of STDP-based learning, and it is not obvious how STDP as it is currently understood could be directly responsible for more general forms of learning. One flexible approach to solving this problem is reinforcement learning, where the solution space is often explored stochastically and learning is driven by a simple scalar evaluation of performance. Models of reinforcement learning typically are abstract algorithms not based on explicit neural modeling (Sutton and Barto 1998), although that is beginning to change (Izhikevich 2007; Pfister et al. 2006; Seung 2003; Xie and Seung 2004). Here we present an implementation of reinforcement learning by a biologically plausible neural network, using a simple and novel modification of the SDTP rule. In the most basic version of the approach pursued here, spiking patterns that are rela-

Address for reprint requests and other correspondence: M. A. Farries, Dept. of Biology, Univ. of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249 (E-mail: michael.farries@utsa.edu).

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

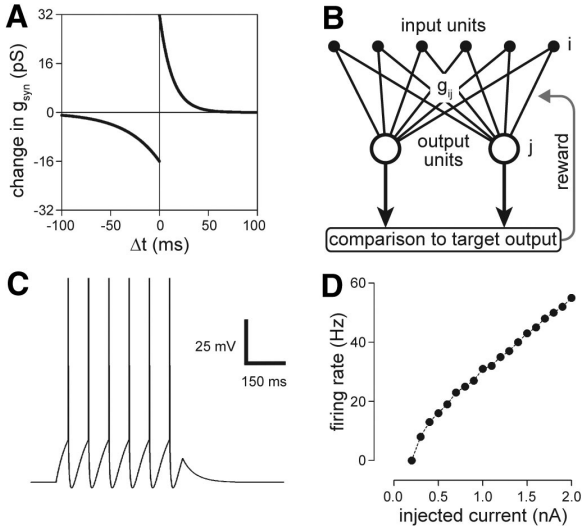

FIG. 1. *A*: spike timing–dependent synaptic plasticity (STDP) plasticity rule. Change in synaptic strength is plotted as a function of timing of postsynaptic action potential (AP) relative to the presynaptic AP: Δt = time of postsynaptic spike − time of presynaptic spike. Experimental studies usually report resulting synaptic change as a fractional change after some standard number of pairings (e.g., 60). In our model, we assume that changes induced by spike pairing at a particular Δt are absolute changes with units of conductance. *B*: structure of model. Network consists of a layer of input units projecting in an all-to-all feedforward fashion to a layer of output units. Synaptic strengths connecting input units i to output units j are represented by the matrix g_{ij}. Activity in output units is compared with some desired "target" output, and a reward signal is calculated from difference between target output and actual output. This reward signal is used to modulate synaptic plasticity. *C*: sample trace from an output layer neuron, injected with a 0.4-nA current pulse. Spikes are marked in this figure by plotting voltages exceeding spike threshold as +50 mV. *D*: firing rate of these neurons as a function of injected current.

tively similar to some "target pattern" of postsynaptic spikes are accompanied by the normal operation of the STDP rule, strengthening the synapses that contributed to the generation of that pattern, while STDP-driven synaptic changes are suppressed after spike trains that are dissimilar to the target pattern. The stochastic exploration of solution space is driven by variations in presynaptic activity. We evaluate this basic idea in a simple yet reasonably biologically plausible feedforward network and identify the factors that are needed to make it work.

METHODS

Cellular model and network architecture

We consider a two layer feedforward network (Fig. 1*B*), where the activity in the input layer is simply modeled as independent inhomogeneous Poisson processes. This input layer projects in an all-to-all pattern to an output layer of explicitly modeled neurons. We sought a model for these output neurons that is reasonably realistic yet "generic," lacking characteristic physiological properties that vary dramatically depending on cell type. We chose a single compartment conductance-based model whose membrane voltage is governed by

$$C\frac{dV}{dt} = g_R(E_{rest} - V) + g_e(E_{syn} - V) + g_{AHP}(E_{AHP} - V)$$

where g_R is the inverse of the input resistance R, g_e is the total active excitatory synaptic conductance, and g_{AHP} is the total active conductance driving the afterhyperpolarization (AHP) that follows each spike. Spike generation was not modeled explicitly; an AP occurred when the membrane voltage crossed a voltage threshold T, and each spike triggered an increment in g_{AHP} by an amount Δg_{AHP}. A spike fired by presynaptic unit i triggered an increment g_{ij} in the g_e of each postsynaptic neuron j, where g_{ij} is the strength of the synapse from input unit i to output neuron j. g_e and g_{AHP} decayed exponentially with time constants τ_{syn} and τ_{AHP}, respectively. Spike refractoriness was ensured by the AHP conductance, which kept interspike intervals above 5 ms over the entire range of synaptic strengths we considered. The cellular parameters used in our simulations were $R = 100$ MΩ, $E_{rest} = -70$ mV, $E_{syn} = 0$ mV, $E_{AHP} = -90$ mV, $\Delta g_{AHP} = 10$ nS, $\tau_{syn} = 3$ ms, $\tau_{AHP} = 10$ ms, and $T = -45$ mV, with the capacitance adjusted to yield a membrane time constant of 50 ms. This produces a model neuron with nearly linear subthreshold responses and a roughly linear spiking response to injected current and active synaptic conductance with no spike rate accommodation (Fig. 1, *C* and *D*).

For the first part of this study, networks consisted of 1,000 input units and a single output neuron. Initial synaptic strengths were chosen from a gaussian distribution with a mean of 0.32 nS and an SD of 0.05 nS. Simulations were divided into epochs ("trials") lasting 1 s. Throughout this paper, time t always denotes the time relative to the onset of the current trial, and thus $0 \leq t \leq 1{,}000$. One half of the input units were governed by homogenous Poisson processes at a rate of 5 Hz, supplying random "background" synaptic input. The remaining input units followed Poisson processes whose rate parameter varied over the course of a trial. For the first 100 ms and final 100 ms of a trial, these units remained largely silent, but for the remainder of the trial, their time-dependent rate parameter consisted of gaussian peaks that placed their spikes at particular times within a trial. The temporal precision of the spiking was controlled by the width of the peaks, set to a SD of 10 ms. Within a given simulation, these rate functions did not change across trials. Thus these units were effectively following the same "script" on each trial, with a different script for each unit. We used two methods for creating rate functions in our simulations. The first method, yielding what we call a "regular" script, generated a random 800-ms spike train (homogeneous Poisson process at 5 Hz) and placed a gaussian centered on each spike in the train. The height of each gaussian was adjusted to give one spike per peak on average, although the actual number of spikes did of course vary from trial to trial. Note that gaussians with peaks centered near the 100- or 900-ms boundaries would give these units a small chance to fire outside of those bounds. An example of a regular script for an input unit is shown in Fig. 2*A*. In the second method, each input unit was randomly assigned a single "burst time" somewhere between 100 and 900 ms into the trial (a different time for each unit), and a 10-ms-wide gaussian was placed at that time. The height of this gaussian was adjusted to yield five spikes on average. Thus the second method yields scripts (called "1-burst" scripts) that cause each input unit to fire a single high-frequency burst of spikes during each trial.

Implementation of baseline synaptic plasticity

For modeling synaptic plasticity, we used the STDP rule described by Froemke and Dan (2002), based on recordings of layer 2/3 pyramidal cells in rat visual cortex. We chose this particular implementation of STDP because of its simplicity, its examination of the effects of whole spike trains rather than just isolated spike pairs, and because it is appropriate for plasticity at corticocortical synapses that could plausibly be involved in the kind of learning that we are interested in. However, studies of STDP at other synapses in the isocortex and hippocampus have revealed substantial differences in the factors controlling the induction of synaptic plasticity. For example, induction of LTP at synapses connecting pairs of layer 5 pyramidal neurons requires higher frequency pairing (>10 Hz) in both visual (Sjöström et al. 2001) and somatosensory (Markram et al. 1997) cortices of rats, whereas Froemke and Dan (2002) could induce LTP with pairing at 0.2 Hz. In the hippocampus, induction of LTP

requires both higher frequency pairing (~5 Hz) and burst firing in the postsynaptic cell (Magee and Johnston 1997; Pike et al. 1999). Because we could not choose one STDP model that incorporates and consolidates these disparate findings, we simply selected one of them, that of Froemke and Dan (2002), with the understanding that our results may not apply to synapses where this specific formulation is not accurate. On the other hand, our model does require that STDP be modulated or gated, conditions that of course are not a part of the Froemke and Dan (2002) formulation. As we argue in the DISCUSSION, additional induction requirements, such as the need for postsynaptic burst firing, may supply the mechanism for this modulation.

The basic rule for STDP-based changes (Fig. 1*A*) is given by

$$F(\Delta t) = A_+ e^{-|\Delta t|/\tau_+} \quad \text{if} \quad \Delta t > 0; \quad F(\Delta t) = A_- e^{-|\Delta t|/\tau_-} \quad \text{if} \quad \Delta t < 0 \quad (1)$$

where Δt is the timing of the postsynaptic spike relative to the presynaptic spike. If the pre- and postsynaptic spikes are perfectly synchronous ($\Delta t = 0$), we assume that the synaptic strength does not change. To this basic formulation, Froemke and Dan (2002) added a spike suppression model, where the effectiveness of a pre- or postsynaptic spike at inducing synaptic changes is suppressed by preceding spikes. Each spike is assigned an "efficacy" $\varepsilon = 1 - e^{-t_{spike}/\tau_s}$, where t_{spike} is the time to the preceding spike. The final change in synaptic strength between units i and j induced by a single spike pair is given by $\Delta g_{ij} = \varepsilon_i^{pre}\varepsilon_j^{post}F(\Delta t_{ij})$, where ε^{pre} and ε^{post} are the separate pre- and postsynaptic efficacies, governed by distinct time constants τ_s^{pre} and τ_j^{post}. Froemke and Dan (2002) fixed the values of these parameters by fitting this model to their data. We adopt most of those values for our model, and under their "additive model," those values are (rounded to the nearest millisecond, well within the SD for all measurements) $\tau_+ = 13$ ms, $\tau_- = 35$ ms, $\tau_i^{pre} = 28$ ms, and $\tau_i^{post} = 88$ ms. We did not directly adopt their values for A_+ and A_-, because they are expressed as the *percentage* change in synaptic strength after 60–80 pairings. It is not clear if synaptic changes generally scale in that way (where stronger synapses experience larger absolute changes in conductance), and because most models of STDP to date have expressed the changes in absolute terms, we continue that practice. Thus our A_+ and A_- have units of conductance, rather than dimensionless fractional changes as in Froemke and Dan (2002). The values we selected for A_+ and A_- were 32 and −16 pS, respectively.

Although we do not implement synaptic changes as a percentage of current strength, we do consider the possibility that the size of changes scales in some way with the strength of the synapse. We impose maximum and minimum strengths on the synapses, g_{max} and g_{min}, equal to 10 times (3.2 nS) and 1/10th (0.032 nS) of the average initial synaptic strength, respectively. Under our "additive" model, the g_{ij} are simply clipped to their maximum/minimum value if the application of STDP would push them outside of that range. However, some STDP modeling studies use a rescaling in which the size of Δg_{ij} is reduced as g_{ij} approaches its limits (Gütig et al. 2003; Rubin et al. 2001; van Rossum et al. 2000). This kind of rescaling is sometimes called a "multiplicative rule" (Gütig et al. 2003; Rubin et al. 2001), although this does not correspond to the additive/multiplicative terminology of Froemke and Dan (2002), and the multiplicative rule given here differs from the one described in Kepecs et al. (2002). We study the behavior of our model under both the additive rule described above and a multiplicative rule that rescales potentiating changes by the factor $(g_{max} - g_{ij})/(g_{max} - g_{min})$ and depressing changes by the factor $(g_{ij} - g_{min})/(g_{max} - g_{min})$, which is simply a generalization of the method of Rubin et al. (2001) to cases in which $g_{max} \neq 1$ and $g_{min} \neq 0$.

In most of our simulations, we incorporated activity-dependent scaling of synaptic strength, modeling the phenomenon reported in pyramidal neurons cultured from rat visual cortex (Turrigiano et al. 1998). Our model of activity-dependent scaling was based on the approach of van Rossum et al. (2000), with some modifications. Postsynaptic activity in each output neuron j was tracked using a variable $a_j(t)$ obeying the equation

$$\tau_a \frac{da_j}{dt} = -a_j + \sum_k \delta(t - t_{kj})$$

where t_{ij} are the spike times in neuron j. van Rossum et al. (2000) made changes in synaptic strength proportional to the difference between $a_j(t)$ and some equilibrium activity level a_{goal}, so that any deviation from this particular activity level triggered synaptic scaling. Rather than insist on the maintenance of a single activity level, we allowed neurons to remain within a range of activity levels without triggering synaptic scaling. If activity in output neuron j wandered outside of that range, synapses were altered as follows

$$\Delta g_{ij} = \beta g_{ij}(a_{min} - a_j), \quad \text{if} \quad a_j < a_{min};$$

$$\Delta g_{ij} = \beta g_{ij}(a_{max} - a_j), \quad \text{if} \quad a_j > a_{max} \; \forall i$$

where a_{min} and a_{max} are the minimum and maximum equilibrium activity levels, respectively, and β is a parameter controlling the rate of activity-dependent scaling. The Δg_{ij} were calculated at the end of each trial. Because we were not trying to maintain activity at one specific level a_{goal}, we did not need to use the "integral controller" correction used by van Rossum et al. (2000). In a few simulations, we implemented activity-dependent changes in intrinsic excitability instead of synaptic scaling. We modeled that process by driving changes in the AP threshold of a given output neuron by the activity level as follows

$$\Delta T = \beta(a - a_{min}), \quad \text{if} \quad a < a_{min}; \quad \Delta T = \beta(a - a_{max}), \quad \text{if} \quad a > a_{max}$$

In all simulations that used activity-dependent scaling, we used the following parameter values: $\tau_a = 10$ s, $a_{max} = 100$ (because $\langle a \rangle = \tau_a \times$ rate, this sets the maximum firing rate to roughly 10 Hz), and $\beta = 10^{-3}$ (synaptic scaling) or $\beta = 10^{-2}$ mV (excitability changes). In most simulations, $a_{min} = 9.5$ (a minimum firing rate of just under 1 Hz). However, for simulations in which output neurons were trained to produce specific spike trains containing more than one spike (i.e., the results shown in Fig. 5), $a_{min} = 9.5 \times N_{spikes}$, where N_{spikes} is the number of spikes in the target spike train.

Reinforcement learning through modulation of synaptic plasticity

Initially, reinforcement learning was implemented by choosing "target" spike trains for each output unit (representing the goal of the training), calculating the difference between those target spike trains and the network's actual output, and transforming that difference into a reward signal that modulated synaptic plasticity. The difference $\Delta_j(t)$ between the actual and target spike trains for neuron j as a function of time t in the current trial was determined by convolving the spike trains (represented by a temporal series of 1s where spikes occur and 0s otherwise) with a gaussian of unit height and SD σ (typically 10 ms) and subtracting one of the smoothed spike trains from the other. The reward signal $Rwd(\Delta_j)$ is

$$Rwd = \langle e^{-\alpha|\Delta_j(t)|} \rangle \quad (2)$$

which maps $|\Delta| \in [0, \infty)$ into the interval (0, 1], with $Rwd(0) = 1$. The angled brackets denote an average over all output neurons j. We used $\alpha = 3$ for all simulations. One should note that if the interspike intervals in the actual output and target output are substantially greater than the smoothing parameter σ, as they were for most of our simulations, the maximum value Δ can take is ~1, and thus the minimum possible reward is $e^{-\alpha}$. If one wanted the minimum reward to be 0, one could redefine the reward as $Rwd = \langle e^{-\alpha|\Delta j(t)|} \rangle - e^{-\alpha}$, but because learning performance is not qualitatively improved by this

definition, we did not adopt it for the simulations presented in this paper. Initially, reward-dependent modulation of STDP was implemented by setting the change in synaptic strength to the product of the reward signal and the change that would be produced by unmodulated STDP. Hence, for synaptic changes triggered by a postsynaptic spike in output unit j occurring at time t, $\Delta g_{ij} = Rwd(t)\varepsilon_i^{pre}\varepsilon_j^{post}F(\Delta t_{ij})$.

However, in most simulations we implemented an adaptation of the temporal difference algorithm for reinforcement learning where adaptive changes are driven by the difference δ_R between the reward received and the reward expected (Sutton and Barto 1998). In the most general implementation of this algorithm, the system's task is to adopt a policy that leads it to choose actions that will maximize its total future reward given the current state of its environment, $s(n)$, where n is the trial number. It uses a "value function" $V[s(n)]$ to estimate the future reward given the current environment $s(n)$: $V[s(n)] = E[Rwd(n) + \gamma Rwd(n+1) + \gamma^2 Rwd(n+2) + \cdots]$, where $E[Rwd(n)]$ is the expected reward resulting from the action triggered by the current state $s(n)$ under the system's current policy, and γ is a "discount factor" ($0 \leq \gamma \leq 1$) that assigns smaller weights to expected rewards further in the future. The "temporal difference error" used to improve the current policy is the sum of the actual reward and the updated expected future reward resulting from the chosen action minus the total future reward expected before that action was taken: $\delta_R = Rwd(n) + \gamma V[s(n+1)] - V[s(n)]$ (Sutton and Barto 1998).

Translating our model into the language of the temporal difference algorithm, the environmental state $s(n)$ is the input pattern presented on trial n, the "policy" is determined by the synaptic strengths, and the action chosen is the set of output spike trains. Our model constitutes a special case in which future states s are independent of the action chosen, so the only reward prediction possible is the average reward given the network's current "policy." Thus $V[s(n)] = \langle Rwd\rangle + \gamma\langle Rwd\rangle + \gamma^2\langle Rwd\rangle + \cdots$ and $\delta_R = Rwd(n) + \gamma\langle Rwd\rangle\Sigma_{m=0}^{\infty}\gamma^m - \langle Rwd\Sigma_{m=0}^{\infty}\gamma^m = Rwd\,(n) - \langle Rwd\rangle$. In our model, both the "environmental state" s (spike trains provided by the input units) and the "action chosen" (spike trains generated by the output units) are functions of time t in the trial. Thus the reward, average reward, and temporal difference error are all functions of time in the trial: $\delta_R(t) = Rwd(t) - \langle Rwd(t)\rangle$. Ideally, $\langle Rwd(t)\rangle$ would be the reward received under a fixed "policy" (fixed synaptic strengths), averaged over many trials. Because the synaptic strengths change on every trial, this ideal is unobtainable and $\langle Rwd(t)\rangle$ is instead a running average of the reward recently received. At the end of each trial, after $\delta_R(t)$ has been calculated, $\langle Rwd(t)\rangle$ is updated as follows

$$\langle Rwd(t)\rangle_{new} = \frac{9}{10}\langle Rwd(t)\rangle_{old} + \frac{1}{10}Rwd(t)$$

It is important to keep in mind that this averaging is conducted over trial number n, not over trial time t, and hence the "average reward" is still a function of time in trial.

To use the temporal difference error to drive learning in our model, we simply multiply the synaptic changes of the unmodulated STDP rule by $\delta_R(t)$ instead of $Rwd(t)$

$$\Delta g_{ij} = \delta_R(t)\varepsilon_i^{pre}\varepsilon_j^{post}F(\Delta t_{ij}), \quad \text{where} \quad \delta_R(t) = Rwd(t) - \langle Rwd(t)\rangle \quad (3)$$

Because $\delta_R(t)$ can be negative, this learning rule permits anti-Hebbian synaptic plasticity, where pre–post pairings induce LTD and post–pre pairings yield LTP. It is not difficult to envision circumstances under which formerly LTP-triggering patterns of activity are made to induce LTD instead (see RESULTS), but we feel that the conversion of LTD into LTP is less plausible. For this reason, *Eq. 3* is applied with the following exception: if $\delta_R < 0$ and $F(\Delta t_{ij}) < 0$, $\Delta g_{ij} = 0$.

We quantified model performance using a modified version of the reward signal. To obtain a performance metric that did not depend on the number of spikes in the target spike train, we normalized the difference $\Delta_j(t)$ between the target spike train and the actual spike train by the number of spikes in the target train, N_j^{spikes}, thus replacing $\Delta_j(t)$ in *Eq. 2* with $\Delta_j^*(t) \equiv \Delta_j(t)/N_j^{spikes}$. To obtain a single number characterizing the performance of the network over a trial, we averaged this modified reward measure (denoted *Rwd**) over the time in trial. Unfortunately, this results in a performance measure that is restricted to a relatively narrow range of values—performance in random networks is already ~0.65, and networks that do not fire at all get an average modified reward of 0.88–0.92, depending on the target pattern. We therefore scaled the performance measure to range between 0 and 1: *Rwd**: *Performance* = $(\langle Rwd^*\rangle - 0.6) \times 2.5$. This performance measure was used only to quantify the success at learning target patterns; it was never used to modulate plasticity or drive the learning process.

After exploring the capabilities of networks containing a single output neuron, we considered networks with multiple output neurons. At first, multineuron reinforcement learning was implemented as described above: each output neuron was assigned a distinct target spike train to reproduce, but all output units received the same reinforcement signal, which was simply derived from an average of the individual neuron rewards. As shown in RESULTS, this was not particularly successful for networks containing more than three or four output neurons. In subsequent multineuron training, the target output activity no longer took the form of distinct spike trains assigned to specific output neurons; instead, the target activity was expressed as the fraction of output neurons that were to fire at different times in the trial. For example, if the target pattern specifies that 25% of the output neurons be active at a particular time, the network's performance is evaluated without regard to which output neurons are firing; only the number active is relevant. To implement this idea, we define $O_j(t)$ as the spike train generated by output neuron j convolved with a gaussian waveform of $\sigma = 10$ ms, and let $G(t)$ denote the goal of learning, the "target pattern" that specifies what fraction of the output neural population should be active as a function of time in trial (naturally, $G(t) \in [0,1]\ \forall t$). The difference between the actual output and the target output is given by $\Delta(t) = \langle O_j(t)\rangle - G(t)$, where the angled brackets denote an average over the output neurons j. Then the reward is again $Rwd(t) = e^{-\alpha|\Delta(t)|}$, and the reinforcement signal used to modulate synaptic plasticity is once again $\delta_R(t) = Rwd(t) - \langle Rwd(t)\rangle$.

This procedure for comparing the output of a neural population to a desired population response $G(t)$ and computing the reinforcement signal $\delta_R(t)$ is fairly straightforward, but it introduces a new complication. The magnitude of $\delta_R(t)$ depends on the magnitude of fluctuations in $Rwd(t)$ across trials, and that in turn depends on the magnitude of fluctuations in $\langle O_j(t)\rangle$. As the number of output neurons increases, the variability in individual output neurons remains the same, and hence the variability of $\langle O_j(t)\rangle$ across trials should decrease as more neurons are included in the average. That will cause $\delta_R(t)$ to grow smaller in networks with more output neurons, and because the amplitude of Δg_{ij} is directly proportional to $\delta_R(t)$, synaptic plasticity will be suppressed in larger networks. The obvious solution is to add a factor to *Eq. 2* that compensates for the shrinkage in $\delta_R(t)$ caused by increasing the number of output neurons N. If the $O_j(t)$ varied independently from trial to trial, an approximate solution to this problem would be obtained by multiplying the Δg_{ij} calculated from *Eq. 2* by the factor $\sqrt{N}$. However, the $O_j(t)$ generally do not vary independently, because variation in $O_j(t)$ is driven by variation in input activity, and all output neurons are driven by the same input units. The precise degree to which trial-to-trial fluctuations in $O_j(t)$ are correlated depends on the synaptic matrix g_{ij}, which makes the appropriate choice of "correction factor" rather complicated. Preliminary simulations indicated that the $\sqrt{N}$ factor overcompensates for diminished Δg_{ij} in networks containing >25 output neurons for most g_{ij} attained over the course of training, and that, on average, $|\Delta g_{ij}|$ was roughly one fifth of the mean magnitude occurring in one-neuron networks for any $N \geq 25$. In view of these results, we

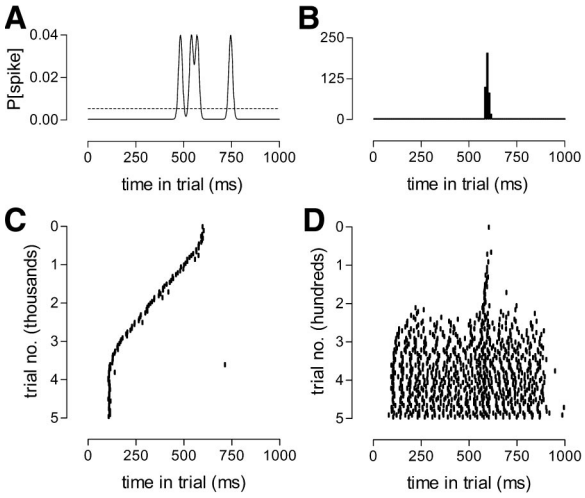

FIG. 2. *A*: solid line, probability of spiking in a 1-ms time bin for 1 of the input units used in simulations depicted in this figure. Functions like this were generated for each temporally patterned input unit by randomly placing gaussians with SD of 10 ms to give an average firing rate of 5 Hz (averaged over all units; individual units could have mean firing rates above or below this value) for an 800-ms period starting 100 ms after trial onset. Dashed line shows uniform probability of spiking for "background" input units, corresponding to an average firing rate of 5 Hz. *B*: peristimulus time histogram (PSTH) of output neuron for network in its initial state, where synaptic weights have been adjusted to make output neuron fire ~600 ms into each trial. This PSTH was generated from 500 repetitions of the stimulus, with synaptic plasticity turned off. Bin size is 10 ms. *C*: raster showing how response of output neuron changes under the influence of STDP. Over many trials, the neuron fires earlier and earlier, until it reaches a limit determined by onset of temporally patterned synaptic input. This simulation uses "additive" implementation of STDP, i.e., there is no rescaling of synaptic changes based on current synaptic strength. This raster only shows activity of every 50th trial. *D*: raster showing how response of the output neuron changes under influence of "multiplicative" STDP, where synaptic changes are rescaled as synaptic strength approaches its upper and lower bounds. This raster only shows activity of every 5th trial. Initial network state is the same as for *C*.

adopted the effective, if inelegant, solution of multiplying the result of *Eq. 3* by five in these simulations

$$\Delta g_{ij} = 5\delta_R(t)\varepsilon_i^{pre}\varepsilon_j^{post}F(\Delta t_{ij}) \quad (4)$$

This overcompensates somewhat in networks containing 10 neurons, but still worked well for all values of N considered here. Networks trained to produce specific spike trains used *Eq. 3* (Figs. 3–5, containing up to five output neurons), whereas networks trained to produce a population response $G(t)$ used *Eq. 4* (Figs. 6–8, containing 10–400 output neurons).

In testing the ability of our model to learn to generate population responses, we considered three kinds of target patterns. The first was simply a broad gaussian centered on the middle of the trial (t = 500 ms) with a peak height of 0.1 and σ = 100 ms. This population response yields an average firing rate among output neurons of 1 Hz. The second type (bursty) consisted of four brief populations bursts (each described by a gaussian of σ = 10 ms) placed randomly in the central 800 ms of the trial, but with a minimum interval of 50 ms between bursts to ensure that the bursts remained distinct. The burst heights were drawn from a normal distribution of mean 5 and variance 1, and the resulting $G(t)$ was normalized to yield an average firing rate of 1 Hz across output neurons. The third type of target pattern was designed to assess our model's ability to learn an "arbitrary" waveform $G(t)$. These "random" target patterns were produced in three stages (Fig. 7*B*, *left*). First, we generated 1,000 ms of zero-mean noise with a gaussian amplitude distribution of unit variance and a correlation time of 100 ms. Second, this noisy waveform was converted into a "probability of spiking." All negative portions of the waveform were set to zero, and regions approaching the bounds of temporally patterned synaptic input (at 100 and 900 ms) were rapidly—but not instantly—forced to zero by multiplication with a sigmoidal envelope $E(t)$: $E(t) = 1 + e^{125\text{ ms} - t})^{-1}(1 + e^{t - 875\text{ ms}})^{-1}$. The result was normalized to give a probability distribution for "potential output spike times." In the third and final stage, N spike times (where N is the number of output neurons) were drawn from this probability distribution, and $G(t)$ became the sum of N gaussians, each of height $\frac{1}{N}$, σ = 10 ms, and centered on the randomly selected spike times. Like the other two types of target pattern, these $G(t)$ correspond to an average firing rate of 1 Hz among the output neurons. Because all $G(t)$ considered in this study specified an average rate of 1 Hz, we could directly adopt as our performance measure in these simulations the value of $Rwd(t)$ averaged over time in trial, without correcting for the number of spikes expected in the output.

All modeling was executed using MATLAB (MathWorks, Natick, MA). All statistical tests were conducted with Prism (GraphPad Software, San Diego, CA).

RESULTS

The synaptic input that neurons receive sometimes takes the form of temporally patterned activity in which presynaptic neurons fire spikes at specific times or vary their firing rate over time in a characteristic way, relative to a sensory stimulus or motor act. This patterned input can be completely stereotyped, as in the songbird vocal system where a premotor nucleus provides a highly stereotypical pattern of activity to a telencephalic vocal motor nucleus every time the bird sings (Hahnloser et al. 2002). Alternatively, the specific pattern of input can vary depending on the qualities of a sensory stimulus but be stereotyped for a given stimulus, so that properties of the stimulus are encoded in the temporal pattern of the input activity (de Ruyter van Steveninck and Bialek 1988; Reinagel and Reid 2000). Neurons receiving such input must reliably generate an appropriate response. We model this situation by having a postsynaptic neuron receive episodic input from presynaptic units whose activity is governed by Poisson processes with rates that can vary over the course of an episode or "trial." During each trial, a subset of the input units fire only around certain times (but at different times for each unit) that remain the same from trial to trial; an example of such a unit is shown in Fig. 2*A*. For the remaining presynaptic units, dubbed "background" units, the probability of spiking is uniform throughout a trial (Fig. 2*A*, dashed line).

Unmodulated STDP destabilizes established mappings between spatiotemporal patterns of input and output activity

Our goal is to explore the possibility of using a modulated version of STDP to train a postsynaptic neuron to produce a desired spike train in response to a specific spatiotemporal pattern of input activity. To help motivate this, we first show the effect of the continuous, unmodulated application of STDP on a neuron that already generates a specific response to its patterned input. Figure 2 shows an example of a neuron that fires a single spike ~600 ms into each trial, receiving input from 500 units that fire in a stereotypical pattern (Fig. 2*A* shows the probability of spiking for 1 such unit over the course of a trial) and 500 background units that fire randomly at an average rate of 5 Hz (uniform spike probability). The response pattern exhibited by the postsynaptic neuron (Fig. 2*B*) was

created by making the synapses of presynaptic units active around 600 ms much stronger than all other synapses.

The operation of the normal STDP rule causes synapses active before 600 ms to grow stronger, so that the postsynaptic neuron eventually begins to fire shortly before the 600-ms mark. That, in turn, causes the depression of the synapses originally responsible for making the neuron fire at 600 ms and the potentiation of other synapses that were active earlier in the trial. In this way, the postsynaptic response occurs earlier and earlier, until it approaches the onset of the temporally patterned activity, 100 ms after the start of the trial (Fig. 2*C*). This phenomenon is well known in the STDP modeling literature and is typically presented as a boon: it is "predictive learning," whereby a neuron learns to respond to synaptic inputs that provide the earliest reliable prediction of its original response (Abbott and Blum 1996; Blum and Abbott 1996; Rao and Sejnowski 2001; Roberts 1999). However, there will inevitably be cases in which such "predictive learning" is not appropriate, and it seems likely that such cases could occur in cortical areas where STDP operates. It seems that some modulation of STDP is necessary simply to maintain stable mappings from presynaptic activity to postsynaptic response.

The simulation described above and shown in Fig. 2*C* assumes that changes in synaptic strength are made "additively," i.e., the magnitude of the change is independent of synaptic strength. This results in a strongly bimodal distribution of synaptic strengths (data not shown) that is characteristic of the additive implementation of STDP (Gütig et al. 2003; Kepecs et al. 2002; Rubin et al. 2001; Song et al. 2000; van Rossum et al. 2000). Some modeling studies of STDP assume that synaptic changes depend on current synaptic strength, with the magnitude of the changes biased toward potentiation or depression as synaptic strength approaches its lower or upper bounds, respectively (Gütig et al. 2003; Rubin et al. 2001). This is sometimes called "multiplicative" STDP (Gütig et al. 2003; Rubin et al. 2001), although that term is also applied to cases in which the magnitudes of both LTP and LTD increase with synaptic strength (Kepecs et al. 2002). Here, we adopt the terminology of Rubin et al. (2001) and Gütig et al. (2003). There is some experimental evidence for this phenomenon, at least for depressing changes in cultured hippocampal neurons (Bi and Poo 1998), and unlike additive STDP, it yields a unimodal distribution of synaptic strengths that can resemble the distribution of quantal amplitudes measured experimentally (Gütig et al. 2003; Rubin et al. 2001; van Rossum et al. 2000). Testing our model with multiplicative STDP, we found that there was still bias toward firing earlier as the simulation proceeded, but response changes were dominated by a large increase in firing rate (Fig. 2*D*). Multiplicative STDP causes an overall increase in synaptic strength, because the initial strengths of most synapses were relatively close to the lower bound. Although the details differ, the continuous application of either additive or multiplicative STDP inevitably destroys any specific patterned response to temporally patterned presynaptic activity.

Simplest implementation of STDP-driven reinforcement learning is only partially successful

We begin with an extremely simple implementation of STDP-driven reinforcement learning. The spike trains generated by the output neurons are compared with some desired "target" output, and from the difference, a reward signal is computed. We calculated the difference $\Delta(t)$ between the target output and the actual output by subtracting smoothed versions of their respective spike trains, generated by convolving the spike trains with a gaussian of an SD of σ (10 ms, in most cases). In choosing a specific form for the reward signal, we required that it depend only on the absolute difference between the target output and the actual output, i.e., it could not convey any "instructive" information about the kinds of changes needed such as whether the probability of firing at a particular time should be raised or lowered. We also wanted the reward function to map differences $\Delta(t)$ onto the interval (0, 1], with $\Delta = 0$ generating a reward signal $Rwd = 1$ and $Rwd \rightarrow 0$ as Δ increases. For networks containing a single output neuron, we chose to define the reward signal as $Rwd(t) = e^{-\alpha|\Delta(t)|}$.

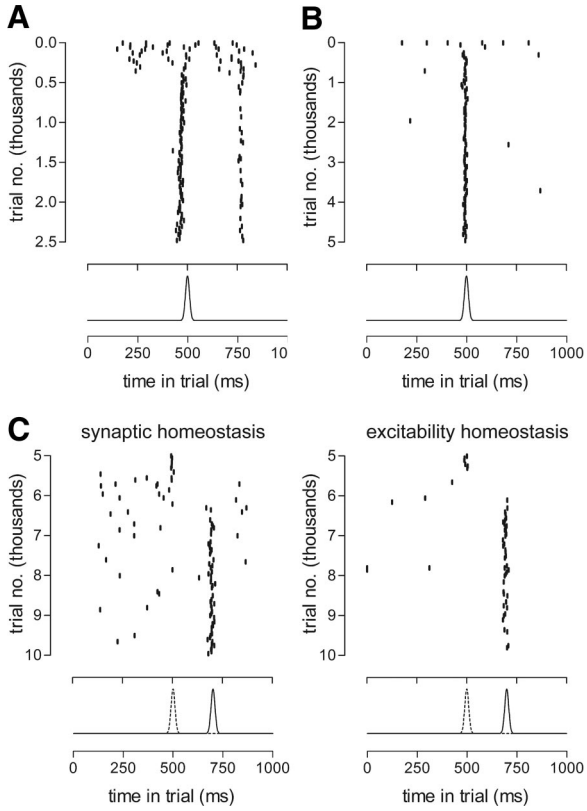

FIG. 3. *A*: *top*: raster showing how the response of an output neuron changes during training. Raster only shows activity of every 25th trial. *Bottom*: target output spike train used in this simulation, convolved with a gaussian with SD of 10 ms. *B*: *top*: raster showing learning when activity-dependent scaling and anti-Hebbian plasticity are included. Initial synaptic strengths, rate functions for input units, and target output are all the same as in *A*. Raster shows every 50th trial. *Bottom*: target output spike train used in this simulation. *C*: after 5,000 trials of training shown in *B*, target output is switched to a single spike at 700 ms. *Top*: output over 5,000 trials with this new target. *Bottom*: original target output (dashed line) and new target (solid line), smoothed with $\sigma = 10$ ms. *D*: as in *C*, but with activity-dependent changes in excitability rather than synaptic scaling. All simulations shown in this figure use additive STDP.

The reward signal was used to modulate synaptic plasticity simply by multiplying the synaptic changes triggered by a postsynaptic spike at time t according to the standard STDP rule by the value of the reward signal at time t. Thus STDP-driven changes are largest during times when the actual output matches the target output, and grow smaller as the difference between them increases. This modulation of STDP could be implemented biologically, for example, by modulation of

N-methyl-D-aspartate (NMDA)-type glutamate receptors (NMDARs); many neuromodulators are known to affect NMDARs (Köles et al. 2001; MacDonald et al. 1998). One should note that this implementation of STDP-driven reinforcement learning requires that the appropriate modulatory signal be present at the same time the output spike train is being generated. That in turn implies that the system providing the modulatory signal must somehow predict how closely the spike train will match the target output before they can be compared directly. This is an onerous task, but it is not impossible. Because variations in the output are driven by variations in the input activity, a modulatory system that monitored activity in the input layer could in principle use that information to predict how well the resulting output will match the target, although such a system would have to constantly adapt as synaptic plasticity changes the mapping between input activity and output activity. Arguments about the plausibility of such a system are reserved for the Discussion.

An example of the performance of this kind of model is shown in Fig. 3*A*, using the "additive" form of STDP. With the network in its initial state, the output neuron fired at a mean rate of 5.13 Hz (averaged over the entire 1-s trial), firing at fairly regular intervals starting shortly after the onset of temporally patterned input. The target spike train was a single spike fired 500 ms into the trial (Fig. 3*A*, *bottom*). Under training, the output neuron came to reliably fire an AP shortly before the 500-ms mark and stopped firing during most other times (Fig. 3*A*, *top*). However, it also fired consistently ~800 ms after trial onset.

This example highlights a fundamental problem with the model in its current form: it has no mechanism to remove "unwanted" spikes, e.g., the second spike fired in many trials shown in Fig. 3*A*. This spike arose because the network in its initial state had a high probability of firing at that time (800 ms), enough to potentiate synapses active just before that time even with minimal reward (see METHODS). If the network already reliably fires a spike at a certain time, there is no guarantee that it will cease to do so under training, even if the reward signal at that time is strictly zero. There is another problem that is not shown in Fig. 3*A*, but which is readily apparent. If the network begins in a state in which it never fires a spike at a particular time, it can never learn to fire at that time no matter how large the "reward"—in STDP, no plastic changes occur in the absence of postsynaptic spikes.

Inclusion of activity-dependent synaptic scaling and anti-Hebbian STDP enables accurate reinforcement learning

There are forms of synaptic plasticity that do not depend on correlations between pre- and postsynaptic activity, such as the activity-dependent scaling of synaptic strength reported in isocortical neurons (Turrigiano et al. 1998). This form of plasticity causes synaptic strength, as measured by the amplitude distribution of miniature excitatory postsynaptic currents, to increase if postsynaptic activity is suppressed by application of tetrodotoxin. Activity-dependent synaptic scaling or other homeostatic mechanisms for maintaining postsynaptic activity could solve one of the problems our current model faces—the inability to learn to fire at times when the starting network never fires. We incorporated activity-dependent scaling of synaptic strength into our model to test this hypothesis. The other major problem with our current model, difficulty in removing unwanted spikes in the output train, might be solved by allowing some form of "anti-Hebbian" STDP to occur under certain conditions. Examples of anti-Hebbian STDP, in which postsynaptic APs following EPSPs induce LTD rather than LTP, has been reported in a cerebellum-like structure in the electric fish (Bell et al. 1997) and at some synapses in the mouse dorsal cochlear nucleus (Tzounopoulos et al. 2004).

If anti-Hebbian STDP is to be used in our model, we must carefully consider how to apply it in a way that supports reinforcement learning. Guidance on this question can be found in the literature on an important algorithm for reinforcement learning known as temporal difference learning (Sutton and Barto 1998). In temporal difference learning, adaptive changes are not directly driven by the reward; rather, they are driven by the difference between the future expected reward at one trial and the actual reward (plus an updated future expected reward) received on the next trial. In our model, this difference, denoted $\delta_R(t)$, is the difference between reward as a function of time in trial (the actual reward) and the average reward received over the last few trials (the expected reward): $\delta_R(t) = Rwd(t) - \langle Rwd(t) \rangle$. This temporal difference signal is used to modulate synaptic plasticity by multiplying the synaptic changes triggered by a postsynaptic spike at time t according to the standard STDP rule by $\delta_R(t)$. Whenever $\delta_R(t) < 0$, i.e., whenever network performance is worse than its average performance over the last few trials, synaptic plasticity is anti-Hebbian, with one wrinkle noted below.

Although one of the first experimental studies of STDP observed anti-Hebbian STDP in the brain stem (Bell et al. 1997), until recently, cortical STDP studies uniformly reported Hebbian plasticity. This raises the question of whether it is plausible enough to be included in a model aiming at a moderate degree of biological realism—could anti-Hebbian STDP be implemented by forms of neuromodulation that have already been observed in the isocortex? The fact that both LTP and LTD are triggered by increases in postsynaptic $[Ca^{2+}]$ suggests that it could. Because LTD is triggered by small increases in $[Ca^{2+}]$, whereas LTP appears with larger Ca^{2+} transients (Cho et al. 2001; Cormier et al. 2001; Ismailov et al. 2004; Yang et al. 1999), reducing the amount of Ca^{2+} influx resulting from pairing EPSPs with postsynaptic APs could make normally potentiating patterns of activity induce LTD instead. Indeed, partial blockade of NMDARs does exactly that (Cummings et al. 1996; Froemke et al. 2005; Nishiyama et al. 2000), indicating that simple modulation of NMDARs, already proposed above, could suffice to implement our STDP-based version of temporal difference learning. However, such a simple mechanism could not support fully anti-Hebbian plasticity—although formerly LTP-inducing pairings would yield LTD, it is hard to see how formerly LTD-inducing pairings could cause potentiation. Furthermore, most recent studies that have found anti-Hebbian STDP in the isocortex report mainly pre–post LTD (Sjöström and Häusser 2006), although some post–pre LTP has been reported at distal synapses, attributed to the delay between the first postsynaptic spike and the time of maximal dendritic depolarization (Letzkus et al. 2006). Consequently, we do not incorporate fully anti-Hebbian plasticity into our model; if $\delta_R(t) < 0$, pairings that would normally trigger synaptic depression do not change synaptic strength at all.

Figure 3*B* shows the performance of our model with activity-dependent synaptic scaling and anti-Hebbian STDP included, where the initial synaptic weights, patterns of input activity, and target output are the same as in Fig. 3*A*. As training progresses, the network now learns to produce the target output, using either additive (Fig. 3*B*) or multiplicative (data not shown) STDP, with quite accurate performance after fewer than 1,000 trials of training. Figure 3*C* shows how activity-dependent synaptic scaling permits the output neuron to learn to generate spikes at times when it originally never fired. The simulation starts after 5,000 trials of training to fire at 500 ms (the point reached at the end of the raster in Fig. 3*B*), but now the target output is switched to a single spike at 700 ms (Fig. 3*C*, *bottom*). Anti-Hebbian STDP causes the neuron to stop firing at 500 ms, at which point synaptic scaling increases overall synaptic strength until new spikes appear, including spikes near 700 ms (Fig. 3*C*, *left*). In addition to synaptic scaling, cortical neurons can also adjust their intrinsic excitability in response to lasting changes in activity level (Desai et al. 1999). We wanted to know if alternative forms of activity homeostasis like this could substitute for synaptic scaling in our model. We modeled excitability homeostasis by having the AP threshold adapt if postsynaptic activity levels remained too low or too high. We found that excitability homeostasis could substitute for synaptic scaling (Fig. 3*C*, *right*); our model requires some form of activity homeostasis, but is not especially sensitive to the specific form it takes. However, excitability homeostasis would ultimately have to be supplemented by some other process, because the AP threshold drops every time the target output is changed and no circumstances normally arise to bring it back up. For this reason, we used synaptic scaling for all subsequent simulations.

Model performance is sensitive to the width of gaussians used to smooth spike trains

As explained above, the reward signal that drives learning is calculated from the difference between a target spike train and the actual spike train generated by the output neuron. To compute that difference, the spike trains are convolved with a Gaussian whose width specifies the temporal precision demanded of the model. Thus far, we have used a gaussian with a SD (σ) of 10 ms. This choice is arbitrary; therefore we examined model performance under different values of σ. One might expect that if σ is too small (if the level of temporal precision demanded is too high), the model will be unable to learn to produce the target spike train. That is indeed the case: performance is substantially degraded if σ is just 5 ms, as shown in Fig. 4*A*. On the other hand, one might expect that model performance would improve—or at least remain unchanged—with larger σ. That is not the case. The "predictive" aspect of the STDP rule shown in Fig. 2 manifests itself as σ increases: rather than firing spikes near the peak of the gaussian, at 500 ms, output neurons learn to fire earlier in the trial with larger σ. If σ is large enough, the network will sometimes come to fire two spikes, neither of which occurs at the target time of 500 ms (Fig. 4*B*). The final average reward achieved after training, our chosen measure of overall performance, is plotted in Fig. 4*C* for different values of σ (see Supplemental Fig. 1 to see how this performance measure is related to the activity generated by the network)[1]. To maintain a consistent measure of model performance, the final rewards plotted in Fig. 4*C* were computed using a 10-ms gaussian, although the reward signal used to drive learning was computed using the specified σ (5–100 ms). Figure 4*D* shows PSTHs showing the output activity of the model after training with different values of σ.

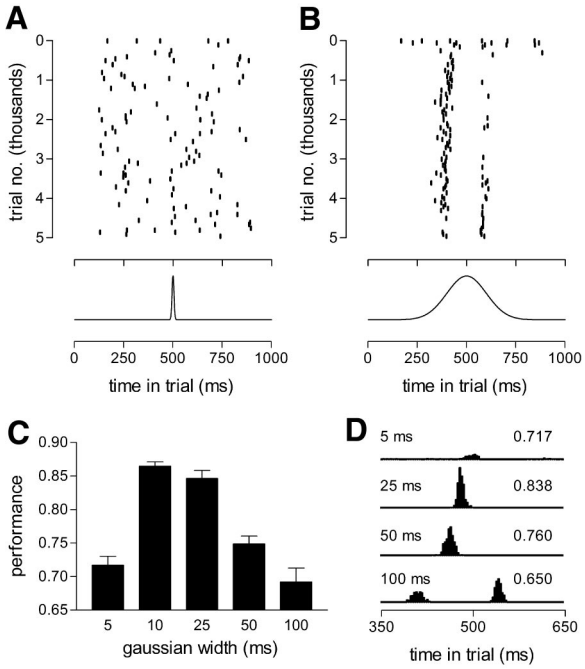

FIG. 4. Effect of spike train smoothing on model performance. *A*: training with target and output spike trains smoothed by a gaussian of $\sigma = 5$ ms. *Top*: raster of output spike trains generated during training; every 50th trial is shown. *Bottom*: smoothed target spike train. *B*: same as *A*, but with spike trains smoothed by a gaussian of $\sigma = 100$ ms. *C*: average final performance after 5,000 trials of training with gaussians of different widths (σ). In all cases, the target spike train consisted of a single spike fired 500 ms into the trial. Bar height represents the average performance of 10 repetitions (using a different initial set of synaptic weights on every repetition) at each σ; error bars denote SD of final performance among the 10 repetitions. "Average final performance" is defined as average reward achieved over the last 500 trials of training. For this purpose, *Rwd* is calculated from spike trains smoothed with $\sigma = 10$ ms to provide a consistent measure of performance. *D*: PSTHs showing examples of performance after training with different σ. Each PSTH was generated by simulating example networks for 500 trials with synaptic plasticity turned off. PSTHs show only the middle 300 ms of trials; there was little or no activity outside of this range after training. Numbers on *left* indicate smoothing σ used; average performance (*Rwd* calculated using $\sigma = 10$ ms) is shown on *right*. Bin size is 3 ms.

Learning arbitrary spike trains

Thus far, we only examined the ability of the model to learn to generate a "spike train" consisting of a single spike. The model can learn to produce more arbitrary spike trains, but not as consistently. Figure 5 shows two examples of the model trained on target trains containing three spikes: one successfully (Fig. 5*A*) and one not (Fig. 5*B*). Figure 5*C* shows the final performance of networks taught to produce randomly generated spike trains containing no more than five APs; each bar represents the average final performance of 10 networks (each with a different target spike train). Average performance de-

[1] The online version of this article contains supplemental data.

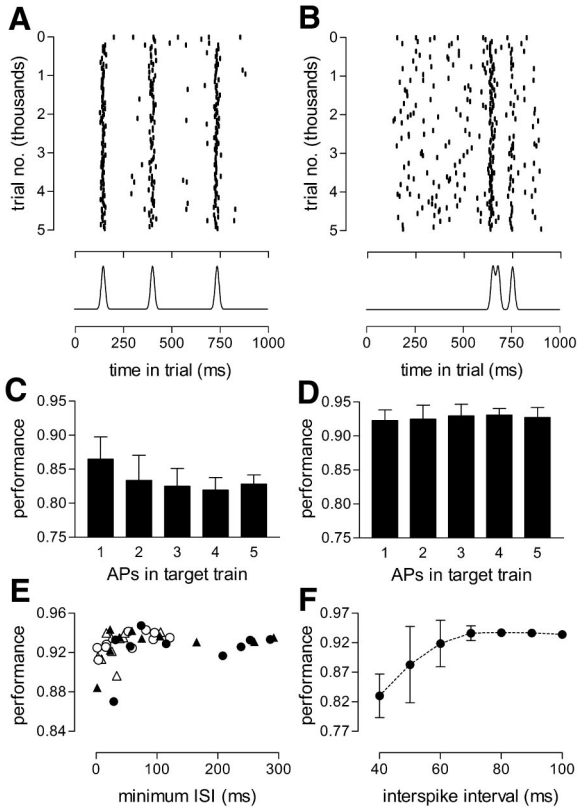

FIG. 5. Learning arbitrary spike trains under additive STDP. *A*: raster showing a successful example of learning a target train containing 3 spikes. Smoothed version of target spike train is show *below*. *B*: example of a failed attempt to learn a 3-spike target train. *C*: average final performance in learning arbitrary spike trains containing different numbers of spikes. As before, this is the average reward obtained during the last 500 trials of training, normalized for spike number (see METHODS). Each column shows mean and SD over 10 repetitions, where each repetition uses a different starting network and different target spike train. *D*: as in *C*, but with presynaptic units firing just 1 high-frequency burst during each trial ("1-burst script") rather than at multiple times during a trial as in all previous simulations ("regular script", example shown in Fig. 2*A*). *E*: final performance during training with 1-burst scripts as a function of the smallest interspike interval occurring in the target spike train. Symbols indicate the total number of spikes in the target train as follows: filled circles, 2 APs; filled triangles, 3 APs; open circles, 4 APs; open triangles, 5 APs. *F*: performance after training to produce spike pairs with interspike intervals varied systematically between 40 and 100 ms in 10-ms increments. Each point plots average of 10 repetitions (each a different starting network and spike pair, but with the same interspike interval); error bars indicate SD.

clines as the number of spikes in the target train increases—the means are significantly different (ANOVA, $P = 0.004$) and there is a trend toward lower performance with more spikes ($P = 0.002$, slope $= -0.0035$, $r^2 = 0.17$). We suggest two possible explanations for this trend. First, spike trains with more APs are more likely to contain short interspike intervals (ISIs), and these could pose a problem because *1*) the AHP makes it more difficult to reach spike threshold again shortly after spiking and *2*) synapses active shortly after the first AP that could help trigger a second AP will be subjected to LTD caused by the depressing portion of the STDP rule. Furthermore, target trains with more spikes may be more difficult to learn as synaptic adjustments that drive spiking at one time may interfere with the model's ability to remain silent at other times. This is because presynaptic units fire at several distinct times throughout the trial, and thus strengthening the synapse of a presynaptic unit because it is active at one time may also increase synaptic drive during times when the output neuron should not fire. As the number of spikes in the target train increases, one might expect this problem to grow less manageable. This problem would be circumvented if each presynaptic unit fired only a single burst during a trial. Indeed, if we use networks receiving this kind of input, performance is improved for all target spike trains ($P < 0.0001$ for all 5 groups, Mann-Whitney test), with spike trains containing more spikes showing the largest improvement (Fig. 5*D*). With this pattern of presynaptic input, there are no longer any significant differences in average performance among target trains with different numbers of spikes (ANOVA, $P = 0.78$). This suggests that short ISIs may not pose any serious difficulty for this model, but when the data in Fig. 5*D* are plotted as a function of the minimum ISI occurring in the target spike train (Fig. 5*E*), one sees that the worst performance occurs with target trains containing shorter ISIs. We systematically explored the effect of ISI on performance using two-spike target trains (Fig. 5*F*) and found that average performance on shorter ISIs (<60 ms) was significantly worse than on longer ISIs (80–100 ms; ANOVA followed by Tukey's multiple comparison test, using $P \leq 0.05$ as the criterion for significance). Results using multiplicative STDP were similar, except that networks trained under multiplicative STDP systematically fired output spikes a few milliseconds earlier in the trial than those trained with additive STDP (Supplemental Fig. 2).

Learning in networks with multiple output neurons

We showed that our reward-modulated version of STDP is capable of training a single output neuron to produce an arbitrary spike train in response to temporally patterned synaptic input and that performance is best when *1*) the input units fire only one burst per trial, *2*) the target spike train does not contain ISIs shorter than 80 ms, and *3*) the additive implementation of STDP is used. However, realistic learning tasks will entail training a population of output neurons to produce some target pattern of activity. This target pattern might specify distinct target spike trains for each output neuron, in a direct extension of our single-neuron model. This task would be trivial if each output neuron received its own individually tailored reinforcement signal, but it proves to be quite difficult if one global reinforcement signal is broadcast to all output neurons, calculated from the average of the rewards that each output neuron would have been assigned had they were being trained individually (Supplemental Fig. 3).

Although the model fails to accurately learn target spike trains with as few as five neurons in the output layer, demanding that every output neuron in the network learn to produce a specific spike train is probably unreasonable and unrealistic. For most realistic tasks, the necessary pattern of output activity can probably be realized by many different sets of specific spike trains generated in the output population. To model this situation, we defined the target output as a time-varying function specifying the fraction of output neurons that should be active over the course of a trial, regardless of which specific output neurons are active at any time. For example, the task being learned could require that output neurons gradually become more active over a trial, peaking in the middle of the trial and declining as the trial concludes. Such a situation is

shown in Fig. 6, where the target pattern of activity is a 100-ms-wide gaussian centered at 500 ms into the trial, with a peak value of 0.1. Network output is represented by smoothing the individual spike trains with a 10-ms-wide gaussian and averaging over all output neurons, yielding a single waveform that gives the fraction of cells active over time in trial. The reinforcement signal is then calculated in a manner directly analogous to the single neuron case: we calculate $Rwd(t)$ by subtracting the "fraction active" waveform from the target pattern, taking the absolute value, and exponentiating the result, which is used to calculate the temporal difference reinforcement signal, $\delta_R(t)$.

In the example shown in Fig. 6, the output layer contains 100 neurons, input units fire only one burst per trial, and additive STDP is used. After 5,000 trials of training, an aggregate PSTH generated by adding together the PSTHs of all output neurons (Fig. 6*A*, *top*, collected over 500 trials) reveals that the output neurons collectively generate a reasonable copy of the target pattern (Fig. 6*A*, *middle*). The ability to learn target patterns like this requires that output neurons be trained together as a population. This is shown by examining the aggregate behavior of an ensemble of 100 networks, each containing one output neuron (Fig. 6*A*, *bottom*). Each of these networks was separately trained on the broad gaussian target pattern, but they could not reproduce this pattern individually (also shown in Fig. 4) or collectively (Fig. 6*A*, *bottom*).

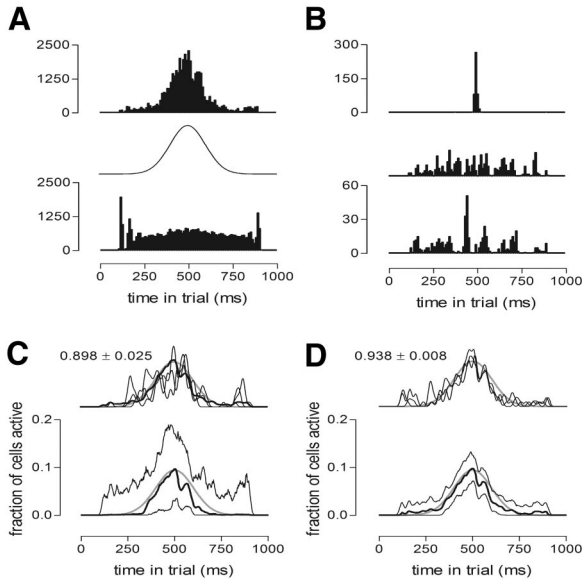

FIG. 6. Learning a broad gaussian population response in a network with multiple output neurons. *A*: *top*: "aggregate PSTH" showing output of this network after training, generated by summing the PSTHs of all the individual output neurons, collected from 500 trials. *Middle*: target pattern of activity used to train this network. *Bottom*: aggregate PSTH showing output of 100 neurons separately trained on the template shown above, again collected from 500 trials of activity after 5,000 trials of training. *B*: example posttraining PSTHs of individual neurons taken from network depicted in *A* (*top*), showing the 3 basic kinds of responses observed. *Top*: example of a neuron that fires at roughly the same time on every trial. The majority of output neurons conform to this pattern. *Middle*: example of a neuron that is capable of firing at most times within a trial. *Bottom*: example of a neuron that can fire at many different times in a trial, but with a propensity to fire around a particular time. Scale of the *middle* PSTH is the same as the *bottom* one. Bin size for all PSTHs in *B* and *C* is 10 ms. *C*: graphs of the fraction of output neurons active over the course of a trial, calculated by smoothing the spike trains with a 10-ms-wide gaussian and averaging over all output neurons. *Top*: 3 examples of activity on individual trials (thin black lines) plotted with mean activity over 500 trials (thick black line) and target activity (thick gray line). *Bottom*: 95% CI for activity over a trial (2.5 and 97.5% percentiles calculated from 500 trials; thin black lines), median activity (thick black line), and target activity (thick gray line). Scale in both graphs is the same. Numbers in *top right corner* give performance (mean ± SD) over the 500 trials used to generate this panel. *D*: same as *C*, but with uncorrelated inputs driving output neurons.

Although the activity of output neurons in this example collectively approximates the target pattern, the individual neurons within the population do not. The majority of output neurons consistently fire at or near a particular, neuron-specific time on every trial after training ("temporally specific" neurons; example shown in Fig. 6*B*, *top*). Other output neurons can fire at almost any time in a trial (Fig. 6*B*, *middle*), whereas a third group combine these two response patterns (Fig. 6*B*, *bottom*) or tends to fire at two or more discrete times during a trial. If we define the "temporal specificity" of a neuron's response pattern as the percentage of spikes it fires within 15 ms of its most probable firing time, we find that there is a bimodal distribution of temporal specificity among output neurons trained on this target pattern (Supplemental Fig. 4B, *top*). If we classify those neurons firing >60% of their spikes within 15 ms of their most probable firing time as "temporally specific," we find that 64% of output neurons can be so designated. The distribution of times at which these temporally specific neurons are most likely to fire reproduces the central part of the gaussian pattern that the network as a whole generates (Supplemental Fig. 4*B*, *bottom*).

The top PSTH in Fig. 6*A* shows how well the output of this network averaged over 500 trials reproduces the target pattern. However, it does not tell us whether the network reproduces this pattern on individual trials; it is possible that the output is highly variable and that the gaussian shape of Fig. 6*A* emerges only after summing the results of many trials, which would not constitute a very successful example of learning. To show the activity on individual trials, we instead plot the smoothed spike trains averaged over all output neurons, i.e., the "fraction active" waveform that is used to compare output activity to target activity. The *top graph* of Fig. 6*C* plots three examples of output activity on individual trials (thin black lines) along with the mean "fraction active" waveform (thick black line) and the target pattern (thick gray line). Although the average activity pattern does closely resemble the target pattern, the activity on individual trials varies considerably from trial to trial. The *bottom graph* of Fig. 6*C* shows the range of activity exhibited on individual trials by plotting the 95% CI bounds (thin black lines) for these waveforms; the thick black line is the median activity, which differs somewhat from the mean activity.

Some trial-to-trial variability in output activity is driven by variations in input activity, and indeed the model's ability to learn depends on this variability. We might also expect this variability to decrease as the number of output neurons increases, because variations in the activity of individual output neurons would make smaller fractional contributions to the overall activity pattern and would tend to average out. However, this is not true under the conditions pertaining to this model (see METHODS). This is because the activity in the output neurons does not vary independently; the all-to-all connectivity pattern between the input and output layers causes correlations in these variations in output neuron activity (to a degree that depends on the synaptic matrix g_{ij}). Although the all-to-all connectivity pattern was a practical choice for modeling pur-

poses, it probably does not reflect a connection pattern common in the vertebrate CNS. Even if, for example, one cortical area projects strongly to another, it is unlikely that all individual neurons in the recipient region receive input from exactly the same set of presynaptic neurons. We tested whether the high trial-to-trial variability of Fig. 6*C* is caused by correlated fluctuations in synaptic drive received by output neurons by "decorrelating" the synaptic input. We generated a separate set of presynaptic spikes for each output unit, but where each set is drawn from the same probability distribution (script). Thus the output neurons receive the same average synaptic input as before, but the trial-to-trial fluctuations in presynaptic activity are now independent across output neurons. When the network shown in Fig. 6*C* is driven by such "uncorrelated" input, the trial-to-trial variability in output activity is greatly reduced and the output more closely matches the target activity on individual trials (Fig. 6*D*).

We now consider the model's capabilities for learning a wider range of target patterns. We begin with a class of target patterns that is quite distinct from the single broad gaussian used thus far: a series of large but brief population bursts. These "bursty" patterns consist of four randomly placed 10-ms-wide gaussians whose height specifies the fraction of neurons that should participate in that burst. Figure 7*A* shows an example of a network trained on such a target pattern, with the *two leftmost graphs* directly paralleling Fig. 6*C*—the *top left graph* plots activity from individual trials (thin black lines) and the mean activity (thick black line), whereas the *bottom left graph* shows the 95% CI for output activity (thin black lines) and the median activity (thick black line); the target pattern is shown on all graphs as a thick gray line. This bursty target pattern is reproduced with considerably greater fidelity than the broad hump of Fig. 6, so much so that it is difficult to distinguish the individual lines on the *leftmost graphs* of Fig. 7*A*. The *middle graphs* of Fig. 7*A* zoom in on the two central bursts, showing how both the height and timing of the bursts are fairly well matched to the target pattern, with comparatively little trial-to-trial variation even with normal "correlated" inputs. If the output neurons are driven by "uncorrelated" inputs as in Fig. 6*D*, the output variability drops further (Fig. 7*A*, *right*).

Having considered target patterns with both widely and narrowly temporally distributed patterns, we now examine our model's ability to reproduce "random" target patterns, shown in the *leftmost column* of Fig. 7*B*. From a 1,000-ms waveform drawn from a gaussian noise distribution with a correlation time of 100 ms (Fig. 7*B*, *top left*), we generate a "probability of spiking" by clipping the portions of the waveform that are negative or that approach the limits of temporally patterned input, 100 and 900 ms into the trial, and normalize the result (Fig. 7*B*, *middle left*). We use this probability waveform to randomly select locations for *N* 10-ms-wide gaussians, each of

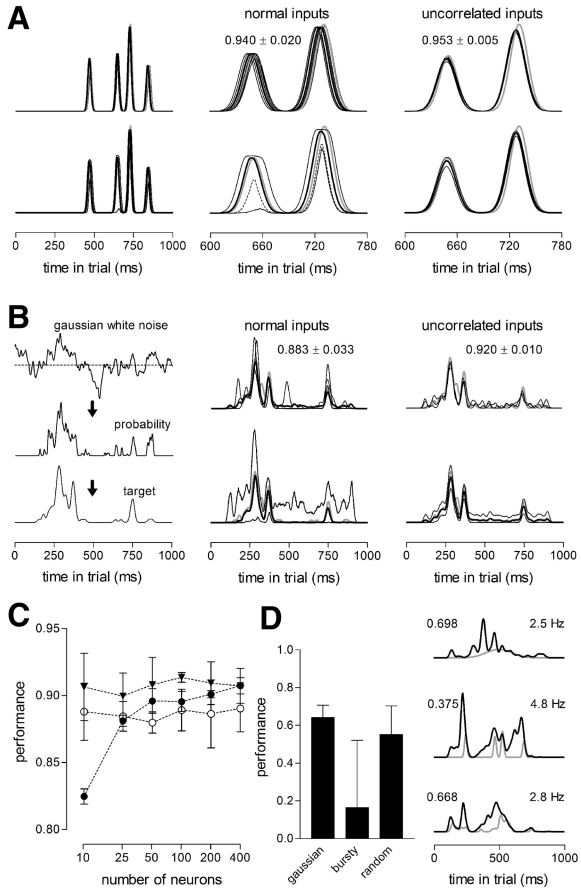

FIG. 7. Learning arbitrary population responses. *A*: example of a network of 100 output neurons trained to produce a series of 4 "population bursts" of activity of varying intensity. *Top left*: 10 examples of activity on individual trials (thin black lines) plotted with mean activity over 500 trials (thick black line) and target activity (thick gray line). *Bottom left*: 95% CI for activity over a trial (thin black lines), median activity (thick black line), and target activity (thick gray line). *Middle*: central portion of the *leftmost graphs* on an expanded time scale, with the 5% activity percentile (thin dashed line) added to the *bottom graph* to show that failures to produce burst at 650 ms, while possible, are rare. *Right*: as in the *middle graphs*, but with network driven by "uncorrelated" inputs and 5% activity percentile omitted. Vertical scale in all graphs in this panel is the same. *B*: example of a network of 100 output neurons trained to produce a "random" pattern of activity. *Leftmost column* of graphs shows how these random target patterns are generated. A 1-s segment of gaussian white noise is created (*top*; dashed line marks zero level) and converted into a probability distribution (*middle*) by setting any negative parts of the waveform to 0, as well as the 1st and last 100 ms of the waveform, and normalizing. This probability distribution is used to select the peak times of *N* 10-ms-wide gaussian bumps that are added together to give a "random" target pattern for a network containing *N* output neurons (*bottom*). *Middle graphs* show how well a network can learn to generate this sample target pattern after 5,000 trials of training; *top graph* shows 3 examples of activity on individual trials (thin black lines) plotted with mean activity over 500 trials (thick black line) and target activity (thick gray line). *Bottom graphs* show 95% CI for activity over a trial (thin black lines), median activity (thick black line), and target activity (thick gray line). *Right*, as in the *center* graphs, but with the network driven by "uncorrelated" inputs. Vertical scales in *center* and *right graphs* are the same. Performance (mean ± SD) attained over the 500 trials used to generate *A* and *B* are shown on relevant graphs. *C*: final performance as a function of number of neurons in output layer for the 3 types of target pattern considered here: filled circles, 100-ms-wide gaussian of peak height 0.1 centered at 500 ms; filled triangles, "bursty" patterns of 4 10-ms-wide gaussian peaks of varying heights; open circles, "random" patterns generated as shown in *B* (*left*). Each symbol represents average final performance over 5 different networks, each with a different target pattern (except for the filled circles, where target pattern is always the same); error bars indicate SD. *D*: performance in 100-neuron networks using multiplicative STDP. *Left*: average final performance over 5 simulations for each of the 3 types of target pattern; error bars indicate SD. *Right*: examples of simulations for each target type. Black lines show mean activity over 50 trials; gray lines plot target patterns. Numbers on *right* of each graph show final performance for each network; average firing rate of neurons in output layer is shown to the *left*. In all cases, target patterns represent an average firing rate of 1 Hz among output neurons.

height $\frac{1}{N}$, where N is the number neurons in the output layer of the network; the sum of these N gaussians gives us the target pattern of activity (Fig. 7*B*, *bottom left*). This last part of the procedure guarantees that the target pattern can actually be generated by N output neurons whose spike trains are smoothed with 10-ms-wide gaussians. Our model can learn to reproduce such patterns on average, but with a substantial amount of trial-to-trial variability (Fig. 7*B*, *middle*). With "uncorrelated" inputs, variability is decreased and network activity on individual trials now more closely resembles the target pattern (Fig. 7*B*, *right*).

We assessed our model's performance on these three types of target pattern (100-ms-wide gaussian, bursty, random) as the number of neurons in the output layer was varied (Fig. 7*C*). A two-way ANOVA showed that both neuron number and pattern type contribute to the variation in final performance and that these two factors interact ($P < 0.0001$ in all cases). The results of Fig. 7*C* suggest that the dependence of performance on neuron number, and its interaction with pattern type, is caused entirely by the fact that networks with fewer output neurons do a relatively poor job of reproducing the broad gaussian target pattern. This is confirmed by rerunning a two-way ANOVA with the 100-ms-wide gaussian data omitted; now only pattern type ($P < 0.0001$) and not neuron number ($P = 0.86$) contributes to performance differences, with no interaction between the two factors ($P = 0.96$). Unlike the bursty and random pattern types, the broad gaussian cannot be precisely mimicked by any network; it can only be approximated, and the best achievable approximation improves as the number of output neurons increases.

Thus far, networks trained to reproduce a "population response" have used only additive STDP, with input units governed only by "one-burst" scripts, i.e., input units that each fire just one short burst of spikes on each trial. When networks using multiplicative STDP were tested on this task, we found that they consistently failed to reproduce any of the target pattern types we considered (Fig. 7*D*). The multiplicative rule's bias toward LTP was probably a factor, because these networks always overshot the target activity pattern (examples shown in Fig. 7*D*, *right*). Networks using "regular scripts," where a given input unit could fire at multiple distinct times within a trial, were also unable to learn most types of population reponses (Supplemental Fig. 5).

In the simulations described above, we trained each network on just one target pattern, as might be the case in, for example, song learning in birds whose repertoire is limited to one song. However, more generally a neural population may learn to produce different responses to different patterns of synaptic input or to generate a continuous mapping between input and output. Although we will not attempt a full exploration of our model's capacity for learning multiple input–output pairings, we did establish that this model can learn to produce at least eight distinct output patterns in response to distinct input patterns (Supplemetal Fig. 6).

Model performance with simplified versions of the STDP rule

The implementation of the STDP rule used in our model is fairly complicated, incorporating not only the relative timing of pre- and postsynaptic spikes, but also a dependence on the firing history of the presynaptic and postsynaptic neurons. We chose this implementation not because of its value for reinforcement learning but because experimental studies suggest that these additional factors influence the synaptic changes induced by STDP protocols (Froemke and Dan 2002; Froemke et al. 2006; Wang et al. 2005; Wittenberg and Wang 2006). Our results showed that this particular form of spike history dependence is not fatal to our model. However, the history dependence used, taken from Froemke and Dan (2002), is not unique; Froemke and Dan themselves published a modified version of this rule (Froemke et al. 2006) for the same kind of synapses. Furthermore, different kinds of synapses could show distinct forms of history dependence. We did not attempt to investigate our model's performance under all reasonable forms of history dependence. Rather, we simply sought to determine whether our model's success depends on the specific form used here. Thus we examined the performance of our model in the absence of any history dependence.

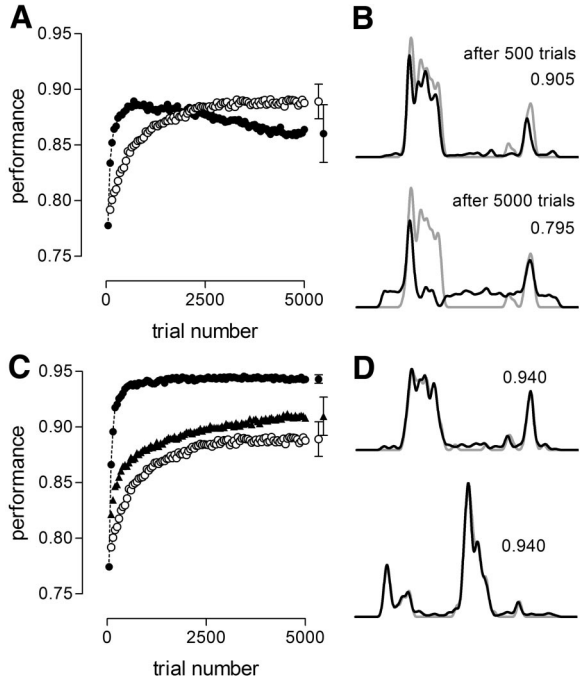

FIG. 8. Learning with a simplified version of the STDP rule. *A*: learning curves showing performance over time averaged over 5 simulations for each curve, with each simulation using a different network and target pattern. Filled circles, simulations lacking spike history dependence (spike suppression mechanism); open circles, control simulations using the full (history-dependent) plasticity rule. Immediately to the *left* of these curves, and plotted on the same vertical scale, are average final performances for each case with error bars showing SD across simulations. These symbols have been nudged horizontally to avoid error bar overlap. *B*: activity generated by an example network trained without spike suppression at 2 stages of training. *Top graph* shows mean network output (black line) over 50 trials using the network as it was after 500 trials of training; *bottom graph* shows mean activity after 5,000 trials. For both graphs, 50 trials used to generate them were run with synaptic plasticity turned off. Gray line shows target pattern, and numbers give mean performance over 50 trials shown. *C*: as in *A*, but filled circles now denote networks lacking both spike suppression and post-before-pre long-term depression (LTD), and filled triangles denote networks with spike suppression but without post-before-pre LTD. Open circles are control networks, the same data as plotted in *A*. *D*: 2 examples of final output generated by networks trained with a plasticity rule lacking both spike suppression and post-before-pre LTD. Black lines are mean output over 50 trials, and gray lines are target patterns in each case. Numbers give mean performance over 50 trials.

The spike history dependence of the STDP rule of Froemke and Dan (2002) is derived from their "spike suppression model," where the efficacy of a spike at inducing synaptic changes is suppressed by the occurrence of preceding spikes in the same neuron. To remove spike history dependence from the STDP rule, we omit the "spike efficacy" factors ε_i^{pre} and ε_j^{post} from the rule (*Eq. 4*). We tested this simplified STDP rule in five networks containing 100 output neurons and trained on "random" target patterns. As shown in Fig. 8*A*, these networks (filled circles) initially learn faster than control networks that include spike suppression (open circles), but this performance peaks after ~500 trials and begins a slow decline, ending after 5,000 trials at a performance level that is lower on average than the control networks. Although this difference is not quite significant ($P = 0.056$, Mann-Whitney test), it is an alarming trend for our model's success. The example shown in Fig. 8*B* illustrates the proximate cause of this steady decline in performance. After 500 trials, this network, lacking the spike suppression mechanism, generates an output pattern that is as good a copy of the target pattern (Fig. 8*B*, *top*) as the full model could generate after 10 times as much training. As training proceeds, however, the network fails to maintain the sustained elevation of activity appearing in the first half of the target pattern; the activity plateau generated by the network gradually shortens and by the end of training is reduced to a brief population burst at the onset of the early activity plateau demanded by the target pattern (Fig. 8*B*, *bottom*). Because this loss of activity would cause many of the output neurons to fire at average rates significantly <1 Hz, the homeostatic plasticity mechanism is engaged and instigates a compensatory increase in baseline firing rate (Fig. 8*B*, *bottom*). The process shown here is repeated in the other simulations using the STDP rule without spike suppression—sustained bouts of activity specified in the target patterns are gradually shortened, accompanied by an increase in baseline firing.

This effect is caused by the fact that our performance-modulated version of the STDP rule is biased toward LTD relative to the traditional unmodulated form of the rule because its anti-Hebbian regimen (applied whenever $\delta_R < 0$) includes only LTD, whereas both LTP and LTD are possible when $\delta_R > 0$. This issue becomes pertinent when the network generates a sustained bout of activity. Most of the spikes in presynaptic bursts that are responsible for maintaining this activity occur before the postsynaptic spikes they trigger, but when the presynaptic activity patterns consist of high-frequency bursts (as in the 1-burst scripts used here), one or two of the later spikes in the burst can occasionally occur after the postsynaptic spike triggered by earlier spikes. If $\delta_R > 0$ at this time on a given trial, these final presynaptic spikes will induce LTD at the relevant synapse. On such trials, this LTD is more than counterbalanced by LTP induced by the majority of spikes in the burst occurring before the postsynaptic spike, but on trials in which $\delta_R < 0$ (i.e., performance is worse than the recent average performance), the presynaptic spikes occurring before the postsynaptic spike trigger LTD without any compensatory LTP induced by the few presynaptic spikes that may appear after the postsynaptic spike. Thus once the network has reached a plateau in performance when δ_R is just as likely to be negative as positive, the changes induced at these synapses averaged over several trials will be slightly depressing, gradually eroding the sustained bout of activity that the network is supposed to generate. The inclusion of the spike suppression mechanism avoids this by suppressing the contributions of later spikes in presynaptic bursts, the only spikes that can trigger LTD when $\delta_R > 0$, since the average ISI in these bursts (6.3 ms) is considerably shorter than the recovery time constant for presynaptic spike suppression (28 ms). With the spike suppression mechanism in place, a plateau in performance ($\langle\delta_R\rangle = 0$) now produces an approximate balance between LTP and LTD.

If the main advantage of spike suppression is to counter the depressing portion of the basic STDP rule, networks lacking both spike suppression and post-before-pre-LTD should perform at least as well as control networks incorporating both features. We tested this by running simulations in which the "spike efficacy" factors ε_i^{pre} and ε_j^{post} are omitted, as above, and the parameter governing the size of post-before-pre LTD (A_-, *Eq. 1*) is set to zero. Now learning is quite rapid (Fig. 8*C*, filled circles), and performance achieves an asymptotic level well above the values attained by control networks ($P = 0.008$, Mann-Whitney test; 2 examples shown in Fig. 8*D*). If spike suppression is used while A_- (Fig. 8*C*, filled triangles), performance is significantly worse ($P = 0.008$, Mann-Whitney test) and is not significantly different from control performance ($P = 0.22$). In summary, one aspect of the spike history dependence of the STDP rule of Froemke and Dan (2002), presynaptic spike suppression, does in fact assist reinforcement learning under the conditions prevailing here (postsynaptic spike suppression is rarely engaged because the average firing rate among output neurons is roughly 1 Hz). However, it does so by mending an imbalance between LTP and LTD caused by combining anti-Hebbian plasticity with a conjunction of post-before-pre LTD and burst firing. The solution, to make only the first spike in a high-frequency presynaptic burst "count" in the induction of synaptic plasticity, is not specific to this particular form of history dependence. In the absence of post-before-pre LTD, this history dependence actually slightly impedes performance. In that sense, the particular form of history dependence used is not integral to the success of our model.

DISCUSSION

Our model attempts to find a biologically plausible solution to a fairly general learning problem—how to train a neural population to generate arbitrary responses to patterned synaptic input—that could be applicable to a wide range of specific neural systems and functional tasks. In this endeavor, it largely succeeds. Our approach can, with some moderate restrictions, teach a neuron to convert temporally patterned synaptic input into an arbitrarily selected spike train. Although it cannot reliably get any but the smallest neural populations to produce distinct spike trains specifically assigned to each neuron, it can train a neural population to generate global response patterns, where neurons spontaneously adopt distinct firing patterns that collectively produce the target population response. Furthermore, these population responses, once trained, are evoked only by input patterns very similar to the ones presented during training, and networks can learn multiple input pattern-population response pairs. These successes were achieved simply by taking an "off the shelf" plasticity rule derived directly from experimental studies and subjecting it to a simple and plausible form of modulation.

These accomplishments do come with a list of requirements and restrictions. First, a form of activity-regulating homeostasis is needed to guarantee the presence of postsynaptic spikes, because STDP alone can do nothing without activity in both the presynaptic and postsynaptic cells. This can be achieved through the inclusion of known physiological processes: homeostatic regulation of either intrinsic excitability or, the choice we favored here, synaptic strength. Another relatively minor requirement is the exclusion of the form of multiplicative STDP examined here. Although this form of strength-dependent synaptic modification is a staple of the STDP modeling literature, it has relatively little experimental support, and it poses a dilemma between two unrealistic alternatives: either synaptic changes must be strongly biased toward LTP ($g_{ij} \ll g_{max}$) or the maximum achievable strength can be only about twice the starting strength ($g_{ij} \approx \frac{1}{2}\, g_{max}$). In addition, our model requires anti-Hebbian synaptic plasticity to ensure that unwanted spikes fired by the postsynaptic cell can always be removed. A more challenging requirement, the need for reward prediction, is discussed in the context of possible biological implementations of the model.

Although our mechanism for reinforcement learning works using the full STDP rule, it is disconcerting that the LTD portion of this rule contributes nothing to the model's success; indeed, it actually impedes performance and would do so disastrously were it not for the spike suppression mechanism built into the STDP rule we used. On the other hand, there is no reason why the two halves of the STDP rule—pre–post LTP and post–pre LTD—should serve the same functions. LTD induced by the recurrence of postsynaptic spikes preceding EPSPs may serve wholly distinct functions that are not represented in our model. There is growing evidence that the LTP and LTD portions of STDP rule are mechanistically quite distinct, at least at some cortical synapses, with post–pre LTD using a different method for detecting coincident pre- and postsynaptic activity (involving postsynaptic endocannabinoid release and presynaptic NMDARs), possibly using different calcium sources for induction (internal stores instead of extracellular calcium admitted through postsynaptic NMDARs), and perhaps with a different site of expression (Bender et al. 2006; Nevian and Sakmann 2006; Sjöström et al. 2003, 2004). This makes it more likely that they can be regulated independently, and a recent study of hippocampal STDP was able to identify induction protocols that could engage these two forms of synaptic plasticity separately (Wittenberg and Wang 2006). We suggest that post–pre LTD might be suppressed in vivo during reinforcement learning.

One of the strengths of our model is the fact that it is based on an experimentally defined form of synaptic plasticity, but it does require additional conjectures concerning the modulation of that plasticity that are not experimentally established. A more parsimonious model that avoided these conjectures while retaining the ability to train a neural population to map its synaptic inputs into a wide range of possible outputs would be preferable. Legenstein et al. (2005) studied the learning capabilities of the unmodulated STDP rule and describe a method whereby a network using this rule can learn a wide range of mappings from input patterns to output activity. A recently published model by Davison and Frégnac (2006) implements a version of this method to model the learning of coordinate transformations between different frames of reference, where the neural population being trained receives all-to-all inputs from an input layer, encoding untransformed coordinates, and topographic inputs from a "training layer" encoding the desired output. Although this model offers a plausible way to learn a coordinate transformation, it cannot supplant our model in the full range of learning tasks we consider. The method Legenstein et al. (2005) describe for learning arbitrary mappings and the Davison and Frégnac (2006) model are both effectively "instructive," since inputs from the training layer directly bias the output layer toward generating the desired output, whereas our training signal is based on merely the similarity between desired output and actual output. Furthermore, the fact that the projections from the training layer are topographic in the Davison and Frégnac (2006) model means that the evaluation signal is not global; local populations in their output layer receive individually tailored training signals. These conditions are reasonable for learning coordinate transformations, but are probably too demanding for all forms of cortical learning.

Two other models have been published recently that are concerned with the marriage of STDP and reinforcement learning. One, proposed by Izhikevich (2007), posits the modulation of STDP by a reward signal mediated by dopamine. In this model, the relative timing of pre- and postsynaptic spikes generates a synaptic "eligibility trace" governed by the STDP rule, but synaptic changes are implemented only if dopamine is delivered before the eligibility trace decays. The Izhikevich (2007) model offers a solution to the problem of delayed reward, whereas we assume that this problem is solved elsewhere by a system that provides a reward prediction to the network in advance of the actual reward. On the other hand, Izhikevich (2007) considers a much more limited set of potential input patterns and desired output patterns. Because Izhikevich (2007) did not use input patterns with strong, long-range temporal correlations, he did not encounter the problems that required the use anti-Hebbian plasticity coupled to a temporal difference learning signal. A second study, by Pfister et al. (2006), calculates the synaptic changes that increase the likelihood of obtaining a set of target output spike trains given the set of input spike trains, thereby deriving STDP-like learning rules. Pfister et al. (2006) note that if the problem is instead cast in the form of maximizing reward, a similar rule can be derived. However, the STDP rules derived by Pfister et al. (2006) are functions of the "desired" spike times of postsynaptic neurons, not their actual spike times. Although Pfister et al. (2006) provided considerable insight into why the STDP rule might take the form it does, they rely on more abstracted (and more analytically tractable) neural models than we do and do not explore the specific issue of reinforcement learning in great detail. Both Izhikevich (2007) and Pfister et al. (2006) offer valuable approaches to the problem of STDP and reinforcement learning, yet are complementary to our model.

Biological implementation of the model

The most challenging characteristic of our model with regard to credible implementation is probably the need for "reward prediction," i.e., the reinforcement signal must arrive at roughly the same time the activity it evaluates is being generated. This problem is not unique to our model and can be viewed as one specialized facet of the general "temporal credit

assignment problem" all models of reinforcement learning face (Sutton and Barto 1998), but nonetheless is a major obstacle to the implementation of our model by a real neural system. The problem might be solved by giving the evaluation system that calculates the reinforcement signal access to the input activity that drives variations in output activity. The evaluation system could, in principle, use the pattern of input activity generated on a particular trial to predict whether the output activity on that trial will be a better or worse match to the target pattern than average, permitting the timely arrival of appropriate reinforcement. This would be an extraordinary feat of neural computation, but there is a neural population known to do something rather like it: the midbrain dopaminergic neurons. These neurons fire bursts in response to unexpected reward and to stimuli that predict reward; these neurons can also signal the absence of predicted reward through pauses in their spontaneous firing (reviewed in Schultz 1998). If these neurons predict reward based on internal factors, like an efference copy of noisy motor commands, as well as external stimuli, then they could potentially provide the kind of reward prediction required by our model.

Dopamine released by midbrain neurons could provide the reinforcement signal for our model, but dopaminergic innervation of the telencephalon is quite heterogeneous, and is most prominent outside of the isocortex, namely in the striatum, the input structure of the basal ganglia. Within the striatum, dopamine does modulate synaptic plasticity, and although the effects of dopamine are still poorly understood and vigorously debated, it may do so in a way consistent with role of δ_R in our model, with increased dopamine promoting LTP at corticostriatal synapses and decreased dopamine promoting LTD (Reynolds and Wickens 2002). This might make our model plausible within the striatum, but how could it apply to the isocortex, which receives far less dopaminergic input? We begin by asking how midbrain dopaminergic cells generate their reward-predicting responses. It seems unlikely that this complex calculation could be performed entirely by these neurons themselves, and of the various potential sources for this information, one of the best candidates is itself the major target of dopaminergic innervation: the basal ganglia. The basal ganglia receive input from virtually the entire cortex, and thus have access to the primary information needed to predict rewards. Many factors affect the activity of basal ganglia neurons, including of course sensory stimuli and motor plans, but these responses are often modulated by reward expectation (Arkadir et al. 2004; Hikosaka et al. 2006). It is not unreasonable to hypothesize that reward-predicting information can be found not just in midbrain dopaminergic neurons, but also in basal ganglia outputs that are relayed to the isocortex. In this way, almost the entire cortex could receive the reward-predicting information demanded by our model.

Basal ganglia output could conceivably reach the cortex via the GABAergic and cholinergic projections of the basal forebrain (Gritti et al. 1997), and acetylcholine has been reported to modulate cortical synaptic plasticity (Rasmusson 2000). However, the most obvious conduit of basal ganglia output to the cortex is the thalamus. That raises the question of how a glutamatergic thalamocortical projection could modulate corticocortical synaptic plasticity. In rats, the primary thalamic relay from basal ganglia to cortex is the ventromedial nucleus (Gerfen 1992; Gerfen et al. 1982; Kha et al. 2001), which projects to almost the entire cortical mantle, but specifically to layer 1 (Herkenham 1979). This is intriguing from the point of view of our model because layer 1 inputs to the dendritic tufts of pyramidal neurons can trigger dendritic calcium spikes, accompanied by bursts of sodium spikes, when combined with action potentials initiated in the soma (Larkum et al. 1999), and such calcium spikes could influence plastic changes induced at the corticocortical synapses that helped initiate the somatic spike. As we noted in our Methods section, and as emphasized in a recent review (Lisman and Spruston 2005), a number of studies report that low frequency pairing of individual pre- and postsynaptic spikes does not suffice to induce synaptic plasticity. High-frequency pairing is evidently necessary to induce LTP at some cortical synapses (Markram et al. 1997), and this may reflect a requirement for sustained depolarization (Sjöström et al. 2001). Dendritic spikes triggered by layer 1 excitation may help meet this requirement, and a recent report indicates that the requirement for high-frequency pairing is waived when the postsynaptic neuron fires bursts rather than individual spikes (Nevian and Sakmann 2006). Another study of layer 5 pyramidal neurons found that EPSPs followed by single APs induced anti-Hebbian LTD, whereas EPSPs followed by high-frequency bursts—triggering large dendritic spikes—induced LTP (Letzkus et al. 2006). In our view, reports that the induction of synaptic plasticity requires more than is accounted for by the basic STDP rule does not necessarily undermine the STDP concept per se; rather, they indicate that STDP is modulated, that this modulation may even encompass the possibility of anti-Hebbian STDP, and that this modulation may be accomplished by a system that is capable of providing the reward-predicting reinforcement signal we require.

Reward-modulated STDP as a model for song learning in oscine birds

This speculative hypothesis would be more plausible if we could identify a specific example featuring a learned behavior with a known neural substrate to which our model might be applied. There is as yet no good example in mammals of a learned behavior whose specific cortical and subpallial substrates have been identified and characterized, but such an example does exist in songbirds. These birds must learn the songs they sing, and the neural substrate for this behavior consists of two well-described forebrain pathways: a "motor pathway" from HVC to the robust nucleus of the acropallium (RA) required for singing per se, and an "anterior forebrain pathway" (AFP) that is required for song learning (for reviews, see Brainard 2004; Farries 2004; Fee et al. 2004). The AFP is hypothesized to evaluate the bird's vocal performance and transmit information to RA that enables learning, and it contains basal ganglia circuitry very similar to that of mammals (Farries and Perkel 2002; Farries et al. 2005). Thus the AFP could play the role of the "evaluation system" in our model, while HVC and RA correspond to the input and output layers, respectively. Furthermore, HVC projects to the AFP and supplies both auditory and premotor information (e.g., Doupe 1997; Hessler and Doupe 1999), giving the AFP the information it would need to predict performance from premotor activity. HVC neurons projecting to RA even fire in the one-burst pattern that works best for our model; these neurons

fire a single high-frequency burst during a song motif (Hahnloser et al. 2002). For these reasons, the song system could be an ideal testing ground for our STDP-based model of reinforcement learning and its implementation via the basal ganglia.

Conversely, the song system does differ in certain critical ways from the basal ganglia-thalamocortical system we propose for mammals. First, feedback from the AFP reaches the motor pathway via a pallial (cortex-like) nucleus, the lateral magnocellular nucleus of the medial nidopallium (LMAN), rather than directly from the thalamus. Furthermore, the avian pallium is not organized into laminae; thus there is no "layer 1" to receive modulatory inputs. Even so, the LMAN-RA projection has an unusual property that could help it play the same functional role as the one we propose for VM's innervation of layer 1: the postsynaptic receptors at LMAN-RA synapses are almost exclusively NMDARs (Mooney and Konishi 1991; Stark and Perkel 1999). This fact has long been touted as a possible link between behavioral plasticity (dependent on LMAN) and synaptic plasticity, which in other systems depends on calcium influx through NMDARs. However, NMDARs are not just conduits for calcium; they are also dendritic voltage-gated ion channels whose availability is controlled extrinsically, by glutamate. As voltage-gated channels, NMDARs might help generate dendritic spikes in RA neurons, as they are known to do in mammalian cortical neurons (Schiller et al. 2000). We suggest that activity in LMAN, controlled by basal ganglia circuitry upstream in the AFP, could influence the occurrence of dendritic spikes in RA neurons, and thereby control the magnitude and polarity of plasticity induced at HVC-RA and intrinsic RA-RA synapses.

This perspective, wherein the AFP's primary role is to evaluate performance and modulate plasticity but not to directly influence behavior, is an old one in the songbird literature, supported by early lesion studies demonstrating that while the AFP is required for song learning, it is *not* required for singing in birds that have already learned their song (Bottjer et al. 1984; Scharff and Nottebohm 1991; Sohrabji et al. 1990). However, this view has been challenged recently by two observations. First, the AFP does in fact influence behavior; specifically, activity in LMAN (the output station of the AFP) contributes to song variability (Kao et al. 2005; Ölveczky et al. 2005). Second, LMAN activity recorded during singing does not appear to be influenced by auditory feedback (Leonardo 2004), as it should if LMAN is transmitting a signal derived from comparing the actual song to an auditory representation of the target song. But our model posits a reinforcement signal that is derived from a *prediction* of performance based on premotor activity—a direct auditory comparison of actual song to target song would arrive too late to be of service in our model. Thus our model is perfectly consistent with Leonardo's (2004) results. Of course, auditory feedback is necessary in the long run to establish and maintain the putative mapping between premotor activity and performance prediction, consistent with the known effects of deafening on the acquisition and maintenance of song (Konishi 1965; Nordeen and Nordeen 1992). As for the behavioral variability, Ölveczky et al. (2005) note that this is just as important for reinforcement learning as the evaluation of the variants, and suggest that the generation of variability may be the prime function of the AFP, with the evaluation performed elsewhere. Although AFP output undeniably enhances behavioral variability, it is possible that this is simply an epiphenomenon, a side effect that occurs as the AFP performs its primary task of modulating plasticity. On the other hand, there is no reason why the AFP could not serve both functions, helping to generate variants and evaluating them. Indeed, if the AFP is able to "predict" which variants will better match the tutor song, then it may well bias variation in a way that accelerates learning, a possibility also raised by Ölveczky et al. (2005). This may prove to be a line of convergence between the roles traditionally ascribed to the songbird AFP (evaluation of behavior) and to the mammalian basal ganglia (control of behavior).

Our model can claim two accomplishments hitherto rare in the modeling literature: *1*) it proposes a mechanism for reinforcement learning employing known physiological phenomena with relatively modest modifications, and *2*) it uses STDP, albeit in modified form, to achieve a general-purpose form of learning. Along the way, we identified a number of requirements which can also be regarded as predictions, i.e., things that must be true of any system that implements the model. The most prominent of these are *1*) STDP can be modulated, *2*) this modulation includes the possibility of anti-Hebbian STDP, *3*) this modulation is "predictive" in the sense discussed above, and *4*) STDP is not multiplicative in the sense of Rubin et al. (2001). This list is necessarily incomplete; there are many things that could impact this model's performance that were not examined. Future studies will have to evaluate the effects of such things as recurrent excitatory connections, nonrandom patterns of connectivity, inhibitory networks, and intrinsic physiological properties. Even with the model as it stands, some important questions remain unanswered, including the number of input-output pairs that can be "stored" in these networks, the factors that control this capacity, and the extent to which these networks can learn continuous mappings between input and output (as opposed to a list of discrete input–output pattern pairs). Independent of particular details of implementation, we hope that this model can serve as a starting point from which we can understand how neural systems learn to generate appropriate responses to the inputs they receive.

REFERENCES

Abbott LF, Blum KI. Functional significance of long-term potentiation for sequence learning and prediction. *Cereb Cortex* 6: 406–416, 1996.

Andersen P, Sundberg SH, Sveen O, Wigström H. Specific long-lasting potentiation of synaptic transmission in hippocampal slices. *Nature* 266: 736–737, 1977.

Arkadir D, Morris G, Vaadia E, Bergman H. Independent coding of movement direction and reward prediction by single pallidal neurons. *J Neurosci* 24: 10047–10056, 2004.

Bell CC, Han VZ, Sugawara Y, Grant K. Synaptic plasticity in a cerebellum-like structure depends on temporal order. *Nature* 387: 278–281, 1997.

Bender VA, Bender KJ, Brasier DJ, Feldman DE. Two coincidence detectors for spike timing-dependent plasticity in somatosensory cortex. *J Neurosci* 26: 4166–4177, 2006.

Bi G-q, Poo M-m. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neurosci* 18: 10464–10472, 1998.

Blum KI, Abbott LF. A model of spatial map formation in the hippocampus of the rat. *Neural Comput* 8: 85–93, 1996.

Bottjer SW, Miesner EA, Arnold AP. Forebrain lesions disrupt development but not maintenance of song in passerine birds. *Science* 224: 901–903, 1984.

Brainard MS. Contributions of the anterior forebrain pathway to vocal plasticity. *Ann NY Acad Sci* 1016: 377–394, 2004.

Cho K, Aggleton JP, Brown MW, Bashir ZI. An experimental test of the role of postsynaptic calcium levels in determining synaptic strength using perirhinal cortex of rat. *J Physiol* 532: 459–466, 2001.

Cormier RJ, Greenwood AC, Connor JA. Bidirectional synaptic plasticity correlated with the magnitude of dentritic calcium transients above a threshold. *J Neurophysiol* 85: 399–406, 2001.

Cummings JA, Mulkey RM, Nicoll RA, Malenka RC. Ca^{2+} signaling requirements for long-term depression in the hippocampus. *Neuron* 16: 825–833, 1996.

Davison AP, Frégnac Y. Learning cross-model spatial transformations through spike timing-dependent plasticity. *J Neurosci* 26: 5604–5615, 2006.

de Ruyter van Steveninck RR, Bialek W. Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequences. *Proc R Soc Lond B* 234: 379–414, 1988.

Debanne D, Gähwiler BH, Thompson SM. Long-term synaptic plasticity between pairs of individual CA3 pyramidal cells in rat hippocampal slice cultures. *J Physiol* 507: 237–247, 1998.

Desai NS, Rutherford LC, Turrigiano GG. Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat Neurosci* 2: 515–520, 1999.

Doupe AJ. Song- and order-selective neurons in the songbird anterior forebrain and their emergence during vocal development. *J Neurosci* 17: 1147–1167, 1997.

Farries MA. The avian song system in comparative perspective. *Ann NY Acad Sci* 1016: 61–76, 2004.

Farries MA, Ding L, Perkel DJ. Evidence for "direct" and "indirect" pathways through the song system basal ganglia. *J Comp Neurol* 484: 93–104, 2005.

Farries MA, Perkel DJ. A telencephalic nucleus essential for song learning contains neurons with physiological characteristics of both striatum and globus pallidus. *J Neurosci* 22: 3776–3787, 2002.

Fee MS, Kozhevnikov AA, Hahnloser RHR. Neural mechanisms of vocal sequence generation in the songbird. *Ann NY Acad Sci* 1016: 153–170, 2004.

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

Froemke RC, Dan Y. Spike-timing-dependent synaptic modifiation induced by natural spike trains. *Nature* 416: 433–438, 2002.

Froemke RC, Poo M-m, Dan Y. Spike-timing-dependent synaptic plasticity depends on dendritic location. *Nature* 434: 221–225, 2005.

Froemke RC, Tsay IA, Raad M, Long JD, Dan Y. Contribution of individual spikes in burst-induced long-term synaptic modification. *J Neurophysiol* 95: 1620–1629, 2006.

Gerfen CR. The neostriatal mosaic: multiple levels of compartmental organization in the basal ganglia. *Annu Rev Neurosci* 15: 285–320, 1992.

Gerfen CR, Staines WA, Arbuthnott GW, Fibiger HC. Crossed connections of the substantia nigra in the rat. *J Comp Neurol* 207: 283–303, 1982.

Gerstner W, Kempter R, van Hemmen JL, Wagner H. A neuronal learning rule for sub-millisecond temporal coding. *Nature* 383: 76–78, 1996.

Gritti I, Mainville L, Mancia M, Jones BE. GABAergic and other noncholinergic basal forebrain neurons, together with cholinergic neurons, project to the mesocortex and isocortex in the rat. *J Comp Neurol* 383: 163–177, 1997.

Gütig R, Aharonov R, Rotter S, Sompolinsky H. Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity. *J Neurosci* 23: 3697–3714, 2003.

Hahnloser RHR, Kozhevnikov AA, Fee MS. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* 419: 65–69, 2002.

Herkenham M. The afferent and efferent connections of the ventromedial thalamic nucleus in the rat. *J Comp Neurol* 183: 487–518, 1979.

Hessler NA, Doupe AJ. Singing-related neural activity in a dorsal forebrain-basal ganglia circuit of adult zebra finches. *J Neurosci* 19: 10461–10481, 1999.

Hikosaka O, Nakamura K, Nakahara H. Basal ganglia orient eyes to reward. *J Neurophysiol* 95: 567–584, 2006.

Ismailov I, Kalikulov D, Inoue T, Friedlander MJ. The kinetic profile of intracellular calcium predicts long-term potentiation and long-term depression. *J Neurosci* 24: 9847–9861, 2004.

Izhikevich EM. Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb Cortex* 17: 2443–2452, 2007.

Kao MH, Doupe AJ, Brainard MS. Contributions of an avain basal ganglia-forebrain circuit to real-time modulation of song. *Nature* 433: 638–643, 2005.

Kempter R, Gerstner W, van Hemmen JL. Hebbian learning and spiking neurons. *Phys Rev E* 59: 4498–4514, 1999.

Kempter R, Gerstner W, van Hemmen JL. Intrinsic stabilization of output rates by spike-based Hebbian learning. *Neural Comput* 13: 2709–2741, 2001.

Kepecs A, van Rossum MCW, Song S, Tegnér J. Spike-timing-dependent plasticity: common themes and divergent vistas. *Biol Cybern* 87: 446–458, 2002.

Kha HT, Finkelstein DI, Tomas D, Drago J, Pow DV, Horne MK. Projections from the substantia nigra pars reticulata to the motor thalamus of the rat: single axon reconstructions and immunohistochemical study. *J Comp Neurol* 440: 20–30, 2001.

Köles L, Wirkner K, Illes P. Modulation of ionotropic glutamate receptor channels. *Neurochem Res* 26: 925–932, 2001.

Konishi M. The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Z Tierpsychol* 22: 770–783, 1965.

Larkum ME, Zhu JJ, Sakmann B. A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature* 398: 338–341, 1999.

Legenstein R, Naeger C, Maas W. What can a neuron learn with spike-timing-dependent plasticity? *Neural Comput* 17: 2337–2382, 2005.

Leonardo A. Experimental test of the birdsong error-correction model. *Proc Natl Acad Sci USA* 101: 16935–16940, 2004.

Letzkus JJ, Kampa BM, Stuart GJ. Learning rules for spike timing-dependent plastcity depend on dendritic synapse location. *J Neurosci* 26: 10420–10429, 2006.

Lisman JE, Spruston N. Postsynaptic depolarization requirements for LTP and LTD: critique of spike timing-dependent plasticity. *Nat Neurosci* 8: 839–841, 2005.

MacDonald JF, Xiong X-G, Lu W-Y, Raouf R, Orser BA. Modulation of NMDA receptors. *Prog Brain Res* 116: 191–208, 1998.

Magee JC, Johnston D. A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. *Science* 275: 209–212, 1997.

Markram H, Lübke J, Frotscher M, Sakaguchi H. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275: 213–215, 1997.

Mooney R, Konishi M. Two distinct inputs to an avian song nucleus activate different glutamate receptor subtypes on individual neurons. *Proc Natl Acad Sci USA* 88: 4075–4079, 1991.

Nevian T, Sakmann B. Spine Ca^{2+} signaling in spike-timing-dependent plasticity. *J Neurosci* 26: 11001–11013, 2006.

Nishiyama M, Hong K, Mikoshiba K, Poo M-m, Kato K. Calcium stores regulate the polarity and input specificity of synaptic modification. *Nature* 408: 584–588, 2000.

Nordeen KW, Nordeen EJ. Auditory feedback is necessary for the maintenance of stereotyped song in adult zebra finches. *Behav Neural Biol* 57: 58–66, 1992.

Ölveczky BP, Andalman AS, Fee MS. Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLoS Biol* 3: 902–909, 2005.

Pfister J-P, Toyoizumi T, Barber D, Gerstner W. Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. *Neural Comput* 18: 1318–1348, 2006.

Pike FG, Meredith RM, Oldingand AWA, Paulsen O. Postsynaptic bursting is essential for 'Hebbian' induction of associative long-term potentiation at excitatory synapses in rat hippocampus. *J Physiol* 518: 571–576, 1999.

Rao RPN, Sejnowski TJ. Spike-timing-dependent Hebbian plasticity as temporal difference learning. *Neural Comput* 13: 2221–2237, 2001.

Rasmusson DD. The role of acetylcholine in cortical synaptic plasticity. *Behav Brain Res* 115: 205–218, 2000.

Reinagel P, Reid RC. Temporal coding of visual information in the thalamus. *J Neurosci* 20: 5392–5400, 2000.

Reynolds JNJ, Wickens JR. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw* 15: 507–521, 2002.

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

Rubin J, Lee DD, Sompolinsky H. Equilibrium properties of temporally asymmetric Hebbian plasticity. *Phys Rev Lett* 86: 364–367, 2001.

Scharff C, Nottebohm F. A comparative study of the behavior deficits following lesions of various parts of the zebra finch song system: implications for vocal learning. *J Neurosci* 11: 2896–2913, 1991.

Schiller J, Major G, Koester HJ, Schiller Y. NMDA spikes in basal dendrites of cortical pyramidal neurons. *Nature* 404: 285–289, 2000.

Schultz W. Predictive reward signal of dopamine neurons. *J Neurophysiol* 80: 1–27, 1998.

Seung HS. Learning in spiking neural networks by reinforcement of stochastic synaptic tranmission. *Neuron* 40: 1063–1073, 2003.
Sjöström PJ, Häusser MA. A cooperative switch determines the sign of synaptic plasticity in distal dendrites of neocortical pyramidal neurons. *Neuron* 51: 227–238, 2006.
Sjöström PJ, Turrigiano GG, Nelson SB. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron* 32: 1149–1164, 2001.
Sjöström PJ, Turrigiano GG, Nelson SB. Neocortical LTD via coincident activation of presynaptic NMDA and cannabinoid receptors. *Neuron* 39: 641–654, 2003.
Sjöström PJ, Turrigiano GG, Nelson SB. Endocannabinoid-dependent neocortical layer-5 LTD in the absence of postsynaptic spiking. *J Neurophysiol* 92: 3338–3343, 2004.
Sohrabji F, Nordeen EJ, Nordeen KW. Selective impairment of song learning following lesions of a forebrain nucleus in juvenile zebra finches. *Behav Neural Biol* 53: 51–63, 1990.
Song S, Abbott LF. Cortical development and remapping through spike timing-dependent plasticity. *Neuron* 32: 339–350, 2001.
Song S, Miller KD, Abbott LF. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nat Neurosci* 3: 919–926, 2000.
Stark LL, Perkel DJ. Two-stage, input-specific synaptic maturation in a nucleus essential for vocal production in the zebra finch. *J Neurosci* 19: 9107–9116, 1999.
Suri RE, Sejnowski TJ. Spike propagation synchronized by temporally asymmetric Hebbian learning. *Biol Cybern* 87: 440–445, 2002.
Sutton RS, Barto AG. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press, 1998.
Tegnér J, Kepecs A. An adaptive spike-timing-dependent plasticity rule. *Neurocomputing* 44–46: 189–194, 2002.
Turrigiano GG, Leslie KR, Desai NS, Rutherford LC, Nelson SB. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* 391: 892–896, 1998.
Tzounopoulos T, Kim Y, Oertel D, Trussell LO. Cell-specific, spike timing-dependent plasticities in the dorsal cochlear nucleus. *Nat Neurosci* 7: 719–725, 2004.
van Rossum MCW, Bi G-q, Turrigiano GG. Stable Hebbian learning from spike timing-dependent plasticity. *J Neurosci* 20: 8812–8821, 2000.
Wang H-X, Gerkin RC, Nauen DW, Bi G-q. Coactivation and timing-dependent integration of synaptic potentiation and depression. *Nat Neurosci* 8: 187–193, 2005.
Wittenberg GM, Wang SS-H. Malleability of spike-timing-dependent plasticity at the CA3-CA1 synapse. *J Neurosci* 26: 6610–6617, 2006.
Xie X, Seung HS. Learning in neural networks by reinforcement of irregular spiking. *Phys Rev E* 69: 041909, 2004.
Yang S-N, Tang Y-G, Zucker RS. Selective induction of LTP and LTD by postsynaptic $[Ca^{2+}]_i$ elevation. *J Neurophysiol* 81: 781–787, 1999.