

Model of Birdsong Learning Based on Gradient Estimation by Dynamic Perturbation of Neural Conductances

Ila R. Fiete,^{1,2} Michale S. Fee,^{3,5} and H. Sebastian Seung^{4,5}

¹Kavli Institute for Theoretical Physics, University of California, Santa Barbara, Santa Barbara; ²Center for Theoretical Biological Physics, University of California, San Diego, La Jolla, California; and ³McGovern Institute for Brain Research, ⁴Howard Hughes Medical Institute, and ⁵Brain and Cognitive Sciences Department, Massachusetts Institute of Technology, Cambridge, Massachusetts

Submitted 14 December 2006; accepted in final form 13 July 2007

Fiete IR, Fee MS, Seung HS. Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *J Neurophysiol* 98: 2038–2057, 2007. First published July 25, 2007; doi:10.1152/jn.01311.2006. We propose a model of songbird learning that focuses on avian brain areas HVC and RA, involved in song production, and area LMAN, important for generating song variability. Plasticity at HVC → RA synapses is driven by hypothetical “rules” depending on three signals: activation of HVC → RA synapses, activation of LMAN → RA synapses, and reinforcement from an internal critic that compares the bird’s own song with a memorized template of an adult tutor’s song. Fluctuating glutamatergic input to RA from LMAN generates behavioral variability for trial-and-error learning. The plasticity rules perform gradient-based reinforcement learning in a spiking neural network model of song production. Although the reinforcement signal is delayed, temporally imprecise, and binarized, the model learns in a reasonable amount of time in numerical simulations. Varying the number of neurons in HVC and RA has little effect on learning time. The model makes specific predictions for the induction of bidirectional long-term plasticity at HVC → RA synapses.

INTRODUCTION

Songbirds hatch not knowing how to sing. At first, a juvenile male is incapable of complex vocalizations and merely listens to its tutor’s song. After the bird begins to vocalize, it gradually learns to produce an accurate copy of the tutor’s song (Immelmann 1969; Price 1979), even when isolated from all other birds including its tutor (Brainard and Doupe 2002). In the latter period, auditory feedback of the bird’s own song is crucial: if the bird is deafened after exposure to the tutor song but before it begins to sing, it cannot learn (Konishi 1965; Marler and Tamura 1964). Together, these findings suggest that the juvenile songbird memorizes a template of the tutor song and afterward learns by comparing its own vocalizations with the template.

A number of song-related avian brain areas have been discovered (Fig. 1A). Song production areas (Fig. 1A, open blue) include HVC (high vocal center) and RA (robust nucleus of the arcopallium), which generate sequences of neural activity patterns and through motoneurons control the muscles of the vocal apparatus during song (Hahnloser et al. 2002; Suthers and Margoliash 2002; Wild 1993, 2004; Yu and Margoliash 1986). Lesion of HVC or RA causes immediate loss of song (Nottebohm et al. 1976; Simpson and Vicario 1990). Other

areas in the anterior forebrain pathway (AFP) appear to be important for song learning but not production (Fig. 1A, filled green), at least in adults. The AFP is regarded as an avian homologue of the mammalian basal ganglia thalamocortical loop (Farries 2004; Perkel 2004; Reiner et al. 2004). In particular, lesion of area LMAN (lateral magnocellular nucleus of the nidopallium) has little immediate effect on song production in adults, but arrests song learning in juveniles (Bottjer et al. 1984; Doupe 1993; Scharff and Nottebohm 1991). These facts suggest that LMAN plays a role in driving song learning, but the locus of plasticity is in brain areas related to song production, such as HVC and RA.

Actor, critic, and experimenter

Nearly a decade ago, Doya and Sejnowski (1998) attempted to place such observations in a schema borrowed from mathematical theories of reinforcement learning. In this schema, learning is based on interactions between an *actor* and a *critic* (Fig. 1B). The critic evaluates the performance of the actor at a desired task. The actor uses this evaluation to change in a way that improves its performance. To learn by trial and error, the actor performs the task differently each time. It generates both good and bad variations, and the critic’s evaluation is used to reinforce the good ones.¹ Ordinarily it is assumed that the actor generates variations by itself. However, Doya and Sejnowski considered a schema in which the source of variation is external to the actor. We will call this source the *experimenter*.

Doya and Sejnowski proceeded to identify the three parts of their schema with specific areas of the avian brain. The actor was identified with HVC, RA, and the motor neurons that control vocalization. They hypothesized that the actor learns through plasticity at the synapses from HVC to RA (Fig. 1C). Based on evidence of structural changes like axonal growth and retraction that take place in the HVC to RA projection during song learning (Herrmann and Arnold 1991; Kittelberger and Mooney 1999; Mooney 1992; Sakaguchi and Saito 1996; Stark and Scheich 1997), this view is widely regarded as plausible. Curiously, no reliable protocols for the induction of activity-dependent plasticity at these synapses in vitro have yet been found (R Mooney, private communication), possibly for

¹ We use the term “critic” here to emphasize connections with previous theories of reinforcement learning (Barto et al. 1983).

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.

Address for reprint requests and other correspondence: I. R. Fiete, California Institute of Technology, Division of Biology 216-76, Pasadena, CA 91125 (E-mail: ilafiete@caltech.edu).

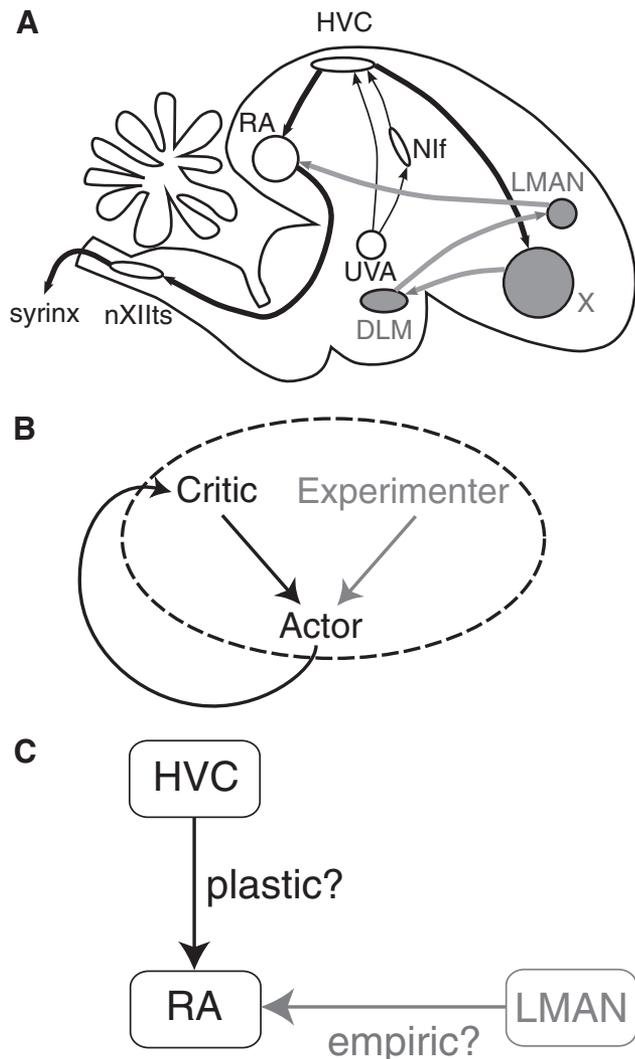


FIG. 1. Avian song pathways and the tripartite hypotheses. *A*: avian brain areas involved in song production and song learning. Premotor pathway (open) includes areas necessary for song production. Anterior forebrain pathway (filled) is required for song learning but not for song production. *B*: tripartite reinforcement learning schema: the *actor* produces behavior; the *experimenter* sends fluctuating input to the actor, producing variability in behavior that is used for trial-and-error learning; the *critic* evaluates the behavior of the actor and sends a reinforcement signal to it. For birdsong, the actor includes premotor song production areas HVC (high vocal center) and RA (robust nucleus of the arcopallium). Doya and Sejnowski hypothesized that the experimenter is LMAN (lateral magnocellular nucleus of the nidopallium). Location of the critic is unknown. *C*: plastic and empiric synapses. RA receives synaptic input from both HVC and LMAN. We will call the HVC synapses “plastic,” in keeping with the hypothesis that these synapses are the locus of plasticity for song learning. Doya and Sejnowski conjectured that LMAN produces song variability by driving slow exploration in HVC \rightarrow RA synaptic strengths. However, recent data indicate that LMAN produces transient song perturbations (Kao et al. 2005) by driving rapid conductance fluctuations in postsynaptic RA neurons (Olviczky et al. 2005). We will refer to the connections from LMAN to RA as “empiric,” in keeping with the hypothesis that they are specialized for experimentation.

good reasons, which we consider in the DISCUSSION. For the experimenter and critic, Doya and Sejnowski turned to the anterior forebrain pathway, hypothesizing that the critic is X and the experimenter is LMAN.

What is the current status of the Doya–Sejnowski tripartite schema? The actor part of their model was on firm ground, but

their ideas about the critic and the experimenter were more speculative. Unfortunately, the location of the critic is still unknown, although it is widely believed to exist. Because the critic has not been found, the nature of its feedback is still unknown. One could imagine a powerful critic, which gives the actor specific instructions about how to improve song. This would place more of the computational burden of the learning problem on the critic. Or one could imagine a weak critic, which simply tells the actor whether performance is good or bad. This would place more of the burden of learning on the actor.

On the other hand, there is increasing support for their general idea of LMAN as an experimenter. First, we review evidence in support of LMAN as an experimenter. Then we argue that recent experiments show important departures from the assumptions of Doya and Sejnowski about the structure and dynamics of LMAN’s input to RA, which call for a different formulation of learning with LMAN experimentation in the songbird system.

During song, LMAN neural spiking is quite variable from trial to trial and more irregular than activity in RA (Hessler and Doupe 1999b; Leonardo 2004). Moreover, mean activity in LMAN correlates with the overall song variability: In adult birds, LMAN activity is low during song directed at females, which tends to be extremely stable and stereotyped, and much higher during the more variable *undirected* song (Hessler and Doupe 1999b; Kao et al. 2005). Although, as noted earlier, LMAN lesions have little effect on adult song, especially during directed bouts, closer inspection reveals that LMAN lesions reduce the slight variability present in adult undirected song (Kao et al. 2005). In juveniles, there is much greater trial-to-trial song variability compared with that of adults; this is dramatically reduced after LMAN lesions (Scharff and Nottebohm 1991). Recently it was shown that reversible pharmacological inactivation of juvenile LMAN with tetrodotoxin (TTX) or muscimol leads to immediate reduction in song variability (Kao et al. 2005; Olviczky et al. 2005). All of this evidence suggests that LMAN generates song variability through its projection to RA.²

But how, mechanistically and functionally, does LMAN drive song variability and learning? Doya and Sejnowski proposed that the role of LMAN input to RA is to produce a fluctuation that is *static* over the duration of a song bout, directly in the *synaptic strengths* from premotor nucleus HVC to RA. From a functional perspective, the model of Doya and Sejnowski is akin to “weight perturbation” (Dembo and Kailath 1990; Seung 2003; Williams 1992) and relatively easy to implement: a temporary but static HVC–RA weight change that lasts the duration of one song causes some change in song performance. If performance is good, the critic sends a reinforcement signal that makes the temporary static perturbation permanent. From a neurobiological perspective their model requires machinery whereby *N*-methyl-D-aspartate (NMDA)–mediated synaptic transmission from LMAN to RA can drive synaptic weight changes that remain static over the 1- to 2-s

² There are other suggestions for the role of LMAN in song learning, and its role in song learning is far from settled: Some have hypothesized that LMAN is a critic, but the appropriate neural signals have not been found there. Others have hypothesized that LMAN provides a permissive signal that gates synaptic plasticity and learning (Brainard and Doupe 2000a; Margoliash 2002; Troyer and Bottjer 2001).

duration of song, in the heterosynaptic HVC–RA connections. However, LMAN activity in the songbird is dynamic and variable throughout song, evolving on a 10- to 100-ms timescale (Hessler and Doupe 1999a,b; Leonardo 2004), at odds with the assumption that at the beginning of song LMAN triggers an instantaneous perturbation in the HVC–RA weights, which is then held constant throughout the song.

Next, in recent experiments, transient stimulation in LMAN leads to transient, subsyllable-long changes in either song pitch or amplitude (Kao et al. 2005). Presumably, local stimulation excites local myotopic ensembles of LMAN neurons; if this LMAN activity led to static perturbations of a set of HVC synapses projecting to a myotopic RA group, it would have produced changes in pitch or amplitude that were not transient, but lasted to produce consistent biases in pitch or amplitude throughout one song iteration. In Olveczky et al. (2005), blocking NMDA receptor currents in RA causes the same reduction in song variability as does LMAN inactivation,³ indicating that the effects of LMAN activity in RA are through ordinary glutamatergic synaptic transmission into RA neurons. In short, LMAN appears to drive fast, transient song fluctuations on a subsyllable level, effected by ordinary excitatory transmission that drives *dynamic postsynaptic membrane conductance fluctuations* in the postsynaptic RA neurons. This picture of rapidly fluctuating glutamatergic input from LMAN driving fast conductance perturbations in RA is quite different, in its neurobiological mechanism and mathematical implications for reinforcement learning, from the Doya and Sejnowski model based on slow modulatory influences on HVC → RA weights.

Finally, for song learning, synapses from different HVC neurons to the same postsynaptic RA neuron must have the flexibility to change in opposite directions. Within the weight-perturbation model of Doya and Sejnowski, this requires that each synapse from HVC onto a single RA neuron receive independent perturbations in different directions, relative to other synapses from different HVC neurons onto the same RA neuron. In neurobiological terms, this could be possible if, for each synapse from a distinct HVC neuron onto a RA neuron, there were a separate LMAN input. However, this seems unlikely considering that each RA neuron receives only about 50 synapses from LMAN (Canady et al. 1988; Hermann and Arnold 1991) compared with about 1,000 synapses from ~200 different HVC neurons (Kittelberger and Mooney 1999).

Next, we describe a learning rule that, like the weight-perturbation scheme used by Doya and Sejnowski, also belongs in the broad category of actor–critic reinforcement learning rules. However, the rule is distinct functionally and in its neurobiological implications from weight-perturbation-like schemes. Applied to the song system, the rule is fully consistent with the physiological and anatomical findings on LMAN input to RA and with the phenomenology of song learning.

Learning with empiric synapses

The goal of this work is to relate the high-level concept of reinforcement learning by the tripartite schema to a biologically realistic lower level of description in terms of microscopic events at synapses and neurons in the birdsong system, to demonstrate song learning in a network of realistic spiking neurons, and to examine the plausibility of reinforcement algorithms in explaining biological fine motor skill learning with respect to learning time in the birdsong network.

The present model is based on many of the same general assumptions that were made by Doya and Sejnowski. We assume a tripartite actor–critic–experimenter schema. The critic is weak, providing only a scalar evaluation signal. The HVC sequence is fixed, and only the map from HVC to the motor neurons is learned, through plasticity at the HVC → RA synapses.⁴ LMAN perturbs song through its inputs to the song premotor pathway. However, the structure and dynamics of LMAN inputs, and their influence on learning, are different, with distinct neurobiological implications. In keeping with our hypothesis that the function of LMAN drive to RA is to perform “experiments” for trial-and-error learning, the connections from LMAN to RA will be called “empiric” synapses (Fig. 1C).

We make a specific theoretical proposal for synaptic reinforcement learning in the case of birdsong, illustrated in Fig. 2. Functionally, our scheme is similar to “node perturbation” (Fiete and Seung 2006; Werfel et al. 2005; Xie and Seung 2004) because it relies on independent perturbations delivered to neurons (rather than to individual plastic synapses, as in weight perturbation). From a neurobiological perspective, this scheme is more realistic, for two reasons. First, it is in better agreement with the microanatomy of LMAN–RA synapses because it only requires one independent LMAN input per RA neuron, rather than per HVC–RA synapse. Second, the perturbation to each neuron in our model is temporally varying on a rapid timescale, not static, during song. This is consistent with activity in LMAN during song production and song learning.

We assume that each RA neuron receives many plastic synaptic inputs from HVC, in addition to a single *empiric* synapse from LMAN that dynamically drives the postsynaptic RA conductance throughout song by ordinary excitatory neurotransmission.⁵ The dynamic postsynaptic conductance perturbations must somehow be translated into appropriate instructions for plasticity in the incoming plastic synapses. Because each RA neuron continually receives conductance inputs from both HVC and LMAN, and both vary with time during song, the challenge is to understand how an RA neuron might use dynamic LMAN perturbations to extract information about the correct long-term weight changes for its HVC inputs.

In our proposal, the conductance of the plastic synapse from neuron j in HVC to neuron i in RA is given by $W_{ij} s_{ij}^{\text{HVC}}(t)$, where the synaptic activation $s_{ij}^{\text{HVC}}(t)$ determines the time course of conductance changes, and the plastic parameter W_{ij} determines their amplitude. Changes in W_{ij} are governed by the plasticity rule

³ This intervention primarily blocks glutamatergic input to RA from LMAN, which is primarily through NMDA receptors (Mooney 1992; Stark and Perkel 1999). It is thought to have little effect on glutamatergic input to RA from HVC, which is primarily through non-NMDA receptors (Mooney 1992; Stark and Perkel 1999).

⁴ It is also important to have a complementary theory of how the HVC sequence is formed, but this is outside the scope of this paper.

⁵ In reality, there may be multiple synapses from LMAN onto an RA neuron, but most likely there are far fewer from LMAN than from HVC (Canady et al. 1988; Gurney 1981; Hermann and Arnold 1991).

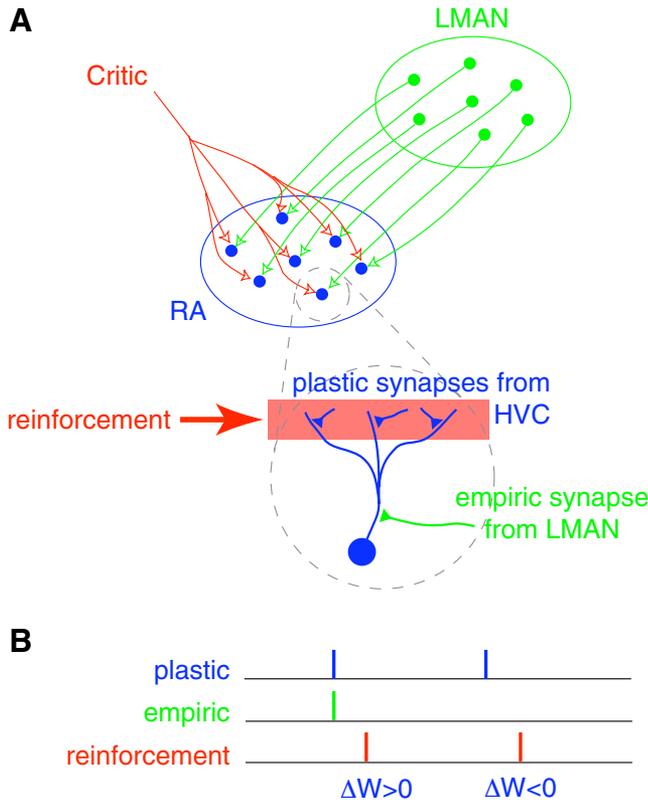


FIG. 2. A proposal for plasticity rules at HVC → RA synapses. Synaptic plasticity rule for gradient estimation by dynamic perturbation of conductances. We use the actor–critic–experimenter schema (Fig. 1B) and distinguish between plastic and empiric synapses (Fig. 1C). A: neurons in the experimenter (LMAN) dynamically perturb the conductances of RA neurons through empiric synapses. Critic signals improvements in performance and globally broadcasts a reinforcement signal to all plastic synapses (HVC → RA). B: if coincident activation of a plastic synapse and empiric synapse onto the same RA neuron is followed by reinforcement, then the plastic synapse is strengthened. If activation of the plastic synapse without the empiric synapse is followed by reinforcement, the plastic synapse is weakened.

$$\frac{dW_{ij}}{dt} = \eta R(t) e_{ij}(t) \tag{1}$$

The positive parameter η , called the learning rate, controls the overall amplitude of synaptic changes. The eligibility trace $e_{ij}(t)$ is a hypothetical quantity present at every plastic synapse. It signifies whether the synapse is “eligible” for modification by reinforcement and is based on the recent activation of the plastic synapse and the empiric synapse onto the same RA neuron

$$e_{ij}(t) = \int_0^t dt' G(t-t') [s_i^{\text{LMAN}}(t') - \langle s_i^{\text{LMAN}} \rangle] s_{ij}^{\text{HVC}}(t') \tag{2}$$

Here $s_i^{\text{LMAN}}(t)$ is the conductance of the empiric (LMAN → RA) synapse onto the i th RA neuron. The temporal filter $G(t)$ is assumed to be nonnegative and its shape determines how far back in time the eligibility trace can “remember” the past.

An important aspect of Eq. 2 is that the instantaneous activation of the empiric synapse is measured relative to its own expected activity $\langle s_i^{\text{LMAN}}(t) \rangle$. This subtraction of average activation in the empiric synapse enables bidirectional synaptic changes, even if the reinforcement signal $R(t)$ is constrained to

be nonnegative.⁶ In our model, each empiric synapse is driven by a Poisson spike train from an LMAN neuron with constant firing rate, so $\langle s_i^{\text{LMAN}} \rangle$ is a fixed constant throughout song and throughout learning (and thus easy to estimate by a simple time average) for every RA neuron.

The preceding equations have the advantage of mathematical precision, but it is helpful to have verbal formulations of the conditions for synaptic strengthening and weakening, illustrated in Fig. 2B. Suppose an empiric synapse and a plastic synapse onto the same RA neuron are activated at the same time. By Eq. 2, the eligibility trace tends to be positive for some time after. If positive reinforcement arrives during this time interval, then W_{ij} is increased by Eq. 1. Therefore the condition for synaptic strengthening can be summarized by the following rule.

- R1: If coincident activation of a plastic (HVC → RA) synapse and empiric (LMAN → RA) synapse onto the same RA neuron is followed by positive reinforcement, then the plastic synapse is strengthened.

Now suppose that the plastic synapse is active without the empiric synapse at time t . Then the eligibility trace tends to be negative for some time after that. If positive reinforcement arrives during this time interval, then W_{ij} is decreased. So the condition for synaptic weakening is summarized by this rule.

- R2: If activation of a plastic synapse without activation of the empiric synapse onto the same RA neuron is followed by positive reinforcement, then the plastic synapse is weakened.

For negative reinforcement, the signs of the synaptic changes in R1 and R2 are reversed.⁷ Having described our model of synaptic plasticity both in equations and words, let us now examine why it is appropriate for improving performance during trial-and-error learning. First, consider the intuitive justification for R1. Activation of an empiric synapse constitutes “extra” input to an RA neuron at a particular time. Subsequent positive reinforcement suggests that this “extra” input is better. However, the activation of the empiric synapse onto that particular neuron at that time was a chance event. To consolidate this chance occurrence, which led to positive reinforcement, and thereby ensure that in future song trials that specific RA neuron fires a little extra at that specific moment in time, the plastic input synapses active at that moment are strengthened. This plasticity rule causes synaptic changes that allow modifications in song to be local both in time (during the song trajectory) and space (at a neuron level).

To understand rule R2, note that each empiric synapse has a nonzero average level of activation, which is determined by the firing rate of the presynaptic LMAN neuron. If the empiric synapse is not active at a particular time, it means that the RA neuron is receiving less input than usual for that moment in time. Subsequent positive reinforcement suggests that this deficit of input is better. This LMAN-driven chance deficit is consolidated for future trials within the HVC–RA pathway by weakening the plastic synapses that were active at that time.

⁶ In all of our simulations, except where specifically noted otherwise (Fig. 6), the reinforcement signal is assumed to be nonnegative.

⁷ Rules R1 and R2 are applicable when the plastic and empiric synapses are both excitatory, as is the case for the synapses onto RA neurons from HVC and LMAN, or are both inhibitory. These rules could also be adapted to mixed excitatory and inhibitory synapses by changing signs (Fiete and Seung 2006).

R1 and R2 describe how the presence or absence of chance LMAN input to RA, if followed by positive reinforcement, causes HVC–RA synapses to undergo either long-term potentiation (LTP, R1) or long-term depression (LTD, R2). Because the presence or absence of empiric (LMAN) input determines the sign of synaptic change when reinforcement is present, LMAN's role in the preceding rules might be mistaken as supervisory. We note, however, that in our theoretical formulation and birdsong model, output performance does not affect patterns of activity in the empiric (LMAN) input, which would be a requirement if LMAN were sending supervisory signals to RA based on output performance. Furthermore, if reinforcement is held constant, or if it varies independently of eligibility, then rules R1 and R2 produce no net (average) change in synaptic weights: over many trials, synaptic strengthening and weakening due to R1 and R2 cancels, even when LMAN is active. This can readily be seen from *Eq. 2*, where the average of synaptic eligibility alone is always zero. It is only when reinforcement actually covaries with fluctuations of the synaptic eligibility that there is a net nonzero change in synaptic weight.

Let us more closely examine how the demands of the desired trajectory—reflected in the reinforcement signal—set the balance between R1 and R2 to determine the actual direction of net synaptic change. Consider a scenario where overall performance would improve with an increase in the activity of RA neuron *A* at time *t* in the trajectory, a decrease in its activity at time *t'* in the trajectory, and be unaffected by changes in its activity at time *t''*. How do plasticity rules R1 and R2 combine to produce these changes? In this hypothetical scenario the network will tend to receive positive reinforcement in song trials where neuron *A* happens to be more active at time *t* than usual for that time, due to chance input from an empiric LMAN synapse. In trials where the empiric synapse to neuron *A* is quiescent at *t*, the network will tend to get less or no positive reinforcement because the neuron is less active than usual for that time. In short, for this scenario reinforcement is greater after empiric input to neuron *A* at time *t* than without, causing R1 to dominate over R2 and resulting in a net LTP of those regular inputs to neuron *A* that were active at time *t*. Conversely, because the trajectory would be better with less-than-usual activity in neuron *A* at time *t'*, reinforcement will be larger in trials where LMAN inputs to *A* are quiescent at *t'*, meaning that R2 will dominate and produce a net LTD of HVC synapses to *A* that were active at *t'*. Finally, because reinforcement does not depend on the activity of neuron *A* at *t''*, then reinforcement will arrive with equal likelihood after quiescence or activity in the LMAN input at *t''*, and the effects of R1 and R2 will cancel, resulting in zero average synaptic change for inputs to *A* that were active at *t''*.

Gradient learning

In the preceding text our synaptic plasticity rules were justified with intuitive arguments. They can also be understood using a formal mathematical theory developed elsewhere (Fiete and Seung 2006). Under reasonable assumptions, the rules—based on dynamic conductance perturbations of the actor neurons—perform stochastic gradient ascent on the expected

value of the reinforcement signal.⁸ The antagonism between plasticity rules R1 and R2 ensures that they compute the subtraction that is the essence of the definition of a gradient. This means that song performance as evaluated by the critic is guaranteed to improve on average. The guarantee holds even if the synapses are embedded in a network that is very complex: for example, the network may be recurrent and consist of conductance-based spiking neurons with synapses that display short-term plasticity. The guarantee is also broadly independent of model details or parameter choices.

Gradient learning can be regarded as a method for (approximately) solving a computational problem: finding a configuration of synaptic strengths that optimizes the performance of a network as evaluated by a critic. In general, this optimization problem is nontrivial. The performance of the network is determined by the collective effects of a large number of synapses and neurons. The role of any given synapse in performance may not be obvious, given that its effect may be exerted through multiple polysynaptic or even recurrent pathways involving both excitation and inhibition. Furthermore, this role may shift over time as the network changes during learning.

Is the principle of gradient learning also used by the brain? One might be skeptical that such a formal principle is relevant for neurobiology. However, gradient learning has a property that is important for brains: it is very robust. Even when properties of the actor and critic are varied, the plasticity rules are still guaranteed to improve average network performance.

The role of numerical simulations

This paper contains the results of many numerical simulations, which might seem irrelevant given that the principle of gradient learning guarantees that the plasticity rules will improve performance. Why are the simulations important? Although there are mathematical guarantees that gradient learning will improve performance, there is no assurance about how *fast* these improvements will be. If learning turns out to take longer than the lifetime of a zebra finch, then our model of learning, based on the general principle of random single-neuron experimentation and global reinforcement, could be rejected. Thus learning speed is the main issue explored in our numerical simulations. We explore how learning time scales with the number of neurons (to obtain an estimate of learning speed in a realistically sized song network) and with the precision and delay of the reinforcement signal.

Reinforcement learning in its essence is a parallel blind local search in the space of plastic parameters to climb a hill (the reinforcement function, which reflects overall performance on the desired task). The number of search dimensions equals the number of independently perturbed parameters. In algorithms based on synaptic weight perturbation (Dembo and Kailath 1990; Seung 2003), the search dimension is the number of weights, whereas in algorithms based on node perturbation

⁸ Strictly speaking, the proof of gradient ascent is for episodic learning, in which the conductance perturbations are rapidly time-varying within each episode, but reinforcement is delivered and synaptic changes are made only at the end of each episode. Our birdsong model uses on-line learning, in which in addition to rapidly evolving conductance perturbations, reinforcement is delivered continuously throughout the song, and synaptic changes are made throughout. In this case, the plasticity rules perform an approximation to stochastic gradient ascent. Also see footnote 11.

(Fiete and Seung 2006; Xie and Seung 2004), like the one proposed here, the search dimension is the number of perturbed neurons multiplied by the number of independent time steps in the trajectory. Because optimization by blind multiparameter local search is slow, reinforcement learning might similarly be too slow. Indeed, previous theoretical work on reinforcement learning algorithms shows that in certain feedforward networks, learning time scales proportionally with the number of plastic parameters or with the dimensionality of the input perturbations (Cauwenberghs 1993; Werfel et al. 2005).

Existing models of song learning are far from biologically realistic in network size, output degrees of freedom, neural dynamics, and characteristics of the reinforcement signal (temporal delay or broadening), and do not explore how convergence speed and final error would be affected if these properties were made to approach those found in the actual songbird. In fact, even in a small, simplified neural network model with small numbers of output degrees of freedom, Doya and Sejnowski (2000) reported that learning with independent random perturbations from LMAN resulted in relatively poor convergence to the tutor song. To remedy this situation, they assumed that LMAN computes and carries an instructive gradient signal for HVC–RA synaptic change, in addition to a random component. In addition, learning with a weight-perturbation scheme can be significantly slower and scale more poorly with network size than node-perturbation-like rules such as ours, as demonstrated in a network similar to the birdsong network (Werfel et al. 2005). Thus existing work provides few results on the possibility or accuracy of song learning based on uncorrelated random perturbations from LMAN in full-scale, realistic network models of birdsong acquisition.

In the bird there are as many as 8,000 RA neurons (and therefore as many potentially independent exploratory perturbations) and $20,000 \times 8,000 \sim 10^8$ plastic HVC–RA weights. We show that even in such large networks, it is possible at least in principle for independent random neural perturbation to produce biologically realistic learning.

To challenge our plasticity rules, we have made our model of song production quite complex. Unlike any existing models of sensorimotor learning in the song pathway, the model neurons in HVC and RA are biophysically realistic, generating spikes and interacting through synaptic conductances. The spiking activity of the network is converted into an acoustic signal by a simple model of the vocal organ. To further challenge our plasticity rules, we have intentionally “crippled” the critic’s reinforcement signal, to make it more difficult to learn from. The critic is modeled as a template matcher that compares the acoustic signal with a template drawn from real zebra finch song. The critic’s signal reaches the actor only after a temporal delay, is temporally imprecise, and is binary rather than analog. These features could be realistic if the critic’s signal is broadcast by secretion of a neuromodulator. The question is whether the plasticity rules will still be able to learn in a reasonable amount of time.

Although our models of song production and evaluation are highly complex, one should not forget that the underlying model of synaptic plasticity is extremely simple: it consists of the two equations (Eqs. 1 and 2). It is this simple model that is being tested herein. The complexities are there to make the test challenging.

METHODS

The model

ACTOR. In our model of song production, a model neural network controls a source–filter model of the avian vocal organ. Neurons interact through synaptic conductances and generate spikes, unlike past models based on nonspiking neurons (Doya and Sejnowski 1998; Troyer and Doupe 2000).

The network is composed of layers that represent HVC, RA, and motor neurons (Fig. 1). The connectivity of the network is feedforward, except for weak global inhibition in RA. Two output units represent motor neuron pools. They low-pass filter and sum the synaptic currents from RA, to produce a pair of time-varying control signals for the vocal organ.

In zebra finches, each RA-projecting HVC neuron generates a single burst of spikes at a stereotyped time during a song motif (Hahnloser et al. 2002). The burst onset times of the population of neurons are distributed throughout the song motif. To simulate these short bursts, we stimulate each HVC neuron in our model with a single current pulse during the song. This pattern of activity remains unchanged during learning.

Our source–filter model of the syrinx, the avian vocal organ, is mathematically similar to digital models of speech production (Rabiner and Schafer 1978). Oscillatory motions of the syrinx are driven by air flow, yielding an acoustic output of a set of harmonically related frequencies. The pitch or fundamental frequency of the harmonics is adjusted by muscles that control the tension of the syringeal fold (Goller and Larsen 1997; Suthers et al. 1999; Warner 1972; Wild 1997), whereas amplitude is partially controlled by air flow. The *source* in our source–filter model is a pulse train, yielding an acoustic output of a set of harmonically related frequencies, with pitch and amplitude controlled by the two time-varying outputs of the motor network. In the bird, the vocal tract and beak filter the broad spectral content of the syringeal output, and may also directly affect the syringeal oscillations (Beckers et al. 2003; Nowicki 1987; Suthers et al. 1999). The *filter* in our source–filter model is based on ten linear predictive coefficients, which are generated from zebra finch song recordings to produce a broad spectral envelope similar to that of real songs. For simplicity, the filter is static over the duration of the simulated song and does not change with learning.

Our use of the source–filter model is a compromise between simplicity and realism. More realistic models of the syrinx have relied on physics-based simulations (Fletcher 1988; Titze 1988), and display both quasiperiodic or chaotic behaviors. The quasiperiodic behaviors are similar to that of our source–filter model, but are much more time consuming to simulate.

CRITIC. The critic compares the pitch and amplitude of the generated song against those of the template, which is a recording of real zebra finch song, and sends a delayed comparison of the two back to the song network. At every instant in time, the error of the model song with respect to the template is computed as the sum of the squares of the pitch and amplitude differences. The critic’s signal is “crippled” in several ways to make learning more difficult and thus to test the capabilities of our model: First, the critic’s signal is binarized, rather than analog. Whenever the error is below a similarity threshold, then the critic provides a reinforcement of strength one; otherwise its signal is zero. Second, the signal is temporally delayed by 50 ms. Third, the signal is temporally broadened in some simulations.

There is a similarity threshold for each moment of song, set by the average performance at that moment in the last few trials. This adaptive threshold ensures that the critic gives positive reinforcement roughly 50% of the time. If the threshold were set improperly, then the critic would be hypercritical (never reinforcing anything) or uncritical (reinforcing everything). Our use of an adaptive threshold is similar to baseline comparison in reinforcement learning, which can result in faster learning and lower final error (Dayan 1990).

In our model, the critic's signal reaches HVC \rightarrow RA synapses after a delay of $T_{delay} = 50$ ms relative to the RA neural activities that gave rise to it. This number was inferred as follows. First, the delay from RA activity to acoustic output is estimated to lie in the range from 20 ms (Fee et al. 2004) to 45 ms (Troyer and Doupe 2000). The lower of these two numbers, when added to an estimated auditory processing delay of 30 ms (Troyer and Doupe 2000), yields $T_{delay} = 50$ ms.

In some simulations, the critic's signal is temporally broadened in addition to being delayed. This is done by low-pass filtering with a 50 ms time constant (see *Numerical details*).

EXPERIMENTER. In each time interval $[t, t + dt]$ during song, LMAN neurons fire a spike with probability $p = \lambda dt$, with firing rate $\lambda = 80$ Hz chosen to be consistent with the averaged spiking rate of putative RA-projecting single LMAN units recorded in the singing bird (Leonardo 2004). This underlying firing rate is taken to be constant throughout song and over learning. LMAN spike trains are regenerated, and thus vary, from iteration to iteration.

Synaptic plasticity

As described earlier, the reinforcement signal $R(t)$ is delayed by 50 ms after the neural events that gave rise to the song that it evaluates. Therefore reinforcement starts 50 ms after the song has begun and ends 50 ms after the song has ended. Equation 1 is applied during this period. The learning rate is $\eta = 0.0002$.

The temporal filter $G(t) = t^n e^{-t/\tau_e}$ was used in Eq. 2, with $\tau_e = 10$ ms and $n = 5$. The peak of this filter is at $T_{delay} = n\tau_e$, so the eligibility trace can be regarded as a version of the instantaneous eligibility that is delayed by $T_{delay} = 50$ ms to match the time delay in the reinforcement signal. However, to be realistic we assume that delaying the eligibility trace comes at the cost of introducing temporal imprecision. The width of the filter, defined as the time between the two inflection points flanking the delta-function response peak, is $2\sqrt{T_{delay}\tau_e}$, so a temporal imprecision of 45 ms is introduced by filtering to produce a 50 ms delay. In the simulations, the time average $\langle s_i^{LMAN} \rangle$ is computed by averaging the LMAN spike train of the current trial. It could be implemented instead by a low-pass filter at every LMAN \rightarrow RA synapse.

There is no clear experimental evidence for plasticity in the RA \rightarrow motor output connections, although it is possible these weights are also learned. In addition, the rules described in R1 and R2 could be used in the recurrent RA synapses at the same time as in the HVC \rightarrow RA synapses, and would drive gradient learning on the whole network. We have focused our attention on the HVC \rightarrow RA synapses because they are widely expected to be involved in song learning (Herrmann and Arnold 1991; Kittelberger and Mooney 1999; Mooney 1992; Sakaguchi and Saito 1996; Stark and Scheich 1997).

Numerical details

VOLTAGE AND CONDUCTANCE DYNAMICS. The membrane potentials V of all neurons in HVC and RA are governed by

$$C_m \frac{dV_i}{dt} = -g_L(V_i - V_L) - g_{E_i}(V_i - V_E) - g_{I_i}(V_i - V_I) \quad (3)$$

with intrinsic leak conductance g_L so that C_m/g_L defines the membrane time constant, and with excitatory and inhibitory synaptic conductances g_E and g_I , respectively. The reset condition is $V_i \rightarrow V_{reset}$ when V_i crosses the threshold voltage V_θ ; this threshold-reset event represents a voltage spike followed by repolarization. Following a spike in the i th neuron in HVC or RA, the synaptic activation $s_{ki}(t)$ in the synapse from neuron i to neuron k is incremented by one. Between spikes it decays with time constant τ_s

$$\frac{ds_{ki}(t)}{dt} = -\frac{s_{ki}(t)}{\tau_s} \quad (4)$$

In our simulations, $s_{ki}(t) = s_i(t)$. For notational clarity, we denote synaptic activations in HVC, RA, and LMAN by $s_i^{HVC}(t)$, $s_i^{RA}(t)$, and $s_i^{LMAN}(t)$, respectively. Note that although we have used integrate-and-fire neurons and relatively simple time courses for synaptic dynamics, the learning rule is guaranteed to perform stochastic gradient ascent on the reinforcement R even for more complicated neuron models (e.g., Hodgkin-Huxley) and synaptic time courses (Fiete and Seung 2006).

RA neurons receive excitatory synaptic inputs from HVC and LMAN, and global (recurrent) inhibitory inputs due to activity in RA.

Two nonspiking motor output units with time constant τ_m and tonic activations b_i sum the synaptic activations from RA, through a fixed set of RA-output weights A

$$\tau_m \frac{dm_i(t)}{dt} + m_i(t) = \sum_j A_{ij} s_j^{RA}(t) + b_i$$

The weights A are chosen so that the RA neurons have myotopic connections to the outputs and have push-pull control over each output.

PREMOTOR NETWORK PARAMETERS. For all HVC and RA neurons, $C_m = 1 \mu\text{F}/\text{cm}^2$, $V_L = -60$ mV, $V_E = 0$ mV, and $V_I = -70$ mV. The leak conductance is $g_L = 0.3$ mS/cm² for HVC neurons and $g_L = 0.44$ mS/cm² for RA neurons. The threshold membrane potential is $V_\theta = -50$ mV, and $V_{reset} = -55$ mV. The synaptic time constant is $\tau_s = 5$ ms for HVC \rightarrow RA, LMAN \rightarrow RA, and RA \rightarrow motor output connections. We also assume $\tau_m = 5$ ms. In all simulations, the time grain is $dt = 0.2$ ms, so Eqs. 3 and 4 are discretized, and $\delta(t - t_i^j) \rightarrow \delta_{i,t_i^j}$. There are N_{HVC} HVC neurons, N_{RA} RA neurons, and N_{LMAN} LMAN neurons in our simulations. In all cases, $N_{LMAN} = N_{RA}$. The synaptic conductances in HVC are $g_{L,i}(t) = 0$ for all neurons at all times; $g_{E,i}(t) = 0$ for all neurons at most times in the motif, except for one brief excitatory pulse of duration 6 ms and magnitude 0.13 mS/cm² per neuron per motif. The onset times for the pulses for different HVC neurons are distributed evenly across the simulated motif, and this pattern of HVC inputs stays fixed throughout learning. In RA, the synaptic conductances are $g_{E,i}(t) = 0.0024[\sum_j W_{ij} s_j^{HVC}(t) + s_i^{LMAN}(t)]$, and $g_{I,i}(t) = (0.2/N_{RA}) \sum_j s_j^{RA}(t)$ for all i . With these numerical values, the average excitatory drive to each RA neuron is approximately eightfold stronger than the average inhibitory drive from global inhibition. However, results reported here do not depend on the existence of global inhibition in RA; we have performed simulations with no inhibition in RA, and the results remain qualitatively unchanged. The HVC \rightarrow RA synaptic weights W are initialized randomly with uniform probability on the interval $[0, 1.5]$ in all the simulations shown herein. RA-output weight matrix A : half of all RA neurons, randomly chosen, project to m_1 ; the other half project to m_2 . Of the set projecting to m_1 , half the weights are of uniform strength $440/N_{RA}$ and half are $-440/N_{RA}$. Similarly, of the set projecting to m_2 , half the weights are uniformly $640/N_{RA}$ and the other half are uniformly $-640/N_{RA}$. These values were chosen to be large enough so that the maximum range of the network outputs could span the amplitudes and pitches present in the recorded tutor song. The opposing signs of the weights A to the output pools are meant to represent bidirectional muscle control from some resting position (Suthers et al. 1999)—rather than literal excitatory or inhibitory synapses. The strengths scale inversely with N_{RA} to keep the mean output drive the same when N_{RA} is varied. The baseline or "resting" values of the outputs in the absence of any drive from RA are $b_1 = 60$ and $b_2 = 40$.

All microscopic parameters such as individual neural leak conductances, time constants, and so forth are kept fixed, while scaling the size of the network and generating learning curves for the scaled

network. To do this correctly, we have to scale some other macroscopic parameters together with network size. For example, if the RA layer is scaled up by a factor of 4 in size, then all weights from RA to the motor outputs are globally scaled downward by the same factor of 4 to keep the maximum summed drive to the output units, and thus the range of allowed vocal pitch and amplitude, fixed. Such scaling is described in both the preceding and subsequent text.

The total length of the simulated song motif is $T = 300$ ms in Fig. 4. In Fig. 6, we study the effects of song length and HVC size on learning time. To make the comparison reasonable, we change song length and HVC size while keeping total HVC drive per song-moment constant, so we scale N_{HVC} with song length. Both are reduced fourfold, so $T = 75$ ms and $N_{\text{HVC}} = 180$; all other parameters are kept unchanged. In Fig. 7, we study the effects of scaling RA size on learning time. Because of the result that song learning does not depend on song length and HVC size, and because it is currently infeasible to run simulations with larger networks, both curves are trained with the short-duration song ($T = 75$ ms) with small HVC ($N_{\text{HVC}} = 180$). In one curve, $N_{\text{RA}} = 200$; in the other, RA size is increased fourfold, to $N_{\text{RA}} = 800$. N_{LMAN} and the weights A rescale automatically as described earlier. To keep the total variance of the output motor pools fixed as N_{RA} is scaled, we rescale the size of the experimental pulses from LMAN to be larger by a factor of $\sqrt{N_{\text{RA}}}$. The learning rate η is empirically adjusted in both cases to give the fastest possible stable (monotonically nonincreasing on a coarse scale) learning curves for each case.

SOUND GENERATOR. Due to the 0.2-ms time discretization used to integrate the preceding network dynamics, the outputs $m_1(t)$ and $m_2(t)$ are generated at a resolution of 5 kHz only. We linearly interpolate these output trains to generate a pair of output command signals, $\bar{m}_1(t)$ and $\bar{m}_2(t)$, sampled at 44 kHz. $\bar{m}_1(t)$ specifies the delta-pulse spacing (pitch period); for period to pulse conversion, a counter sums $1/\bar{m}_1(t)$ until it crosses 1, which triggers a pulse of duration $(1/44,000)$ s, and the counter is reset to 0. The height of each pulse is specified by the value of $\bar{m}_2(t) \times 10^{-3}$ at the time of the pulse. We use a fixed 10-parameter linear predictive coding (lpc) filter derived from a concatenated sample of three arbitrarily selected zebra finch song recordings. The filter parameters are static and do not change over the course of the song or over the course of song learning. The real part of the filtered pulse train is the student song.

CRITIC. Pitch extraction: The songs are windowed into overlapping segments by multiplication with a 300-sample (6.8-ms) Hanning window that shifts by 10 samples (0.23 ms) at a time until the entire length of simulated song is covered. To obtain a value for the pitch from each windowed segment, we compute the autocorrelation of that segment; the pitch period is assigned to be the number of samples between the highest peak (at zero time lag) and the second-highest peak, so long as this value is between 12 and 80; if outside this range, the distance to the next-highest peak is computed, until a value is found that falls in the allowed range. The middle 10 samples of the current windowed segment are assigned this value of estimated pitch. This procedure is repeated for each segment. The beginning of the first windowed segment and the end of the last windowed segment of the song are assigned the same pitch values as their closest assigned neighbors. **Amplitude extraction:** The songs are windowed into 100-sample (2.3-ms) disjoint segments. All 100 samples of each disjoint segment are assigned an amplitude of $0.3 \times \max |\text{song segment}|$. Let $p(t)$, $a(t)$ represent the student song pitch and amplitude, and let $\bar{p}(t)$, $\bar{a}(t)$ represent the tutor song pitch and amplitude.

The reinforcement signal R is computed by thresholding the delayed estimate of performance

$$R(t) = \Theta[D(t) - \bar{D}(t)] \quad (5)$$

In simulations where the evaluation signal is temporally broadened, it is low-pass filtered according to

$$[dR(t)/dt] = -[R(t)/\tau_R] + \Theta[D(t) - \bar{D}(t)] \quad (6)$$

with $\tau_R = 50$ ms.

In the preceding expressions, $\Theta[D(t) - \bar{D}(t)]$ is 0 when the performance $D(t)$ is worse than a threshold $\bar{D}(t)$, and is 1 when it is better. To mimic delays inherent in the transformation of network activity into vocal output and auditory processing, we assume that $D(t)$ is itself a delayed measure of network performance: at time t , it reflects the performance of the network outputs at $t - T_{\text{delay}}$. It is given by $D(t + T_{\text{delay}}) = -\{[\bar{p}(t) - p(t)]^2/c_p^2 + [\bar{a}(t) - a(t)]^2/c_a^2\}$ when the tutor song is nonsilent, and is $D(t + T_{\text{delay}}) = -2[\bar{a}(t) - a(t)]^2/c_a^2$ during silent intervals in the tutor song. The parameters c_p and c_a equalize the importance given by the critic to pitch and amplitude; $c_p = 60$, $c_a = 80 \times 10^{-3}$, and $T_{\text{delay}} = 50$ ms. The critic threshold $\bar{D}(t)$ adapts as the model birdsong network learns song, and is time-varying within the song. For each time t_0 in the motif, $\bar{D}(t_0)$ is obtained by linearly low-pass filtering $D(t_0)$ over the past five motif iterations. In all the simulations except Fig. 6, $\tau_R = 0$ ms; in other words, the reinforcement is delayed while eligibility is correspondingly delayed and broadened (temporally imprecise), although the reinforcement signal is not itself not broadened.

RESULTS

Our simulations used model networks of spiking neurons (Fig. 3). A layer of HVC neurons drives a layer of RA neurons. The RA neurons drive two output units, which represent the population activity of two motor neuron pools. The network controls a source-filter model of the avian vocal tract, which consists of a pulse train exciting a linear filter to yield simulated birdsong. The frequency and amplitude of the pulse train are controlled by the two output units of the network. The model network together with the model vocal tract constitute the “actor” of the actor-critic-experimenter schema.

Figure 3 (*right*) depicts the activity of the network during a simulation. Dynamical variables include the membrane voltages of HVC and RA neurons (shown), as well as synaptic conductances (not shown). Spikes in these neurons are modeled using a leaky integrate-and-fire mechanism. Because the output units represent the population activity of motor neuron pools, rather than single neurons, they are nonspiking, carrying signals that vary smoothly in time.

The songs of the model network before and after learning are compared in Fig. 4A. The network learned to approximate the song template shown in Fig. 4A (*left*), which was a 300-ms segment of song recorded from a real zebra finch. Before learning, the simulated song looks nothing like the template. After learning, the simulated song is a good approximation to the template (sound files included in Supplemental Materials).⁹

Before learning, the strengths of the synapses from HVC to RA were initialized randomly. During the learning process, the strengths of these synapses were changed according to *Eqs. 1* and *2*. The spatiotemporal pattern of HVC neural activity was assumed to remain constant. Changes in the synapses from HVC to RA caused the formation of a “premotor map” that translates HVC spiking into a sequence of vocal commands appropriate for generating song.

Dynamics of learning

The start and end of the learning process are depicted in Fig. 4A. The process did not occur suddenly, but rather happened

⁹ The online version of this article contains supplemental data.

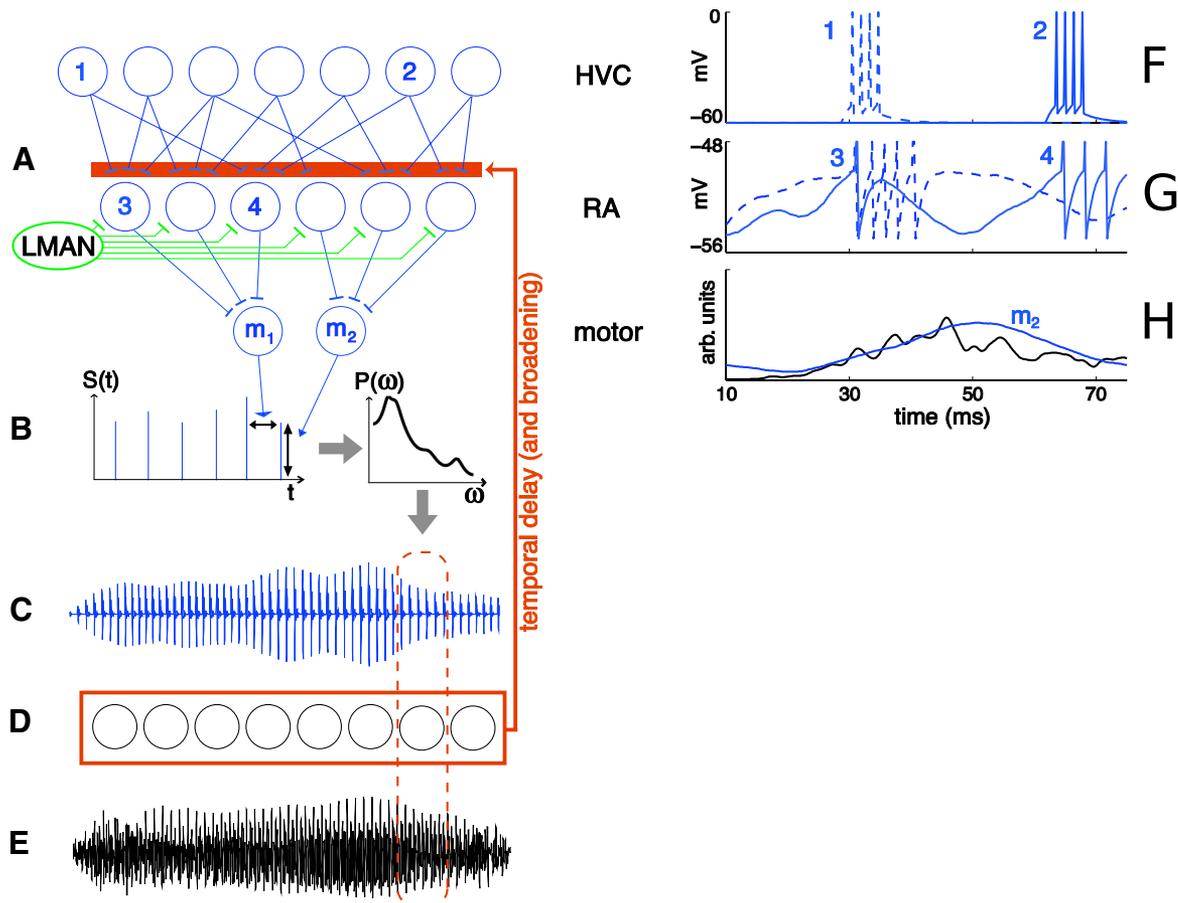


FIG. 3. A spiking network model of birdsong learning. HVC and RA are layers of integrate-and-fire neurons (A), controlling the pitch and amplitude of a simple model of avian vocalization (B). LMN sends empiric synapses to RA neurons. Acoustic output signal (C) is compared by a critic (D) to a template (E) that is a recording of an actual zebra finch “tutor” song. When the match is good, the critic signals the plastic synapses in the RA layer, which change their strengths according to Eq. 1 given in the text. F: activity of 2 typical model HVC neurons, driven by brief current pulses, shown for a segment of the simulated song motif. G: activity of 2 typical model RA neurons, receiving HVC and LMN inputs, after 1,000 iterations of learning (more traces are available in APPENDIX A). H: amplitude of the tutor song (black) and activity in the motor output controlling model song amplitude (blue).

incrementally. The network generated simulated songs for thousands of trials. During each trial, it received reinforcement signals from the critic, which compared the simulated song with the song template. Whenever the match between the two was good, the critic sent a positive reinforcement signal. This happened many times per song because the critic continuously evaluated the song throughout each trial. Because the threshold for good performance was set by the average over recent trials, the threshold became higher as performance improved.

The “learning curve” of Fig. 4B is a graph of song error versus the number of trials. This error is the mismatch in pitch and amplitude between the simulated song and the real song. It starts high and then converges to a low value within about 2,000 iterations. Is this convergence time fast or slow? It has been estimated that a juvenile zebra finch may practice its song up to 100,000 times over the course of learning (Johnson et al. 2002). Therefore the model learns relatively quickly, compared with a real zebra finch. As will be seen later, the learning time of the model may change if the properties of the reinforcement signal are changed.

After convergence there is a residual error that does not vanish. The residual could arise from several sources. First, the network may have converged to the vicinity of a local minimum of the error, rather than a global minimum. Second, even

a global minimum might have nonzero error. Third, even if the network converged to a global minimum, such convergence would be probabilistic. As long as the synaptic strengths are governed by the learning rules, they would continue to fluctuate around their optimal values. Fourth, even if the synaptic strengths were frozen at their optimal values, the simulated song would fluctuate randomly because the network continues to be perturbed by random synaptic input from LMN from trial to trial.

RA size

If many (N) neurons collectively drive the output of a network, the share of any one neuron’s activity in the total output and reinforcement is small ($\sim 1/N$). If all neurons fluctuate independently and simultaneously, any one neuron’s contribution to the overall output fluctuations is swamped by all other neural contributions. A neuron would have to correlate its own activity with the output for many trials to determine the sign of its effect on the output. Therefore when learning is based on the correlation of individual neural fluctuations with a global reinforcement signal in large networks, learning may be expected to be quite slow.

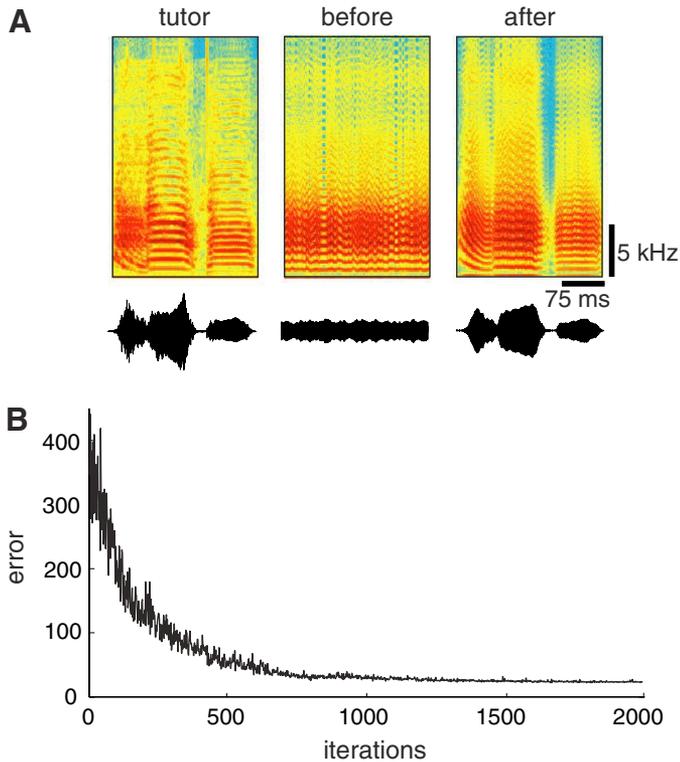


FIG. 4. Song spectrograms, song pressure waves, and the learning curve. *A*: song spectrograms and sound pressure waves. *Left*: template is a 300-ms recording of an actual zebra finch song. *Middle*: before learning, the song is a harmonic stack with randomly varying pitch and amplitude. *Right*: at 1,200 iterations, the model produces a reasonable copy of the template song. *B*: time course of song learning in the model song network. Song learning has neared its asymptotic value at around 1,000 iterations.

In the simulations of Fig. 4, our model learned song substantially faster than a real zebra finch. However, the model network was composed of just 720 HVC neurons and 200 RA neurons. The HVC and RA of a real zebra finch are estimated to contain about 20,000 HVC neurons and 8,000 RA neurons, or 10–100 times more neurons and 500–5,000 times more synapses than in the model. Each RA neuron receives parallel, independent, time-varying perturbations from LMAN. RA neural activities sum to drive the motor pools; thus correlations between conductance fluctuations in a single RA neuron with the reinforcement signal diminish with increasing RA size. What is the learning time in a realistically large birdsong network? Unfortunately, numerical simulations of a model network of this size are currently impractical. Instead, we have taken the approach of varying the size of HVC and RA in our model to empirically determine how learning time scales with network size. This allows us to extrapolate learning time for network sizes larger than we can simulate.

We performed numerical simulations to investigate the dependence of learning time on RA size. Figure 5*B* shows that the learning curve changes little even if RA size is increased by a factor of 4.

This result in the full spiking network is consistent with analytical and numerical results in a reduced model of the birdsong network (APPENDIX B). We find that in a network of linear neurons, if reinforcement is computed relative to a baseline and if RA–output connections are myotopic (each RA neuron projects to just one motor pool), then learning time is

independent of the size of the RA layer and thus also does not increase with the number of independent perturbations injected into the system.

These results may be surprising, when compared with theoretical studies indicating that the learning time for a feedforward network can scale linearly with its size, if trained by a reinforcement learning algorithm (Cauwenberghs 1993; Werfel et al. 2003).

Why is it that learning does not slow down with increasing RA size? In the birdsong network, individual RA–output (and thus RA–reinforcement) correlations do diminish with RA size. If the learning problem depended on each HVC–RA synapse attaining a specific desired value, learning would indeed have slowed down considerably. However, what matters for song production is the summed output from several RA neurons to each motor pool, not the individual contribution of

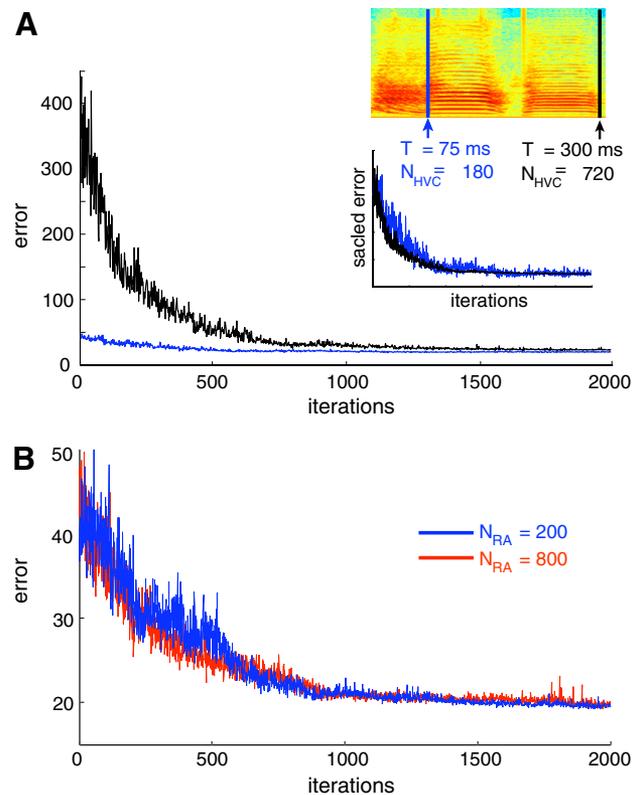


FIG. 5. Learning time does not scale with network size. *A*: learning time is independent of HVC size and the length of trained song. Learning curves for the model with a long song and large HVC (black), and 4-fold shorter song and smaller HVC (blue). *Inset 1*: tutor song spectrogram. First, the model song network is trained on 300 ms of tutor song, with 720 RA-projecting neurons in HVC. Data are the same as in Fig. 4. Next, the HVC size of the model network and the length of the template song are reduced 4-fold, to 180 neurons and 75 ms, whereas all other parameters of the network are unchanged. HVC size and song length are scaled together to keep the summed HVC drive at each moment of song fixed. *Inset 2*: to better compare the time course of learning, both curves are shifted to zero baseline error, then scaled so their initial errors match. Time course of learning is the same for these 2 cases, despite the 4-fold change in trained song length and HVC size. This happens because moments of song separated by ≥ 50 ms are produced and learned in parallel by independent sets of HVC \rightarrow RA synapses. *B*: learning time does not scale as a function of RA size. Learning curves for 200 (blue) and 800 (red) RA neurons. Larger network learns at roughly the same rate as the smaller because its extra neural degrees of freedom are redundant for the fixed task, and therefore do not slow down learning. Shorter 75-ms song template fragment from *A* was used for both curves.

each RA neuron. Consequently there are many configurations of synaptic strengths that will lead to good performance. In other words, the model network is a degenerate or redundant representation. Because it is so large, it has more neurons than necessary to perform the task. Thus although there are more synaptic strengths to learn in a large network, each can be learned more sloppily. These two effects compensate for each other, so that learning time is unchanged.

HVC size and song duration

In Fig. 5A, the dependence of learning time on HVC size is addressed. In our model, HVC size is equivalent to song duration. This is because each HVC neuron bursts only once during song [in accord with experimental findings (Hahnloser et al. 2002)], and a fixed number of HVC neurons is assumed to be active at any given moment. Therefore we have scaled song duration in tandem with HVC size.

Learning curves for two model networks are shown. The first network has 720 HVC neurons and is trained on 300 ms of song. The second network has 180 HVC neurons and is trained on 75 ms of song. The learning curves look about the same. This suggests that learning time is independent of HVC size/song duration.

What is the reason for this independence? Because each HVC neuron bursts only once during song, moments of song separated by ≥ 10 ms are driven by completely separate sets of HVC neurons. Further, the critic evaluates each moment of song, delivering its evaluation continuously in time. This means that the learning of each moment of song occurs independently and in parallel. As a result, when measured in number of trials, learning time has no dependence on HVC size/song duration.¹⁰ If the critic delivered a single evaluation for the whole song rather than separate evaluations for each moment,¹¹ then we expect that learning time would become dependent on song length. However, we find it plausible that the critic compares song output with the template continuously throughout time.

Analytical and numerical results in a reduced model of the birdsong network (APPENDIX B) are consistent with the full spiking model network results. In the reduced model as in the

¹⁰ This argument is not strictly valid for very short song durations because the critic's signal is temporally broadened and delayed by 50 ms, causing coupling between song moments sufficiently close in time. However, the statement is true for song moments separated in time by >50 ms. Thus once song is longer than about 50 ms, increasing HVC size/song duration does not affect learning time.

¹¹ Learning will still converge, albeit with differences in learning time and residual error. In fact, the proof that the synaptic plasticity rules perform stochastic gradient ascent on the reinforcement (Fiete and Seung 2006) is strictly for learning in batch mode, in which reinforcement arrives at the end of the entire trajectory, and weight updates are performed at that point. Extending such a batch learning algorithm to an on-line setting by using a discounted, on-line eligibility trace, as we have done here, is in general a heuristic technique that results in a biased estimate of the gradient of the expected value of reinforcement R (Baxter and Bartlett 2001). If network dynamics are Markov, with a mixing time that is short compared to the memory of the eligibility trace, then the bias may be small. Our birdsong model with on-line feedback learns the feedforward $HVC \rightarrow RA \rightarrow$ motor weights in a network that, except for weak global inhibition in RA, is essentially feedforward (HVC dynamics may well be fully recurrent, but HVC activity is an input to the $HVC \rightarrow RA \rightarrow$ motor network). Given the HVC input, the network mixing time is very short. That is why on-line learning may provide an accurate approximation to gradient following.

spiking network, increasing song length/HVC size has no effect on learning time. This is true only if reinforcement is delivered on-line, and if HVC activity is unary, with each neuron firing exactly once per motif. We find that if the encoding of different time steps in HVC is statistically orthogonal but not unary, learning time will grow linearly with the number of HVC neurons.

Number of muscle groups or output degrees of freedom

It is difficult to systematically vary the complexity or dimensionality of the model sound generator, which uses two network-driven control variables (pitch and amplitude) to produce output sounds that can resemble a recorded finch song. We would encounter the same difficulty if the sound generator were constructed from physics-based parameterized models of the songbird syrinx (Elemans et al. 2004; Fletcher 1988; Titze 1988). Instead, in a reduced model of the song network (APPENDIX B), we can systematically vary the number of output units that independently contribute to performance and thus to the reinforcement, and analytically compute the dependence of learning time on the number of output degrees of freedom.

In this complementary approach (APPENDIX B), we find that learning time grows linearly with the number of outputs that must be independently controlled and that independently contribute to the reinforcement signal. In contrast with scaling of the RA layer, doubling or quadrupling the number of independent outputs affects network size only slightly because the total number of outputs is small compared with the total number of neurons in the song network. However, scaling the number of outputs makes the task more complex: there are now additional, nonredundant degrees of freedom that need to be learned.

Extrapolating from this result, we expect that if the number of output pools were increased from two in our present model to eight, the estimated number of independent muscle groups controlling song production, then learning time should increase by a factor of 4. The estimated learning time with this scaling is still well within the 100,000 trials made by zebra finches before their songs crystallize (Johnson et al. 2002). Because the present model can learn to produce a good imitation of the tutor song in far fewer iterations than actual finches, there is much room for the addition of complexity in the generation of song from motor output pools.

Temporal delay of reinforcement

In our simulations, we have assumed that the reinforcement signal arrives in RA 50 ms after the neural activity that caused the moment of song that it evaluates (see METHODS for a justification of this number). To learn from the reinforcement signal, a plastic synapse must "remember" whether it was activated and whether there was input from an empiric synapse 50 ms ago. In the model, this is enabled by a lingering eligibility trace at each synapse. In the spirit of handicapping the learning process, we have assumed that the eligibility trace maintains information over timescales of 50 ms, but the information is also temporally imprecise over this timescale (this is implemented in the simulations with a low-pass filter). The temporal imprecision of the eligibility trace leads to a significant slowing of learning, and this slowdown has been included

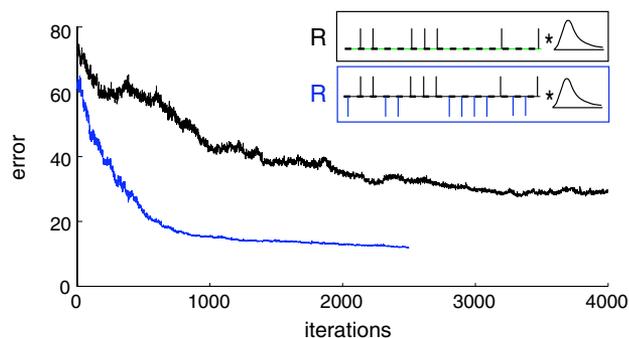


FIG. 6. Learning time and the reinforcement signal. Learning time and baseline error increase with temporally imprecise reinforcement and decrease when reinforcement has a small mean. In all preceding simulations, the reinforcement signal of 0's and 1's is delayed but temporally precise. If it is temporally broadened by 50 ms to mimic the effects of a temporally imprecise neuromodulatory signal, then learning suffers a significant slowdown even after the learning rate parameter is adjusted to find the fastest stable learning curve (*top learning curve*, black). Baseline error also increases. Temporally broadened reinforcement delivers less-specific information about song performance, further exacerbating, by a roughly equal amount, the temporal credit assignment problem already incurred in all previous simulations, from broadening of the delayed synaptic eligibilities. If the song evaluation of 0's and 1's (*top box*) is translated to -1 's and 1's (*bottom box*) before being temporally broadened, the resulting reinforcement signal has a far smaller mean value. *Bottom learning curve* (blue) shows the dramatic effects of this simple mean-subtraction operation on the reinforcement signal: learning time even with temporally broadened reinforcement is fast, converging in about 1,000 iterations and the baseline error is in fact lower than past simulations where the reinforcement signal was temporally precise but consisted of 0's and 1's—compare with learning curves from Fig. 5, *A* and *B*.

in all our learning curves. The slowdown is similar to the one that additionally results from temporal imprecision in the reinforcement signal, which is discussed next.

Temporal precision of reinforcement

The reinforcement signal would be temporally imprecise, for example, if it were delivered as a neuromodulator. Temporal imprecision reduces the amount of useful information about song performance available at any given time, and would tend to slow down learning. Therefore we repeated the numerical simulations of Fig. 4, this time assuming that the raw reinforcement signal is not only delayed and sent through a binary bottleneck as before, but also low-pass filtered in time by 50 ms. In the presence of such temporal imprecision, the network learns in about 3,000 iterations (Fig. 6, *top*, black curve). Because this is still much faster than in a real zebra finch, we expect that our model could tolerate even more temporal imprecision in reinforcement.¹²

From our analysis in a reduced model of the birdsong network (APPENDIX B), we find that increasing temporal imprecision in the reinforcement should lead to linearly proportional growth in learning time.

The true temporal imprecision of neuromodulator signals is still not really known. Neuromodulator concentrations have been observed to decay with time constants as short as 100–500 ms (Suaud-Chagny et al. 1995). However, the effective

time constant of a neuromodulatory signal could be shorter, if the molecular machinery for its detection were attuned to transients.

Subtracting a reinforcement baseline

In our simulations, we have assumed that the raw reinforcement signal from the critic is binary (0 or 1), before temporal broadening by the low-pass filter. This restriction was made to handicap the learning (it is easier to learn from an analog reinforcement signal because it contains more information).

However, if the critic is allowed to output both positive and negative values (± 1), learning can be faster, as demonstrated in the bottom curve of Fig. 6. The critic's 0–1 signal has been replaced by a ± 1 signal, before the low-pass filtering. (Subtracting a constant baseline from the filtered reinforcement signal to center it about zero is equivalent.) Learning time and residual error are significantly reduced.

It is unknown whether the critic in the avian brain produces negative, nonnegative, or both positive and negative reinforcement. Our results indicate that reinforcement of only a single sign leads to substantially slower learning. This is in accord with previous studies of reinforcement baseline subtraction (Dayan 1990).

DISCUSSION

We have proposed a model of birdsong learning based on the idea that LMAN dynamically perturbs the conductances of RA neurons to generate song variability that is used for trial-and-error learning. Our model is based on synaptic plasticity rules that estimate the gradient of the expected value of a reinforcement signal provided by a critic. In numerical simulations, the plasticity rules enable a spiking network of model neurons to learn to sing a tutor's song. To make learning as challenging as possible for our hypothetical plasticity rules, the reinforcement signal is delayed, temporally imprecise, and binarized. Furthermore, the eligibility trace at each synapse is assumed to maintain information about recent activity in a temporally imprecise way. In spite of these handicaps, the learning proceeds rapidly relative to a real zebra finch. In our simulations, learning time has little dependence on the number of HVC and RA neurons. Our results on birdsong acquisition show that reinforcement and experimentation even on the level of single neurons could be used for trial-and-error motor learning.

Predictions for synaptic physiology

If the location of the hypothetical critic or the identity of the reinforcement signal were identified in the avian brain,¹³ rules R1 and R2 could be tested directly through experiments. We predict that bidirectional long-term plasticity at HVC \rightarrow RA synapses is inducible by manipulation of three signals (HVC, LMAN, and reinforcement). Coincident activation of HVC and LMAN inputs onto an RA neuron followed by reinforcement should produce long-term potentiation of the activated HVC \rightarrow RA synapses (rule R1). Activation of HVC inputs alone followed by reinforcement should cause long-term depression of

¹² The only complication is that the residual error after convergence becomes higher. It may be possible to fix this by gradually annealing the learning rate η to smaller values during the learning process. Alternatively, mean-subtracting the reinforcement signal after temporal broadening can sharply reduce learning time and residual error, as subsequently described.

¹³ RA responds to neuromodulatory inputs such as acetylcholine (Shea and Margoliash 2003) and norepinephrine (Mello et al. 1998; Solis and Perkel 2006). Alternatively, reinforcement may be conveyed through ordinary glutamatergic synaptic transmission.

the activated HVC \rightarrow RA synapses (rule R2). Intriguingly, thus far no one has found a reliable protocol for inducing long-term plasticity of these synapses *in vitro* (Mooney, private communication). Rules R1 and R2 suggest a new heterosynaptic induction protocol, based on computational principles, that could be tested experimentally.

The plasticity rules require that a plastic synapse (HVC \rightarrow RA) be able to detect the activation of an empiric synapse (LMAN \rightarrow RA) converging onto the same neuron. In other words, LMAN input must have some special property that makes it distinguishable from HVC input, by the postsynaptic RA neurons. At the present time, based on limited experimental investigations, we can only speculate as to what that property might be. One notable difference between the two inputs to RA is that LMAN synapses have a much higher ratio of NMDA to AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid) currents than do HVC synapses (Mooney 1992; Stark and Perkel 1999). Furthermore, 70% of LMAN synapses are on shafts (Canady et al. 1988), whereas HVC synapses terminate almost exclusively on dendritic spines.

Simulated LMAN lesion/inactivation

In our model, LMAN neurons fire at a constant rate, so that the average activation of the empiric synapse $\langle s_i^{\text{LMAN}} \rangle$ is fixed in time. A biological implementation of the plasticity rules would require a mechanism that estimates the average $\langle s_i^{\text{LMAN}} \rangle$. A simple possibility is just low-pass filtering of $s_i^{\text{LMAN}}(t)$ with a long time constant. Then, as long as the underlying LMAN firing rates vary slowly, the plasticity rules should satisfactorily estimate the gradient. Another possibility is that the signal $s_i^{\text{LMAN}}(t)$ is subjected to some kind of high-pass filtering before it affects the plasticity mechanisms, which is basically the same as subtracting $\langle s_i^{\text{LMAN}} \rangle$. Some examples of almost ideal high-pass filtering are known in biology (Alon et al. 1999).

The model can be used to predict the effects of removal of LMAN input from RA. The conductances $s_i^{\text{LMAN}}(t)$ of the empiric synapses vanish in the eligibility trace. At first, the eligibility trace is expected to become negative, by Eq. 2. If $\langle s_i^{\text{LMAN}} \rangle$ is computed by low-pass filtering, it should also vanish. When both $s_i^{\text{LMAN}}(t)$ and $\langle s_i^{\text{LMAN}} \rangle$ are zero, the eligibility trace vanishes completely and all plasticity comes to a halt.¹⁴

This feature of song learning in our model is consistent with behavioral studies that show stoppage of learning and premature song crystallization after LMAN lesions in juvenile birds (Bottjer et al. 1984; Scharff and Nottebohm 1991). It is also consistent with LMAN lesion studies in the adult: deafening induces plasticity in the song system, presumably due to the effects of altered auditory feedback on the bird's evaluation of its song performance, but LMAN lesions abolish such plasticity (Brainard and Doupe 2000b).

¹⁴ This argument was based on the mathematical formulation of our plasticity rules. In terms of the verbal formulation of the learning rules, R1 immediately becomes inoperative without LMAN spiking, but R2 is still operative. However, the amplitude of the weakenings induced by R2 decreases to zero as $\langle s_i^{\text{LMAN}} \rangle$ goes to zero. Therefore, R2 eventually becomes inoperative when deprived of LMAN input and all plasticity comes to a halt.

Robustness of gradient learning

We have found that learning in our model is quite robust to details of the song production network. This is because the learning rule is based on the principle of gradient search, for which such details are irrelevant. We have tried many different neural network models of song production, with widely diverging model neuron properties and synaptic organizations. In all cases, our plasticity rules drove the network in the direction of improving song performance. We found that only two aspects were important for achieving robust learning.

First, the overall scale of the synaptic changes made by our plasticity rules is an adjustable parameter, sometimes called the *learning rate*. If the learning rate is chosen too large, then performance does not improve with time. This is because gradient learning is a local search algorithm. Although a small change in the direction of the gradient is guaranteed to improve performance, a large change has no such guarantee. On the other hand, if the learning rate is too small, performance improves very slowly. We adjusted the learning rate to an intermediate value that roughly optimized the speed of learning.

Second, it was important for the critic to measure song performance relative to an adaptive threshold based on recent performance. The need for this adaptive threshold is intuitively obvious. If the threshold were fixed at a low value, then eventually the network would improve enough that its performance always exceeds the threshold. In this case, the critic would always give positive reinforcement. Its signal would become completely uninformative and the network would cease to learn. If the threshold were fixed at a high value, then the network would never exceed it. It would never receive any positive reinforcement and would be unable to learn. An adaptive threshold ensures that neither of these situations occurs. The adaptive threshold is similar to baseline comparison in reinforcement learning, which is known to result in faster learning and lower final error (Dayan 1990).

How could an adaptive threshold be implemented biologically? One can imagine a critic composed of auditory neurons with selective responses to the tutor song and the bird's own song. The preferred stimulus of each neuron is a particular moment of the tutor's song. The firing of such a neuron in response to a particular moment of the bird's own song is proportional to how closely it resembles the corresponding moment of the template.¹⁵ Then the similarity threshold is set by the degree of tutor song specificity of the neuron. For an adaptive similarity threshold, the excitability of each neuron could be homeostatically regulated so that its average firing rate remains constant even as the bird's own song improves. Thus each neuron would become more critical with improving performance, but the average rate of reinforcement would remain constant.

Why should the brain use extrinsic perturbations for trial-and-error learning?

Extrinsic sources of variability may have some advantages relative to intrinsic sources.

¹⁵ This requires that the critic keep track of time elapsed during song, which could be implemented through communication with HVC.

If experimentation is intrinsic to the actor (HVC–RA) network, for example generated solely through neural or synaptic stochasticity within neurons of the same network, then the statistics of perturbation are inextricably related to the statistics of activity in the actor network, and it is not possible to independently modulate the size and statistics of variations in the actor network without also qualitatively changing the actor activity patterns. If the experimenter is independent of and external to the actor, however, then the statistics of perturbation added to the actor may be easily modulated in time and space according to need without qualitatively affecting the dynamics in the actor network, merely by regulating the overall level of activity in the experimenter.

For example, birdsong is more variable during undirected practice than when it is directed at other males or at females. This increased experimentation in a context where immediate performance is not critical could facilitate learning and is made possible by an upward modulation of overall LMAN activity (Hessler and Doupe 1999a,b). If perturbations were solely derived from within the HVC–RA network, a similar functionality would require modulatory influences that affect the variability of activity in the network and thus of song without altering the average song trajectory—a difficult feat.

Random experimentation

Our theoretical statement—that the synaptic plasticity rules will perform stochastic gradient learning on any reinforcement signal derived from any combination of neural activities—assumed that different actor neurons receive uncorrelated empiric inputs (Fiete and Seung 2006). To fulfill this assumption in our model, LMAN neurons project to RA neurons in a one-to-one manner and the spiking of LMAN neurons is assumed Poisson. Let us examine these idealizations more carefully.

In the songbird, neuroanatomical findings suggest that although the projection from LMAN to RA is not actually one to one, it does show relatively little convergence and divergence. There are roughly 50 LMAN synapses onto an RA neuron (Canady et al. 1988; Herrmann and Arnold 1991), possibly originating from a handful of colocalized LMAN neurons (Gurney 1981). In contrast, the HVC → RA pathway has much greater convergence and divergence: an RA neuron receives up to 1,000 HVC synapses from up to 200 different HVC neurons (Kittelberger and Mooney 1999). In theory, for a fixed neural architecture where each myotopic group of RA neurons projects convergently to a single motor pool, and performance depends on the independent activity not of single RA neurons but of each motor pool, myotopically related RA neurons could receive correlated perturbations without restricting exploration in the space of song output. Thus RA neurons could in principle receive correlated empiric inputs to drive learning, so long as different myotopic groups are perturbed independently.

Next, LMAN spiking in the songbird appears to have some temporal correlations with the ongoing song. Thus it is not truly Poisson, but it is much more irregular than RA spiking. In adult birds, some systematic patterns can be detected in LMAN spike trains, although there is also a great deal of variability across bouts of song (Hessler and Doupe 1999a,b; Kao et al. 2005; Leonardo 2004). In juvenile birds, the systematic com-

ponent of these spike trains is even weaker (unpublished data). We expect our plasticity rules to be robust to the presence of weak correlations. They will probably change synapses in a direction that is similar to the gradient, although not exactly equal.

In principle, temporal structure in the experimenter might actually facilitate learning, rather than being a hindrance. For example, if an upward perturbation from LMAN to RA at t in the last song trial led to positive reinforcement, it may be advantageous for learning if LMAN activity at t in the next trial were biased in the upward direction. Such functionality would require that LMAN neurons be able to learn specific, time-varying trajectories: the same primary task and computational burden as the actor network (HVC, RA) is faced with in the first place for learning song.

Thus it is possible that LMAN may be devoted to providing not purely random, but “intelligent” perturbation to the song network, and the role of the anterior forebrain pathway may be to devise these strategies. To use terminology from machine learning, where an “adaptive critic” describes a critic that itself learns to modify its output as a function of recent external rewards to the actor network, LMAN may play the role of an “adaptive experimenter,” modifying the perturbations it injects into RA as a function of recent reinforcements. Computationally and neurobiologically, it remains to be seen whether and how LMAN might learn good perturbation trajectories. On the computational side, our model provides a conceptual and theoretical framework on which to build and explore these possibilities.

Other neural models of reinforcement learning

Many existing neural models of reinforcement learning are based on the principle of gradient estimation (Dembo and Kailath 1990; Seung 2003; Williams 1992; Xie and Seung 2004). Could any of these other models be applied to birdsong learning?

Before we discuss the differences between these models, it is best to emphasize their similarities. All models of this category assume a scalar reinforcement signal from a critic. The models estimate a gradient through the covariance between the reinforcement signal and a large number of fluctuating network variables. The models have all been criticized for similar reasons. A scalar reinforcement signal seems quite impoverished, and learning is likely to be too slow; one can imagine supervisory signals that contain more detailed information about how the network can improve its performance.

Our numerical simulations demonstrate in a model of a well-characterized premotor network, consisting of realistic neurons, driving a natural behavior, that reinforcement learning models are viable contenders for helping to explain the goal-directed learning of song in zebra finches. We show that learning time has little dependence on the number of HVC and RA neurons. We have obtained similar learning times after replacing the plasticity rules used herein with other neural models of reinforcement learning (results not shown). Therefore a major result of this paper is that it is not possible to categorically reject neural models of reinforcement learning on the grounds that they are too slow to account for biological learning.

Having emphasized the similarities between neural models of reinforcement learning, let us consider their differences in the context of birdsong. It is helpful to use two dichotomies in making distinctions between the models. First, in these models either *neurons* or *synapses* are subject to random fluctuations. Second, these fluctuations are either *intrinsic* to the neurons or synapses or they are provided by perturbations from some *extrinsic* source.

These two dichotomies define four model classes. The associative-reward inaction algorithm involves intrinsic fluctuations of neurons (Barto and Anandan 1985; Williams 1992; Xie and Seung 2004). The hedonistic synapse involves intrinsic fluctuations of synapses (Seung 2003). The Doya–Sejnowski model proposed extrinsic perturbation of synapses (Dembo and Kailath 1990; Doya and Sejnowski 1998). The model proposed herein relies on extrinsic perturbation of neurons.

As explained in the INTRODUCTION, LMAN appears to be an *extrinsic* source of fluctuations in the activity of RA *neurons*. This is why our model seems more plausible at present than the others. Note that this argument is based on empirical data specific to birdsong, rather than general theoretical grounds.

How do the synaptic plasticity rules compare with Hebbian learning?

Note that Hebbian rules have usually been formulated and tested in the context of unsupervised associative learning, based on temporally contiguous activity between pairs of neurons and, unlike reinforcement learning rules, there is no a priori reason to expect them to be capable of performing hill climbing or gradient learning on a reinforcement signal in arbitrary recurrent networks. However, it is possible to modify the synaptic plasticity rules described here to make them appear more Hebbian and to ask whether these Hebbian variants are also capable of gradient learning.

Suppose that the empiric (LMAN) input is very strong, so that activity in the empiric input always drives the target actor neuron to fire, and that the regular synapses (HVC–RA) in the actor network are very weak, so they are incapable of producing spikes in the postsynaptic actor neuron. In this limit, spikes of the postsynaptic actor neuron (RA) can be identified with activity in the empiric synapse (LMAN–RA). The heterosynaptic coincidence of empiric and regular inputs to the postsynaptic neuron required for learning in R1 is now equivalent to a coincidence of presynaptic (HVC) and postsynaptic (RA) neural activity within the actor network, making R1 Hebbian. The weakening of R2 is equivalent to a weight decay term proportional to activity in the regular (HVC–RA) synapse, weighted by average postsynaptic neural activity (RA). Within the limit of very weak synaptic connectivity within the actor and very strong empiric drive, our learning rule looks Hebbian, with a specific activity-dependent form of weight decay, and will perform hill climbing on the reinforcement signal.

Over the course of learning, however, as some of the regular synaptic weights in the network become strong enough to drive autonomous activity within the actor network, firing events in the postsynaptic neuron will begin to reflect not just empiric inputs, but regular inputs as well. The Hebbian (modified R1) component of the learning rule described earlier will drive upward weight changes even when the empiric synapse is

inactive because of coincidences in neural activity within the actor network. In other words, there will be more synaptic strengthening events than produced by the original R1, and there are no existing guarantees that learning will perform hill climbing on the reinforcement. In our numerical simulations of the model birdsong network, we tested this Hebbian variant of our rule and found that, early on, when the HVC → RA weights were weak, learning proceeded smoothly with decreasing error between tutor and model; however, in the second half of learning, when the HVC → RA weights reached a critical strength, learning became unstable and the match between tutor and model songs diverged.

It is possible to imagine other Hebb-like variants of our synaptic plasticity rule, and it is possible (but remains to be shown) that some Hebb-like learning rule may work to reduce error and drive song learning in a realistic model of the birdsong network. In the end, deciding between Hebbian and non-Hebbian plasticity rules is not possible on purely theoretical grounds, but requires synaptic physiology experiments like those described earlier to test different possible plasticity induction paradigms in the songbird.

Does birdsong learning really look like a slow climb up a gradient?

The idea of gradual trial-and-error learning might seem incompatible with recent studies of zebra finches in which the learner's pitch slowly increases to twice that of the tutor, apparently diverging from the target pitch, and then suddenly drops by half (Tchernichovski et al. 2001). Could not this sudden transition be an “aha” moment of insight? Although such sudden transitions may seem inconsistent with the gradual, roughly monotonic decrease in error of our simulated learning curves, this is not the case. What is nonmonotonic according to an artificial critic could well be monotonic when judged by a biological critic. Humans typically judge tones separated by a factor of 2 in pitch (octave) to be more similar than tones separated by a factor of 1.5. If the avian song critic is similar, it might regard an increase of pitch to twice the target value to be an improvement in performance, rather than a deterioration. Furthermore, the motor changes involved in this apparently discontinuous process may also be smooth: a sudden drop in pitch can arise from gradual changes in motor control parameters, through period-doubling bifurcation of the nonlinear oscillations of the syrinx (Fee et al. 1998).

Where is the critic?

Some have suggested that the template and critic may reside in the motor or premotor periphery. For example, if the song template is stored as a representation of the motor activations necessary to produce the tutor song, then a reinforcement signal may be generated based on the match between actual motor activity and the template. However, this scenario requires that the bird be able translate the tutor song, an auditory signal, and store it as the set of motor commands needed to produce the song accurately; then, the bird must have a mechanism for comparing the actual motor commands with the desired ones to issue a criticism. Although our model has been agnostic about identity of the critic, we find this scenario

unlikely, and instead propose that the song template and critic may be rooted in the auditory periphery.

A set of neurons with specific auditory sensitivity to both the tutor song and the bird's own song (BOS) would in principle satisfy all the characteristics necessary for both template storage and song criticism in our model. Neurons with joint tutor- and BOS sensitivity respond to current BOS, with enhanced firing whenever BOS resembles tutor song. A simple activity-dependent homeostatic mechanism (Leslie et al. 2001) could ensure that as BOS evolves to resemble the tutor song, the threshold for firing in these neurons increases, so that the neurons become activated only when the similarity between BOS and tutor song further improves.

This mechanism would predict that joint tutor- and BOS-specific auditory neurons in the juvenile songbird initially respond to a broad range of auditory stimuli, but as motor learning progresses, become more selective to only the tutor song and the current BOS, as has been observed in experiments (Nick and Konishi 2005).

Interestingly, a hypothetical critic consisting of joint tutor- and BOS-specific auditory neurons with homeostatic machinery can explain the highly counterintuitive observation that during song learning, increasing the juvenile bird's exposure to the tutor song hinders learning (Tchernichovski et al. 1999). Exposure to tutor song drives the joint tutor- and BOS-specific neurons to fire vigorously, and homeostasis would raise their thresholds. As a result, these neurons would fire less in response to BOS, even when it is good compared with recent trials. The resulting lack of positive reinforcement from these critic neurons for such trials would hinder song learning in our model.

APPENDIX A

Neural activity in the premotor network

Although HVC activity is fixed during learning, RA activity is changed by the synaptic learning rules. The changes in RA activity are appropriate for improving song performance. However, they are not explicitly constrained in any other way by the synaptic learning rules: RA activity is an "emergent" property of the model. Therefore it is interesting to examine whether RA activity in our model network resembles RA activity in birds.

Aligned to a spectrogram of simulated song after 1,200 learning trials (Fig. A1) are activities of ten randomly selected model neurons from HVC. These were explicitly specified in the model to resemble songbird HVC neural activities. From among the 200 model RA neurons, the activities of 24 are graphed in Fig. A1, C–F. The voltage traces are drawn equally from four functional groups: those that exert control, through motor projections, on pitch or amplitude, in the upward or downward directions.

These model RA activities exhibit three characteristics that are qualitatively in accord with measurements in zebra finches. 1) Activity of RA neurons is less temporally sparse compared with that of HVC neurons. 2) Activity of RA neurons is apparently uncorrelated with features of the output song. In other words, the RA activity patterns at two instants of song that are acoustically similar need not be similar. 3) Different RA neurons appear to be relatively uncorrelated with each other (Leonardo and Fee 2005), even if they belong to the same functional group, i.e., project to the same motor pool. These three properties result from the fact that motor output is driven by the summed activities of large numbers of RA neurons. A single neuron contributes only weakly to motor output, so good performance does

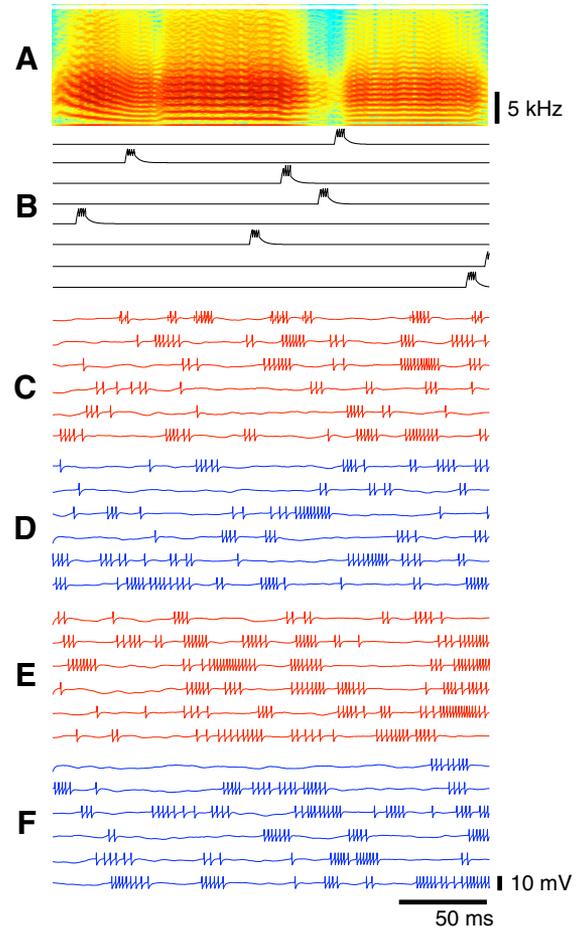


FIG. A1. Sample neural activities in the model network. A: spectrogram of the model network's output song after 1,200 iterations of learning, as in Fig. 4. B: voltage traces of HVC neurons. C–F: voltage traces of RA neurons. Spikes are omitted for better resolution of the subthreshold voltages. B: voltage traces of 10 randomly selected neurons HVC neurons; these HVC neural activities are enforced inputs to the song-learning network. C–F: voltage traces of neurons from each of the 4 RA neuron pools that project to the 2 output motor pools, after learning is complete. C: voltage trains of 6 different neurons in RA that project to the half of motor pool m_1 responsible for driving song pitch in one direction. D: traces from RA neurons projecting to the other half of motor pool m_1 , which drives song pitch in the opposite direction. E: traces from RA neurons that drive motor pool m_2 , responsible for song amplitude, in one direction. F: traces from neurons driving motor pool m_2 , and song amplitude, in the opposite direction. Simulations shown here include recurrent inhibition in RA to mimic the functional connectivity of the bird song pathway (see METHODS), but similar results are obtained if such inhibition is removed entirely (not shown).

not require the activity of any single neuron to be well correlated with song.

Although these three properties are satisfied qualitatively, quantitative aspects of RA activity are dependent on details of the simulations. For example, the final degree of temporal sparseness in RA activity depends on the initial conditions for the strengths of synapses from HVC to RA. This is because the network is a highly degenerate, or redundant, representation. There are many configurations of synaptic strengths that lead to good song performance. If there were only a single such configuration, then RA activity would be tightly constrained, but this is not the case.

Also, if the synapses from RA to motor pools were made weaker, motor activity would require a larger fraction of the RA neural population to be active at any time. This could cause their activities to be more strongly correlated with one another, and with the output.

APPENDIX B

Analysis of the scaling of learning time in a reduced model

We have used numerical simulations of spiking neural networks to investigate how learning time depends on network size and other factors. Neural nonlinearities make it difficult to derive the functional relationships between learning time and network properties. Exhaustive simulations could help illuminate certain functional dependences, but are limited to the parameter space tested. Here we follow a complementary approach, using a simplified model that is amenable to mathematical analysis. The simplified model is a linear, nonspiking network with an architecture similar to that of our spiking network model of birdsong learning. It is trained by the analogous learning rule for nonspiking networks. Learning time is provably independent of the number of hidden neurons, similar to the independence shown earlier in Fig. 5. We show that if coding in the input units (HVC) is unary, with on-line reinforcement, then learning time is independent of the number of inputs and the duration of song. If the input activity patterns were statistically independent (indeed orthogonal) in time but not unary, then learning time would grow with song length. Further calculations show that learning time is linear in the number of output neurons, and linear in temporal broadening of the reinforcement signal.

LEARNING BY NODE PERTURBATION IN NONSPIKING NETWORKS. Consider the two networks shown in Fig. B1, *A* and *B*. The left network (Fig. B1A) has two layers of synaptic weights, whereas the right network (Fig. B1B) has a single layer. The following is a description of how to train the networks using an algorithm, called *node perturbation*, that is the nonspiking network analogue to the rules R1 and R2.

Suppose that network performance is quantified by $R(x, y)$, which is a deterministic function of the input x and output y . The network has two modes of operation: noiseless and noisy. The reinforcement in the noiseless mode is R_0 , whereas the reinforcement in the noisy mode is R_ξ .

The left network computes $y_i = f[\sum_j A_{ij}f(W_{jk}x_k)]$ in the noiseless mode and $y_i = f[\sum_j A_{ij}f(W_{jk}x_k + \zeta_j)]$ in the noisy mode, where ζ_j is white noise. At each time t , an input vector $x(t)$ is drawn from a distribution. The network is run in both modes, to determine the difference in reinforcements $R_\xi(t) - R_0(t)$. Then the update

$$\Delta W_{jk}(t) = \eta[R_\xi(t) - R_0(t)]\zeta_j(t)x_k(t) \quad (B1)$$

is made. Here the weights A_{ij} are assumed to be fixed, but they could be trained similarly.

The right network computes $y_i = f(\sum_j U_{ij}x_j)$ in its noiseless mode and $y_i = f(\sum_j U_{ij}x_j + \xi_i)$ in its noisy mode. The weights are updated by

$$\Delta U_{ij}(t) = \eta[R_\xi(t) - R_0(t)]\xi_i(t)x_j(t) \quad (B2)$$

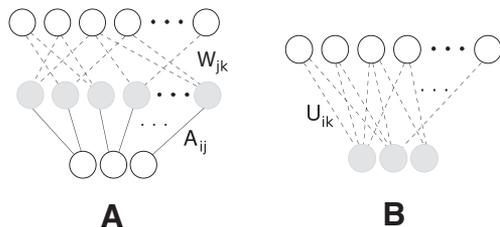


FIG. B1. Analysis of learning time in a simplified network model. *A*: network with input, hidden, and output layers. Matrix W represents the synapses from the input to the hidden layer, whereas A represents the synapses from the hidden layer to the output layer. Dotted lines: plastic weights. Gray circles: neurons that receive perturbations. *B*: equivalent network with input and output layers only, and synaptic strengths given by $U \equiv AW$. Once again, dotted lines represent plastic weights and gray circles represent the perturbed neurons.

Temporal imprecision in the reinforcement signal can also be included by modifying the learning rule to

$$\Delta W_{ij}(t) = \eta \sum_{\tau=0}^k [R_\xi(t - \tau) - R_0(t - \tau)]\zeta_i(t)x_j(t) \quad (B3)$$

Here the $\tau = 0$ term corresponds to Eq. B1. The other terms with $\tau \neq 0$ have zero mean, but slow down the learning rule by adding variance. A similar learning rule can be written for U_{ij} .

The following calculations and the scaling results subsequently derived depend on the assumption of an ideal subtraction $R_\xi - R_0$ between the noisy and noiseless reinforcements.

LEARNING TIME IS INDEPENDENT OF THE NUMBER OF HIDDEN NEURONS. We now show that training of the left network of Fig. B1 is equivalent to training the right network, provided that the neurons are linear and the numbers of input and output neurons are held fixed. This means that learning time is independent of the size of the hidden layer in the left network.

First, suppose that the hidden layer is at least as large as the output layer. Then the left network has the same representational power as the right network. To demonstrate this, note that the left network computes $y = AWx$, whereas the right network computes $y = Ux$. Therefore if $AW = U$, the networks are equivalent.

By increasing the size of the hidden layer, it is possible to make the left network much larger than the right network. The extra degrees of freedom are said to be redundant because they do not increase representational power. However, it is conceivable that they could slow down learning. The following argument proves this is not the case.

More precisely, suppose that A is fixed and satisfies $AA^T = I$. The weights W are trained according to Eq. B1. Multiplying this learning rule by A yields Eq. B2, where $U = AW$ and $\xi = A\zeta$. If ζ is white noise, and $AA^T = I$, then ξ is also white noise. Therefore the two learning rules (Eq. B1 and Eq. B2) are equivalent. In other words, node perturbation of the hidden neurons of the left network is equivalent to node perturbation of the output neurons of the right network.

Because of this equivalence, the addition of hidden neurons does not affect learning time, if the hidden layer is already as large as the output layer. The equivalence extends to the case where the reinforcement signal is temporally broadened.

LEARNING CURVES. The full learning curve of the networks can be calculated and yields the dependence of learning time on the number of input and output neurons, as well as on the temporal imprecision in the reinforcement signal.

Suppose that the network of Fig. B1B is being trained to give the same output as a “teacher” network with the same architecture and weight vector U^* . Therefore the reinforcement signal is the difference between the teacher output U^*x and the network output y

$$R(x, y) = -\frac{1}{2}|U^*x - y|^2$$

This is equal to $R_\xi = -\frac{1}{2}|(U - U^*)x + \xi|^2$ and $R_0 = -\frac{1}{2}|(U - U^*)x|^2$ in the noisy and noiseless modes of operation.

The inputs are assumed to be drawn from one of two distributions. 1) A Gaussian distribution with zero mean. These patterns are roughly orthogonal but not sparse, and have considerable overlap in their use of neurons and synapses. 2) The set of all unary input vectors (with a single component equal to one and all the rest equal to zero). These mimic sparsely active sequences similar to those found in songbird area HVC. For both of these distributions, the mean of R_0 over the input distribution is $-\frac{1}{2}\|U - U^*\|^2$, where the matrix norm is defined as $\|A\|^2 = \sum_{ij} A_{ij}^2$. This squared difference measures the performance of the learner at duplicating the teacher, and the learning rule performs stochastic gradient ascent on this function. The graph of this quantity

versus time will be called the *learning curve*, and will be calculated in the following. To simplify notation, the origin of weight space will be relocated to U^* , so that U is substituted for $U - U^*$.

Including the effect of temporal broadening of the reinforcement signal, the weight update after pattern t is given by Eq. B3

$$\Delta U_{ij}(t) = \eta \sum_{\tau=0}^k [R_{\xi}(t - \tau) - R_0(t - \tau)] \xi_{\tau}(t) x_j(t) \quad (B4)$$

A recursion relation can be obtained from this by squaring $U_{ij}(t + 1) = U_{ij}(t) + \Delta U_{ij}(t)$ and averaging first over perturbations ξ and second over inputs x , as in Werfel et al. (2003). The average over the perturbations yields

$$\begin{aligned} \langle U_{ij}(t + 1)^2 \rangle_{\xi} &= U_{ij}(t)^2 - 2\eta\sigma^2 U_{ij}(t) [U(t)x(t)]_{i,x_j(t)} \\ &+ \eta^2\sigma^4 \{2[U(t)x(t)]_{i,x_j(t)}^2 + \sum_m [U(t)x(t)]_{m,x_j(t)}^2 \\ &+ \sum_m \sum_{\tau=0}^k [U(t)x(t - \tau)]_{m,x_j(t)}^2\} + \text{const} \end{aligned}$$

where the additive constant is independent of U .

The average over the input patterns x depends on the choice between the two input distributions mentioned earlier. For uncorrelated Gaussian noise with unity variance

$$\langle \|U(t)\|^2 \rangle_{\xi x} = \{1 - 2\eta\sigma^2 + \eta^2\sigma^4[(M + 2)(N + 2) + kMN]\}^t \|U(0)\|^2 + \text{const} \quad (B5)$$

In the second case, the inputs are unary vectors, so that $x_i(t) = \delta_{i(N),t}$ where tN is defined to be t modulo N . This results in

$$\langle \|U(t)\|^2 \rangle_{\xi x} = \{1 - 2\eta\sigma^2 + \eta^2\sigma^4[(M + 2) + kM]\}^t \sum_{ij} \delta_{i(N),j} \|U(0)\|^2 + \text{const} \quad (B6)$$

BEST LEARNING TIMES AND SCALING. In Eqs. B5 and B6, the error $\|U(t)\|^2$ is exponentially decreasing, provided that the multipliers are < 1 . The rate of decrease depends on the parameter η . The optimal choice of η , defined as the value that leads to the fastest possible decrease in error per iteration, yields the learning curves

$$\langle \|U(t)\|^2 \rangle_{\xi x} = \|U(0)\|^2 \left(1 - \frac{1}{(k + 1)MN}\right)^t + \text{const} \quad \text{Gaussian inputs} \quad (B7)$$

$$\langle \|U(t)\|^2 \rangle_{\xi x} = \|U(0)\|^2 \left(1 - \frac{1}{(k + 1)M}\right)^t + \text{const} \quad \text{unary inputs} \quad (B8)$$

for large M and N . The additive constant does not depend on time and scales like $\eta^2\sigma^6$. This residual term can be made arbitrarily small by choosing a small variance (σ^2) for the perturbing test inputs. Note that the rates of convergence of the optimal learning curves are independent of the parameters σ or η .

In both cases, learning time scales linearly with the number of output neurons M and with the temporal imprecision k of the reinforcement signal. When the number of output neurons increases, the difficulty of the task increases, so it is not surprising that learning takes longer. This is different from changing the number of hidden neurons, which does not change the difficulty of the task. When there is temporal imprecision in the reinforcement signal, there is interference between the perturbations and reinforcements at different times, which slows down learning.

Why does learning time depend on the number of input neurons in the first case, but not in the second? First, we notice that in the

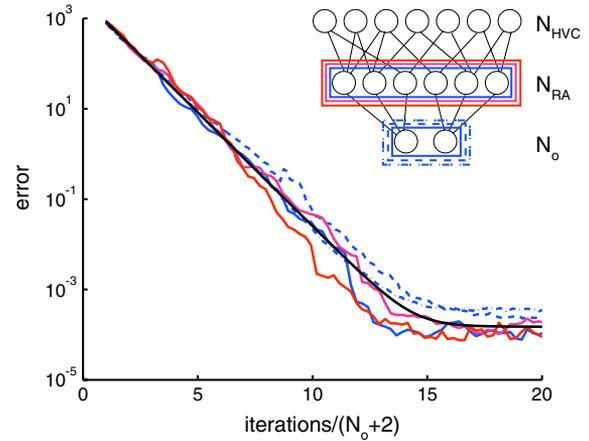


FIG. B2. Learning time in a simplified model birdsong network is independent of the size of the hidden layer and number of independent perturbations, and grows linearly with the size of the output layer. Best-case learning curves are plotted as a function of iteration number divided by $(N_o + 2)$, where N_o is the number of output neurons. Solid lines: number of output neurons is fixed ($N_o = 2$), and the number N_{RA} of hidden neurons, each receiving independent random perturbations, is scaled from 20 (magenta), to 200 (blue), to 2,000 (red). Learning time does not depend on the number of hidden neurons or independent perturbations. Blue lines: number of hidden neurons is fixed at 200, and the number of output neurons is scaled from 2 (solid blue), to 5 (dashed blue), to 10 (dot-dashed blue). Learning time scales linearly with the number of independent outputs and is proportional to $N_o + 2$. Black curve: analytical prediction for scaling with N_{RA} and N_o , derived in APPENDIX B, matches numerical simulations.

case of Gaussian input patterns, even though the patterns are statistically orthogonal, it is nevertheless the case that all input neurons and all their outgoing synapses are involved in each pattern. Thus despite the fact that the patterns are essentially orthogonal, there is still interference in the weight updates of the different synapses. In the case of binary orthogonal patterns, each neuron is used in only one pattern; the same is true for the synapses. Thus there is no interference in the learning of separate patterns and separate synaptic weights. This latter case is the relevant one for birdsong: patterns in HVC are binary and orthogonal. Individual HVC neurons, and their synapses, participate in only one part of the song, eliminating synaptic interference. This is consistent with a separate analysis of learning time and sparse firing in HVC (Fiete et al. 2004). Thus learning time does not depend on the number of HVC neurons or length of the learned song.

Our analytical calculation of the learning curves is compared with numerical simulations of learning according to Eq. B1 in the left network of Fig. B2. As shown in Fig. B2, learning time is independent of the number of hidden neurons and linear in the number of output neurons.

ACKNOWLEDGMENTS

We are grateful to R. Hahnloser and J. Werfel for discussions.

GRANTS

This work was supported by National Science Foundation Grant PHY 99-07949 to I. R. Fiete.

REFERENCES

Alon U, Surette M, Barkai N, Leibler S. Robustness in bacterial chemotaxis. *Nature* 397: 168–171, 1999.
 Barto AG, Anandan P. Pattern-recognizing stochastic learning automata. *IEEE Trans Syst Man Cybern* 15: 360–375, 1985.
 Barto AG, Sutton RS, Anderson CW. Neuronlike adaptive elements that can solve difficult learning control-problems. *IEEE Trans Syst Man Cybern* 13: 834–846, 1983.

- Baxter J, Bartlett P.** Infinite-horizon policy-gradient estimation. *J Artif Intell Res* 15: 319–350, 2001.
- Beckers G, Suthers R, ten Cate C.** Pure-tone birdsong by resonance filtering of harmonic overtones. *Proc Natl Acad Sci USA* 100: 7372–7376, 2003.
- Bottjer S, Miesner E, Arnold A.** Forebrain lesions disrupt development but not maintenance of song in passerine birds. *Science* 224: 901–903, 1984.
- Brainard M, Doupe A.** Auditory feedback in learning and maintenance of vocal behaviour. *Nat Rev Neurosci* 1: 31–40, 2000a.
- Brainard M, Doupe A.** Interruption of a basal ganglia-forebrain circuit prevents plasticity of learned vocalizations. *Nature* 404: 762–766, 2000b.
- Brainard M, Doupe A.** What songbirds teach us about learning. *Nature* 417: 351–358, 2002.
- Canady R, Burd G, Devoogd T, Nottebohm F.** Effect of testosterone on input received by an identified neuron type of the canary song system: a Golgi/electron microscopy/degeneration study. *J Neurosci* 8: 3770–3784, 1988.
- Cauwenberghs G.** A fast stochastic error-descent algorithm for supervised learning and optimization. In: *Advances in Neural Information Processing Systems*. San Mateo, CA: Morgan Kaufmann, 1993, vol. 5, p. 244–251.
- Dayan P.** Reinforcement comparison. In: *Proceedings of the 1990 Connectionist Models Summer School*, edited by Touretzky DS, Elman JL, Sejnowski TJ, Hinton GE. San Mateo, CA: Morgan Kaufmann, 1990, p. 45–51.
- Dembo A, Kailath T.** Model-free distributed learning. *IEEE Trans Neural Netw* 1: 58–70, 1990.
- Doupe A.** A neural circuit specialized for vocal learning. *Curr Opin Neurobiol* 3: 104–111, 1993.
- Doya K, Sejnowski T.** A computational model of birdsong learning by auditory experience and auditory feedback. In: *Central Auditory Processing and Neural Modeling*, edited by Brugge J, Poon P. New York: Plenum, 1998, p. 77–88.
- Doya K, Sejnowski T.** A computational model of avian song learning. In: *The New Cognitive Neurosciences*, edited by Gazzaniga M. Cambridge, MA: MIT Press, 2000, p. 469–482.
- Elemans C, Larsen O, Hoffmann M, van Leeuwen J.** Quantitative modeling of the biomechanics of the avian syrinx. *Anim Biol* 53: 183–193, 2004.
- Farries M.** The avian song system in comparative perspective. *Ann NY Acad Sci* 1016: 61–76, 2004.
- Fee M, Kozhevnikov A, Hahnloser R.** Neural mechanisms of vocal sequence generation in the songbird. *Ann NY Acad Sci* 1016: 153–170, 2004.
- Fee M, Shraiman B, Pesaran B, Mitra P.** The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird. *Nature* 395: 67–71, 1998.
- Fiete I, Hahnloser R, Fee M, Seung H.** Temporal sparseness of the premotor drive is important for rapid learning in a neural network model of birdsong. *J Neurophysiol* 92: 2274–2282, 2004.
- Fiete I, Seung H.** Gradient learning in spiking neural networks by dynamic perturbation of conductances. *Phys Rev Lett* 97: 048104, 2006.
- Fletcher NH.** Bird song—a quantitative acoustic model. *J Theor Biol* 135: 455–481, 1988.
- Goller F, Larsen O.** A new mechanism of sound generation in songbirds. *Proc Natl Acad Sci USA* 94: 14787–14791, 1997.
- Gurney M.** Hormonal control of cell form and number in the zebra finch song system. *J Neurosci* 1: 658–673, 1981.
- Hahnloser R, Kozhevnikov A, Fee M.** An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* 419: 65–70, 2002.
- Herrmann K, Arnold A.** The development of afferent projections to the robust archistriatal nucleus in male zebra finches: a quantitative electron microscopic study. *J Neurosci* 11: 2063–2074, 1991.
- Hessler N, Doupe A.** Singing-related neural activity in a dorsal forebrain-basal ganglia circuit of adult zebra finches. *J Neurosci* 19: 10461–10481, 1999a.
- Hessler N, Doupe A.** Social context modulates singing-related neural activity in the songbird forebrain. *Nat Neurosci* 2: 209–211, 1999b.
- Immelmann K.** Song development in the zebra finch and in other estrildid finches. In: *Bird Vocalizations*, edited by Hinde RA. New York: Cambridge Univ. Press, 1969, p. 61–74.
- Johnson F, Soderstrom K, Whitney O.** Quantifying song bout production during zebra finch sensory-motor learning suggests a sensitive period for vocal practice. *Behav Brain Res* 131: 57–65, 2002.
- Kao M, Doupe A, Brainard M.** Contributions of an avian basal ganglia forebrain circuit to real-time modulation of song. *Nature* 433: 638–643, 2005.
- Kittelberger J, Mooney R.** Lesions of an avian forebrain nucleus that disrupt song development alter synaptic connectivity and transmission in the vocal premotor pathway. *J Neurosci* 19: 9385–9398, 1999.
- Konishi M.** The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Z Tierpsychol* 22: 770–783, 1965.
- Leonardo A.** Experimental test of the birdsong error-correction model. *Proc Natl Acad Sci USA* 101: 16935–16940, 2004.
- Leonardo A, Fee M.** Ensemble coding of vocal control in birdsong. *J Neurosci* 25: 652–661, 2005.
- Leslie K, Nelson S, Turrigiano G.** Postsynaptic depolarization scales quantal amplitude in cortical pyramidal neurons. *J Neurosci* 21: RC170, 2001.
- Margoliash D.** Evaluating theories of bird song learning: implications for future directions. *J Comp Physiol A Sens Neural Behav Physiol* 188: 851–866, 2002.
- Marler P, Tamura M.** Culturally transmitted patterns of vocal behavior in sparrows. *Science* 146: 1483–1486, 1964.
- Mello C, Pinaud R, Ribeiro S.** Noradrenergic system of the zebra finch brain: immunocytochemical study of dopamine-beta-hydroxylase. *J Comp Neurol* 400: 207–228, 1998.
- Mooney R.** Synaptic basis for developmental plasticity in a birdsong nucleus. *J Neurosci* 12: 2464–2477, 1992.
- Nick T, Konishi M.** Neural auditory selectivity develops in parallel with song. *J Neurobiol* 62: 469–481, 2005.
- Nottebohm F, Stokes T, Leonard C.** Central control of song in the canary, *Serinus canarius*. *J Comp Neurol* 165: 457–486, 1976.
- Nowicki S.** Vocal tract resonances in oscine bird sound production: evidence from birdsongs in a helium atmosphere. *Nature* 325: 53–55, 1987.
- Olveczky B, Andalman A, Fee M.** Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLoS Biol* 3: e153, 2005.
- Perkel D.** Origin of the anterior forebrain pathway. *Ann NY Acad Sci* 1016: 736–748, 2004.
- Price P.** Developmental determinants of structure in zebra finch song. *J Comp Physiol Psychol* 93: 268–277, 1979.
- Rabiner L, Schafer R.** *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
- Reiner A, Perkel D, Mello C, Jarvis E.** Songbirds and the revised avian brain nomenclature. *Ann NY Acad Sci* 1016: 77–108, 2004.
- Sakaguchi H, Saito N.** Developmental changes in axon terminals visualized by immunofluorescence for the growth-associated protein, gap-43, in the robust nucleus of the archistriatum of the zebra finch. *Brain Res Dev Brain Res* 95: 245–251, 1996.
- Scharff C, Nottebohm F.** A comparative study of the behavioral deficits following lesions of various parts of the zebra finch song system: implications for vocal learning. *J Neurosci* 11: 2896–2913, 1991.
- Seung H.** Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron* 40: 1063–1073, 2003.
- Shea S, Margoliash D.** Basal forebrain cholinergic modulation of auditory activity in the zebra finch song system. *Neuron* 40: 1213–1226, 2003.
- Simpson H, Vicario D.** Brain pathways for learned and unlearned vocalizations differ in zebra finches. *J Neurosci* 10: 1541–1556, 1990.
- Solis M, Perkel D.** Noradrenergic modulation of activity in a vocal control nucleus in vitro. *J Neurophysiol* 95: 2265–2276, 2006.
- Stark H, Scheich H.** Dopaminergic and serotonergic neurotransmission systems are differentially involved in auditory cortex learning: a long-term microdialysis study of metabolites. *J Neurochem* 68: 691–697, 1997.
- Stark L, Perkel D.** Two-stage, input-specific synaptic maturation in a nucleus essential for vocal production in the zebra finch. *J Neurosci* 19: 9107–9116, 1999.
- Snaud-Chagny M, Dugast C, Chergui K, Mshghina M, Gonon F.** Uptake of dopamine released by impulse flow in the rat mesolimbic and striatal systems in vivo. *J Neurochem* 65: 2603–2611, 1995.
- Suthers R, Goller F, Pytte C.** The neuromuscular control of birdsong. *Philos Trans R Soc Lond* 354: 927–939, 1999.
- Suthers R, Margoliash D.** Motor control of birdsong. *Curr Opin Neurobiol* 12: 684–690, 2002.
- Tchernichovski O, Lints T, Mitra P, Nottebohm F.** Vocal imitation in zebra finches is inversely related to model abundance. *Proc Natl Acad Sci USA* 96: 12901–12904, 1999.
- Tchernichovski O, Mitra P, Lints T, Nottebohm F.** Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science* 291: 2564–2569, 2001.
- Titze I.** The physics of small-amplitude oscillation of the vocal folds. *J Acoust Soc Am* 83: 1536–1552, 1988.

- Troyer T, Bottjer S.** Birdsong: models and mechanisms. *Curr Opin Neurobiol* 11: 721–726, 2001.
- Troyer T, Doupe A.** An associational model of birdsong sensorimotor learning. I. Efference copy and the learning of song syllables. *J Neurophysiol* 84: 1204–1223, 2000.
- Warner RW.** The anatomy of the syrinx in passerine birds. *J Zool (Lond)* 168: 381–393, 1972.
- Werfel J, Xie X, Seung HS.** Learning curves for stochastic gradient descent in linear feedforward networks. In: *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2003, vol. 16, p. 1197–1204.
- Werfel J, Xie X, Seung HS.** Learning curves for stochastic gradient descent in linear feedforward network. *Neural Comput* 17: 12699–12718, 2005.
- Wild J.** Descending projections of the songbird nucleus robustus archistriatalis. *J Comp Neurol* 338: 225–241, 1993.
- Wild J.** Neural pathways for the control of birdsong production. *J Neurobiol* 33: 653–670, 1997.
- Wild J.** Functional neuroanatomy of the sensorimotor control of singing. *Ann NY Acad Sci* 1016: 438–462, 2004.
- Williams R.** Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learn* 8: 229–256, 1992.
- Xie X, Seung H.** Learning in neural networks by reinforcement of irregular spiking. *Phys Rev E Stat Nonlin Soft Matter Phys* 69: 041909, 2004.
- Yu A, Margoliash D.** Temporal hierarchical control of singing in birds. *Science* 273: 1871–1875, 1986.