Neuron, Vol. 40, 1063–1073, December 18, 2003, Copyright ©2003 by Cell Press

Learning in Spiking Neural Networks by Reinforcement of Stochastic Synaptic Transmission

Viewpoint

H. Sebastian Seung
Howard Hughes Medical Institute and
Brain and Cognitive Sciences Department
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Summary

It is well-known that chemical synaptic transmission is an unreliable process, but the function of such unreliability remains unclear. Here I consider the hypothesis that the randomness of synaptic transmission is harnessed by the brain for learning, in analogy to the way that genetic mutation is utilized by Darwinian evolution. This is possible if synapses are "hedonistic," responding to a global reward signal by increasing their probabilities of vesicle release or failure, depending on which action immediately preceded reward. Hedonistic synapses learn by computing a stochastic approximation to the gradient of the average reward. They are compatible with synaptic dynamics such as short-term facilitation and depression and with the intricacies of dendritic integration and action potential generation. A network of hedonistic synapses can be trained to perform a desired computation by administering reward appropriately, as illustrated here through numerical simulations of integrate-and-fire model neurons.

Introduction

Many types of learning can be regarded as optimizations. For example, operant conditioning can be viewed as a process by which animals adapt their behaviors so as to maximize reward. The adage that "practice makes perfect" refers to the iterative improvement of complex motor skills like playing the piano or serving a tennis ball. It is widely believed that learning is based at least in part on plasticity of the synaptic organization of the brain. Therefore, it seems plausible that there are types of synaptic plasticity that are tailored for the function of optimizing neural circuits.

What specific forms could such synaptic plasticity take? To stimulate the imagination, it is helpful to draw inspiration from evolution, the best-known example of an optimizing process in biology. A fascinating aspect of evolution is that it requires imperfect genetic replication. Such unreliability might otherwise seem undesirable, but random mutation and recombination are actually essential for generating variation, which allows evolution to search for improved genotypes.

An unreliable process also lies at the heart of neural computation: synaptic transmission. When depolarized by an action potential, a presynaptic terminal may release neurotransmitter, or it may fail to release (Stevens, 1993). At first glance, such unreliability may seem surprising and potentially detrimental to brain function. But another possibility is that synaptic unreliability is used by the brain for the purposes of learning (Minsky, 1954; Hinton, 1989), in analogy to the way in which unreliable genetic replication is used for evolution.

Here I propose a specific implementation of this idea. According to the proposal, synapses are "hedonistic," responding to a global reward signal by increasing their probabilities of release or failure, depending on which action immediately preceded reward. Remarkably, if each synapse in a network behaves hedonistically, selfishly seeking reward, then the network as a whole behaves hedonistically, learning to increase its average reward by generating appropriate collective actions. This statement can be formulated and justified mathematically and defines the sense in which hedonistic synapses serve the function of optimization.

The concept of the hedonistic synapse is potentially relevant to any brain area in which a reinforcement or supervisory signal is broadcast globally. For example, there is evidence that the neuromodulator octopamine functions as a reward signal in the mushroom bodies, a locus of olfactory learning in the insect brain (Menzel, 2001). Octopamine is delivered by the VUM_{mx1} neuron, which arborizes diffusely over the mushroom bodies. Similarly, the vertebrate striatum receives dense projections from dopamine neurons in the substantia nigra, which appear to carry a common reward signal (Montague et al., 1996; Schultz, 2002). Climbing fiber input to the cerebellum may provide an error signal for the adaptation of gaze-stabilizing behaviors such as the vestibuloocular reflex (Ito, 2001).

In brain areas that receive such a global reinforcement signal, it is plausible that synaptic plasticity is driven by interactions between the global signal and other signals that are local to the synapse. Finding the exact rules governing these interactions is an important challenge. Hypotheses like the hedonistic synapse may prove useful in the search for learning rules that combine global and local signals in the brain.

It should be noted that numerous methods of optimizing the synaptic connectivity of a model neural network have been explored in the field of machine learning. A famous example is the backpropagation algorithm, which computes the gradient of an objective function with respect to the synaptic strengths of a network (Rumelhart et al., 1986). Many alternatives to backpropagation have also been proposed (Barto et al., 1983; Narendra and Thathachar, 1989; Mazzoni et al., 1991; Williams, 1992; Jabri and Flower, 1992; Cauwenberghs, 1993; Unnikrishnan and Venugopal, 1994).

However, all of these learning rules were formulated for model networks that fail to incorporate two basic neurobiological facts. First, biological synapses are driven by presynaptic action potentials and modulate the membrane conductances of their postsynaptic targets. Second, the efficacy of synaptic transmission varies dynamically over time from spike to spike, due to short-term facilitation and depression (Thomson, 2000).

*Correspondence: seung@mit.edu

The learning rules studied in this paper are compatible with both of these features of biological synapses.

By its nature, the present work is speculative. The particular form of plasticity hypothesized here may turn out to exist in the brain. Even if it does not, the general concept of biological synapses that rely on microscopic randomness for the purposes of optimization could still be correct. One can imagine many possible realizations of the concept, of which the hedonistic synapse is just one. The wider goal of this paper is to stimulate theorists to imagine these possibilities and experimentalists to search for them.

Results

In operant conditioning, rewarding an animal tends to increase the future probability of actions that were performed immediately prior to reward. According to one interpretation, this phenomenon occurs because animals are hedonistic, or reward-seeking.

What if synapses, like animals, were hedonistic? To elaborate on the analogy, suppose that the possible "actions" performed by a synapse are only two in number. When stimulated by a presynaptic spike, the synapse either releases a vesicle of neurotransmitter, or it fails to release. A hedonistic synapse obeys the following learning rule:

> (R1) The release probability is increased if reward follows release and is decreased if reward follows failure.

The rule can be generalized to negative reinforcement, here called "punishment" for convenience:

> (R2) The release probability is decreased if punishment follows release and is increased if punishment follows failure.

To learn from reinforcement, a hedonistic synapse must maintain a record of its recent releases and failures. This is provided by a dynamical variable called the *eligibility trace* (Klopf, 1982). Learning at each synapse is driven by the product of the local eligibility trace with the global reinforcement signal (Figure 1).

The full description of the hedonistic synapse model is given in the Experimental Procedures and is somewhat more complex than described in Figure 1, as it incorporates dynamic effects such as short-term facilitation and depression. These effects are ignored for simplicity in Figures 1–4 but are included in the final example of Figure 5.

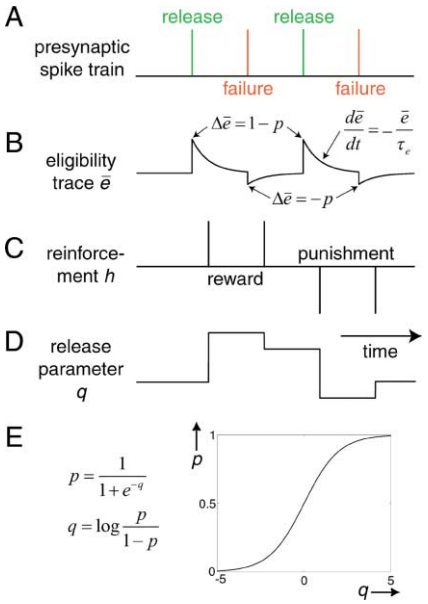

Figure 1. A Hedonistic Synapse Responds to Reinforcement by Changing Its Release Probability Based on Its Recent Actions

For simplicity, the model is depicted with the assumption that there is no short-term facilitation or depression. The full model, including short-term plasticity, is described in the Experimental Procedures. (A) Suppose, for example, that a synapse responds to four presynaptic spikes with the sequence of release, failure, release, failure. (B) The eligibility trace is a record of the synapse's recent actions. It jumps upward by $1 - p$ with every release and downward by $-p$ with every failure and otherwise decays exponentially. (C) Suppose, for example, that reinforcement is administered following each of the four presynaptic spikes. Reward is delivered twice and then punishment twice. (D) Plasticity is driven by the product of the eligibility trace $\bar{e}(t)$ and the reinforcement signal $h(t)$, as in Equation 4. (E) The release parameter q is monotonically related to the release probability p by the sigmoidal function of Equation 1, drawn here for the simple case of no short-term facilitation ($c = 0$).

Training a Multilayer Network

Now consider a network of hedonistic synapses. The network can be trained by rewarding desired behavior and punishing undesired behavior. This is illustrated in Figure 2 for the particular case of a network with a multilayer perceptron architecture. The 60 input neurons are Poisson spike trains, while the 60 hidden neurons and 1 output neuron are of the integrate-and-fire variety (see Experimental Procedures). The network was trained to perform the exclusive-or (XOR) computation on two binary variables. Each binary variable was encoded in the firing rates of half of the input neurons (Figure 2A). The XOR computation is a classic benchmark for neural network training, because a simple, single-layer perceptron cannot represent it; a multilayer perceptron is necessary.

The training was accomplished by presenting the inputs and then rewarding or punishing the synapses depending on the activity of the output neuron. More specifically, when the input was "01" or "10," the synapses were rewarded for every output spike. When the input was "00" or "11," the synapses were punished for every output spike. In other words, the reinforcement signal was either the spike train of the output neuron or its negative. Prior to training, the network responded on average with more spikes to "11" than to "01" or "10." After cycling through 200 presentations of each input pattern, the network learned to respond to "01" and "10" but to suppress almost all spiking to "11" (Figure 2B).

It is important to note that a single time-varying reward signal sufficed to train the entire network, including not only the synapses feeding directly into the output neuron but also the synapses from the input neurons to the hidden neurons. This nonlocal spread of a supervisory signal is reminiscent of backpropagation learning

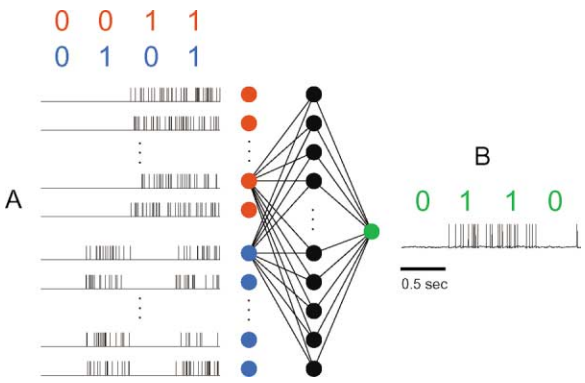

Figure 2. A Network of Hedonistic Synapses and Integrate-and-Fire Neurons Learns the XOR Function of Two Binary Variables

In the multilayer perceptron architecture, two layers of hedonistic synapses make feedforward connections between three layers of neurons: 60 input neurons (red and blue), 60 hidden neurons (black), and 1 output neuron (green). Neurons are chosen to be either excitatory or inhibitory at random. All neurons in the input layer project to all neurons in the hidden layer, though only some connections are depicted for clarity. (A) Spike trains of input neurons. The two binary variables are encoded in the firing rates of two neural populations (red and blue), each of size 30. A subset of four neurons from each population is shown. The value "1" was represented by Poisson spiking at 40 Hz, while the value "0" was represented by no spiking. In each learning epoch, the four input patterns "00," "01," "10," and "11" were presented for 500 ms each. Input spike trains are depicted for a subset of four neurons in each of two populations. (B) Spike train of output neuron (green). After learning for 200 presentations of all four input patterns, the network has learned the XOR function. It suppresses almost all output spiking in response to the input "11," while still responding to "10" or "01." More than 90% of the time, convergence to a good solution occurred within a few hundred iterations.

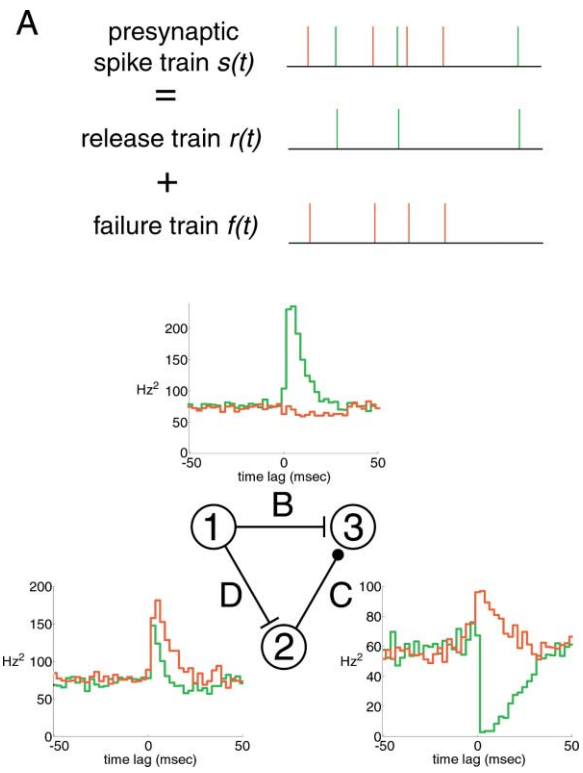

Figure 3. The Differential Correlation of Release and Failure with Reward Provides an Appropriate Learning Signal

This is illustrated through simulation of a simple circuit consisting of (1) an excitatory input neuron, (2) an inhibitory interneuron, and (3) an output neuron. The input neuron was a Poisson spike train at 20 Hz, while the other two were integrate-and-fire neurons. The release probabilities of all synapses were held fixed at 0.5, and the simulation was run for 1000 s of simulated time. The output spike train was considered as the reward. (A) Each presynaptic spike train was separated into release and failure trains. For each of the three synapses (B, C, and D) of the simple circuit, two cross-correlations are graphed. One is of the release train with reward (green), while the other is of the failure train with reward (red). By comparing these two cross-correlations, each synapse can determine whether to increase or decrease its release probability, as in Equation 11. The most interesting example is synapse D. Although it drives the inhibitory interneuron, both its releases and failures are positively correlated with reward, because the monosynaptic excitatory path B is stronger than the disynaptic path. Nevertheless, failure is more strongly correlated with reward than is release, so the synapse knows to decrease its release probability. This example illustrates that antagonistic effects of release and failure on plasticity can be important for deriving the correct learning signal.

(Rumelhart et al., 1986), but there is an important difference. The backpropagated error signal spreads by a computation that depends in a detailed way on the architecture of the network and on the precise values of its synaptic strengths. In contrast, broadcast of a reward signal to an assembly of hedonistic synapses requires communication but no computation at all.

Release-Failure Antagonism

In Figure 2, the end result of the training was that the synapses changed their release probabilities so as to increase the reward received by the network, thereby improving its performance on the XOR computation. How were the synapses able to determine the changes in release probabilities appropriate for increasing reward? The short answer is that a hedonistic synapse learns by comparing two cross-correlations, one between release and reward and the other between failure and reward. The difference between these two correlations tells the synapse how its actions are causally related to reward.

To illustrate this point, the model circuit of Figure 3 was simulated. The circuit retains some of the basic elements of the larger network trained in Figure 2. The input neuron generates a Poisson spike train, while the others are of the integrate-and-fire variety. The spike train of the output neuron is regarded as the reward signal. For each synapse, the release-reward and failure-reward correlation functions are graphed. The difference between them indicates whether the release probability should be increased or decreased to maximize reward. It is straightforward to verify this for each synapse, because the circuit is so simple.

In particular, one synapse most strikingly illustrates that the learning signal resides in the difference between the correlations rather than in either correlation alone. This is the synapse from the input neuron to the interneuron. Because release is positively correlated with reward, one might erroneously conclude that the release probability should be increased. However, the positive correlation is due to the other pathway from the input neuron to the output neuron. It turns out that failure is

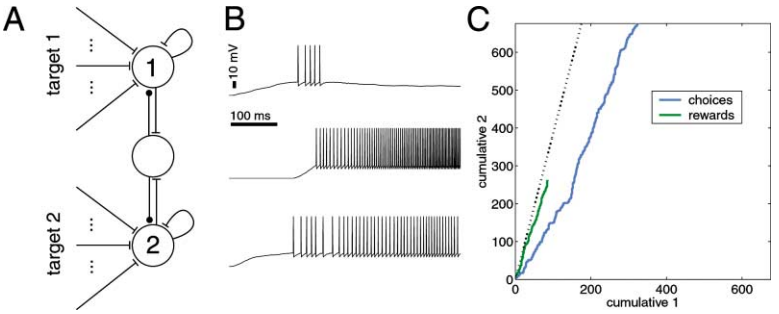

Figure 4. When a Circuit of Hedonistic Synapses Receives Probabilistic Reward for Repeated Choices between Two Targets, Its Behavior Can Obey the Matching Law Studied by Psychologists

(A) The decision-making circuit consists of three integrate-and-fire model neurons, two excitatory ("1" and "2") and one inhibitory (unlabeled). Each excitatory neuron receives 100 feedforward excitatory synaptic inputs, as well as recurrent synaptic input from itself and from the global inhibitory interneuron. During each 500 ms trial, the feedforward synapses are driven by 10 Hz Poisson spike trains. The feedforward synapses are hedonistic, while the recurrent synapses are deterministic and not plastic.

(B) In each trial, the circuit "chooses" one of the two targets, through "winner-take-all" behavior. One of the excitatory neurons is suppressed (top), while the other remains active (bottom). The winner-take-all behavior is mediated by the global inhibitory neuron (middle). The choice is probabilistic, because the synaptic input received by neurons 1 and 2 fluctuates.

(C) Before each trial, the targets are baited with probabilities $p_1 = 0.1$ and $p_2 = 0.3$. The decision-making circuit chooses one target and harvests reward if it is available. If there is unharvested reward at the end of a trial, it remains for the next trial. The circuit learns a probabilistic strategy in which target 2 is favored over target 1, as is evident in the plot of cumulative choices of target 2 versus cumulative choices of target 1 (blue). For comparison, the cumulative rewards harvested at each target are also plotted (green). The slopes of the two lines become approximately equal to each other, which corresponds to matching behavior. The dotted line indicates the theoretical slope that would hold for perfect matching.

also positively correlated with reward and even more strongly so. Hence, the release probability should be decreased to maximize reward. This example demonstrates that it can be important for release and failure to have antagonistic effects on plasticity.

While the difference between the correlations provides an appropriate learning signal, it should be noted that this signal is obscured by a learning noise, the random fluctuations corrupting any estimate of correlations based on a finite time interval. This means that dynamics of learning executes a random walk in the parameter space, which is biased in a direction that increases the average reward. A picturesque term for such behavior is "hill-climbing," which comes from visualizing the average reward as the height of a landscape over the parameter space. The formal term is "stochastic gradient ascent," as explained in the Experimental Procedures.

The Matching Law

In the simulation of Figure 2, the reinforcement was a deterministic function of the network behavior, but hedonistic learning is also applicable to contexts in which reinforcement is probabilistic. Such contexts have been studied extensively by psychologists in research on the matching law (Davison and McCarthy, 1988; Gallistel, 1994). When animals are presented with repeated choices between competing alternatives, they distribute their choices so that the returns from the alternatives are approximately the same. Return is defined as the total reward obtained from an alternative divided by the number of times it was chosen.

Figure 4 illustrates that a neural circuit containing hedonistic synapses can learn matching behavior when it chooses repeatedly between two targets that are probabilistically baited with reward. A discrete-time version of the concurrent variable-interval reward schedule is

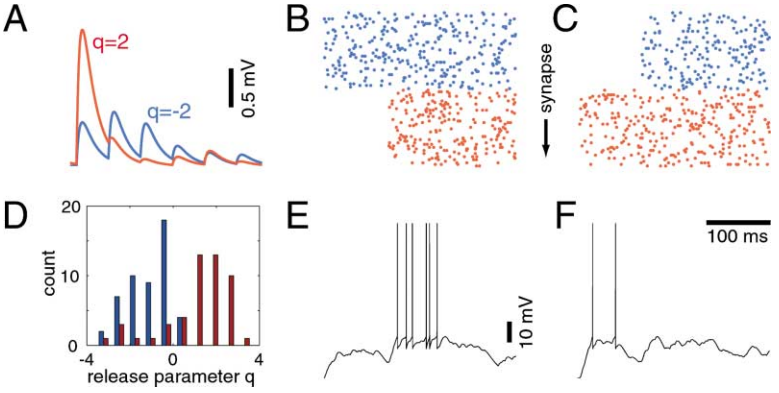

Figure 5. Hedonistic Learning Applied to Synapses with Short-Term Facilitation and Depression

A single integrate-and-fire neuron was trained to become selective to the temporal order of its 100 synaptic inputs. (A) Postsynaptic voltage response of a model synapse to 20 Hz spike trains, averaged over 100 trials. Depending on the release probability for the first spike in the train, the efficacy of the synapse exhibits different time courses. The initial release probability is specified as $1/(1 + \exp(-q))$, where q is called the release parameter. For $q = 2$ (red) the initial release probability is high and the synapses depresses. For $q = -2$ (blue) the initial release probability is low and the synapse facilitates first and then depresses. (B) The 100 synapses are driven by 20 Hz Poisson spike trains. The stimulus onset is 100 ms earlier for the blue group than the red group. Output spikes to this stimulus were rewarded during training. (C) The stimulus onset is 100 ms earlier for the red synapses than the blue synapses. Output spikes to this stimulus were punished during training. (D) All release parameters were initialized to be zero, but the two groups of synapses ended up learning different dynamical behaviors. After training with 500 presentations of each stimulus, the release parameters of the blue group were typically less than those of the red group, so that the red synapses tended to depress more strongly than the blue. (E and F) After training, the output neuron became selective to temporal order, because its synapses learned the appropriate dynamical behaviors. The mean responses to stimuli B and C were six and three spikes, respectively. Typical voltage traces are shown; the response fluctuated from trial to trial.

used, as explained in the Experimental Procedures. Whenever the decision-making circuit of Figure 4A chooses a baited target, it harvests reward, which induces plasticity at its hedonistic synapses. It chooses probabilistically between the two targets, due to the stochasticity of its feedforward synapses. Initially, it chooses with roughly equal odds, but over time it learns a preference that approximately satisfies the matching law (Figure 4C).

Dynamic Synapses

For any synaptic learning rule to be biologically plausible, it must be able to deal with the fact that the efficacy of a biological synapse is not static but changes dynamically from spike to spike (Tsodyks et al., 1998). Although this feature of synaptic transmission was neglected in the preceding examples for simplicity, it is compatible with the hedonistic synapse model. The implementation of short-term plasticity is explained in the Experimental Procedures, and its behavior is illustrated in Figure 5A, which depicts simulations of the average postsynaptic response to a tetanus of presynaptic spikes. Comparison of the two traces reveals that the probability of vesicle release to the first spike in the train affects the responses to later spikes. If the initial release probability is high, succeeding responses are depressed. If the initial release probability is low, then facilitation is visible at first, after which depression sets in. This is qualitatively in accord with measurements on biological synapses (Markram and Tsodyks, 1996).

If these different dynamical behaviors are available, hedonistic synapses can learn to use them for computation, as shown in the rest of Figure 5. The synapses onto an integrate-and-fire model neuron were divided into two groups. During training, the two groups were activated sequentially, in either order. Spikes in response to one order were rewarded, while spikes in response to the other order were punished. After training, the average spiking response of the model neuron was greater for one temporal order than for the other. Selectivity of response was achieved because the two groups of synapses learned different dynamical behaviors. One group learned to depress quickly, so that its peak response was immediate. The other group learned to facilitate, effectively introducing a time lag into its peak response. Such a time lag is known to be sufficient to give rise to selectivity to temporal ordering (Chance et al., 1998; Buchs and Senn, 2002).

Postsynaptic Voltage Dependence

At certain synapses, induction of long-term plasticity requires depolarization of the postsynaptic neuron, in addition to presynaptic spiking (Malenka and Nicoll, 1999). This property is generally regarded as the hallmark of Hebbian learning, but in fact it is also compatible with hedonistic learning.

Although the basic learning rules R1 and R2 do not depend on the state of the postsynaptic neuron (except in the special case where postsynaptic spiking is the reward signal), such a dependence is a natural modification. To understand why, recall that a hedonistic synapse compares the release-reward correlation with the failure-reward correlation in order to determine how its actions are causally related to reward. During periods when the postsynaptic neuron is far below threshold, release and failure have no effect on the rest of the network and therefore have no effect on reward. A hedonistic synapse is better off ignoring such time periods, because they contribute only learning noise and no learning signal.

Therefore, it is advantageous to modify rules R1 and R2 so that they are not applied if the postsynaptic voltage is well below threshold for a window in time surrounding the presynaptic spike. One caveat should be noted: according to this modification, plasticity at a hedonistic synapse is dependent on postsynaptic depolarization but does not require a postsynaptic spike. According to some experiments, this stronger condition is necessary for long-term potentiation (Magee and Johnston, 1997).

Postsynaptic Locus of Plasticity

It is known that long-term plasticity of a biological synapse can occur through changes in the probability of vesicle release (Dale et al., 1988; Stevens and Wang, 1994; Bolshakov and Siegelbaum, 1995). Such a presynaptic locus of plasticity is consistent with the basic model of a hedonistic synapse. However, long-term plasticity can also occur through changes in the amplitude of the postsynaptic conductance elicited by vesicle release (Malenka and Nicoll, 1999). This suggests the following modification of rules R1 and R2: make the changes in the amplitude of the postsynaptic conductance rather than in the release probability.

The modified rules will still tend to increase average reward, provided it is true that changing the amplitude of a postsynaptic conductance and changing its probability have similar effects on the behavior of the postsynaptic neuron. There are situations where this statement can be violated. For example, suppose that stimulation of a synapse drives the postsynaptic neuron above spiking threshold. Changing the amplitude of the synaptic conductance will change the timing of the spike, while changing the release probability will change the probability of the spike. If the computation of the network depends critically on precise spike timing, then these two changes could have different effects on reward.

Temporal Antagonism

The relative timing of presynaptic and postsynaptic spiking is important in the induction of long-term potentiation at hippocampal and cortical synapses (Markram et al., 1997; Bi and Poo, 1998). In particular, reversing the temporal order causes a reversal in the sign of plasticity.

Temporal antagonism can also be incorporated into the hedonistic synapse model. According to rules R1 and R2, plasticity is induced when reinforcement follows activation of the synapse. Now suppose that plasticity of the opposite sign is induced when reinforcement *precedes* activation of the synapse:

(R3) The release probability is decreased if reward precedes release and is increased if reward precedes failure.

(R4) The release probability is increased if punish-

ment precedes release and is decreased if punishment precedes failure.

As explained in the Experimental Procedures, these rules can be implemented using another variable called the *reward trace*. It is also shown that temporal antagonism can be advantageous, because it has no effect on the learning signal but can reduce the learning noise. To understand why, it is helpful to reexamine the cross-correlograms of Figure 3. Recall that the difference between the green and red correlograms is the signal that drives learning according to rules R1 and R2, due to release-failure antagonism. But there is also a learning noise, because any estimation of correlations based on a finite time interval is imperfect. Adding rules R3 and R4 has the effect of implementing a second difference operation, that between the right (positive lag) and left (negative lag) halves of each correlogram. This extra difference can help suppress the effects of fluctuations in the correlogram baselines.

Discussion

The following imaginary dialog addresses frequently asked questions about the hedonistic synapse hypothesis. Its organization should be convenient for the reader who is only interested in some of the questions.

Hedonistic synapses are just a mechanism for stochastic gradient learning, a topic that has been studied extensively in the field of neural networks. What is new here?

The present work is not intended as an advance in the mathematical theory of stochastic gradient learning. Rather, it is an attempt to make this theory relevant to neurobiology. Although the idea of gradient learning has a long history in computer science, so far it has had little or no impact on synaptic physiology. The present model is intended to be concrete enough to be experimentally testable in neural systems. Finding forms of synaptic plasticity related to gradient learning would forge a new link between neurobiology and computer science.

I'm a synaptic physiologist. How can I look for hedonistic synapses in the brain?

To look for hedonistic synapses in the brain, one must be able to manipulate a global reinforcement signal and have the capability of resolving release and failure events at one or a few synapses. If a synapse is hedonistic, the most effective way of inducing long-term plasticity would be to make reward contingent on its actions. For a hedonistic synapse, pairing release with reward would have the opposite effect as pairing failure with reward. Some plausible candidates for global reinforcement signals are mentioned in the Introduction. In addition, some examples of heterosynaptic plasticity may be ripe for explanation in terms of global reinforcement signals.

There are many sources of randomness in the brain. Why do you single out stochastic vesicle release as the basis for stochastic gradient learning?

Singling out a specific source of randomness is essential for transforming stochastic gradient learning from an abstract mathematical concept into a scientific hypothesis that can be confirmed or refuted.

While I have focused on stochastic vesicle release, I should make clear that there are many other possible ways of learning by correlating a reward signal with a source of microscopic noise. For example, learning could be based on fluctuations in quantal size that are postsynaptic in origin, leading to a learning rule that is similar to the weight perturbation algorithms studied in machine learning (Jabri and Flower, 1992; Cauwenberghs, 1993). Alternatively, learning could be based on fluctuations due to irregular action potential firing, as explored by a number of previous models (Williams, 1992; Barto et al., 1983; Mazzoni et al., 1991). On a slower time scale, stochastic gradient learning could be based on the creation and destruction of synapses.

In this paper, stochastic vesicle release was considered rather than other sources of randomness, for several reasons. The first was pedagogical: hedonistic synapses are simple and easy to understand. Second, stochastic vesicle release is a basic and universal property of chemical synaptic transmission, and the possibility of helping explain its function is surely an exciting one. Third, hedonistic synapses are generally applicable to realistic model neurons, however complex the processes of action potential generation and dendritic integration may be.

In contrast, previous proposals to base stochastic gradient learning on irregular spiking have fallen short of implementation with biophysically realistic model neurons. Instead, these proposals have modeled spiking as an intrinsically random Bernoulli (Williams, 1992; Barto et al., 1983; Mazzoni et al., 1991) or Poisson process (X. Xie and S.S., unpublished data). In contrast, biological neurons appear to fire action potentials in an essentially deterministic way, judging from studies in vitro (Mainen and Sejnowski, 1995). Irregular spiking observed in vivo is believed to arise from fluctuations in synaptic drive produced by the dynamics of neural networks (van Vreeswijk and Sompolinsky, 1998). It remains a challenge to implement stochastic gradient learning for such network models. There is strong motivation to meet this challenge, as theoretical studies suggest that learning can be faster if based on irregular spiking rather than on synaptic noise (Werfel et al., 2004).

You've demonstrated learning with hedonistic synapses for some toy problems, but it will never scale up to really large networks

This is true, but let's be fair: the difficulty of scale up is not peculiar to hedonistic synapses but is a disease common to all learning methods that use local search to optimize an objective function. Such "hill-climbing" methods as backpropagation have had impressive successes, but none has scaled up to the challenge of creating artificial intelligence that rivals the human brain in capability. At this point, no one believes that a com-

plete theory of intelligence can be based on hill-climbing alone.

Early theorists in machine learning argued that solving the scale up problem depended on finding a general method of decomposing a complex goal into simpler subgoals (Minsky, 1961). This method could be applied recursively to achieve arbitrarily complex goals. Such a "divide-and-conquer" strategy could be executed by a system composed of many agents, each of which learns from reinforcement to accomplish its subgoal. Some subgoals might be innate, but truly intelligent behavior would require the subgoals themselves to be learned.

If one takes these "society of mind" theories seriously (Minsky, 1988; Baum, 1999), then it seems plausible that the brain is a collection of tiny modules, each of which has the capability to learn from reinforcement. Each module would be a neural circuit that is small enough to be trained by a simple hill-climbing method. The modules would exchange many reinforcement signals, in a manner that is more decentralized than the way neuromodulatory systems are thought to operate.

Computer scientists may prefer to focus on the problem of how to organize a "society of mind" and endow its agents with interactions that lead to human-like artificial intelligence. But neurobiologists can proceed differently: they can try to identify the brain's hill-climbing mechanisms, without waiting for the more daunting problem of intelligence to be solved.

Even if stochastic gradient learning were applied to small neural circuits in the brain, wouldn't it still be too slow?

This issue has never been investigated in a serious way, although it deserves to be. One must identify a learned behavior that is executed by a small brain module and construct a model neural network with an architecture based on experimental data. Then one can determine the time it takes the model to acquire the behavior through stochastic gradient learning and compare this time with the learning time of the biological system. For example, the zebra finch practices its song tens of thousands of times in the course of song learning (Johnson et al., 2002). Could a network model of the avian song generation nuclei acquire song in a comparable amount of time through stochastic gradient learning? Until this type of careful comparison has been made, it seems hasty to rule out stochastic gradient learning as a brain mechanism.

What if the reward signal is delayed in time? Won't that be catastrophic for the learning time of hedonistic synapses?

A fixed delay in reward has no effect on learning, provided that the same delay is added to the eligibility trace. A variable delay is more problematic: it requires that the time constant of the eligibility trace be made as long as the delay fluctuations. This could lead to a slowdown in learning.

A modest variable delay was present in the example of learning matching behavior (Figure 4). The learning was segmented into 1 s episodes. The reward signal was delivered only at the end of the episode, while the vesicle release and failure events critical for reward occurred throughout the episode. Therefore, the delay was variable, ranging from 0 to 1 s. Accordingly, the time constant of the eligibility trace was chosen to be 1 s, and learning was able to proceed. This example shows that modest delays are not catastrophic for hedonistic learning, under certain conditions.

In general, one expects temporal delays to slow down hedonistic learning. To be fair, delays are also problematic for other types of learning. For example, delays can cause Hebbian synapses to change in the wrong direction, according to a recent model of oculomotor learning (Raymond and Lisberger, 1998).

The problem of delayed reward is of fundamental importance in the field of reinforcement learning. In games like checkers or backgammon, there is a huge temporal delay separating the win-lose-draw reinforcement signal and the moves that lead up to it. Such long delays are handled with a value function that estimates future reward, and techniques for learning the value function, such as temporal difference learning (Tesauro, 1992). Similar strategies may be implemented in the brain via the mesencephalic dopamine system (Montague et al., 1996) and could be used in conjunction with hedonistic learning to handle very long delays.

Do you believe that hedonistic synapses are the explanation of operant conditioning?

In the example of Figure 4, a simple circuit of hedonistic synapses learns matching behavior through a process resembling operant conditioning. However, this example should not be mistaken for a complete theory of operant conditioning. It is conceivable that hedonistic synapses could be sufficient to account for operant conditioning in animals with small nervous systems. But large nervous systems would rely on additional elements besides a simple hill-climbing mechanism, because of the scale up problem. In short, hedonistic synapses might be involved in operant conditioning, but they are not intended as a complete explanation of the phenomenon.

Could temporal antagonism alone be sufficient for stochastic gradient learning?

Consider learning based on rules R1 through R4, but modified by removing all dependence on failures, so that there is no release-failure antagonism. In Figure 3, this would mean that learning is driven by the green release-reward correlogram only. The sign of plasticity would be given by subtracting the left half of the correlogram (negative lag) from the right half (positive lag). For synapses B and C of Figure 3, this difference is an appropriate learning signal, but for synapse D it is not. This example shows that temporal antagonism alone is not always sufficient for extracting a proper gradient signal.

In your example, two out of three synapses would change in the right direction. Couldn't the circuit end up increasing its average reward anyway, in spite of the errant synapse?

Yes, this possibility illustrates that stochastic gradient ascent is sufficient but not necessary for hill-climbing. In stochastic gradient ascent, the average change of

every synapse is in the direction appropriate for increasing reward. A weaker condition is that the vector of average changes point within 90° of the gradient vector. This condition is typically sufficient to insure the hill-climbing property that the expected reward tends to increase. In principle, the search for learning rules should be widened to include all hill-climbing algorithms, and more research is needed along these lines.

Do hedonistic synapses require that reward and punishment be balanced so that the average reinforcement is zero?

In the simulation of matching behavior (Figure 4), only reward is administered, demonstrating that it is possible to learn from a nonnegative reward signal alone. However, there can be advantages to measuring reinforcement relative to a baseline, so the average reinforcement signal is zero. Use of a baseline has been shown to enhance the speed of learning in certain machine learning contexts (Sutton, 1984; Dayan, 1990). One could imagine that the reinforcement signal for hedonistic synapses is given by a nonnegative neuromodulatory signal minus its baseline value. The subtraction of a baseline has some relation to the temporal antagonism discussed previously.

What do you regard as the greatest weakness of your model?

In order to implement release-failure antagonism, Equation 2 is designed so that the eligibility trace has exactly zero mean. The equation requires a hedonistic synapse to keep track of its release probability, which varies dynamically because of short-term facilitation. A biological synapse might be able to implement Equation 2 approximately, but not perfectly. Violation of the zero mean condition would produce bias in the gradient estimate, which could hinder learning. Problems with bias tend to be less severe if the reinforcement signal has zero mean or temporal antagonism is added to the learning rule.

Experimental Procedures

The hedonistic synapse model is described and then derived from the general theory of REINFORCE learning (Williams, 1992; Baxter and Bartlett, 2001). Conditions under which learning by hedonistic synapses approximates stochastic gradient ascent are discussed, along with its relationship to other stochastic gradient learning rules. Details of the numerical simulations are given.

Hedonistic Synapse Model

A synapse is modeled as having two states, available (A) and refractory (R). When the synapse is available, a presynaptic spike stimulates vesicle release ($A \rightarrow R$ transition) with probability

$$p = \frac{1}{1 + e^{-q-c}}, \tag{1}$$

which is a function of the *release parameter* q, and the calcium-like variable c. To model calcium dynamics at the presynaptic terminal, c jumps by Δ_c for each presynaptic spike and otherwise decays exponentially, $dc/dt = -c/\tau_c$. This model for c is convenient, but other models could be substituted with no change in the learning rule, as long as Equation 1 still holds. After releasing a vesicle and entering the refractory state, the synapse recovers ($R \rightarrow A$) with time constant $1/\tau_r$ (Fuhrmann et al., 2002). While refractory, the synapse cannot release a vesicle.

Note that p is defined as the probability of vesicle release, conditioned on the synapse being in the A state. The overall release probability is the product of p and the probability that the synapse is in the A state. The conditional probability p goes up with successive spikes in a tetanus, due to increasing c. But the probability of being in the A state goes down, due to refractoriness. Therefore, the overall release probability shows a mix of facilitation and depression. The relative strength of the two effects depends on model parameters.

Implementation of the learning rule requires that the synapse maintain some trace of its recent actions. This is provided by a hypothetical quantity $\bar{e}(t)$, which jumps by

$$\Delta\bar{e} = \begin{cases} 1 - p, & \text{release,} \\ -p, & \text{failure} \end{cases} \tag{2}$$

at an available synapse in response to a presynaptic spike. There is no jump ($\Delta\bar{e} = 0$) for a refractory synapse. During the intervals between presynaptic spikes, it decays exponentially,

$$\frac{d\bar{e}}{dt} = -\frac{\bar{e}}{\tau_e}. \tag{3}$$

The time constant τ_e sets the time scale over which the synapse remembers its past actions.

The quantity $\bar{e}(t)$ is called an *eligibility trace*, because it signifies when the synapse is eligible for reinforcement by a reward signal $h(t)$ (Klopf, 1982). Plasticity is driven by the product of the reward signal and the eligibility trace,

$$\frac{dq}{dt} = \eta h(t)\bar{e}(t), \tag{4}$$

where $\eta > 0$ is a learning rate.

In the examples of Figures 2–4, short-term plasticity was eliminated from the model for simplicity, by not allowing the calcium-like variable to increase ($\Delta_c = 0$) and making recovery from the refractory state instantaneous ($\tau_r = 0$). In the example of Figure 5, short-term plasticity was modeled using the parameter values $\Delta_c = 1$, $\tau_c = 500$ ms, and $\tau_r = 800$ ms.

Note that several events are modeled as taking place instantaneously in response to a presynaptic spike. They are simulated as taking place in the following sequence: vesicle release or failure, change in $\bar{e}$, change in c.

REINFORCE Learning

"REINFORCE" is an acronym invented by Williams to refer to a class of learning rules that have the form of "REward Increment = Nonnegative Factor × Offset Reinforcement × Characteristic Eligibility" (Williams, 1992). The particular variant of REINFORCE learning used here is due to Kimura et al. (1995) and Baxter and Bartlett (2001). Suppose that samples x_t are generated by a Markov chain $P_\theta(x_t|x_{t-1})$ parametrized by θ, and reward $h(x_t)$ is received at every time step. Define the *eligibility*

$$e(x_t, x_{t-1}) = \nabla_\theta \log P_\theta(x_t|x_{t-1}), \tag{5}$$

and the *eligibility trace*

$$\bar{e}_t = \beta\bar{e}_{t-1} + e(x_t, x_{t-1}). \tag{6}$$

Then the REINFORCE learning rule is

$$\Delta\theta_t = \eta h(x_t)\bar{e}_t, \tag{7}$$

where $\eta > 0$ is the learning rate.

REINFORCE Learning for Stochastic Synapses

The network models studied in computational neuroscience are often written as systems of ordinary differential equations. Let the state vector x stand for all the dynamical variables of the network. In a numerical simulation, both time t and state x take on discrete values, so that the dynamics is describable in terms of a Markov transition matrix $P_\theta(x_t|x_{t-1})$. Given such a description, a REINFORCE learning rule can be derived by computing the gradient in Equation 5. Here this is done for a simple model of stochastic synapses.

The eligibility of the synapse from neuron j to neuron i is given by

$$e_{ij}(x_t, x_{t-1}) = \frac{\partial}{\partial q_{ij}} \log P_\theta(x_t|x_{t-1}).$$

This vanishes for any time step in which neuron j does not spike, because then $P_\theta(x_t|x_{t-1})$ is independent of q_{ij}. It also vanishes for any time step in which the synapse is refractory. If the synapse is available and neuron j does spike, then

$$\log P_\theta(x_t|x_{t-1}) = \begin{cases} \log p_{ij} + \ldots, & \text{release,} \\ \log(1 - p_{ij}) + \ldots, & \text{failure.} \end{cases}$$

Terms depending on the release probabilities of other synapses are omitted, because they vanish when taking the derivative with respect to q_{ij}. It follows that

$$e_{ij}(x_t, x_{t-1}) = \begin{cases} 1 - p_{ij}, & \text{release,} \\ -p_{ij}, & \text{failure,} \end{cases} \quad (8)$$

after applying the chain rule and the identity $dp/dq = p(1 - p)$. Taking the continuous time limit $dt \to 0$, $\beta \to 1$ of Equation 6 with $\tau_e = dt/(1 - \beta)$ held fixed yields Equations 2 and 3.

REINFORCE as Stochastic Gradient Learning

For those readers with interest in the theory of reinforcement learning, a brief summary of the mathematical foundations of Equations 5–7 is given here. A comprehensive treatment can be found in Baxter and Bartlett (2001). The goal of REINFORCE learning is to maximize the average reward. When the learning process is segmented into distinct episodes that are statistically independent of each other, the property of stochastic gradient ascent can be proven rigorously (Williams, 1992). Then convergence to the maximum of the expected reward follows from stochastic approximation theory, provided that the learning rate $\eta \to 0$ as $t \to \infty$ in the appropriate way (Kushner and Clark, 1978).

For fully online learning rules like Equation 7, the theoretical situation is less satisfactory, and only partial guarantees can be given (Baxter and Bartlett, 2001). For an ergodic Markov chain, the average reward can be written either as a time average or as an average over the equilibrium distribution $\pi_\theta(x)$,

$$H(\theta) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} h(x_t) = \sum_x h(x)\pi_\theta(x).$$

For fixed θ, it can shown that the time average of the right hand side of Equation 7 is a biased estimate of the gradient of the average reward,

$$\nabla_\theta H(\theta) \approx \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} h(x_t)\bar{e}_t. \quad (9)$$

The safe way to use this formula for learning would be to estimate the gradient by holding θ fixed for T time steps, computing the sum in Equation 9 for finite T, and then updating θ. If T is much longer than the mixing time, then updating θ every T steps should result in a good approximation to stochastic gradient ascent. Updating θ at every time step as in Equation 7 has less theoretical justification but often seems to work in practice.

Bias-Variance Tradeoff

There is a bias or approximation error in Equation 9, even in the limit $T \to \infty$. To understand this error, note that Equations 9 and 6 are equivalent to

$$\nabla_\theta H(\theta) \approx \sum_{\tau=0}^{\infty} \beta^\tau C(\tau), \quad (10)$$

where

$$C(\tau) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} e(x_t, x_{t-1})h(x_{t+\tau})$$

is the correlation between the eligibility and the reward $h(x_{t+\tau})$ at time lag τ. It can be shown that Equation 10 is an exact equality when $\beta = 1$, but only an approximation when $\beta < 1$. Unfortunately, the $\beta = 1$ formula is unusable for finite T, the case relevant for learning. This is because $C(\tau)$ vanishes when τ becomes much longer than the mixing time of the Markov chain, but the fluctuations in the estimate of $C(\tau)$ do not. In other words, $C(\tau)$ contributes noise but no signal to the gradient estimate for large τ. Therefore it is advantageous to choose $\beta < 1$ to reduce the variance of the gradient estimate, although it increases the bias. It can be shown that the bias is small when the time constant $1/(1 - \beta)$ is longer than the mixing time (Baxter and Bartlett, 2001).

To apply these ideas to the specific case of hedonistic synapses, note that the time average of the right hand side of Equation 4 is

$$\langle h(t)\bar{e}(t)\rangle_t = \int_0^\infty d\tau e^{-\tau/\tau_e}[(1 - p)C_{rh}(\tau) - pC_{fh}(\tau)], \quad (11)$$

where $C_{rh}(\tau) = \langle r(t)h(t + \tau)\rangle_t$ and $C_{fh}(\tau) = \langle f(t)h(t + \tau)\rangle_t$ are the correlations of release and failure with reward. This formula is valid if short-term synaptic plasticity is neglected, so that the release probability p does not vary dynamically. The release train is defined as $r(t) = \Sigma_i \delta(t - t_i)$, where t_i is the time of the ith vesicle release at the synapse, and the failure train $f(t)$ is defined similarly. The sum $r(t) + f(t)$ is equivalent to the presynaptic spike train (see Figure 3A).

Equation 11 reveals that plasticity is driven by a weighted difference between $C_{rh}(\tau)$ and $C_{fh}(\tau)$ for positive time lags τ. The weightings $1 - p$ and p compensate for the fact that unequal probabilities of release and failure affect the baseline values of the correlations. From the difference between the correlations, the synapse can determine whether release or failure is more likely to lead to reward.

The time constant τ_e of the eligibility trace (the continuum analog of β) should be chosen long enough so that the integral in Equation 11 captures the correlations in $C_{rh}(\tau)$ and $C_{fh}(\tau)$. Further lengthening of τ_e is counterproductive, as it adds noise to the learning update without adding any signal.

Temporal Antagonism

The learning rule of Equation 7 is a stochastic approximation to the gradient. Other rules sharing this property can be generated by adding extra terms with zero mean, which can be helpful if they have the effect of reducing fluctuations. One interesting modification is

$$\Delta\theta_t = \eta h(x_t)\bar{e}_t - \eta\bar{h}_{t-1}e(x_t, x_{t-1}), \quad (12)$$

where the *reward trace* $\bar{h}$ is defined by

$$\bar{h}_t = \beta\bar{h}_{t-1} + h(x_t). \quad (13)$$

Because $\bar{h}_{t-1}$ depends on past rewards, it is uncorrelated with the present eligibility $e(x_t, x_{t-1})$, which has zero mean. Therefore $\langle\bar{h}e\rangle = \langle\bar{h}\rangle\langle e\rangle = 0$, so that the extra term in Equation 12 adds no bias to the gradient estimate. However, it can reduce the variance of the estimate in certain situations. For example, consider the special case where h is a constant, so that the gradient of $\langle h\rangle$ is zero. Then Equation 7 would cause θ to randomly walk away from its initial condition, whereas θ does not stray very far if it obeys Equation 12. The subtraction in Equation 12 is closely related to reinforcement comparison, which has been studied in the field of machine learning (Sutton, 1984; Dayan, 1990).

The right hand side of Equation 12 is antisymmetric in h and e, resulting in temporal antagonism in the interaction between these quantities. When Equation 12 is applied to hedonistic synapses, temporal antagonism leads to the contrast between rules R1 and R3 and between R2 and R4.

Related Learning Rules

Minsky proposed using stochastic synaptic transmission as the basis for learning in his SNARC, an artificial neural network (Minsky, 1954). His learning rule was based on the product of reinforcement and release. It did not include release-failure antagonism, which is important for the property of stochastic gradient ascent (see Figure 3). Later, Hinton (1989) proposed applying the linear reward-inaction (L_{R-I}) algorithm (Narendra and Thathachar, 1989) to neural networks. This algorithm updates p directly, rather than q. The q parametrization is theoretically pleasing because it results in stochastic gradient ascent, rather than a diagonally rescaled version. Furthermore, it is convenient when formulating the learning rule for dynamic synapses.

There have been numerous other proposals for learning by correlating synaptic noise with reward in artificial neural networks (Jabri and Flower, 1992; Cauwenberghs, 1993; Unnikrishnan and Venugo-

pal, 1994), but they have not employed the stochastic, all-or-none transmission seen in biological synapses. In another class of learning rules, reward is correlated with the noise of random spike generation (Barto et al., 1983; Mazzoni et al., 1991; Williams, 1992; Bartlett and Baxter, 2000; Xie and Seung, unpublished data).

Variable-Interval Reward Schedule

The matching law of Herrnstein (not to be confused with probability matching) is commonly studied using variable-interval (VI) schedules running concurrently at two targets (Davison and McCarthy, 1988; Gallistel, 1994). In a VI schedule, reward remains at a target until it is harvested. Upon harvesting, the next reward appears after a time interval that is chosen randomly from an exponential distribution.

A discrete-time analog of the classical continuous-time VI schedule is used in Figure 4. The simulation is broken into discrete trials, and the waiting time for the next reward is chosen randomly from a geometric distribution. This is implemented by tossing a biased coin for each unbaited target at the beginning of the trial. The result of the coin toss determines whether or not the target is rebaited with reward. At the end of each trial, unharvested reward remains at a target for the next trial.

For this particular task, it can be proven that matching behavior maximizes reward, within the class of memoryless policies. But there are other tasks for which matching deviates from maximizing. Animals have been observed to learn matching behavior even when it does not maximize reward. This has led to the proposal that animals follow a principle of "melioration" rather than maximization (Herrnstein and Prelec, 1991). It can be proven that REINFORCE learners, including the neural circuit of Figure 4, do something like melioration when the time constant of the eligibility trace is short (S.S., unpublished data).

Numerical Simulations

The input neurons produce Poisson spike trains. All other neurons are modeled with the integrate-and-fire equations

$$C\frac{dV_i}{dt} = -g_L(V_i - V_L) - \sum_j G_{ij}(V_i - V_{ij}) + I_{tonic},$$

with $V_L = -74$, $g_L = 25$ nS, and $C = 500$ pF. This and other differential equations are integrated using an exponential Euler update with a time step of 0.5 ms. When the membrane potential V_i reaches the threshold value of $V_\theta = -54$ mV, it is reset to $V_{reset} = -60$ mV. To simulate the random barrage of tonic input from sources outside the network, there is an additional term I_{tonic}. In Figure 2, this has mean 425 pA and standard deviation 200 pA. In Figure 3, it has mean 450 pA and standard deviation 300 pA. No tonic input is added in Figures 4 and 5.

The reversal potential V_{ij} of the synapse from neuron j to i is set at either 0 or -70 mV, depending on whether the synapse is excitatory or inhibitory. For every spike of neuron j, the release variables r_{ij} for all i are chosen randomly, taking the value 1 with probability p_{ij} and 0 otherwise. For dynamic synapses, this procedure is modified as described earlier. The synaptic conductances are updated via $\Delta G_{ij} = W_{ij} r_{ij}$. In the absence of presynaptic spikes, the G_{ij} decay exponentially with time constant $\tau_s = 5$ ms, except for the excitatory synapses of Figure 4, which have a time constant of 100 ms. The W_{ij} do not change in time. In Figure 2, the excitatory W_{ij} are chosen randomly from an exponential distribution with mean 2.4 nS. The inhibitory W_{ij} are chosen similarly, but with a mean of 45 nS. In Figure 3, the conductances are (b) 10, (c) 20, and (d) 3 nS. In Figure 4, the maximal conductances are 1 nS for the excitatory autapses, 2 nS for the synapses onto the inhibitory neuron, 25 nS for the inhibitory synapses onto the excitatory neurons, and 0.25 nS for the feedforward synapses. In Figure 5, the maximal conductances are 4 nS.

The time constant of the eligibility trace is $\tau_e = 20$ ms in Figures 2 and 5, and the reward signal is the spike train of the output neuron or its negative. Therefore, each spike of the output neuron leads to a change $\Delta q_{ij} = \pm\eta e_{ij}$ depending on whether the spike counts as reward or punishment. In Figure 4, the time constant of the eligibility trace is 1000 ms.

The learning rate is $\eta = 0.3$ in Figure 2, and $\eta = 0.1$ in Figures 4 and 5. Bound constraints of ± 3 are imposed on q_{ij} in Figure 2, for a slight improvement in learning performance.

Acknowledgments

I am grateful to Winfried Denk, Zach Mainen, Bill Newsome, Whitman Richards, Mark Schnitzer, and Xiao-Jing Wang for their comments on drafts of this paper. I also thank Geoff Hinton for pointing me to early work on reinforcement learning; Misha Tsodyks for advice about short-term synaptic plasticity; Greg Corrado for help with the matching law; and Ila Fiete and Xiao-Hui Xie for many thought-provoking discussions.

Received: September 11, 2003
Revised: November 10, 2003
Accepted: November 13, 2003
Published: December 17, 2003

References

Bartlett, P.L., and Baxter, J. (2000). A biologically plausible and locally optimal learning algorithm for spiking neurons. Technical Report Australian National University.

Barto, A.G., Sutton, R.S., and Anderson, C.W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans. Syst. Man Cybern. *13*, 834–846.

Baum, E.B. (1999). Toward a model of intelligence as an economy of agents. Mach. Learn. *35*, 155–185.

Baxter, J., and Bartlett, P.L. (2001). Infinite-horizon policy-gradient estimation. Journal of Artificial Intelligence Research *15*, 319–350.

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

Bolshakov, V.Y., and Siegelbaum, S.A. (1995). Regulation of hippocampal transmitter release during development and long-term potentiation. Science *269*, 1730–1734.

Buchs, N.J., and Senn, W. (2002). Spike-based synaptic plasticity and the emergence of direction selective simple cells: simulation results. J. Comput. Neurosci. *13*, 167–186.

Cauwenberghs, G. (1993). A fast stochastic error-descent algorithm for supervised learning and optimization. Adv. Neural Info. Proc. Syst. *5*, 244–251.

Chance, F.S., Nelson, S.B., and Abbott, L.F. (1998). Synaptic depression and the temporal response characteristics of v1 cells. J. Neurosci. *18*, 4785–4799.

Dale, N., Schacher, S., and Kandel, E.R. (1988). Long-term facilitation in aplysia involves increase in transmitter release. Science *239*, 282–285.

Davison, M., and McCarthy, D. (1988). The Matching Law: A Research Review (Hillsdale, NJ: Erlbaum).

Dayan, P. (1990). Reinforcement comparison. In Proceedings of the 1990 Connectionist Models Summer School, D.S. Touretzky, J.L. Elman, T.J. Sejnowski, and G.E. Hinton, eds, (San Mateo, CA: Morgan Kaufmann), pp. 45–51.

Fuhrmann, G., Segev, I., Markram, H., and Tsodyks, M. (2002). Coding of temporal information by activity-dependent synapses. J. Neurophysiol. *87*, 140–148.

Gallistel, C.R. (1994). Foraging for brain stimulation: toward a neurobiology of computation. Cognition *50*, 151–170.

Herrnstein, R.J., and Prelec, D. (1991). Melioration: a theory of distributed choice. J. Econ. Perspect. *5*, 137–156.

Hinton, G.E. (1989). Connectionist learning procedures. Artif. Intell. *40*, 185–234.

Ito, M. (2001). Cerebellar long-term depression: characterization, signal transduction, and functional roles. Physiol. Rev. *81*, 1143–1195.

Jabri, M., and Flower, B. (1992). Weight perturbation - an optimal architecture and learning technique for analog VLSI feedforward and recurrent multilayer networks. IEEE Trans. Neural Netw. *3*, 154–157.

Johnson, F., Soderstrom, K., and Whitney, O. (2002). Quantifying song bout production during zebra finch sensory-motor learning

suggests a sensitive period for vocal practice. Behav. Brain Res. *131*, 57–65.

Kimura, H., Yamamura, M., and Kobayashi, S. (1995). Reinforcement learning by stochastic hill climbing on discounted reward. In Proceedings of the 12th International Conference on Machine Learning, A. Prieditis and S. Russell, eds. (San Francisco: Morgan Kaufmann), pp. 295–303.

Klopf, A.H. (1982). The Hedonistic Neuron (Washington: Hemisphere).

Kushner, H.J., and Clark, D.S. (1978). Stochastic Approximation Methods for Constrained and Unconstrained Systems (New York: Springer-Verlag).

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

Mainen, Z.F., and Sejnowski, T.J. (1995). Reliability of spike timing in neocortical neurons. Science *268*, 1503–1506.

Malenka, R.C., and Nicoll, R.A. (1999). Long-term potentiation—a decade of progress? Science *285*, 1870–1874.

Markram, H., and Tsodyks, M. (1996). Redistribution of synaptic efficacy between neocortical pyramidal neurons. Nature *382*, 807–810.

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

Mazzoni, P., Anderson, R.A., and Jordan, M.I. (1991). A more biologically plausible learning rule for neural networks. Proc. Natl. Acad. Sci. USA *88*, 4433–4437.

Menzel, R. (2001). Searching for the memory trace in a mini-brain, the honeybee. Learn. Mem. *8*, 53–62.

Minsky, M.L. (1954). Theory of neural-analog reinforcement systems and its application to the brain-model problem. PhD thesis, Princeton University, Princeton, New Jersey.

Minsky, M. (1961). Steps toward artificial intelligence. Proc. IRE *49*, 8–30.

Minsky, M. (1988). Society of Mind (New York: Simon & Schuster).

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

Narendra, K.S., and Thathachar, M.A.L. (1989). Learning Automata: An Introduction (Englewood Cliffs, NJ: Prentice Hall).

Raymond, J.L., and Lisberger, S.G. (1998). Neural learning rules for the vestibulo-ocular reflex. J. Neurosci. *18*, 9112–9129.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning internal representations by back-propagating errors. Nature *323*, 533–536.

Schultz, W. (2002). Getting formal with dopamine and reward. Neuron *36*, 241–263.

Stevens, C.F. (1993). Quantal release of neurotransmitter and long-term potentiation. Cell Suppl. *72*, 55–63.

Stevens, C.F., and Wang, Y. (1994). Changes in reliability of synaptic function as a mechanism for plasticity. Nature *371*, 704–707.

Sutton, R.S. (1984). Temporal credit assignment in reinforcement learning. PhD thesis, University of Massachusetts, Amherst, Amherst, Massachusetts.

Tesauro, G. (1992). Practical issues in temporal difference learning. Mach. Learn. *8*, 257–277.

Thomson, A.M. (2000). Facilitation, augmentation and potentiation at central synapses. Trends Neurosci. *23*, 305–312.

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. Neural Comput. *10*, 821–835.

Unnikrishnan, K.P., and Venugopal, K.P. (1994). Alopex: a correlation based learning algorithm for feed-forward and recurrent neural networks. Neural Comput. *6*, 469–490.

van Vreeswijk, C., and Sompolinsky, H. (1998). Chaotic balanced state in a model of cortical circuits. Neural Comput. *10*, 1321–1371.

Werfel, J.K., Xie, X., and Seung, H.S. (2004). Learning curves for stochastic gradient descent in linear feedforward networks. Adv. Neural Info. Proc. Syst., in press.

Williams, R.J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. *8*, 229–256.