Biol Cybern (2008) 99:517–523
DOI 10.1007/s00422-008-0265-6

ORIGINAL PAPER

Biological
Cybernetics

An implementation of reinforcement learning based on spike timing dependent plasticity

Patrick D. Roberts · Roberto A. Santiago · Gerardo Lafferriere

Received: 21 February 2008 / Accepted: 19 September 2008 / Published online: 22 October 2008

Abstract An explanatory model is developed to show how synaptic learning mechanisms modeled through spike-timing dependent plasticity (STDP) can result in long-term adaptations consistent with reinforcement learning models. In particular, the reinforcement learning model known as temporal difference (TD) learning has been used to model neuronal behavior in the orbitofrontal cortex (OFC) and ventral tegmental area (VTA) of macaque monkey during reinforcement learning. While some research has observed, empirically, a connection between STDP and TD, there has not been an explanatory model directly connecting TD to STDP. Through analysis of the learning dynamics that results from a general form of a STDP learning rule, the connection between STDP and TD is explained. We further demonstrate that a STDP learning rule drives the spike probability of a reward predicting neuronal population to a stable equilibrium. The equilibrium solution has an increasing slope where the steepness of the slope predicts the probability of the reward, similar to the results from electrophysiological recordings suggesting a different slope that predicts the value of the anticipated reward of Montague and Berns [Neuron 36(2):265–284, 2002]. This connection begins to shed light into more recent data gathered from VTA and OFC which are not well modeled by TD. We suggest that STDP provides the underlying mechanism for explaining reinforcement learning and other higher level perceptual and cognitive function.

Keywords Computational neuroscience · Learning · Synaptic plasticity · Spiking neuron model

1 Introduction

The theory of reinforcement learning states that behavior followed by reward leads to greater likelihood of that same behavior occurring again (Sutton and Barto 1998). Thus, any mechanism that implements reinforcement learning, whether biological or mechanical, must have the capability to predict future reward. The learning aspect of reinforcement learning refers to the ability to associate a present state with a reward which will be received later. This basic principle drove the development of the temporal difference (TD) learning algorithm (Sutton and Barto 1998) which has become a crucial element in many computational models of reinforcement learning. However, the biological implementation of TD, in terms of biological circuitry and learning mechanisms, is as yet unresolved (Wörgötter and Porr 2005). We present a biological model of reinforcement learning that is more parsimonious than previous models and relies on known synaptic learning rules.

Recently, significant progress has been made in connecting theories of reinforcement learning to observed adaptation of neuronal processing. In particular, the activity of dopaminergic (DA) cells in the ventral tegmental area (VTA) and Substantia Nigra pars compacta (SNc) of macaque monkey during conditioning protocols seems to correspond strongly

This material is based upon work supported by the National Science Foundation under Grants No. IOB-0445648 (PDR) and DMS-0408334 (GL) and by a Career Support grant from Portland State University (GL).

P. D. Roberts (✉)
Department of Science and Engineering, Oregon Health and Science University, Portland, OR 97239, USA
e-mail: robertpa@ohsu.edu

R. A. Santiago
Systems Science Program, Portland State University, Portland, OR 97207, USA

G. Lafferriere
Department of Mathematics and Statistics, Portland State University, Portland, OR 97207, USA

to a model of reward prediction. Specifically, Schultz et al. (1997) have proposed that the response patterns of DA cells reflect a process of learning to predict future reward through the TD learning algorithm. The correspondence between the model of TD learning algorithm and recordings from DA cells in macaque is quite extraordinary, leading to the implication that the macaque uses the TD algorithm to predict future reward. The experimental results by Schultz et al. (1997) provide a foundation for understanding reinforcement learning that links the neuronal level with the behavioral level. Indeed, the TD algorithm also models behavior as well as the DA neuron responses of macaque, although some debate has arisen over recordings from DA cells in macaque during conditioning protocols where the rewarding event occurs probabilistically (Daw and Dayan 2004; Niv et al. 2005; Morris et al. 2006). The present project explores a new explanatory model of these DA responses which relies directly on spike timing dependent plasticity (STDP) and finds that the TD algorithm is a derivative result.

1.1 Temporal derivatives in the DA system

Previous applications of the TD model to the DA system of macaque have relied on a physiological interpretation which has caused some confusion. The TD model is well named since algorithmically it depends on the differences in neuronal behavior between two time steps. Montague et al. (1996) posit the existence of neurons which are able to calculate a temporal derivative (neuron D in Fig. 1a), although no known neuron or neuronal mechanism is proposed. To see this we return to the basic definition of the model as proposed in the article by Schultz et al. (1997) and its supporting model framework (Montague et al. 1996).

The TD model of the DA system relies upon a serial stimulus representation (SSR), ie. an internal representation of a stimulus which is "played back" repeatedly for a finite number of intervals. A stimulus arrives into the system and is represented by the spiking of neurons at time $x_i(t)$ during stimulus cycle t. Through a series of delays a set of neurons output a spike at fixed intervals after the first spike, where the spike-time of the ith delayed neurons is denoted as $x_i(t)$. Due to the presence of many cumulative delays in the nervous system, the existence of such SSRs through delay lines seems biologically plausible.

The model of Montague et al. (1996) posits that the set of spikes, at times $x_i(t)$, arrive as input to a neuron causing changes in its membrane potential, V, and whose output is the temporal derivative of its membrane potential, $\dot{V}$. Figure 1a diagrammatically shows the convergence of the SSR onto a single neuron D. Each element of the SSR is modified by a weight w_i which models the strength of synaptic communication of input x_i onto D. The output of neuron D, ($\dot{V}$), can be roughly approximated by the quantity $V(t-1) - V(t)$. Thus it is claimed (Montague et al. 1996; Schultz et al. 1997) that the output of the neuron is the TD $\dot{V} \approx V(t) - V(t-1)$. This TD signal becomes the fundamental driving term for the adjustments of the weights onto D, w_i. The model describes how the inputs to neuron D drive its membrane potential and how the TD output of D drives synaptic weight change. The model correctly predicts the output of the system (Schultz et al. 1997), but here we stop to critique this mechanism, in particular its existence as a mechanism separate from neurons in VTA.

Neuron D is not supported by any direct evidence, and is meant to denote an intermediate region between cortical projections and dopamine neurons in the midbrain (Montague et al. 1996). In the case presented in Schultz et al. (1997), the idea of an "intermediate region" has been replaced with the statement that the model poses the hypothesis that such a neuron or mechanism exists. The justification for positing its existence is due to the nature of the recordings taken from VTA whose output could be explained consistently with the TD algorithm if they received inputs proportional to $\dot{V}$. Moreover, the adaptation of the output of these VTA neurons seems also to be driven by the term $\dot{V}$. Since physically, changes in neuron activity are associated with changes in synaptic efficacy, the researchers naturally posit that the $\dot{V}$ signal must be driving the changes in synaptic efficacy somewhere upstream of VTA. Thus, D plays the critical role of calculating $\dot{V}$ by some unknown mechanism. We show in the following that the function of D can be explained by physiologically realistic synaptic mechanisms with a simple neural model.

1.2 Reinforcement learning and STDP

We propose that the functionality represented by neuron D is actually a STDP mechanism within the neurons of the cortex, the striatum or the VTA (C in Fig. 1b). The TD algorithm can be seen as a discrete approximation of differential Hebbian learning. Early work in computational learning models (Klopf 1998; Kosko 1986) showed that differential Hebbian learning could implement basic classical conditioning. In these models, changes in synaptic efficacy are proportional to the time derivative of the membrane voltage, $\dot{V}(t)$. Since $\dot{V}(t)$ can be approximated discretely as the difference of the membrane voltage at two time points $V(t) - V(t-k)$, the operative component of TD, many researchers have noted this connection between differential Hebbian learning and TD (for a review, see Wörgötter and Porr 2005). Another mechanism exists that is more akin to differential Hebbian learning, seems more biologically plausible, and is supported by currently available research. Specifically, we will now discuss how STDP implements differentiable Hebbian learning (Wörgötter and Porr 2005).

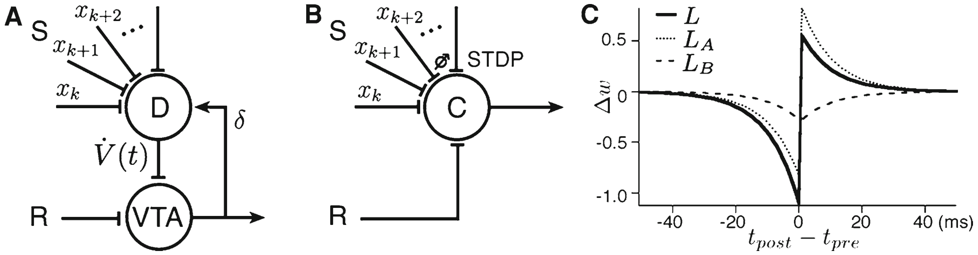

Fig. 1 **a** Model network proposed in Montague et al. (1996). The output from VTA provides an error signal (δ) to instruct the input neurons for reinforcement learning. **b** Proposed network leads to a prediction of future rewards represented in the responses of neuron C based strictly on STDP. **c** Spike-timing dependent learning rule (Bi and Mu-Ming 1998; Feldman 2000) used in spiking simulations (*solid black trace*). The learning rule decomposes into an antisymmetric part (*dotted trace*) and a symmetric part (*dashed trace*)

STDP has been shown to implement differential Hebbian learning (Roberts 1999). Several researchers have shown how STDP can be modelled with TD (Rao and Sejnowski 2000) and TD-like rules or differential Hebbian learning (Wörgötter and Porr 2005), but none of these researchers have shown how STDP implements TD and TD-like rules (however, see Izhikevich 2007 for a complementary approach). By treating the fast time scale (activity during the stimulus cycle, x) independently from the slow time scale (changes caused by plasticity, t), TD-like rules emerge from weight changes generated by the anti-symmetric temporal learning rule (Fig. 1c) and are related to differential Hebbian learning (Roberts 1999). Specifically, $\Delta w \approx \frac{\mathrm{d}V}{\mathrm{d}t} = \dot{V}$. We will now extend this relation to find a solution for the spike probability of cortical, striatal, or VTA neurons in anticipation of a reward following a stimulus.

2 Methods and models

2.1 Model of spike timing dependent plasticity

We denote the spike probability of a neuron C in Fig. 1b as, $P(x_m, t_n)$, the spike probability function. We will assume that this function takes values between 0 and 1, and is a monotonically increasing sigmoidal function of the membrane potential, $V(x_m, t_n)$. We will give a specific example in Sect. 3. The membrane potential is the sum of a the reward, $R(x_m)$, and the SSR postsynaptic potentials, $E(x_m)$, scaled by adaptive weights $w(x_p, t_n)$,

$$V(x_m, t_n) = R(x_m) + \sum_p w(x_p, t_n) E(x_m - x_p). \tag{1}$$

The reward is expressed here as a deterministic input to the membrane potential, and is zero except for at the time of a reward, x_R, when $R(x_R) = 1$. We will later generalize this definition to include a probability of a reward, $R(x_R) = P_R \leq 1$. The weights change as functions of the timing difference between pre- and postsynaptic spikes. Let $S^{\text{post}}(x_s)$ be a collection of random variables, each of which is 1 if there is a postsynaptic spike at time x_s, and zero otherwise. The weights change additively with the occurrence of postsynaptic spikes, $\Delta w(x_p, t_n) = \sum_s L(x_s - x_p) S^{\text{post}}(x_s)$, where the learning function, $L(x)$, defines the amount and direction of change for each spike pair. The average weight change is then, $\langle \Delta w(x_p, t_n) \rangle = \sum_m L(x_m - x_p) P(x_m, t_n)$ (Roberts 1999). We will decompose the learning rule into antisymmetric and symmetric parts, $L(x_m) = L_A(x_m) + L_B(x_m)$.

We are now in the position to solve for the average change in synaptic efficacy during each stimulus-reward cycle. In the continuous limit, the spike probability function is expanded in powers of $z = x_m - x$ to yield,

$$\begin{aligned}\langle \Delta w(x, t) \rangle &= \int_{-\infty}^{\infty} L(z) \left[P(x, t) + z \frac{\partial}{\partial x} P(x, t) \right. \\ &\qquad \left. + \frac{z^2}{2!} \frac{\partial^2}{\partial x^2} P(x, t) + \cdots \right] \mathrm{d}z \\ &= \sum_k L_k \frac{\partial^k}{\partial x^k} P(x, t), \end{aligned} \tag{2}$$

where the moments of the temporal learning rule are defined by $L_k = \frac{1}{k!} \int_{-\infty}^{\infty} z^k L(z) \mathrm{d}z$. Using the decomposition of the learning rule we get $L_k = \frac{1}{k!} (\int z^k L_A(z) \mathrm{d}z + \int z^k L_B(z) \mathrm{d}z)$, and the integral limits are infinite. The first two terms of the Taylor expansion (Eq. 2) yield: $\langle \Delta w \rangle = L_0 P(x, t) + L_1 P_x(x, t)$, where $L_0 = \int L_A(z) \mathrm{d}z + \int L_B(z) \mathrm{d}z$ and $P_x(x, t)$ is the derivative of $P(x, t)$ with respect to x. Since L_A is antisymmetric, the first term vanishes. We have chosen to let $L_B < 0$ for reasons of stability (Song et al. 1993) and following the physiological evidence in Feldman (2000), leading to $L_0 = \int L_B(z) \mathrm{d}z = -\alpha$. Similarly, $L_1 = \int z L_A(z) \mathrm{d}z = \beta$, so that

$$\langle \Delta w \rangle = -\alpha P(x, t) + \beta P_x(x, t). \tag{3}$$

This expression provides a general form for synaptic weight changes as a result of STDP with SSR. Note that the expression is independent of the specific form of the learning rule and only depends on the relative magnitude of the spike-timing symmetric and anti-symmetric parts. Thus, the learning dynamics could also apply to deformations of the STDP learning rule caused by pairings of multiple spikes.

2.2 Simulation of learning dynamics

As a non-trivial confirmation of the conclusions that we draw from Eq. 3, we constructed a numerical model of spiking neurons with the STDP learning rule and circuitry shown in Fig. 1b. The spike probability of the model neuron was calculated from the membrane potential, $V(x_m, t_n)$, the sum of all inputs, was normalized to a maximum value of unity. The spike-probability function was defined as $P(x_m, t_n) = (1+\exp(-\mu(V(x_m, t_n)-\theta)))^{-1}$, where we let the threshold $\theta = 30$ and the noise parameter $\mu = 100$ (Roberts and Bell 2000; Roberts 2004). During each time-step in (x_m, t_n), the probability of a spike was calculated and a pseudo-random number generator was used to determine whether to assign a spike. At the beginning of each cycle in t_n, the assigned spikes were used to update the weights using the STDP learning rule. The spike times of the model were binned for each time step to generate histograms that represent the time average of the spike probability. As a comparison, we simulated an ensemble of 100 identical spiking models to calculate the ensemble average of the spike probability.

3 Results

3.1 Approximate temporal difference learning

To recover TD learning, we approximate the probability of postsynaptic spiking as proportional to the membrane voltage, $P(V) = \mu V$. If we let $T = t + x$, we get $\langle \Delta w\rangle \approx -\alpha\mu V(T) + \beta\mu V_x(T)$, where $V_x(T)$ is the implicit derivative of $V(T)$ with respect to x. Using discrete approximation of $V_x(T)$, and rescaling our STDP learning parameters, $\hat{\alpha} = \mu\alpha$ and $\hat{\beta} = \mu\beta$, we finally arrive at the TD-like formula,

$$\begin{aligned}\langle \Delta w\rangle &\approx \hat{\beta}(V(T) - V(T-1)) - \hat{\alpha}V(T)\\ &= \hat{\beta}(\gamma V(T) - V(T-1)),\end{aligned} \tag{4}$$

where $\gamma = 1 - \hat{\alpha}/\hat{\beta} = 1 - \alpha/\beta$. In this last expression, $\hat{\beta}$ is interpreted as the learning rate and γ is the discount factor of TD learning (Sutton and Barto 1998). Thus, our average weight change, derived from STDP, can implement TD learning. The learning rate depends on the noise parameter because in a noisy system (small μ), there will be many random jumps before the system converges, and learning will be slow.

3.2 Continuum approximation

To obtain an analytic expression for the learning dynamics, we investigate the continuum limit of the prediction of future reward by linearizing the spike probability function,

$$w_t = -\hat{\alpha} w * E + \hat{\beta}(w * E)_x, \tag{5}$$

where E is the EPSP kernel. Here, $w = w(x,t)$, and subscripts denote partial derivatives so that $w_t = \partial w/\partial t$ and $(w * E)_x = \partial(w * E)/\partial x$, and the convolution is with respect to the x variable.

The initial and boundary conditions are

$$\text{(I.C.)} \quad w(x,0) = \varphi(x), \tag{6}$$

$$\text{(B.C.)} \quad w(R,t) = P_R, \tag{7}$$

where R denotes the time at which the reward occurs and P_R is the amount of the reward represented by a constant value of the weight at the time of reward. We have here replaced the reward input ($R(x_m)$ in Eq. 1) by a probability of a reward, and represented it as a fixed weight at time $x = R$. The function φ represents an initial distribution of weights.

3.3 Stability of the weights

To analyze the stability of the weights under the differential equation, we begin with a change in w to adjust the boundary conditions. Define $\tilde{w}(x,t) = w(x,t) - p_e(x)$, so that $p_e(x)$ is the equilibrium solution (such a solution exists as a generalized function). The boundary conditions of the new function $\tilde{w}$ satisfies $\tilde{w}(R,t) = 0, \forall t \geq 0$, and the initial condition is transformed into $\tilde{w}(x,0) = \tilde{\varphi}(x) = \varphi(x) - p_e(x)$. Substituting back into Eq. 5 we get: $\tilde{w}_t = \hat{\beta}(\tilde{w}*E)_x - \hat{\alpha}(\tilde{w}*E)$. Thus, $\tilde{w}$ satisfies the same PDE but with zero boundary conditions.

We next look for solutions of the form $\tilde{w} = f(x)g(t)$, where f and g are smooth and f is periodic of period R. Then $\tilde{w}_t = f(x)g'(t)$ and $(\tilde{w} * E)_x = (f * E)'(x)g(t)$. Substituting this form of $\tilde{w}$ into Eq. 5 and dividing both sides by $\tilde{w}$ we may conclude that $g(t) \propto \mathrm{e}^{\lambda t}$, where λ is a constant, and

$$\hat{\beta}(f * E)'(x) - \hat{\alpha} f * E(x) = \lambda f(x). \tag{8}$$

We introduce operator notation and denote by $\mathcal{E}$ the convolution operator ($\mathcal{E}(f) = f * E$), by $\mathcal{D}_x$ the derivative operator ($\mathcal{D}_x(f) = f'$), and by $\mathcal{I}$ the identity ($\mathcal{I}(f) = f$). Thus, f must be an eigenfunction of $(\hat{\beta}\mathcal{D}_x - \hat{\alpha}\mathcal{I})\mathcal{E}$. Such eigenfunctions are of the form $\mathrm{e}^{i2\pi kx/R}$, for k any integer. Notice also that

$$\mathcal{E}(\mathrm{e}^{i2\pi kx/R}) = \hat{E}(k)\mathrm{e}^{i2\pi kx/R}, \tag{9}$$

where we define $\hat{E}(k)$ by $\int_0^R E(y)\mathrm{e}^{-i2\pi ky/R}\mathrm{d}y$. We will refer to $\hat{E}(k)$ as the kth Fourier coefficient of E. Substituting into

Eq. 8 we obtain the possible values of λ: $\lambda = (\hat{\beta}\frac{i2\pi k}{R} - \hat{\alpha})\hat{E}(k)$.

We finally arrive at the solution to the weight configuration that results from STDP. Allowing for superpositions of functions of the form $f(x)g(t)$ we get the general form for $\tilde{w}$,

$$\tilde{w}(x,t) = \sum_{k\in\mathbb{Z}} \tilde{A}_k \mathrm{e}^{(\hat{\beta}\frac{i2\pi k}{R} - \hat{\alpha})\hat{E}(k)t}\mathrm{e}^{\frac{i2\pi k}{R}x}. \tag{10}$$

Convergence to the desired equilibrium as $t \to \infty$ will occur if $\lambda = (\hat{\beta}\frac{i2\pi k}{R} - \hat{\alpha})\hat{E}(k)$ in the exponent has negative real part.

To illustrate these estimates we restrict our analysis to the special case of $E(x) = \mathrm{e}^{-\sigma x}$, with $\sigma > 0$. In this case, for each integer k we have

$$\hat{E}(k) = \frac{1 - \mathrm{e}^{-\sigma R}}{\sigma + \frac{2\pi k i}{R}}. \tag{11}$$

The real part of λ is negative if and only if

$$k^2 < \gamma_c \frac{\sigma}{\left(\frac{2\pi}{R}\right)^2}, \tag{12}$$

where $\gamma_c = \hat{\alpha}/\hat{\beta} = \alpha/\beta$. Therefore convergence can be guaranteed in this case if all the coefficients $\tilde{A}_k$ are zero, except (possibly) those for which k satisfies (12). The coefficients $\tilde{A}_k$ need to be determined from a Fourier series expansion of the initial condition $\varphi(x)$, thus convergence can be guaranteed if φ does not have high frequency components.

The membrane potential V can be recovered from the weights w via the convolution formula $V = w * E$. Therefore, due to the continuity of the convolution operation, the membrane potential also converges under the low frequency conditions described above.

3.4 Numerical simulation of continuous model

We used the following discretization scheme to calculate numerical solutions to the fully nonlinear equation

$$w_t = -\alpha P(x,t) + \beta P_x(x,t), \tag{13}$$

where $P(x,t) = P(V(x,t))$ as in Sect. 2.2. Both derivatives w_t and P_x were replaced by their forward increment quotients:

$$w_t \approx \frac{w(x,t+\ell) - w(x,t)}{\ell}, \tag{14}$$

$$P_x(x,t) \approx \frac{P(x+h,t) - P(x,t)}{h}. \tag{15}$$

Moreover, the value of $V = w * E$ is approximated by the finite sum $\sum_{j=0}^{N} w(jh,t)E(x - jh)$.

After substituting such approximations in (13) and rearranging the terms we obtain the following recursions:

$$V(x,t) = \sum_{j=0}^{N} w(jh,t)E(x - jh), \tag{16}$$

$$\begin{aligned} w(x,t+\ell) &= w(x,t) - (\beta\rho + \alpha\ell)P(x,t) \\ &\quad + \beta\rho P(x+h,t), \end{aligned} \tag{17}$$

where $\rho = \ell/h$. Results of the numerical simulation are presented in Fig. 2a, b. A cross section of the solutions' time-evolution is presented in Fig. 2c.

3.5 Numerical simulations of the DA system

We can use Eq. 3 to predict the expected spike probability of the output (C) neurons in Fig. 1b after an association has been learned between a stimulus and a reward. When the weights reach an equilibrium, $\langle \Delta w \rangle = 0$, so that $\alpha P(x,t) =$

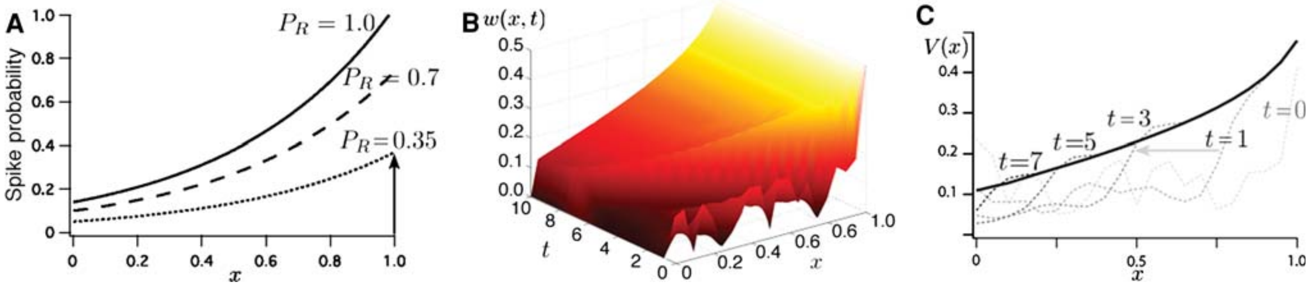

Fig. 2 **a** Numerical solution of our PDE in the linear case predicts the probability of a reward (P_R) at $x = 1.0$ with $\alpha = 1$ and $\beta = 2$. There is a unique solution for the spike probability at any time prior to the reward for a given reward probability. **b** Time course of the weight change leading to equilibrium in the nonlinear case after starting with a random weight configuration. The spike probability function has a threshold, $\theta = 0.3$ and noise, $\mu = 1$. The t-step, $h = 0.01$, and the x-step, $l = 0.05$, and there are 20 input neurons with 1,000 trials. The probability of the reward is $P_R = 0.5$. The edge that runs diagonally and borders the smooth exponential is the leading edge of the wave moving towards equilibrium. **c** The membrane potential in the nonlinear case approaches equilibrium (*black solid trace*) following repeated trials ($\Delta t = 0.01$, $E(x) = \exp(-x/\tau)$, $\tau = 20$). Shown in *gray* are various stages of progression towards equilibrium as time slices. In contrast to the exponential solutions of the linear case, the membrane potential is flattened due to the spike probability function, $P(V(x,t))$

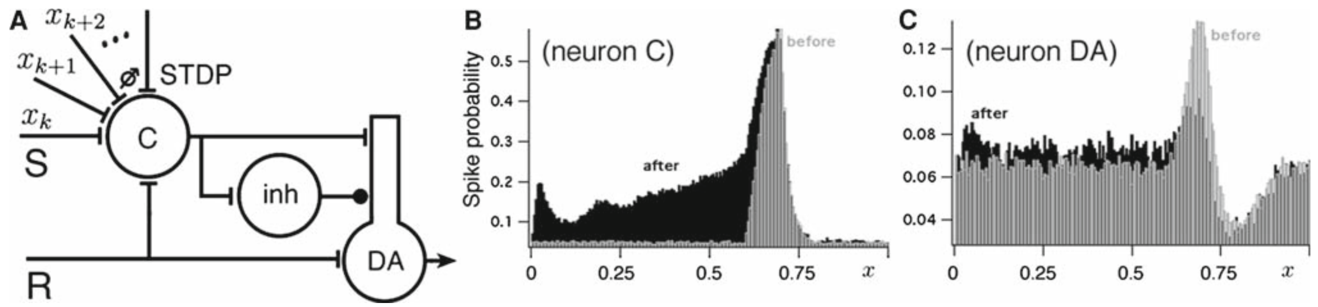

Fig. 3 **a** Proposed network to explain phasic responses of dopamine neurons (*DA*) to rewards. The DA neuron is excited by the C neuron that predicts future reward and inhibited by an interneuron. **b** Post stimulus spike histogram of C from a spiking neuron simulation of the network in **a** with a reward at $x = 0.7$. The reward causes a strong response late in the cycle (*gray histogram*, before), and the learning rule increases the weights to generate a ramp that predicts the probability of the reward (*black histogram*, after). **c** Post stimulus spike histogram for DA in the same simulation as **b**. Before learning, DA is modulated only at the reward (*gray histogram*, before), but following repeated presentation of the reward, DA responds with a phasic pulse immediately following the initial stimulus at time $x = 0.0$, and the response to the reward is reduced

$\beta P'(x, t)$, with the solution,

$$f(w(x,t)) = A(t)\mathrm{e}^{\gamma_c x}. \tag{18}$$

This exponential solution is consistent with the neurons recorded in the striatum and the cortex (Suri and Schultz 2001; Montague and Berns 2002; Tremblay and Schultz 1999).

The exponential prediction of future reward is also consistent with our numerical simulation (Fig. 3), a simulation that uses a single spiking neuron model and 200 input neurons with synapses that change according to the above STDP learning rule (Feldman 2000; Roberts 1999). We extended our spiking model to include a parsimonious description of how dopaminergic neuron (DA) activity could be generated with a simple circuit. The reward determines the initial height of the prediction of reward (neuron C) given the time until reward and the learning parameters. If the reward is removed, then there would be a dip in the DA response because, in our simple model, the DA response is monotonically related to the sum of the reward input and the inhibition. The response to the reward after training could be further reduced by increasing the strength of the inhibitory connection. Unlike the model of Montague et al. (1996), in the model presented here, the DA neuron does not directly instruct the learning in the TD sense, but follows the combination of excitatory and inhibitory inputs. However, in the biological system, DA responses may modulate the STDP learning rule during phasic activity (Otani et al. 2003; Pawlak and Kerr 2008).

In our simulation, a dip in DA activity follows the reward response (Fig. 3c). This dip is caused by the inhibition outlasting the excitatory input from the direct reward pathway. Although a dip is not apparent in Suri and Schultz (2001), such lingering inhibition is visible in other published recordings of DA neurons (Waelti et al. 2001). Our simple representation of neurons that we used for our analytic results could not take into account some of the more subtle biological mechanisms that might modify the precise spike patterns. For instance, if the inhibitory neurons project to the presynaptic terminal of the reward inputs onto the DA neurons, then the reward response would be reduced without causing the dip in response (Houk et al. 1995).

4 Discussion

Our results that the STDP leads to a stable spike probability that anticipates future rewards leads to several predictions. First, STDP with a learning rule found in cortical neurons (Markram et al. 1997; Feldman 2000) could also be found in VTA. In addition, plasticity in cortical neurons could drive VTA activity to generate the prediction of future rewards by the STDP learning rule that has already been characterized between pyramidal cells (Markram et al. 1997; Feldman 2000). The key element, in addition to the STDP learning rule, is a series of inputs that arrive with a consistent time delay following the initial cue.

The second prediction follows from recent recordings from conditioning protocols with probabilistic reward that could be parsimoniously explained by the STDP model. Because the STDP learning dynamics change weights as a function of the *time-derivative* of postsynaptic activity, the circuitry predicted by the model does not require specialized time-derivative detecting neurons as has previously been suggested (Schultz et al. 1997; Montague et al. 1996). The circuitry can generate both the striatal and dopamine responses to reward prediction as shown in Fig. 3. The final output of dopamine neurons is then a feed forward interaction of excitation and inhibition. The recurrence of dopamine neurons is not necessarily signaling a prediction error in the TD, but can be used to gate the plasticity of inputs to allow learning to occur when a reward is present.

Finally, these results predict that any part of the brain that has an STDP learning rule with an antisymmetric component engages in forecasting the future. There is also a symmetric component in our model that is necessary for the stability of the learning dynamics. Because the dynamics presented here are generic and not dependent on the specific parameters of the learning rule, any similar STDP rule will lead to prediction of future events if the timing information is available. We have based these results on STDP involving the pairing of only one presynaptic spike with one postsynaptic spike. However, physiological spike trains seldom result in only pairing of single pre- and postsynaptic spikes, but rather spike patterns are paired with presumably nonlinear interactions (Froemke and Dan 2002) not incorporated in our model. We expect that our results will generalize to multiple spike pairings whenever the average symmetric part of the learning rule is negative and greater then the average antisymmetric part of the learning rule because of the generality of the analysis. Prediction is an essential function of the nervous system, and this analysis suggests that the synaptic learning rules that many synapses express are specifically tuned to generate predictions of future events.

References

Bi Q, Mu-Ming P (1998) Precise spike timing determines the direction and extent of synaptic modifications in cultured hippocampal neurons. J Neurosci 18:10,464–10,472

Daw ND, Dayan P (2004) Neuroscience. Matchmaking. Science 304(5678):1753–1754

Feldman DE (2000) Timing-based LTP and LTD at vertical inputs to layer II/III—pyramids in rat cortex. Neuron 27:45–56

Froemke RC, Dan Y (2002) Spike-timing-dependent synaptic modification induced by natural spike trains. Nature 416(6879):433–438

Houk J, Davis J, Beiser D (1995) Models of information processing in the basal ganglia. MIT Press, Cambridge

Izhikevich EM (2007) Solving the distal reward problem through linkage of stdp and dopamine signaling. Cereb Cortex 17(10):2443–2452

Klopf A (1988) A neuronal model for classical conditioning. Psychobiology 16:85–125

Kosko B (l986) Differential Hebbian learning. In: Denker JS (ed) AIP Conference Proceedings 151: Neural Networks for Computing. American Institute of Physics, New York, pp 277–288

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

Montague P, Berns G (2002) Neural economics and the biological substrates of valuation. Neuron 36(2):265–284

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16(5):1936–1947

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9:1057–1063

Niv Y, Duff MO, Dayan P (2005) Dopamine, uncertainty and TD learning. Behav Brain Funct 1:6

Otani S, Daniel H, Roisin MP, Crepel F (2003) Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. Cereb Cortex 13(11):1251–1256

Pawlak V, Kerr J (2008) Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. J Neurosci 28(10):2435

Rao RPN, Sejnowski TJ (2000) Predictive sequence learning in recurrent neocortical circuits. In: Solla SA, Leen TK, Muller KR (eds) Advances in neural information processing systems, vol 12. MIT Press, Cambridge pp 164–170

Roberts PD (1999) Computational consequences of temporally asymmetric learning rules: I. Differential Hebbian learning. J Compu Neurosci 7:235–246

Roberts PD (2004) Recurrent biological neural networks: the weak and noisy limit. Phys Rev E 69:031910

Roberts PD, Bell CC (2000) Computational consequences of temporally asymmetric learning rules: II. Sensory image cancellation. J Compu Neurosci 9:67–83

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1598

Song S, Miller KD, Abbott LF (1993) Competitive hebbian learning through spike-timing-dependent synaptic plasticity. Nature Neurosci 3:919–926

Suri R, Schultz W (2001) Temporal difference model reproduces anticipatory neural activity. Neural Comput 13:841–862

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge

Tremblay L, Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. Nature 398(6729):661–663

Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. Nature 412: 43–48

Wörgötter F, Porr B (2005) Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. Neural Comput 17(2):245–319