LETTER **Communicated by Patrick Roberts**

On the Asymptotic Equivalence Between Differential Hebbian and Temporal Difference Learning

Christoph Kolodziejski
kolo@bccn-goettingen.de
Bernstein Center for Computational Neuroscience, University of Göttingen, 37073 Göttingen, Germany

Bernd Porr
B.Porr@elec.gla.ac.uk
Department of Electronics and Electrical Engineering, University of Glasgow, Glasgow, Scotland

Florentin Wörgötter
worgott@bccn-goettingen.de
Bernstein Center for Computational Neuroscience, University of Göttingen, 37073 Göttingen, Germany

In this theoretical contribution, we provide mathematical proof that two of the most important classes of network learning—correlation-based differential Hebbian learning and reward-based temporal difference learning—are asymptotically equivalent when timing the learning with a modulatory signal. This opens the opportunity to consistently reformulate most of the abstract reinforcement learning framework from a correlation-based perspective more closely related to the biophysics of neurons.

1 Introduction

Network learning subdivides into supervised (SL), reinforcement (RL), and correlation based (Hebbian, CL) according to the existence and role of an error signal for controlling the learning (Doya, 2000). In SL, explicit error signals exist, in RL networks learn from unspecific reinforcement signals (rewards), whereas in CL, learning takes place as a self-organization process relying on multiplicative signal correlations only, hence without an error signal. Convergence proofs exist for some generic cases of these learning mechanisms (Widrow & Hoff, 1960; Oja, 1982; Hopfield, 1982; Linsker, 1988; Dayan & Sejnowski, 1994). The equivalence of the different rules, however, remains a pressing question as it would allow extending conclusions on convergence across different mechanisms.

Recently there have been several contributions toward solving this question (Izhikevich, 2007; Roberts, Santiago, & Lafferriere, 2008; Florian, 2007; Potjans, Morrison, & Diesmann, 2009), which presented specific solutions to be discussed later (see section 4). Thus, there is more and more evidence emerging that Hebbian learning and reinforcement learning can be brought together under a more unifying framework. Such an equivalence would have substantial influence on our understanding of network learning, as these two types of learning could be interchanged under these conditions.

Thus, the goal of this study is to prove that the most influential form of RL, which relies on the temporal difference (TD) learning rule (Sutton, 1988), is asymptotically equivalent to CL and convergent over wide parameter ranges when using a third factor as a gating signal, together with a differential Hebbian emulation of CL.

The idea of differential Hebbian learning was first used by Klopf (1988) to describe classical conditioning relating to the stimulus substitution model of Sutton and Barto (1981). Around the same time, Kosco (1986) used differential Hebbian learning in a machine learning context and examined its features. One of its most important features is the implicit introduction of negative weight changes (LTD), which leads to intrinsic stabilization properties in networks. Earlier approaches had to explicitly introduce negative weight changes into the learning rule, for example, by way of a threshold (Oja, 1982). Negative weight changes (LTD) had first been discovered physiologically by Dudek and Bear (1992) and later in conjunction with the timing of pre- and postsynaptic activity by Markram, Lübke, Frotscher, & Sakmann (1997) (spike-timing-dependent plasticity).

One drawback of reinforcement learning algorithms, like temporal difference learning, is their use of discrete time and discrete nonoverlapping states. In real neural systems, time is continuous, and the state space can be represented only by the activity of neurons, many of which will be active at the same time and for the same space. This creates a rather continuous state-space representation in real systems. In order to allow overlapping states or for generalizing over a wider range of input regions, RL algorithms are usually extended by value function approximation methods (Sutton & Barto, 1998). However, while biologically more realistic (Tamosiunaite et al., 2008), this makes initially elegant RL algorithms often quite opaque, and convergence cannot be guaranteed (Tsitsiklis & Van Roy, 1997; Wiering, 2004). Here we are not concerned with function approximation, but instead address the question of how to transform an RL algorithm (TD-learning) to continuous time using differential Hebbian learning with a third factor and remaining fully compatible with neuronally plausible operations. Only a few other approaches to formulate RL in continuous time and space exist (Baird, 1993; Doya, 1996, 2000). The differences to our approach are discussed in section 4.

Biophysical considerations about how such a third factor might be implemented in real neural tissue are of secondary importance for this study.

At this stage, we are concerned with a formal proof only. Some biophysical aspects are treated in section 4, though.

This letter proceeds as follows. First, we describe the basics of TD and differential Hebbian learning. This will lead us to a general signal (= "state") structure that is necessary to put both frameworks together. Next, we take a close look at the differential Hebbian learning rule and its differential equation. This gives us definitions that we will use for the formal analysis of the equivalence. A simulated network will then be investigated, and some technical constraints will be evaluated before we finish with a general discussion. Details of analytical calculations used in the main text are given in the appendix.

1.1 Emulating RL by Temporal Difference Learning. Reinforcement learning maximizes the rewards $r(s)$ an agent will receive in the future when following a policy π traveling along states s. The return R is defined as the sum of the future rewards: $R(s_i) = \sum_k \gamma^k r(s_{i+k+1})$, where future rewards are discounted by a factor $0 < \gamma < 1$. One central goal of RL is to determine the values $V(s)$ for each state given by the average expected return $E^\pi\{R\}$, which can be obtained when following policy π. Many algorithms exist to determine the values, almost all of which rely on the temporal difference (TD) learning rule, equation 1.1 (Sutton, 1988).

Every time the agent encounters a state s_i, it updates the value $V(s_i)$ with the discounted value $V(s_{i+1})$ and the reward $r(s_{i+1})$ of the next state that is associated with the consecutive state s_{i+1}:

$$V(s_i) \to (1-\alpha)V(s_i) + \alpha(r(s_{i+1}) + \gamma V(s_{i+1})), \tag{1.1}$$

where α is the learning rate. This rule is called TD($\lambda = 0$), shortened as TD(0), as it only evaluates adjacent states. For values of $\lambda \neq 0$, more of the recently visited states are used for value-function update. TD(0) is by far the most influential RL learning rule as it is the simplest way to ensure optimality of learning (Dayan & Sejnowski, 1994; Sutton & Barto, 1998).

1.2 Differential Hebbian Learning with a Third Factor. In traditional Hebbian learning, the change of a weight ρ relies on the correlation between input $u(t)$ and output $v(t)$ of a neuron: $\rho'(t) = \tilde{\alpha} \cdot u(t) \cdot v(t)$, where $\tilde{\alpha}$ is the learning rate and prime denotes the temporal derivative. If we consider the change of the postsynaptic signal and therefore replace $v(t)$ with $v'(t)$, we will arrive at differential Hebbian learning. Then, negative weight changes are also possible, and this yields properties similar to experimental neurophysiological observations (spike-timing-dependent plasticity; Markram et al., 1997).

In order to achieve the equivalence (see section 4 for a discussion), we introduce a local third modulatory factor $M(t)$ responsible for controlling

the learning (Porr & Wörgötter, 2007). The three-factor differential Hebbian learning rule is then

$$\rho_k'(t) = \tilde{\alpha} \cdot u_k(t) \cdot v'(t) \cdot M(t), \tag{1.2}$$

where $u_k(t)$ is the considered presynaptic signal and

$$v(t) = \sum_k \rho_k(t) u_k(t), \tag{1.3}$$

the postsynaptic activity of a model neuron with weights $\rho_k(t)$. We will assume in the following that our modulatory signal $M(t)$ is either 1 or 0, thus represented by a step function. By means of the learning rate $\tilde{\alpha}$, we can set the ratio between the weight change over the weight ρ'/ρ to be significantly smaller than the state change over the state value u'/u.

2 Analytical Derivation

We are going to analyze the weight change of weight $\rho_i(t)$ when considering three consecutive signals $u_{i-1}(t)$, $u_i(t)$, and $u_{i+1}(t)$, with the index i representing a temporal (and not, e.g., a spatial) ordering. The third factor $M(t)$ opens a time window for the considered weight $\rho_i(t)$ in which changes can occur. Although this time window could be located anywhere depending on the output $v(t)$, it should be placed at times where a state change occurs, and thus at the beginning or end of the state $s_i(t)$, as it makes sense only if states correlate with temporally neighboring states.

The relation between state $s(t)$ and input $u(t)$ is determined by a convolution: $u(t) = \int_0^\infty s(z)h(t-z)dz$ with filter function $h(t)$, which is identical for all states. As we are using only states that are either on or off during visiting duration S, the input functions $u(t)$ do not differ between states. Therefore, we will use $u_i(t)$ (with index i) having a particular state in mind and $u(t)$ (without index i) when pointing to functional development.

Furthermore, we define the time period between the end of a state $s_i(t)$ and the beginning of the next state $s_{i+1}(t)$ as T ($T < 0$ in case of overlapping states). Concerning the modulatory third factor $M(t)$, we define its length as L and the time period between the beginning of $M(t)$ and the end of the state $s_i(t)$ as O. These four parameters (L, O, T, and S) are constant over states and are displayed in detail in Figure 1B.

2.1 Differential Equation. For the following analysis, we need to substitute equation 1.3 in equation 1.2 and solve this differential equation, which consists of a homogeneous and an inhomogeneous part:

$$\begin{aligned}\rho_i'(t) = {} & \tilde{\alpha} \cdot M(t) \cdot u_i(t)[u_i(t) \cdot \rho_i(t)]' \\ & + \tilde{\alpha} \cdot M(t) \cdot u_i(t) \left[\sum_{j \neq i} u_j(t) \cdot \rho_j(t) \right]',\end{aligned} \tag{2.1}$$

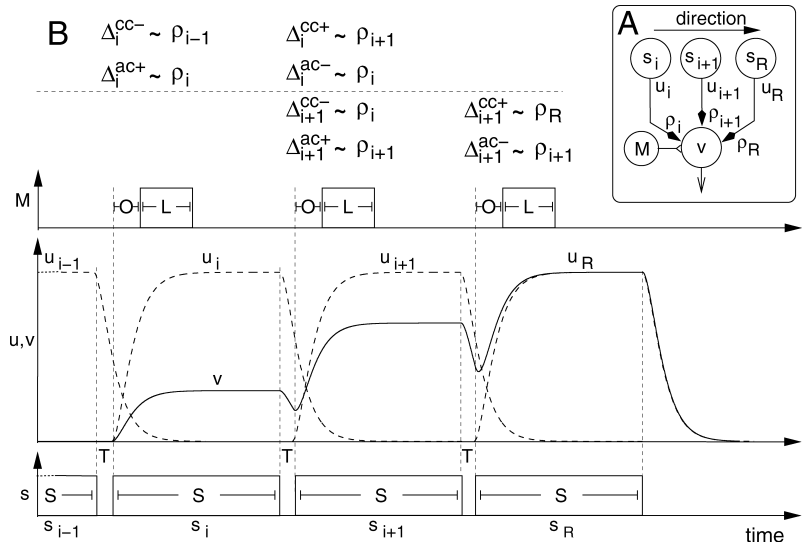

Figure 1: (A) The setup. Three states, including the rewarded state, converge on the neuron, which learns according to equation 1.2. The modulatory factor M will influence learning at synapse ρ_i. The states s will be active according to the direction arrow. (B) The signal structure. The lower part shows the states s_i, which have a duration of length S. We assume that the duration for the transition between two states is T. In the middle, the output v and the signals u are depicted. Here u is given by $u(t) = \int_0^S (e^{-a(t-z)} - e^{b(t-z)})\, dz$. The third factor M is released for the duration L after a time delay of O and is shown in the upper part. For each state, the weight change separated into autocorrelation $\Delta^{ac\pm}$ and cross-correlation $\Delta^{cc\pm}$ and their dependence on the weights according to equations 2.4 and 2.8 are indicated.

where the modulator $M(t)$ is defining the integration boundaries (see equations 2.7 and A.2). The first summand leads to the homogeneous solution, which we will define as autocorrelation $\rho^{ac}(t)$. The second summand(s) will lead to the inhomogeneous solution, and this we will define as cross-correlation $\rho^{cc}(t)$. Together we have $\rho(t) = \rho^{ac}(t) + \rho^{cc}(t)$.

In general, the overall change of the weight $\rho_i(t)$ after integrating over the visiting duration of $s_i(t)$ and $s_{i+1}(t)$ and using the modulatory signal $M(t)$ is $\Delta\rho_i =: \Delta_i = \Delta_i^{ac} + \Delta_i^{cc}$

Without restrictions, we can now limit further analysis of equation 2.1, in particular of the cross-correlation term, to the case of $j = i \pm 1$ as the modulatory factor affects only the weights of the last and the following state.

Weight changes in general are slow, and we can assume a quasi-static process ($\frac{u_i'}{u_i} \gg \frac{\rho_i'}{\rho_i}$, $\tilde{\alpha} \to 0$). As a consequence, the derivatives of ρ on the right-hand side of equation 2.1 can be neglected.

The solution of the autocorrelation $\rho_i^{ac}(t)$ is then in general (see section A.1)

$$\rho_i^{ac}(t) = \rho_i^{ac}(t_0) e^{\tilde{\alpha} \cdot M(t) \cdot \frac{1}{2}[u_i^2(t) - u_i^2(t_0)]}, \tag{2.2}$$

and the overall weight change with the third factor being present between $t = O$ and $t = O + L$ as well as between $t = O + S + T$ and $t = O + S + T + L$ (see Figure 1B) is therefore

$$\Delta_i^{ac} = \rho_i \left(e^{\tilde{\alpha} \frac{1}{2}[u_i^2(O+L) - u_i^2(O) + u_i^2(O+S+T+L) - u_i^2(O+S+T)]} - 1 \right). \tag{2.3}$$

Using again the argument of a quasi-static process ($\tilde{\alpha} \to 0$), we can expand the exponential function to the first order:

$$\begin{aligned} \Delta_i^{ac} &:= -\tilde{\alpha} \rho_i \kappa - o(\tilde{\alpha}^2) \\ &= -\tilde{\alpha} \rho_i \frac{1}{2} \big[u_i^2(O) - u_i^2(O+L) \\ &\quad + u_i^2(O+S+T) - u_i^2(O+S+T+L) \big] + o(\tilde{\alpha}^2), \end{aligned} \tag{2.4}$$

where we have defined κ to be positive:

$$\begin{aligned} \kappa(L, O, T, S) &= \frac{1}{2}(u^2(O) - u^2(O+L)) \\ &\quad + \frac{1}{2}(u^2(O+S+T) + -u^2(O+S+T+L)) \\ &= -(\kappa^+(L, O) + \kappa^-(L, O, T, S)). \end{aligned} \tag{2.5}$$

Here we have left out the index i as all states are identical and split κ into κ^+ and κ^- representing the first (positive slope of signal u; see Figure 1B) and the second (negative slope of signal u; see Figure 1B) occurrence of the third factor. To this end, we also have to split Δ_i^{ac} into $\Delta_i^{ac+} = \tilde{\alpha} \rho_i \kappa^+$ and $\Delta_i^{ac+} = \tilde{\alpha} \rho_i \kappa^-$.

Next we investigate the cross-correlation $\rho^{cc}(t)$ again under the assumption of a quasi-static process. This leads us to

$$\begin{aligned} \rho_i^{cc}(t) = \rho_i^{cc}(t_0) + \tilde{\alpha} \Bigg[& \rho_{i-1} \int_0^t M(z) \cdot u_i(z) u'_{i-1}(z)\, dz \\ & + \rho_{i+1} \int_0^t M(z) \cdot u_i(z) u'_{i+1}(z)\, dz \Bigg], \end{aligned} \tag{2.6}$$

where we can use the third factor $M(t)$ to define the integration boundaries. The third factor is present between $t = O$ and $t = O + L$ for the first summand and $t = O + S + T$ and $t = O + S + T + L$ for the second summand (see Figure 1B). Furthermore, as all signals are identical, we can shift between signals by $t = S + T$. In detail, this is $u_{i-1}(t) = u_i(t + S + T)$ and $u_{i+1}(t) = u_i(t - S - T)$. The overall weight change can then be split into Δ_i^{cc-} and Δ_i^{cc+}:

$$\Delta_i^{cc} = \tilde{\alpha} \cdot \rho_{i-1} \int_O^{O+L} u_i(z) u_i'(z + S + T)\, dz + \tilde{\alpha} \cdot \rho_{i+1} \int_{O+S+T}^{O+S+T+L} u_i(z) u_i'(z - S - T)\, dz \tag{2.7}$$

$$:= \underbrace{\tilde{\alpha} \cdot \rho_{i-1} \cdot (-\tau^-)}_{\Delta_i^{cc-}} + \underbrace{\tilde{\alpha} \cdot \rho_{i+1} \cdot \tau^+}_{\Delta_i^{cc+}}. \tag{2.8}$$

Here we defined $\tau^\pm$ as being positive:

$$\tau^-(L, O, T, S) = -\int_O^{O+L} u(z) u'(z + S + T)\, dz \tag{2.9}$$

$$\tau^+(L, O, T, S) = \int_O^{O+L} u(z + T + S) u'(z)\, dz, \tag{2.10}$$

which also is independent of i.

Both $\tau^\pm$ and κ depend on the actual signal shape $u(t)$ used and the values for the parameters L, O, T, and S.

2.2 Asymptotic Equivalence. Without restrictions, we can now limit the discussion to the situation in Figure 1A, where we have one intermediate state transition (from s_i to s_{i+1}) and a final one (from s_{i+1} to s_R), which leads to the reward. It represents the fact that rewards are usually associated with unconditioned stimuli, which will lead to strong, insuppressible responses. This indicates that a strong prewired connection from the receptor of the unconditioned stimulus onto the target neurons exists. For simplicity, this structure here is incorporated in the network as state s_R. Thus, three-factor differential Hebbian will influence two synaptic connections ρ_i and ρ_{i+1} of states s_i and s_{i+1} respectively, which directly project onto neuron v.

Figure 1B shows a realistic situation of state transitions leaving the old state s_{i-1} and entering the new state s_i and so on. The signals as such could be considered as membrane voltages or firing rates of neurons.

The lower part of the signal traces indicates the times when the states s are visited and the upper part when the third factor M is either on or

off. The part above the signal traces highlights the different contributions (Δ-values) to the overall weight change defined in the previous section.

First, we consider the weight change of ρ_i. This results from the transition between s_{i-1} and s_i from the visiting state s_i itself and from the transition between s_i and s_{i+1}. The short learning period at the beginning of the signal u_i will cause a negative weight change Δ_i^{cc-} because of the correlation between the negative derivative of u_{i-1} and the positive value of u_i. Additionally, there is a weight change Δ_i^{ac+} caused by the signal itself. Due to the positive slope of the signal u_i at the beginning of the state, the contribution will be positive. The next learning interval occurs when the state s_i has been left, and the signal u_i already decays. Thus, having a negative slope, the autocorrelation causes a weight change $\Delta_i^{ac-} < 0$. The fourth contribution yields a positive weight change Δ_i^{cc+} because the positive derivative of the next state signal u_{i+1} correlates with the positive value of signal u_i of state s_i. The same sequence exists when the next state transition occurs, yielding contributions for the Δ_{i+1}-values. During the first trial (where all weights are zero), only the cross-correlation Δ_{i+1}^{cc} yields a contribution due to the finding of the reward.

In general the weight after a single trial is the sum of the old weight ρ_i and the four Δ_i-values:

$$\rho_i \to \rho_i + \Delta_i^{ac-} + \Delta_i^{ac+} + \Delta_i^{cc-} + \Delta_i^{cc+}. \tag{2.11}$$

Using equations 2.4 and 2.8, we can reformulate equation 2.11 into

$$\rho_i \to \rho_i + \tilde{\alpha} \cdot (\kappa^+ + \kappa^-) \cdot \rho_i - \tilde{\alpha} \cdot \tau^- \cdot \rho_{i-1} + \tilde{\alpha} \cdot \tau^+ \cdot \rho_{i+1}. \tag{2.12}$$

Substituting $\kappa = -(\kappa^- + \kappa^+)$, $\alpha = \tilde{\alpha} \cdot \kappa$ and $\gamma^\pm = \tau^\pm/\kappa$, we get

$$\rho_i \to (1-\alpha) \cdot \rho_i - \alpha \cdot \gamma^- \cdot \rho_{i-1} + \alpha \cdot \gamma^+ \cdot \rho_{i+1}. \tag{2.13}$$

The convergence on $\rho_i = \gamma \rho_{i+1}$ is a property of these kinds of equations (see section A.2). Therefore, we can use equation A.15 in the appendix with $\mu = \gamma$, $\varepsilon_1 = \tau^+/\kappa = \gamma^+$, and $\varepsilon_2 = \tau^-/\kappa = \gamma^-$, although only if the values, namely κ and $\tau^\pm$, are strictly positive (will be discussed in section 2.3). This gives us

$$\frac{1}{\gamma} = \frac{1}{2\gamma^+} + \sqrt{\frac{1}{(2\gamma^+)^2} + \frac{\gamma^-}{\gamma^+}}, \tag{2.14}$$

and our weight development can be simplified to

$$\rho_i \to (1-\alpha) \cdot \rho_i + \alpha \cdot \gamma \cdot \rho_{i+1}. \tag{2.15}$$

At this point, we can make the transition from weights ρ_i (differential Hebbian learning) to states $V(s_i)$ (temporal difference learning). Additionally, we note that sequences terminate only at $i + 1$; thus, this index will capture the reward state s_R and its value $r(s_{i+1})$, while this is not the case for all other indices (see section 4 for a detailed discussion of rewards at nonterminal states). Consequently this gives us an equation almost identical to equation 1.1:

$$V(s_i) \to (1 - \alpha)V(s_i) + \alpha \cdot \gamma[r(s_{i+1}) + V(s_{i+1})], \tag{2.16}$$

where one small difference arises, as in equation 2.16, the reward is scaled by γ. However, this has no influence, as numerical reward values are arbitrary. Thus, if learning follows this third-factor differential Hebbian rule, weights will converge to the optimal estimated TD values. This proves that under some conditions for the signal shape and the parameters S, L, O, and T (which influence whether $\kappa > 0$ and $\tau^{\pm} > 0$), TD(0) and the proposed three-factor differential Hebbian learning are indeed asymptotically equivalent.

2.3 Analysis of γ (as Well as κ and $\tau^{\pm}$). Here we will take a closer look at the signal shape and the parameters (L, O, T, and S) which influence the values of κ (see equation 2.5) and $\tau^{\pm}$ (see equations 2.9 and 2.10), and therefore γ (see equation 2.14). For guaranteed convergence, these values are constrained by two conditions (see section A.2): $\tau^{\pm} \geq 0$ and $\kappa > 0$ (where $\kappa = 0$ is allowed in case of $\tau^{\pm} = 0$). A nonpositive value of κ would lead to divergent weights ρ and negative values of $\tau^{\pm}$ to oscillating weight pairs (ρ_i, ρ_{i+1}). However, even if fulfilled, these conditions will not always lead to meaningful weight developments. In particular, $\tau^{\pm}$-values of 0 leave all weights at their initial weight value, and discount factors, which are represented by γ-values exceeding 1, are usually not considered in reinforcement learning (Sutton & Barto, 1998). Thus, it makes sense to introduce more rigorous conditions and demand that $0 < \gamma \leq 1$ and $\kappa > 0$.

Furthermore, as these conditions depend on the signal shape, the following theoretical considerations need to be guided by biophysics. Hence, we will discuss neuronally plausible signals that can arise at a synapse. This constrains u to functions that possess only one maximum and divide the signal into a rising and a falling phase.

One quite general possibility for the shape of the signal u is the function used in Figure 1, for which we investigate the area of convergence. We have three parameters to be varied, as we do not have to consider the parameter S if we take this value to be large compared to $|T|$, L, or O. For this, Figure 2 shows the γ-value in 3 different panels. In each panel, we varied the parameters O and T from minus to plus 2 P, where P is the time the signal u needs to reach the maximum. In each of the panels, we plot γ-values for a particular fraction of L/P.

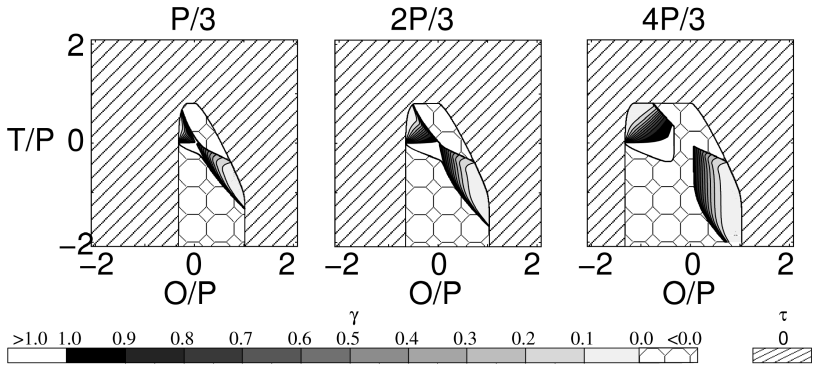

Figure 2: Shown are γ-values dependent on the ratio O/P and T/P for different values of L/P (1/3, 2/3, and 4/3). Here, P is the length of the rising as well as the falling phase. The shape of the signal u is identical to the one used in Figure 1 and is given by $u(t) = \int_0^S (e^{-a(t-z)} - e^{b(t-z)})\, dz$ with parameters $a = 0.006$ and $b = 0.066$. The individual figures are subdivided into a patterned area where the weights will diverge ($\kappa \leq 0$; see equation 2.4), a striped area where no overlap between both signals and the third factor exists, and into a white area that consists of γ-values, which, however, are beyond a meaningful range ($\gamma > 1$). The detailed gray shading represents γ-values ($0 < \gamma \leq 1$) for which convergence is fulfilled.

The gray shading displays in detail the γ-values for which the condition is fulfilled, whereas white represents those areas for which we receive $\gamma > 1$. The striped area indicates parameter configurations for which no overlap between two consecutive signals and the third factor exists ($\tau = 0$), and for the patterned regions, κ is smaller than zero.

If the L-value is greater than $P-O-T$, the area of convergence does not depend on L anymore, as the third factor then reaches a plateau as well as covers the whole falling phase of the signal u. On the contrary, if the L-value reaches the rising phase of the consecutive state, the area of convergence decreases again (not shown).

For positive O-values, there exist γ-values that are independent of (negative) T-values. Hence, if states overlap ($T < 0$), the γ-value is invariant with respect to the degree of overlap. This is an important aspect, as value function approximation methods often use overlapping kernels to represent features. In a biological context, this corresponds to overlapping receptive fields providing the input to the system. We find that in these cases, γ remains unaffected by the degree of (receptive field) overlap, which in general is different for any two input units.

To extend these considerations to more general but smoother shapes, we can therefore Taylor expand both the rising and the falling phases to the second order. With these constraints, γ can be calculated analytically (see

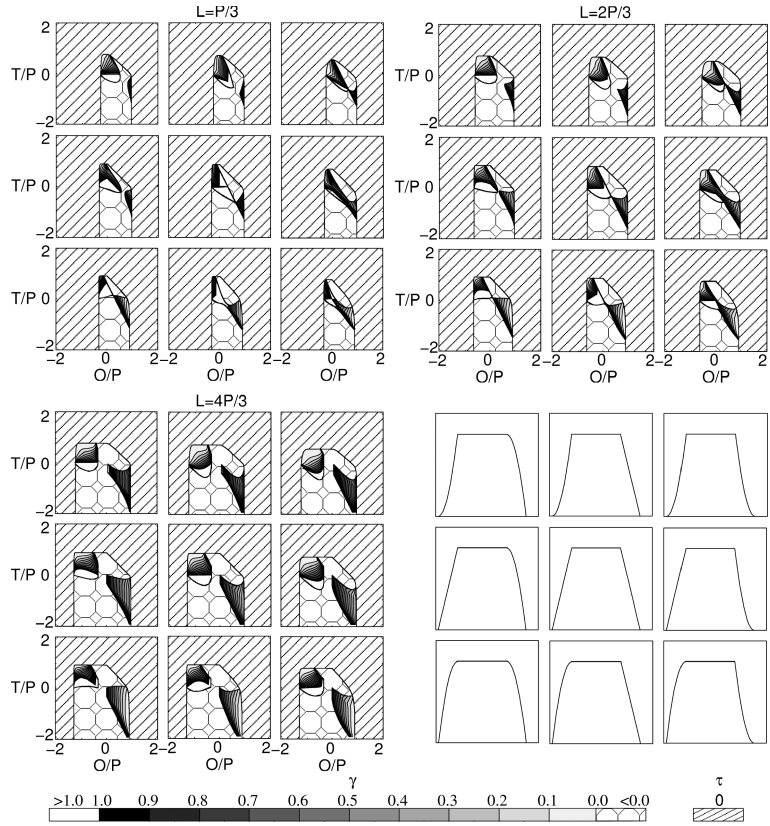

Figure 3: Shown are γ-values for different shapes of the signal u dependent on the ratio O/P and T/P for three different values of L/P. The upper left panel is for $L/P = 1/3$, the upper right for $L/P = 2/3$, and the lower left panel for $L/P = 4/3$, where P is the length of the rising as well as the falling phase. The different shapes are shown in the lower right, and the relevant equation (equation A.16), is given in section A.3. The rows represent different η-values (top to bottom: 0, 1, and 2) and the columns different ξ-values (left to right: 0, 1, and 2). The individual figures are subdivided into a patterned area where the weights will diverge ($\kappa \leq 0$; see equation 2.4), a striped area where no overlap between both signals and the third factor exists, and into a white area that consists of γ-values, which, however, are beyond a meaningful range ($\gamma > 1$). The detailed gray shading represents γ-values ($0 < \gamma \leq 1$) for which convergence is fulfilled.

section A.3) and is then plotted against O and T in Figure 3 for nine input functions, shown in the lower right. On the upper left, the ratio between the duration of the third factor and P was set to one-third, in the upper right to two-thirds, and on the lower left to four-thirds. Analogous to the

exponential function, the area of convergence increases with increasing L-values.

If comparing the results of Figure 3 with Figure 9 (in section A.3), both figures match each other, as we have used the same derivations as for Figure 3 in the appendix. However, even if more general filter functions are used, like the exponential one used for Figure 2, both figures still match quite well, especially in the regions where the system diverges and where it stays constant.

In summary, Figures 2, 3, and 9 show clearly that the area of convergence changes only gradually and the area as such is increasing with increasing duration of the third factor. Altogether, it shows that for a general neuronally plausible signal shape u, the condition for asymptotic equivalence between temporal difference learning and differential Hebbian learning with a third factor is fulfilled for a wide parameter range and thus all realistic relative timing intervals between state activations and third factor.

3 Simulation of a Small Network

In this section, we show that we can reproduce the behavior of TD-learning in a small, linear network of neurons designed according to our algorithm. Obtained weights of the differential Hebbian learning neuron represent the corresponding TD-value (see Figure 4A, inset). It is known that in a linear TD-learning system, values at the end of learning will follow an exponential function with a decay rate given by the discount factor γ. This is shown in Figure 4A. In Figure 4B, we also investigate the assumption of a quasi-static process.

Details of this simulation are as follows. The network consists of N states s, which are connected to a neuron v, which uses differential Hebbian learning. The modulatory signal is added by an additional neuron M. The network is shown in the inset of Figure 4A. The states are indexed such that the state closest to the reward has index 1 (hence, the reward has the index 0). At the beginning of learning, all weights are initialized to 0 except the weight connected with the reward. Each trial begins with state N approaching the reward at which a trial is terminated; thus, each state becomes active once.

The weights of the states connected to the differential Hebbian learning neuron are shown in Figure 4A for three different γ-values after learning. States indexed with higher numbers (hence, farther away from the reward) have smaller weights, and the relation $\rho_{i+1} = \gamma\rho_i$ where i indicates the distance to the reward holds for each γ-value. This is indicated by an exponential fit. It also should be noted that the weights at states far away from the reward deviate from the exponential fit, but only for the highest γ-value. This is an effect caused by the finite number of states and at the same time by a γ^+-value that is higher than 1 (see section 4.1 for details).

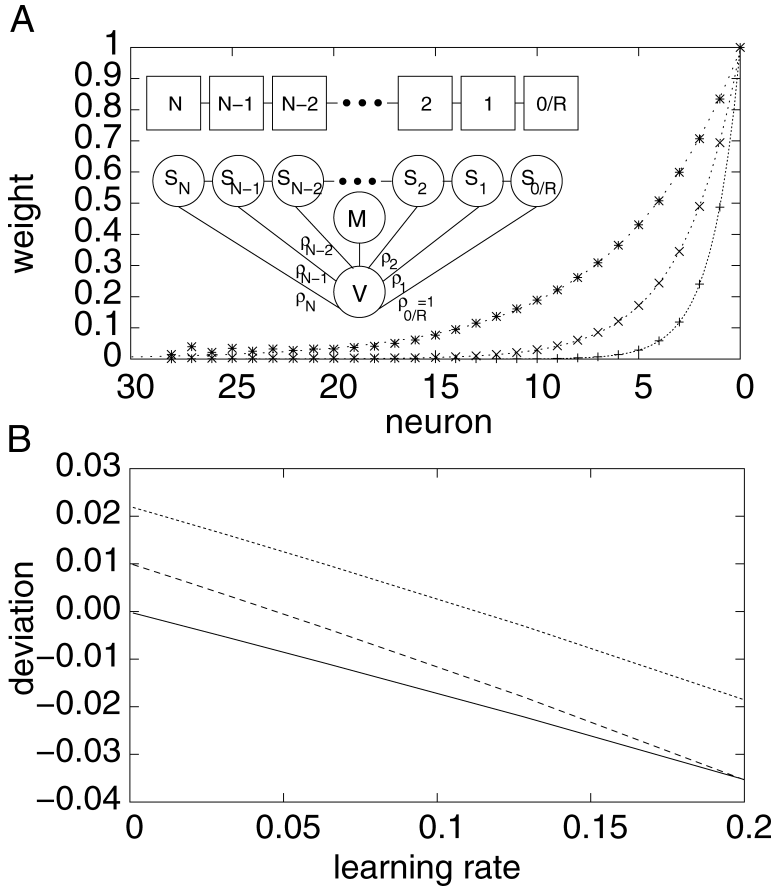

Figure 4: Weights of a differential Hebbian learning neuron. (A) The inset shows the arrangement of the states (top) and the network used for the simulations (bottom). We used $N = 30$ for our simulations. The weights of the network and their exponential fit for three γ-values are plotted. (B) The dependence of the weights to the learning rate. The difference of the weight closest to the reward ($\rho_1 = \gamma \cdot \rho_0 = \gamma \cdot 1$) and the calculated γ-value are plotted here and can be fitted by a logarithmic function [$f(x) \propto \log(1 - x)$]. The γ-values used are ($*$, dotted) $\gamma = 0.835697$ [$S = 3000$, $T = 330$, $L = 650$, $O = -220$], (x, dashed) $\gamma = 0.710166$ [$S = 3000$, $T = 300$, $L = 650$, $O = -220$], (+, solid) $\gamma = 0.507729$ [$S = 3000$, $T = 300$, $L = 550$, $O = -220$]. The shape of the filter used here is identical to the shape used in Figure 2, and the learning rate used in A is 0.12.

In these systems, learning rates are usually in the range of 10^{-5} to 10^{-2} (Porr & Wörgötter, 2003, 2007). The question arises whether in this range, the assumption of a quasi-static process will hold. If it holds, we would expect that the weight closest to the reward (ρ_1) will exactly reach the value of γ after learning. In Figure 4B, the deviation from this expectation given by $\rho_1 - \gamma$ is plotted against the learning rate. As indicated by equation 2.5, the deviation increases with increasing learning rates, but remains small up to a rate of 10^{-1}, which is well in the range of useful learning rates. The actual shape of the curves is a consequence of different interacting processes depending, for example, on the total number of states (see section 4.1) and others.

Looking at higher γ-values, it is apparent that the effect of a finite number of states behaves antagonistically to the deviation caused by the increased learning rate—the weight after learning is shifted to higher values (independent of the learning rate). Therefore, if higher γ-values (or a smaller number of states) are used, the simulated weight ρ_1 may be modified such that it will be identical to the calculated γ-value even if finite learning rates larger than 0 are used. However, this will not correct the simulated γ-value as such since the weights are then no longer arranged exponentially (indicated by Figure 4A).

4 Discussion

4.1 Technical Discussion. Most realistic learning examples are not affected by these constraints. Thus, readers who are not interested in these special aspects may choose to skip this section.

4.1.1 Quasi-Static Process, $\alpha \ll 1$. In this proof, the assumption of a quasi-static process has been used three times. First, we used this assumption for solving the differential equation, equation 2.1, of the weight change. More precisely, we neglected the derivative of the weight on the right side of that equation. If we considered this term as well, we would get an inverse square root function $1/\sqrt{1-\alpha\,\nu}$ instead of the exponential function $e^{\alpha\frac{1}{2}\nu}$ (see equation 2.2), which we assumed during our proof, where the parameter ν is defined here as

$$\nu := u_i^2(O+L) - u_i^2(O) + u_i^2(O+S+T+L) - u_i^2(O+S+T). \quad (4.1)$$

The inverse square root function, however, has similar properties and expands equally with respect to the first order around small values of α compared to the exponential function, justifying constraint $\alpha \ll 1$ here.

Second, we truncated the expansion of equation 2.3 after the first order, which is allowed only for $\alpha \ll 1$. Would the necessary condition $\kappa > 0$ be affected if we had not truncated the expansion? Considering ν as defined in equation 4.1 and equation 2.5, we observe that a positive value of κ

corresponds to a negative value of ν and will lead to a negative weight change of Δ_i^{ac}. Hence, given a negative value of ν (a necessary condition for $\kappa > 0$ if taking only the first-order terms of the expansion into account), this leads directly to a negative weight change of Δ_i^{ac} in equation 2.3. This is due to the properties of the exponential function ($e^x - 1 < 0 \quad \forall \quad x < 1$) or of the inverse square root function ($1/\sqrt{1-x} - 1 < 0 \quad \forall \quad x < 1$), and, as a consequence, constraining to $\alpha \ll 1$ is allowed here as well.

Third, we neglected, due to $\alpha \ll 1$, the variability of the homogeneous solution (ρ^{ac}; see equation 2.2) in order to calculate the inhomogeneous solution (ρ^{cc}; see equation 2.6) of the weight ρ. However, taking the variability into consideration will not affect the linearity with respect to ρ. This is because equation 2.8 can be directly split into $\tau^{\pm}$ and ρ, and the additional homogeneous solution will change only the integral (see equations 2.10 and 2.9), which leads to $\tau^{\pm}$ and, as the solution does not rely on ρ, will not complicate the linear decomposition.

4.1.2 Reward Only at the End of a Sequence. In most physiological experiments (Schultz, Apicella, Scarnati, & Ljungberg, 1992; Montague, Dayan, & Sejnowski, 1996; Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006) the reward is given at the end of the stimulus sequence. Our assumption that the reward state is a terminating state and is therefore only at the end of the learning sequence conforms to this paradigm. However, for TD in general, we cannot assume that the reward is provided only at the end. Differential Hebbian learning will then lead to a slightly different solution compared to TD-learning. This solution has already been discussed in a another context (Dayan, 2002). Specifically, the difference in our case is the final result for the state-value after convergence for states that provide a reward. We get $V(s) \to \gamma V(s_{i+1}) + r(s_{i+1}) - r(s_i)$ compared to TD-learning: $V(s) \to \gamma V(s_{i+1}) + r(s_{i+1})$). It would be interesting to assess with physiological and or behavioral experiments which of the two equations more closely represents experimental reality. To do so, one has to guarantee that the reward given at the end is worth the costs of the animal incurred until reaching it (Hassani, Cromwell, & Schultz, 2001).

4.1.3 Finite Number of States. If we consider just a finite number of states without periodic boundary conditions and assume that the same state neuron s_0 always starts the whole sequence, the corresponding weight will not converge to $\rho_0 = \gamma\ \rho_1$ but to $\rho_0 = \gamma^+\ \rho_1$ due to the missing ρ_{-1} weight (see equations 2.15 and A.11). However, in this case, the gradual order of the weight values is disturbed only if the γ^+-value is larger than 1.

4.1.4 Stochastically Uncertain Environments. It is known that in stochastically uncertain environments, the TD-values converge with probability one only when the learning rate decreases (Kushner & Clark, 1978; Dayan & Sejnowski, 1994). In our implementation, the signal $M(t)$ is constant. If

it were instead implemented to diminish during repeated encounters with the same state, it would immediately incorporate the property of decreasing learning rates too.

4.1.5 Non-Markov. In the algorithm presented, each TD-value arises from the interaction of three states. Hence, for a considered state, it is of relevance from where it had been reached. Thus, strictly speaking, the algorithm is history dependent and violates the Markov property required for TD-learning (Sutton & Barto, 1998). This has no deeper consequence, as the standard trick for non-Markovian systems, which often are also encountered in conventional RL problems, can be applied here in the same way. It is possible, without much effort, to design a network where in a higher network layer, states are being concatenated into new larger states, which now obey the Markov property. In many practical applications, this is not even required as the value-gradient field will build up toward the reward regardless of the non-Markovian algorithm presented here. An implementation of SARSA (Singh, Jaakkola, Littman, & Szepesvári, 2000) using our algorithm behaves in this way (data not shown).

4.2 General Discussion. The TD-rule has become the most influential algorithm in reinforcement learning because of its tremendous simplicity and proven convergence to the optimal value function (Sutton & Barto, 1998). It had been successfully transferred to control problems too in the form of Q- or SARSA learning (Watkins & Dayan, 1992; Singh et al., 2000), which uses the same algorithmic structure while maintaining similar advantageous mathematical properties (Watkins & Dayan, 1992).

In this study, we have shown that TD(0)-learning and differential Hebbian learning modulated by a third factor are asymptotically equivalent under certain conditions. This proof relies on commonly applicable, fairly general assumptions, thus rendering a generic result that does not constrain the design of larger networks. It has long been suspected that RL in neuronal tissue would have to rely on the use of a third factor in a Hebb rule (Schultz, 1998), and several earlier results have pointed to the possibility of an equivalence between RL and CL. Izhikevich (2007) solved the distal reward problem using a spiking neural network, yet with fixed exponential functions (Gerstner, Kempter, van Hemmen, & Wagner, 1996) to emulate differential Hebbian characteristics. His approach is related to neurophysiological findings on spike-timing-dependent plasticity (STDP; Markram et al., 1997). Each synapse learned the correlation between conditioned stimuli and unconditioned stimuli (e.g., a reward) through STDP and a third signal. Furthermore Roberts, Santiago, and Lafferriere (2008) showed that asymmetrical STDP and TD-learning are related. In our differential Hebbian learning model, in contrast to the work described above, STDP emerges automatically because of the use of the derivative in the postsynaptic potential (see equation 1.2). The relation between STDP and differential Hebbian

learning and its asymptotic equivalence when using serial states was discussed in Roberts (1999). Rao and Sejnowski (2001) showed that using the temporal difference will directly lead to STDP, but they could not provide a rigorous proof for the equivalence. Recently it has been shown that the online policy-gradient RL algorithm (OLPOMDP) developed by Baxter, Bartlett, and Weaver (2001) can be emulated by spike-timing-dependent plasticity (Florian, 2007), however, in a complex way using a global reward signal. On the other hand, the observations reported here provide a rather simple, equivalent correlation-based implementation of TD and support the importance of three-factor learning for providing a link between conventional Hebbian approaches and reinforcement learning.

Additionally, our approach is intrinsically based on continuous states as well as on continuous time. Several attempts to shape RL algorithms, which usually have discrete states and time, into continuous systems exist but lack biological motivation. In particular, Baird (1993) extended Q-learning by the "advantage updating" method, and Doya (2000) performed the transformation from a discrete sum to a continuous integral for the calculation of the return R. In his case, every value function V consists of a state representation and a corresponding weight. These weights need to be adjusted in order to let the delta error converge to zero. This is done by a gradient descent algorithm, which results in an update rule that demands a weight derivative, which is difficult to emulate in a biologically realistic way.

Our results rely in a fundamental way on the third factor M, and the analysis performed in this study indicates that the third factor is necessary for the emulation of TD-learning by a differential Hebb rule. To explain the reason requires a closer look at the TD-learning rule. We find that the TD-rule requires a leakage term $-\alpha \cdot V(s)$. If this term does not exist, values would diverge. It can be shown that in differential Hebbian learning without a third factor, the autocorrelation part, the source of the leakage needed (see equations 2.11 and 2.4), is nonexistent (Kolodziejski, Porr, & Wörgötter, 2008). This shows that only through a well-timed third factor is the ratio between the cross-correlation and autocorrelation term correctly adjusted. This ratio is at the end responsible for the γ-value we will get using differential Hebbian learning to emulate TD-learning.

This study is mainly concerned with showing the formal equivalence between TD and differential Hebbian learning. Possible links to biophysical mechanisms play a minor role here. Nonetheless, one could consider neuromodulators for the role of the third factor M. The required reliability of timing, however, makes it unlikely that dopamine could take on this role, as the timing of these signals does not seems reliable enough (Redgrave & Gurney, 2006), although Pawlak and Kerr (2008) could show that LTP in the striatum emerges only in the presence of dopamine. The attributed, albeit still much discussed, role of the dopaminergic responses from neurons in the substatia nigra (pars compacta) or the ventral tegmental Area (VTA) as possibly representing the δ-error of TD learning (Schultz et al., 1992;

Montague et al., 1996) is thus neither questioned nor supported in our study. A very good alternative for the role of a well-timed third factor, however, seems to be the response characteristic of the cholinergic tonically active neurons (TAN). Their response, which is a reduction of activity, is exceedingly well timed and occurs together with conditioned stimuli (Graybiel, 1998; Morris, Arkadir, Nevet, Vaadia, & Bergman, 2004). The fact that TANs cease to fire would require an additional inversion to make this compatible with our M factor, but when considering possible disinhibitory effects, this should not pose a fundamental problem.

Thus, it is also important that we were able to show that our algorithm is stable across a wider range of possible biological signals as different temporal profiles exist, for example, for synapse and channel activation (compare AMPA versus NMDA characteristics). This is required, as it is not yet clear which signals are involved in any three-factor learning, and this might also depend on the considered cell type and brain structure.

Appendix: Derivations and Analytical Calculations

A.1 Solving the Homogeneous Part. We start by taking the homogeneous part of the differential equation 2.1 where we neglect the index i:

$$\rho'(t) = \frac{d\rho(t)}{dt} = \alpha \cdot u(t)\, u'(t)\, \rho(t)\, M(t). \tag{A.1}$$

Next we separate the variables and integrate both sides from zero to infinity:

$$\int_{\rho_0}^{\rho} \frac{d\rho}{\rho} = \alpha \cdot \int_0^{\infty} u(t)\, u'(t)\, M(t)\, dt. \tag{A.2}$$

We mentioned in the beginning that $M(t)$ is either 1 or 0; thus, we can model this function as a sum of Heaviside functions $\Theta(t)$:

$$M(t) = \sum_m \Theta\big(t - b_m^{lower}\big)\, \Theta\big(b_m^{upper} - t\big). \tag{A.3}$$

As the Heaviside functions determine the integration boundaries of the right-hand side of equation A.2, this can be simplified to

$$\begin{aligned} &\int_{t_0}^{t} u(t)\, u'(t) \sum_m \Theta\big(t - b_m^{lower}\big)\, \Theta\big(b_m^{upper} - t\big)\, dt \\ &\quad = \sum_m \int_{b_m^{lower}}^{b_m^{upper}} u(t)\, u'(t)\, dt. \end{aligned} \tag{A.4}$$

The left side of equation A.2 solves to a logarithmic function, and for its right side, we use following derivative:

$$\frac{du^2(t)}{dt} = 2\,u(t)\,u'(t). \tag{A.5}$$

All in all, this leads to

$$\ln \frac{\rho}{\rho_0} = \frac{\alpha}{2} \sum_m \left[u^2\left(b_m^{upper}\right) - u^2\left(b_m^{lower}\right)\right], \tag{A.6}$$

where we have to invert the logarithmic function,

$$\rho = \frac{\alpha}{2} \cdot \rho_0 \cdot \exp \sum_m \left[u^2\left(b_m^{upper}\right) - u^2\left(b_m^{lower}\right)\right], \tag{A.7}$$

and we get the weight ρ after integrating over a pulse pair if having weight ρ_0 before.

A.2 Solving the Difference Equation. First, we need to show to which value equation 2.15 converges, and then we can solve the resulting difference equation. The equation for which convergence needs to be calculated can be simplified to

$$x_{n+1} = (1 - \alpha)\,x_n + \alpha\,y. \tag{A.8}$$

This difference equation can be solved in a simpler way as a differential equation:

$$\frac{dx(t)}{dt} = -\alpha\,x(t) + \alpha\,y. \tag{A.9}$$

The homogeneous part solves to an exponential function with exponent $-\alpha t$, and the inhomogeneous solution can be found by the method of variation of parameters:

$$\begin{aligned}
x_{inhom}(t) &= x_{hom}(t) \int_0^\infty x_{hom}^{-1}(z)\,\alpha\,y\,dz \\
x_{inhom}(t) &= \exp(-\alpha\,t) \int_0^t \exp(\alpha\,z)\,\alpha\,y\,dz \\
x_{inhom}(t) &= \exp(-\alpha\,t)\,\alpha\,y \left[\frac{1}{\alpha}\,\exp(\alpha\,z)\right]_0^t = y\,(1 - \exp(-\alpha\,t)).
\end{aligned} \tag{A.10}$$

This gives us for the convergence,

$$\begin{aligned} x(t) &= (C - y)\, \exp(-\alpha\, t) + y \\ \lim_{t\to\infty} x(t) &= y, \end{aligned} \tag{A.11}$$

where C is a constant. This shows that difference equations like equation A.8 or differential equations like equation A.9 always converge to y. Additionally, Kushner and Clark (1978) showed that this holds also for a stochastic variable with mean y.

Using this result, we can directly set $x_\infty = y$, and we then have the following difference equation, which needs to be solved:

$$\rho_n = \varepsilon_1\, \rho_{n+1} - \varepsilon_2\, \rho_{n-1}. \tag{A.12}$$

We get the solution using the ansatz $\rho_n = \mu^n$:

$$\mu^n = \varepsilon_1\, \mu^n\, \mu - \varepsilon_2\, \mu^n\, \mu^{-1}. \tag{A.13}$$

Assuming $\mu \neq 0$, we can divide equation A.13 by μ^{n-1} and solve the following quadratic equation:

$$\begin{aligned} 0 &= \mu^2 - \frac{1}{\varepsilon_1}\, \mu - \frac{\varepsilon_2}{\varepsilon_1} \\ \mu_{1/2} &= \frac{1}{2\,\varepsilon_1} \pm \sqrt{\frac{1}{(2\,\varepsilon_1)^2} + \frac{\varepsilon_2}{\varepsilon_1}}. \end{aligned} \tag{A.14}$$

A negative solution, however, would lead to an oscillation; therefore, we consider only the positive sign, which leads to a positive solution,

$$\mu = \frac{1}{2\,\varepsilon_1} + \sqrt{\frac{1}{(2\,\varepsilon_1)^2} + \frac{\varepsilon_2}{\varepsilon_1}}, \tag{A.15}$$

that, potentiated by n, gives us the value to which ρ_n will converge. Therefore, we can set $\rho_n = \mu^{-1}\, \rho_{n+1}$, and equation A.12 is simplified.

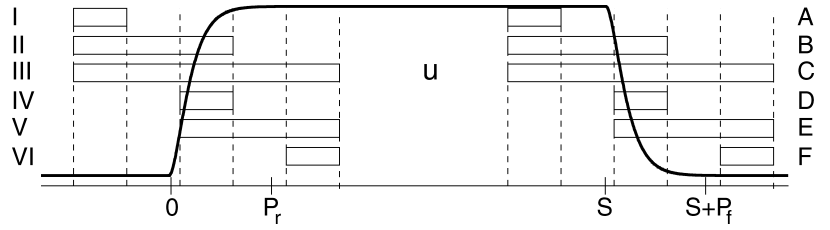

Figure 5: The filter functions with all possible regions for both the rising and the falling phases. The relevant intervals are defined by I to VI and A to F.

A.3 Analytical Calculation of γ Using First- and Second-Order Terms.

A.3.1 Taylor Expansion of the Filter Function. The Taylor expansion to the second order of an arbitrary filter function with a plateau is described as

$$u(t) = U \cdot \begin{cases} 0 & \text{if } t < 0 \\ (1-\eta)\left(\frac{t}{P^r}\right)^2 + \eta\,\frac{t}{P^r} & \text{if } t \geq 0 \cap t \leq P^r \\ 1 & \text{if } t > P^r \cap t \leq S \\ 1-(1-\xi)\left(\frac{t-S}{P^f}\right)^2 - \xi\,\frac{t-S}{P^f} & \text{if } t > S \cap t \leq S + P^f \\ 0 & \text{if } t > S + P^f, \end{cases} \tag{A.16}$$

where U is the height of the plateau, P^r and P^f the length of the rising and the falling phase, respectively, S the length of a state, and η and ξ the degree of the second-order term for the rising and the falling phase, respectively. Here, $\eta = 1$ or $\xi = 1$ leads to a linear slope, $\eta = 0$ or $\xi = 2$ to a convex, and $\eta = 2$ or $\xi = 0$ to a concave slope.

A.3.2 Intervals. Having defined the actual shape of the filter function, we now have to distinguish between different occurrence times of the third factor. Figure 5 shows the six essential regions of both the rising and the falling phase. For this, we defined intervals I to VI for the rising phase and A to F for the falling phase (see Table 1).

Additionally we need to define another four intervals used for the correlation of the signals when calculating $\tau^{\pm}$ (see Table 2).

A.3.3 Analytical Calculation of κ. Next we calculate κ using the filter function, equation A.16. For the following, we assume that $O + L < S$ and $O + S + T > P_r$ which holds if S is sufficiently larger than O, L, and $|T|$.

Table 1: Intervals Used to Discriminate Between the Different Occurrences of the Third Factor.

I	$O < 0 \cap O + L < 0$	A	$O + T < S \cap O + L + T < S$
II	$O < 0 \cap O + L \geq 0 \cap O + L < P_r$	B	$O + T < S \cap O + L + T \geq S \cap O + L + T < P_f$
III	$O < 0 \cap O + L \geq P_r$	C	$O + T < S \cap O + L + T \geq P_f$
IV	$O \geq 0 \cap O < P_r \cap O + L < P_r$	D	$O + T \geq S \cap O + T < P_f \cap O + L + T < P_f$
V	$O \geq 0 \cap O < P_r \cap O + L \geq P_r$	E	$O + T \geq S \cap O + T < P_f \cap O + L + T \geq P_f$
VI	$O \geq P_r \cap O + L \geq P_r$	F	$O + T \geq P_f \cap O + L + T \geq P_f$

Note: See Figure 5 for a more intuitive representation.

Table 2: Intervals Needed for the Correlation of Two consecutive Signals with a Time Delay of T.

a	$-T < 0$
b	$-T + P_f < 0$
c	$-T < P_r$
d	$-T + P_f < P_r$

These assumptions prevent cases where the third factor would affect the signal after the next signal, which is nonsensical. To simplify the cases, we split κ into κ^+ and κ^-, as done in the main text: $\kappa = -(\kappa^+ + \kappa^-)$. In Table 3 the analytical results for $\kappa^\pm$ are stated using equations A.17 and A.18. These equations result from equation 2.5, where we included the filter function, equation A.16. In detail, equation A.17 represents the squared rising phase, and equation A.18 the squared falling phase. It is important to mention that both functions $\Phi(t)$ and $\Psi(t)$ are bounded between 0 and 1:

$$\Phi(t) = \left(\frac{t}{P_r}\right)^2 \cdot \left(\eta + (1-\eta)\frac{t}{P_r}\right)^2 \tag{A.17}$$

$$0 \leq \Phi \leq 1 \, \forall \eta : 0 \leq \eta \leq 2 \cap \forall t : 0 \leq t \leq P_r$$

$$\Psi(t) = \left(1 - \frac{t-S}{P_f}\right)^2 \cdot \left(1 + (1-\xi)\frac{t-S}{P_f}\right)^2 \tag{A.18}$$

$$0 \leq \Psi \leq 1 \, \forall \xi : 0 \leq \xi \leq 2 \cap \forall t : S \leq t \leq S + P_f .$$

The results can be summarized by plotting the areas, which represent definitive divergent areas where we simplify the rising and falling time to identical values: $P = P_r = P_f$. These areas are composed of intervals in which the sum of κ^+ and κ^- is always positive. For instance, for interval A conjoined with either of the intervals I to VI, the sum of κ^+ and κ^- is always greater than zero and thus divergent. The same holds for interval F. Both areas are indicated in Figure 6 with different shades of gray. There is an additional divergent area III that is there only if the value of L is greater than P.

Table 3: Analytical Result of $\kappa^{\pm}$.

Interval	$\kappa^{+}/(\frac{U^2}{2})$	Interval	$\kappa^{-}/(\frac{U^2}{2})$
I	0	A	0
II	$\Phi(O+L)$	B	$\Psi(O+S+T+L)-1$
III	1	C	-1
IV	$\Phi(O+L)-\Phi(O)$	D	$\Psi(O+S+T+L)-\Psi(O+S+T)$
V	$1-\Phi(O)$	E	$-\Psi(O+S+T)$
VI	0	F	0

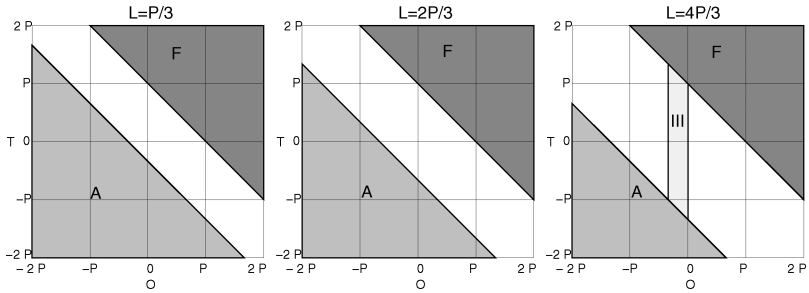

Figure 6: The divergent regions for different values of L. The intervals indicated by the gray can be found in Table 1. Dark gray represents regions independent of L, whereas the gray regions depend on L. The light gray region also depends on L; however, this area is there only for $L > P$.

There would also be an interval C (not shown) for which the sum of κ^{+} and κ^{-} is always less than zero if conjoined with interval I to VI except III. However, this shows up only if L is greater than P, but then interval III becomes valid. This can be resolved by using different values for P_r and P_f where $P_f < P_r$.

Only in the white area can there be convergence; however, the shape of the actual filter function $u(t)$ determines whether a certain area converges. Additionally, it is important to include the values for $\tau^{\pm}$ into the considerations as these values in different areas can be 0, which means that there is no overlap between two consecutive signals and the third factor. This will be investigated next.

A.3.4 Analytical Calculation of τ^{+} and τ^{-}. Here we calculate the values for $\tau^{\pm}$ using the filter function, equation A.16. We likewise assume that there is no overlap of the signal with the signal after its next state. In Table 4, the analytical results for $\tau^{\pm}$ are stated using equations A.19 to A.22. These equations result from equations 2.9 and 2.10, where we included the filter function, equation A.16. It is important to mention that these functions, $\zeta(t)$,

Table 4: Analytical Results of $\tau^{\pm}$ and γ.

Interval	$\tau^+/(\frac{U^2}{2})$	$-\tau^-/(\frac{U^2}{2})$	γ
I	0	0	0
II $\cap$ A	$\zeta(O+L)-\zeta(0)$	0	∞
II $\cap$ B $\cap$ a	$\psi(O+L)-\psi(0)$	$\varphi(O+L)-\varphi(0)$	$\gtrless 0$
II $\cap$ B $\cap$ $\bar{a}$	$\zeta(-T)-\zeta(0)+\psi(O+L)-\psi(-T)$	$\varphi(O+L)-\varphi(-T)$	$\gtrless 0$
II $\cap$ C $\cap$ a $\cap$ b	0	0	0
II $\cap$ C $\cap$ a $\cap$ $\bar{b}$	$\psi(-T+P_f)-\psi(0)$	$\varphi(-T+P_f)-\varphi(0)$	$\gtrless 0$
II $\cap$ C $\cap$ $\bar{a}$	$\zeta(-T)-\zeta(0)+\psi(-T+P_f)-\psi(-T)$	$\varphi(-T+P_f)-\varphi(-T)$	$\gtrless 0$
II $\cap$ D	$\psi(O+L)-\psi(O)$	$\varphi(O+L)-\varphi(O)$	$\gtrless 0$
II $\cap$ E $\cap$ b	0	0	0
II $\cap$ E $\cap$ $\bar{b}$	$\psi(-T+P_f)-\psi(0)$	$\varphi(-T+P_f)-\varphi(0)$	$\gtrless 0$
II $\cap$ F	0	0	0
III $\cap$ A	$\zeta(P_r)-\zeta(0)$	0	∞
III $\cap$ B $\cap$ a	$\psi(P_r)-\psi(0)$	$\varphi(P_r)-\varphi(0)+\chi(O+L)-\chi(P_r)$	∞
III $\cap$ B $\cap$ $\bar{a}$ $\cap$ c	$\zeta(-T)-\zeta(0)+\psi(P_r)-\psi(-T)$	$\varphi(P_r)-\varphi(-T)+\chi(O+L)-\chi(P_r)$	∞
III $\cap$ B $\cap$ $\bar{c}$	$\zeta(P_r)-\zeta(0)$	$\chi(O+L)-\chi(P_r)$	∞
III $\cap$ C $\cap$ b	0	0	0
III $\cap$ C $\cap$ a $\cap$ $\bar{b}$ $\cap$ d	$\psi(-T+P_f)-\psi(0)$	$\varphi(-T+P_f)-\varphi(0)$	∞
III $\cap$ C $\cap$ a $\cap$ $\bar{d}$	$\psi(P_r)-\psi(0)$	$\varphi(P_r)-\varphi(0)+\chi(-T+P_f)-\chi(P_r)$	∞
III $\cap$ C $\cap$ $\bar{a}$ $\cap$ d	$\zeta(-T)-\zeta(0)+\psi(-T+P_f)-\psi(-T)$	$\varphi(-T+P_f)-\varphi(-T)$	∞
III $\cap$ C $\cap$ $\bar{a}$ $\cap$ $\bar{d}$	$\zeta(-T)-\zeta(0)+\psi(P_r)-\psi(-T)$	$\varphi(P_r)-\varphi(-T)+\chi(-T+P_f)-\chi(P_r)$	∞
III $\cap$ C $\cap$ $\bar{c}$	$\zeta(P_r)-\zeta(0)$	$\chi(-T+P_f)-\chi(-T)$	∞
III $\cap$ D	$\psi(P_r)-\psi(0)$	$\varphi(P_r)-\varphi(0)+\chi(O+L)-\chi(P_r)$	∞
III $\cap$ E $\cap$ b	0	0	0
III $\cap$ E $\cap$ $\bar{b}$ $\cap$ d	$\psi(-T+P_f)-\psi(0)$	$\varphi(-T+P_f)-\varphi(0)$	∞
III $\cap$ E $\cap$ $\bar{d}$	$\psi(P_r)-\psi(0)$	$\varphi(P_r)-\varphi(0)+\chi(-T+P_f)-\chi(P_r)$	∞
III $\cap$ F	0	0	0

Table 4: Continued.

Interval	$\tau^+/(\frac{U^2}{2})$	$-\tau^-/(\frac{U^2}{2})$	γ
IV ∩ A	$\zeta(O+L)-\zeta(O)$	0	∞
IV ∩ B	$\zeta(-T)-\zeta(O)+\psi(O+L)-\psi(-T)$	$\varphi(O+L)-\varphi(-T)$	≥ 0
IV ∩ C	$\zeta(-T)-\zeta(O)+\psi(-T+P_f)-\psi(-T)$	$\varphi(-T+P_f)-\varphi(-T)$	≥ 0
IV ∩ D	$\psi(O+L)-\psi(O)$	$\varphi(O+L)-\varphi(O)$	≥ 0
IV ∩ E	$\psi(-T+P_f)-\psi(O)$	$\varphi(-T+P_f)-\varphi(O)$	≥ 0
IV ∩ F	0	0	0
V ∩ A	$\zeta(P_r)-\zeta(O)$	0	∞
V ∩ B ∩ c	$\zeta(-T)-\zeta(O)+\psi(P_r)-\psi(-T)$	$\varphi(P_r)-\varphi(-T)+\chi(O+L)-\chi(P_r)$	≥ 0
V ∩ B ∩ $\bar{\text{c}}$	$\zeta(P_r)-\zeta(O)$	$\chi(O+L)-\chi(-T)$	≥ 0
V ∩ C ∩ d	$\zeta(-T)-\zeta(O)+\psi(-T+P_r)-\psi(-T)$	$\varphi(-T+P_f)-\varphi(-T)$	≥ 0
V ∩ C ∩ c ∩ $\bar{\text{d}}$	$\zeta(-T)-\zeta(O)+\psi(P_r)-\psi(-T)$	$\varphi(P_r)-\varphi(-T)+\chi(-T+P_f)-\chi(P_r)$	≥ 0
V ∩ C ∩ $\bar{\text{c}}$	0	$\chi(-T+P_f)-\chi(-T)$	0
V ∩ D	$\psi(P_r)-\psi(O)$	$\varphi(P_r)-\varphi(O)+\chi(O+L)-\chi(P_r)$	≥ 0
V ∩ E ∩ d	$\psi(-T+P_f)-\psi(O)$	$\varphi(-T+P_f)-\varphi(O)$	≥ 0
V ∩ E ∩ $\bar{\text{d}}$	$\psi(P_r)-\psi(O)$	$\varphi(P_r)-\varphi(O)$	≥ 0
V ∩ F	0	0	0
VI ∩ A	0	0	0
VI ∩ B	0	$\chi(O+L)-\chi(-T)$	0
VI ∩ C	0	$\chi(-T+P_f)-\chi(-T)$	0
VI ∩ D	0	$\chi(O+L)-\chi(O)$	0
VI ∩ E	0	$\chi(-T+P_f)-\chi(O)$	0
VI ∩ F	0	0	0

$\psi(t)$, $\chi(t)$, and $\varphi(t)$ are always greater than 0.

$$\zeta(t) = \frac{2}{U^2}\int U\,u'(t)\,dt = 2\left((1-\eta)\left(\frac{t}{P^r}\right)^2 + \eta\,\frac{t}{P^r}\right) \tag{A.19}$$

$$\zeta \geq 0\,\forall\eta : 0 \leq \eta \leq 2$$

$$\begin{aligned}\psi(t) &= \frac{2}{U^2}\int u'(t)\,u(t+S+T)\,dt \\ &= 2\left((1-\eta)\left(\frac{t}{P_r}\right)^2 + \eta\frac{t}{P_r}\right) - (1-\eta)(1-\xi) \\ &\quad\times\left(\left(\frac{t}{P_f}\right)^2\left(\frac{t}{P_r}\right)^2 + \frac{8}{3}\left(\frac{t}{P_f}\right)^2\frac{t}{P_r}\frac{T}{P_r} + 2\left(\frac{t}{P_f}\right)^2\left(\frac{T}{P_r}\right)^2\right) \\ &\quad-\eta\,(1-\xi)\left(\frac{2}{3}\left(\frac{t}{P_f}\right)^2\frac{t}{P_r} + 2\left(\frac{t}{P_f}\right)^2\frac{T}{P_r} + 2\left(\frac{T}{P_f}\right)^2\frac{t}{P_r}\right) \\ &\quad-(1-\eta)\,\xi\left(\frac{4}{3}\frac{t}{P_f}\left(\frac{t}{P_r}\right)^2 + 2\frac{T}{P_f}\left(\frac{t}{P_r}\right)^2\right) \\ &\quad-\eta\,\xi\left(\frac{t}{P_f}\frac{t}{P_r} + 2\frac{t}{P_f}\frac{T}{P_r}\right)\end{aligned} \tag{A.20}$$

$$\psi \geq 0\,\forall\eta : 0 \leq \eta \leq 2 \cap \forall\xi : 0 \leq \xi \leq 2.$$

$$\begin{aligned}\chi(t) &= \frac{2}{U^2}\int U\,u'(t+S+T)\,dt \\ &= -2\left((1-\xi)\left(\left(\frac{t}{P_f}\right)^2 + 2\frac{t}{P_f}\frac{T}{P_f}\right) + \xi\,\frac{t}{P_f}\right)\end{aligned} \tag{A.21}$$

$$\chi \leq 0\,\forall\xi : 0 \leq \xi \leq 2$$

$$\begin{aligned}\varphi(t) &= \frac{2}{U^2}\int u(t)\,u'(t+S+T)\,dt \\ &= -(1-\eta)(1-\xi)\left(\left(\frac{t}{P_f}\right)^2\left(\frac{t}{P_r}\right)^2 + \frac{4}{3}\left(\frac{t}{P_f}\right)^2\frac{t}{P_r}\frac{T}{P_r}\right) \\ &\quad-\eta\,(1-\xi)\left(\frac{4}{3}\left(\frac{t}{P_f}\right)^2\frac{t}{P_r} + 2\left(\frac{t}{P_f}\right)^2\frac{T}{P_r}\right) \\ &\quad-(1-\eta)\,\xi\left(\frac{1}{3}\frac{t}{P_f}\left(\frac{t}{P_r}\right)^2\right) \\ &\quad-\eta\,\xi\left(\frac{t}{P_f}\frac{t}{P_r}\right)\end{aligned} \tag{A.22}$$

$$\varphi \leq 0\,\forall\eta : 0 \leq \eta \leq 2 \cap \forall\xi : 0 \leq \xi \leq 2.$$

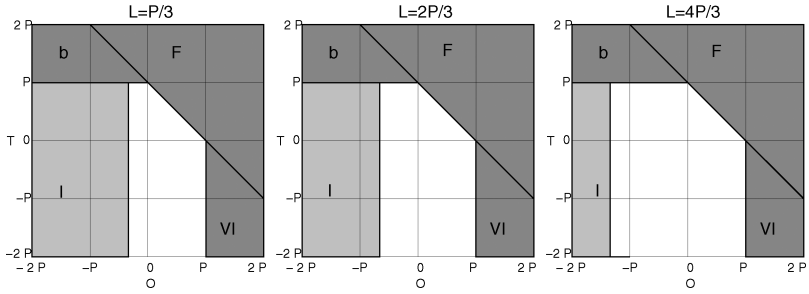

Figure 7: The regions of zero τ^+ values for different values of L. The intervals indicated by gray shading can be found in Table 1. See Figure 6 for explanations of the shades of gray.

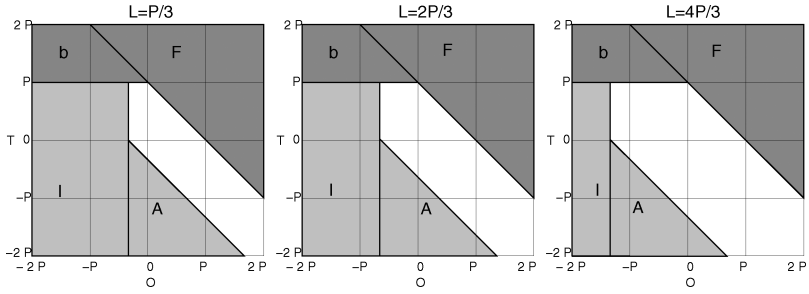

Figure 8: The regions that result in zero τ^- values for different values of L. The intervals indicated by gray shading can be found in Table 1. See Figure 6 for explanations of the shades of gray.

For these calculations, we also simplify the rising and falling time to have identical values: $P = P_r = P_f$. The results can be summarized by plotting the areas for which $\tau^\pm$ are zero (see Figures 7 and 8). For instance, similar to κ, interval F conjoined with intervals I to VI results in both τ^+ and τ^- being equal to zero. In case of τ^-, this observation can also be made for interval A, and in case of τ^+, the interval VI gives zero. There are two additional intervals, I and b, which result in a value of zero for both τ measures. These areas are indicated with different shades of gray in Figure 7 for τ^+ and in Figure 8 for τ^-.

As $\tau^\pm$ will not affect convergence, all areas would yield convergence. However, only for the white area would γ result in a value unequal to zero.

A.3.5 Analytical Calculation of γ. Finally we can calculate the value of γ using equation 2.14. This is not done explicitly here; however, we indicate in the last column of Table 4 whether γ is zero (no overlap between two consecutive signals and the third factor), greater than zero, or infinite.

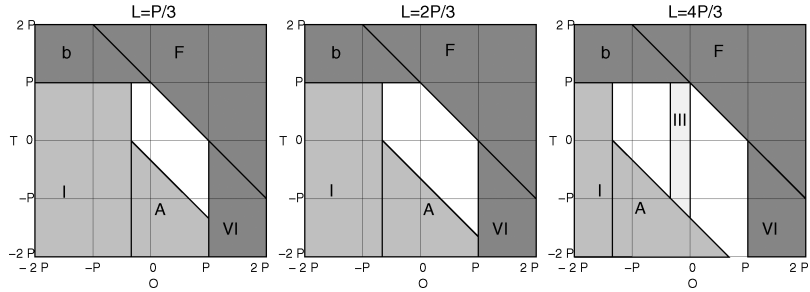

Figure 9: The regions of meaningless γ values where there is either no overlap between two consecutive signals and the third factor ($\tau^{\pm} = 0$), or γ diverges ($\kappa < 0$). The areas are shown for different values of L. The intervals indicated by gray shading can be found in Table 1. See Figure 6 for explanations of the shades of gray. It is now possible to compare this figure with Figures 2 and 3.

Furthermore we can combine the considerations made for κ and $\tau^{\pm}$, which are illustrated in Figures 6 to 8 into Figure 9. If we compare Figure 9 with Figures 2 and 3, we find that the areas of divergence map exactly, and only in the white areas convergence can be found. Therefore the considerations about convergence and nonzero $\tau^{\pm}$ values can be transfered from the filter function, equation A.16, to all possible functions that possess only one plateau.

Acknowledgments

F.W. acknowledges the support of the European Commission, IP-Project PACO-PLUS. Helpful discussions with Minija Tamosiunaite are gratefully acknowledged.

References

Baird, L. (1993). *Advantage updating* (Tech. Rep. WL-TR-93-1146). Ohio: Wright Laboratory, Wright-Patterson Air Force Base.

Baxter, J., Bartlett, P. L., & Weaver, L. (2001). Experiments with infinite-horizon, policy-gradient estimation. *Journal of Artificial Intelligence Research, 15*, 351–381.

Dayan, P. (2002). Matters temporal. *Trends in Cognitive Sciences, 6*(3), 105–106.

Dayan, P., & Sejnowski, T. (1994). TD(λ) converges with probability 1. *Mach. Learn., 14*(3), 295–301.

Doya, K. (1996). Temporal difference learning in continuous time and space. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems, 8* (pp. 1073–1079). Cambridge, MA: MIT Press.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology, 10*(6), 732–739.

Dudek, S., & Bear, M. (1992). Homosynaptic long-term depression in area CA1 of hippocampus and effects of N-methyl-D-aspartate receptor blockade. *Proceedings of the National Academy of Sciences, 89*(10), 4363–4367.

Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation, 19*, 1468–1502.

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

Graybiel, A. (1998). The basal ganglia and chunking of action repertoires. *Neurobiol. Learn. Mem., 70*(1–2), 119–136.

Hassani, O. K., Cromwell, H. C., & Schultz, W. (2001). Influence of expectation of different rewards on behavior-related neuronal activity in the striatum. *J. Neurophysiol., 85*(6), 2477–2489.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*, 2554–2558.

Izhikevich, E. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex, 17*, 2443–2452.

Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiol., 16*(2), 85–123.

Kolodziejski, C., Porr, B., & Wörgötter, F. (2008). Mathematical properties of neuronal TD-rules and differential Hebbian learning: A comparison. *Biological Cybernetics, 98*(3), 259–272.

Kosco, B. (1986). Differential Hebbian learning. In J. S. Denker (Ed.), *Neural networks for computing: AIP Conference Proc. proceedings* (Vol. 151). New York: American Institute of Physics.

Kushner, H. K., & Clark, D. S. (1978). *Stochastic approximation for constrained and unconstrained systems*. Berlin: Springer-Verlag.

Linsker, R. (1988). Self-organisation in a perceptual network. *Computer, 21*(3), 105–117.

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. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience, 76*(5), 1936–1947.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., & Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. *Neuron, 43*(1), 133–143.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience, 9*(8), 1057–1063.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biol., 15*(3), 267–273.

Pawlak, V., & Kerr, J. N. D. (2008). Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *J. Neurosci., 28*(10), 2435–2446.

Porr, B., & Wörgötter, F. (2003). Isotropic-sequence-order learning in a closed-loop behavioural system. *Phil. Trans. R. Soc. Lond. A, 361*, 2225–2244.

Porr, B., & Wörgötter, F. (2007). Learning with "relevance": Using a third factor to stabilise Hebbian learning. *Neural Comp., 19*, 2694–2719.

Potjans, W., Morrison, A., & Diesmann, M. (2009). A spiking neural network model of an actor-critic learning agent. *Neural Computation, 21*(2), 301–339.

Rao, R., & Sejnowski, T. (2001). Spike-timing-dependent Hebbian plasticity as temporal difference learning. *Neural Computation, 13*, 2221–2237.

Redgrave, P., & Gurney, K. (2006). The short-latency dopamine signal: A role in discovering novel actions? *Nature Reviews Neuroscience, 7*, 967–975.

Roberts, P. (1999). Computational consequences of temporally asymmetric learning rules: I. Differential Hebbian learning. *J. Comput. Neurosci., 7*(3), 235–246.

Roberts, P., Santiago, R., & Lafferriere, G. (2008). An implementation of reinforcement learning based on spike-timing dependent plasticity. *Biological Cybernetics, 99*(6), 517–523.

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

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci., 12*(12), 4595–4610.

Singh, S. P., Jaakkola, T., Littman, M. L., & Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning, 38*(3), 287–308.

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

Sutton, R., & Barto, A. (1981). Towards a modern theory of adaptive networks: Expectation and prediction. *Psychol. Review, 88*, 135–170.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tamosiunaite, M., Ainge, J., Kulvicius, T., Porr, B., Dudchenko, P., & Wörgötter, F. (2008). Path-finding in real and simulated rats: On the usefulness of forgetting and frustration for navigation learning. *J. Comp. Neuroscience, 25*, 562–582.

Tsitsiklis, J. N., & Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control, 42*(5), 674–690.

Watkins, C., & Dayan, P. (1992). Technical note: Q-learning. *Mach. Learn., 8*, 279–292.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In *IRE WESCON Convention Record* (pp. 96–104). New York: Institute of Radio Engineers.

Wiering, M. (2004). Convergence and divergence in standard averaging reinforcement learning. In J. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Proceedings of the 15th European Conference on Machine learning ECML'04* (pp. 477–488). Berlin: Springer-Verlag.

Received April 4, 2008; accepted August 15, 2008.