Dynamics of Temporal Difference Learning

Andreas Wendemuth
Otto-von-Guericke-University, 39016 Magdeburg, Germany
Cognitive Systems Group
andreas.wendemuth@e-technik.uni-magdeburg.de

Abstract

In behavioural sciences, the problem that a sequence of stimuli is followed by a sequence of rewards $r(t)$ is considered. The subject is to learn the full sequence of rewards from the stimuli, where the prediction is modelled by the Sutton-Barto rule. In a sequence of n trials, this prediction rule is learned iteratively by temporal difference learning. We present a closed formula of the prediction of rewards at trial time t within trial n. From that formula, we show directly that for $n \to \infty$ the predictions converge to the real rewards. In this approach, a new quality of correlation type Toeplitz matrices is proven. We give learning rates which optimally speed up the learning process.

1 Temporal Difference Learning

We consider here a mathematical treatment of a problem in behavioural biology. This problem has been described e.g. in [Dayan and Abbott 2001] as learning to predict a reward. It is Pavlovian in the sense that classical conditioning is adressed, however reward is not immediate, but after a series of stimuli there is a latency time, followed by a series of rewards. Using a simple linear rule, we will show that the subject is able to predict the remaining rewards by that rule, repeating the same stimuli-reward pattern over a series of trials. A review of learning theory related to these problems can be found in [Sutton and Barto 1998].

There are recent experimental biological correlates with this mathematical model, e.g. in the activity of primate dopamine cells during appetitive conditioning tasks, together with the psychological and pharmacological rationale for studying these cells. A review was given in [Schultz 1998], the connection to temporal difference learning can be found in [Montague *et al.* 1996].

In this paper, the focus is on two mathematical issues: a) to provide a direct *constructive* proof of convergence by giving an explicit dependence of the prediction error over trials, b) to *minimize* the learning time by giving a formula for optimal setting of the learning rate. Hence, the paper contributes as well to *temporal difference learning* as a purely mathematical issue, which may be valuable also without reference to behavioural biology and reinforcement learning. It can also be understood in a Dynamic Programming sense, see [Watkins 1989] and later work, e.g. [Gordon 2001] and [Szepesvari and Smart 2004]. We adopt the following notation in the course of this paper, following [Dayan and Abbott 2001]:

- Stimulus $u(t)$
- Future reward $r(t)$
- Sum of future rewards $R(t)$
- Weights $w(k)$
- Predicted reward $v(t)$

We want to compute, for all trials n of duration T, and for any time of trial t, the predicted reward $v^n(t)$. The (extended) stimulus u(t) is given at times $t_{\mathrm{u,min}} \dots t_{\mathrm{u,max}}$, the (extended) reward $r(t)$ is presented at times $t_{\mathrm{r,min}} \dots t_{\mathrm{r,max}}$. Stimulus and reward do not overlap, i.e. $t_{\mathrm{r,min}} > t_{\mathrm{u,max}}$.

The subject is to learn the total remaining reward at time t, $R(t) = \left\langle \sum_{\tau=0}^{T-t} r(t+\tau) \right\rangle$ (only after stimulus onset!)

The brackets $\langle\rangle$ refer, in general, to stochastic values of $R(t)$, if not exactly the same rewards are given at each trial, but when there are fluctuations.

All previous stimuli $u(t)$ are weighted to give a linear prediction model [Sutton and Barto 1990]:

$$v(t) = \sum_{\tau=0}^{t} w(\tau) u(t-\tau) \tag{1}$$

An update rule for the Sutton-Barto-formula can easily be derived from the mean square error [Dayan and Abbott 2001]:

$$\langle R(t) - v(t) \rangle^2 = \left\langle \sum_{\tau=0}^{T-t} r(t+\tau) - \sum_{\tau=0}^{t} w(\tau) u(t-\tau) \right\rangle^2 \tag{2}$$

using the partial derivative with respect to any weight $w(\alpha)$,

$$\frac{\partial \langle R(t) - v(t) \rangle^2}{\partial w(\alpha)} = -2 \left\langle \sum_{\tau=0}^{T-t} r(t+\tau) - \sum_{\tau=0}^{t} w(\tau) u(t-\tau) \right\rangle u(t-\alpha)$$

Updates are made in discrete steps in the direction of the negative gradient, providing the following rule:

$$\begin{aligned} \Delta \mathbf{w}(\tau) &= \varepsilon \delta(t) \mathbf{u}(t-\tau) \\ \delta(t) &= \left\langle \sum_{\tau=0}^{T-t} r(t+\tau) \right\rangle - v(t) = \langle R(t) \rangle - v(t) \end{aligned} \tag{3}$$

Obviously, the total future reward $R(t)$ is not known to the subject at time t. We can use $R(t) = r(t) + R(t+1)$. In this formulation, again $R(t+1)$ is not known. The subject can approximate the value by his prediction $v(t+1)$. This will provide the so-called

Temporal difference rule at time t:

$$\begin{aligned} \Delta \mathbf{w}(\tau) &= \varepsilon \delta(t) \mathbf{u}(t-\tau) \ , \quad \forall \tau = 0...t \\ \delta(t) &= r(t) + v(t+1) - v(t) \end{aligned} \tag{4}$$

The updates can be made sequentially, after each time step t, or in parallel, sampling the weight updates for all times in a given trial, and then updating at the end of the trial. The two alternatives do not substantially differ in the final result after many trials. For computational reasons, we use the parallel version of the update rule. Other learning schedules may be adopted, however we use eq. (4) since this is widely accepted in the community, following [Dayan and Abbott 2001]. Let us denote the trial number by superscripts n.

2 Dynamics of temporal difference learning

The parallel temporal difference rule eq. (4) was obtained, using an approximation to gradient descent. Therefore, convergence is not guaranteed and shall be proven in the course of this paper, where a compact notation will be introduced. We proceed as follows:

- Incorporate the rule into the prediction formula, yielding a recursive formula for all predictions $v^n(t)$ at trial n.
- Make this a closed formula $v^n(t)$.
- Show convergence $v^n(t) \to R(t)$ for large N.
- Choose an optimal learning rate ε for maximally fast convergence.

We obtain for the predictions $v^{n+1}(t)$ at trial $n+1$ the following recursive formula. Summation limits explicitely state situations where stimuli and reward phases may overlap, we will restrict to non-overlapping cases later.

$$\begin{aligned} &\frac{v^{n+1}(t) - v^n(t)}{\varepsilon} \underline{\underline{\ t \geq t_{u,\min}\ }} \sum_{k=t_{u,\min}}^{\min(t,t_{u,\max})} \frac{\Delta w^n(t-k) u(k)}{\varepsilon} \\ &= \sum_{k=t_{u,\min}}^{\min(t,t_{u,\max})} u(k) \sum_{x=t_{u,\min}}^{\min(t_{r,\max}-t+k,t_{u,\max})} \delta^n(x+t-k) u(x) \\ &= \sum_{y=t_{u,\min}-\min(t,t_{u,\max})}^{\min(t_{r,\max}-t,t_{u,\max}-t_{u,\min})} \delta^n(t+y)\, g(t,y) \end{aligned} \tag{5}$$

where $\delta^n(t) = r(t) + v^n(t+1) - v^n(t)$ and

$$g(t,y) = \sum_{k=\max(t_{u,\min},t_{u,\min}-y)}^{\min(\min(t,t_{u,\max}),t_{u,\max}-y)} u(k) u(y+k) \tag{6}$$

Since we are interested whether the subject will make proper predictions after all stimuli have been given, we will restrict our analysis to times $t > t_{u,\max}$. Then $g(t,y)$ will become independent of t, giving

$$g(y) = g(t,y) = \sum_{k=\max(t_{u,\min},t_{u,\min}-y)}^{\min(t_{u,\max},t_{u,\max}-y)} u(k) u(y+k) \tag{7}$$

This can be written in matrix notation as follows, with $\mathbf{E}$ the unit matrix:

$$\begin{pmatrix} v^{n+1}(t_{u,max}) \\ v^{n+1}(t_{u,max}+1) \\ \vdots \\ \\ v^{n+1}(t_{r,\max}) \end{pmatrix} = (\mathbf{E} - \varepsilon \mathbf{G}\ \mathbf{A}) \times$$

$$\times \begin{pmatrix} v^{n}(t_{u,max}) \\ v^{n}(t_{u,max}+1) \\ \vdots \\ \\ v^{n}(t_{r,\max}) \end{pmatrix} + \varepsilon \mathbf{G} \begin{pmatrix} 0 \\ \vdots \\ r(t_{r,\min}) \\ \vdots \\ r(t_{r,\max}) \end{pmatrix}$$

where we shall now use the total time $T = t_{r,\max} - t_{u,\max} + 1$:

$$g(y) = \sum_{k=t_{u,\min}}^{t_{u,\max}-y} u(k)\ u(y+k) \ , \quad \forall\, y = 0...k \tag{8}$$

with any real values $u(k)$ and $u(t_{u,\max}) \neq 0$, and $T \times T$-matrices

$$\mathbf{G} = \begin{pmatrix} g_0 & g_1 & \cdots & g_k & \mathbf{0} \\ g_1 & g_0 & g_1 & \cdots & \ddots\, g_k \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ g_k & \cdots & g_1 & g_0 & g_1 \\ \mathbf{0} & \ddots\, g_k & \cdots & g_1 & g_0 \end{pmatrix} \text{ and}$$

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{pmatrix} \tag{9}$$

where the "$\mathbf{0}$" in $\mathbf{G}$ refers to the full remaining upper and lower triangle, respectively, and in $\mathbf{A}$ all entries except for the two diagonals are 0. We can further write this in compact notation (vectors have component index t):

$$\mathbf{v}^{n+1} = (\mathbf{E} - \varepsilon \mathbf{G}\ \mathbf{A})\ \mathbf{v}^n + \varepsilon \mathbf{G}\ \mathbf{r}$$

With $\mathbf{v}^0 = \mathbf{0}$ one has

$$\begin{aligned} \mathbf{v}^{N+1} &= \varepsilon \sum_{n=0}^{N} (\mathbf{E} - \varepsilon \mathbf{G}\ \mathbf{A})^n\ \mathbf{G}\ \mathbf{r} \\ &= \varepsilon (\varepsilon \mathbf{G}\mathbf{A})^{-1} (\mathbf{E} - (\mathbf{E} - \varepsilon \mathbf{G}\mathbf{A})^{(N+1)}) \mathbf{G}\mathbf{r} \\ &= \mathbf{A}^{-1} \Big(\mathbf{E}\ - \mathbf{G}^{-1}\ (\mathbf{E} - \varepsilon \mathbf{G}\ \mathbf{A})^{(N+1)}\ \mathbf{G} \Big) \mathbf{r} \end{aligned} \tag{10}$$

Only if $(-\mathbf{G}\ \mathbf{A})$ is a Hurwitz (stability) matrix (all eigenvalues have negative real part), a suitable ε can be chosen such that $(\mathbf{E} - \varepsilon \mathbf{G}\ \mathbf{A})^{(N+1)}$ will vanish with large N. A similar approach, relying on Hurwitz matrices, was followed in [Sutton 1988], however there the Hurwitz property was not shown immediately on the matrix structure of eq. (9). Our aim here is to provide an explicit proof, which will establish

as an interesting result in its own right a general property of Toeplitz matrices with correlation entries.

We will show the Hurwitz property later, and we will give values for ε for which convergence is assured and for which it is optimal. Continuing with the large N-behaviour, we obtain $\mathbf{v}^{N+1} \stackrel{N\to\infty}{\longrightarrow} \mathbf{A}^{-1}\,\mathbf{r}$. This holds no matter what the stimulus! With

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 & 1 \\ & & \ddots & \ddots & \vdots \\ & & & 1 & 1 \\ & & & & 1 \end{pmatrix}$$

we obtain

$$v^{\infty}(t) = \sum_{\tau=0}^{T-t} r(t+\tau) = R(t)$$

which is the desired result. This proves convergence in full generality. We need to show that $(-\mathbf{G}\ \mathbf{A})$ is a Hurwitz matrix, i.e. that all eigenvalues of that matrix have negative real part. Further, we want to give values for ε which provide maximum speed of convergence.

3 Proof of convergence

In previous approaches in the literature, temporal difference (TD) learning was extended to TD(λ) learning where the parameter λ refers to an exponential weighting with recency. Learning and convergence issues were considered e.g. in [Dayan 1992] and [Dayan and Sejnowski 1994]. In the TD(λ) framework, what we consider here is TD(0)-learning where in contrast to the mentioned literature we concentrate on a direct proof of convergence by establishing a general property of Toeplitz matrices with correlation entries.

The proof proceeds as follows: we will show first that $\mathbf{G}$ is positive definite. Then, we will show that all eigenvalues of $(\mathbf{G}\ \mathbf{A})$ have positive real part. Hence, we will have established the Hurwitz property required for convergence.

In order to see that $\mathbf{G}$ is positive definite, we consider any nonzero real vector $\mathbf{x}$ and show that $\mathbf{x^T G x}$ is always positive. We first extract from $\mathbf{G}$ the diagonal matrix $\mathbf{G_D}$ and the upper triangular matrix $\mathbf{G}_+$. Then we have

$$\begin{aligned} \mathbf{x^T G x} &= \mathbf{x^T G_D x} + \mathbf{x^T G_+ x} + \mathbf{x^T G_+^T x} \\ &= \mathbf{x^T G_D x} + \mathbf{x^T G_+ x} + (\mathbf{x^T G_+^T x})^T \\ &= \mathbf{x^T G_D x} + 2\,\mathbf{x^T G_+ x} \\ &= \sum_{i=0}^{N} g(0)\,x_i^2 + 2\sum_{i=0}^{N} x_i \sum_{j=1}^{k} g(j)\,x_{i+j} \end{aligned} \quad (11)$$

where in order to keep summation limits easy it is understood that $x_i = 0$ for $i < 0$ and $i > N$ ("zero padding"). There are $2k+1$ Toeplitz bands, k being the number of stimuli. Shifting the time index in eq. (8) from $t_{u,\min}$ to 0, $g(j)$ arises from the k stimuli for $u(k) \neq 0,\ u(t)$ real as

$$g(j) = \sum_{t=0}^{k-|j|} u(t)\,u(t+|j|)\ , \qquad |j| = 0 \ldots k\ , \forall\, t \quad (12)$$

hence $\mathbf{x^T G x}$ can be written as

$$-\sum_{i=0}^{N}\sum_{t=0}^{k} u(t)^2\,x_i^2 + 2\sum_{i=0}^{N}\sum_{j=0}^{k}\sum_{t=0}^{k-j} u(t)\,u(t+j)\,x_i\,x_{i+j} \quad (13)$$

Rearranging the inner sums with

$$\sum_{j=0}^{k}\sum_{t=0}^{k-j} = \sum_{t=0}^{k}\sum_{j=0}^{k-t} \quad (14)$$

and using vectors $\mathbf{u}$ of dimension $(k+1)$ with components $u_t = u(t)$ leads to

$$\mathbf{x^T G x} = -\sum_{i=0}^{N}\sum_{t=0}^{k} u(t)^2\,x_i^2 + 2\sum_{i=0}^{N} \mathbf{u^T}\tilde{\mathbf{X}}^{(i)}\mathbf{u} \quad (15)$$

where the matrices $\tilde{\mathbf{X}}^{(i)}$ are $(k+1)\times(k+1)$ band matrices with entries $\tilde{\mathbf{X}}^{(i)}_{m,m+n} = x_i x_{i+n}$ for all m and $n \geq 0$, and 0 otherwise. Symmetrizing the matrix products, we can write

$$2\sum_{i=0}^{N} \mathbf{u^T}\tilde{\mathbf{X}}^{(i)}\mathbf{u} = \mathbf{u^T}\left[\sum_{i=0}^{N}\tilde{\mathbf{X}}^{(i)}\right]\mathbf{u} + \mathbf{u^T}\left[\sum_{i=0}^{N}(\tilde{\mathbf{X}}^{(i)})^T\right]\mathbf{u} \quad (16)$$

The matrix elements of the second term are, for $-k \leq n \leq 0$:

$$\left[\sum_{i=0}^{N}(\tilde{\mathbf{X}}^{(i)})^T\right]_{m,m+n} = \sum_{j=0}^{N} x_j x_{j-n} = \sum_{i=0}^{N} x_i x_{i+n} \quad (17)$$

where the summation limits in the last result were taken, for convenience, larger than necessary for covering all cases of n for which the summand is nonzero (note that n is nonpositive), the extra terms being zero due to the zero padding convention for the x_i. As a result, the matrix elements of the first and the second term in eq. (16) have identical format, for nonnegative and nonpositive n, respectively.

Making use of eq. (16) and eq. (17), and incorporating the first sum in eq. (15) leads to

$$\mathbf{x^T G x} = \mathbf{u^T}\left[\sum_{i=0}^{N}\hat{\mathbf{X}}^{(i)}\right]\mathbf{u} \quad (18)$$

where the matrices $\hat{\mathbf{X}}^{(i)}$ are band matrices with entries $\hat{\mathbf{X}}^{(i)}_{m,m+n} = x_i x_{i+n}$ for positive and negative n.

We can now write the sum of matrices in eq. (18) in a different way. Denote vectors $\tilde{\mathbf{x}}^{(\mathbf{i})}$ of dimension $(k+1)$ with components $\tilde{x}^{(i)}_m = x_{i+m}$, $m = 0 \ldots k$, where, as before, it is understood that $x_j = 0$ for $j < 0$ and $j > N$. Hence, these vectors will have nonzero components only for $i = -k \ldots N$.

Notice that the following covariance matrices are generated by these vectors:

$$\left[\tilde{\mathbf{x}}^{(\mathbf{i})}(\tilde{\mathbf{x}}^{(\mathbf{i})})^T\right]_{(m,n)} = x_{i+m}\,x_{i+n} \quad (19)$$

Then we have for $(m = 0 \ldots k, n = -m \ldots k-m)$:

$$\left[\sum_{i=-k}^{N}\tilde{\mathbf{x}}^{(\mathbf{i})}(\tilde{\mathbf{x}}^{(\mathbf{i})})^T\right]_{(m,n)} = \sum_{i=-k+m}^{N+m} x_i\,x_{i+n-m} \quad (20)$$

and

$$\left[\sum_{i=-k}^{N} \tilde{\mathbf{x}}^{(\mathbf{i})}(\tilde{\mathbf{x}}^{(\mathbf{i})})^T\right]_{(m,m+n)} = \sum_{i=-k+m}^{N+m} x_i\, x_{i+n} \qquad (21)$$

which establishes that

$$\sum_{i=-k}^{N} \tilde{\mathbf{x}}^{(\mathbf{i})}(\tilde{\mathbf{x}}^{(\mathbf{i})})^T = \sum_{i=0}^{N} \hat{\mathbf{X}}^{(i)} \qquad (22)$$

Then we can rewrite eq. (18), using eq. (22):

$$\mathbf{x^T G x} = \sum_{i=-k}^{N} \mathbf{u^T}\tilde{\mathbf{x}}^{(\mathbf{i})}(\tilde{\mathbf{x}}^{(\mathbf{i})})^T\mathbf{u} = \sum_{i=-k}^{N} |\mathbf{u^T}\tilde{\mathbf{x}}^{(\mathbf{i})}|^2 \qquad (23)$$

This sum is certainly nonnegative. We will now show that it cannot become zero.

The proof is by contradiction. Suppose the sum is zero. Then all individual terms must be zero. Start with the first term. We have from the definition

$$\mathbf{u^T}\tilde{\mathbf{x}}^{(-\mathbf{k})} = \sum_{j=0}^{k} u_j x_{-k+j} = u_k x_0 \qquad (24)$$

This becomes zero, since $u_k \neq 0$ (eq. (12)), only for $x_0 = 0$. Now proceed with the second term:

$$\mathbf{u^T}\tilde{\mathbf{x}}^{(-\mathbf{k+1})} = \sum_{j=0}^{k} u_j x_{-k+1+j} = u_{k-1}x_0 + u_k x_1 = u_k x_1 \qquad (25)$$

where the previous result $x_0 = 0$ was used. This becomes zero, since $u_k \neq 0$ (eq. (12)), only for $x_1 = 0$. Continuing in this vein, after N steps, we end with the product

$$\mathbf{u^T}\tilde{\mathbf{x}}^{(-\mathbf{k+N})} = \sum_{j=0}^{k} u_j x_{-k+N+j}$$
$$= u_0 x_{N-k} + \ldots + u_{k-1}x_{N-1} + u_k x_N = u_k x_N$$

where all the previous results were used. Hence $x_N = 0$, which in total means that $\mathbf{x} = \mathbf{0}$. This is a contradiction, since $\mathbf{x}$ must be a nonzero vector.

With this result, we have established that $\mathbf{x^T G x} > 0$, hence we have established:

Symmetric Toeplitz matrices of the structure of $\mathbf{G}$ *after eq. (9), with correlation entries* g_j *after eq. (12), are positive definite.*

We now turn our attention to showing that all eigenvalues of $(\mathbf{G}\ \mathbf{A})$ have positive real part. We will look at any real $\mathbf{A}$ and any $\mathbf{G}$ which is real, symmetric and positive definite. (We will not require the stronger condition that the structure of $\mathbf{G}$ is after eq. (9), and we say nothing about the entries g_j.)

Under these premises, the proof will first give a sufficient condition for the sign of the real part of the eigenvalues of $(\mathbf{G}\ \mathbf{A})$. Then we will show that this sign is indeed positive for the $\mathbf{A}$ at hand.

Let $\mathbf{y}$ be any eigenvector of $(\mathbf{G}\ \mathbf{A})$ and λ be the corresponding eigenvalue. Since $(\mathbf{G}\ \mathbf{A})$ is real, $\mathbf{y}$ and λ are either both real or both complex. In the complex case, there exists a further eigenvector $\mathbf{y}^*$ and corresponding eigenvalue λ^*, where $()^*$ denotes complex conjugated and transposed.

Let us denote $\mathbf{z} = \mathbf{A}\,\mathbf{y}$ and write

$$\mathbf{z}^*\mathbf{G}\mathbf{z} = \mathbf{y}^*\mathbf{A}^*\mathbf{G}\mathbf{A}\mathbf{y} = \lambda\mathbf{y}^*\mathbf{A}^*\mathbf{y} \qquad (26)$$

Also, with $\mathbf{G}$ real and symmetric, we have $\mathbf{G} = \mathbf{G}^*$ and

$$\mathbf{z}^*\mathbf{G}\mathbf{z} = (\mathbf{z}^*\mathbf{G}\mathbf{z})^* = \lambda^*\mathbf{y}^*\mathbf{A}\mathbf{y} \qquad (27)$$

The summation of eqns. 26 and 27 gives

$$2\mathbf{z}^*\mathbf{G}\mathbf{z} = \mathbf{y}^*\left[\lambda^*\mathbf{A} + \lambda\mathbf{A}^*\right]\mathbf{y} \qquad (28)$$

With $\mathbf{G}$ real, symmetric and positive definite, we have for any complex $\mathbf{z}$ that $\mathbf{z}^*\mathbf{G}\mathbf{z}$ is real and positive. This can be used as follows.

Case 1: $\mathbf{y}$ and λ are both real.
Then we have, with real $\mathbf{A}$, from eq. (28) immediately

$$0 < \mathrm{Re}(2\mathbf{z}^*\mathbf{G}\mathbf{z}) = \lambda\mathbf{y^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{y} \qquad (29)$$

hence

$$\mathrm{sign}(\mathrm{Re}(\lambda)) = \mathrm{sign}(\lambda) = \mathrm{sign}(\mathbf{y^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{y}) \qquad (30)$$

Case 2: $\mathbf{y}$ and λ are both complex.
Let us write $\mathbf{y} = \mathbf{v} + i\mathbf{w}$ and $\lambda = g + ih$. Then from eq. (28), with real $\mathbf{A}$, we obtain

$$\begin{aligned} 0 \;<\; & \mathrm{Re}(2\mathbf{z}^*\mathbf{G}\mathbf{z}) = g\,\mathbf{v^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{v} \qquad (31)\\ & + g\,\mathbf{w^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{w} - 2\,h\,\mathbf{w^T}\left[\mathbf{A^T} - \mathbf{A}\right]\mathbf{v} \\ 0 \;=\; & \mathrm{Im}(2\mathbf{z}^*\mathbf{G}\mathbf{z}) = h\,\mathbf{v^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{v} \qquad (32)\\ & + h\,\mathbf{w^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{w} + 2\,g\,\mathbf{w^T}\left[\mathbf{A^T} - \mathbf{A}\right]\mathbf{v} \end{aligned}$$

Combining eqns. 31 and 32 gives

$$0 < g(g^2 + h^2)\left\{\mathbf{v^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{v} + \mathbf{w^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{w}\right\} \qquad (33)$$

from which follows with $\mathrm{sign}(\mathrm{Re}(\lambda)) = \mathrm{sign}(g)$:

$$\mathrm{sign}(\mathrm{Re}(\lambda)) = \mathrm{sign}\left\{\mathbf{v^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{v} + \mathbf{w^T}\left[\mathbf{A^T} + \mathbf{A}\right]\mathbf{w}\right\} \qquad (34)$$

The results for both cases, eqns. 30 and 34, allow the following sufficient condition:
Let $\mathbf{G}$ *be a real, symmetric and positive definite matrix, and* $\mathbf{A}$ *be a real square matrix. If the matrix* $\left[\mathbf{A^T} + \mathbf{A}\right]$ *is positive definite, then the real parts of the eigenvalues of* $(\mathbf{G}\ \mathbf{A})$ *are positive.*

Let us now turn our attention to the specific matrix $\mathbf{A}$ given in eq. (9). We have

$$\mathbf{A^T} + \mathbf{A} = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & \end{pmatrix} \qquad (35)$$

It can be shown that the matrix $\mathbf{Q} = \mathbf{A^T} + \mathbf{A}$ is positive definite, by using the previous result of this paper: matrices of the structure of $\mathbf{G}$ are positive definite. Noting that $\mathbf{Q}$ has the structure of $\mathbf{G}$, using eq. (12) with $k = 1$ and $\mathbf{u}$ = (1,-1), immediately gives the desired result. This completes our proof that $(-\mathbf{G}\ \mathbf{A})$ is a Hurwitz matrix.

4 Learning rate ε for fast convergence

The convergence behaviour of the temporal difference rule is known from eq. (10). For large number of trials, convergence speed will be dominated by the eigenvalue of $(\mathbf{E} - \varepsilon \mathbf{G}\ \mathbf{A})$ with the largest modulus, for which convergence speed is slowest. We have for the eigenvalues (ev):

$$\mathrm{ev}(\mathbf{E} - \varepsilon \mathbf{G}\ \mathbf{A}) = 1 - \varepsilon\ \mathrm{ev}(\mathbf{G}\ \mathbf{A}) = 1 - \varepsilon\ (g + \ i\, h) \quad (36)$$

where the same notation for the eigenvalues of $\mathbf{G}\ \mathbf{A}$ as in the previous section, $\lambda = g + \ i\, h$, was used.

Hence, the first task to do is to compute the eigenvalues of $\mathbf{G}\ \mathbf{A}$ which are only dependent on the given stimuli $u(k)$ and on the total time T. Note that $\mathbf{G}\ \mathbf{A}$ defines the special structure of an unsymmetric Toeplitz matrix with $2k + 2$ bands. It is well known that a closed solution for the eigenvalues of Toeplitz matrices exists only for not more than three bands, they are given for symmetric and asymmetric cases e.g. in [Beam and Warming 1993]. Hence only for $k = 0$ can a closed solution been given, which we do in sec. 5. For $k > 0$ numerical values must be obtained, special methods for the solution, and for the asymptotic structure of the eigenvalues for large T, are given in [Beam and Warming 1993].

The square modulus m of the eigenvalues is given by

$$m(\varepsilon) = |\ \mathrm{ev}(\mathbf{E} - \varepsilon \mathbf{G}\ \mathbf{A})\ |^2 = (1 - \varepsilon\ g)^2 + \varepsilon^2\, h^2 \quad (37)$$

We will first give a choice for ε for which convergence is always assured and which also serves as a good rough value for setting ε without much computational effort.

Note first that $m(\varepsilon = 0) = 1$ for all eigenvalues. Further, since $(-\mathbf{G}\ \mathbf{A})$ is Hurwitz, $g > 0$ for all eigenvalues, and for small $\varepsilon > 0$, $m(\varepsilon) < 1$ for all eigenvalues due to eq. (37). Lastly, for large ε, $m(\varepsilon)$ will increase again beyond all bounds for all eigenvalues, since $m(\varepsilon) \propto \varepsilon^2$ for large ε and all eigenvalues.

After this sketch of the general behaviour of $m(\varepsilon)$, it is clear that for each eigenvalue k there is a $\varepsilon(k) \neq 0$ for which $m(\varepsilon(k)) = 1$ and $m(\varepsilon(k))$ is rising with $\varepsilon(k)$. Computing these values from eq. (37) for eigenvalues $\lambda_k = g_k + ih_k$ gives

$$\varepsilon(k) = \frac{2\, g_k}{g_k^2 + h_k^2} \quad (38)$$

Hence, $m(\varepsilon) < 1$ for $0 < \varepsilon < \min_{\mathrm{k}}\ \varepsilon(k)$, and a good choice $\tilde{\varepsilon}$ for convergent behaviour is therefore the mean value of the bounds,

$$\tilde{\varepsilon} = \min_k\ \frac{g_k}{g_k^2 + h_k^2} \quad (39)$$

Note again that whenever $h_k \neq 0$, there is an additional eigenvalue $g_k - \ i\, h_k$ for which the minimum in eq. (39) is identical and needs not be computed.

We now turn to finding the optimal value ε^* for fastest convergence. This is given by

$$\varepsilon^* = \mathrm{argmin}_\varepsilon\ \max_k\ m(\varepsilon(k)) \quad (40)$$

This optimization problem has no closed solution. However, a simple line search algorithm can be given to find the solution in finitely many (less than T) steps.

The general idea is to start at $\varepsilon = 0$ and increase ε until we reach ε^*. In all of this process, we select a k and go along curves $m(\varepsilon(k))$, ensuring the condition that our chosen k satisfies $\max_k\ m(\varepsilon(k))$.

At $\varepsilon = 0$, all $m(\varepsilon(k)) = 1$ and decreasing with ε. The slope is given by $-2g_k < 0$. We start by choosing the k^* for which $k^* = \mathrm{argmin}_k\ g_k$, which ensures that the condition $\max_k\ m(\varepsilon(k))$ is satisfied by our choice of k^*. Our current position is $\varepsilon_c = 0$. Then we apply the following procedure:

[START]

Continuing on curve k^*, we would reach the minimum for

$$\varepsilon_{k^*}^* = \frac{g_{k^*}}{g_{k^*}^2 + h_{k^*}^2} \quad (41)$$

if everywhere between our current position ε_c and the proposed position $\varepsilon_{k^*}^*$ the maximum condition $\max_k\ m(\varepsilon(k))$ is satisfied for k^*. In order to check this, we compute the intersections of our curve under inspection k^* to all other curves k.

After eq. (37), these intersections are given at values ε_k according to

$$(1 - \varepsilon_k\ g_k)^2 + \varepsilon_k^2\, h_k^2 = (1 - \varepsilon_k\ g_{k^*})^2 + \varepsilon_k^2\, h_{k^*}^2 \quad (42)$$

or

$$\varepsilon_k = \ 2\ \frac{g_k - g_{k^*}}{g_k^2 - g_{k^*}^2 + h_k^2 - h_{k^*}^2} \quad (43)$$

Now we inspect the intersection which is closest to our current position of ε_c, i.e. for which $\hat{k} = \mathrm{argmin}_k\{\ \varepsilon_k,\ \varepsilon_k > \varepsilon_c\}$.

If $\varepsilon_{\hat{k}} > \varepsilon_{k^*}^*$, then $\varepsilon_{k^*}^*$ after eq. (41) is indeed the solution (*free minimum*). [STOP]

Else, there is an intersection prior to reaching $\varepsilon_{k^*}^*$. If the curve that is intersecting is *rising* at $\varepsilon(\hat{k})$, this means that we are done, since we actually have reached the condition of eq. (40). The solution is

$$\varepsilon(\hat{k}) = \min_k \left\{ \varepsilon(k) = \ 2\ \frac{g_k - g_{k^*}}{g_k^2 - g_{k^*}^2 + h_k^2 - h_{k^*}^2}\ ,\ \varepsilon(k) > \varepsilon_c \right\} \quad (44)$$

(*bounded minimum*). [STOP]

Else, if the curve that is intersecting is *falling* at $\varepsilon(\hat{k})$, we must continue on that new curve which now satisfies the maximum condition in eq. (40). To this end, we set $\varepsilon_c = \varepsilon(\hat{k})$ and $k^* = \hat{k}$ and continue (*change in falling curves with maximum modulus*). [GO TO START]

Fig. 1 illustrates the behaviour with free minimum, fig. 2 with bounded minimum. Both figures have a change in falling curves with maximum modulus.

This algorithm terminates after finitely many steps, either at the minimum of the curve k^* or at an intersection of a falling with a rising curve. It is clear that one of the two possibilities exist since all curves eventually rise with large ε.

5 One stimulus

We look at the special case where only one stimulus u is present, hence $k = 0$. Then, eq. (10) takes a particularly simple form, where $(\mathbf{E} - \mathbf{A})$ is the unit shift matrix:

$$\mathbf{A}\mathbf{v}^{N+1} = \mathbf{r} - \sum_{j=0}^{T} \binom{N}{j} (1 - \varepsilon\, u^2)^{N-j} \left[\varepsilon\, u^2\, (\mathbf{E} - \mathbf{A})\right]^j \mathbf{r} \quad (45)$$

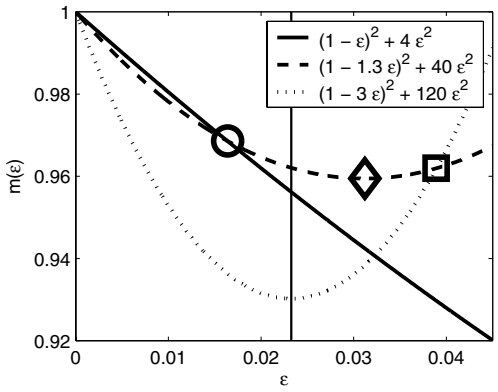

Figure 1: *Free minimum*: starting at $\varepsilon = 0$, the solid curve has maximum modulus (k^*). At the circle, another falling curve (dashed) is intersecting and takes the maximum role of k^*. The global minimum of that curve (diamond) is as well the total minimum after eq. (40), with $m = 0.959$. The next intersection with the rising (dotted) curve (square) is of no importance since it occurs at a higher ε. The vertical line indicates the approximate value $\tilde{\varepsilon}$ after eq. (39) with non-optimal $m(\tilde{\varepsilon}) = 0.962$.

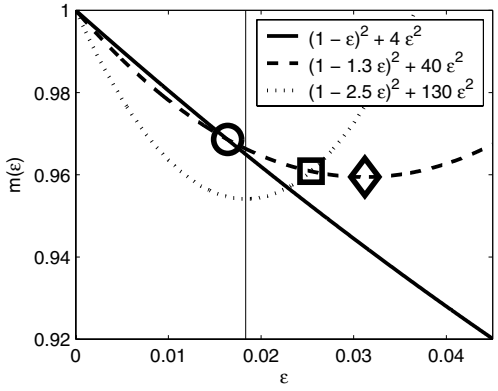

Figure 2: *Bounded minimum*: As in fig. 1, starting at $\varepsilon = 0$, first the solid curve and then the dashed curve have maximum modulus (k^*). The next intersection (square) occurs with the rising (dotted) curve, hence this intersection is the total minimum after eq. (40), with $m = 0.961$. The global minimum of the dashed curve (diamond) is of no importance since it occurs at a higher ε. The vertical line indicates the approximate value $\tilde{\varepsilon}$ after eq. (39) with non-optimal $m(\tilde{\varepsilon}) = 0.966$.

After eq. (38), any $0 < \varepsilon < 2u^{-2}$ will ensure convergence with exponential error decay over time, except for the case when choosing the optimal learning rate $\varepsilon\, u^2 = 1$. In this case we obtain for $N \leq T$:

$$\mathbf{v}^{N+1} = \mathbf{A}^{-1}\big(\mathbf{E} - (\mathbf{E} - \mathbf{A})^N\big)\mathbf{r} \tag{46}$$

and for $N > T$ we have $\mathbf{v}^{N+1} = \mathbf{A}^{-1}\mathbf{r}$. Hence, after finitely many (T) steps, convergence is reached when choosing the optimal learning rate. This is due to the fact that $(\mathbf{G}\ \mathbf{A})$ has only one degenerate eigenvalue, and T generalized eigenvectors of algebraic multiplicity $1 \ldots T$.

6 Conclusion

We have presented a closed formula of the prediction of rewards at trial time t within trial n and shown its convergence, where two general formulae for Toeplitz matrices and eigenvalues of products of matrices were derived and utilized. We have given learning rates which optimally speed up the learning process.

References

[Beam and Warming 1993] Beam, R and Warming, R (1993). The Asymptotic Spectra of Banded Toeplitz and Quasi-Toeplitz Matrices. SIAM Jour. Scientific Computing, 14 (4), 971–1006.

[Dayan and Abbott 2001] Dayan, P and Abbott, LF (2001). Theoretical Neuroscience. Cambridge, MA: MIT Press.

[Dayan and Sejnowski 1994] Dayan, P and Sejnowski, TJ (1994). TD(λ) converges with probability 1. Machine Learning, 14, 295-301.

[Dayan 1992] Dayan, P (1992). The convergence of TD(λ) for general λ. Machine Learning, 8, 341-362.

[Gordon 2001] Gordon, GJ (2001) Reinforcement learning with function approximation converges to a region. Neural Information Processing Systems, pages 1040-1046 (NIPS'01).

[Montague *et al.* 1996] Montague, PR, Dayan, P and Sejnowski, TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience, 16, 1936-1947.

[Schultz 1998] Schultz, W (1998) Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80, 1–27.

[Sutton and Barto 1998] Sutton, RS and Barto, AG (1998) Reinforcement Learning. Cambridge, MA: MIT Press.

[Sutton and Barto 1990] Sutton, RS and Barto, AG: Time-derivative models of Pavlovian reinforcement. In M. Gabriel and J. Moore, editors: Learning and Computational Neuroscience: Foundations of Adaptive Networks, pages 497–537. MIT Press, 1990.

[Sutton 1988] Sutton, RS (1988). Learning to predict by the method of temporal differences. Machine Learning, 3, 9–44.

[Szepesvari and Smart 2004] Szepesvari, C and Smart, WD (2004) Convergent value function approximation methods. International Conference on Machine Learning (ICML04).

[Watkins 1989] Watkins, CJCH (1989) Learning from Delayed Rewards. PhD thesis, Kings College, Cambridge, England.