

True Online Temporal-Difference Learning

Harm van Seijen^{†‡}
A. Rupam Mahmood[†]
Patrick M. Pilarski[†]
Marlos C. Machado[†]
Richard S. Sutton[†]

HARM.VANSEIJEN@MALUUBA.COM
ASHIQUE@UALBERTA.CA
PATRICK.PILARSKI@UALBERTA.CA
MACHADO@UALBERTA.CA
SUTTON@CS.UALBERTA.CA

[†]*Reinforcement Learning and Artificial Intelligence Laboratory
University of Alberta
2-21 Athabasca Hall, Edmonton, AB
Canada, T6G 2E8*

[‡]*Maluuba Research
2000 Peel Street, Montreal, QC
Canada, H3A 2W5*

Editor: George Konidaris

Abstract

The temporal-difference methods $\text{TD}(\lambda)$ and $\text{Sarsa}(\lambda)$ form a core part of modern reinforcement learning. Their appeal comes from their good performance, low computational cost, and their simple interpretation, given by their forward view. Recently, new versions of these methods were introduced, called true online $\text{TD}(\lambda)$ and true online $\text{Sarsa}(\lambda)$, respectively (van Seijen & Sutton, 2014). Algorithmically, these true online methods only make two small changes to the update rules of the regular methods, and the extra computational cost is negligible in most cases. However, they follow the ideas underlying the forward view much more closely. In particular, they maintain an exact equivalence with the forward view at all times, whereas the traditional versions only approximate it for small step-sizes. We hypothesize that these true online methods not only have better theoretical properties, but also dominate the regular methods empirically. In this article, we put this hypothesis to the test by performing an extensive empirical comparison. Specifically, we compare the performance of true online $\text{TD}(\lambda)/\text{Sarsa}(\lambda)$ with regular $\text{TD}(\lambda)/\text{Sarsa}(\lambda)$ on random MRPs, a real-world myoelectric prosthetic arm, and a domain from the Arcade Learning Environment. We use linear function approximation with tabular, binary, and non-binary features. Our results suggest that the true online methods indeed dominate the regular methods. Across all domains/representations the learning speed of the true online methods are often better, but never worse than that of the regular methods. An additional advantage is that no choice between traces has to be made for the true online methods. Besides the empirical results, we provide an in-depth analysis of the theory behind true online temporal-difference learning. In addition, we show that new true online temporal-difference methods can be derived by making changes to the online forward view and then rewriting the update equations.

Keywords: temporal-difference learning, eligibility traces, forward-view equivalence

1. Introduction

Temporal-difference (TD) learning is a core learning technique in modern reinforcement learning (Sutton, 1988; Kaelbling et al., 1996; Sutton & Barto, 1998; Szepesvari, 2010). One of the main challenges in reinforcement learning is to make predictions, in an initially unknown environment, about the (discounted) sum of future rewards, the return, based on currently observed feature values and a certain behaviour policy. With TD learning it is possible to learn good estimates of the expected return quickly by bootstrapping from other expected-return estimates. $\text{TD}(\lambda)$ (Sutton, 1988) is a popular TD algorithm that combines basic TD learning with eligibility traces to further speed learning. The popularity of $\text{TD}(\lambda)$ can be explained by its simple implementation, its low-computational complexity and its conceptually straightforward interpretation, given by its forward view. The forward view of $\text{TD}(\lambda)$ states that the estimate at each time step is moved towards an update target known as the λ -return, with λ determining the fundamental trade-off between bias and variance of the update target. This trade-off has a large influence on the speed of learning and its optimal setting varies from domain to domain. The ability to improve this trade-off by adjusting the value of λ is what underlies the performance advantage of eligibility traces.

Although the forward view provides a clear intuition, $\text{TD}(\lambda)$ closely approximates the forward view only for appropriately small step-sizes. Until recently, this was considered an unfortunate, but unavoidable part of the theory behind $\text{TD}(\lambda)$. This changed with the introduction of true online $\text{TD}(\lambda)$ (van Seijen & Sutton, 2014), which computes exactly the same weight vectors as the forward view at any step-size. This gives true online $\text{TD}(\lambda)$ full control over the bias-variance trade-off. In particular, true online $\text{TD}(1)$ can achieve fully unbiased updates. Moreover, true online $\text{TD}(\lambda)$ only requires small modifications to the $\text{TD}(\lambda)$ update equations, and the extra computational cost is negligible in most cases.

We hypothesize that true online $\text{TD}(\lambda)$, and its control version true online Sarsa(λ), not only have better theoretical properties than their regular counterparts, but also dominate them empirically. We test this hypothesis by performing an extensive empirical comparison between true online $\text{TD}(\lambda)$, regular $\text{TD}(\lambda)$ (which is based on accumulating traces), and the common variation based on replacing traces. In addition, we perform comparisons between true online Sarsa(λ) and Sarsa(λ) (with accumulating and replacing traces). The domains we use include random Markov reward processes, a real-world myoelectric prosthetic arm, and a domain from the Arcade Learning Environment (Bellemare et al., 2013). The representations we consider range from tabular values to linear function approximation with binary and non-binary features.

Besides the empirical study, we provide an in-depth discussion on the theory behind true online $\text{TD}(\lambda)$. This theory is based on a new online forward view. The traditional forward view, based on the λ -return, is inherently an offline forward view meaning that updates only occur at the end of an episode, because the λ -return requires data up to the end of an episode. We extend this forward view to the online case, where updates occur at every time step, by using a bounded version of the λ -return that grows over time. Whereas $\text{TD}(\lambda)$ approximates the traditional forward view only at the end of an episode, we show that $\text{TD}(\lambda)$ approximates this new online forward view at all time steps. True online $\text{TD}(\lambda)$ is equivalent to this new online forward view at all time steps. We prove this by deriving the true online $\text{TD}(\lambda)$ update equations directly from the online forward

view update equations. This derivation forms a blueprint for the derivation of other true online methods. By making variations to the online forward view and following the same derivation as for true online TD(λ), we derive several other true online methods.

This article is organized as follows. We start by presenting the required background in Section 2. Then, we present the new online forward view in Section 3, followed by the presentation of true online TD(λ) in Section 4. Section 5 presents the empirical study. Furthermore, in Section 6, we present several other true online methods. In Section 7, we discuss in detail related papers. Finally, Section 8 concludes.

2. Background

Here, we present the main learning framework. As a convention, we indicate scalar-valued random variables by capital letters (e.g., S_t, R_t), vectors by bold lowercase letters (e.g., θ, ϕ), functions by non-bold lowercase letters (e.g., v), and sets by calligraphic font (e.g., \mathcal{S}, \mathcal{A}).¹

2.1 Markov Decision Processes

Reinforcement learning (RL) problems are often formalized as *Markov decision processes* (MDPs), which can be described as 5-tuples of the form $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, where \mathcal{S} indicates the set of all states; \mathcal{A} indicates the set of all actions; $p(s'|s, a)$ indicates the probability of a transition to state $s' \in \mathcal{S}$, when action $a \in \mathcal{A}$ is taken in state $s \in \mathcal{S}$; $r(s, a, s')$ indicates the expected reward for a transition from state s to state s' under action a ; the discount factor γ specifies how future rewards are weighted with respect to the immediate reward.

Actions are taken at discrete time steps $t = 0, 1, 2, \dots$ according to a *policy* $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, which defines for each action the selection probability conditioned on the state. The *return* at time t is defined as the discounted sum of rewards, observed after t :

$$G_t := R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i},$$

where R_{t+1} is the reward received after taking action A_t in state S_t . Some MDPs contain special states called *terminal states*. After reaching a terminal state, no further reward is obtained and no further state transitions occur. Hence, a terminal state can be interpreted as a state where each action returns to itself with a reward of 0. An interaction sequence from the initial state to a terminal state is called an *episode*.

Each policy π has a corresponding state-value function v_π , which maps any state $s \in \mathcal{S}$ to the expected value of the return from that state, when following policy π :

$$v_\pi(s) := \mathbb{E}\{G_t | S_t = s, \pi\}.$$

In addition, the action-value function q_π gives the expected return for policy π , given that action $a \in \mathcal{A}$ is taken in state $s \in \mathcal{S}$:

$$q_\pi(s, a) := \mathbb{E}\{G_t | S_t = s, A_t = a, \pi\}.$$

1. An exception to this convention is the TD error, a scalar-valued random variable that we indicate by δ_t .

Because no further rewards can be obtained from a terminal state, the state-value and action-values for a terminal state are always 0.

There are two tasks that are typically associated with an MDP. First, there is the task of determining (an estimate of) the value function v_π for some given policy π . The second, more challenging task is that of determining (an estimate of) the optimal policy π_* , which is defined as the policy whose corresponding value function has the highest value in each state:

$$v_{\pi_*}(s) := \max_{\pi} v_{\pi}(s), \quad \text{for each } s \in \mathcal{S}.$$

In RL, these two tasks are considered under the condition that the reward function r and the transition-probability function p are unknown. Hence, the tasks have to be solved using samples obtained from interacting with the environment.

2.2 Temporal-Difference Learning

Let's consider the task of learning an estimate V of the value function v_π from samples, where v_π is being estimated using linear function approximation. That is, V is the inner product between a feature vector $\phi(s) \in \mathbb{R}^n$ of s , and a weight vector $\theta \in \mathbb{R}^n$:

$$V(s, \theta) = \theta^\top \phi(s).$$

If s is a terminal state, then by definition $\phi(s) := \mathbf{0}$, and hence $V(s, \theta) = 0$.

We can formulate the problem of estimating v_π as an error-minimization problem, where the error is a weighted average of the squared difference between the value of a state and its estimate:

$$E(\theta) := \frac{1}{2} \sum_i d_\pi(s_i) (v_\pi(s_i) - \theta^\top \phi(s_i))^2,$$

with d_π the stationary distribution induced by π . The above error function can be minimized by using stochastic gradient descent while sampling from the stationary distribution, resulting in the following update rule:

$$\theta_{t+1} = \theta_t - \alpha \frac{1}{2} \nabla_\theta (v_\pi(S_t) - \theta^\top \phi_t)^2,$$

using ϕ_t as a shorthand for $\phi(S_t)$. The parameter α is called the *step-size*. Using the chain rule, we can rewrite this update as:

$$\begin{aligned} \theta_{t+1} &= \theta_t + \alpha (v_\pi(S_t) - \theta^\top \phi_t) \nabla_\theta (\theta^\top \phi_t), \\ &= \theta_t + \alpha (v_\pi(S_t) - \theta^\top \phi_t) \phi_t. \end{aligned}$$

Because v_π is in general unknown, an estimate U_t of $v_\pi(S_t)$ is used, which we call the *update target*, resulting in the following general update rule:

$$\theta_{t+1} = \theta_t + \alpha (U_t - \theta^\top \phi_t) \phi_t. \tag{1}$$

There are many different update targets possible. For an unbiased estimator the full return can be used, that is, $U_t = G_t$. However, the full return has the disadvantage that its

variance is typically very high. Hence, learning with the full return can be slow. Temporal-difference (TD) learning addresses this issue by using update targets based on other value estimates. While the update target is no longer unbiased in this case, the variance is typically much smaller, and learning much faster. TD learning uses the Bellman equations as its mathematical foundation for constructing update targets. These equations relate the value of a state to the values of its successor states:

$$v_\pi(s) = \sum_a \pi(s, a) \sum_{s'} p(s'|s, a) (r(s, a, s') + \gamma v_\pi(s')) .$$

Writing this equation in terms of an expectation yields:

$$v_\pi(s) = \mathbb{E}\{R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s\}_{\pi, p, r} .$$

Sampling from this expectation, while using linear function approximation to approximate v_π , results in the update target:

$$U_t = R_{t+1} + \gamma \boldsymbol{\theta}^\top \boldsymbol{\phi}_{t+1} .$$

This update target is called a one-step update target, because it is based on information from only one time step ahead. Applying the Bellman equation multiple times results in update targets based on information further ahead. Such update targets are called multi-step update targets.

2.3 TD(λ)

The TD(λ) algorithm implements the following update equations:

$$\delta_t = R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t , \quad (2)$$

$$\boldsymbol{e}_t = \gamma \lambda \boldsymbol{e}_{t-1} + \boldsymbol{\phi}_t , \quad (3)$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \delta_t \boldsymbol{e}_t , \quad (4)$$

for $t \geq 0$, and with $\boldsymbol{e}_{-1} = \mathbf{0}$. The scalar δ_t is called the *TD error*, and the vector \boldsymbol{e}_t is called the *eligibility-trace* vector. The update of \boldsymbol{e}_t shown above is referred to as the *accumulating-trace* update. As a shorthand, we will refer to this version of TD(λ) as ‘accumulate TD(λ)’, to distinguish it from a slightly different version that is discussed below. While these updates appear to deviate from the general, gradient-descent-based update rule given in (1), there is a close connection to this update rule. This connection is formalized through the forward view of TD(λ), which we discuss in detail in the next section. Algorithm 1 shows the pseudocode for accumulate TD(λ).

Accumulate TD(λ) can be very sensitive with respect to the α and λ parameters. Especially, a large value of λ combined with a large value of α can easily cause divergence, even on simple tasks with bounded rewards. For this reason, a variant of TD(λ) is sometimes used that is more robust with respect to these parameters. This variant, which assumes binary features, uses a different trace-update equation:

$$\boldsymbol{e}_t[i] = \begin{cases} \gamma \lambda \boldsymbol{e}_{t-1}[i], & \text{if } \boldsymbol{\phi}_t[i] = 0; \\ 1, & \text{if } \boldsymbol{\phi}_t[i] = 1, \end{cases} \quad \text{for all features } i .$$

Algorithm 1 accumulate TD(λ)

INPUT: $\alpha, \lambda, \gamma, \boldsymbol{\theta}_{init}$
 $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_{init}$
Loop (over episodes):
obtain initial ϕ
 $e \leftarrow \mathbf{0}$
While terminal state has not been reached, do:
obtain next feature vector ϕ' and reward R
 $\delta \leftarrow R + \gamma \boldsymbol{\theta}^\top \phi' - \boldsymbol{\theta}^\top \phi$
 $e \leftarrow \gamma \lambda e + \phi$
 $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \delta e$
 $\phi \leftarrow \phi'$

where $\mathbf{x}[i]$ indicates the i -th component of vector \mathbf{x} . This update is referred to as the *replacing-trace* update. As a shorthand, we will refer to the version of TD(λ) using the replacing-trace update as ‘replace TD(λ)’.

3. The Online Forward View

The traditional forward view relates the TD(λ) update equations to the general update rule shown in Equation (1). Specifically, for small step-sizes the weight vector at the end of an episode computed by accumulate TD(λ) is approximately the same as the weight vector resulting from a sequence of Equation (1) updates (one for each visited state) using a particular multi-step update target, called the λ -return (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996). The λ -return for state S_t is defined as:

$$G_t^\lambda := (1 - \lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_t^{(n)} + \lambda^{T-t-1} G_t, \quad (5)$$

where T is the time step the terminal state is reached, and $G_t^{(n)}$ is the n -step return, defined as:

$$G_t^{(n)} := \sum_{k=1}^n \gamma^{k-1} R_{t+k} + \gamma^n V(S_{t+n} | \boldsymbol{\theta}_{t+n-1}).$$

We call a method that updates the value of each visited state at the end of the episode an *offline* method; we call a method that updates the value of each visited state immediately after the visit (i.e., at the time step after the visit) an *online* method. TD(λ) is an online method. The update sequence of the traditional forward view, however, corresponds with an offline method, because the λ -return requires data up to the end of an episode. This leaves open the question of how to interpret the weights of TD(λ) *during* an episode. In this section, we provide an answer to this long-standing open question. We introduce a bounded version of the λ -return that only uses information up to a certain horizon and we use this to construct an online forward view. This online forward view approximates the weight vectors of accumulate TD(λ) at *all* time steps, instead of only at the end of an episode.

3.1 The Online λ -Return Algorithm

The concept of an online forward view contains a paradox. On the one hand, multi-step update targets require data from time steps far beyond the time a state is visited; on the other hand, the online aspect requires that the value of a visited state is updated immediately. The solution to this paradox is to assign a sequence of update targets to each visited state. The first update target in this sequence contains data from only the next time step, the second contains data from the next two time steps, the third from the next three time steps, and so on. Now, given an initial weight vector and a sequence of visited states, a new weight vector can be constructed by updating each visited state with an update target that contains data up to the current time step. Below, we formalize this idea.

We define the *interim λ -return* for state S_k with horizon $h \in \mathbb{N}^+, h > k$ as follows:

$$G_k^{\lambda|h} := (1 - \lambda) \sum_{n=1}^{h-k-1} \lambda^{n-1} G_k^{(n)} + \lambda^{h-k-1} G_k^{(h-k)}. \quad (6)$$

Note that this update target does not use data beyond the horizon h . $G_k^{\lambda|h}$ implicitly defines a sequence of update targets for S_k : $\{G_k^{\lambda|k+1}, G_k^{\lambda|k+2}, G_k^{\lambda|k+3}, \dots\}$. As time increases, update targets based on data further away become available for state S_k . At a particular time step t , a new weight vector is computed by performing an Equation (1) update for each visited state using the interim λ -return with horizon t , starting from the initial weight vector θ_{init} . Hence, at time step t , a sequence of t updates occurs. To describe this sequence mathematically, we use weight vectors with two indices: θ_k^t . The superscript indicates the time step at which the updates are performed (this value corresponds with the horizon of the interim λ -returns that are used in the updates). The subscript is the iteration index of the sequence (it corresponds with the number of updates that have been performed at a particular time step). As an example, the update sequences for the first three time steps are:

$$t = 1 : \quad \theta_1^1 = \theta_0^1 + \alpha(G_0^{\lambda|1} - (\theta_0^1)^\top \phi_0)\phi_0,$$

$$\begin{aligned} t = 2 : \quad \theta_1^2 &= \theta_0^2 + \alpha(G_0^{\lambda|2} - (\theta_0^2)^\top \phi_0)\phi_0, \\ \theta_2^2 &= \theta_1^2 + \alpha(G_1^{\lambda|2} - (\theta_1^2)^\top \phi_1)\phi_1, \end{aligned}$$

$$\begin{aligned} t = 3 : \quad \theta_1^3 &= \theta_0^3 + \alpha(G_0^{\lambda|3} - (\theta_0^3)^\top \phi_0)\phi_0, \\ \theta_2^3 &= \theta_1^3 + \alpha(G_1^{\lambda|3} - (\theta_1^3)^\top \phi_1)\phi_1, \\ \theta_3^3 &= \theta_2^3 + \alpha(G_2^{\lambda|3} - (\theta_2^3)^\top \phi_2)\phi_2, \end{aligned}$$

with $\theta_0^t := \theta_{init}$ for all t . More generally, the update sequence at time step t is:

$$\theta_{k+1}^t := \theta_k^t + \alpha(G_k^{\lambda|t} - (\theta_k^t)^\top \phi_k)\phi_k, \quad \text{for } 0 \leq k < t. \quad (7)$$

We define θ_t (without superscript) as the final weight vector of the update sequence at time t , that is, $\theta_t := \theta_t^t$. We call the algorithm implementing Equation (7) the *online λ -return*

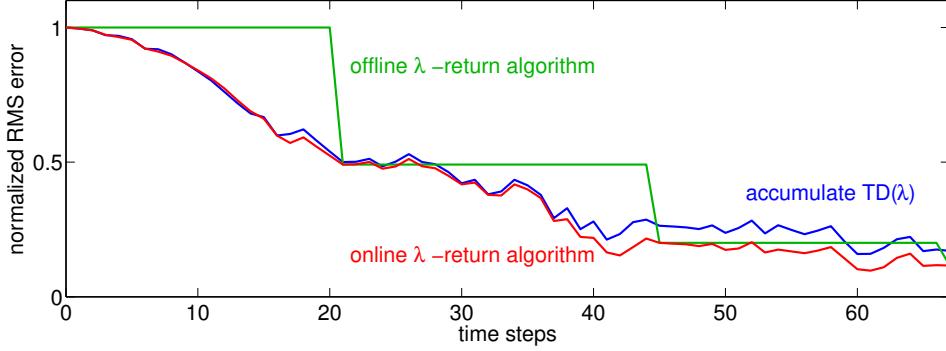


Figure 1: RMS error as function of time for the first 3 episodes of a random walk task, for $\lambda = 1$ and $\alpha = 0.2$. The error shown is the RMS error over all states, normalized by the initial RMS error.

algorithm. By contrast, we call the algorithm that implements the traditional forward view the *offline λ -return algorithm*.

The update sequence performed by the online λ -return algorithm at time step T (the time step that a terminal state is reached) is very similar to the update sequence performed by the offline λ -return algorithm. In particular, note that $G_t^{\lambda|T}$ and G_t^λ are the same, under the assumption that the weights used for the value estimates are the same. Because these weights are in practise not exactly the same, there will typically be a small difference.²

Figure 1 illustrates the difference between the online and offline λ -return algorithm, as well as accumulate $\text{TD}(\lambda)$, by showing the RMS error on a random walk task. The task consists of 10 states laid out in a row plus a terminal state on the left. Each state transitions with 70% probability to its left neighbour and with 30% probability to its right neighbour (or to itself in case of the right-most state). All rewards are 1 and $\gamma = 1$. Furthermore, $\lambda = 1$ and $\alpha = 0.2$. The right-most state is the initial state. Whereas the offline λ -return algorithm only makes updates at the end of an episode, the online λ -return algorithm, as well as accumulate $\text{TD}(\lambda)$, make updates at every time step.

The comparison on the random walk task shows that accumulate $\text{TD}(\lambda)$ behaves similar to the online λ -return algorithm. In fact, the smaller the step-size, the smaller the difference between accumulate $\text{TD}(\lambda)$ and the online λ -return algorithm. This is formalized by Theorem 1. The proof of the theorem can be found in Appendix A. The theorem uses the term Δ_i^t , which is defined as:

$$\Delta_i^t := (\bar{G}_i^{\lambda|t} - \boldsymbol{\theta}_0^\top \boldsymbol{\phi}_i) \boldsymbol{\phi}_i,$$

with $\bar{G}_i^{\lambda|t}$ the interim λ -return for state S_i with horizon t that uses $\boldsymbol{\theta}_0$ for all value evaluations. Note that Δ_i^t is independent of the step-size.

2. If $\lambda = 1$ there is never a difference because there is no bootstrapping.

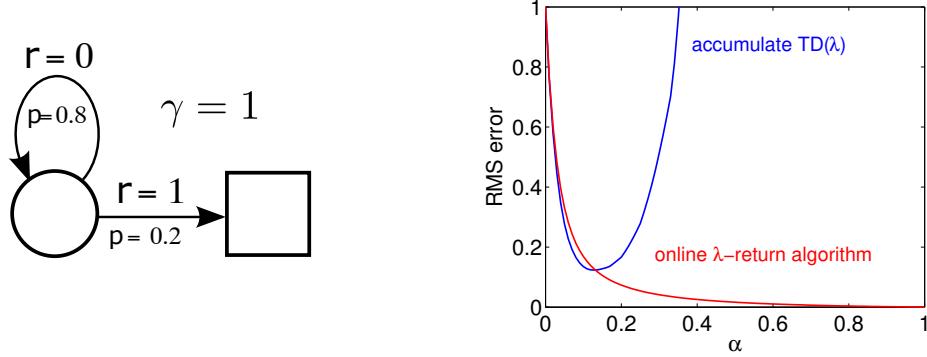


Figure 2: *Left:* One-state example (the square indicates a terminal state). *Right:* The RMS error of the state value at the end of an episode, averaged over the first 10 episodes, for $\lambda = 1$.

Theorem 1 Let $\boldsymbol{\theta}_0$ be the initial weight vector, $\boldsymbol{\theta}_t^{td}$ be the weight vector at time t computed by accumulate $TD(\lambda)$, and $\boldsymbol{\theta}_t^\lambda$ be the weight vector at time t computed by the online λ -return algorithm. Furthermore, assume that $\sum_{i=0}^{t-1} \Delta_i^t \neq \mathbf{0}$. Then, for all time steps t :

$$\frac{\|\boldsymbol{\theta}_t^{td} - \boldsymbol{\theta}_t^\lambda\|}{\|\boldsymbol{\theta}_t^{td} - \boldsymbol{\theta}_0\|} \rightarrow 0, \quad \text{as } \alpha \rightarrow 0.$$

Theorem 1 generalizes the traditional result to arbitrary time steps. The traditional result states that the difference between the weight vector at the end of an episode computed by the offline λ -return algorithm and the weight vector at the end of an episode computed by accumulate $TD(\lambda)$ goes to 0, if the step-size goes to 0 (Bertsekas & Tsitsiklis, 1996).

3.2 Comparison to Accumulate $TD(\lambda)$

While accumulate $TD(\lambda)$ behaves like the online λ -return algorithm for small step-sizes, small step-sizes often result in slow learning. Hence, higher step-sizes are desirable. For higher step-sizes, however, the behaviour of accumulate $TD(\lambda)$ can be very different from that of the online λ -return algorithm. And as we show in the empirical section of this article (Section 5), when there is a difference, it is almost exclusively in favour of the online λ -return algorithm. In this section, we analyze why the online λ -return algorithm can outperform accumulate $TD(\lambda)$, using the one-state example shown in the left of Figure 2.

The right of Figure 2 shows the RMS error over the first 10 episodes of the one-state example for different step-sizes and $\lambda = 1$. While for small step-sizes accumulate $TD(\lambda)$ behaves indeed like the online λ -return algorithm—as predicted by Theorem 1—, for larger step-sizes the difference becomes huge. To understand the reason for this, we derive an analytical expression for the value at the end of an episode.

First, we consider accumulate $TD(\lambda)$. Because there is only one state involved, we indicate the value of this state simply by V . The update at the end of an episode is $V_T =$

$V_{T-1} + \alpha e_{T-1} \delta_{T-1}$. In our example, $\delta_t = 0$ for all time steps t , except for $t = T - 1$, where $\delta_{T-1} = 1 - V_{T-1}$. Because δ_t is 0 for all time steps except the last, $V_{T-1} = V_0$. Furthermore, $\phi_t = 1$ for all time steps t , resulting in $e_{T-1} = T$. Substituting all this in the expression for V_T yields:

$$V_T = V_0 + T\alpha(1 - V_0), \quad \text{for accumulate TD}(\lambda). \quad (8)$$

So for accumulate TD(λ), the total value difference is simply a summation of the value difference corresponding to a single update.

Now, consider the online λ -return algorithm. The value at the end of an episode, V_T , is equal to V_T^T , resulting from the update sequence:

$$V_{k+1}^T = V_k^T + \alpha(G_k^{\lambda|T} - V_k^T), \quad \text{for } 0 \leq k < T.$$

By incremental substitution, we can directly express V_T in terms of the initial value, V_0 , and the update targets:

$$V_T = (1 - \alpha)^T V_0 + \alpha(1 - \alpha)^{T-1} G_0^{\lambda|T} + \alpha(1 - \alpha)^{T-2} G_1^{\lambda|T} + \cdots + \alpha G_{T-1}^{\lambda|T}.$$

Because $G_k^{\lambda|T} = 1$ for all k in our example, the weights of all update targets can be added together and the expression can be rewritten as a single pseudo-update, yielding:

$$V_T = V_0 + (1 - (1 - \alpha)^T) \cdot (1 - V_0), \quad \text{for the online } \lambda\text{-return algorithm.} \quad (9)$$

The term $1 - (1 - \alpha)^T$ in (9) acts like a pseudo step-size. For larger α or T this pseudo step-size increases in value, but as long as $\alpha \leq 1$ the value will never exceed 1. By contrast, for accumulate TD(λ) the pseudo step-size is $T\alpha$, which can grow much larger than 1 even for $\alpha < 1$, causing divergence of values. This is the reason that accumulate TD(λ) can be very sensitive to the step-size and it explains why the optimal step-size for accumulate TD(λ) is much smaller than the optimal step-size for the online λ -return algorithm in Figure 2 ($\alpha \approx 0.15$ versus $\alpha = 1$, respectively). Moreover, because the variance on the pseudo step-size is higher for accumulate TD(λ) the performance at the optimal step-size for accumulate TD(λ) is worse than the performance at the optimal step-size for the online λ -return algorithm.

3.3 Comparison to Replace TD(λ)

The sensitivity of accumulate TD(λ) to divergence, demonstrated in the previous subsection, has been known for long. In fact, replace TD(λ) was designed to deal with this. But while replace TD(λ) is much more robust with respect to divergence, it also has its limitations. One obvious limitation is that it only applies to binary features, so it is not generally applicable. But even in domains where replace TD(λ) can be applied, it can perform poorly. The reason is that replacing previous trace values, rather than adding to it, reduces the multi-step characteristics of TD(λ).

To illustrate this, consider the two-state example shown in the left of Figure 3. It is easy to see that the value of the left-most state is 2 and of the other state is 0. The state representation consists of only a single, binary feature that is 1 in both states and 0 in the terminal state. Because there is only a single feature, the state values cannot be represented

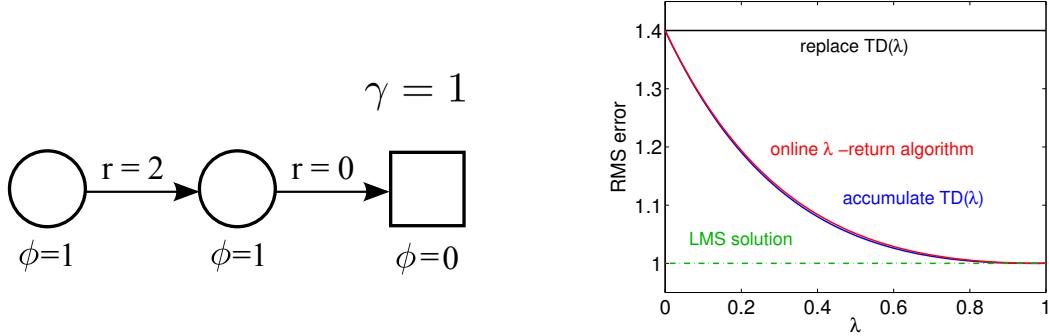


Figure 3: *Left:* Two-state example. *Right:* The RMS error after convergence for different λ (at $\alpha = 0.01$). We consider values to be converged if the error changed less than 1% over the last 100 time steps.

exactly. The weight that minimizes the mean squared error assigns a value of 1 to both states, resulting in an RMS error of 1. Now consider the graph shown in the right of Figure 3, which shows the asymptotic RMS error for different values of λ . The error for accumulate $\text{TD}(\lambda)$ converges to the least mean squares (LMS) error for $\lambda = 1$, as predicted by the theory (Dayan, 1992). The online λ -return algorithm has the same convergence behaviour (due to Theorem 1). By contrast, replace $\text{TD}(\lambda)$ converges to the same value as $\text{TD}(0)$ for any value of λ . The reason for this behaviour is that because the single feature is active at all time steps, the multi-step behaviour of $\text{TD}(\lambda)$ is fully removed, no matter the value of λ . Hence, replace $\text{TD}(\lambda)$ behaves exactly the same as $\text{TD}(0)$ for any value of λ at all time steps. As a result, it also behaves like $\text{TD}(0)$ asymptotically.

The two-state example very clearly demonstrates that there is a price payed by replace $\text{TD}(\lambda)$ to achieve robustness with respect to divergence: a reduction in multi-step behaviour. By contrast, the online λ -return algorithm, which is also robust to divergence, does not have this disadvantage. Of course, the two-state example, as well as the one-state example from the previous section, are extreme examples, merely meant to illustrate what can go wrong. But in practise, a domain will often have some characteristics of the one-state example and some of the two-state example, which negatively impacts the performance of both accumulate and replace $\text{TD}(\lambda)$.

4. True Online $\text{TD}(\lambda)$

The online λ -return algorithm is impractical on many domains: the memory it uses, as well as the computation required per time step increases linearly with time. Fortunately, it is possible to rewrite the update equations of the online λ -return algorithm to a different set of update equations that can be implemented with a computational complexity that is independent of time. In fact, this alternative set of update equations differs from the update equations of accumulate $\text{TD}(\lambda)$ only by two extra terms, each of which can be computed

Algorithm 2 true online TD(λ)

INPUT: $\alpha, \lambda, \gamma, \theta_{init}$
 $\theta \leftarrow \theta_{init}$
Loop (over episodes):
obtain initial ϕ
 $e \leftarrow \mathbf{0}; V_{old} \leftarrow 0$
While terminal state has not been reached, do:
obtain next feature vector ϕ' and reward R
 $V \leftarrow \theta^\top \phi$
 $V' \leftarrow \theta^\top \phi'$
 $\delta \leftarrow R + \gamma V' - V$
 $e \leftarrow \gamma \lambda e + \phi - \alpha \gamma \lambda (e^\top \phi) \phi$
 $\theta \leftarrow \theta + \alpha (\delta + V - V_{old}) e - \alpha (V - V_{old}) \phi$
 $V_{old} \leftarrow V'$
 $\phi \leftarrow \phi'$

efficiently. The algorithm implementing these equations is called true online TD(λ) and is discussed below.

4.1 The Algorithm

For the online λ -return algorithm, at each time step a sequence of updates is performed. The length of this sequence, and hence the computation per time step, increases over time. However, it is possible to compute the weight vector resulting from the sequence at time step $t+1$ directly from the weight vector resulting from the sequence at time step t . This results in the following update equations (see Appendix B for the derivation):

$$\delta_t = R_{t+1} + \gamma \theta_t^\top \phi_{t+1} - \theta_t^\top \phi_t, \quad (10)$$

$$e_t = \gamma \lambda e_{t-1} + \phi_t - \alpha \gamma \lambda (e_{t-1}^\top \phi_t) \phi_t, \quad (11)$$

$$\theta_{t+1} = \theta_t + \alpha \delta_t e_t + \alpha (\theta_t^\top \phi_t - \theta_{t-1}^\top \phi_t) (e_t - \phi_t), \quad (12)$$

for $t \geq 0$, and with $e_{-1} = \mathbf{0}$. Compared to accumulate TD(λ), both the trace update and the weight update have an additional term. We call a trace updated in this way a *dutch trace*; we call the term $\alpha(\theta_t^\top \phi_t - \theta_{t-1}^\top \phi_t)(e_t - \phi_t)$ the *TD-error time-step correction*, or simply the δ -correction. Algorithm 2 shows pseudocode that implements these equations.³

In terms of computation time, true online TD(λ) has a (slightly) higher cost due to the two extra terms that have to be accounted for. While the computation-time complexity of true online TD(λ) is the same as that of accumulate/replace TD(λ)— $\mathcal{O}(n)$ per time step with n being the number of features—, the actual computation time can be close to twice as much in some cases. In other cases (for example if sparse feature vectors are used), the computation time of true online TD(λ) is only a fraction more than that of

3. When using a time-dependent step-size (e.g., when annealing the step-size) use the pseudocode from Section 6.1. For reasons explained in that section this requires a modified trace update. That pseudocode is the same as the pseudocode from van Seijen & Sutton (2014).

accumulate/replace TD(λ). In terms of memory, true online TD(λ) has the same cost as accumulate/replace TD(λ).

4.2 When Can a Performance Difference be Expected?

In Section 3, a number of examples were shown where the online λ -return algorithm outperforms accumulate/replace TD(λ). Because true online TD(λ) is simply an efficient implementation of the online λ -return algorithm, true online TD(λ) will outperform accumulate/replace TD(λ) on these examples as well. But not in all cases will there be a performance difference. For example, it follows from Theorem 1 that when appropriately small step-sizes are used, the difference between the online λ -return algorithm/true online TD(λ) and accumulate TD(λ) is negligible. In this section, we identify two other factors that affect whether or not there will be a performance difference. While the focus of this section is on performance *difference* rather than performance *advantage*, our experiments will show that true online TD(λ) performs always at least as well as accumulate TD(λ) and replace TD(λ). In other words, our experiments suggest that whenever there is a performance difference, it is in favour of true online TD(λ).

The first factor is the λ parameter and follows straightforwardly from the true online TD(λ) update equations.

Proposition 1 *For $\lambda = 0$, accumulate TD(λ), replace TD(λ) and the online λ -return algorithm / true online TD(λ) behave the same.*

Proof For $\lambda = 0$, the accumulating-trace update, the replacing-trace update and the dutch-trace update all reduce to $e_t = \phi_t$. In addition, because $e_t = \phi_t$, the δ -correction of true online TD(λ) is 0. \blacksquare

Because the behaviour of TD(λ) for small λ is close to the behaviour of TD(0), it follows that significant performance differences will only be observed when λ is large.

The second factor is related to how often a feature has a non-zero value. We start again with a proposition that highlights a condition under which the different TD(λ) versions behave the same. The proposition makes use of an accumulating trace at time step $t - 1$, e_{t-1}^{acc} , whose non-recursive form is:

$$e_{t-1}^{acc} = \sum_{k=0}^{t-1} (\gamma\lambda)^{t-1-k} \phi_k. \quad (13)$$

Furthermore, the proposition uses $x[i]$ to denote the i -th element of vector x .

Proposition 2 *If for all features i and at all time steps t*

$$e_{t-1}^{acc}[i] \cdot \phi_t[i] = 0, \quad (14)$$

then accumulate TD(λ), replace TD(λ) and the online λ -return algorithm / true online TD(λ) behave the same (for any λ).

Proof Condition (14) implies that if $\phi_t[i] \neq 0$, then $e_{t-1}^{acc}[i] = 0$. From this it follows that for binary features the accumulating-trace update can be written as a replacing-trace

update at every time step:

$$\begin{aligned}\mathbf{e}_t^{acc}[i] &:= \gamma\lambda\mathbf{e}_{t-1}^{acc}[i] + \phi_t[i], \\ &= \begin{cases} \gamma\lambda\mathbf{e}_{t-1}^{acc}[i], & \text{if } \phi_t[i] = 0; \\ 1, & \text{if } \phi_t[i] = 1. \end{cases}\end{aligned}$$

Hence, accumulate TD(λ) and replace TD(λ) perform exactly the same updates.

Furthermore, condition (14) implies that $(\mathbf{e}_{t-1}^{acc})^\top \phi_t = 0$. Hence, the accumulating-trace update can also be written as a dutch trace update at every time step:

$$\begin{aligned}\mathbf{e}_t^{acc} &:= \gamma\lambda\mathbf{e}_{t-1}^{acc} + \phi_t, \\ &= \gamma\lambda\mathbf{e}_{t-1}^{acc} + \phi_t - \alpha\gamma\lambda((\mathbf{e}_{t-1}^{acc})^\top \phi_t) \phi_t.\end{aligned}$$

In addition, note that the δ -correction is proportional to $\theta_t^\top \phi_t - \theta_{t-1}^\top \phi_t$, which can be written as $(\theta_t - \theta_{t-1})^\top \phi_t$. The value $(\theta_t - \theta_{t-1})^\top \phi_t$ is proportional to $(\mathbf{e}_{t-1}^{acc})^\top \phi_t$ for accumulate TD(λ). Because $(\mathbf{e}_{t-1}^{acc})^\top \phi_t = 0$, accumulate TD(λ) can add a δ -correction at every time step without any consequence. This shows that accumulate TD(λ) makes the same updates as true online TD(λ). ■

An example of a domain where the condition of Proposition 2 holds is a domain with tabular features (each state is represented with a unique standard-basis vector), where a state is never revisited within the same episode.

The condition of Proposition 2 holds approximately when the value $|\mathbf{e}_{t-1}^{acc}[i] \cdot \phi_t[i]|$ is close to 0 for all features at all time steps. In this case, the different TD(λ) versions will perform very similarly. It follows from Equation (13) that this is the case when there is a long time delay between the time steps that a feature has a non-zero value. Specifically, if there is always at least n time steps between two subsequent times that a feature i has a non-zero value with $\gamma\lambda^n$ being very small, then $|\mathbf{e}_{t-1}^{acc}[i] \cdot \phi_t[i]|$ will always be close to 0. Therefore, in order to see a large performance difference, the same features should have a non-zero value often and within a small time frame (relative to $\gamma\lambda$).

Summarizing the analysis so far: in order to see a performance difference α and λ should be sufficiently large, and the same features should have a non-zero value often and within a small time frame. Based on this summary, we can address a related question: on what type of domains will there be a performance difference between true online TD(λ) with optimized parameters and accumulate/replace TD(λ) with optimized parameters. The first two conditions suggest that the domain should result in a relatively large optimal α and optimal λ . This is typically the case for domains with a relatively low variance on the return. The last condition can be satisfied in multiple ways. It is for example satisfied by domains that have non-sparse feature vectors (that is, domains for which at any particular time step most features have a non-zero value).

4.3 True Online Sarsa(λ)

TD(λ) and true online TD(λ) are policy evaluation methods. However, they can be turned into control methods in a straightforward way. From a learning perspective, the main difference is that the prediction of the expected return should be conditioned on the state

Algorithm 3 true online Sarsa(λ)

INPUT: $\alpha, \lambda, \gamma, \theta_{init}$
 $\theta \leftarrow \theta_{init}$

Loop (over episodes):

- obtain initial state S
- select action A based on state S (for example ϵ -greedy)
- $\psi \leftarrow$ features corresponding to S, A
- $e \leftarrow \mathbf{0}; Q_{old} \leftarrow 0$
- While terminal state has not been reached, do:
 - take action A , observe next state S' and reward R
 - select action A' based on state S'
 - $\psi' \leftarrow$ features corresponding to S', A' (if S' is terminal state, $\psi' \leftarrow \mathbf{0}$)
 - $Q \leftarrow \theta^\top \psi$
 - $Q' \leftarrow \theta^\top \psi'$
 - $\delta \leftarrow R + \gamma Q' - Q$
 - $e \leftarrow \gamma \lambda e + \psi - \alpha \gamma \lambda (e^\top \psi) \psi$
 - $\theta \leftarrow \theta + \alpha(\delta + Q - Q_{old}) e - \alpha(Q - Q_{old})\psi$
 - $Q_{old} \leftarrow Q'$
 - $\psi \leftarrow \psi'; A \leftarrow A'$

and action, rather than only on the state. This means that an estimate of the action-value function q_π is being learned, rather than of the state-value function v_π .

Another difference is that instead of having a fixed policy that generates the behaviour, the policy depends on the action-value estimates. Because these estimates typically improve over time, so does the policy. The (on-policy) control counterpart of TD(λ) is the popular Sarsa(λ) algorithm. The control counterpart of true online TD(λ) is ‘true online Sarsa(λ)’. Algorithm 3 shows pseudocode for true online Sarsa(λ).

To ensure accurate estimates for all state-action values are obtained, typically some exploration strategy has to be used. A simple, but often sufficient strategy is to use an ϵ -greedy behaviour policy. That is, given current state S_t , with probability ϵ a random action is selected, and with probability $1 - \epsilon$ the greedy action is selected:

$$A_t^{greedy} = \arg \max_a \theta_t^\top \psi(S_t, a),$$

with $\psi(s, a)$ an action-feature vector, and $\theta_t^\top \psi(s, a)$ a (linear) estimate of $q_\pi(s, a)$ at time step t . A common way to derive an action-feature vector $\psi(s, a)$ from a state-feature vector $\phi(s)$ involves an action-feature vector of size $n|\mathcal{A}|$, where n is the number of state features and $|\mathcal{A}|$ is the number of actions. Each action corresponds with a block of n features in this action-feature vector. The features in $\psi(s, a)$ that correspond to action a take on the values of the state features; the features corresponding to other actions have a value of 0.

5. Empirical Study

This section contains our main empirical study, comparing TD(λ), as well as Sarsa(λ), with their true online counterparts. For each method and each domain, a scan over the step-size

α and the trace-decay parameter λ is performed such that the optimal performance can be compared. In Section 5.4, we discuss the results.

5.1 Random MRPs

For our first series of experiments we used randomly constructed Markov reward processes (MRPs).⁴ An MRP can be interpreted as an MDP with only a single action per state. Consequently, there is only one policy possible. We represent a random MRP as a 3-tuple (k, b, σ) , consisting of k , the number of states; b , the branching factor (that is, the number of next states with a non-zero transition probability); and σ , the standard deviation of the reward. An MRP is constructed as follows. The b potential next states for a particular state are drawn from the total set of states at random, and without replacement. The transition probabilities to those states are randomized as well (by partitioning the unit interval at $b - 1$ random cut points). The expected value of the reward for a transition is drawn from a normal distribution with zero mean and unit variance. The actual reward is drawn from a normal distribution with a mean equal to this expected reward and standard deviation σ . There are no terminal states.

We compared the performance of $\text{TD}(\lambda)$ on three different MRPs: one with a small number of states, $(10, 3, 0.1)$, one with a larger number of states, $(100, 10, 0.1)$, and one with a larger number of states but a low branching factor and no stochasticity for the reward, $(100, 3, 0)$. The discount factor γ is 0.99 for all three MRPs. Each MRP is evaluated using three different representations. The first representation consists of *tabular* features, that is, each state is represented with a unique standard-basis vector of k dimensions. The second representation is based on *binary* features. This binary representation is constructed by first assigning indices, from 1 to k , to all states. Then, the binary encoding of the state index is used as a feature vector to represent that state. The length of a feature vector is determined by the total number of states: for $k = 10$, the length is 4; for $k = 100$, the length is 7. As an example, for $k = 10$ the binary feature vectors of states 1, 2 and 3 are $(0, 0, 0, 1), (0, 0, 1, 0)$ and $(0, 0, 1, 1)$, respectively. Finally, the third representation uses non-binary features. For this representation each state is mapped to a 5-dimensional feature vector, with the value of each feature drawn from a normal distribution with zero mean and unit variance. After all the feature values for a state are drawn, they are normalized such that the feature vector has unit length. Once generated, the feature vectors are kept fixed for each state. Note that replace $\text{TD}(\lambda)$ cannot be used with this representation, because replacing traces are only defined for binary features (tabular features are a special case of this).

In each experiment, we performed a scan over α and λ . Specifically, between 0 and 0.1, α is varied according to 10^i with i varying from -3 to -1 with steps of 0.2, and from 0.1 to 2.0 (linearly) with steps of 0.1. In addition, λ is varied from 0 to 0.9 with steps of 0.1 and from 0.9 to 1.0 with steps of 0.01. The initial weight vector is the zero vector in all domains. As performance metric we used the mean-squared error (MSE) with respect to the LMS solution during early learning (for $k = 10$, we averaged over the first 100 time steps; for k

4. The code for the MRP experiments is published online at: <https://github.com/armahmood/totd-rndmdp-experiments>. The process we used to construct the MRPs is based on the process used by Bhatnagar, Sutton, Ghavamzadeh and Lee (2009).

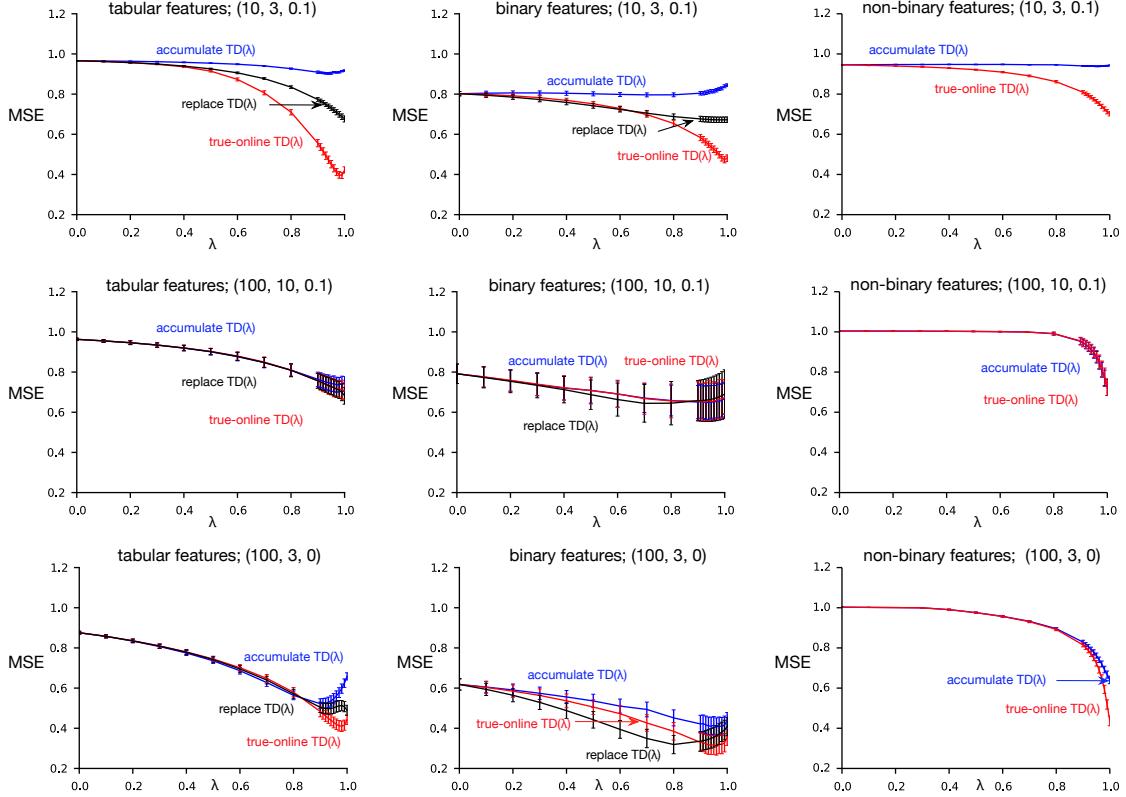


Figure 4: MSE error during early learning for three different MRPs, indicated by (k, b, σ) , and three different representations. The error shown is at optimal α value.

$= 100$, we averaged over the first 1000 time steps). We normalized this error by dividing it by the MSE under the initial weight estimate.

Figure 4 shows the results for different λ at the best value of α . In Appendix C, the results for all α values are shown. The optimal performance of true online $\text{TD}(\lambda)$ is at least as good as the optimal performance of accumulate $\text{TD}(\lambda)$ and replace $\text{TD}(\lambda)$, on all domains and for all representations. A more in-depth discussion of these results is provided in Section 5.4.

5.2 Predicting Signals From a Myoelectric Prosthetic Arm

In this experiment, we compared the performance of true online $\text{TD}(\lambda)$ and $\text{TD}(\lambda)$ on a real-world data-set consisting of sensorimotor signals measured during the human control of an electromechanical robot arm. The source of the data is a series of manipulation tasks performed by a participant with an amputation, as presented by Pilarski et al. (2013). In this study, an amputee participant used signals recorded from the muscles of their residual

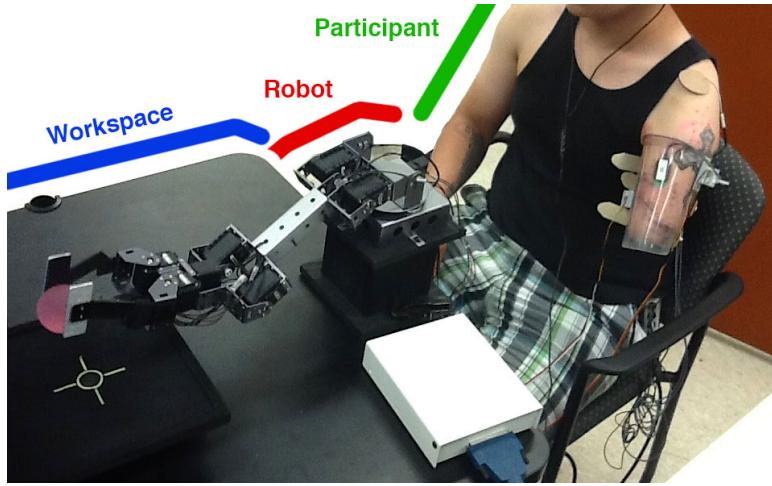


Figure 5: Source of the input data stream and predicted signals used in this experiment: a participant with an amputation performing a simple grasping task using a myoelectrically controlled robot arm, as described in Pilarski et al. (2013). More detail on the subject and experimental setting can be found in Hebert et al. (2014).

limb to control a robot arm with multiple degrees-of-freedom (Figure 5). Interactions of this kind are known as *myoelectric control* (see, for example, Parker et al., 2006).

For consistency and comparison of results, we used the same source data and prediction learning architecture as published in Pilarski et al. (2013). In total, two signals are predicted: grip force and motor angle signals from the robot’s hand. Specifically, the target for the prediction is a discounted sum of each signal over time, similar to return predictions (see general value functions and nexting; Sutton et al., 2011; Modayil et al., 2014). Where possible, we used the same implementation and code base as Pilarski et al. (2013). Data for this experiment consisted of 58,000 time steps of recorded sensorimotor information, sampled at 40 Hz (i.e., approximately 25 minutes of experimental data). The state space consisted of a tile-coded representation of the robot gripper’s position, velocity, recorded gripping force, and two muscle contraction signals from the human user. A standard implementation of tile-coding was used, with ten bins per signal, eight overlapping tilings, and a single active bias unit. This results in a state space with 800,001 features, 9 of which were active at any given time. Hashing was used to reduce this space down to a vector of 200,000 features that are then presented to the learning system. All signals were normalized between 0 and 1 before being provided to the function approximation routine. The discount factor for predictions of both force and angle was $\gamma = 0.97$, as in the results presented by Pilarski et al. (2013). Parameter sweeps over λ and α are conducted for all three methods. The performance metric is the mean absolute return error over all 58,000 time steps of learning, normalized by dividing by the error for $\lambda = 0$.

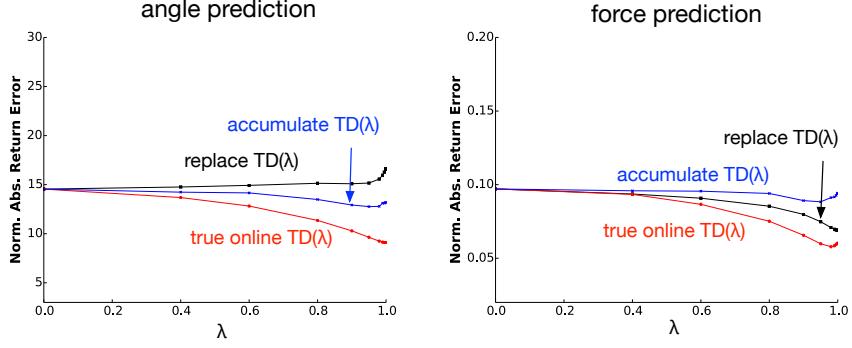


Figure 6: Performance as function of λ at the optimal α value, for the prediction of the servo motor angle (left), as well as the grip force (right).

Figure 6 shows the performance for the angle as well as the force predictions at the best α value for different values of λ . In Appendix D, the results for all α values are shown. The relative performance of replace TD(λ) and accumulate TD(λ) depends on the predictive question being asked. For predicting the robot’s grip force signal—a signal with small magnitude and rapid changes—replace TD(λ) is better than accumulate TD(λ) at all λ values larger than 0. However, for predicting the robot’s hand actuator position, a smoothly changing signal that varies between a range of $\sim 300\text{--}500$, accumulate TD(λ) dominates replace TD(λ). On both prediction tasks, true online TD(λ) dominates accumulate TD(λ) and replace TD(λ).

5.3 Control in the ALE Domain Asterix

In this final experiment, we compared the performance of true online Sarsa(λ) with that of accumulate Sarsa(λ) and replace Sarsa(λ), on a domain from the Arcade Learning Environment (ALE) (Bellemare et al., 2013; Defazio & Graepel, 2014; Mnih et al., 2015), called Asterix. The ALE is a general testbed that provides an interface to hundreds of Atari 2600 games.⁵

In the Asterix domain, the agent controls a yellow avatar, which has to collect ‘potion’ objects, while avoiding ‘harp’ objects (see Figure 7 for a screenshot). Both potions and harps move across the screen horizontally. Every time the agent collects a potion it receives a reward of 50 points, and every time it touches a harp it loses a life (it has three lives in total). The agent can use the actions *up*, *right*, *down*, and *left*, combinations of two directions, and a *no-op* action, resulting in 9 actions in total. The game ends after the agent has lost three lives, or after 5 minutes, whichever comes first.

We use linear function approximation using features derived from the screen pixels. Specifically, we use what Bellemare et al. (2013) call the *Basic* feature set, which “encodes

5. We used ALE version 0.4.4 for our experiments. The code for the Asterix experiments is published online at: <https://github.com/mcmachado/TrueOnlineSarsa>.

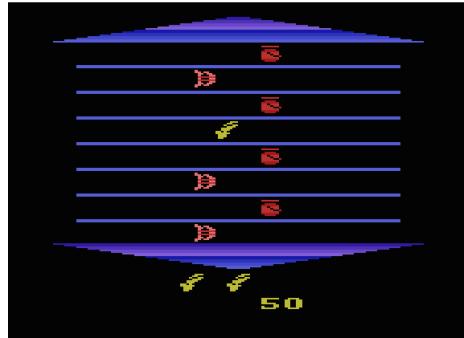


Figure 7: Screenshot of the game ASTERIX.

the presence of colours on the Atari 2600 screen.” It is obtained by first subtracting the game screen background (see Bellemare et al., 2013, sec. 3.1.1) and then dividing the remaining screen in to 16×14 tiles of size 10×15 pixels. Finally, for each tile, one binary feature is generated for each of the 128 available colours, encoding whether a colour is active or not in that tile. This generates 28,672 features (plus a bias term).

Because episode lengths can vary hugely (from about 10 seconds all the way up to 5 minutes), constructing a fair performance metric is non-trivial. For example, comparing the average return on the first N episodes of two methods is only fair if they have seen roughly the same amount of samples in those episodes, which is not guaranteed for this domain. On the other hand, looking at the total reward collected for the first X samples is also not a good metric, because there is no negative reward associated to dying. To resolve this, we look at the return per episode, averaged over the first X samples. More specifically, our metric consists of the average score per episode while learning for 20 hours (4,320,000 frames). In addition, we averaged the resulting number over 400 independent runs.

As with the evaluation experiments, we performed a scan over the step-size α and the trace-decay parameter λ . Specifically, we looked at all combinations of $\alpha \in \{0.20, 0.50, 0.80, 1.10, 1.40, 1.70, 2.00\}$ and $\lambda \in \{0.00, 0.50, 0.80, 0.90, 0.95, 0.99\}$ (these values were determined during a preliminary parameter sweep). We used a discount factor $\gamma = 0.999$ and ϵ -greedy exploration with $\epsilon = 0.01$. The weight vector was initialized to the zero vector. Also, as Bellemare et al. (2013), we take an action at each 5 frames. This decreases the algorithms running time and avoids “super-human” reflexes. The results are shown in Figure 8. On this domain, the optimal performance of all three versions of $\text{Sarsa}(\lambda)$ is similar.

Note that the way we evaluate a domain is computationally very expensive: we perform scans over λ and α , and use a large number of independent runs to get a low standard error. In the case of Asterix, this results in a total of $7 \cdot 6 \cdot 400 = 16,800$ runs per method. This rigorous evaluation prohibits us unfortunately to run experiments on the full suite of ALE domains.

5.4 Discussion

Figure 9 summarizes the performance of the different $\text{TD}(\lambda)$ versions on all evaluation domains. Specifically, it shows the error for each method at its best settings of α and λ . The error is normalized by dividing it by the error at $\lambda = 0$ (remember that all versions of

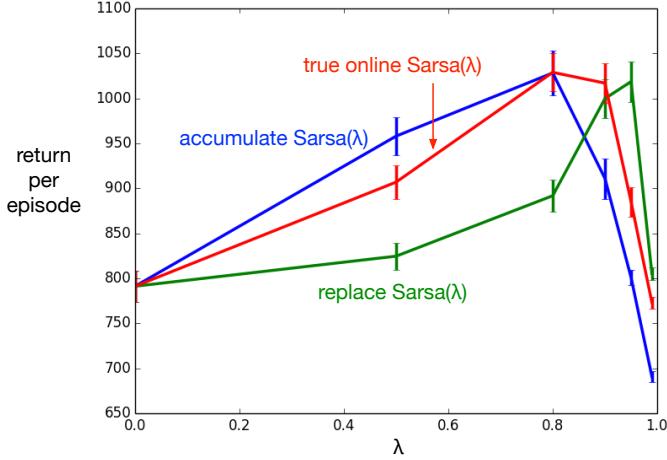


Figure 8: Return per episode, averaged over the first 4,320,000 frames as well as 400 independent runs, as function of λ , at optimal α , on the Asterix domain.

$\text{TD}(\lambda)$ behave the same for $\lambda = 0$). Because $\lambda = 0$ lies in the parameter range that is being optimized over, the normalized error can never be higher than 1. If for a method/domain the normalized error is equal to 1, this means that setting λ higher than 0 either has no effect, or that the error gets worse. In either case, eligibility traces are not effective for that method/domain.

Overall, true online $\text{TD}(\lambda)$ is clearly better than accumulate $\text{TD}(\lambda)$ and replace $\text{TD}(\lambda)$ in terms of optimal performance. Specifically, for each considered domain/representation, the error for true online $\text{TD}(\lambda)$ is either smaller or equal to the error of accumulate/replace $\text{TD}(\lambda)$. This is especially impressive, given the wide variety of domains, and the fact that the computational overhead for true online $\text{TD}(\lambda)$ is small (see Section 4.1 for details).

The observed performance differences correspond well with the analysis from Section 4.2. In particular, note that MRP (10, 3, 0.1) has less states than the other two MRPs, and hence the chance that the same feature has a non-zero value within a small time frame is larger. The analysis correctly predicts that this results in larger performance differences. Furthermore, MRP (100, 3, 0) is less stochastic than MRP (100, 10, 0.1), and hence it has a smaller variance on the return. Also here, the experiments correspond with the analysis, which predicts that this results in a larger performance difference.

On the Asterix domain, the performance of the three $\text{Sarsa}(\lambda)$ versions is similar. This is in accordance with the evaluation results, which showed that the size of the performance difference is domain dependent. In the worst case, the performance of the true online method is similar to that of the regular method.

The optimal performance is not the only factor that determines how good a method is; what also matters is how easy it is to find this performance. The detailed plots in appendices C and D reveal that the parameter sensitivity of accumulate $\text{TD}(\lambda)$ is much higher than

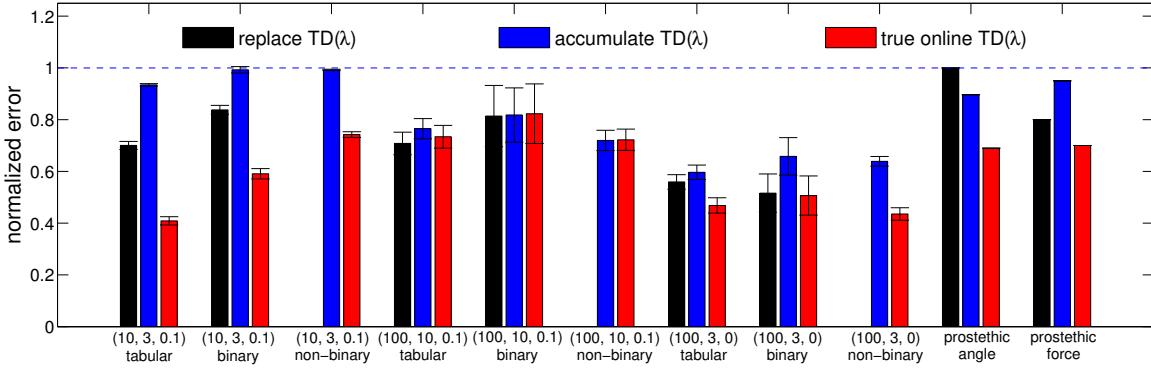


Figure 9: Summary of the evaluation results: error at optimal (α, λ) -settings for all domains/representations, normalized with the $\text{TD}(0)$ error.

that of true online $\text{TD}(\lambda)$ or replace $\text{TD}(\lambda)$. This is clearly visible for MRP (10, 3, 0.1) (Figure 10), as well as the experiments with the myoelectric prosthetic arm (Figure 13).

There is one more thing to take away from the experiments. In MRP (10, 3, 0.1) with non-binary features, replace $\text{TD}(\lambda)$ is not applicable and accumulate $\text{TD}(\lambda)$ is ineffective. However, true online $\text{TD}(\lambda)$ was still able to obtain a considerable performance advantage with respect to $\text{TD}(0)$. This demonstrates that true online $\text{TD}(\lambda)$ expands the set of domains/representations where eligibility traces are effective. This could potentially have far-reaching consequences. Specifically, using non-binary features becomes a lot more interesting. Replacing traces are not applicable to such representations, while accumulating traces can easily result in divergence of values. For true online $\text{TD}(\lambda)$, however, non-binary features are not necessarily more challenging than binary features. Exploring new, non-binary representations could potentially further improve the performance for true online $\text{TD}(\lambda)$ on domains such as the myoelectric prosthetic arm or the Asterix domain.

6. Other True Online Methods

In Appendix B, it is shown that the true online $\text{TD}(\lambda)$ equations can be derived directly from the online forward view equations. By using different online forward views, new true online methods can be derived. Sometimes, small changes in the forward view, like using a time-dependent step-size, can result in surprising changes in the true online equations. In this section, we look at a number of such variations.

6.1 True Online $\text{TD}(\lambda)$ with Time-Dependent Step-Size

When using a time-dependent step-size in the base equation of the forward view (Equation 7) and deriving the update equations following the procedure from Appendix B, it turns out that a slightly different trace definition appears. We indicate this new trace using a ‘+’ superscript: e_t^+ . For fixed step-size, this new trace definition is equal to:

$$e_t^+ = \alpha e_t, \quad \text{for all } t.$$

Algorithm 4 true online TD(λ) with time-dependent step-size

INPUT: $\lambda, \theta_{init}, \alpha_t$ for $t \geq 0$
 $\theta \leftarrow \theta_{init}; t \leftarrow 0$
Loop (over episodes):
 obtain initial ϕ
 $e^+ \leftarrow \mathbf{0}; V_{old} \leftarrow 0$
 While terminal state is not reached, do:
 obtain next feature vector ϕ' and reward R
 $V \leftarrow \theta^\top \phi$
 $V' \leftarrow \theta^\top \phi'$
 $\delta' \leftarrow R + \gamma V' - V_{old}$
 $e^+ \leftarrow \gamma \lambda e^+ + \alpha_t \phi - \alpha_t \gamma \lambda ((e^+)^\top \phi) \phi$
 $\theta \leftarrow \theta + \delta' e^+ - \alpha_t (V - V_{old}) \phi$
 $V_{old} \leftarrow V'$
 $\phi \leftarrow \phi'$
 $t \leftarrow t + 1$

Of course, using e_t^+ instead of e_t also changes the weight vector update slightly. Below, the full set of update equations is shown:

$$\begin{aligned}\delta_t &= R_{t+1} + \gamma \theta_t^\top \phi_{t+1} - \theta_t^\top \phi_t, \\ e_t^+ &= \gamma \lambda e_{t-1}^+ + \alpha_t \phi_t - \alpha_t \gamma \lambda ((e_{t-1}^+)^\top \phi_t) \phi_t, \\ \theta_{t+1} &= \theta_t + \delta_t e_t^+ + (\theta_t^\top \phi_t - \theta_{t-1}^\top \phi_t) (e_t^+ - \alpha_t \phi_t).\end{aligned}$$

In addition, $e_{-1}^+ := 0$. We can simplify the weight update equation slightly, by using

$$\delta'_t = \delta_t + \theta_t^\top \phi_t - \theta_{t-1}^\top \phi_t,$$

which changes the update equations to:⁶

$$\delta'_t = R_{t+1} + \gamma \theta_t^\top \phi_{t+1} - \theta_{t-1}^\top \phi_t, \quad (15)$$

$$e_t^+ = \gamma \lambda e_{t-1}^+ + \alpha_t \phi_t - \alpha_t \gamma \lambda ((e_{t-1}^+)^\top \phi_t) \phi_t, \quad (16)$$

$$\theta_{t+1} = \theta_t + \delta'_t e_t^+ - \alpha_t (\theta_t^\top \phi_t - \theta_{t-1}^\top \phi_t) \phi_t. \quad (17)$$

Algorithm 2 shows the corresponding pseudocode. Of course, this pseudocode can also be used for constant step-size.

6.2 True Online Version of Watkins's Q(λ)

So far, we just considered *on-policy* methods, that is, methods that evaluate a policy that is the same as the policy that generates the samples. However, the true online principle can also be applied to *off-policy* methods, for which the evaluation policy is different from the behaviour policy. As a simple example, consider Watkins's Q(λ) (Watkins, 1989). This is an off-policy method that evaluates the greedy policy given an arbitrary behaviour policy.

6. These equations are the same as in the original true online paper (van Seijen & Sutton, 2014).

Algorithm 5 true online version of Watkins's Q(λ)

INPUT: $\alpha, \lambda, \gamma, \theta_{init}, \Psi$ $\theta \leftarrow \theta_{init}$

Loop (over episodes):

 obtain initial state S select action A based on state S (for example ϵ -greedy) $\psi \leftarrow$ features corresponding to S, A $e \leftarrow \mathbf{0}; Q_{old} \leftarrow 0$

While terminal state has not been reached, do:

 take action A , observe next state S' and reward R select action A' based on state S' $A^* \leftarrow \arg \max_a \theta^\top \psi(S', a)$ (if A' ties for the max, then $A^* \leftarrow A'$) $\psi' \leftarrow$ features corresponding to S', A^* (if S' is terminal state, $\psi' \leftarrow \mathbf{0}$) $Q \leftarrow \theta^\top \psi$ $Q' \leftarrow \theta^\top \psi'$ $\delta \leftarrow R + \gamma Q' - Q$ $e \leftarrow \gamma \lambda e + \psi - \alpha \gamma \lambda (e^\top \psi) \psi$ $\theta \leftarrow \theta + \alpha(\delta + Q - Q_{old}) e - \alpha(Q - Q_{old})\psi$ if $A' \neq A^*$: $e \leftarrow \mathbf{0}$ $Q_{old} \leftarrow Q'$ $\psi \leftarrow \psi'; A \leftarrow A'$

It does this by combining accumulating traces with a TD error that uses the maximum state-action value of the successor state:

$$\delta_t = R_{t+1} + \max_a Q(S_t, a) - Q(S_t, A_t).$$

In addition, traces are reset to 0 whenever a non-greedy action is taken.

From an online forward-view perspective, the strategy of Watkins's Q(λ) method can be interpreted as a growing update target that stops growing once a non-greedy action is taken. Specifically, let τ be the first time step *after* time step t that a non-greedy action is taken, then the interim update target for time step t can be defined as:

$$U_t^h := (1 - \lambda) \sum_{n=1}^{z-t-1} \lambda^{n-1} \tilde{G}_t^{(n)} + \lambda^{z-t-1} \tilde{G}_t^{(z-t)}, \quad z = \min\{h, \tau\},$$

with

$$\tilde{G}_t^{(n)} = \sum_{k=1}^n \gamma^{k-1} R_{t+k} + \gamma^n \max_a \theta_{t+n-1}^\top \psi(S_{t+n}, a).$$

Algorithm 5 shows the pseudocode for the true online method that corresponds with this update target definition. A problem with Watkins's Q(λ) is that if the behaviour policy is very different from the greedy policy traces are reset very often, reducing the overall effect of the traces. In Section 7, we discuss more advanced off-policy methods.

Algorithm 6 tabular true online TD(λ)

```

initialize  $v(s)$  for all  $s$ 
Loop (over episodes):
    initialize  $S$ 
     $e(s) \leftarrow 0$  for all  $s$ 
     $V_{old} \leftarrow 0$ 
    While  $S$  is not terminal, do:
        obtain next state  $S'$  and reward  $R$ 
         $\Delta V \leftarrow V(S) - V_{old}$ 
         $V_{old} \leftarrow V(S')$ 
         $\delta \leftarrow R + \gamma V(S') - V(S)$ 
         $e(S) \leftarrow (1 - \alpha)e(S) + 1$ 
        For all  $s$ :
             $V(s) \leftarrow V(s) + \alpha(\delta + \Delta V)e(s)$ 
             $e(s) \leftarrow \gamma\lambda e(s)$ 
         $V(S) \leftarrow V(S) - \alpha\Delta V$ 
         $S \leftarrow S'$ 

```

6.3 Tabular True Online TD(λ)

Tabular features are a special case of linear function approximation. Hence, the update equations for true online TD(λ) that are presented so far also apply to the tabular case. However, we discuss it here separately, because the simplicity of this special case can provide extra insight.

Rewriting the true online update equations (equations 10 – 12) for the special case of tabular features results in:

$$\begin{aligned}
\delta_t &= R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t), \\
e_t(s) &= \begin{cases} \gamma\lambda e_{t-1}(s), & \text{if } s \neq S_t; \\ (1 - \alpha)\gamma\lambda e_{t-1}(s) + 1, & \text{if } s = S_t, \end{cases} \\
V_{t+1}(s) &= \begin{cases} V_t(s) + \alpha(\delta_t + V_t(S_t) - V_{t-1}(S_t)) e_t(s), & \text{if } s \neq S_t; \\ V_t(s) + \alpha(\delta_t + V_t(S_t) - V_{t-1}(S_t)) e_t(s) - \alpha(V_t(S_t) - V_{t-1}(S_t)), & \text{if } s = S_t. \end{cases}
\end{aligned}$$

What is interesting about the tabular case is that the dutch-trace update reduces to a particularly simple form. In fact, for the tabular case, a dutch-trace update is equal to the weighted average between an accumulating-trace update and a replacing-trace update, with the weight of the former $(1 - \alpha)$ and the latter α . Algorithm 6 shows the corresponding pseudocode.

7. Related Work

Since the first publication on true online TD(λ) (van Seijen & Sutton, 2014), several related papers have been published, extending the underlying concepts and improving the presentation. In sections 7.1, 7.2 and 7.3, we review those papers. In Section 7.4, we discuss other variations of TD(λ).

7.1 True Online Learning and Dutch Traces

As mentioned before, the traditional forward view, which is based on the λ -return, is inherently an offline forward view, because the λ -return is constructed from data up to the end of an episode. As a consequence, the work regarding equivalence between a forward view and a backward view traditionally focused on the final weight vector $\boldsymbol{\theta}_T$. This changed in 2014, when two papers introduced an *online* forward view with a corresponding backward view that has an exact equivalence at each moment in time (van Seijen & Sutton, 2014; Sutton et al., 2014). While both papers introduced an online forward view, the two forward views presented are very different from each other. One difference is that the forward view introduced by van Seijen & Sutton is an on-policy forward view, whereas the forward view by Sutton et al. is an off-policy forward view. However, there is an even more fundamental difference related to how the forward views are constructed. In particular, the forward view by van Seijen & Sutton is constructed in such a way that at each moment in time the weight vector can be interpreted as the result of a sequence of updates of the form:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha(U_k - \boldsymbol{\theta}_k^\top \phi_k) \phi_k, \quad \text{for } 0 \leq k < t. \quad (18)$$

By contrast, the forward view by Sutton et al. gives the following interpretation:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \alpha \sum_{k=0}^{t-1} \delta_k \phi_k, \quad (19)$$

with δ_k some multi-step TD error. Of course, the different forward views also result in different backward views. Whereas the backward view of Sutton et al. uses a generalized version of an accumulating trace, the backward view of van Seijen & Sutton introduced a completely new type of trace.

The advantage of a forward view based on (18) instead of (19) is that it results in much more stable updates. In particular, the sensitivity to divergence of accumulate $\text{TD}(\lambda)$ is a general side-effect of (19), whereas (18) is much more robust with respect to divergence. As a result, true online $\text{TD}(\lambda)$ not only has the property that it has an exact equivalence with an online forward view at all times, it consistently dominates $\text{TD}(\lambda)$ empirically.

The strong performance of true online $\text{TD}(\lambda)$ motivated van Hasselt et al. (2014) to construct an off-policy version of the forward view of true online $\text{TD}(\lambda)$. The corresponding backward view resulted in the algorithm true online GTD(λ), which empirically outperforms GTD(λ). They also introduced the term ‘dutch traces’ for the new eligibility trace.

Van Hasselt & Sutton (2015) showed that dutch traces are not only useful for TD learning. In an offline setting with no bootstrapping using dutch traces can result in certain computational advantages. To understand why, consider the Monte Carlo algorithm (MC), which updates state values at the end of an episode using (18), with the full return as update target. MC requires storing all the feature vectors and rewards during an episode, so the memory complexity is linear in the length of the episode. Moreover, the required computation time is distributed very unevenly: during an episode almost no computation is required, but at the end of an episode there is a huge peak in computation time due to all the updates that need to be performed. With dutch traces an alternative implementation can be made that results in the same final weight vector but that does not require storing all

the feature vectors and where the required computation time is spread out evenly over all the time steps. Van Hasselt & Sutton refer to this appealing property as span-independence: the memory and computation time required per time step is constant and independent of the span of the prediction.⁷

7.2 Backward View Derivation

The task of finding an efficient backward view that corresponds exactly with a particular online forward view is not easy. Moreover, there is no guarantee that there exists an efficient implementation of a particular online forward view. Often, minor changes in the forward view determine whether or not an efficient backward view can be constructed. This created the desire to somehow automate the process of constructing an efficient backward view.

Van Seijen & Sutton (2014) did not provide a direct derivation of the backward view update equations; they simply proved that the forward view and the backward view equations result in the same weight vectors. Sutton et al. (2014) were the first to attempt to come up with a general strategy for deriving a backward view (although for forward views based on Equation 19). Van Hasselt et al. (2014) took the approach of providing a theorem that proves equivalence between a general forward view and a corresponding general backward view. They showed that the forward/backward view of true online TD(λ) is a special case of this general forward/backward view. They showed the same for the off-policy method that they introduced—true online GTD(λ). Recently, Mahmood & Sutton (2015) extended this theorem further by proving equivalence between an even more general forward view and backward view.

Furthermore, van Hasselt & Sutton (2015) derived backward views for a series of increasingly complex forward views. The derivation of the true online TD(λ) equations in Appendix B is similar to those derivations.

7.3 Extension to Non-Linear Function Approximation

The linear update equations of the online forward view presented in Section 3.1 can be easily extended to the case of non-linear function approximation. Unfortunately, it appears to be impossible to construct an efficient backward view with exact equivalence in the case of non-linear function approximation. The reason is that the derivation in Appendix B makes use of the fact that the gradient with respect to the value function is independent of the weight vector; this does not hold for non-linear function approximation.

Fortunately, van Seijen (2016) shows that many of the benefits of true online learning can also be achieved in the case of non-linear function approximation by using an alternative forward view (but still based on Equation 18). While this alternative forward view is not fully online (there is a delay in the updates), it can be implemented efficiently.

7.4 Other Variations on TD(λ)

Several variations on TD(λ) other than those treated in this article have been suggested in the literature. Schapire & Warmuth (1996) introduced a variation of TD(λ) for which

7. The span of the prediction refers to the time difference between the first prediction and the moment its target is known (e.g., for episodic tasks it corresponds to the length of an episode).

upper and lower bounds on performance can be derived and proven. Konidaris et al. (2011) introduced TD_γ , a parameter-free alternative to $\text{TD}(\lambda)$ based on a multi-step update target called the γ -return. TD_γ is an offline algorithm with a computational cost proportional to the episode-length. Furthermore, Thomas et al. (2015) proposed a method based on a multi-step update target, which they call the Ω -return. The Ω -return can account for the correlation of different length returns, something that both the λ -return and the γ -return cannot. However, it is expensive to compute and it is open question whether efficient approximations exist.

8. Conclusions

We tested the hypothesis that true online $\text{TD}(\lambda)$ (and true online $\text{Sarsa}(\lambda)$) dominates $\text{TD}(\lambda)$ (and $\text{Sarsa}(\lambda)$) with accumulating as well as with replacing traces by performing experiments over a wide range of domains. Our extensive results support this hypothesis. In terms of learning speed, true online $\text{TD}(\lambda)$ was often better, but never worse than $\text{TD}(\lambda)$ with either accumulating or replacing traces, across all domains/representations that we tried. Our analysis showed that especially on domains with non-sparse features and a relatively low variance on the return a large difference in learning speed can be expected. More generally, true online $\text{TD}(\lambda)$ has the advantage over $\text{TD}(\lambda)$ with replacing traces that it can be used with non-binary features, and it has the advantage over $\text{TD}(\lambda)$ with accumulating traces that it is less sensitive with respect to its parameters. In terms of computation time, $\text{TD}(\lambda)$ has a slight advantage. In the worst case, true online $\text{TD}(\lambda)$ is twice as expensive. In the typical case of sparse features, it is only fractionally more expensive than $\text{TD}(\lambda)$. Memory requirements are the same. Finally, we outlined an approach for deriving new true online methods, based on rewriting the equations of an online forward view. This may lead to new, interesting methods in the future.

Acknowledgments

The authors thank Hado van Hasselt for extensive discussions leading to the refinement of these ideas. Furthermore, the authors thank the anonymous reviewers for their valuable suggestions, resulting in a substantially improved presentation. This work was supported by grants from Alberta Innovates – Technology Futures and the National Science and Engineering Research Council of Canada. Computing resources were provided by Compute Canada through WestGrid.

Appendix A. Proof of Theorem 1

Theorem 1 Let $\boldsymbol{\theta}_0$ be the initial weight vector, $\boldsymbol{\theta}_t^{td}$ be the weight vector at time t computed by accumulate TD(λ), and $\boldsymbol{\theta}_t^\lambda$ be the weight vector at time t computed by the online λ -return algorithm. Furthermore, assume that $\sum_{i=0}^{t-1} \Delta_i^t \neq \mathbf{0}$. Then, for all time steps t :

$$\frac{\|\boldsymbol{\theta}_t^{td} - \boldsymbol{\theta}_t^\lambda\|}{\|\boldsymbol{\theta}_t^{td} - \boldsymbol{\theta}_0\|} \rightarrow 0, \quad \text{as } \alpha \rightarrow 0.$$

Proof We prove the theorem by showing that $\|\boldsymbol{\theta}_t^{td} - \boldsymbol{\theta}_t^\lambda\|/\|\boldsymbol{\theta}_t^{td} - \boldsymbol{\theta}_0\|$ can be approximated by $\mathcal{O}(\alpha)/(C + \mathcal{O}(\alpha))$ as $\alpha \rightarrow 0$, with $C > 0$. For readability, we will not use the ‘td’ and ‘ λ ’ superscripts; instead, we always use weights with double indices for the online λ -return algorithm and weights with single indices for accumulate TD(λ).

The update equations for accumulate TD(λ) are:

$$\begin{aligned}\delta_t &= R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t, \\ \boldsymbol{e}_t &= \gamma \lambda \boldsymbol{e}_{t-1} + \boldsymbol{\phi}_t, \\ \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha \delta_t \boldsymbol{e}_t.\end{aligned}$$

By incremental substitution, we can write $\boldsymbol{\theta}_t$ directly in terms of $\boldsymbol{\theta}_0$:

$$\begin{aligned}\boldsymbol{\theta}_t &= \boldsymbol{\theta}_0 + \alpha \sum_{j=0}^{t-1} \delta_j \boldsymbol{e}_j, \\ &= \boldsymbol{\theta}_0 + \alpha \sum_{j=0}^{t-1} \delta_j \sum_{i=0}^j (\gamma \lambda)^{j-i} \boldsymbol{\phi}_i, \\ &= \boldsymbol{\theta}_0 + \alpha \sum_{j=0}^{t-1} \sum_{i=0}^j (\gamma \lambda)^{j-i} \delta_j \boldsymbol{\phi}_i.\end{aligned}$$

Using the summation rule $\sum_{j=k}^n \sum_{i=k}^j a_{i,j} = \sum_{i=k}^n \sum_{j=i}^n a_{i,j}$ we can rewrite this as:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \alpha \sum_{i=0}^{t-1} \sum_{j=i}^{t-1} (\gamma \lambda)^{j-i} \delta_j \boldsymbol{\phi}_i. \quad (20)$$

As part of the derivation shown in Appendix B, we prove the following (see Equation 26):

$$G_i^{\lambda|h+1} = G_i^{\lambda|h} + (\lambda \gamma)^{h-i} \delta_h',$$

with

$$\delta'_h := R_{h+1} + \gamma \boldsymbol{\theta}_h^\top \boldsymbol{\phi}_{h+1} - \boldsymbol{\theta}_{h-1}^\top \boldsymbol{\phi}_h.$$

By applying this sequentially for $i+1 \leq h < t$, we can derive:

$$G_i^{\lambda|t} = G_i^{\lambda|i+1} + \sum_{j=i+1}^{t-1} (\gamma \lambda)^{j-i} \delta'_j. \quad (21)$$

Furthermore, $G_i^{\lambda|i+1}$ can be written as:

$$\begin{aligned} G_i^{\lambda|i+1} &= R_{i+1} + \gamma \boldsymbol{\theta}_i^\top \boldsymbol{\phi}_{i+1}, \\ &= R_{i+1} + \gamma \boldsymbol{\theta}_i^\top \boldsymbol{\phi}_{i+1} - \boldsymbol{\theta}_{i-1}^\top \boldsymbol{\phi}_i + \boldsymbol{\theta}_{i-1}^\top \boldsymbol{\phi}_i, \\ &= \delta'_i + \boldsymbol{\theta}_{i-1}^\top \boldsymbol{\phi}_i. \end{aligned}$$

Substituting this in (21) yields:

$$G_i^{\lambda|t} = \boldsymbol{\theta}_{i-1}^\top \boldsymbol{\phi}_i + \sum_{j=i}^{t-1} (\gamma \lambda)^{j-i} \delta'_j.$$

Using that $\delta'_j = \delta_j + \boldsymbol{\theta}_j^\top \boldsymbol{\phi}_j - \boldsymbol{\theta}_{j-1}^\top \boldsymbol{\phi}_j$, it follows that

$$\sum_{j=i}^{t-1} (\gamma \lambda)^{j-i} \delta_j = G_i^{\lambda|t} - \boldsymbol{\theta}_{i-1}^\top \boldsymbol{\phi}_i - \sum_{j=i}^{t-1} (\gamma \lambda)^{j-i} (\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1})^\top \boldsymbol{\phi}_j.$$

As $\alpha \rightarrow 0$, we can approximate this as:

$$\begin{aligned} \sum_{j=i}^{t-1} (\gamma \lambda)^{j-i} \delta_j &= G_i^{\lambda|t} - \boldsymbol{\theta}_{i-1}^\top \boldsymbol{\phi}_i + \mathcal{O}(\alpha), \\ &= \bar{G}_i^{\lambda|t} - \boldsymbol{\theta}_0^\top \boldsymbol{\phi}_i + \mathcal{O}(\alpha), \end{aligned}$$

with $\bar{G}_i^{\lambda|t}$ the interim λ -return that uses $\boldsymbol{\theta}_0$ for all value evaluations. Substituting this in (20) yields:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \alpha \sum_{i=0}^{t-1} (\bar{G}_i^{\lambda|t} - \boldsymbol{\theta}_0^\top \boldsymbol{\phi}_i + \mathcal{O}(\alpha)) \boldsymbol{\phi}_i. \quad (22)$$

For the online λ -return algorithm, we can derive the following by sequential substitution of Equation (7):

$$\boldsymbol{\theta}_t^t = \boldsymbol{\theta}_0 + \alpha \sum_{i=0}^{t-1} (G_i^{\lambda|t} - (\boldsymbol{\theta}_i^t)^\top \boldsymbol{\phi}_i) \boldsymbol{\phi}_i.$$

As $\alpha \rightarrow 0$, we can approximate this as:

$$\boldsymbol{\theta}_t^t = \boldsymbol{\theta}_0 + \alpha \sum_{i=0}^{t-1} (\bar{G}_i^{\lambda|t} - \boldsymbol{\theta}_0^\top \boldsymbol{\phi}_i + \mathcal{O}(\alpha)) \boldsymbol{\phi}_i. \quad (23)$$

Combining (22) and (23), it follows that as $\alpha \rightarrow 0$:

$$\frac{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^t\|}{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|} = \frac{\|(\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^t)/\alpha\|}{\|(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)/\alpha\|} = \frac{\mathcal{O}(\alpha)}{C + \mathcal{O}(\alpha)},$$

with

$$C = \left\| \sum_{i=0}^{t-1} (\bar{G}_i^{\lambda|t} - \boldsymbol{\theta}_0^\top \boldsymbol{\phi}_i) \boldsymbol{\phi}_i \right\| = \left\| \sum_{i=0}^{t-1} \Delta_i^t \right\|.$$

From the condition $\sum_{i=0}^{t-1} \Delta_i^t \neq \mathbf{0}$ it follows that $C > 0$. ■

Appendix B. Derivation Update Equations

In this subsection, we derive the update equations of true online $\text{TD}(\lambda)$ directly from the online forward view, defined by equations (6) and (7) (and $\boldsymbol{\theta}_0^t := \boldsymbol{\theta}_{init}$). The derivation is based on expressing $\boldsymbol{\theta}_{t+1}^{t+1}$ in terms of $\boldsymbol{\theta}_t^t$.

We start by writing $\boldsymbol{\theta}_t^t$ directly in terms of the initial weight vector and the interim λ -returns. First, we rewrite (7) as:

$$\boldsymbol{\theta}_{k+1}^t = (\mathbf{I} - \alpha \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top) \boldsymbol{\theta}_k^t + \alpha \boldsymbol{\phi}_k G_k^{\lambda|t},$$

with \mathbf{I} the identity matrix. Now, consider $\boldsymbol{\theta}_k^t$ for $k = 1$ and $k = 2$:

$$\begin{aligned}\boldsymbol{\theta}_1^t &= (\mathbf{I} - \alpha \boldsymbol{\phi}_0 \boldsymbol{\phi}_0^\top) \boldsymbol{\theta}_{init} + \alpha \boldsymbol{\phi}_0 G_0^{\lambda|t}, \\ \boldsymbol{\theta}_2^t &= (\mathbf{I} - \alpha \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^\top) \boldsymbol{\theta}_1^t + \alpha \boldsymbol{\phi}_1 G_1^{\lambda|t}, \\ &= (\mathbf{I} - \alpha \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^\top)(\mathbf{I} - \alpha \boldsymbol{\phi}_0 \boldsymbol{\phi}_0^\top) \boldsymbol{\theta}_{init} + \alpha (\mathbf{I} - \alpha \boldsymbol{\phi}_1 \boldsymbol{\phi}_1^\top) \boldsymbol{\phi}_0 G_0^{\lambda|t} + \alpha \boldsymbol{\phi}_1 G_1^{\lambda|t}.\end{aligned}$$

For general $k \leq t$, we can write:

$$\boldsymbol{\theta}_k^t = \mathbf{A}_0^{k-1} \boldsymbol{\theta}_{init} + \alpha \sum_{i=0}^{k-1} \mathbf{A}_{i+1}^{k-1} \boldsymbol{\phi}_i G_i^{\lambda|t},$$

where \mathbf{A}_i^j is defined as:

$$\mathbf{A}_i^j := (\mathbf{I} - \alpha \boldsymbol{\phi}_j \boldsymbol{\phi}_j^\top)(\mathbf{I} - \alpha \boldsymbol{\phi}_{j-1} \boldsymbol{\phi}_{j-1}^\top) \dots (\mathbf{I} - \alpha \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top), \quad \text{for } j \geq i,$$

and $\mathbf{A}_{j+1}^j := \mathbf{I}$. We are now able to express $\boldsymbol{\theta}_t^t$ as:

$$\boldsymbol{\theta}_t^t = \mathbf{A}_0^{t-1} \boldsymbol{\theta}_{init} + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^{t-1} \boldsymbol{\phi}_i G_i^{\lambda|t}, \tag{24}$$

Because for the derivation of true online $\text{TD}(\lambda)$, we only need (24) and the definition of $G_i^{\lambda|t}$, we can drop the double indices for the weight vectors and use $\boldsymbol{\theta}_t := \boldsymbol{\theta}_t^t$.

We now derive a compact expression for the difference $G_i^{\lambda|t+1} - G_i^{\lambda|t}$:

$$\begin{aligned}
 G_i^{\lambda|t+1} - G_i^{\lambda|t} &= (1-\lambda) \sum_{n=1}^{t-i} \lambda^{n-1} G_i^{(n)} + \lambda^{t-i} G_i^{(t+1-i)}, \\
 &\quad - (1-\lambda) \sum_{n=1}^{t-i-1} \lambda^{n-1} G_i^{(n)} - \lambda^{t-i-1} G_i^{(t-i)}, \\
 &= (1-\lambda) \lambda^{t-i-1} G_i^{(t-i)} + \lambda^{t-i} G_i^{(t+1-i)} - \lambda^{t-i-1} G_i^{(t-i)}, \\
 &= \lambda^{t-i} G_i^{(t+1-i)} - \lambda^{t-i} G_i^{(t-i)}, \\
 &= \lambda^{t-i} (G_i^{(t+1-i)} - G_i^{(t-i)}), \\
 &= \lambda^{t-i} \left(\sum_{k=1}^{t+1-i} \gamma^{k-1} R_{i+k} + \gamma^{t+1-i} \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \sum_{k=1}^{t-i} \gamma^{k-1} R_{i+k} - \gamma^{t-i} \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t \right), \\
 &= \lambda^{t-i} \left(\gamma^{t-i} R_{t+1} + \gamma^{t+1-i} \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \gamma^{t-i} \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t \right), \\
 &= (\lambda\gamma)^{t-i} (R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t).
 \end{aligned}$$

Note that the difference $G_i^{\lambda|t+1} - G_i^{\lambda|t}$ is naturally expressed using a term that looks like a TD error but with a modified time step. We call this the modified TD error, δ'_t :

$$\delta'_t := R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t.$$

The modified TD error relates to the regular TD error, δ_t , as follows:

$$\begin{aligned}
 \delta'_t &= R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t, \\
 &= R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t + \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t, \\
 &= \delta_t + \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t.
 \end{aligned} \tag{25}$$

Using δ'_t , the difference $G_i^{\lambda|t+1} - G_i^{\lambda|t}$ can be compactly written as:

$$G_i^{\lambda|t+1} - G_i^{\lambda|t} = (\lambda\gamma)^{t-i} \delta'_t. \tag{26}$$

To get the update rule, $\boldsymbol{\theta}_{t+1}$ has to be expressed in terms of $\boldsymbol{\theta}_t$. This is done below, using (24), (25) and (26):

$$\begin{aligned}
 \boldsymbol{\theta}_{t+1} &= \mathbf{A}_0^t \boldsymbol{\theta}_0 + \alpha \sum_{i=0}^t \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i G_i^{\lambda|t+1}, \\
 &= \mathbf{A}_0^t \boldsymbol{\theta}_0 + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i G_i^{\lambda|t+1} + \alpha \boldsymbol{\phi}_t G_t^{\lambda|t+1}, \\
 &= \mathbf{A}_0^t \boldsymbol{\theta}_0 + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i G_i^{\lambda|t} + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (G_i^{\lambda|t+1} - G_i^{\lambda|t}) + \alpha \boldsymbol{\phi}_t G_t^{\lambda|t+1}, \\
 &= (\mathbf{I} - \alpha \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top) \left(\mathbf{A}_0^{t-1} \boldsymbol{\theta}_0 + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^{t-1} \boldsymbol{\phi}_i G_i^{\lambda|t} \right) \\
 &\quad + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (G_i^{\lambda|t+1} - G_i^{\lambda|t}) + \alpha \boldsymbol{\phi}_t G_t^{\lambda|t+1}, \\
 &= (\mathbf{I} - \alpha \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top) \boldsymbol{\theta}_t + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (G_i^{\lambda|t+1} - G_i^{\lambda|t}) + \alpha \boldsymbol{\phi}_t G_t^{\lambda|t+1}, \\
 &= (\mathbf{I} - \alpha \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top) \boldsymbol{\theta}_t + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (\gamma \lambda)^{t-i} \delta'_t + \alpha \boldsymbol{\phi}_t (R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1}), \\
 &= \boldsymbol{\theta}_t + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (\gamma \lambda)^{t-i} \delta'_t + \alpha \boldsymbol{\phi}_t (R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t), \\
 &= \boldsymbol{\theta}_t + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (\gamma \lambda)^{t-i} \delta'_t \\
 &\quad + \alpha \boldsymbol{\phi}_t (R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t + \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t), \\
 &= \boldsymbol{\theta}_t + \alpha \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (\gamma \lambda)^{t-i} \delta'_t + \alpha \boldsymbol{\phi}_t \delta'_t - \alpha (\boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t) \boldsymbol{\phi}_t, \\
 &= \boldsymbol{\theta}_t + \alpha \sum_{i=0}^t \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (\gamma \lambda)^{t-i} \delta'_t - \alpha (\boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t) \boldsymbol{\phi}_t, \\
 &= \boldsymbol{\theta}_t + \alpha \mathbf{e}_t \delta'_t - \alpha (\boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t) \boldsymbol{\phi}_t, \quad \text{with } \mathbf{e}_t := \sum_{i=0}^t \mathbf{A}_{i+1}^t \boldsymbol{\phi}_i (\gamma \lambda)^{t-i}, \\
 &= \boldsymbol{\theta}_t + \alpha \mathbf{e}_t (\delta_t + \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t) - \alpha (\boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t) \boldsymbol{\phi}_t, \\
 &= \boldsymbol{\theta}_t + \alpha \mathbf{e}_t \delta_t + \alpha (\boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t - \boldsymbol{\theta}_{t-1}^\top \boldsymbol{\phi}_t) (\mathbf{e}_t - \boldsymbol{\phi}_t). \tag{27}
 \end{aligned}$$

We now have the update rule for θ_t , in addition to an explicit definition of e_t . Next, using this explicit definition, we derive an update rule to compute e_t from e_{t-1} :

$$\begin{aligned}
e_t &= \sum_{i=0}^t \mathbf{A}_{i+1}^t \phi_i (\gamma \lambda)^{t-i}, \\
&= \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^t \phi_i (\gamma \lambda)^{t-i} + \phi_t, \\
&= (\mathbf{I} - \alpha \phi_t \phi_t^\top) \gamma \lambda \sum_{i=0}^{t-1} \mathbf{A}_{i+1}^{t-1} \phi_i (\gamma \lambda)^{t-i-1} + \phi_t, \\
&= (\mathbf{I} - \alpha \phi_t \phi_t^\top) \gamma \lambda e_{t-1} + \phi_t, \\
&= \gamma \lambda e_{t-1} + \phi_t - \alpha \gamma \lambda (e_{t-1}^\top \phi_t) \phi_t. \tag{28}
\end{aligned}$$

Equations (27) and (28), together with the definition of δ_t , form the true online TD(λ) update equations.

Appendix C. Detailed Results Random MRPs

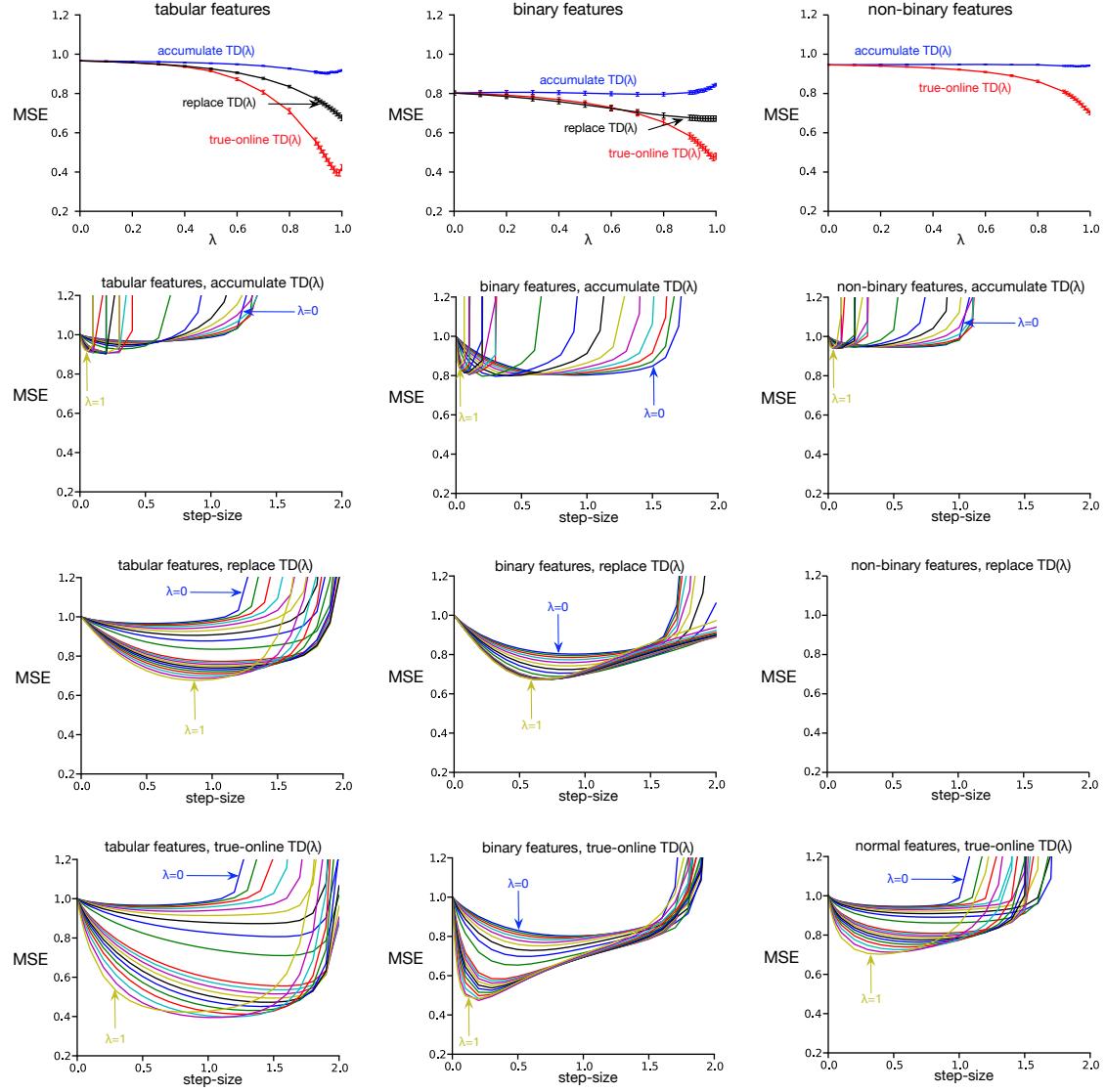


Figure 10: Results on a random MRP with $k = 10$, $b = 3$ and $\sigma = 0.1$. MSE is the mean squared error averaged over the first 100 time steps, as well as 50 runs, and normalized using the initial error. The top graphs summarize the results from the graphs below it; they show the MSE error, for each λ , at the best step-size.

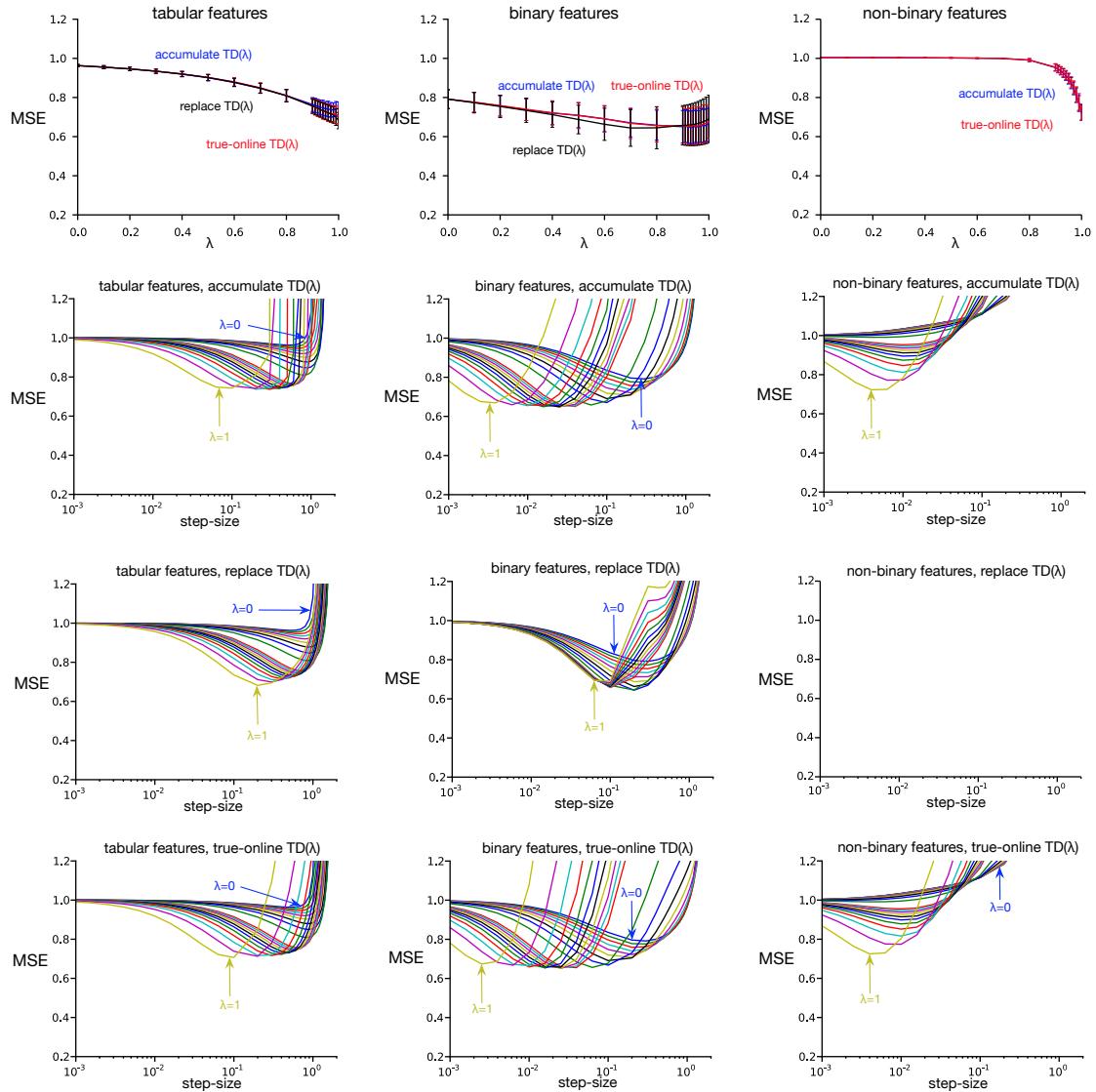


Figure 11: Results on a random MRP with $k = 100$, $b = 10$ and $\sigma = 0.1$. MSE is the mean squared error averaged over the first 1000 time steps, as well as 50 runs, and normalized using the initial error.

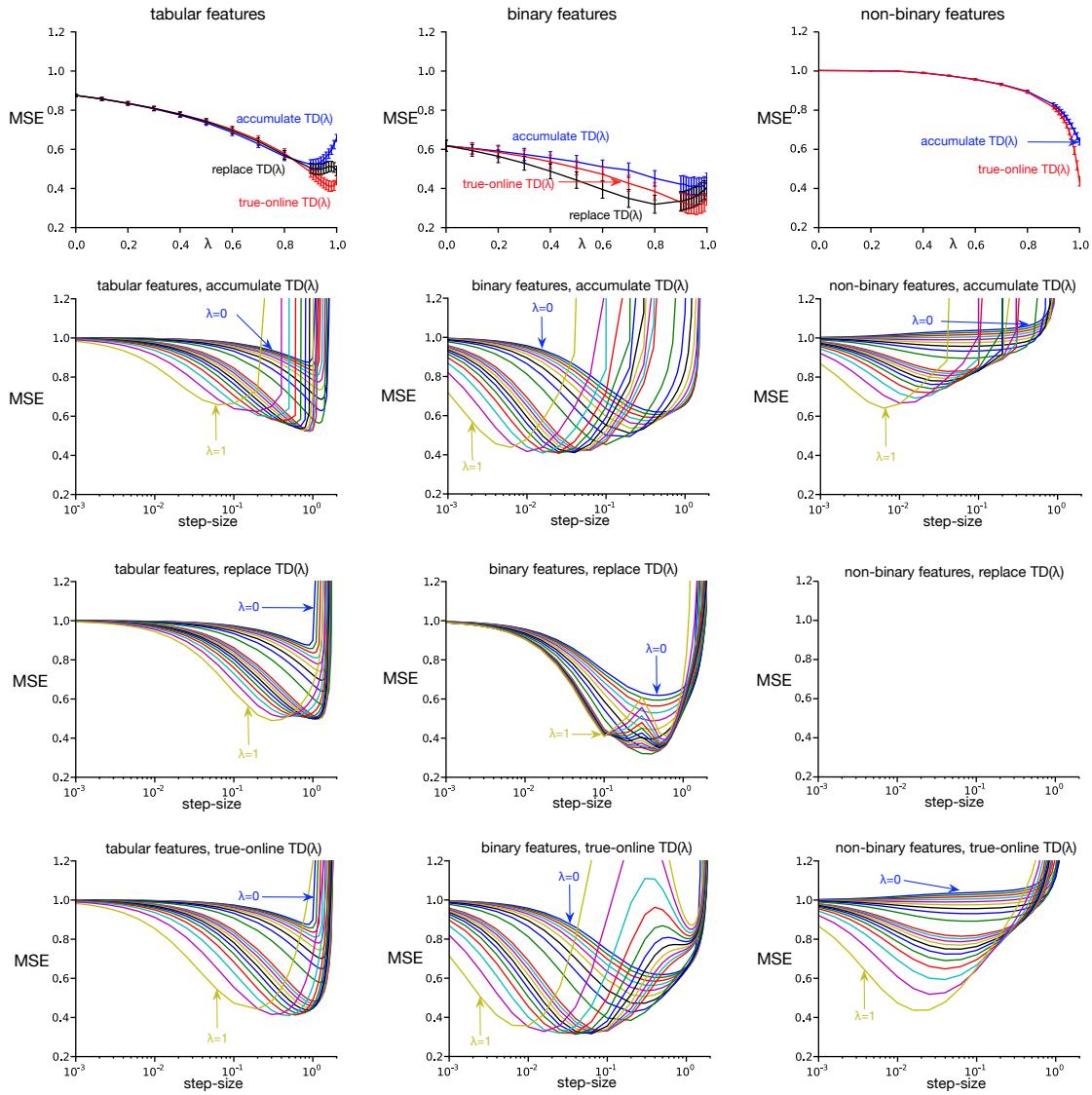


Figure 12: Results on a random MRP with $k = 100$, $b = 3$ and $\sigma = 0$. MSE is the mean squared error averaged over the first 1000 time steps, as well as 50 runs, and normalized using the initial error.

Appendix D. Detailed Results for Myoelectric Prosthetic Arm

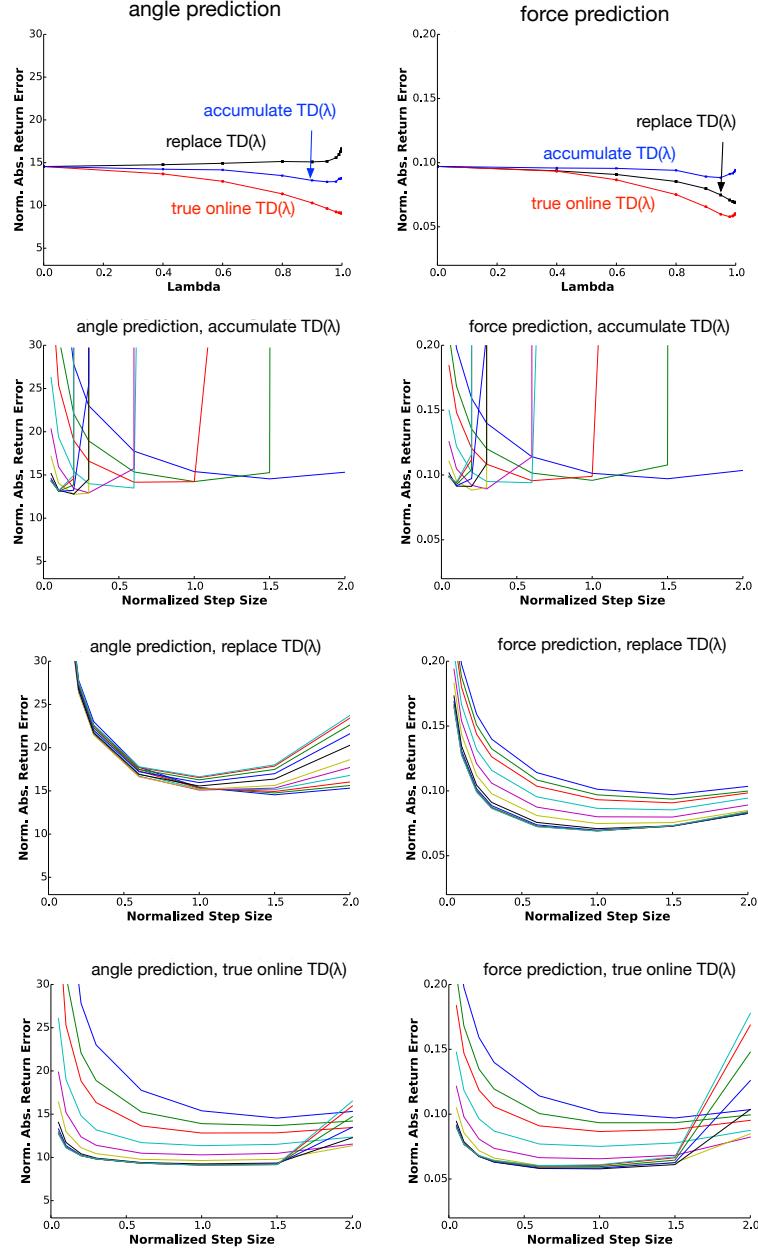


Figure 13: Results on prosthetic data from the single amputee subject described in Pilarski et al. (2013), for the prediction of servo motor angle (*left column*) and grip force (*right column*) as recorded from the amputee's myoelectrically controlled robot arm during a grasping task.

References

- Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279.
- Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.
- Dayan, P. (1992). The convergence of TD(λ) for general λ . *Machine Learning*, 8(3):341–362.
- Defazio, A. and Graepel, T. (2014). A comparison of learning algorithms on the arcade learning environment. *arXiv:1410.8620*.
- Hebert, J. S., Olson, J. L., Morhart, M. J., Dawson, M. R., Marasco, P. D., Kuiken, T. A., and Chan, K. M. (2014). Novel targeted sensory reinnervation technique to restore functional hand sensation after transhumeral amputation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(4):763–773.
- Kaelbling, L. P., Littman, M. L., and Moore, A. P. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285.
- Konidaris, G., Niekum, S., and Thomas, P. S. (2011). TD $_{\gamma}$: Re-evaluating complex backups in temporal difference learning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 2402–2410.
- Maei, H. R. (2011). *Gradient Temporal-Difference Learning Algorithms*. PhD thesis, University of Alberta, Canada.
- Mahmood, A. R. and Sutton, R. S. (2015). Off-policy learning based on weighted importance sampling with linear computational complexity. In *Proceedings of the 31th Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., Kumaran, H. K. D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518:529–533.
- Modayil, J., White, A., and Sutton, R. S. (2014). Multi-timescale nexting in a reinforcement learning robot. *Adaptive Behavior*, 22(2):146–160.
- Parker, P., Englehart, K. B., , and Hudgins, B. (2006). Myoelectric signal processing for control of powered limb prostheses. *Journal of Electromyography and Kinesiology*, 16(6):541–548.
- Pilarski, P. M., Dawson, M. R., Degris, T., Carey, J. P., Chan, K. M., Hebert, J. S., and Sutton, R. S. (2013). Adaptive artificial limbs: A real-time approach to prediction and anticipation. *IEEE Robotics & Automation Magazine*, 20(1):53–64.
- Schapire, R. E. and Warmuth, M. K. (1996). On the worst-case analysis of temporal-difference learning algorithms. *Machine Learning*, 22(1/2/3):95–121.

- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44.
- Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge.
- Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. (2009a). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 993–1000.
- Sutton, R. S., Maei, H. R., and Szepesvári, C. (2009b). A convergent $\mathcal{O}(n)$ algorithm for off-policy temporal-difference learning with linear function approximation. In *Proceedings of Advances in Neural Information Processing Systems 21 (NIPS)*, pages 1609–1616.
- Sutton, R. S., Mahmood, A. R., Precup, D., and van Hasselt, H. (2014). A new $Q(\lambda)$ with interim forward view and Monte Carlo equivalence. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*.
- Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., and Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 761–768.
- Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*. Morgan and Claypool.
- Thomas, P. S., Niekum, S., Theocharous, G., and Konidaris, G. (2015). Policy evaluation using the Ω -return. In *Proceedings of Advances in Neural Information Processing Systems 28 (NIPS)*, pages 334–342.
- van Hasselt, H., Mahmood, A. R. and Sutton, R. S. (2014). Off-policy $TD(\lambda)$ with a true online equivalence. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*.
- van Hasselt, H. and Sutton, R. S. (2015). Learning to predict independent of span. *arXiv:1508.04582*.
- van Seijen, H. H. (2016). Effective multi-step temporal-difference learning for non-linear function approximation. *arXiv:1608.05151*.
- van Seijen, H. H. and Sutton, R. S. (2014). True online $TD(\lambda)$. In *Proceedings of the 31th International Conference on Machine learning (ICML)*.
- Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, England.