Dual Temporal Difference Learning

Min Yang
Dept. of Computing Science
University of Alberta
Edmonton, Alberta
Canada T6G 2E8

Yuxi Li
Dept. of Computing Science
University of Alberta
Edmonton, Alberta
Canada T6G 2E8

Dale Schuurmans
Dept. of Computing Science
University of Alberta
Edmonton, Alberta
Canada T6G 2E8

Abstract

Recently, researchers have investigated novel dual representations as a basis for dynamic programming and reinforcement learning algorithms. Although the convergence properties of classical dynamic programming algorithms have been established for dual representations, temporal difference learning algorithms have not yet been analyzed. In this paper, we study the convergence properties of temporal difference learning using dual representations. We make significant progress by proving the convergence of dual temporal difference learning with eligibility traces. Experimental results suggest that the dual algorithms seem to demonstrate empirical benefits over standard primal algorithms.

1 Introduction

Algorithms for dynamic programming (DP) and reinforcement learning (RL) are usually formulated with respect to value functions for states or state-action pairs (Bertsekas and Tsitsiklis 1996; Sutton and Barto 1998). However, linear programming approaches demonstrate that the value function representation is not an indispensable component for solving DP and RL problems. Instead, investigations into the dual representation show that the notion of state or state-action visit distribution can replace the concept of value functions (Wang et al. 2007, 2008). In particular, the dual representation provides an equivalent but distinct approach to solving DP and RL problems (Wang et al. 2007, 2008). It is known that there exist dual forms for standard DP and RL algorithms, including policy evaluation, policy improvement and off-policy control. It is also known from previous work that dual DP algorithms possess advantageous convergence properties over their primal counterparts: For off-policy control with function approximation, where the gradient-based updates in the primal form often diverge, the dual DP algorithms empirically remain stable. However, no previous work has been done on the dual form temporal difference (TD) learning.

In this paper, we contribute progress to research on dual representations for DP and RL algorithms by presenting new theoretical results for the dual representation of reinforcement learning algorithms; in particular, temporal difference learning with eligibility traces. We also show the convergence property empirically.

The remainder of this paper is organized as follows. First we provide some brief background on existing TD algorithms, both in the primal and dual representations, in Section 2. After presenting the dual form of TD algorithm with eligibility traces in Section 3, and covering necessary preliminaries, we show the convergence property theoretically in Section 4. We then present an empirical study of convergence in Section 5 before concluding.

2 Background

Reinforcement learning is an approach to finding an optimal policy for a sequential decision making problem when one only has access to the environment by choosing actions and observing state transitions and rewards (Bertsekas and Tsitsiklis 1996; Puterman 1994; Sutton and Barto 1998). We consider a Markov decision process (MDP), defined by $(S, A, P, \mathbf{r}, \gamma)$, where S is a finite or countably infinite set of states; A is a set of actions; P is a transition $|S||A| \times |S|$ matrix, whose entry $P_{(sa,s')}$ specifies the conditional probability of transitioning to state s' starting from state s and taking action a; $\mathbf{r}$ is a bounded $|S||A| \times 1$ reward vector,

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

whose entry $\mathbf{r}_{(sa)}$ specifies the reward obtained when taking action a in state s; and $0 < \gamma < 1$ is a discount factor.[1] A policy $\boldsymbol{\pi}$ is represented by an $|S||A| \times 1$ vector, whose entry $\boldsymbol{\pi}_{(sa)}$ specifies the probability of taking action a in state s, that is, $\sum_a \boldsymbol{\pi}_{(sa)} = 1$. It is convenient to represent a policy as an $|S| \times |S||A|$ matrix Π, where $\Pi_{s,s'a} = \boldsymbol{\pi}_{(sa)}$ if $s' = s$, otherwise 0. Note the same definition for Π is also used in Lagoudakis and Parr (2003). The $|S| \times |S|$ matrix ΠP gives the state to state transition probabilities induced by $\boldsymbol{\pi}$. Let $\{\alpha_t\}_{t=0}^{\infty}$ be any positive and non-increasing step-size sequence such that $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$. We assume the MDP induces an irreducible and aperiodic Markov chain.

2.1 Primal TD(0)

TD learning is a traditional approach to policy evaluation (Sutton and Barto 1998) that has been proved to converge to an exact representation of the value function in the tabular case and to a bounded error in the linear function approximation case (Tsitsiklis and Van Roy 1997). To express TD algorithms in the primal representation, recall that the state value function can be specified by an $|S| \times 1$ vector

$$\mathbf{v} = \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r}$$

which satisfies

$$\mathbf{v} = \Pi \mathbf{r} + \gamma \Pi P \mathbf{v}.$$

The first temporal difference algorithm, TD(0), is given by the simple update rule in the tabular case

$$\mathbf{v}(s_t) \quad = \quad \mathbf{v}(s_t) + \alpha_t [r_t + \gamma \mathbf{v}(s_{t+1}) - \mathbf{v}(s_t)] \quad (1)$$

where r_t is the observed reward at time t.

This algorithm can be extended to cope with large state spaces by introducing a function approximator to take the place of an exact representation of $\mathbf{v}$. In linear function approximation, one defines $\phi^{\mathsf{T}}(s) = \{\phi_1(s), \cdots, \phi_k(s)\}$, where k is the number of basis functions. The approximate value function is then given by $\hat{\mathbf{v}}(s) = \boldsymbol{\phi}(s)^{\mathsf{T}} \mathbf{w}$, where $\mathbf{w}$ is the weight vector. The update procedure for approximate TD(0) is then

$$\begin{aligned} \mathbf{w}_{t+1} \quad = \quad & \mathbf{w}_t + \alpha_t \phi(s_t)[r_t + \gamma \phi^{\mathsf{T}}(s_{t+1}) \mathbf{w}_t \\ & - \phi^{\mathsf{T}}(s_t) \mathbf{w}_t] \end{aligned} \quad (2)$$

[1] The analysis in this paper can be extended to the continuous case, which requires the transition probability matrices to be replaced by probability transition kernels.

2.2 Dual TD(0)

Traditionally, the primal value function plays an essential role in DP and RL algorithms. However, it is demonstrated in Wang et al. (2007, 2008) that the classical DP and RL algorithms, namely, policy evaluation, policy improvement, Q(0), Sarsa(0), have natural duals expressed with state or state-action distributions.

To develop a dual form of state policy evaluation, one considers the $|S| \times |S|$ matrix

$$M = (1 - \gamma) \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i$$

which satisfies the linear system

$$M = (1 - \gamma) I + \gamma \Pi P M$$

Each row of M is a probability distribution and the entries $M_{(s,s')}$ correspond to the probability of discounted state visits to s' for a policy $\boldsymbol{\pi}$ starting in state s. We have $(1 - \gamma)\mathbf{v} = M \Pi \mathbf{r}$. That is, given M we can easily recover the state values of $\boldsymbol{\pi}$.

In the tabular dual representation, the TD(0) update rule can be expressed by

$$\begin{aligned} M(s_t, :) \quad = \quad & M(s_t, :) + \alpha [(1 - \gamma) \mathbf{e}_{s_t}^{\mathsf{T}} \\ & + \gamma M(s_{t+1}, :) - M(s_t, :)] \end{aligned} \quad (3)$$

where $\mathbf{e}_{s_t}$ is a column vector of all zeros except for a 1 for the s_t^{th} entry (Wang et al. 2008).

Analogous to the primal case, linear function approximation can be formulated in the dual representation as follows. Let $\boldsymbol{\Upsilon} = (\Upsilon^{(1)}, \ldots, \Upsilon^{(k)})$ be a set of k basis matrices such that each $\Upsilon^{(i)}$ is an $|S| \times |S|$ matrix satisfying the constraints $\Upsilon^{(i)} \mathbf{1} = \mathbf{1}$ and $\Upsilon^{(i)} \geq 0$. Clearly, these are bounded functions. Assume furthermore that these basis matrices are linearly independent; that is, no one of them can be expressed as a linear combination of the others. Define Ψ as an $|S|^2 \times k$ matrix of basis distributions, so that $\Psi(:, i) = \mathrm{vec}(\Upsilon^{(i)})$. Define the operator *reshape* to be the inverse of vec, in that it converts the vector $\Psi \mathbf{w}$, whose dimension is $|S|^2 \times 1$, back into an $|S| \times |S|$ matrix. Then a linear approximation in the dual representation, $\hat{M}$, can be expressed as:

$$\hat{M} = \sum_{i=1}^{k} w_i \Upsilon^{(i)} = reshape(\Psi \mathbf{w}),$$

where $\mathbf{w} \geq 0, \mathbf{w}^{\mathsf{T}} \mathbf{1} = 1$. Note that by these definitions it follows that $\hat{M}$ is a nonnegative row normalized matrix; such that $\hat{M} \geq 0$, and $\hat{M} \mathbf{1} = 1$.

Given this form of linear approximation, the TD(0) algorithm can be expressed by

$$\begin{aligned}\mathbf{w}_{t+1} &= \mathbf{w}_t + \alpha[(1-\gamma)r + \gamma\Gamma(s_{t+1},:)\mathbf{w}_t \\ &\quad -\Gamma(s_t,:)\mathbf{w}_t]\Gamma^\mathsf{T}(s_t,:) \end{aligned} \quad (4)$$

where $\Gamma = ((\Pi\mathbf{r})^\mathsf{T} \otimes I)\Psi$ is an $|S| \times k$ matrix and $\otimes$ is the Kronecker product. To better understand the Γ object, which will be required below, note that by the above definition we have

$$\begin{aligned}\Gamma &= ((\Pi\mathbf{r})^\mathsf{T} \otimes I)[\Psi_{(:,1)}, \cdots, \Psi_{(:,k)}] \\ &= [((\Pi\mathbf{r})^\mathsf{T} \otimes I)\Psi_{(:,1)}, \cdots, ((\Pi\mathbf{r})^\mathsf{T} \otimes I)\Psi_{(:,k)}] \\ &= [((\Pi\mathbf{r})^\mathsf{T} \otimes I)\text{vec}(\Upsilon^{(1)}), \\ &\quad \cdots, ((\Pi\mathbf{r})^\mathsf{T} \otimes I)\text{vec}(\Upsilon^{(k)})] \\ &= [\Upsilon^{(1)}(\Pi\mathbf{r}), \cdots, \Upsilon^{(k)}(\Pi\mathbf{r})]\end{aligned}$$

Thus Γ has a reasonably intuitive interpretation. Each column of Γ is associated with a single basis matrix $\Upsilon^{(k)}$. The column is a scaled version of the value function one would obtain if $\Upsilon^{(k)}$ were the correct M matrix for the policy and the transition dynamics for the domain. That is, we have the relationship $(1-\gamma)v = M\Pi r$.

3 Dual TD(λ)

In this paper, we introduce the dual form of TD(λ) for general λ and analyze its convergence properties. Our analysis below will address both the tabular and function approximation cases.

Recall that in the primal representation with linear function approximation, TD(λ), $\lambda \in [0,1]$, can be expressed by the update rule

$$\begin{aligned} d_t &= r + \gamma\phi^\mathsf{T}(s_{t+1})\mathbf{w}_t - \phi^\mathsf{T}(s_t)\mathbf{w}_t \\ z_t &= \gamma\lambda z_{t-1} + \phi(s_t) \\ \mathbf{w}_{t+1} &= \mathbf{w}_t + \alpha_t d_t z_t \end{aligned} \quad (5)$$

where d_t is the temporal difference, z_t is the eligibility trace (Sutton and Barto 1998; Tsitsiklis and Van Roy 1997).

Our first contribution is to introduce a dual version of TD(λ), in the linear function approximation case, and show how it arises from the combination of a projection and temporal difference operator.

Given the dual linear approximation representation introduced in the previous section, an analogous TD(λ) update to the primal case can be expressed as

$$\begin{aligned} d_t &= (1-\gamma)r + \gamma\Gamma(s_{t+1},:)\mathbf{w}_t - \Gamma(s_t,:)\mathbf{w}_t \\ z_t &= \gamma\lambda z_t + \Gamma^\mathsf{T}(s_t,:) \\ \mathbf{w}_{t+1} &= \mathbf{w}_t + \alpha_t d_t z_t \end{aligned} \quad (6)$$

This algorithm arises from the combination of a temporal difference operator and a projection operator as follows. First define the dual $TD(\lambda)$ operator as

$$\begin{aligned} T^{(\lambda)}M = (1-\lambda)\sum_{m=0}^{\infty}\lambda^m \\ \left[(1-\gamma)\sum_{t=0}^{m}\gamma^t(\Pi P)^t + (\gamma\Pi P)^{m+1}M\right] \end{aligned} \quad (7)$$

$$T^{(1)}M = (1-\gamma)\sum_{i=0}^{\infty}\gamma^i(\Pi P)^i = M^* \quad (8)$$

(We will show that the dual $TD(\lambda)$ operator is a contraction in Lemma 4 below.)

Second, to model the effect of linear approximation we will need to introduce a linear projection operator. Let D be the diagonal matrix with entries from the steady state distribution $\boldsymbol{\mu}$. Define $||\cdot||_D = \sqrt{\langle\cdot,\cdot\rangle_D}$ as the norm on the inner product space, $\langle x, y\rangle_D = x^\mathsf{T}Dy$. Define the pseudo-norm $||M||^2_{D,\Pi\mathbf{r}} = ||M\Pi\mathbf{r}||^2_D = (M\Pi\mathbf{r})^\mathsf{T}D(M\Pi\mathbf{r})$. Define $L_2(S,D,\Pi\mathbf{r})$ as the set of matrices $\{M \in \mathcal{R}^{|S|\times|S|} \,|\, ||M||_{D,\Pi\mathbf{r}} < \infty\}$.

We define the projection operator $\mathcal{P}$ with respect to $||\cdot||_{D,\Pi\mathbf{r}}$ as follows.

$$\mathcal{P}M = \underset{\hat{M}\in\text{col-span}(\Psi)}{\text{argmin}} \left|\left|M - \hat{M}\right|\right|^2_{D,\Pi\mathbf{r}} \quad (9)$$

where M is the true state visit distribution and $\hat{M}$ is an approximation for it. The above equation can be rewritten as

$$\begin{aligned} \mathcal{P}M &= \underset{\hat{M}\in\text{col-span}(\Psi)}{\text{argmin}} \left|\left|M - \hat{M}\right|\right|^2_{D,\Pi\mathbf{r}} \\ \text{subject to } \hat{M} &= \textit{reshape}(\Psi\mathbf{w}) \text{ for some } \mathbf{w} \end{aligned} \quad (10)$$

Manipulating the objective function J_M, we have,

$$\begin{aligned} J_M &= \left|\left|M - \hat{M}\right|\right|^2_{D,\Pi\mathbf{r}} \\ &= \left|\left|M\Pi\mathbf{r} - \hat{M}\Pi\mathbf{r}\right|\right|^2_D \\ &= \left|\left|\text{vec}(M\Pi\mathbf{r}) - \text{vec}(\hat{M}\Pi\mathbf{r})\right|\right|^2_D \\ &= \left|\left|((\Pi\mathbf{r})^\mathsf{T} \otimes I)(\text{vec}(M) - \text{vec}(\hat{M}))\right|\right|^2_D \\ &= ||M\Pi\mathbf{r} - \Gamma\mathbf{w}||^2_D \end{aligned} \quad (11)$$

The optimization problem (10) can be expressed with respect to $\mathbf{w}$ as

$$\begin{aligned} \mathcal{P}M &= \textit{reshape}(\Psi\mathbf{w}^*) \\ \text{subject to } \mathbf{w}^* &= \underset{\mathbf{w}}{\text{argmin}}\, ||M\Pi\mathbf{r} - \Gamma\mathbf{w}||^2_D \end{aligned} \quad (12)$$

We can obtain the gradient,

$$\nabla_{\mathbf{w}} J_M = -2\Gamma^\mathsf{T} D(M\Pi\mathbf{r} - \Gamma\mathbf{w}) \quad (13)$$

Set $\nabla_{\mathbf{w}} J_M = -2\Gamma^\mathsf{T} D(M\Pi\mathbf{r} - \Gamma\mathbf{w}) = 0$, then $\mathbf{w} = (\Gamma^\mathsf{T} D\Gamma)^{-1}\Gamma^\mathsf{T} DM\Pi\mathbf{r}$. With $\hat{M}\Pi\mathbf{r} = \Gamma\mathbf{w}$ and $\hat{M} = \mathcal{P}M$, we have $\mathcal{P}M\Pi\mathbf{r} = \Gamma\mathbf{w}$. Thus, we obtain the projection matrix $\mathcal{P}$ as:

$$\mathcal{P} = \Gamma(\Gamma^\mathsf{T} D\Gamma)^{-1}\Gamma^\mathsf{T} D. \quad (14)$$

To make $\hat{M}$ a basis distribution, we need to further confine it in Simplex(Ψ), by two more projections: $\mathcal{P}_{\mathbf{1}}$, onto the subspace of row normalized matrices span(Ψ) $\cap \{M|M\mathbf{1} = 1\}$, and $\mathcal{P}_+$, onto the subspace of non-negative matrices span(Ψ) $\cap \{M|M\mathbf{1} = 1$ and $M \geq 0\}$.[2]

Now we make the connection between the stochastic gradient update (6) and gradient operator (13). To achieve this, construct a Markov process $X_t = (s_t, s_{t+1}, z_t)$ and define a function u

$$u(\mathbf{w}, X_t) = ((1-\gamma)r_t + \gamma\Gamma(s_{t+1}, :)\mathbf{w} - \Gamma(s_t, :)\mathbf{w})\, z_t$$

Then the dual TD(λ) update becomes:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t u(\mathbf{w}_t, X_t).$$

As we will show later, for any $\mathbf{w}$, $u(\mathbf{w}, X_t)$ has a well-defined steady-state expectation denoted by $E_0[u(\mathbf{w}, X_t)]$, and as will be shown in Lemma 8,

$$E_0[u(\mathbf{w}, X_t)] = \Gamma^\mathsf{T} D(T^{(\lambda)}(\Gamma\mathbf{w}) - \Gamma\mathbf{w})$$

where, with a slight abuse of the notation, $T^{(\lambda)}(\Gamma\mathbf{w})$ denotes $T^{(\lambda)}(\hat{M})\Pi\mathbf{r}$. Thus, as one can see

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{1}{2}\alpha_t \nabla_{\mathbf{w}} J_M$$

is a steepest descent iteration in solving the problem of minimizing $\left|\left|M - \hat{M}\right|\right|^2_{D,\Pi\mathbf{r}}$.

Note that the update (6) requires only $O(k)$ computation if access to Γ is available. If Γ is not explicitly available then, given access to a reward evaluator for computing $r(s,a)$ the required entries in Γ can be efficiently estimated by sampling random states and actions according to Υ and Π, and evaluating $r(s,a)$ pointwise. To keep this paper simple, we assume that the reward function r is known. If r is not known, it is possible to efficiently estimate entries in Γ by sampling. However we did not run any experiments to test this.

In principle, the dual policy evaluation algorithm could be combined with policy improvement steps to obtain a version of dual-Sarsa and dual-Q-learning, as suggested in the existing publication (Wang et al. 2007). But we have not yet analyzed these algorithms.

4 Theoretical results

In this section, we study the convergence property of the dual TD(λ). We present a series of lemmas which underly the theorem for convergence.

Lemma 1 *For any* $M \in L_2(S, D, \Pi\mathbf{r})$, $||\Pi PM||_{D,\Pi\mathbf{r}} \leq ||M||_{D,\Pi\mathbf{r}}$

Proof: The proof involves Jensen's inequality, the Tonelli-Fubini theorem for the interchange of summations and the property that $\boldsymbol{\mu}$ is the steady state distribution.

$$\begin{aligned}
||\Pi PM||^2_{D,\Pi\mathbf{r}} &= (\Pi PM\Pi\mathbf{r})^\mathsf{T} D(\Pi PM\Pi\mathbf{r}) \\
&= \sum_{i=1}^{|S|} \boldsymbol{\mu}(i) \left[\sum_{j=1}^{|S|} (\Pi P)(i,j)(M\Pi\mathbf{r})(j)\right]^2 \\
&\leq \sum_{i=1}^{|S|} \boldsymbol{\mu}(i) \sum_{j=1}^{|S|} (\Pi P)(i,j) \left[(M\Pi\mathbf{r})(j)\right]^2 \\
&= \sum_{j=1}^{|S|} \sum_{i=1}^{|S|} \boldsymbol{\mu}(i)(\Pi P)(i,j) \left[(M\Pi\mathbf{r})(j)\right]^2 \\
&= \sum_{j=1}^{|S|} \boldsymbol{\mu}(j) \left[(M\Pi\mathbf{r})(j)\right]^2 \\
&= (M\Pi\mathbf{r})^\mathsf{T} D(M\Pi\mathbf{r}) \\
&= ||M||^2_{D,\Pi\mathbf{r}} \qquad \blacksquare
\end{aligned}$$

Lemma 2 *M^* as defined in (8) is in $L_2(S, D, \Pi\mathbf{r})$.*

Proof: We use Jensen's inequality for the first inequality and Lemma 1 for the second:

$$\begin{aligned}
||M^*||_{D,\Pi\mathbf{r}} &= \left|\left|(1-\gamma)\sum_{i=0}^{\infty} \gamma^i (\Pi P)^i\right|\right|_{D,\Pi\mathbf{r}} \\
&\leq (1-\gamma)\sum_{i=0}^{\infty} \gamma^i \left|\left|(\Pi P)^i\right|\right|_{D,\Pi\mathbf{r}} \\
&\leq (1-\gamma)\sum_{i=0}^{\infty} \gamma^i ||I||_{D,\Pi\mathbf{r}} = ||I||_{D,\Pi\mathbf{r}} \qquad \blacksquare
\end{aligned}$$

[2] We may apply the operators $\mathcal{P}_{\mathbf{1}}$ and $\mathcal{P}_+$ to $\mathbf{w}$ in (6) and (4) to make it normalized and non-negative at every step, at the end of every N steps or at the end of the experiment. Our analysis focuses on the case the two operators are applied at the last step of the experiment. We believe it holds for the other cases.

Lemma 3 *For any* $M \in L_2(S, D, \Pi\mathbf{r}), \lambda \in [0,1], T^{(\lambda)}M \in L_2(S, D, \Pi\mathbf{r})$.

Proof: Lemma 2 proves the case $\lambda = 1$. Now we prove the case $\lambda \in [0, 1)$.

$$\begin{aligned}
& T^{(\lambda)}M \\
= \; & (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\Big[(1-\gamma)\sum_{t=0}^{m}\gamma^t(\Pi P)^t \\
& +(\gamma\Pi P)^{m+1}M\Big] \\
\leq \; & \Big|\Big|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(1-\gamma)\sum_{t=0}^{m}\gamma^t(\Pi P)^t\Big|\Big|_{D,\Pi\mathbf{r}} \\
& +\Big|\Big|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\gamma\Pi P)^{m+1}M\Big|\Big|_{D,\Pi\mathbf{r}} \\
\leq \; & (1-\lambda)\sum_{m=0}^{\infty}\lambda^m(1-\gamma)\sum_{t=0}^{m}\gamma^t\left|\left|(\Pi P)^t\right|\right|_{D,\Pi\mathbf{r}} \\
& +(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\left|\left|M\right|\right|_{D,\Pi\mathbf{r}} \\
\leq \; & (1-\lambda)\sum_{m=0}^{\infty}\lambda^m(1-\gamma)\sum_{t=0}^{m}\gamma^t\left|\left|I\right|\right|_{D,\Pi\mathbf{r}} \\
& +\gamma\frac{1-\lambda}{1-\gamma\lambda}\left|\left|M\right|\right|_{D,\Pi\mathbf{r}} < \infty
\end{aligned}$$

The next lemma will be useful for establishing the error bound.

Lemma 4 *For any* $M, \overline{M} \in L_2(S, D, \Pi\mathbf{r})$, *and* $\lambda \in [0,1]$, *we have*

$$\begin{aligned}
& \left|\left|T^{(\lambda)}M - T^{(\lambda)}\overline{M}\right|\right|_{D,\Pi\mathbf{r}} \\
\leq \; & \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\left|\left|M-\overline{M}\right|\right|_{D,\Pi\mathbf{r}} \leq \gamma\left|\left|M-\overline{M}\right|\right|_{D,\Pi\mathbf{r}}
\end{aligned}$$

Proof: The case of $\lambda = 1$ is trivial. The result for $\lambda \in [0, 1)$ follows from Lemma 1:

$$\begin{aligned}
& \left|\left|T^{(\lambda)}M - T^{(\lambda)}\overline{M}\right|\right|_{D,\Pi\mathbf{r}} \\
= \; & \left|\left|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\gamma\Pi P)^{m+1}(M-\overline{M})\right|\right|_{D,\Pi\mathbf{r}} \\
\leq \; & (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\left|\left|(M-\overline{M})\right|\right|_{D,\Pi\mathbf{r}} \\
= \; & \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\left|\left|(M-\overline{M})\right|\right|_{D,\Pi\mathbf{r}} \qquad \blacksquare
\end{aligned}$$

Lemma 5 *For* $\lambda \in [0,1]$, M^* *is the fixed point of* $T^{(\lambda)}$, *that is,* $T^{(\lambda)}M^* = M^*$.

Proof: For the case of $\lambda = 1$, the proof follows from the definition of $T^{(1)}$. For $\lambda \in [0, 1)$,

$$\begin{aligned}
& T^{(\lambda)}M^* \\
= \; & (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\Big[(1-\gamma)\sum_{t=0}^{m}\gamma^t(\Pi P)^t \\
& +(\gamma\Pi P)^{m+1}M^*\Big] \\
= \; & (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\Big[(1-\gamma)\sum_{t=0}^{m}\gamma^t(\Pi P)^t \\
& +(\gamma\Pi P)^{m+1}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t(\Pi P)^t\Big] \\
= \; & (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\Big[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t(\Pi P)^t\Big] \\
= \; & (1-\lambda)\sum_{m=0}^{\infty}\lambda^m M^* \\
= \; & M^* \qquad \blacksquare
\end{aligned}$$

The next lemma establishes that the composition $\mathcal{P}T^{(\lambda)}$ is a contraction, and the fixed point, denoted as M_+, must lie in the space $\{M|M = reshape(\Psi\mathbf{w})\}$, which is a subspace of $L_2(S, D, \Pi\mathbf{r})$.

Lemma 6 $\mathcal{P}T^{(\lambda)}$ *is a contractor, and it has a unique fixed point of the form* $M_+ = reshape(\Psi\mathbf{w}^*)$ *for a unique choice of* $\mathbf{w}^*$. *Furthermore,*

$$\left|\left|M_+ - M^*\right|\right|_{D,\Pi\mathbf{r}} \leq \frac{(1-\lambda\gamma)}{1-\gamma}\left|\left|\mathcal{P}M^* - M^*\right|\right|_{D,\Pi\mathbf{r}}$$

Proof: From the fact that the projection $\mathcal{P}$ is non-expansive, and by Lemma 4, $T^{(\lambda)}$ is a contractor, the composition $\mathcal{P}T^{(\lambda)}$ is a contractor. By Lemma 5, $T^{(\lambda)}$ has a unique fixed point of the form $M^* = T^{(\lambda)}M^*$. Now we establish the approximation error bound between M_+ and M^*. Since $M+ = \mathcal{P}T^{(\lambda)}M_+$, $\mathcal{P}$ is non-expansive, $M^* = T^{(\lambda)}M^*$, and from Lemma 4, we have,

$$\begin{aligned}
\left|\left|M_+ - \mathcal{P}M^*\right|\right|_{D,\Pi\mathbf{r}} &= \left|\left|\mathcal{P}T^{(\lambda)}M_+ - \mathcal{P}M^*\right|\right|_{D,\Pi\mathbf{r}} \\
&\leq \left|\left|T^{(\lambda)}M_+ - M^*\right|\right|_{D,\Pi\mathbf{r}} \\
&= \left|\left|T^{(\lambda)}M_+ - T^{(\lambda)}M^*\right|\right|_{D,\Pi\mathbf{r}} \\
&\leq \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\left|\left|M_+ - M^*\right|\right|_{D,\Pi\mathbf{r}}
\end{aligned}$$

With the Pythagorean theorem for the first inequality and Lemma 4 for the second, we have

$$\begin{aligned} & ||M_+ - M^*||_{D,\Pi\mathbf{r}} \\ = \; & ||M_+ - \mathcal{P}M^* + \mathcal{P}M^* - M^*||_{D,\Pi\mathbf{r}} \\ \leq \; & ||M_+ - \mathcal{P}M^*||_{D,\Pi\mathbf{r}} + ||\mathcal{P}M^* - M^*||_{D,\Pi\mathbf{r}} \\ \leq \; & \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\,||M_+ - M^*||_{D,\Pi\mathbf{r}} + ||\mathcal{P}M^* - M^*||_{D,\Pi\mathbf{r}} \end{aligned}$$

The desired result then follows. ∎

Next, we study the expected behavior of the dual TD(λ) algorithm (6) in the steady state. To study a process already in the steady state, we redefine $z_t = \sum_{m=-\infty}^{t}(\gamma\lambda)^{t-m}\Gamma^\mathsf{T}(s_m,:)$. We study the property of $E_0[u(\mathbf{w}, X_t)]$. Recall $X_t = (s_t, s_{t+1}, z_t)$.

Lemma 7 *The following relations hold, and each of them is well defined and finite:*

1) $E_0\left[\Gamma^\mathsf{T}(s_t,:)\Gamma(s_{t+m},:)\right] = \Gamma^\mathsf{T}D(\Pi P)^m\Gamma$
2) $\left|\left|E_0\left[\Gamma^\mathsf{T}(s_t,:)\Gamma(s_{t+m},:)\right]\right|\right| < \infty$
3) $E_0\left[z_t\Gamma(s_t,:)\right] = \sum_{m=0}^{\infty}(\gamma\lambda)^m\Gamma^\mathsf{T}D(\Pi P)^m\Gamma$
4) $E_0\left[z_t\Gamma(s_{t+1},:)\right] = \sum_{m=0}^{\infty}(\gamma\lambda)^m\Gamma^\mathsf{T}D(\Pi P)^{m+1}\Gamma$
5) $E_0\left[z_t r\right] = \sum_{m=0}^{\infty}(\gamma\lambda)^m\Gamma^\mathsf{T}D(\Pi P)^m\Pi\mathbf{r}$

Proof: For any $M, \overline{M} \in L_2(S, D, \Pi\mathbf{r})$, we have

$$\begin{aligned} & E_0\left[(M(s_t,:)\Pi\mathbf{r})(\overline{M}(s_{t+m},:)\Pi\mathbf{r})\right] \\ = \; & \sum_{i\in S}\boldsymbol{\mu}(i)\sum_{j\in S}(\Pi P)(i,j)(M(i,:)\Pi\mathbf{r}\overline{M}(j,:)\Pi\mathbf{r} \\ = \; & \sum_{i\in S}\boldsymbol{\mu}(i)(M(i,:)\Pi\mathbf{r}[(\Pi P)^m\overline{M}](i,:)\Pi\mathbf{r} \\ = \; & (M\Pi\mathbf{r})^\mathsf{T}D(\Pi P)^m(\overline{M}\Pi\mathbf{r}) \end{aligned}$$

Note, $(\Pi P)^m M \in L_2(S, D, \Pi\mathbf{r})$. For any $M\Pi\mathbf{r} = \Gamma\mathbf{w}$ and $\overline{M}\Pi\mathbf{r} = \Gamma\overline{\mathbf{w}}$, we have,

$$\begin{aligned} & E_0\left[(M(s_t,:)\Pi\mathbf{r})(\overline{M}(s_{t+m},:)\Pi\mathbf{r})\right] \\ = \; & E_0\left[\mathbf{w}^\mathsf{T}\Gamma^\mathsf{T}(s_t,:)\Gamma(s_{t+m},:)\overline{\mathbf{w}}\right] \\ = \; & \mathbf{w}^\mathsf{T}\Gamma^\mathsf{T}D(\Pi P)^m\Gamma\overline{\mathbf{w}} \end{aligned}$$

Since $\mathbf{w}$ and $\overline{\mathbf{w}}$ are arbitrary, it follows that,

$$E_0\left[\Gamma^\mathsf{T}(s_t,:)\Gamma(s_{t+m},:)\right] = \Gamma^\mathsf{T}D(\Pi P)^m\Gamma$$

$$\begin{aligned} \left|\left|\Gamma^\mathsf{T}D(\Pi P)^m\Gamma\right|\right| \; & \leq \; k^2\max_{i,j}\left|\Gamma_i^\mathsf{T}D(\Pi P)^m\Gamma_j\right| \\ & = \; k^2\max_{i,j}\left|\Gamma_i^\mathsf{T}D^{\frac{1}{2}}D^{\frac{1}{2}}(\Pi P)^m\Gamma_j\right| \\ & \leq \; k^2\max_{i,j}||\Gamma_i||_D\,||(\Pi P)^m\Gamma_j||_D \\ & \leq \; k^2\max_i||\Gamma_i||_D^2 \\ & = \; k^2\max_i E_0[\Gamma_i^2] < \infty \end{aligned}$$

where $\Gamma_i = \Gamma(:, i)$, the ith column vector. This completes the proof for 1) and 2). Now we prove 3).

$$\begin{aligned} E_0\left[z_0\Gamma(s_0,:)\right] \; & = \; E_0\left[\sum_{k=-\infty}^{0}(\gamma\lambda)^{-k}\Gamma^\mathsf{T}(s_k,:)\Gamma(s_0,:)\right] \\ & = \; \sum_{k=-\infty}^{0}(\gamma\lambda)^{-k}E_0\left[\Gamma^\mathsf{T}(s_k,:)\Gamma(s_0,:)\right] \\ & = \; \sum_{m=0}^{\infty}(\gamma\lambda)^m E_0\left[\Gamma^\mathsf{T}(s_{-m},:)\Gamma(s_0,:)\right] \\ & = \; \sum_{m=0}^{\infty}(\gamma\lambda)^m\Gamma^\mathsf{T}D(\Pi P)^m\Gamma \end{aligned}$$

Note that $E_0[z_t\Gamma(s_t,:)]$ is the same for all t. The results of 4) and 5) can be proved similarly. ∎

Lemma 8 *The following expectation is well defined and finite for any finite* $\mathbf{w}$*,*

$$E_0\left[u(\mathbf{w}, X_t)\right] = \Gamma^\mathsf{T}D(T^{(\lambda)}(\Gamma\mathbf{w}) - \Gamma\mathbf{w})$$

Proof: By Lemma 7, we have,

$$\begin{aligned} & E_0\left[u(\mathbf{w}, X_t)\right] \\ = \; & E_0\left[(1-\gamma)z_t r + \gamma z_t\Gamma(s_{t+1},:)\mathbf{w} - z_t\Gamma(s_t,:)\mathbf{w}\right] \\ = \; & \Gamma^\mathsf{T}D\sum_{m=0}^{\infty}(\gamma\lambda\Pi P)^m((1-\gamma)\Pi\mathbf{r} + \gamma\Pi P\Gamma\mathbf{w} - \Gamma\mathbf{w}) \end{aligned}$$

For $\lambda = 1$, it follows that

$$\begin{aligned} & E_0\left[u(\mathbf{w}, X_t)\right] \\ = \; & \Gamma^\mathsf{T}D\sum_{m=0}^{\infty}(\gamma\Pi P)^m((1-\gamma)\Pi\mathbf{r} + \gamma\Pi P\Gamma\mathbf{w} - \Gamma\mathbf{w}) \\ = \; & \Gamma^\mathsf{T}D\left(M^*\Pi\mathbf{r} + \sum_{m=0}^{\infty}(\gamma\Pi P)^m(\gamma\Pi P\Gamma\mathbf{w} - \Gamma\mathbf{w})\right) \\ = \; & \Gamma^\mathsf{T}D(M^*\Pi\mathbf{r} - \Gamma\mathbf{w}) \end{aligned}$$

For $\lambda \in [0, 1)$, we have $\sum_{m=0}^{\infty}(\gamma\lambda\Pi P)^m M = (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}(\gamma\Pi P)^t M$, therefore,

$$\begin{aligned} & E_0\left[u(\mathbf{w}, X_t)\right] \\ = \; & \Gamma^\mathsf{T}D((1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}(\gamma\Pi P)^t(1-\gamma)\Pi\mathbf{r} \end{aligned}$$

$$
\begin{aligned}
&+(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}(\gamma\Pi P)^t(\gamma\Pi P\Gamma\mathbf{w}-\Gamma\mathbf{w}))\\
=\;&\Gamma^\mathsf{T}D((1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}(\gamma\Pi P)^t(1-\gamma)\Pi\mathbf{r}\\
&+((1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\gamma\Pi P)^{m+1}-I)\Gamma\mathbf{w})\\
=\;&\Gamma^\mathsf{T}D(T^{(\lambda)}(\Gamma\mathbf{w})-\Gamma\mathbf{w})
\end{aligned}
$$

By Lemma 7, each expectation is well defined and finite. ∎

Lemma 9 *We have,* $(\mathbf{w}-\mathbf{w}^*)^\mathsf{T}E_0\left[u(\mathbf{w},X_t)\right] < 0, \forall\mathbf{w}\neq\mathbf{w}^*$.

Proof: From (14), $\mathcal{P}=\Gamma(\Gamma^\mathsf{T}D\Gamma)^{-1}\Gamma^\mathsf{T}D$, we have, $\Gamma^\mathsf{T}D\mathcal{P}=\Gamma^\mathsf{T}D$. Thus,

$$
\begin{aligned}
&(\mathbf{w}-\mathbf{w}^*)^\mathsf{T}\Gamma^\mathsf{T}D(T^{(\lambda)}(\Gamma\mathbf{w})-\Gamma\mathbf{w})\\
=\;&(\mathbf{w}-\mathbf{w}^*)^\mathsf{T}\Gamma^\mathsf{T}D((I-\mathcal{P})T^{(\lambda)}(\Gamma\mathbf{w})\\
&+\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})-\Gamma\mathbf{w})\\
=\;&(\mathbf{w}-\mathbf{w}^*)^\mathsf{T}\Gamma^\mathsf{T}D(\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})-\Gamma\mathbf{w})
\end{aligned}
$$

$$
\begin{aligned}
&\left|\left|\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})-\Gamma\mathbf{w}^*\right|\right|_D\\
=\;&\left|\left|\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})-\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w}^*)\right|\right|_D\\
\leq\;&\left|\left|T^{(\lambda)}(\Gamma\mathbf{w})-T^{(\lambda)}(\Gamma\mathbf{w}^*)\right|\right|_D\\
\leq\;&\gamma\left|\left|\Gamma\mathbf{w}-\Gamma\mathbf{w}^*\right|\right|_D
\end{aligned}
$$

where the first inequality is from Lemma 6 and the second from Lemma 4.

$$
\begin{aligned}
&(\mathbf{w}-\mathbf{w}^*)^\mathsf{T}\Gamma^\mathsf{T}D(T^{(\lambda)}(\Gamma\mathbf{w})-\Gamma\mathbf{w})\\
=\;&(\Gamma\mathbf{w}-\Gamma\mathbf{w}^*)^\mathsf{T}D(\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})-\Gamma\mathbf{w})\\
=\;&(\Gamma\mathbf{w}-\Gamma\mathbf{w}^*)^\mathsf{T}D(\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})-\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w}^*)\\
&+\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w}^*)-\Gamma\mathbf{w})\\
=\;&(\Gamma\mathbf{w}-\Gamma\mathbf{w}^*)^\mathsf{T}D(\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})-\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w}^*)\\
&+\Gamma\mathbf{w}^*-\Gamma\mathbf{w})\\
\leq\;&\left|\left|\Gamma\mathbf{w}-\Gamma\mathbf{w}^*\right|\right|_D\left|\left|\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})-\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w}^*)\right|\right|_D\\
&-\left|\left|\Gamma\mathbf{w}-\Gamma\mathbf{w}^*\right|\right|_D^2\leq(\gamma-1)\left|\left|\Gamma\mathbf{w}-\Gamma\mathbf{w}^*\right|\right|_D^2
\end{aligned}
$$

where we use the Cauchy-Schwartz inequality. The result follows since $\gamma<1$. ∎

Theorem 1 *For any $\lambda\in[0,1]$, the dual TD(λ) algorithm with linear function approximation converges. The limit of convergence $\mathbf{w}^*$ is the unique solution of the equation $\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w})=\Gamma\mathbf{w}$. Furthermore, $\mathbf{w}^*$ satisfies $\left|\left|M_+-M^*\right|\right|_{D,\Pi\mathbf{r}}\leq\frac{(1-\lambda\gamma)}{1-\gamma}\left|\left|\mathcal{P}M^*-M^*\right|\right|_{D,\Pi\mathbf{r}}$.*

Proof: Let $u(\mathbf{w},X_t)=A(X_t)\mathbf{w}_t+b(X_t)$, thus $A(X_t)=z_t(\gamma\Gamma(s_{t+1},:)-\Gamma(s_t,:))$ and $b(X_t)=(1-\gamma)z_tr$. By Lemma 7, $A=E_0\left[A(X_t)\right]$ and $b=E_0\left[b(X_t)\right]$ are well defined and finite.

We now prove A is negative definite. By Lemma 6, we have $\mathcal{P}T^{(\lambda)}(\Gamma\mathbf{w}^*)=\Gamma\mathbf{w}^*$. From (14), we have $\Gamma^\mathsf{T}D\mathcal{P}=\Gamma^\mathsf{T}D$. Thus, $\Gamma^\mathsf{T}DT^{(\lambda)}(\Gamma\mathbf{w}^*)=\Gamma^\mathsf{T}D\Gamma\mathbf{w}^*$. By Lemma 8, $E_0\left[u(\mathbf{w}^*,X_t)\right]=\Gamma^\mathsf{T}D(T^{(\lambda)}(\Gamma\mathbf{w}^*)-\Gamma\mathbf{w}^*)=0$. We have, $A(\mathbf{w}-\mathbf{w}^*)=E_0[u(\mathbf{w},X_t)]-E_0[u(\mathbf{w}^*,X_t)]=E_0[u(\mathbf{w},X_t)]$. By Lemma 9, $(\mathbf{w}-\mathbf{w}^*)^\mathsf{T}A(\mathbf{w}-\mathbf{w}^*)<0$.

Following the line of proof in Tsitsiklis and Van Roy Tsitsiklis and Van Roy (1997), and examining the correspondence between ϕ in the primal representation and Γ in the dual representation, it is not difficult to prove the required "degree of stability" (conditions 5 and 6 of Theorem 2 in Tsitsiklis and Van Roy (1997)).

Therefore, with all the conditions satisfied in Theorem 2 in Tsitsiklis and Van Roy (1997), $\mathbf{w}_t$ converges to $\mathbf{w}^*$, which solves $A\mathbf{w}+b=0$. Since $A\mathbf{w}+b=E_0\left[s(\mathbf{w},X_t)\right]$, from Lemma 8, we have, $\Gamma^\mathsf{T}D(T^{(\lambda)}(\Gamma\mathbf{w})-\Gamma\mathbf{w})=0$. With the fact that $\Gamma^\mathsf{T}D$ has a full row rank, $\mathbf{w}^*$ uniquely satisfies this equation. Lemma 6 implies that $\mathbf{w}^*$ is the unique fixed point of $\mathcal{P}T^{(\lambda)}$ and provides the desired error bound. ∎

5 Empirical results

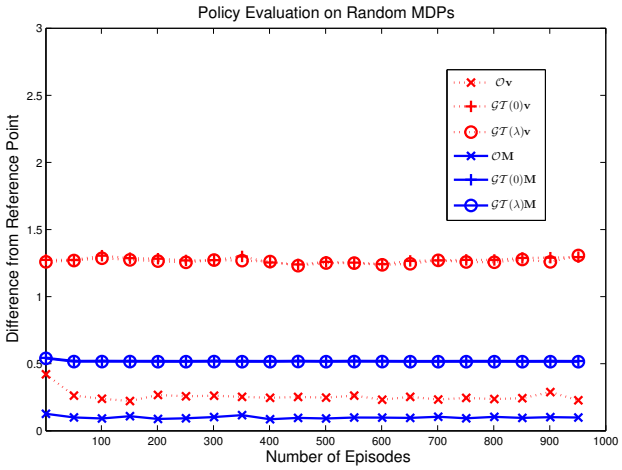
We evaluate the convergence property of the algorithms on two tasks: randomly synthesized MDPs and the mountain car problem. We study three dual TD(λ) algorithms: dual tabular TD(0), dual TD(0) and dual TD(λ), denoted as $\mathcal{O}M$, $\mathcal{GT}(0)M$ and $\mathcal{GT}(\lambda)M$, whose updating procedures are presented in (3), (4) and (6), respectively. We compare their performance with the primal algorithms: tabular TD(0), TD(0) and TD(λ), denoted as $\mathcal{O}\mathbf{v}$, $\mathcal{GT}(0)\mathbf{v}$ and $\mathcal{GT}(\lambda)\mathbf{v}$, whose updating procedures are presented in (1), (2) and (5), respectively.

For the synthetic MDPs, the transition model P follows the uniform distribution $U[0,1]$ and the reward function $\mathbf{r}$ follows the standard normal distribution $N(0,1)$. We report the results for 100 states and 5 actions. The mountain car problem has continuous state and action spaces. We discretize it with 222 states and 3 actions. The reward is -1 everywhere except that the reward is 100 at the right top (the target). Our focus is to study the convergence property of the algorithms, so we do not attempt to refine the choice of basis functions. For both tasks, we choose 5 random bases. For the primal algorithms, we generate random basis functions following $N(0,1)$. For the dual

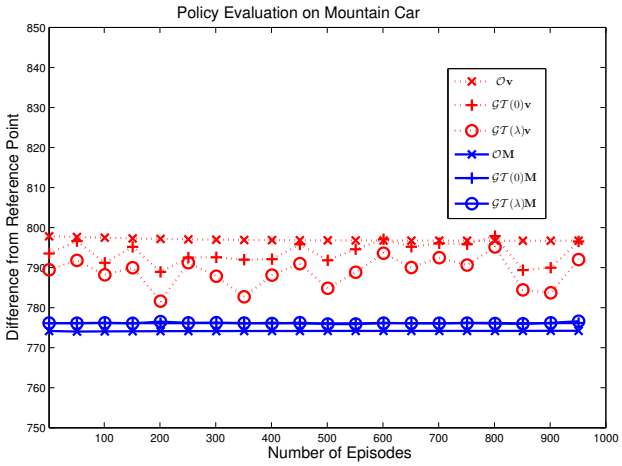
algorithms, we choose basis functions randomly from $U(0,1)$ and normalize them.

The results are averaged over 30 runs, for a fixed random policy and a fixed set of bases. Each run has 1000 episodes and each episode has 1000 steps. In the mountain car problem, an episode terminates once the target is reached. The reference point is calculated offline by solving a DP problem, which is an optimal value. Figure 1 shows the results of which the operators $\mathcal{P}_{\mathbf{1}}$ and $\mathcal{P}^+$ are applied at the end of each episode to make the weight vector normalized and non-negative. For the random MDPs, dual TD(0) and dual TD(λ) have lower errors than TD(0) and TD(λ), and dual tabular TD(0) has lower error than tabular TD(0). Moreover, the curves for dual algorithms are stable. However the curves for primal algorithms are slightly choppy. The dual and primal TD(0) plots overlap with their TD(λ) counterparts respectively in Figure 1(a).

For the mountain car problem, dual algorithms converge fast (in less than 100 episodes). However, primal algorithms TD(0) and TD(λ) are much less stable. The dual algorithms in general achieve lower errors. We have similar results for the cases the operators $\mathcal{P}_{\mathbf{1}}$ and $\mathcal{P}^+$ are applied in every step or at the end of the whole experiments. The dual TD(0) plot overlaps with the TD(λ) plot in Figure 1(b).



(a) Random MDP



(b) Mountain Car

Figure 1: Experimental results. ($\lambda = 0.5$ and $\gamma = 0.9$)

6 Conclusion

Recently, researchers have investigated novel dual representations as a basis for dynamic programming and reinforcement learning algorithms. Although the convergence properties of classical dynamic programming algorithms have been established using dual representations, temporal difference learning algorithms have not yet been analyzed. In this paper, we study the convergence properties of temporal difference learning using dual representations. We contribute significant progress by proving the convergence of dual temporal difference learning with eligibility traces. Experimental results on random MDPs and the mountain car problems suggest that the dual algorithms seem to demonstrate empirical benefits over standard primal algorithms.

References

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107 – 1149, December 2003.

Martin L. Puterman. *Markov decision processes : discrete stochastic dynamic programming*. John Wiley & Sons, New York, 1994.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

John. N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997.

Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (IEEE ADPRL-07)*, April 2007.

Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Stable dual dynamic programming. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2008.