A Worst-Case Comparison between Temporal Difference and Residual Gradient with Linear Function Approximation

Lihong Li LIHONG@CS.RUTGERS.EDU
Department of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ, USA 08854

Abstract

Residual gradient (RG) was proposed as an alternative to TD(0) for policy evaluation when function approximation is used, but there exists little formal analysis comparing them except in very limited cases. This paper employs techniques from online learning of linear functions and provides a worst-case (non-probabilistic) analysis to compare these two types of algorithms when linear function approximation is used. No statistical assumptions are made on the sequence of observations, so the analysis applies to non-Markovian and even adversarial domains as well. In particular, our results suggest that RG may result in smaller temporal differences, while TD(0) is more likely to yield smaller prediction errors. These phenomena can be observed even in two simple Markov chain examples that are non-adversarial.

1. Introduction

Reinforcement learning (RL) is a learning paradigm for optimal sequential decision making (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998) and has been successfully applied to a number of challenging problems. In the RL framework, the *agent* interacts with the *environment* in discrete timesteps by repeatedly observing its current state, taking an action, receiving a real-valued reward, and transitioning to a next state. A *policy* is a function that maps states to actions; semantically, it specifies what action to take given the current state. The goal of an agent is to optimize its policy in order to maximize the expected long-term return, namely, the discounted sum of rewards it receives by following the policy.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

An important step in this optimization process is *policy evaluation*—the problem of evaluating expected returns of a *fixed* policy. This problem is often the most challenging step in approximate policy-iteration algorithms (Bertsekas & Tsitsiklis, 1996; Lagoudakis & Parr, 2003). *Temporal difference* (TD) is a family of algorithms for policy evaluation (Sutton, 1988) and has received a lot of attention from the community. Unfortunately, it is observed (*e.g.*, Baird (1995)) that TD methods may diverge when they are combined with *function approximation*. An alternative algorithm known as *residual gradient* (RG) was proposed by Baird (1995) and enjoys guaranteed convergence to a local optimum. Since RG is similar to TD(0), a particular instance of the TD family, we will focus on RG, TD(0), and a variant of TD(0) in this paper.

Despite convergence issues, little is known that compares RG and TD(0). Building on previous work on *online learning of linear functions* (Cesa-Bianchi et al., 1996) and a similar analysis by Schapire and Warmuth (1996), we provide a worst-case (non-probabilistic) analysis of these algorithms and focus on two evaluation metrics: (i) total squared prediction error, and (ii) total squared temporal difference. The former measures *accuracy* of the predictions, while the latter measures *consistency* and is closely related to the *Bellman error* (Sutton & Barto, 1998).

Either metric may be preferred over the other in different situations. For instance, Lagoudakis and Parr (2003) argue that TD solutions tend to preserve the shape of the value function and is more suitable for approximate policy iteration, while there is evidence that minimizing squared Bellman errors is more robust in general (Munos, 2003). Our analysis suggests that TD can make more accurate predictions, while RG can result in smaller temporal differences. All terms will be made precise in the next section. Although our theory focuses on worst-case upper bounds, we also provide numerical evidence and expect the resulting insights to give useful guidance to RL practitioners in deciding which algorithm best suits their purposes.

2. Preliminaries

Fully observable environments in RL are often modelled as *Markov decision processes* (Puterman, 1994), which are equivalent to induced *Markov chains* when controlled by a fixed policy. Here, however, we consider a different model that is suitable for worst-case analysis, as introduced in the next subsection. This model makes no statistical assumption about the observations, and thus our results apply to much more general situations including *partially* observable or adversarial environments that subsume Markov chains.

Some notation is in order. We use bold-face, lower-case letters to denote real-valued column vectors such as $\mathbf{v}$. Their components are denoted by the corresponding letter with subscripts such as v_t. We use $\|\cdot\|$ to denote the Euclidean, or ℓ_2-norm: $\|\mathbf{v}\| = \sqrt{\mathbf{v}^\top \mathbf{v}}$ where $\mathbf{v}^\top$ is the transpose of $\mathbf{v}$. For a square matrix M, the set of eigenvalues of M, known as the *spectrum* of M, is denoted $\sigma(M)$. If M is symmetric, its eigenvalues must be real, and its largest eigenvalue is denoted $\rho(M)$.

2.1. The Sequential Online Learning Model

Our learning model is adopted from Schapire and Warmuth (1996) and is an extension of the online-learning model to sequential prediction problems. Let k be the dimension of input vectors. The agent maintains a weight vector of the same dimension and uses it to make predictions. In RL, input vectors are often feature vectors of states or state–action pairs, and are used to approximate value functions (Sutton & Barto, 1998). Learning proceeds in discrete timesteps and terminates after T steps. The agent starts with an initial input vector $\mathbf{x}_1 \in \mathbb{R}^k$ and an initial weight vector $\mathbf{w}_1 \in \mathbb{R}^k$. At timestep $t \in \{1, 2, 3, \cdots, T\}$:

- The agent makes a prediction $\hat{y}_t = \mathbf{w}_t^\top \mathbf{x}_t \in \mathbb{R}$, where $\mathbf{w}_t$ is the weight vector at time t. Throughout the paper, assume $\|\mathbf{x}_t\| \le X$ for some known constant $X > 0$.
- The agent then observes an *immediate reward* $r_t \in \mathbb{R}$ and the next input vector $\mathbf{x}_{t+1}$. Based on this information, it updates its weight vector whose new value is denoted $\mathbf{w}_{t+1}$. The change in weight is $\Delta\mathbf{w}_t = \mathbf{w}_{t+1} - \mathbf{w}_t$.

By convention, $r_t = 0$ and $\mathbf{x}_t = \mathbf{0}$ for $t > T$. Define the *return* at time t by $y_t = \sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau$, where $\gamma \in [0, 1)$ is the *discount factor*. Since $\gamma < 1$, it effectively diminishes future rewards exponentially fast. A quick observation is that $y_t = r_t + \gamma y_{t+1}$, which is analogous to the Bellman equation for Markov chains (Sutton & Barto, 1998). The agent attempts to mimic y_t by its prediction $\hat{y}_t$, and the *prediction error* is $e_t = y_t - \hat{y}_t$. Our first evaluation metric is the *total squared prediction error*: $\ell_{\mathcal{P}} = \sum_{t=1}^{T} e_t^2 = \|\mathbf{e}\|^2$.

Another useful metric in RL is the *temporal differences* (also known as *TD errors*), which measures how consistent the predictions are. In particular, the temporal difference at time t is $d_t = r_t + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t$, and the *total squared temporal difference* is $\ell_{\mathcal{TD}} = \sum_{t=1}^{T} d_t^2 = \|\mathbf{d}\|^2$.

2.2. Previous Work

Previous convergence results of TD and RG often rely heavily on certain stochastic assumptions of the environment such as the assumption that the sequence of observations, $[(\mathbf{x}_t, r_t)]_{t\in\mathbb{N}}$, are generated by an irreducible and aperiodic Markov chain. Tsitsiklis and Van Roy (1997) first proved convergence of TD with linear function approximation, while they also pointed out the potential divergence risk when nonlinear approximation is used.

To resolve the instability issue of TD(0), Baird (1995) proposed the RG algorithm, but also noted that RG may converge more slowly than TD(0) in some problems. Such an observation was later proved by Schoknecht and Merke (2003), who used spectral analysis to compare the asymptotic convergence rates of the two algorithms. Although their results are interesting, they only apply to quite limited cases where, for example, a certain matrix associated with TD updates has real eigenvalues only (which does not hold in general). More importantly, they study *synchronous* updates while TD and RG are often applied *asynchronously* in practice. Furthermore, their results assume that the value function is represented by a lookup table, but the initial motivation of studying RG was to develop a provably convergent algorithm when function approximation is used.

Schapire and Warmuth (1996) were also concerned with similar worst-case behavior of TD-like algorithms within the model described in Subsection 2.1. They defined a new class of algorithms called TD$^*(\lambda)$, which is very similar to the TD(λ) algorithms of Sutton (1988). They developed worst-case bounds for the total squared prediction error of TD$^*(\lambda)$, but not the total squared temporal difference.

2.3. Algorithms

The algorithms we consider all update the weight vector incrementally and differ only in the update rules. TD(0) uses the following rule:

$$\Delta\mathbf{w}_t = \eta d_t \mathbf{x}_t, \tag{1}$$

where $\eta \in (0,1)$ is the *step-size* parameter controlling aggressiveness of the update. Although TD(0) is widely used in practice, analysis turns out to be easier with a close relative of it, TD*(0). This algorithm differs from TD(0) in that it adapts the step-size based on the input vectors (Schapire and Warmuth (1996) defined TD*(0) in a different, but equivalent, form):

$$\Delta \mathbf{w}_t = \frac{\eta d_t \mathbf{x}_t}{1 - \gamma\eta \mathbf{x}_t^\top \mathbf{x}_{t+1}}. \tag{2}$$

Due to space limitation, we only provide results for TD*(0), but similar results hold for TD(0). It is expected, and also supported by the numerical evidence in Section 4, that TD(0) and TD*(0) have similar behavior and performance in practice. For this reason, we refer to both algorithms as TD in the rest of the paper if there is no risk of confusion. In contrast, RG uses the following update rule:

$$\Delta \mathbf{w}_t = \eta d_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right). \tag{3}$$

3. Main Results

This section contains the main theoretical results. We will first describe how to evaluate an algorithm in the worst-case scenario. For completeness, we also summarize the squared prediction error bounds for TD*(0) due to Schapire and Warmuth (1996). Then, we analyze total squared temporal difference bounds and RG.

Our analysis makes a few uses of matrix theory (see, *e.g.*, Horn and Johnson (1986)), and several technical lemmas are found in the appendix. Two basic facts about $\rho(M)$ will be used repeatedly: (i) if M is negative-definite, then $\rho(M) < 0$; and (ii) the Rayleigh-Ritz theorem (Horn & Johnson, 1986, Theorem 4.2.2) states that $\rho(M) = \max_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^\top M \mathbf{v}}{\mathbf{v}^\top \mathbf{v}}$.

3.1. Evaluation Criterion

Analogous to other online-learning analysis, we treat $\ell_{\mathcal{P}}$ and $\ell_{\mathcal{TD}}$ as *total losses*, and compare the total loss of an algorithm to that of an arbitrary weight vector, $\mathbf{u}$. We wish to prove that this difference is small for *all* $\mathbf{u}$, including the *optimal* (in any well-defined sense) but unknown vector $\mathbf{u}^*$.

The prediction using vector $\mathbf{u}$ at time t is $y_t^{\mathbf{u}} = \mathbf{u}^\top \mathbf{x}_t$. Accordingly, the prediction error and temporal difference at time t are $e_t^{\mathbf{u}} = y_t - y_t^{\mathbf{u}}$ and $d_t^{\mathbf{u}} = r_t + \gamma \mathbf{u}^\top \mathbf{x}_{t+1} - \mathbf{u}^\top \mathbf{x}_t$, respectively. The total squared prediction error and total squared temporal difference of $\mathbf{u}$ are $\ell_{\mathcal{P}}^{\mathbf{u}} = \|\mathbf{e}^{\mathbf{u}}\|^2 = \sum_{t=1}^T \left(y_t - \mathbf{u}^\top \mathbf{x}_t\right)^2$ and $\ell_{\mathcal{TD}}^{\mathbf{u}} = \|\mathbf{d}^{\mathbf{u}}\|^2 = \sum_{t=1}^T \left(r_t + \gamma \mathbf{u}^\top \mathbf{x}_{t+1} - \mathbf{u}^\top \mathbf{x}_t\right)^2$, respectively.

3.2. Squared Prediction Errors of TD*(0)

Using step-size $\eta = \frac{1}{X^2+1}$, Schapire and Warmuth (1996) showed a worst-case upper bound:

$$\ell_{\mathcal{P}} \leq \frac{\left(1+X^2\right)\left(\ell_{\mathcal{P}}^{\mathbf{u}} + \|\mathbf{w}_1 - \mathbf{u}\|_2^2\right)}{1-\gamma^2}.$$

Furthermore, if E and W are known beforehand such that $\ell_{\mathcal{P}}^{\mathbf{u}} \leq E$ and $\|\mathbf{w}_1 - \mathbf{u}\| \leq W$, then the step-size η can be optimized by $\eta = \frac{W}{X\sqrt{E}+X^2 W}$ to yield an asymptotically better bound:

$$\ell_{\mathcal{P}} \leq \frac{\ell_{\mathcal{P}}^{\mathbf{u}} + 2WX\sqrt{E} + X^2W^2}{1-\gamma^2}. \tag{4}$$

3.3. Squared Temporal Differences of TD*(0)

We will extend the analysis of Schapire and Warmuth (1996) to the new loss function $\ell_{\mathcal{TD}}$ by examining how the potential function, $\|\mathbf{w}_t - \mathbf{u}\|^2$, evolves when a single update is made at time t. It can be shown (Schapire & Warmuth, 1996, Eqn 8) that $-\|\mathbf{w}_1 - \mathbf{u}\|^2 \leq \eta^2 X^2 \mathbf{e}^\top D^\top D \mathbf{e} + 2\eta \mathbf{e}^\top D^\top (\mathbf{e}^{\mathbf{u}} - \mathbf{e})$, where

$$D = \begin{pmatrix} 1 & -\gamma & 0 & \cdots & 0 & 0 \\ 0 & 1 & -\gamma & \cdots & 0 & 0 \\ & & \ddots & & & \\ 0 & 0 & \cdots & 1 & -\gamma & 0 \\ 0 & 0 & \cdots & 0 & 1 & -\gamma \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

Define $\mathbf{f} = D\mathbf{e}$. According to Lemma A.1(1), $\mathbf{d}^{\mathbf{u}} = D\mathbf{e}^{\mathbf{u}}$, and hence the inequality above is rewritten as:

$$-\|\mathbf{w}_1 - \mathbf{u}\|^2 \leq \eta^2 X^2 \mathbf{f}^\top \mathbf{f} - 2\eta \mathbf{f}^\top D^{-1} \mathbf{f} + 2\eta \mathbf{f}^\top D^{-1} \mathbf{d}^{\mathbf{u}}.$$

Using the fact that $2\mathbf{p}^\top \mathbf{q} \leq \|\mathbf{p}\|^2 + \|\mathbf{q}\|^2$ for $\mathbf{p} = \frac{\eta}{\sqrt{b}} D^{-\top}\mathbf{f}$, $\mathbf{q} = \sqrt{b}\mathbf{d}^{\mathbf{u}}$, and arbitrary $b > 0$, the inequality becomes $-\|\mathbf{w}_1 - \mathbf{u}\|^2 \leq \mathbf{f}^\top M_1 \mathbf{f} + b\ell_{\mathcal{TD}}^{\mathbf{u}}$, where

$$M_1 = \eta^2 X^2 I + \frac{\eta^2}{b} D^{-1} D^{-\top} - \eta(D^{-1} + D^{-\top}) \tag{6}$$

is a symmetric matrix. Since $\rho(M_1)$ is the largest eigenvalue of M_1, we have $\mathbf{f}^\top M_1 \mathbf{f} \leq \rho(M_1)\|\mathbf{f}\|^2$, and hence, $-\|\mathbf{w}_1 - \mathbf{u}\|^2 \leq \|\mathbf{f}\|^2 \rho(M_1) + b\ell_{\mathcal{TD}}^{\mathbf{u}}$. Combining this with Lemma A.4, we have that $\|\mathbf{f}\|^2$ is at most

$$(1+\gamma)^2 \left(X^2 + \frac{1}{b(1-\gamma)^2}\right)\left(b\ell_{\mathcal{TD}}^{\mathbf{u}} + \|\mathbf{w}_1 - \mathbf{b}\|^2\right),$$

when the step-size is

$$\eta = \frac{1}{(1+\gamma)\left(X^2 + \frac{1}{b(1-\gamma)^2}\right)}. \tag{7}$$

Due to Lemma A.1 (2), we have

$$\begin{aligned} d_t^2 &= \left(1-\gamma\eta\mathbf{x}_t^\top\mathbf{x}_{t+1}\right)^2 f_t^2 \\ &\le \left(1+\gamma\eta X^2\right)^2 f_t^2 \;\le\; \frac{(1+2\gamma)^2}{(1+\gamma)^2} f_t^2. \end{aligned}$$

Therefore, $\ell_{\mathcal{TD}}$ is at most

$$(1+2\gamma)^2\left(X^2+\frac{1}{b(1-\gamma)^2}\right)\left(b\ell_{\mathcal{TD}}^{\mathbf{u}}+\|\mathbf{w}_1-\mathbf{u}\|^2\right).$$

Using $b=1$, we have thus proved the first main result.

Theorem 3.1. *Let η be given by Eqn 7 using $b=1$, then the following holds for TD*(0):*

$$\ell_{\mathcal{TD}} \le (1+2\gamma)^2\left(X^2+\frac{1}{(1-\gamma)^2}\right)\left(\ell_{\mathcal{TD}}^{\mathbf{u}}+\|\mathbf{w}_1-\mathbf{u}\|^2\right).$$

Theorem 3.2. *If E and W are known beforehand such that $\ell_{\mathcal{TD}}^{\mathbf{u}} \le E$ and $\|\mathbf{w}_1-\mathbf{u}\| \le W$, then η can be optimized in TD*(0) so that*

$$\ell_{\mathcal{TD}} \;\le\; (1+2\gamma)^2 \left(\frac{\ell_{\mathcal{TD}}^{\mathbf{u}}}{(1-\gamma)^2}+\frac{2XW\sqrt{E}}{1-\gamma}+X^2\|\mathbf{w}_1-\mathbf{u}\|^2\right). \quad (8)$$

Proof. Previous analysis for Theorem 3.1 yields

$$\begin{aligned} \ell_{\mathcal{TD}} &\le (1+2\gamma)^2\left(\left(bX^2\ell_{\mathcal{TD}}^{\mathbf{u}}+\frac{\|\mathbf{w}_1-\mathbf{u}\|^2}{b(1-\gamma)^2}\right)+\left(\frac{\ell_{\mathcal{TD}}^{\mathbf{u}}}{(1-\gamma)^2}+X^2\|\mathbf{w}_1-\mathbf{u}\|^2\right)\right) \\ &\le (1+2\gamma)^2\left(\left(bX^2E+\frac{W^2}{b(1-\gamma)^2}\right)+\left(\frac{\ell_{\mathcal{TD}}^{\mathbf{u}}}{(1-\gamma)^2}+X^2\|\mathbf{w}_1-\mathbf{u}\|^2\right)\right) \end{aligned}$$

for any $b>0$. We may simply choose $b=\frac{W}{X(1-\gamma)\sqrt{E}}$, and the step-size in Eqn 7 becomes

$$\eta=\frac{1}{(1+\gamma)\left(X^2+\frac{X\sqrt{E}}{W(1-\gamma)}\right)}. \quad \square$$

3.4. Squared Prediction Errors of RG

By the update rule in Eqn 3 and simple algebra,

$$\begin{aligned} \Delta\mathbf{w}_t^\top(\mathbf{w}_t-\mathbf{u}) &= \eta d_t(\mathbf{x}_t-\gamma\mathbf{x}_{t+1})^\top(\mathbf{w}_t-\mathbf{u}) \\ &= \eta d_t\left(\left(\mathbf{w}_t^\top\mathbf{x}_t-\gamma\mathbf{w}_t^\top\mathbf{x}_{t+1}-r_t\right)-\left(\mathbf{u}^\top\mathbf{x}_t-\gamma\mathbf{u}^\top\mathbf{x}_{t+1}-r_t\right)\right) \\ &= \eta d_t(d_t^{\mathbf{u}}-d_t), \\ \|\Delta\mathbf{w}_t\|^2 &= \eta^2 d_t^2\|\mathbf{x}_t-\gamma\mathbf{x}_{t+1}\|_2^2 \\ &\le \eta^2 d_t^2 X^2(1+\gamma)^2. \end{aligned}$$

Similar to the previous section, we use the potential function $\|\mathbf{w}_t-\mathbf{u}\|^2$ to measure progress of learning:

$$\begin{aligned} -\|\mathbf{w}_1-\mathbf{u}\|^2 &\le \sum_{t=1}^T\left(\|\mathbf{w}_{t+1}-\mathbf{u}\|^2-\|\mathbf{w}_t-\mathbf{u}\|^2\right) \\ &= \sum_{t=1}^T\left(2\Delta\mathbf{w}_t^\top(\mathbf{w}_t-\mathbf{u})+\Delta\mathbf{w}_t^\top\Delta\mathbf{w}_t\right) \\ &\le \sum_{t=1}^T\left(2\eta d_t(d_t^{\mathbf{u}}-d_t)+\eta^2 d_t^2X^2(1+\gamma)^2\right) \\ &= 2\eta\mathbf{d}^\top\mathbf{d}^{\mathbf{u}}-2\eta\mathbf{d}^\top\mathbf{d}+\eta^2X^2(1+\gamma)^2\mathbf{d}^\top\mathbf{d}. \end{aligned}$$

According to Lemma A.1 (1) and using the fact that $2\mathbf{p}^\top\mathbf{q}\le\|\mathbf{p}\|^2+\|\mathbf{q}\|^2$ for $\mathbf{p}=\frac{\eta}{\sqrt{b}}D^\top\mathbf{d}$, $\mathbf{q}=\sqrt{b}\mathbf{e}^{\mathbf{u}}$, and arbitrary $b>0$, the inequality above is written as:

$$\begin{aligned} -\|\mathbf{w}_1-\mathbf{u}\|^2 \;\le\; & b\|\mathbf{e}^{\mathbf{u}}\|^2+\frac{\eta^2}{b}\mathbf{d}^\top DD^\top\mathbf{d}+ \\ & \left(\eta^2X^2(1+\gamma)^2-2\eta\right)\|\mathbf{d}\|^2 \end{aligned}$$

Due to Lemma A.1 (3), $\mathbf{d}=\Sigma D\mathbf{e}$, where

$$\begin{aligned} \Sigma = \operatorname{diag}\Bigg(&\frac{1}{1+\gamma\eta(\mathbf{x}_1-\gamma\mathbf{x}_2)^\top\mathbf{x}_2}, \\ &\frac{1}{1+\gamma\eta(\mathbf{x}_2-\gamma\mathbf{x}_3)^\top\mathbf{x}_3},\cdots, \\ &\frac{1}{1+\gamma\eta(\mathbf{x}_{T-1}-\gamma\mathbf{x}_T)^\top\mathbf{x}_T},1\Bigg). \end{aligned} \quad (9)$$

Then, the inequality above becomes:

$$-\|\mathbf{w}_1-\mathbf{u}\|^2 \le b\|\mathbf{e}^{\mathbf{u}}\|^2+\mathbf{e}^\top M_2\mathbf{e},$$

where

$$M_2=D^\top\Sigma\left(\frac{\eta^2}{b}DD^\top+\left(\eta^2X^2(1+\gamma)^2-2\eta\right)I\right)\Sigma D. \quad (10)$$

Since $\mathbf{e}^\top M_2\mathbf{e}\le\rho(M_2)\|\mathbf{e}\|^2$, Lemma A.5 implies the following theorems when the step-size is

$$\eta=\frac{1}{(1+\gamma)^2\left(X^2+\frac{1}{b}\right)}. \quad (11)$$

Theorem 3.3. *Let η be given by Eqn 11 using $b=1$, then the following holds for RG:*

$$\ell_{\mathcal{P}} \le \frac{(1+2\gamma)^2\left(X^2+1\right)}{(1-\gamma)^2}\left(\ell_{\mathcal{P}}^{\mathbf{u}}+\|\mathbf{w}_1-\mathbf{u}\|^2\right).$$

Theorem 3.4. *If E and W are known beforehand such that $\ell_{\mathcal{P}}^{\mathbf{u}} \le E$ and $\|\mathbf{w}_1-\mathbf{u}\| \le W$, then η can be optimized in RG so that*

$$\ell_{\mathcal{P}} \le \frac{(1+2\gamma)^2}{(1-\gamma)^2}\left(\ell_{\mathcal{P}}^{\mathbf{u}}+2XW\sqrt{E}+X^2\|\mathbf{w}_1-\mathbf{u}\|^2\right). \quad (12)$$

Proof. Previous analysis in this subsection yields

$$\begin{aligned}\ell_{\mathcal{P}} &\leq \frac{(1+2\gamma)^2}{(1-\gamma)^2}\left(\left(\ell_{\mathcal{P}}^{\mathbf{u}} + X^2\left\|\mathbf{w}_1-\mathbf{u}\right\|^2\right) + \left(X^2 b\ell_{\mathcal{P}}^{\mathbf{u}} + \frac{\left\|\mathbf{w}_1-\mathbf{u}\right\|^2}{b}\right)\right)\\ &\leq \frac{(1+2\gamma)^2}{(1-\gamma)^2}\left(\left(\ell_{\mathcal{P}}^{\mathbf{u}} + X^2\left\|\mathbf{w}_1-\mathbf{u}\right\|^2\right) + \left(X^2 bE + \frac{W^2}{b}\right)\right).\end{aligned}$$

We simply choose $b = \frac{W}{X\sqrt{E}}$ and accordingly the step-size in Eqn 11 becomes

$$\eta = \frac{1}{(1+\gamma)^2\left(X^2 + \frac{X\sqrt{E}}{W}\right)}. \quad \square$$

3.5. Squared Temporal Differences of RG

It is most convenient to turn this problem into one of analyzing the total squared prediction error in the original online-learning-of-linear-function framework (Cesa-Bianchi et al., 1996). In particular, define $\mathbf{z}_t = \mathbf{x}_t - \gamma\mathbf{x}_{t+1}$ and thus $\|\mathbf{z}_t\| \leq (1+\gamma)X$. Now, RG can be viewed as a gradient descent algorithm operating over the sequence of data $[(\mathbf{z}_t, r_t)]_{t\in\{1,2,\cdots,T\}}$. Due to Theorem IV.1 of Cesa-Bianchi et al. (1996), we immediately have

$$\ell_{\mathcal{TD}} \leq 2.25\left(\ell_{\mathcal{TD}}^{\mathbf{u}} + X^2(1+\gamma)^2\left\|\mathbf{u}\right\|^2\right),$$

for any $\mathbf{u}$ when the step-size is $\eta = \frac{2}{3X^2(1+\gamma)^2}$. If E and W are known beforehand so that $\ell_{\mathcal{TD}}^{\mathbf{u}} \leq E$ and $\|\mathbf{u}\| \leq W$, then η can be optimized (Theorem IV.3 of Cesa-Bianchi et al. (1996)) by $\eta = \frac{W}{X(1+\gamma)(WX(1+\gamma)+\sqrt{E})}$ to obtain the following improved bound:

$$\ell_{\mathcal{TD}} \leq \ell_{\mathcal{TD}}^{\mathbf{u}} + 2WX(1+\gamma)\sqrt{E} + (1+\gamma)^2W^2X^2. \quad (13)$$

3.6. Discussions

Based on Eqns 4, 8, 12, and 13, Table 1 summarizes the asymptotic upper bounds (when $T \to \infty$) assuming E and W are known beforehand to optimize η.[1] Although our bounds are all upper bounds, results in the table *suggest* that, in worst cases, TD*(0) (and also TD(0)) tend to make smaller prediction errors, while RG tends to make smaller temporal differences. The gaps between corresponding bounds increase as $\gamma \to 1$. On the other extreme where $\gamma = 0$, all these asymptotic bounds coincide, which is not surprising as TD(0), TD*(0), and RG are all identical when $\gamma = 0$.

[1] Strictly speaking, the validity of these asymptotic results relies on the assumptions that (i) $\sqrt{E} = o(\ell_{\mathcal{P}}^{\mathbf{u}})$, and (ii) W and X remain constant as $T \to \infty$. Both assumptions are reasonable in practice.

Table 1. Asymptotic upper bounds for total squared prediction error and total squared temporal difference of TD*(0) and RG.

	$\ell_{\mathcal{P}}/\ell_{\mathcal{P}}^{\mathbf{u}}$	$\ell_{\mathcal{TD}}/\ell_{\mathcal{TD}}^{\mathbf{u}}$
TD*(0)	$\frac{1}{1-\gamma^2} + o(1)$	$\frac{(1+2\gamma)^2}{(1-\gamma)^2} + o(1)$
RG	$\frac{(1+2\gamma)^2}{(1-\gamma)^2} + o(1)$	$1 + o(1)$

Since it is unknown whether the leading constants in Table 1 are optimal, the next section will provide numerical evidence to support our claims about the relative strengths of these algorithms.

It is worth mentioning that in sequential prediction or decision problems, the factor $\frac{1}{1-\gamma}$ often plays a role similar to the *decision horizon* (Puterman, 1994). Therefore, in some sense, our bounds also characterize how prediction errors and temporal differences may scale with decision horizon, in the worst-case sense.

When $\ell_{\mathcal{P}}$ or $\ell_{\mathcal{TD}}$ are relatively small, the asymptotic bounds in Table 1 are less useful as the $\|\mathbf{w}_1-\mathbf{u}\|^2$ in the bounds dominate $\ell_{\mathcal{P}}$ or $\ell_{\mathcal{TD}}$. However, we still get similar qualitative results by comparing the constant factors of the term $\|\mathbf{w}_1-\mathbf{u}\|^2$ in the bounds.

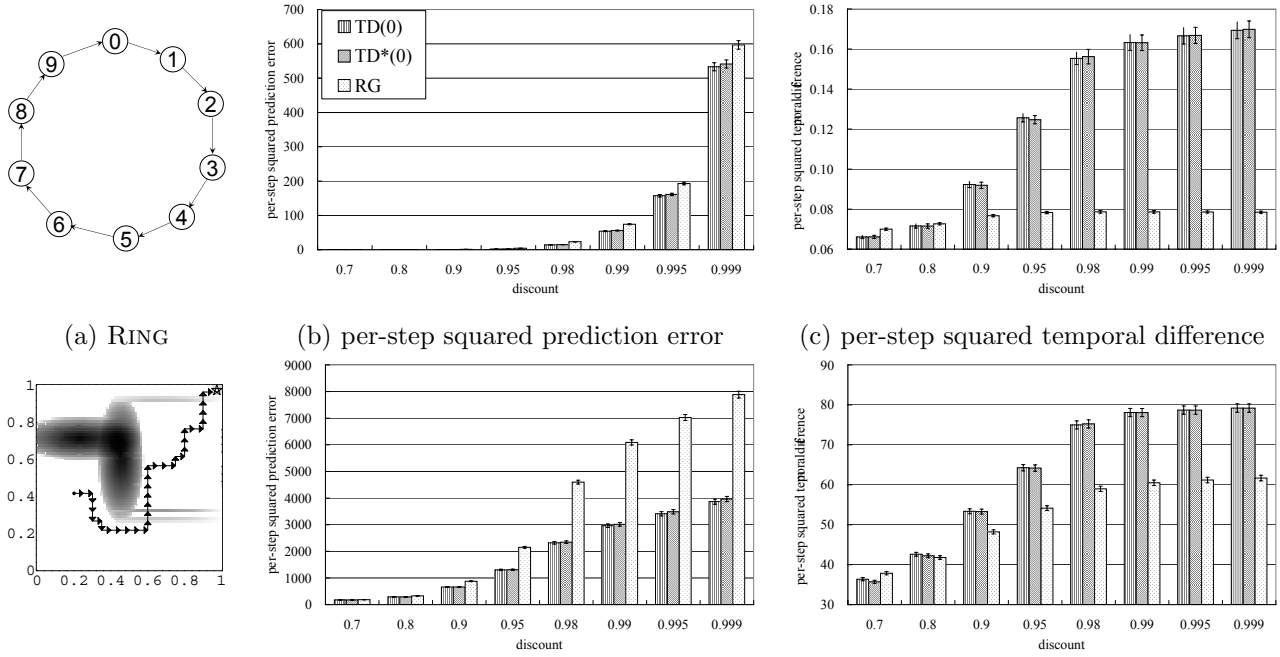
Since our setting is quite different from that of Schoknecht and Merke (2003), our results are not comparable to theirs.

4. Experiments

This section presents empirical evidence in two Markov chains that supports our claims in Section 3.6.

The first is the RING Markov chain (Figure 1 (a)), a variant of the HALL problem introduced by Baird (1995) in which RG was observed to converge to the optimal weights more slowly than TD(0). The state space is a ring consisting of 10 states numbered from 0 through 9. Each state is associated with a randomly selected feature vector of dimension $k = 5$: $\mathbf{x}^{(0)}, \cdots, \mathbf{x}^{(9)} \in \mathbb{R}^k$. Transitions are deterministic and are indicated by arrows. The reward in every state is stochastic and is distributed uniformly in $[-0.1, 0.1]$. As in HALL , the value of every state is exactly 0.

The second problem is a benchmark problem known as PUDDLEWORLD (Boyan & Moore, 1995). The state space is a unit square (Figure 1 (d)), and a start state of an episode is randomly selected in $[0, 0.2] \times [0, 0.2]$.



(a) RING (b) per-step squared prediction error (c) per-step squared temporal difference

(d) PUDDLEWORLD (e) per-step squared prediction error (f) per-step squared temporal difference

Figure 1. Two Markov chains we used: (a) RING and (d) PUDDLEWORLD (Boyan & Moore, 1995). All results are averaged over 500 runs, with 99% confidence intervals plotted. RING and PUDDLEWORLD results are in (b,c) and (e,f), respectively.

The agent adopts a fixed policy that goes north or east with probability 0.5 each. Every episode takes about 40 steps to terminate. The reward is -1 unless the agent steps into the puddles and receives penalty for that; the smallest possible reward is -41. We used 16 RBF features of width 0.3, whose centers were evenly distributed in the state space. We also tried a degree-two polynomial feature: for a state $\mathbf{s} = (s_1, s_2)^\top$, the feature vector had six components: $\mathbf{x_s} = \left(1, s_1, s_2, s_1 s_2, s_1^2, s_2^2\right)^\top$. Since the results are similar to those for RBF features, they are not included.

We ran three algorithms in the experiments: TD(0), TD*(0), and RG. For a fair comparison, all algorithms started with the all-one weight vector and were given the same sequence of $(\mathbf{x}_t, r_t)$ for learning. The procedure was repeated 500 times. For RING , each run used a different realization of feature $\mathbf{x}^{(s)}$ and $T = 500$; for PUDDLEWORLD , each run consisted of 50 episodes (yielding slightly less than 2000 steps in total). A wide range of step-sizes were tried, and the best choices for each discount-factor–algorithm combination were used to evaluate $\ell_{\mathcal{P}}$ and $\ell_{\mathcal{TD}}$, respectively. Figure 1 (b,c,e,f) gives the average per-step squared prediction errors and squared temporal differences for these two problems, with 99% confidence intervals plotted.

These results are consistent with our analysis: TD(0) and TD*(0) tended to make more accurate predictions, while RG did a better job at minimizing temporal differences; the differences between these algorithms were even larger as the discount factor γ approached 1.[2] Finally, as a side effect, it is verified that TD(0) and TD*(0) had essentially identical performance, although their best learning rates might differ.

5. Conclusion

We have carried out a worst-case analysis to compare two policy-evaluation algorithms, TD and RG, when linear function approximation is used. Together with previously known results due to Schapire and Warmuth (1996) and Cesa-Bianchi et al. (1996), our results suggest that, although the TD algorithms may make more accurate predictions, RG may be a better choice when small temporal differences are desired. This claim is supported by empirical evidence in two simple Markov chains. Although the analysis is purely mathematical, we expect the implications to deepen the understanding of these two types of algorithms and can provide useful insights to RL practitioners.

[2]This effect was less obvious when γ got too close to 1. This was because the trajectories in our experiments were not long enough for such γ to have full impacts.

There has been relatively little attention to this sort of online-learning analysis within the RL community. Our analysis shows that this kind of analysis may be helpful and provide useful insights. A few directions are worth pursuing. First, we have focused on worst-case upper bounds, but it remains open whether matching lower bounds can be found. More extensive empirical studies are also necessary to see if such worst-case behavior can be observed in realistic problems. Second, we wish to generalize the analysis of total squared temporal difference from TD(0) and TD*(0) to TD(λ) and TD*(λ), respectively. Finally, we would like to mention that, in their original forms, both TD and RG use *additive* updates. Another class of updates known as *multiplicative* updates (Kivinen & Warmuth, 1997) has been useful when the number of features (*i.e.*, the k in Subsection 2.1) is large but only a few of them are relevant for making predictions. Such learning rules have potential uses in RL (Precup & Sutton, 1997), but it remains open whether these algorithms converge or whether worst-case error bounds similar to the ones given in this paper can be obtained.

A. Lemmas and Proofs

Lemma A.1. *This lemma collects a few basic facts useful in our analysis (D is given in Eqn 5):*

1. *In all three algorithms,* $\mathbf{d^u} = D\mathbf{e^u}$.
2. *In TD*(0),* $d_t = (1 - \gamma\eta\mathbf{x}_t^\top\mathbf{x}_{t+1})(e_t - \gamma e_{t+1})$.
3. *In RG,* $d_t = \frac{e_t - \gamma e_{t+1}}{1+\gamma\eta(\mathbf{x}_t - \gamma\mathbf{x}_{t+1})^\top\mathbf{x}_{t+1}}$.

Proof. 1. Since $y_t = r_t + \gamma y_{t+1}$, we have

$$\begin{aligned} d_t^{\mathbf{u}} &= r_t + \gamma\mathbf{u}^\top\mathbf{x}_{t+1} - \mathbf{u}^\top\mathbf{x}_t \\ &= \left(y_t - \mathbf{u}^\top\mathbf{x}_t\right) - \left(y_t - r_t - \gamma\mathbf{u}^\top\mathbf{x}_{t+1}\right) \\ &= \left(y_t - \mathbf{u}^\top\mathbf{x}_t\right) - \gamma\left(y_{t+1} - \mathbf{u}^\top\mathbf{x}_{t+1}\right) \\ &= e_t^{\mathbf{u}} - \gamma e_{t+1}^{\mathbf{u}}. \end{aligned}$$

In matrix form, this is $\mathbf{d^u} = D\mathbf{e^u}$.

2. Since $\mathbf{w}_t = \mathbf{w}_{t+1} - \Delta\mathbf{w}_t$ and $y_t = r_t + \gamma y_{t+1}$,

$$\begin{aligned} d_t &= r_t + \gamma\mathbf{w}_t^\top\mathbf{x}_{t+1} - \mathbf{w}_t^\top\mathbf{x}_t \\ &= r_t + \gamma(\mathbf{w}_{t+1} - \Delta\mathbf{w}_t)^\top\mathbf{x}_{t+1} - \mathbf{w}_t^\top\mathbf{x}_t + \\ &\quad (y_t - r_t - \gamma y_{t+1}) \\ &= (y_t - \mathbf{w}_t^\top\mathbf{x}_t) - \gamma(y_{t+1} - \mathbf{w}_{t+1}^\top\mathbf{x}_{t+1}) - \\ &\quad \gamma\Delta\mathbf{w}_t^\top\mathbf{x}_{t+1} \\ &= e_t - \gamma e_{t+1} - \frac{\gamma\eta d_t\mathbf{x}_t^\top\mathbf{x}_{t+1}}{1 - \gamma\eta\mathbf{x}_t^\top\mathbf{x}_{t+1}}. \end{aligned}$$

Reorganizing terms will complete the proof.

3. Similar to the proof for part (2) except that $\Delta\mathbf{w}_t$ is computed by Eqn 3. □

Two technical lemmas are useful to prove Lemma A.4. It should be noted that the bounds they give are tight.

Lemma A.2. *For D given in Eqn 5, let A be $D^\top D$ or $DD^\top$, and B be $D^{-1}D^{-\top}$ or $D^{-\top}D^{-1}$. Then, $\sigma(A) \subseteq \left[(1-\gamma)^2, (1+\gamma)^2\right]$ and $\sigma(B) \subseteq \left[(1+\gamma)^{-2}, (1-\gamma)^{-2}\right]$.*

Proof. It can be verified that $D^\top D$ equals

$$\begin{pmatrix} 1 & -\gamma & 0 & \cdots & 0 \\ -\gamma & 1+\gamma^2 & -\gamma & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & 1+\gamma^2 & -\gamma \\ 0 & 0 & \cdots & -\gamma & 1+\gamma^2 \end{pmatrix}.$$

Since $D^\top D$ is symmetric, $\sigma\left(D^\top D\right) \subset \mathbb{R}$. It follows from Geršgorin's theorem (Horn & Johnson, 1986, Theorem 6.1.1) that $\sigma\left(D^\top D\right) \subseteq \left[(1-\gamma)^2, (1+\gamma)^2\right]$. The same holds for $\sigma\left(DD^\top\right)$. The second part follows immediately by observing that $D^{-1}D^{-\top} = \left(D^\top D\right)^{-1}$ and $D^{-\top}D^{-1} = \left(DD^\top\right)^{-1}$. □

Lemma A.3. *Let D be given by Eqn 5, then*

$$\sigma(D^{-1} + D^{-\top}) \subseteq \left[\frac{2}{1+\gamma}, \frac{2}{1-\gamma}\right].$$

Proof. It can be verified that $D^{-1} + D^{-\top}$ equals

$$G = \begin{pmatrix} 2 & \gamma & \gamma^2 & \cdots & \gamma^{T-2} & \gamma^{T-1} \\ \gamma & 2 & \gamma & \cdots & \gamma^{T-3} & \gamma^{T-2} \\ & & \ddots & & & \\ \gamma^{T-3} & \gamma^{T-4} & \cdots & 2 & \gamma & \gamma^2 \\ \gamma^{T-2} & \gamma^{T-3} & \cdots & \gamma & 2 & \gamma \\ \gamma^{T-1} & \gamma^{T-2} & \cdots & \gamma^2 & \gamma & 2 \end{pmatrix},$$

and that $(G-I)^{-1}$ equals

$$\begin{pmatrix} \frac{1}{1-\gamma^2} & \frac{-\gamma}{1-\gamma^2} & 0 & \cdots & 0 & 0 \\ \frac{-\gamma}{1-\gamma^2} & \frac{1+\gamma^2}{1-\gamma^2} & \frac{-\gamma}{1-\gamma^2} & \cdots & 0 & 0 \\ & & \ddots & & & \\ 0 & 0 & \cdots & \frac{1+\gamma^2}{1-\gamma^2} & \frac{-\gamma}{1-\gamma^2} & 0 \\ 0 & 0 & \cdots & \frac{-\gamma}{1-\gamma^2} & \frac{1+\gamma^2}{1-\gamma^2} & \frac{-\gamma}{1-\gamma^2} \\ 0 & 0 & \cdots & 0 & \frac{-\gamma}{1-\gamma^2} & \frac{1}{1-\gamma^2} \end{pmatrix}.$$

Clearly, $(G-I)^{-1}$ is symmetric, and it follows from Geršgorin's theorem that

$$\sigma\left((G-I)^{-1}\right) \subseteq \left[\frac{1-\gamma}{1+\gamma}, \frac{1+\gamma}{1-\gamma}\right].$$

Therefore,

$$\begin{aligned} \sigma(G-I) &\subseteq \left[\left(\frac{1+\gamma}{1-\gamma}\right)^{-1}, \left(\frac{1-\gamma}{1+\gamma}\right)^{-1}\right] \\ &= \left[\frac{1-\gamma}{1+\gamma}, \frac{1+\gamma}{1-\gamma}\right]. \end{aligned}$$

Consequently,

$$\sigma(G) \subseteq \left[1 + \frac{1-\gamma}{1+\gamma}, 1 + \frac{1+\gamma}{1-\gamma}\right] = \left[\frac{2}{1+\gamma}, \frac{2}{1-\gamma}\right]. \quad \square$$

We are now ready to prove the following lemma.

Lemma A.4. *$\rho(M_1) \leq -\frac{2\eta}{1+\gamma} + \eta^2\left(X^2 + \frac{1}{b(1-\gamma)^2}\right)$ where M_1 is given in Eqn 6.*

Proof. By Weyl's theorem (Horn & Johnson, 1986, Theorem 4.3.1),

$$\begin{aligned}\rho(M_1) \quad &\leq \quad \rho\left(\eta^2 X^2 I\right) + \rho\left(\frac{\eta^2}{b} D^{-1} D^{-\top}\right) + \\ &\qquad \rho\left(-\eta\left(D^{-1} + D^{-\top}\right)\right).\end{aligned}$$

The lemma then follows immediately from Lemmas A.2 and A.3. $\square$

Lemma A.5. *Let M_2 be defined by Eqn 10 and suppose the step-size is given by Eqn 11, then*

$$\rho(M_2) \leq -\frac{(1-\gamma)^2}{(1+2\gamma)^2\left(X^2 + \frac{1}{b}\right)}.$$

Proof. Let $\alpha = \frac{\eta^2}{b}$ and $\beta = \eta^2 X^2 (1+\gamma)^2 - 2\eta$, then $M_2 = D^\top \Sigma \left(\alpha D D^\top + \beta I\right) \Sigma D$. It is known that

$$\rho(M_2) = \max_{\mathbf{v}_1 \neq \mathbf{0}} \frac{\mathbf{v}_1^\top M_2 \mathbf{v}_1}{\mathbf{v}_1^\top \mathbf{v}_1}.$$

Define $\mathbf{v}_2 = D\mathbf{v}_1$ and we have:

$$\begin{aligned}\rho(M_2) \quad &= \quad \max_{\mathbf{v}_2 \neq \mathbf{0}} \frac{\mathbf{v}_2 \Sigma \left(\alpha D D^\top + \beta I\right) \Sigma \mathbf{v}_2}{\mathbf{v}_2^\top D^{-\top} D^{-1} \mathbf{v}_2} \\ &\leq \quad \max_{\mathbf{v}_2 \neq \mathbf{0}} \frac{(1-\gamma)^2 \mathbf{v}_2 \Sigma \left(\alpha D D^\top + \beta I\right) \Sigma \mathbf{v}_2}{\mathbf{v}_2^\top \mathbf{v}_2},\end{aligned}$$

where the last step is due to Lemma A.2 and the fact that M_2 is negative-definite for $\eta \ll 1$. Similarly, we define $\mathbf{v}_3 = \Sigma \mathbf{v}_2$ and use the fact that

$$0 \leq \mathbf{v}_2^\top \mathbf{v}_2 = \mathbf{v}_3^\top \Sigma^{-2} \mathbf{v}_3 \leq \left(1 + \gamma(1+\gamma)\eta X^2\right)^2 \|\mathbf{v}_3\|^2$$

to obtain:

$$\begin{aligned}\rho(M_2) \quad &\leq \quad \max_{\mathbf{v}_3 \neq \mathbf{0}} \frac{(1-\gamma)^2 \mathbf{v}_3^\top \left(\alpha D D^\top + \beta I\right) \mathbf{v}_3^\top}{\left(1 + \gamma(1+\gamma)\eta X^2\right)^2 \mathbf{v}_3^\top \mathbf{v}_3} \\ &= \quad \frac{(1-\gamma)^2 \rho\left(\alpha D D^\top + \beta I\right)}{\left(1 + \gamma(1+\gamma)\eta X^2\right)^2} \\ &\leq \quad \frac{(1-\gamma)^2\left(\alpha(1+\gamma)^2 + \beta\right)}{\left(1 + \gamma(1+\gamma)\eta X^2\right)^2}.\end{aligned}$$

If we choose η as in Eqn 11, then the lemma follows immediately from the fact that

$$1 + \frac{\gamma X^2}{(1+\gamma)\left(X^2 + \frac{1}{b}\right)} \leq 1 + \frac{\gamma}{1+\gamma} = \frac{1+2\gamma}{1+\gamma}. \quad \square$$

Acknowledgment

We thank Michael Littman, Hengshuai Yao, and the anonymous reviewers for helpful comments that improved the presentation of the paper. The author is supported by NSF under grant IIS-0325281.

References

Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. *Proceedings of the Twelfth International Conference on Machine Learning (ICML-95)* (pp. 30–37).

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming.* Athena Scientific.

Boyan, J. A., & Moore, A. W. (1995). Generalization in reinforcement learning: Safely approximating the value function. *Advances in Neural Information Processing Systems 7 (NIPS-94)* (pp. 369–376).

Cesa-Bianchi, N., Long, P. M., & Warmuth, M. (1996). Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks*, *7*, 604–619.

Horn, R. A., & Johnson, C. R. (1986). *Matrix analysis.* Cambridge University Press.

Kivinen, J., & Warmuth, M. K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, *132*, 1–63.

Lagoudakis, M. G., & Parr, R. (2003). Least-squares policy iteration. *Journal of Machine Learning Research*, *4*, 1107–1149.

Munos, R. (2003). Error bounds for approximate policy iteration. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-03)* (pp. 560–567).

Precup, D., & Sutton, R. S. (1997). Exponentiated gradient methods for reinforcement learning. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97)* (pp. 272–277).

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming.* New York: Wiley-Interscience.

Schapire, R. E., & Warmuth, M. K. (1996). On the worst-case analysis of temporal-difference learning algorithms. *Machine Learning*, *22*, 95–122.

Schoknecht, R., & Merke, A. (2003). TD(0) converges provably faster than the residual gradient algorithm. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-03)* (pp. 680–687).

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

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

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