On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence

Nathaniel Korda NATHANIEL.KORDA@ENG.OX.AC.UK
MLRG, University of Oxford, UK.

Prashanth L.A. PRASHANTH.LA@INRIA.FR
INRIA Lille - Nord Europe, Team SequeL, FRANCE.

Abstract

We provide non-asymptotic bounds for the well-known temporal difference learning algorithm TD(0) with linear function approximators. These include high-probability bounds as well as bounds in expectation. Our analysis suggests that a step-size inversely proportional to the number of iterations cannot guarantee optimal rate of convergence unless we assume knowledge of the mixing rate for the Markov chain underlying the policy considered. This problem is alleviated by employing the well-known Polyak-Ruppert averaging scheme, leading to optimal rate of convergence without any knowledge of the mixing rate. Furthermore, we propose a variant of TD(0) with linear approximators that incorporates a centering sequence, and we establish that it exhibits an exponential rate of convergence in expectation. We demonstrate the usefulness of our bounds on two synthetic experimental settings.

1. Introduction

Many stochastic control problems can be cast within the framework of Markov decision processes (MDP). Reinforcement learning (RL) is a popular approach to solve MDPs, when the underlying transition mechanism is unknown. An important problem in RL is to estimate the *value function* V^π for a given stationary policy π. We focus on discounted reward MDPs with a high-dimensional state space $\mathcal{S}$. In this setting, one can only hope to estimate the value function approximately and this constitutes the *policy evaluation* step in several approximate policy iteration methods, for e.g. actor-critic algorithms (Konda & Tsitsiklis, 2003), (Bhatnagar et al., 2009).

Proceedings of the 31^{st} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

Temporal difference learning (Sutton & Barto, 1998) (TD(0)) is a well-known policy evaluation algorithm that is online and works with a single sample path obtained by simulating the underlying MDP. However, the classic TD(0) algorithm uses full-state representations (i.e. it stores an entry for each state $s \in \mathcal{S}$) and hence, suffers from the curse of dimensionality. A standard trick to alleviate this problem is to approximate the value function within a linearly parameterized space of functions, i.e., $V^\pi(s) \approx \theta^\intercal \phi(s)$. Here θ is a tunable parameter and $\phi(s)$ is a column feature vector with dimension $d << |S|$. This approximation allows for efficient implementation of TD(0) even on large state spaces.

The update rule for TD(0) that incorporates linear function approximators is as follows: Starting with an arbitrary θ_0,

$$\theta_{n+1} = \theta_n + \gamma_n \big(r(s_n, \pi(s_n)) + \beta \theta_n^\intercal \phi(s_{n+1}) - \theta_n^\intercal \phi(s_n)\big)\phi(s_n). \tag{1}$$

In the above, the quantities γ_n are *step sizes* that are chosen in advance and satisfy standard stochastic approximation conditions (see assumption (A5)). Further, $r(s, a)$ is the instantaneous reward in state s on choosing action a and $\beta \in (0, 1)$ is the discount factor.

In (Tsitsiklis & Van Roy, 1997), the authors establish that θ_n governed by (1) converges almost surely to the fixed point, θ^*, of the *projected Bellman equation* given by

$$\Phi\theta^* = \Pi \mathcal{T}^\pi(\Phi\theta^*). \tag{2}$$

In the above, $\mathcal{T}^\pi$ is the Bellman operator, Π is the orthogonal projection onto the linearly parameterized space within which we approximate the value function, and Φ is the feature matrix with rows $\phi(s)^\intercal, \forall s \in \mathcal{S}$ denoting the features corresponding to state $s \in \mathcal{S}$ (see Section 2 for details). Let P denote the transition probability matrix with components $p(s, \pi(s), s')$, r be a vector with components $r(s, \pi(s))$ and Ψ be a diagonal matrix whose diagonal forms the stationary distribution (assuming it exists) of the Markov chain

for the underlying policy π. Then, θ^* can be written as the solution to the following system of equations (see Section 6.3 of (Bertsekas, 2011))

$$A\theta^* = b, \text{ where } A = \Phi^\intercal \Psi (I - \beta P)\Phi \text{ and } b = \Phi^\intercal \Psi r. \quad (3)$$

Our aim is to derive non-asymptotic bounds on $\|\theta_n - \theta^*\|_2$, both in high-probability and in expectation, to quantify the rate of convergence of TD(0) with linear function approximators. To the best of our knowledge, there are no non-asymptotic bounds for TD(0) with function approximation, while there are asymptotic convergence and rate results available. Finite time analysis of TD(0) is challenging for two reasons:
(1) The asymptotic limit θ^* is the fixed point of the Bellman operator, which assumes that the underlying MDP is begun from the stationary distribution Ψ (whose influence is evident in (3)). However, the samples provided to the algorithm come from simulations of the MDP that are not begun from Ψ. This presents a difficulty for a finite time analysis, since we do not know exactly the number of steps after which mixing of the underlying Markov chain has occurred, and TD(0) starts to see samples from the stationary distribution. Moreover, an assumption on the mixing rate amounts to assuming (partial) knowledge of the transition dynamics of the Markov chain underlying the policy π.
(2) Standard results from stochastic approximation theory suggest that in order to obtain the optimal rate of convergence for a step size choice of $\gamma_n = c/(c+n)$, one has to chose the constant c carefully. In the case of TD(0), we derive this condition and point out the optimal choice for c requires knowledge of the mixing rate of the underlying Markov chain for policy π.
We handle the first problem by establishing that under a mixing assumption (the same as that used to establish asymptotic convergence for TD(0) in (Tsitsiklis & Van Roy, 1997)), the mixing error can be handled in the non-asymptotic bound. This assumption is broad enough to encompass a reasonable range of MDP problems. We alleviate the second problem by using iterate averaging. In both cases, we are obliged to include a projection step in order to bound the effect of the error due to sampling.

One inherent problem with iterative schemes that use a single sample to update the iterate at each time step, is that of variance. This is the reason why it is necessary to carefully choose the step-size sequence: too large and the variance will force divergence; too small and the algorithm will converge, but not to the solution intended. Indeed, iterate averaging is a technique that aims to allow for larger step-sizes, while producing the same overall rate of convergence (and we show that it succeeds in eliminating the necessity to know properties of the mixing time). A more direct approach is to center the updates, and this was pioneered recently for stochastic gradient descent in convex optimization (Johnson & Zhang, 2013). We propose a variant of TD(0) that uses this approach. We give a finite-time analysis, and show that the algorithm results in an exponential convergence rate, while not requiring a projection step to bound the iterates.

Our contributions can be summarized as follows:
(1) Under assumptions similar to (Tsitsiklis & Van Roy, 1997), we provide non-asymptotic bounds, both in high probability as well as in expectation, that quantify the convergence rate of TD(0) with function approximation.
(2) We also propose a variant of TD(0) that incorporates a centering sequence, that can easily be used in approximate policy iteration schemes, and we show that it converges faster than the regular TD(0) algorithm in expectation.

The key insights from our finite-time analysis are:
(1) With a step-size $\gamma_n = c/(c+n)$ where $(1-\beta)^2\mu c \in (1/2, \infty)$, we obtain the optimal rate of convergence of the order $O\left(1/\sqrt{n}\right)$ for the bound in expectation. Here μ is the smallest eigenvalue of the matrix $\Phi^\intercal \Psi \Phi$ (see Theorem 1).
(2) To obtain the optimal rate in the high-probability bound, the choice of c requires the knowledge of the mixing rate of the underlying Markov chain for policy π (see Theorem 1). As pointed out earlier, this is problematic as it implies (partial) knowledge about the transition dynamics of the MDP.
(3) With iterate averaging, one can get rid of the dependency of c on the mixing rate and still obtain the optimal rate of convergence, both in high probability as well as in expectation (see Theorem 2).
(4) For the centered variant of TD(0), we obtain an exponential convergence rate when the underlying Markov chain mixes fast (see Theorem 3).
(5) We illustrate the usefulness of our bounds on two simple synthetic experimental setups. In particular, using the step-sizes suggested by our bounds in Theorems 1–3, we are able to establish convergence empirically for TD(0), its averaging as well as centered variants.

Deriving convergence rate results for TD(0), especially of non-asymptotic nature, requires sophisticated machinery. We base our approach on that proposed in (Frikha & Menozzi, 2012) (and later expanded to include iterate averaging in (Fathi & Frikha, 2013)). We would like to remark that asymptotic convergence rate results for TD(λ) are available in (Konda, 2002). The authors establish there that TD(λ) converges asymptotically to a multi-variate Gaussian distribution $\mathcal{N}(0, \Sigma)$, where Σ is a covariance matrix that is a function of the matrix A. This rate result in the form of a central limit theorem holds true for TD(λ) when combined with iterate averaging, while the non-averaged case does not result in the optimal rate of convergence. Our results are consistent with this observation, as we establish from a finite time analysis that the non-averaged TD(0) can result in optimal convergence only if the step-size constant

c in $\gamma_n = c/(c+n)$ is set carefully (as a function of the mixing time), while one can get rid of this dependency and still obtain the optimal rate with iterate averaging. Least squares temporal difference methods are popular alternatives to the classic TD(λ). Asymptotic convergence rate results for LSTD(λ) and LSPE(λ), two popular least squares methods, are available in (Konda, 2002) and (Yu & Bertsekas, 2009), respectively. However, to the best of our knowledge, there are no concentration bounds that quantify the rate of convergence through a finite time analysis. A related work in this direction is the finite time bounds for LSTD in (Lazaric et al., 2010). However, the analysis there is under a fast mixing rate assumption, while we provide non-asymptotic rate results without making any such assumption. We note here that assuming a mixing rate implies partial knowledge of the transition dynamics of the MDP under a stationary policy and in typical RL settings, this information is not available.

2. TD(0) with Linear Approximation

We consider an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$. The aim is to estimate the value function V^π for any given stationary policy $\pi : \mathcal{S} \to \mathcal{A}$, where

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t r(s_t, \pi(s_t)) \mid s_0 = s\right]. \quad (4)$$

In the above, s_t denotes the state of the MDP at time t, $\beta \in (0,1)$ is the discount factor, and $r(s,a)$ denotes the instantaneous reward obtained in state s under action a. The expectation is with respect to the transition dynamics that specify the probability of transitioning from state s to s' under action a for any $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$ (we denote this probability by $p(s,a,s')$). It is well-known that the value function V^π is the solution to the fixed point relation $V = \mathcal{T}^\pi(V)$, where the Bellman operator $\mathcal{T}^\pi$ is defined as

$$\mathcal{T}^\pi(V)(s) := r(s,\pi(s)) + \beta \sum_{s'} p(s,\pi(s),s')V(s'), \quad (5)$$

TD(0) (Sutton & Barto, 1998) performs a fixed point-iteration using stochastic approximation: Starting with an arbitrary V_0, update

$$V_n(s_n) := V_{n-1}(s_n) + \gamma_n\big(r(s_n,\pi(s_n)) + \beta V_{n-1}(s_{n+1}) - V_{n-1}(s_n)\big), \quad (6)$$

where γ_n are step-sizes that satisfy standard stochastic approximation conditions.

As discussed in the introduction, while TD(0) algorithm is simple and provably convergent to the fixed point of $\mathcal{T}^\pi$ for any policy, it suffers from the curse of dimensionality associated with high-dimensional state spaces. A popular approach is to parameterize the value function using a linear function approximator, i.e. for every $s \in \mathcal{S}$, approximate $V^\pi(s) \approx \phi(s)^\intercal\theta$. Here $\phi(s)$ is a d-dimensional feature vector with $d << |\mathcal{S}|$, and θ is a tunable parameter. Incorporating function approximation, an update rule for the TD(0) analogous to (6) is given in (1). For the purposes of analysis, we also incorporate a projection step into the algorithm, so that, for all n, $\|\theta_n\| \le H$.

3. Concentration bounds for TD(0)

3.1. Assumptions

(A1) *Ergodicity*: The Markov chain induced by the policy π is irreducible and aperiodic. Moreover, there exists a stationary distribution $\Psi(=\Psi_\pi)$ for this Markov chain. Let $\mathbb{E}_\Psi$ denote the expectation w.r.t. this distribution.
(A2) *Bounded rewards*: $|r(s,\pi(s))| \le 1$, for all $s \in \mathcal{S}$.
(A3) *Linear independence*: The feature matrix Φ has full column rank. This assumption implies that the matrix $\Phi^\intercal\Psi\Phi$ has smallest eigenvalue $\mu > 0$.
(A4) *Bounded features*: $\|\phi(s)\|_2 \le 1$, for all $s \in \mathcal{S}$.
(A5) The step sizes satisfy $\sum_n \gamma_n = \infty$, and $\sum_n \gamma_n^2 < \infty$.
(A6) *Bounded mixing time*: There exists a non-negative function $B(\cdot)$ such that: For all $s_0 \in \mathcal{S}$ and $m \ge 0$,

$$\sum_{\tau=0}^{\infty} \|\mathbb{E}(r(s_\tau,\pi(s_\tau))\,\phi(s_\tau)|\,s_0) - \mathbb{E}_\Psi(r(s_\tau,\pi(s_\tau))\phi(s_\tau))\| \le B(s_0), \quad (7)$$

$$\sum_{\tau=0}^{\infty} \|\mathbb{E}[\phi(s_\tau)\phi(s_{\tau+m})^\intercal \mid s_0] - \mathbb{E}_\Psi[\phi(s_\tau)\phi(s_{\tau+m})^\intercal]\| \le B(s_0), \quad (8)$$

where the function $B(\cdot)$ is assumed to be bounded such that, for any $q > 1$, there exists a $K_q < \infty$ such that

$$\mathbb{E}[B^q(s) \mid s_0] \le K_q B^q(s_0). \quad (9)$$

For finite state settings, it is easy to establish the following mixing rate:

$$P(s_t = s \mid s_0) - \psi(s)| \le C\rho^t. \quad (10)$$

Here, it can be shown that $B(s_0) = \Theta(1/(1-\rho))$, where ρ is an unknown quantity that relates to the second eigenvalue of the transition probability matrix. However, for obtaining optimum convergence rates for TD(0) with step-size $\gamma_n = c/(c+n)$, the constant c has to be set using $B(s_0)$ (see Theorem 1 below).

The above assumptions are similar in nature to those made in (Tsitsiklis & Van Roy, 1997) for establishing asymptotic convergence of TD(0) with linear function approximators. In particular, assumptions (A1), (A3), (A5) and (A6) have exact counterparts in (Tsitsiklis & Van Roy, 1997), while (A2) and (A4) are simplified versions of corresponding boundedness assumptions in (Tsitsiklis & Van Roy, 1997).

3.2. Non-averaged case

Theorem 1. *Under (A1)-(A6), we have the following:*

(i) **Bound in expectation:** *With* $\gamma_n = \frac{c}{(c+n)}$, *where* c *is chosen such that* $(1-\beta)^2\mu c > 1/2$, *we have,*

$$\mathbb{E}\left\|\theta_n - \theta^*\right\|_2 \leq \frac{K_1(n)}{\sqrt{n+c}}, \text{ where}$$

$$K_1(n) = \frac{2\sqrt{c}\|\theta_0-\theta^*\|_2}{(n+c)^{2(1-\beta)^2\mu c-1/2}} + \frac{c(1-\beta)(3+6H)}{\sqrt{2(1-\beta)^2\mu c-1}},$$

(ii) **High-probability bound:** *With* $\gamma_n = \frac{c}{(c+n)}$, *where* c *is chosen such that* $(\mu(1-\beta)/2 + 3B(s_0))\, c \in (1,\infty)$, *we have, for any* $\delta > 0$

$$\mathbb{P}\left(\left\|\theta_n - \theta^*\right\|_2 \leq \frac{K_2(n)}{\sqrt{n+c}}\right) \geq 1-\delta,$$

where $K_2(n) := \frac{(1-\beta)c\sqrt{\ln(1/\delta)(1+9B(s_0)^2)}}{(\mu(1-\beta)/2+3B(s_0)^2)c-1} + K_1(n)$

Proof. See Section 5.1. □

Remark 1. *$K_1(n)$ and $K_2(n)$ above are $O(1)$, i.e., they can be upper bounded by a constant. Thus, one can indeed get the optimal rate of convergence of the order $O\left(1/\sqrt{n}\right)$ with a step-size $\gamma_n = \frac{c}{(c+n)}$. However, this rate is contingent upon on the constant c in the step-size being chosen correctly. This is problematic because the right choice of c requires the knowledge of eigenvalue μ for expectation bound and both μ and mixing bound $B(s_0)$ for high-probability bound. Knowing μ and $B(s_0)$ would imply knowledge about the transition probability matrix of the underlying Markov chain and the latter information is unavailable in a typical RL setting.*

3.3. Iterate Averaging

The idea here is to employ larger step-sizes $\gamma_n = (1-\beta)(c/(c+n))^\alpha$ and combine it with averaging of the iterates, i.e., $\bar{\theta}_{n+1} := (\theta_1 + \ldots + \theta_n)/n$. This principle was introduced independently by Ruppert (Ruppert, 1991) and Polyak (Polyak & Juditsky, 1992), for accelerating stochastic approximation schemes. The main advantage for us is that one obtains the optimal rate of convergence without any constraint on the step-size constant c:

Theorem 2. *Under (A1)-(A6), choosing* $\gamma_n = \frac{(1-\beta)}{2}\left(\frac{c}{c+n}\right)^\alpha$, *with* $\alpha \in (1/2, 1)$ *and* $c \in (0,\infty)$, *we have, for any* $\delta > 0$,

$$\mathbb{E}\left\|\bar{\theta}_{n+1} - \theta^*\right\|_2 \leq \frac{K_1^A(n)}{(n+c)^{\alpha/2}} \tag{11}$$

and $\mathbb{P}\left(\left\|\bar{\theta}_{n+1} - \theta^*\right\|_2 \leq \frac{K_2^A(n)}{(n+c)^{\alpha/2}}\right) \geq 1-\delta,$ (12)

where

$$K_1^A(n) = \frac{2\sqrt{c}\left\|\theta_0-\theta^*\right\|_2}{(n+c)^{\frac{1-\alpha}{2}}} + \frac{(1-\beta)c^\alpha(3+6H)}{(\mu c^\alpha(1-\beta)^2)^{\frac{\alpha(1+2\alpha)}{2(1-\alpha)}}},$$

$$K_2^A(n) = \frac{\left(\frac{2\alpha}{\mu\left[\frac{1-\beta}{2}+B(s_0)\right]c^\alpha} + \frac{2(3^\alpha)}{\alpha}\right)^{\frac{1}{2}}}{\mu\left[\frac{1}{2}+\frac{B(s_0)}{1-\beta}\right]n^{(1-\alpha)/2}} + K_1(n).$$

Proof. See Section 5.2. □

Remark 2. *The step-size exponent α can be chosen arbitrarily close to 1, resulting in a convergence rate of the order $O\left(1/\sqrt{n}\right)$. However although the constants $K_1^A(n)$ and $K_2^A(n)$ remain $O(1)$, there is a minor tradeoff here since a choice of α close to 1 would result in their bounding constants blowing up. One cannot choose c too large or too small for the same reasons.*

4. TD(0) with Centering (CTD)

4.1. The Algorithm

Let $X_n = (s_n, s_{n+1})$. Then, the TD(0) algorithm can be seen to perform the following fixed-point iteration:

$$\theta_n = \theta_{n-1} + \gamma_n f_{X_n}(\theta_n). \tag{13}$$

where $f_{X_n}(\theta) := (r(s_n,\pi(s_n)) + \beta\theta^\intercal\phi(s_{n+1}) - \theta^\intercal\phi(s_n))\phi(s_n)$. The limit of (13) is the solution, θ^*, of $F(\theta) = 0$, where $F(\theta) := \Pi T^\pi(\Phi\theta) - \Phi\theta$. The idea behind the CTD algorithm is to use reduce the variance of the increments $f_{X_n}(\theta_n)$, in order that larger step sizes can be used. This is achieved by choosing an extra iterate $\bar{\theta}_n$, centred over the previous θ_n, and using an increment of the form $f_{X_n}(\theta_n) - f_{X_n}(\bar{\theta}_n) + F(\bar{\theta}_n)$.

This approach is inspired by a recently proposed algorithm SVRG in (Johnson & Zhang, 2013) for optimizing a strongly-convex function that is a finite sum of smooth functions. However, the setting for TD(0) with function approximation that we have is considerably more complicated owing to the following reasons:
(i) Unlike (Johnson & Zhang, 2013), we do not have a function that is finite-sum of smooth functions. Instead, we have the value function which is an infinite (discounted) sum, with the individual functions making up the sum being made available in an online fashion (i.e. as new samples are generated from the simulation of the underlying MDP for policy π).
(ii) The centering term in SVRG directly uses $F(\cdot)$, which in our case is a limit function that is neither directly accessible nor can be simulated for any given θ.
(iii) Obtaining exponential convergence rate is also difficult owing to the fact that TD(0) does not initially see samples from the stationary distribution and there is an underlying

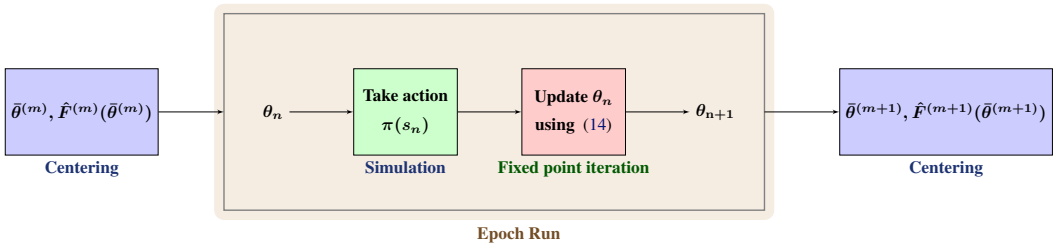


Figure 1. Illustration of centering principle in CTD algorithm.

mixing term that affects the rate.

(iv) Finally, there are extra difficulties owing to the fact that we have a fixed point iteration, while the corresponding algorithm in (Johnson & Zhang, 2013) is stochastic gradient descent (SGD).

The CTD algorithm that we propose overcomes the difficulties mentioned above and the overall scheme of this epoch-based algorithm is presented in Figure 1. At the start of the m^{th} epoch, a random iterate is picked from the previous epoch, i.e. $\bar{\theta}^{(m)} = \theta_{i_n}$, where i_n is drawn uniformly at random in $\{(m-1)M, \ldots, mM\}$. Thereafter, for the epoch length M, CTD performs the following iteration: Set $\theta_{mM} = \bar{\theta}^{(m)}$ and for $n = mM, \ldots, (m+1)M - 1$ update

$$\theta_{n+1} = \theta_n + \gamma\Big(f_{X_{i_n}}(\theta_n) - f_{X_{i_n}}(\bar{\theta}^{(m)}) + \hat{F}^{(m)}(\bar{\theta}^{(m)})\Big), \tag{14}$$

where $\hat{F}^{(m)}(\theta) := M^{-1}\sum_{i=(m-1)M}^{mM} f_{X_i}(\theta)$. Unlike TD(0), one can choose a large (constant) stepsize γ in (14). This choice in conjunction with iterate averaging via the choice of $\bar{\theta}^{(m)}$ results in an exponential convergence rate for CTD (see Remark 3 below).

4.2. Finite time bound

Theorem 3. *Assume (A1)-(A4) and (A6) and let θ^* denote the solution of $F(\theta) = 0$. Let the epoch length M of the CTD algorithm (14) be chosen such that $C_1 < 1$, where*

$$C_1 := ((2\mu\gamma M)^{-1} + \gamma d^2/2)/((1-\beta) - d^2\gamma/2))$$

***(i) Finite state spaces:** Here the Markov chain underlying policy π mixes fast (see (10)) and we obtain*[1]

$$\|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2 \le C_1^m \left(\|\Phi(\bar{\theta}^{(0)} - \theta^*)\|_\Psi^2\right) + CMC_2H(5\gamma+4)\max\{C_1, \rho^M\}^{(m-1)} \tag{15}$$

(ii) General state spaces:

$$\|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2 \le C_1^m \left(\|\Phi(\bar{\theta}^{(0)} - \theta^*)\|_\Psi^2\right) + C_2H(5\gamma+4)\sum_{k=1}^{m-1} C_1^{(m-2)-k} B_{(k-1)M}^{kM}(s_0), \tag{16}$$

where $C_2 = \gamma/(2M((1-\beta) - d^2\gamma/2))$, and $B_{(k-1)M}^{kM}$ is an upper bound on the partial sums $\sum_{i=(k-1)M}^{kM}(\mathbb{E}(\phi(s_i) \mid s_0) - \mathbb{E}_\Psi(\phi(s_i)))$ and $\sum_{i=(k-1)M}^{kM}(\mathbb{E}(\phi(s_i)\phi(s_{i+l}) \mid s_0) - \mathbb{E}_\Psi(\phi(s_i)\phi(s_{i+l})^\intercal))$, for $l = 0, 1$.

Proof. See Section 5.3. □

For finite state space settings, we obtain exponential convergence rate using (15), while for MDPs that do not mix exponentially fast, the second (mixing) term in (16) will dominate and decide the rate of the CTD algorithm.

Remark 3. *Combining the result in (15) with the bound in statement (4) of Theorem 1 in (Tsitsiklis & Van Roy, 1997), we obtain*

$$\|\Phi\bar{\theta}^{(m)} - V^\pi\|_\Psi \le \frac{1}{1-\beta}\|\Pi V^\pi - V^\pi\|_\Psi + C_1^{m/2}\left(\|\Phi(\bar{\theta}^{(0)} - \theta^*)\|_\Psi\right) + \sqrt{CC_2}\max\{C_1, \rho\}^{(m-1)/2}.$$

The first term on the RHS above is an artifact of function approximation, while the second and third terms reflect the convergence rate of the CTD algorithm.

Remark 4. *As a consequence of the fact that $(\bar{\theta}^{(m)} - \theta^*)^T I(\bar{\theta}^{(m)} - \theta^*) \le \frac{1}{\mu}(\bar{\theta}^{(m)} - \theta^*)^T\Phi^T\Psi\Phi(\bar{\theta}^{(m)} - \theta^*)$, one can obtain the following bound on the parameter error for CTD:*

$$\|\bar{\theta}^{(m)} - \theta^*\|_2 \le (1/\mu)\Bigg(C_1^m\left(\|\Phi(\bar{\theta}^{(0)} - \theta^*)\|_\Psi^2\right) + C_2H(5\gamma+4)\sum_{k=1}^{m-1} C_1^{(m-2)-k} B_{(k-1)M}^{kM}(s_0)\Bigg).$$

Comparing the above bound with those in Theorems 1–2, we can infer that CTD exhibits an exponential convergence rate of order $O(C_1^m)$, while TD(0) with/without averaging can converge only at a sublinear rate of order $O(n^{-1/2})$.

[1] For any $v \in \mathbb{R}^d$, we take $\|v\|_\Psi := \sqrt{v^\intercal\Psi v}$.

5. Analysis

5.1. Non-averaged case: Proof of Theorem 1

We split the analysis in two, first considering the bound in expectation, and second the bound in high probability. Both bounds involve a martingale decomposition, the former of the iteration (1), and the latter directly of the centered error.

Bound in expectation First we state a theorem bounding the expected error for general step-size sequences:

Theorem 4. *Under (A1)-(A5), we have,*

$$\mathbb{E}\left\|\theta_n-\theta^*\right\|_2 \leq \Bigg[\underbrace{2\exp(-(1-\beta)\mu\Gamma_n)\left\|\theta_0-\theta^*\right\|_2}_{\textit{initial error}}.+ \underbrace{\Bigg(\sum_{k=1}^{n-1}(3+6H)^2\gamma_{k+1}^2\exp(-2(1-\beta)\mu(\Gamma_n-\Gamma_{k+1})\Bigg)^{\frac{1}{2}}}_{\textit{sampling and mixing error}}\Bigg],$$

where $\Gamma_k := \sum_{i=1}^k \gamma_i$ *and* $\left\|\theta_n\right\|_2 \leq H, \forall n$.

The initial error depends on the initial point θ_0 of (1). The mixing error arises due to the fact that we don't supply samples to the TD(0) algorithm from the stationary distribution of the underlying Markov chain for policy π, while the sampling error arises out of a martingale difference sequence. These error components can be clearly seen from the first step of the proof in the sketch provided below.

Proof sketch of Theorem 4. Recall that $f_{X_n}(\theta) := [r(s_n,\pi(s_n)) + \beta\theta_{n-1}^\intercal\phi(s_{n+1}) - \theta_{n-1}^\intercal\phi(s_n)]\phi(s_n)$. The first step is to rewrite the recursion (1) as follows:

$$\theta_{n+1} = \theta_n + \gamma_n\left[\mathbb{E}_\Psi(f_{X_n}(\theta_n)) + \epsilon_n + \Delta M_n\right], \quad (17)$$

where $\epsilon_n := \mathbb{E}(f_{X_n}(\theta_n) \mid \mathcal{F}_n) - \mathbb{E}_\Psi(f_{X_n}(\theta_n))$ is the mixing error term, while $\Delta M_n := f_{X_n}(\theta_n) - \mathbb{E}(f_{X_n}(\theta_n) \mid \mathcal{F}_n)$ is a martingale sequence (recall that $\mathcal{F} = \{\mathcal{F}_n := \sigma(\theta_1,\ldots,\theta_n)\}$ is the filtration generated by the iterates $\theta_n, n \geq 0$).

The next step is to unroll (17) as follows:

$$\begin{aligned} z_{n+1} =& (I-\gamma_n A)z_n + \gamma_n(\epsilon_n + \Delta M_n) \\ =& \Pi_n z_0 + \sum_{k=1}^n \gamma_k \Pi_n \Pi_k^{-1}(\epsilon_k + \Delta M_k), \end{aligned}$$

where $A := \Phi^\intercal\Psi(I-\beta P)\Phi$ and $\Pi_n := \prod_{k=1}^n (I-\gamma_k A)$.

The mixing error can be bounded using (A6). However, bounding $\left\|\Delta M_n\right\|_2$ is tricky since it requires the iterate θ_n to be bounded as well and the latter is complicated owing to the form of TD(0) update (1). We workaround this by assuming that the iterate is projected, i.e., $\left\|\theta_n\right\|_2 \leq H$. The reader is referred to Appendix B.1 for the detailed proof. □

High probability bound Now we state a theorem bounding the error with high probability for general step-sizes:

Theorem 5. *Under (A1)-(A6), we have,*

$$P(\left\|\theta_n-\theta^*\right\|_2 - \mathbb{E}\left\|\theta_n-\theta^*\right\|_2 \geq \epsilon) \leq e^{-\epsilon^2\left(2\sum_{i=1}^n L_i^2\right)^{-1}},$$

where $L_i := \gamma_i[\prod_{j=i+1}^n(1-2\gamma_j(\mu(1-\beta-\frac{\gamma_j}{2}) + [1+\beta(3-\beta)]B(s_0)))]^{1/2}$.

This theorem decomposes the problem of bounding $\left\|\theta_n-\theta^*\right\|_2$ into bounding the deviation from its mean $\mathbb{E}\left\|\theta_n-\theta^*\right\|_2$ in high probability and the size of the mean itself. The latter is already bounded by Theorem 4.

Proof Sketch of Theorem 5. Recall that $z_n := \theta_n - \theta^*$.
Step 1: We rewrite $\left\|z_n\right\|_2^2 - E\left\|z_n\right\|_2^2$ as a telescoping sum of martingale differences as follows:

$$\left\|z_n\right\|_2 - \mathbb{E}\left\|z_n\right\|_2 = \sum_{i=1}^n g_i - \mathbb{E}[g_i\,|\mathcal{F}_{i-1}] = \sum_{i=1}^n D_i,$$

where $D_i := g_i - \mathbb{E}[g_i\,|\mathcal{F}_{i-1}]$, $g_i := \mathbb{E}[\left\|z_n\right\|_2\,|\theta_i]$.
Step 2: We establish that the functions g_i, conditioned on $\mathcal{F}_{i-1}$, are Lipschitz continuous in $f_{X_i}(\theta_{i-1})$ with constants L_i.
Step 3: We invoke a standard martingale concentration bound using the L_i-Lipschitz property of the g_i functions and the assumption (A3) to obtain:

$$P(\left\|z_n\right\|_2 - \mathbb{E}\left\|z_n\right\|_2 \geq \epsilon) \leq \exp\left(\frac{\alpha\lambda^2}{2}\sum_{i=1}^n L_i^2 - \lambda\epsilon\right).$$

The result follows by optimizing over λ. The detailed proof is provided in Appendix B.2. □

Rates In order to obtain the rates and constants presented in Theorem 1, we specialize Theorems 4 5 for the choice of step-size sequence, $\gamma_n = (1-\beta)c/(c+n)$.

Supposing that c is chosen so that $(1-\beta)^2\mu c > 1/2$, then from the bound in expectation in Theorem 4 we have:

$$\begin{aligned} &\sum_{k=1}^{n-1}\gamma_{k+1}^2\exp(-2(1-\beta)\mu(\Gamma_n-\Gamma_{k+1})) \\ &\quad\leq \frac{(1-\beta)^2c^2}{(n+c)^{2(1-\beta)^2\mu c}}\sum_{k=1}^n(c+k)^{-(2-2(1-\beta)^2\mu c)} \\ &\quad\leq \frac{(1-\beta)^2c^2}{1-2(1-\beta)^2\mu c}\frac{1}{c+n} \end{aligned}$$

where, in the last inequality, we have compared the sum with an integral. Similarly

$$\exp(-(1-\beta)\mu\Gamma_n) \leq \left(\frac{c}{n+c}\right)^{2(1-\beta)^2\mu c} \leq \left(\frac{c}{n+c}\right)^{\frac{1}{2}}.$$

So we have

$$\mathbb{E}\left\|\theta_n - \theta^*\right\|_2 \leq \frac{2\sqrt{c}\left\|\theta_0 - \theta^*\right\|_2 + c(1-\beta)(3+6H)}{(c+n)^{1/2}},$$

and the result in Theorem 1 now follows.

For the bound in high probability, a calculation (see the Appendix B.3) shows that choosing c so that $(\mu(1-\beta)/2 + 3B(s_0))\, c \in (1,\infty)$, we have

$$\sum_{i=1}^n L_i^2 \leq \frac{(1-\beta)^2 c^2}{((\mu(1-\beta)/2 + 3B(s_0))\, c - 1)}(n+c)^{-1}$$

and the result in Theorem 1 follows.

5.2. Iterate Averaging: Proof of Theorem 2

In order to prove the results in Theorem 2 we again consider the case of a general step sequence. Recall that $\bar{\theta}_{n+1} := (\theta_1 + \ldots + \theta_n)/n$ and let $z_n = \bar{\theta}_{n+1} - \theta^*$. First, we directly give a bound on the error in high probability for the averaged iterates (the bound in expectation can be obtained directly from the bound in Theorem 4):

Theorem 6. *Under (A1)-(A6) we have, for all $\epsilon \geq 0$ and $\forall n \geq 1$,*

$$P(\|z_n\|_2 - \mathbb{E}\|z_n\|_2 \geq \epsilon) \leq e^{-\epsilon^2\left(2\sum_{i=1}^n L_i^2\right)^{-1}},$$

where $L_i := \frac{\gamma_i}{n}(1 + \sum_{l=i+1}^{n-1}\prod_{j=i}^{l}(1 - 2\gamma_j(\mu(1-\beta-\frac{\gamma_j}{2}) + [1+\beta(3-\beta)]B(s_0))))$.

Rates In order to obtain the rates in Theorem 2, we again specialise the general results to the choice of step-size: $\gamma_n = (1-\beta)(c/(c+n))^\alpha$. To bound the expected error we directly average the errors of the non-averaged iterates:

$$\mathbb{E}\left\|\bar{\theta}_{n+1} - \theta^*\right\|_2 \leq \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left\|\theta_k - \theta^*\right\|_2,$$

and directly applying the bounds in expectation given in Theorem 4.

For the rate of the bound in high probability we need to specialise the bounds for the bound in expectation in Theorem 6 for the new choice of step-size sequence. In particular, we compute the value of the Lipschitz L_i constants for our choice of step-sizes to obtain:

$$\sum_{i=1}^n L_i^2 \leq \frac{\left[\frac{2\alpha}{\mu\left[\frac{1-\beta}{2} + B(s_0)\right]c^\alpha} + \frac{5^\alpha}{\alpha}\right]^2}{\mu^2\left[\frac{1}{2} + \frac{B(s_0)}{1-\beta}\right]^2}\frac{1}{n} \tag{18}$$

The above inequality together with Theorem 6 completes the result. The reader is referred to Appendix C for detailed proofs.

5.3. TD(0) with centering: Proof of Theorem 3

Step 1: One-step expansion of the recursion (14). Let

$$\bar{f}_{X_{i_n}}(\theta_n) := f_{X_{i_n}}(\theta_n) - f_{X_{i_n}}(\bar{\theta}^{(m)}) + \mathbb{E}(f_{X_{i_n}}(\bar{\theta}^{(m)}) \mid \mathcal{F}_n).$$

Then, using (14), we obtain

$$\begin{aligned}&\|\theta_{n+1} - \theta^*\|_2^2 \\ \leq &\|\theta_n - \theta^*\|_2^2 + 2\gamma(\theta_n - \theta^*)^\intercal \bar{f}_{X_{i_n}}(\theta_n) + \gamma^2\left\|\bar{f}_{X_{i_n}}(\theta_n)\right\|_2^2\end{aligned} \tag{19}$$

Step 2: Bounding the variance of centred updates.

$$\begin{aligned}&\mathbb{E}\left(\left\|\bar{f}_{X_{i_n}}(\theta_n)\right\|_2^2\middle|\mathcal{F}_n\right) \leq \mathbb{E}\Bigg(\left\|f_{X_{i_n}}(\theta_n) - f_{X_{i_n}}(\theta^*)\right\|_2^2 \\ &+ \left\|f_{X_{i_n}}(\bar{\theta}^{(m)}) - f_{X_{i_n}}(\theta^*) - \mathbb{E}(f_{X_{i_n}}(\bar{\theta}^{(m)}))\right\|_2^2\middle|\mathcal{F}_n\Bigg) \\ &\leq \mathbb{E}\left(\left\|f_{X_{i_n}}(\theta_n) - f_{X_{i_n}}(\theta^*)\right\|_2^2\middle|\mathcal{F}_n\right) \\ &+ \mathbb{E}\left(\left\|f_{X_{i_n}}(\bar{\theta}^{(m)}) - f_{X_{i_n}}(\theta^*)\right\|_2^2\middle|\mathcal{F}_n\right),\end{aligned} \tag{20}$$

where we have used that for any random variable ξ, $\mathbb{E}(\|x - \mathbb{E}x\|^2) \leq \mathbb{E}(\|x\|^2)$. Let $\epsilon_n(\theta) = \mathbb{E}\left(\left\|f_{X_{i_n}}(\theta) - f_{X_{i_n}}(\theta^*)\right\|_2^2\middle|\mathcal{F}_n\right) - \mathbb{E}_\Psi(\left\|f_{X_{i_n}}(\theta) - f_{X_{i_n}}(\theta^*)\right\|_2^2)$. The second term in $\epsilon_n(\theta)$ can be bounded as follows: For any θ, we have

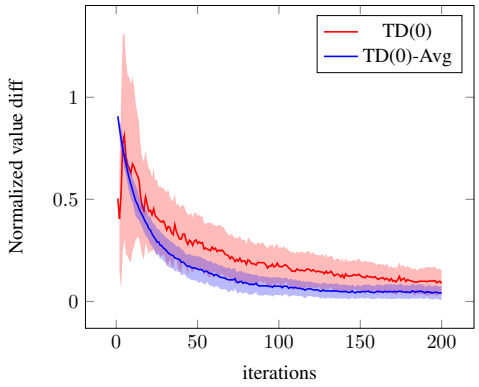
$$\begin{aligned}&\mathbb{E}_\Psi(\left\|f_{X_{i_n}}(\theta) - f_{X_{i_n}}(\theta^*)\right\|_2^2) \\ &= (\theta-\theta^*)^\intercal(\Phi^\intercal(I-\beta P)\Psi\Phi)^\intercal\Phi^\intercal\Psi(I-\beta P)\Phi(\theta-\theta^*) \\ &\leq (\theta-\theta^*)^\intercal(\Phi^\intercal\Psi\Phi)^\intercal\Phi^\intercal\Psi\Phi(\theta-\theta^*) \leq d^2\|\Phi(\theta-\theta^*)\|_\Psi^2.\end{aligned}$$

In the final inequality, we have used $\|\Phi^\intercal\Psi\Phi\|_2 \leq d^2$. Plugging the above in (20), we obtain

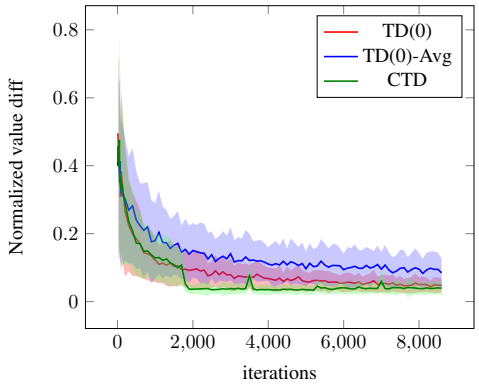
$$\begin{aligned}\mathbb{E}\left(\left\|\bar{f}_{X_{i_n}}(\theta_n)\right\|_2^2\middle|\mathcal{F}_n\right) &\leq \epsilon_n(\theta_n) + \epsilon_n(\bar{\theta}^{(m)}) \\ &d^2\left(\|\Phi(\theta_n-\theta^*)\|_\Psi^2 + \|\Phi(\bar{\theta}^{(m)}-\theta^*)\|_\Psi^2\right)\end{aligned} \tag{21}$$

Step 3: Analysis for a particular epoch.
Let $e_n(\theta) := \mathbb{E}\left[f_{X_{i_n}}(\theta)\middle|\mathcal{F}_n\right] - \mathbb{E}_\Psi\left[f_{X_{i_n}}(\theta)\right]$. Notice that $\mathbb{E}_\Psi\left[f_{X_{i_n}}(\theta^*)\right] = 0$ and hence $\mathbb{E}_\Psi\left[f_{X_{i_n}}(\theta_n)\right] = \mathbb{E}_\Psi((\beta\phi(s_{n+1}) - \phi(s_n))\phi(s_n)^\intercal)(\theta_n - \theta^*) = (\Phi^\intercal\Psi(I - \beta P)\Phi)(\theta_n - \theta^*)$. Thus, we can rewrite (19) as follows:

$$\begin{aligned}&\mathbb{E}\left(\|\theta_{n+1} - \theta^*\|_2^2\middle|\mathcal{F}_n\right) \leq \|\theta_n - \theta^*\|_2^2 \\ &- 2\gamma\left((1-\beta) - \frac{d^2\gamma}{2}\right)\|\Phi(\theta_n - \theta^*)\|_\Psi^2 \\ &+ \gamma^2 d^2\|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2 + 2\gamma(\theta_n - \theta^*)^\intercal e_n(\theta_n) \\ &+ \gamma^2(\epsilon_n(\theta_n) + \epsilon_n(\bar{\theta}^{(m)}))\end{aligned} \tag{22}$$



(a) Example 1



(b) Example 2

Figure 2. Empirical illustration of TD(0), TD(0) with averaging and CTD algorithms. The normalised value difference is defined to be $\|\Phi(\theta_n - \theta^*)\|_\Psi / \|\Phi(\theta^*)\|_\Psi$.

Notice that $(\bar{\theta}^{(m)} - \theta^*)^\intercal I(\bar{\theta}^{(m)} - \theta^*) \leq \mu^{-1}(\bar{\theta}^{(m)} - \theta^*)^\intercal \Phi^\intercal \Psi \Phi(\bar{\theta}^{(m)} - \theta^*)$ and hence we obtain the following by unrolling (22) and then setting $\theta_0 = \bar{\theta}^{(m)}$:

$$
\begin{aligned}
&2\gamma M((1-\beta) - \frac{d^2\gamma}{2})\mathbb{E}\left(\|\Phi(\bar{\theta}^{(m+1)} - \theta^*)\|_\Psi^2 \middle| \mathcal{F}_{mM}\right) \\
\leq\ & \left(\frac{1}{\mu} + M\gamma^2 d^2\right)\|\bar{\theta}^{(m)} - \theta^*\|_2^2 \\
&+ \gamma^2 \sum_{n=mM}^{(m+1)M-2} \mathbb{E}\left(\epsilon_n(\theta_n) + \epsilon_n(\bar{\theta}^{(m)}) \middle| \mathcal{F}_{mM}\right) \\
&+ \mathbb{E}\left(2\gamma \sum_{n=mM}^{(m+1)M-1} (\theta_n - \theta^*)^\intercal e_n(\theta_n) \middle| \mathcal{F}_{mM}\right)
\end{aligned}
$$

Step 4: Combining across epochs.
Finally, we obtain (16) by unrolling (across epochs) the final recursion in the previous step.

6. Numerical Experiments

We test the performance of TD(0), TD(0) with averaging and CTD algorithms (see Appendix E for full description).

Example 1. This is a two-state toy example, which is borrowed from (Yu & Bertsekas, 2009). The setting has the transition structure $P = [0.2, 0.8; 0.3, 0.7]$ and the rewards given by $r(1, j) = 1, r(2, j) = 2$, for $j = 1, 2$. The features are one-dimensional, i.e., $\Phi = (1 \quad 2)^\intercal$.

Fig. 2(a) presents the results obtained on this example. For setting the step-sizes of TD(0), we used the guideline from Theorem 1. Note that this results in convergence for TD(0), with the caveat that setting the step-size constant c requires knowledge of underlying transition structure through μ. It is evident that TD(0) with averaging gives performance on par with TD(0) and unlike TD(0), the setting of c is not constrained here. Given that convergence is rapid for TD(0) on this example, we do not plot CTD in Fig 2(a) as the epoch length suggested by Theorem 3 is 100 and this is already enough for TD(0) itself to converge. CTD resulted in a normalized value difference of about 0.03 on this example, but the effect of averaging across epochs for CTD will be seen better in the next example.

Example 2. Here the number of states are 100, the transitions are governed by a random stochastic matrix and the rewards are random and bounded between 0 and 1. Features are 3-dimensional and are picked randomly in $(0, 1)$. The results obtained for the three algorithms are presented in Fig. 2(b). It is evident that all algorithms converge, with CTD showing the lowest variance. As in example 1, the setting parameters for TD(0) was dictated by Theorem 1, while for CTD, the step-size and epoch length were set such that the constant C_1 in Theorem 3 is < 1.

7. Conclusions

TD(0) with linear function approximators is a well-known policy evaluation algorithm. While asymptotic convergence rate results are available for this algorithm, there are no finite-time bounds that quantify the rate of convergence. In this paper, we derived non-asymptotic bounds, both in high-probability as well as in expectation. From our results, we observed that iterate averaging is necessary to obtain the optimal $O\left(1/\sqrt{n}\right)$ rate of convergence. This is because, to obtain the optimal rate with the classic step-size choice $\propto 1/n$, it is necessary to know the mixing rate of the underlying Markov chain. We also proposed a fast variant of TD(0) that incorporates a centering sequence and established that the rate of convergence of this algorithm is exponential. We established the practicality of our bounds by using them to guide the step-size choices in two synthetic experimental setups.

References

Bertsekas, Dimitri P. Approximate dynamic programming. 2011.

Bhatnagar, S., Sutton, R., Ghavamzadeh, M., and Lee, M. Natural actor-critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Fathi, Max and Frikha, Noufel. Transport-entropy inequalities and deviation estimates for stochastic approximation schemes. *arXiv preprint arXiv:1301.7740*, 2013.

Frikha, Noufel and Menozzi, Stéphane. Concentration Bounds for Stochastic Approximations. *Electron. Commun. Probab.*, 17:no. 47, 1–15, 2012.

Johnson, Rie and Zhang, Tong. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 315–323, 2013.

Konda, Vijay R. *Actor-Critic Algorithms*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2002.

Konda, Vijay R and Tsitsiklis, John N. On Actor-Critic Algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

Lazaric, Alessandro, Ghavamzadeh, Mohammad, and Munos, Rémi. Finite-sample analysis of lstd. In *ICML*, pp. 615–622, 2010.

Polyak, Boris T and Juditsky, Anatoli B. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Ruppert, David. Stochastic approximation. *Handbook of Sequential Analysis*, pp. 503–529, 1991.

Sutton, Richard S and Barto, Andrew G. *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press, 1998.

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

Yu, Huizhen and Bertsekas, Dimitri P. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54 (7):1515–1531, 2009.