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 (partial) knowledge of the stationary distribution for the Markov chain underlying the policy considered. We also provide bounds for the iterate averaged TD(0) variant, which gets rid of the step-size dependency while exhibiting the optimal rate of convergence. Furthermore, we propose a variant of TD(0) with linear approximators that incorporates a centering sequence, and 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, e.g. actor-critic algorithms (Konda & Tsitsiklis, 2003), (Bhatnagar et al., 2009).

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

Temporal difference learning is a well-known policy evaluation algorithm that is both 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*, chosen in advance ,and satisfying standard stochastic approximation conditions (see assumption (A5)). Further, $r(s, a)$ is the reward recieved in state s on choosing action a, $\beta \in (0, 1)$ is a discount factor, and s_n is the state of the MDP at time n.

Asymptotic convergence of TD(0). 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 more details). Let P denote the transition probability matrix with components $p(s, \pi(s), s')$ that denote the probability of transitioning from state s to s' under the action

$\pi(s)$. Let 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 work. We 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 is challenging because:
(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 is a problem 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 the distribution of the samples that TD(0) sees has become the stationary distribution. Moreover, an assumption on this 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.

Variance reduction. 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 stationary distribution of the underlying Markov chain). A more direct approach is to center the updates, and this was pioneered recently for solving batch problems via stochastic gradient descent in convex optimization (Johnson & Zhang, 2013). We propose a variant of TD(0) that uses this approach, though our setting is considerably more complicated as samples arrive online and the function being optimized is not accessible directly.

Our contributions can be summarized as follows:
(1) ***Concentration bounds.*** Under assumptions similar to (Tsitsiklis & Van Roy, 1997), we provide non-asymptotic bounds, both in high probability as well as in expectation and these quantify the convergence rate of TD(0) with function approximation.
(2) ***Centered TD.*** We propose a variant of TD(0) that incorporates a centering sequence 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) Choosing $\gamma_n = \frac{c_0 c}{(c+n)}$, with $c_0 < \mu(1-\beta)/(2(1+\beta)^2)$ and c such that $\mu(1-\beta)c_0 c > 1$, we obtain the optimal rate of convergence of the order $O\left(1/\sqrt{n}\right)$, both in high-probability as well as in expectation. Here μ is the smallest eigenvalue of the matrix $\Phi^\intercal \Psi \Phi$ (see Theorem 1). However, obtaining this rate is problematic as it implies (partial) knowledge (via μ) of the transition dynamics of the MDP.
(2) With iterate averaging, one can get rid of the step-size dependency and still obtain the optimal rate of convergence, both in high probability as well as in expectation (see Theorem 2).
(3) For the centered variant of TD(0), we obtain an exponential convergence rate when the underlying Markov chain mixes fast (see Theorem 3).
(4) 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), and both its averaging, as well as centered variants.

Related work. Concentration bounds for general stochastic approximation schemes have been derived in (Frikha & Menozzi, 2012) and later expanded to include iterate averaging in (Fathi & Frikha, 2013). Unlike the aforementioned reference, deriving convergence rate results for TD(0), especially of non-asymptotic nature, requires sophisticated machinery as it involves Markov noise that impacts the mixing rate of the underlying Markov chain. An asymptotic normality result for TD(λ) is available in (Konda, 2002). The authors establish there that TD(λ) converges asymptotically to a multi-variate Gaussian distribution with a covariance matrix that depends on A (see (3)). This rate result 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 a certain quantity that depends on the stationary distribution - see (A3) below), 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

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

Recall that $\beta \in (0,1)$ is the discount factor, s_t denotes the state of the MDP at time t, and $r(s,a)$ denotes the reward obtained in state s under action a. The expectation in (4) is taken with respect to the transition dynamics P. It is well-known that 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, and popular method to allieviate this 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 TD(0) analogous to (6) is given in (1).

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) ***Combined step size and mixing assumption***: There exists a non-negative function $B'(\cdot)$ such that: For all $s \in \mathcal{S}$ and $m \ge 0$,

$$\sum_{\tau=0}^{\infty} e^{3(1+\beta)\sum_{j=1}^{\tau-1}\gamma_\tau} \|\mathbb{E}(r(s_\tau,\pi(s_\tau))\phi(s_\tau) \mid s_0 = s) - \mathbb{E}_\Psi(r(s_\tau,\pi(s_\tau))\phi(s_\tau))\| \le B'(s),$$

$$\sum_{\tau=0}^{\infty} e^{3(1+\beta)\sum_{j=1}^{\tau-1}\gamma_\tau} \|\mathbb{E}(\phi(s_\tau)\phi(s_{\tau+m})^\intercal \mid s_0 = s) - \mathbb{E}_\Psi(\phi(s_\tau)\phi(s_{\tau+m})^\intercal)\| \le B'(s),$$

(A6') ***Uniform mixing bound***: (A6) holds, and there exists a constant B' that is an uniformly bound on $B(s), \forall s \in \mathcal{S}$.

In comparison to the assumptions in (Tsitsiklis & Van Roy, 1997), (A1), (A3), (A5) have exact counterparts in (Tsitsiklis & Van Roy, 1997), while (A2), (A4) and (A6) are simplified versions of the corresponding boundedness assumptions in (Tsitsiklis & Van Roy, 1997).

Remark 1. ***(Geometric ergodicity)*** *A Markov chain is mixing at a geometric rate if*

$$P(s_t = s \mid s_0) - \psi(s)| \le C\rho^t. \quad (7)$$

For finite state space settings, the above condition holds and hence (A7) is easily satisfied. Moreover, $B' = \Theta\left(1/(1-(1-\rho)^{1-\epsilon}\right)$, for any $\epsilon > 0$. Here ρ is an unknown quantity that relates to the second eigenvalue of the transition probability matrix. See Chapters 15 and 16 of (Meyn & Tweedie, 2009) for a detailed treatment of the subject matter.

3.2. Non-averaged case

Theorem 1. *Under **(A1)-(A6)**, choosing $\gamma_n = \frac{c_0 c}{(c+n)}$, with $c_0 < \mu(1-\beta)/(2(1+\beta)^2)$ and c such that $\mu(1-\beta)c_0 c > 1$, we have,*

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

*In addition, assuming **(A6')**, 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

$$\boldsymbol{K_1(n)} := \left(\frac{c(\|\theta_0-\theta^*\|_2+C)}{(n+c)^{\mu(1-\beta)c_0c-1}} + \frac{(1+\|\theta^*\|_2)c_0^2c^2+Cc_0c}{\mu(1-\beta)c_0c-1}\right)^{\frac{1}{2}},$$

$$\boldsymbol{K_2(n)} := \frac{c_0cB'(2[2+c_0c][1+\beta(3-\beta)]\ln(1/\delta))^{\frac{1}{2}}}{(\mu(1-\beta)c_0c-1)^{\frac{1}{2}}} + K_1(n),$$

and $C := 6dB(s_0)\left(\frac{\|\theta_0\|_2+d+\|\theta^*\|_2}{1-\beta}\right)^2$.

Proof. See Section 5.1. □

Remark 2. *$K_1(n)$ 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 and knowing μ would imply knowledge about the transition probability matrix of the underlying Markov chain. The latter information is unavailable in a typical RL setting. The next section derives bounds for the iterate averaged variant that overcomes this problematic step-size dependency.*

3.3. Iterate Averaging

The idea here is to employ larger step-sizes 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 following theorem establishes that iterate averaging results in the optimal rate of convergence without any step-size dependency:

Theorem 2. *Under **(A1)-(A6)**, choosing $\gamma_n = c_0\left(\frac{c}{c+n}\right)^\alpha$, with $\alpha \in (1/2, 1)$ and $c \in (0, \infty)$, we have, for all $n > n_0 := (c\mu(1-\beta)/(2c_0(1+\beta)^2))^{-1/\alpha}$,*

$$\mathbb{E}\left\|\bar{\theta}_n - \theta^*\right\|_2 \leq \frac{K_1^{IA}(n)}{(n+c)^{\alpha/2}}.$$

*In addition, assuming **(A6)'**, we have, for any $\delta > 0$,*

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

where $\boldsymbol{K_1^{IA}(n)} :=$

$$\frac{((1+dc_0c^\alpha(c+n_0)^{1-\alpha})e^{(1+\beta)c_0c^\alpha(c+n_0)^{1-\alpha}}+\|\theta^*\|+C)C''}{(n+c)^{(1-\alpha)/2}}$$
$$+ \frac{n_0\left[(1+dc_0c^\alpha(c+n_0)^{1-\alpha})e^{(1+\beta)c_0c^\alpha(c+n_0)^{1-\alpha}}+\|\theta^*\|\right]}{n^{1-\alpha/2}}$$
$$+ c^\alpha c_0\left[1+\|\theta^*\|_2^{\frac{1}{2}} + \left(\frac{C}{c^\alpha c_0}\right)^{\frac{1}{2}}\right]\left(\frac{\mu(1-\beta)c_0c^\alpha}{1-\alpha}\right)^{-\frac{\alpha+2\alpha^2}{2(1-\alpha)}},$$

$$\boldsymbol{K_2^{IA}(n)} := \frac{4\sqrt{(1+C')B'}}{\mu(1-\beta)}\frac{C'''}{n^{(1-\alpha)/2}} + K_1^{IA}(n-n_0),$$

$$C' := \left(3^\alpha + \left[\frac{4\alpha}{\mu(1-\beta)c_0c^\alpha} + \frac{2^\alpha}{\alpha}\right]^2\right)^{\frac{1}{2}},\ C'' := \sum_{k=1}^{\infty} k^{-2\alpha},$$

and $C''' := \sum_{k=1}^{\infty} e^{-\frac{\mu c^\alpha(1-\beta)c_0}{2(1-\alpha)}((n+c)^{1-\alpha}-((c+n_0)^{1-\alpha})}$.

Proof. See Section 5.2. □

Remark 3. *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 $K_1^{IA}(n)$ and $K_2^{IA}(n)$ 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)

CTD is a *control variate* solution to reduce the variance of the updates of normal TD(0). This is achieved by adding a zero-mean, ***centering*** term to the TD(0) update.

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

$$\theta_n = \theta_{n-1} + \gamma_n f_{X_n}(\theta_n). \tag{8}$$

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 (8) 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 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 approximating $f_{X_n}(\theta_n) - f_{X_n}(\bar{\theta}_n) + F(\bar{\theta}_n)$. The intuitive motivation for this choice is that when the CTD algorithm arrives close to θ^*, the centering term alone ensures the updates become small, while with regular TD(0), one has to rely on a decaying step size to keep the iterates close to θ^*.

The approach is inspired by the SVRG algorithm, proposed in (Johnson & Zhang, 2013), for a optimising a strongly-convex function. 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 are not optimising a function that is a finite-sum of smooth functions in a batch setting. Instead, we are estimating a value function

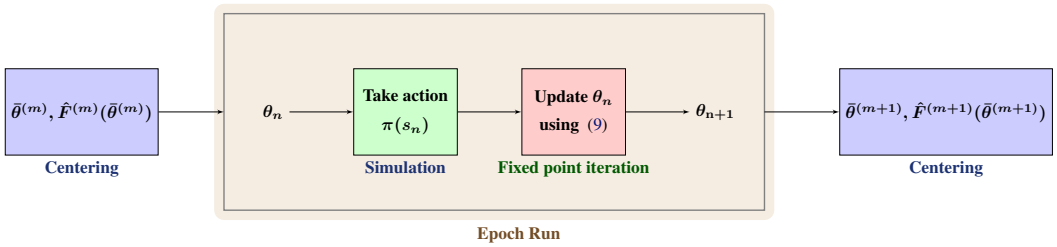


Figure 1. Illustration of centering principle in CTD algorithm.

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 the 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 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

$$\theta_{n+1} = \Upsilon\Big(\theta_n + \gamma\Big(f_{X_{i_n}}(\theta_n) - f_{X_{i_n}}(\bar{\theta}^{(m)}) + \hat{F}^{(m)}(\bar{\theta}^{(m)})\Big)\Big), \tag{9}$$

where $\hat{F}^{(m)}(\theta) := M^{-1}\sum_{i=(m-1)M}^{mM} f_{X_i}(\theta)$ and Υ is a projection operator that ensures that the iterates stay within a H-ball. Unlike TD(0), one can choose a large (constant) stepsize γ in (9). This choice in conjunction with iterate averaging via the choice of $\bar{\theta}^{(m)}$ results in an exponential convergence rate for CTD (see Remark 4 below).

4.1. Finite time bound

Theorem 3 below presents a finite time bound in expectation for CTD under the following mixing assumption:
(A6") There exists a non-negative function $B'(\cdot)$ such that: For all $s \in \mathcal{S}$ and $m \geq 0$,

$$\sum_{\tau=0}^{\infty} \|\mathbb{E}(r(s_\tau, \pi(s_\tau))\phi(s_\tau) \mid s_0 = s) - \mathbb{E}_\Psi(r(s_\tau, \pi(s_\tau))\phi(s_\tau))\| \leq B'(s),$$

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

The above is weaker than assumption (A6) used earlier for regular TD(0), and this is facilitated by the fact that we project the CTD iterates onto a H-ball.

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 (9) 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) Geometrically ergodic chains: *Here the Markov chain underlying policy π mixes fast (see (7)) and we obtain*[1]

$$\|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2 \leq C_1^m \left(\|\Phi(\bar{\theta}^{(0)} - \theta^*)\|_\Psi^2\right) + CMC_2H(5\gamma+4)\max\{C_1, \rho^M\}^{(m-1)}, \tag{10}$$

where $C_2 = \gamma/(M((1-\beta) - d^2\gamma/2))$.

(ii) General Markov chains:

$$\|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2 \leq C_1^m \left(\|\Phi(\bar{\theta}^{(0)} - \theta^*)\|_\Psi^2\right) + C_2H(5\gamma+4)\sum_{k=1}^{m-1} C_1^{(m-2)-k} B_{(k-1)M}^{kM}(s_0), \tag{11}$$

where $B_{(k-1)M}^{kM}$ 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. □

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

For finite state space settings, we obtain exponential convergence rate (10) since they are geometrically ergodic, while for MDPs that do not mix exponentially fast, the second (mixing) term in (11) will dominate and decide the rate of the CTD algorithm.

Remark 4. *Combining the result in* (10) *with the bound in statement* (4) *of Theorem* 1 *in (Tsitsiklis & Van Roy, 1997), we obtain*

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

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

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

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* $O(n^{-1/2})$.

5. Analysis

5.1. Non-averaged case: Proof of Theorem 1

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

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

Theorem 4. *Under (A1)-(A5) and (A6'), we have,*

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

where $L_i := \gamma_i[e^{-\mu(1-\beta)\sum_{k=i}^n\gamma_k}(1 + [\gamma_i + \sum_{k=i}^{n-1}[\gamma_k - \gamma_{k+1}]e^{\mu(1-\beta)\sum_{j=i}^{k+1}\gamma_j}][1+\beta(3-\beta)]B')]^{\frac{1}{2}}$.

Proof Sketch of Theorem 4. Recall that $z_n := \theta_n - \theta^*$.
Step 1: We rewrite $\|z_n\|_2^2 - E\|z_n\|_2^2$ as a telescoping sum of martingale differences as follows:

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

where $D_i := g_i - \mathbb{E}[g_i|\mathcal{F}_{i-1}]$, $g_i := \mathbb{E}[\|z_n\|_2|\theta_i]$.
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:

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

The result follows by optimizing over λ. The detailed proof is provided in (Korda & Prashanth, 2014). □

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

Theorem 5. *Under (A1)-(A6) and assuming that* $\gamma_n \leq \mu(1-\beta)/(2(1+\beta)^2)$ *for all n, we have,*

$$\mathbb{E}(\|\theta_{n+1} - \theta^*\|_2|s_0) \leq \Bigg[e^{-\mu(1-\beta)\sum_{k=1}^n\gamma_k}(\|z_0\|_2 + C) + (1+\|\theta^*\|_2)\sum_{k=1}^n\gamma_k^2 e^{-\mu(1-\beta)\sum_{j=k}^n\gamma_j} + C\sum_{k=1}^{n-1}(\gamma_{k+1}-\gamma_k)e^{-\mu(1-\beta)\sum_{j=k+1}^n\gamma_j}\Bigg]^{\frac{1}{2}} \quad (12)$$

where $C = 2(2+\beta)(d+4)B(s_0)\left(\frac{\|\theta_0\|_2 + d + \|\theta^*\|_2}{1-\beta}\right)^2$.

Proof sketch of Theorem 5. First we define some notation. Let $a_n := \beta\phi(s_n)\phi(s_{n+1})^\intercal - \phi(s_n)\phi(s_n)^\intercal$, $\epsilon_n := \mathbb{E}(a_n \mid \mathcal{F}_n) - \mathbb{E}_\Psi(a_n)$ and $\Delta M_n := a_n - \mathbb{E}(a_n \mid \mathcal{F}_n))$. Then, we can rewrite TD(0) update as follows:

$$z_{n+1} = [I - \gamma_n(A + \epsilon_n + \Delta M_n)]\, z_n + \gamma_n\epsilon_n'.$$

where $\epsilon_n' = f_{X_n}(\theta^*) - \mathbb{E}_\Psi\left(f_{X_n}(\theta^*)\right)$. ϵ_n, ϵ_n' are the mixing error components that arise due to the fact that TD(0) does not see samples from the stationary distribution of the underlying Markov chain has not mixed, while the martingale difference ΔM_n arises out of a sampling error.

Squaring the error z_n and taking the expectation, we obtain

$$\begin{aligned}&\mathbb{E}\left(\|z_{n+1}\|_2^2|\mathcal{F}_n\right)\\ \leq&[1 - 2\gamma_n(\mu(1-\beta) - 2\gamma_n(1+\beta)^2)]\|z_n\|_2^2\\ &+ \gamma_n z_n^\intercal\mathbb{E}(\epsilon_n \mid \mathcal{F}_n)\, z_n + \gamma_n^2(1 + (1+\beta)\|\theta^*\|_2)^2\\ &+ 2\gamma_n\mathbb{E}\left((\epsilon_n')^\intercal[I - \gamma_n(A + \epsilon_n + \Delta M_n)]\, z_n \mid \mathcal{F}_n\right).\end{aligned}$$

The remaining proof amounts to bounding each of the quantities on the RHS above. Bounding the error terms (both mixing and martingale errors) is tricky because it requires the iterate θ_n to be bounded as well and we do not project the iterates to artificially keep it bounded. Using (A6) coupled with technical arguments we show that the error terms stay bounded after unrolling the TD recursion. The reader is referred to (Korda & Prashanth, 2014) for the detailed proof. □

Rates For the expectation bound, the rate derivation involves bounding each term on the RHS in (12) after choosing step-sizes $\gamma_n = \frac{c_0 c}{(c+n)}$. We sketch the derivation for the second RHS term below:

$$\sum_{k=1}^{n} \gamma_k^2 e^{-\mu(1-\beta)\Gamma_k^n} \leq \sum_{k=1}^{n} \frac{c_0^2 c}{(c+k)^2} e^{-\mu(1-\beta)c_0 c \sum_{i=k}^{n} \frac{1}{c+i}} \leq \frac{c_0^2 c^2}{(\mu(1-\beta)c_0 c - 1)} \frac{1}{c+n}$$

where, in the last inequality, we have compared the sum with an integral and used assumption that $\mu(1-\beta)c_0 c > 1$. The other terms can be bounded in a similar fashion and the result in Theorem 1 follows.

For the bound in high probability, a calculation (see (Korda & Prashanth, 2014)) shows that choosing the step-size as before, we obtain

$$\sum_{i=1}^{n} L_i^2 \leq \frac{4(2+c_0 c)B'}{(c+n)^{\mu(1-\beta)c_0 c}} \sum_{i=1}^{n} \frac{c_0^2 c^2}{(c+i)^{2-\mu(1-\beta)c_0 c}}$$

and the result in Theorem 1 follows after comparing the summation above with an integral.

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 $z_n = \bar{\theta}_{n+1} - \theta^*$. 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 5):

Theorem 6. *Suppose that* $\forall n > n_0$, $\gamma_n \leq \mu(1-\beta)/(2(1+\beta)^2)$. *Then, under (A1)-(A5) and (A6'), and we have,* $\forall \epsilon \geq 0$ *and* $\forall n > n_0$,

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

where $L_i := \frac{\gamma_i}{n}\left(1 + \sum_{l=i+1}^{n-1}\left[e^{-\mu(1-\beta)\sum_{k=i}^{n}\gamma_k}(1 + [\gamma_i + \sum_{k=i}^{n-1}[\gamma_k - \gamma_{k+1}]][1+\beta(3-\beta)]B')\right]^{\frac{1}{2}}\right)$.

Rates To bound the expected error we average the errors of the non-averaged iterates. However, this averaging this is not straightforward as the bound in Theorem 5 holds only if $n > n_0$ (which ensures that γ_n is sufficiently small). Note that n_0 can be easily derived from the specific form of the step sequece. Hence we analyse the initial phase ($n < n_0$) and later phase $n \geq n_0$ separately as follows:

$$\mathbb{E}\|z_n\|_2 \leq \frac{\sum_{k=1}^{n_0} \mathbb{E}\|\theta_k - \theta^*\|_2}{n} + \frac{\sum_{k=n_0+1}^{n} \mathbb{E}\|\theta_k - \theta^*\|_2}{n}$$

The last term on RHS above is bounded using Theorem 1, while the first term is bounded by unrolling TD(0) recursion for the first n_0 steps and bounding the individual terms that arise using (A6).

For the rate of the bound in high probability, one has to again separately bound the influence of the first n_0 steps, and then use the expectation bound together with 6. The reader is referred to (Korda & Prashanth, 2014) for detailed proofs.

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

Step 1: Let $\bar{f}_{X_{i_n}}(\theta_n) := f_{X_{i_n}}(\theta_n) - f_{X_{i_n}}(\bar{\theta}^{(m)}) + \mathbb{E}(f_{X_{i_n}}(\bar{\theta}^{(m)}) \mid \mathcal{F}_n)$.. An one-step expansion of the recursion (9) gives

$$\|\theta_{n+1} - \theta^*\|_2^2 \leq \|\theta_n - \theta^*\|_2^2 + 2\gamma(\theta_n - \theta^*)^\intercal \bar{f}_{X_{i_n}}(\theta_n) + \gamma^2 \left\|\bar{f}_{X_{i_n}}(\theta_n)\right\|_2^2. \tag{13}$$

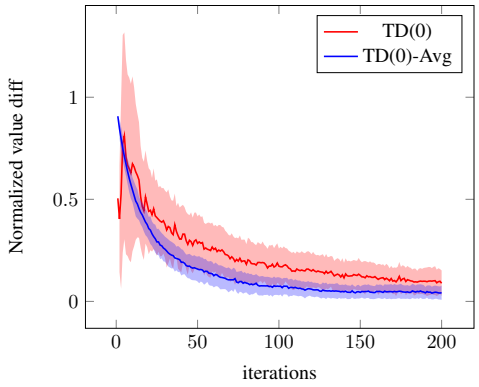
Step 2: We bound the variance of centered updates:

$$\mathbb{E}\left(\left\|\bar{f}_{X_{i_n}}(\theta_n)\right\|_2^2 \middle| \mathcal{F}_n\right) \leq \mathbb{E}\left(\|e_n(\theta^*)\|_2^2 \mid \mathcal{F}_n\right) + \epsilon_n(\bar{\theta}^{(m)}) + \epsilon_n(\theta_n) + d^2\left(\|\Phi(\theta_n - \theta^*)\|_\Psi^2 + \|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2\right),$$

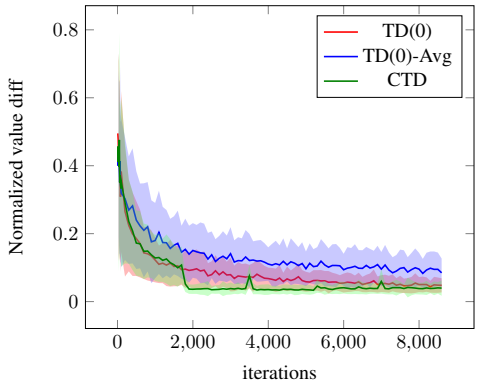
where $\epsilon_n(\theta) = \mathbb{E}\left(\left\|f_{X_{i_n}}(\theta) - f_{X_{i_n}}(\theta^*)\right\|_2^2 \middle| \mathcal{F}_n\right) - \mathbb{E}_{\Psi,\theta_n}(\left\|f_{X_{i_n}}(\theta) - f_{X_{i_n}}(\theta^*)\right\|_2^2)$ and $e_n(\theta) := \mathbb{E}\left[f_{X_{i_n}}(\theta) \middle| \mathcal{F}_n\right] - \mathbb{E}_{\Psi,\theta_n}\left[f_{X_{i_n}}(\theta)\right]$..

Step 3: From (13) we obtain the following recursion within an epoch

$$\begin{aligned} &2\gamma M((1-\beta) - \frac{d^2\gamma}{2})\mathbb{E}\left(\|\Phi(\bar{\theta}^{(m+1)} - \theta^*)\|_\Psi^2 \middle| \mathcal{F}_{mM}\right) \\ &\leq \left(\frac{1}{\mu} + M\gamma^2 d^2\right)\|\bar{\theta}^{(m)} - \theta^*\|_2^2 \\ &+ \gamma^2 \sum_{n=mM}^{(m+1)M-2} \mathbb{E}\left(\epsilon_n(\theta_n) + \epsilon_n(\bar{\theta}^{(m)}) + \|e_n(\theta^*)\|_2^2 \middle| \mathcal{F}_{mM}\right) \\ &+ \mathbb{E}\left(2\gamma \sum_{n=mM}^{(m+1)M-1} (\theta_n - \theta^*)^\intercal e_n(\theta_n) \middle| \mathcal{F}_{mM}\right) \end{aligned} \tag{14}$$



(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 $\|\Phi(\theta_n - \theta^*)\|_\Psi / \|\Phi(\theta^*)\|_\Psi$.

where we have first simplified (13) before using the bound on the variance of the centered term (from Step 2), and also the fact that

$$(\bar{\theta}^{(m)} - \theta^*)^\intercal I(\bar{\theta}^{(m)} - \theta^*) \leq \mu^{-1}(\bar{\theta}^{(m)} - \theta^*)^\intercal \Phi^\intercal \Psi \Phi(\bar{\theta}^{(m)} - \theta^*)$$

Step 4: Finally, we obtain (11) by unrolling (across epochs) (14), and bounding the individual error terms involving $\epsilon(\cdot)$ and $e_n(\cdot)$. For the latter, we use (A6") and the fact that we project the CTD iterates to a H-ball.

6. Numerical Experiments

We test the performance of TD(0), TD(0) with averaging and CTD algorithms.

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 less than 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 properties of the stationary distribution 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.

Acknowledgments The first author was gratefully supported by the EPSRC Autonomous Intelligent Systems project EP/I011587. The second author would like to thank the European Community's Seventh Framework Programme (FP7/2007 – 2013) under grant agreement n^o 270327 for funding the research leading to these results.

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.

Korda, Nathaniel and Prashanth, L.A. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. *arXiv preprint arXiv:1411.3224v2*, 2014.

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

Meyn, Sean P and Tweedie, Richard L. *Markov chains and stochastic stability*. Cambridge university press, 2009.

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.