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

Nathaniel Korda *1 and Prashanth L A †2

1Oxford University, UNITED KINGDOM.
2INRIA 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.

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 \( V^\pi \) for a given stationary policy \( \pi \). We focus on discounted reward MDPs with a high-dimensional state space \( S \). 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 and Tsitsiklis [2003], Bhatnagar et al. [2009].

Temporal difference learning Sutton and 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 \( s \in S \)) 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., \( V^\pi(s) \approx \theta^T \phi(s) \). Here \( \theta \) 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 \( \theta_0 \),

\[
\theta_{n+1} = \theta_n + \gamma_n (r(s_n, \pi(s_n)) + \beta \theta_n^T \phi(s_{n+1}) - \theta_n^T \phi(s_n)) \phi(s_n). 
\]  

(1)

In the above, the quantities \( \gamma_n \) 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 and Van Roy [1997], the authors establish that $\theta_n$ governed by (1) converges almost surely to the fixed point, $\theta^*$, of the projected Bellman equation given by

$$\Phi \theta^* = \Pi T^\pi (\Phi \theta^*).$$

In the above, $T^\pi$ is the Bellman operator, $\Pi$ is the orthogonal projection onto the linearly parameterized space within which we approximate the value function, and $\Phi$ is the feature matrix with rows $\phi(s)^T, \forall s \in S$ denoting the features corresponding to state $s \in S$ (see Section 2 for details). Let $P$ denote the transition probability matrix with components $p(s, \pi(s), s')$ and let $\Psi$ be a diagonal matrix whose diagonal forms the stationary distribution (assuming it exists) of the Markov chain for the underlying policy $\pi$. Then, $\theta^*$ can be written as the solution to the following system of equations (see Section 6.3 of Bertsekas [2011])

$$A\theta^* = b,$$

where $A = \Phi^T \Psi (I - \beta P) \Phi$ and $b = \Phi^T \Psi r$.

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 $\theta^*$ is the fixed point of the Bellman operator, which assumes that the underlying MDP is begun from the stationary distribution $\Psi$ (whose influence is evident in (3)). However, the samples provided to the algorithm come from simulations of the MDP that are not begun from $\Psi$. 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 $\pi$.

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 choose 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 $\pi$.

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 and 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 and 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 a set of assumptions similar to Tsitsiklis and 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 linear function approximators.

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 $O(1/\sqrt{n})$ for the bound in expectation. Here $\mu$ is the smallest eigenvalue of the matrix $\Phi^T \Psi \Phi$ (see
Theorem 3 in Section 2.

(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 $\pi$ (see Theorem 3). 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 4).

(4) For the centered variant of TD(0), we obtain an exponential convergence rate when the underlying Markov chain mixes fast (see Theorem 5).

TD(0) with function approximation is an efficient algorithm that is easy to implement on large state space problems. However, deriving convergence rate results, especially of non-asymptotic nature, requires sophisticated machinery. In particular, we base our approach on that proposed in Frikha and Menozzi [2012] (and later expanded on). We would like to remark that asymptotic convergence rate results for TD(0) are available in [Konda 2002]. The authors establish there that TD(0) converges asymptotically to a multi-variate Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma$ 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(0) 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(0). Asymptotic convergence rate results for LSTD(0) and LSPE(0), two popular least squares methods, are available in [Konda 2002] and [Yu and 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 $S$ and action space $A$. The aim is to estimate the value function $V^\pi$ for any given stationary policy $\pi : S \to A$, where

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

(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 $s, s' \in S$ and $a \in A$ (we denote this probability by $p(s, a, s')$). It is well-known that the value function $V^\pi$ is the solution to the fixed point relation $V = T^\pi(V)$, where the Bellman operator $T^\pi$ is defined as

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

(5)

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

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

(6)

where $\gamma_n$ 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 $T^\pi$ 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 $s \in S$, approximate $V^\pi(s) \approx \phi(s)^T \theta$. Here $\phi(s)$ is a $d$-dimensional feature vector with $d << |S|$, and $\theta$ 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 $n$, $\|\theta_n\| \leq H$.

3 Concentration bounds for TD(0)

3.1 Assumptions

(A1) Ergodicity: The Markov chain induced by the policy $\pi$ is irreducible and aperiodic. Moreover, there exists a stationary distribution $\Psi(= \Psi_\pi)$ for this Markov chain. Let $E_\Psi$ denote the expectation w.r.t. this distribution.

(A2) Bounded rewards: $|r(s, \pi(s))| \leq 1$, for all $s \in S$.

(A3) Linear independence: The feature matrix $\Phi$ has full column rank. This assumption implies that the matrix $\Phi^T \Phi$ has smallest eigenvalue $\mu > 0$.

(A4) Bounded features: $\|\phi(s)\|_2 \leq 1$, for all $s \in S$.

(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 $s_0 \in S$ and $m \geq 0$,

$$\sum_{\tau=0}^{\infty} \|E_\Psi \phi(s_\tau) \mid \mathcal{F}_n - E_\Psi \phi(s_\tau)\| \leq B(s_0),$$

$$\sum_{\tau=0}^{\infty} \|E_\Psi [\phi(s_\tau) \phi(s_{\tau+m})^T \mid s_0] - E_\Psi [\phi(s_\tau) \phi(s_{\tau+m})^T]\| \leq B(s_0),$$

where $\mathcal{F} = \{\mathcal{F}_n := \sigma(\theta_1, \ldots, \theta_n)\}$ is the filtration generated by the iterates $\theta_n$, $n \geq 0$, and the function $B(\cdot)$ is assumed to be bounded such that, for any $q \geq 1$, there exists a $K_q < \infty$ such that

$$E[B^q(s) \mid s_0] \leq K_q B^q(s_0).$$

The above assumptions are similar in nature to those made in [Tsitsiklis and 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 and Van Roy, 1997], while (A2) and (A4) are simplified versions of corresponding boundedness assumptions in [Tsitsiklis and Van Roy, 1997].

3.2 Non-averaged case

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

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

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

where

$$K_1(n) := \sqrt{1 + 9 B(s_0)^2} \left(\frac{2}{(n + c)^2(1 - \beta)^2 \mu c - 1/2} + \frac{2\beta(1 - \beta) c H}{1 - 2(1 - \beta)^2 \mu c}\right).$$

1We assume an upper bound of 1 for both rewards and features for the sake of simplicity and we believe our analysis can be extended (left for future work) to the following generalized variants:

(A2') The rewards satisfy $E_\Psi (r^2(s, \pi(s))) < \infty, \forall s \in S$;

(A4') The feature vector $\phi_k(s)$ for any $k = 1, \ldots, d$ and $s \in S$ satisfies $E_\Psi (\phi_k^2(s)) < \infty$. 

(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 \( \delta > 0 \)
\[
\mathbb{P} \left( \| \theta_n - \theta^* \|_2 \leq \frac{K_2(n)}{\sqrt{n+c}} \right) \geq 1 - \delta,
\]
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 \( O(1/\sqrt{n}) \) 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 mixing bound \( B(s_0) \) and knowing \( B(s_0) \) would imply knowledge about the transition probability matrix of the underlying Markov chain. For finite state space settings, the mixing bound \( B(s_0) \) can be shown to be a constant (see Section VII of [Tsitsiklis and Van Roy 1997]), but one that depends on the second eigenvalue of the transition probability matrix 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., \( \bar{\theta}_{n+1} := (\bar{\theta}_1 + \ldots + \bar{\theta}_n)/n \). This principle was introduced independently by Ruppert [Ruppert 1991] and Polyak [Polyak and 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 \( \delta > 0 \),
\[
\mathbb{E} \left\| \bar{\theta}_{n+1} - \theta^* \right\|_2 \leq \frac{K_3^A(n)}{(n+c)^{\alpha/2}}
\]
and
\[
\mathbb{P} \left( \left\| \bar{\theta}_{n+1} - \theta^* \right\|_2 \leq \frac{K_3^A(n)}{(n+c)^{\alpha/2}} \right) \geq 1 - \delta,
\]
Proof. See Section 5.2.

Remark 2. The step-size exponent \( \alpha \) can be chosen arbitrarily close to 1, resulting in a convergence rate of the order \( O(1/\sqrt{n}) \). 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 \( \alpha \) close to 1 would result in their bounding constants blowing up. One cannot choose \( c \) too large or too small for the same reasons.

Thus, iterate averaging results in the optimal rate of convergence, while having no dependency for the choice of \( c \). This is consistent with the asymptotic convergence rate results from [Konda 2002], where the authors establish the TD(\(\lambda\)) exhibits optimal convergence only with iterate averaging.

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:

\[
\theta_n = \theta_{n-1} + \gamma_n f_{X_n}(\theta_n).
\]

where \( f_{X_n}(\theta) := (r(s_n, \pi(s_n)) + \beta \theta^T\phi(s_{n+1}) - \theta^T\phi(s_n))\phi(s_n) \). The limit of (14) is the solution, \( \theta^* \), 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 \( \theta_n \), 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 and 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 and 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 \( \pi \)).

(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 \( \theta \).

(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 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 and Zhang 2013] is stochastic gradient descent (SGD).

The upshot of the above is that, while centering is an interesting idea for SGD, it is not straightforward to apply it to improve the convergence rate of TD(0).

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. \( \bar{\theta}^{(m)} = \bar{\theta}_n \), where \( \bar{\theta}_n \) is drawn uniformly at random in \( \{ (m-1)M, \ldots, mM \} \). Thereafter, for
the epoch length \( M \), CTD performs the following iteration: Set \( \theta_{nM} = \bar{\theta}^{(m)} \) and for \( n = mM, \ldots, (m+1)M - 1 \) update

\[
\theta_{n+1} = \theta_n + \gamma \left( f_{X_{n}}(\theta_n) - f_{X_{n}}(\bar{\theta}^{(m)}) + \hat{F}^{(m)}(\bar{\theta}^{(m)}) \right),
\]

(15)

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 \( \gamma \) in (15). 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.** Let \( \theta^* \) denote the solution of \( F(\theta) = 0 \). Let the epoch length \( M \) of the CTD algorithm \((15)\) be chosen such that \( C_1 < 1 \), where

\[
C_1 := \left( \frac{1}{2\mu \gamma M ((1 - \beta) - d^2 \gamma)} + \frac{\gamma d^2}{2((1 - \beta) - d^2 \gamma)} \right)
\]

Then, under \((A1)-(A4)\) and \((A6)\), we have,

\[
\| \Phi(\bar{\theta}^{(m)} - \theta^*) \|_\Psi \leq C_1^m \left( \| \Phi(\bar{\theta}^{(0)} - \theta^*) \|_\Psi \right) + C_2 \sum_{k=1}^{m-1} C_1^{(m-2)-k} \sum_{i=(k-1)M}^{kM} \mathbb{E}_{\theta_i} \| \epsilon_i \|_2^2,
\]

(16)

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

In particular, if the Markov chain underlying policy \( \pi \) satisfies the following property:

\[
| P(s_t = s \mid s_0) - \psi(s) | \leq C \rho t^M,
\]

(17)

then

\[
\| \Phi(\bar{\theta}^{(m)} - \theta^*) \|_\Psi \leq C_1^m \left( \| \Phi(\bar{\theta}^{(0)} - \theta^*) \|_\Psi \right) + CC_2 \max \{C_1, \rho\}^{(m-1)}
\]

(18)

**Proof.** See Section 5.3.

**Remark 3.** For finite state space settings, it is easy to see that the fast-mixing assumption \((17)\) holds (see Section VII of [Tsitsiklis and Van Roy, 1997]) and hence one can obtain exponential convergence rate using \((16)\). On the other hand, 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 4.** Combining the result in \((18)\) with the bound in statement \((4)\) of Theorem 1 in [Tsitsiklis and Van Roy, 1997], we obtain

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

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.** To get the optimal rate for the CTD algorithm, we need to know the value of \( \mu \) to set the step-size \( \gamma \). However, we can get rid of this dependency by explicitly regularizing the problem. This would imply that we solve \((A + \mu I)\theta^* = b\) instead of \((3)\), where we choose \( \mu \).

\( \text{For any } v \in \mathbb{R}^d, \text{we take } \| v \|_\Psi := \sqrt{\Psi^T v}. \)
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)-(A6), we have,

\[
\mathbb{E} \|\theta_n - \theta^*\|_2 \leq \sqrt{1 + 9B(s_0)^2 \left[ \exp(-(1 - \beta)\mu \Gamma_n) \|\theta_0 - \theta^*\|_2 \right]}
\]

(19)

\[
+ \left( \sum_{k=1}^{n-1} (1 + 2\beta H)\gamma_{k+1}^2 \exp(-2(1 - \beta)\mu(\Gamma_n - \Gamma_k+1)) \right),
\]

where \(\Gamma_k := \sum_{i=1}^k \gamma_i\) and \(\|\theta_n\|_2 \leq H, \forall n\).

The initial error depends on the initial point \(\theta_0\) 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 \(\pi\), 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 \(f_{X_n}(\theta) := [r(s_n, \pi(s_n)) + \beta \theta_{n-1}^T\phi(s_{n+1}) - \theta_{n-1}^T\phi(s_n)]\phi(s_n)\). The first step is to rewrite the recursion (1) as follows:

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

(20)

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

The next step is to unroll (20) as follows:

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

\[
= \Pi_n z_0 + \sum_{k=1}^{n} \gamma_k \Pi_n \Pi_k^{-1} (\epsilon_k + \Delta M_k),
\]

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

The mixing error can be bounded using (A6). However, bounding \(\|\Delta M_n\|_2\) is tricky since it requires the iterate \(\theta_n\) 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., \(\|\theta_n\|_2 \leq H\). The reader is referred to Appendix A.1 for the detailed proof.

In order to obtain the rates and constants presented in Theorem 1, we now specialize Theorem 4 for a particular 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:

\[
\sum_{k=1}^{n-1} 2 \beta H \gamma^{k+1}_n \exp(-2(1-\beta)\mu(\Gamma_n - \Gamma_{k+1})) \\
\leq \frac{4 \beta^2 (1-\beta)^2 c^2 H^2}{(n+c)^2(1-\beta)^2 \mu c} \sum_{k=1}^{n} (c+k)^{-2(1-\beta)^2 \mu c} \\
\leq \frac{4 \beta^2 (1-\beta)^2 c^2 H^2}{1-2(1-\beta)^2 \mu c} \frac{1}{c+n}
\]

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

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

So we have

\[
E \| \theta_n - \theta^* \|_2 \leq \sqrt{c} \| \theta_0 - \theta^* \|_2 + 2 \beta (1-\beta)cH \\
\]

and the result in Theorem 1 now follows.

**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,

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

where

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

This theorem decomposes the problem of bounding \( \| \theta_n - \theta^* \|_2 \) into bounding the deviation from its mean \( E \| \theta_n - \theta^* \|_2 \) 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 \( z_n := \theta_n - \theta^* \). The first step is to rewrite \( \| z_n \|_2^2 - E \| z_n \|_2^2 \) as a telescoping sum of martingale differences as follows:

\[
\| z_n \|_2 - E \| z_n \|_2 = \sum_{i=1}^{n} g_i - E[g_i | F_{i-1}] = \sum_{i=1}^{n} D_i,
\]

where \( D_i := g_i - E[g_i | F_{i-1}] \), \( g_i := E[\| z_n \|_2 | \theta_i] \). The next step is to establish that the functions \( g_i \) are Lipschitz continuous as follows:

**Lemma 6.** Under the conditions of Theorem 5 conditioned on \( F_{i-1} \), the functions \( g_i \) are Lipschitz continuous in \( f_{X_i}(\theta_{i-1}) \) with constants \( L_i \).

The final step is to invoke a standard martingale concentration bound using the \( L_i \)-Lipschitz property of the \( g_i \) functions and the assumption (A3) to obtain:
Lemma 7. Under the conditions of Theorem 5 we have

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

The result follows by optimizing over $\lambda$. For detailed proofs of the above lemmas, the reader is referred to Appendix A.2.

Now, to derive the rate given in Theorem 1 we again specialise to the choice of step size sequence: $\gamma_n = (1 - \beta)c/(c + n)$. A calculation (see the Appendix A.3) shows that

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

We now find several regimes for the rate of convergence, based on the choice of $c$, however choosing it so that $(\mu(1 - \beta)/2 + 3B(s_0)) < (1, \infty)$ we have

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

(we have used comparisons with integrals to bound the summations), and the result in Theorem 1 now 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 8.** Under (A1)-(A3) we have, for all $\epsilon \geq 0$ and $\forall n \geq 1$,

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

where

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

**Proof.** Let $\zeta_i^1$ be the value of the averaged iterate $\bar{\theta}_{i+1}$ at instant $i$ and $\zeta_i^2$ be the value of the iterate $\theta_i$ at instant $i$. Then as in the previous section, we decompose $\|z_n\|_2 - \mathbb{E}\|z_n\|_2$ into a sum of martingale differences:

$$\|z_n\|_2 - \mathbb{E}\|z_n\|_2 = \sum_{i=1}^{n} D_i,$$

where $D_i := g_i - \mathbb{E}[g_i | \mathcal{F}_{i-1}]$ and $g_i := \sum_{n=1}^{n} \mathbb{E}[\|z_n\|_2 | \zeta_i = (\zeta_i^1, \zeta_i^2)]$.

Now we need to prove that the functions $g_i$ are Lipschitz continuous in the random innovation at time $i$. Let $\tilde{\Theta}_i^j(\zeta)$ denote the mapping that returns the value of $\zeta_i$, given that $\zeta_{i-1} = \zeta_1^1$ and $\zeta_i = \zeta_i^2$. Then by applying Theorem 5 we have

$$\mathbb{E}\left[\|\tilde{\Theta}_i^j(\zeta) - \tilde{\Theta}_i^j(\zeta')\|_2\right] \leq \frac{i + 1}{n} \|\zeta^1 - \zeta'^1\|_2 + \frac{1}{n} \sum_{i=1}^{n-1} \prod_{j=1}^{i} \left(1 - 2\gamma_j \left(\mu \left(1 - \beta - \frac{\gamma_j}{2}\right) + [1 + \beta(3 - \beta)] B(s_0)\right)\right)^{\frac{1}{2}} \|\zeta^2 - \zeta'^2\|_2$$

The rest of the proof follows in a similar manner to the proof of Theorem 5.
Now, in order to obtain the rates in Theorem 3, 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:

\[
E \| \bar{\theta}_{n+1} - \theta^* \|_2 \leq \frac{1}{n} \sum_{k=1}^{n} E \| \theta_k - \theta^* \|_2 ,
\]

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

**Lemma 9.** Under conditions of Theorem 2 we have

\[
E \| \bar{\theta}_{n+1} - \theta^* \|_2 \leq \sqrt{\frac{1 + 9B(s_0)}{n}} \left( \sum_{n=1}^{\infty} \exp(-\mu c(n+c)^{1-\alpha}) \| \theta_0 - \theta^* \|_2 \right.
\]

\[
+ 2\beta H c^\alpha (1-\beta) \left( \mu c^\alpha (1-\beta)^2 \right)^{\alpha} \left( \frac{1+2\alpha}{2\alpha} \right) (n+c)^{-\frac{\alpha}{2}}
\]

**Proof.** See Appendix B.2  

For the rate of the bound in high probability we need to specialise the bounds for the bound in expectation in Theorem 5 for the new choice of step-size sequence. In particular, we must compute the value of the Lipschitz \( L_i \) constants for our choice of step-sizes:

**Lemma 10.** Under conditions of Theorem 2 we have

\[
\sum_{i=1}^{n} L_i^2 \leq \frac{\left[ \frac{\mu^2}{2} \left[ \frac{1}{1 + \frac{B(s_0)}{\mu} c^\alpha \| \Psi \|_2} \right]^2 - \frac{1}{n} \right]}{\| \theta_0 - \theta^* \|_2^2}
\]

**Proof.** See Appendix B.1  

Finally this lemma together with Theorem 5 completes the result.

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

**Step 1: Rewriting the CTD update (15).**
Recall that the last term in (15) is \( F(m)(\tilde{\theta}(m)) = M^{-1} \sum_{i=1}^{mM} \hat{f}_{X_i}(\tilde{\theta}(m)) \). We rewrite this term as follows:

\[
\tilde{F}(m)(\tilde{\theta}(m)) = E\Psi(f_{X_i}(\tilde{\theta}(m))) + \epsilon_n,
\]

where \( \epsilon_n := E(f_{X_i}(\tilde{\theta}(m)) \mid F_m) - E\Psi(f_{X_i}(\tilde{\theta}(m))) \). In the above, we have used the fact that \( E(f)(\tilde{\theta}(m)) - E(f_{X_i}(\tilde{\theta}(m)) \mid F_m) = 0 \), since \( i_n \) is chosen uniformly at random from \( \{m-1,M,\ldots,mM\} \).

**Step 2: Bounding the variance of centred updates.** Let

\[
\hat{f}_{X_i}(\theta_n) := f_{X_i}(\theta_n) - f_{X_i}(\tilde{\theta}(m)) + E\Psi(f_{X_i}(\tilde{\theta}(m))
\]

Then,

\[
E\Psi(\| \hat{f}_{X_i}(\theta_n) \|_2^2) \leq E\Psi \left( \| f_{X_i}(\theta_n) - f_{X_i}(\theta^*) \|_2^2 + \| f_{X_i}(\tilde{\theta}(m)) - f_{X_i}(\theta^*) - E\Psi(f_{X_i}(\tilde{\theta}(m))) \|_2^2 \right)
\]

\[
\leq E\Psi \left( \| f_{X_i}(\theta_n) - f_{X_i}(\theta^*) \|_2^2 + E\Psi \left( \| f_{X_i}(\tilde{\theta}(m)) - f_{X_i}(\theta^*) \|_2^2 \right) \right),
\]

where we have used that for any random variable \( \xi, E(\| x - E_x \|^2) \leq E(\| x \|^2) \).
For any \( \theta \), we have
\[
\mathbb{E}_\psi \left( \| f_{X_{n+1}} (\theta) - f_{X_n} (\theta^*) \|_2^2 \right) = \langle \theta - \theta^* \rangle^\top (\Phi^\top (I - \beta P) \Phi)^\top \Phi^\top (I - \beta P) \Phi (\theta - \theta^*) \\
\leq \langle \theta - \theta^* \rangle^\top (\Phi^\top \Psi \Phi)^\top \Phi^\top \Phi (\theta - \theta^*) \leq d^2 \| \Phi (\theta - \theta^*) \|_\Psi^2 .
\]
In the final inequality, we have used \( \| \Phi^\top \Psi \Phi \|_2 \leq d^2 \). Plugging the above in (26), we obtain
\[
\mathbb{E}_\psi \left( \| f_{X_{n+1}} (\theta_n) \|_2^2 \right) \leq d^2 \left( \| \Phi (\theta_n - \theta^*) \|_\Psi^2 + \| \Phi (\tilde{\theta}^{(m)} - \theta^*) \|_\Psi^2 \right) \tag{27}
\]

**Step 3: Analysis for a particular epoch.**
Taking the expected error of the \( n^{th} \) iterate, we have:
\[
\mathbb{E}_{\theta_n} \| \theta_{n+1} - \theta^* \|_2^2 \leq \| \theta_n - \theta^* \|_2^2 + 2\gamma \mathbb{E}_{\theta_n} \| \epsilon_n \|_2^2 + 2\gamma (\theta_n - \theta^*)^\top \mathbb{E}_{\theta_n} \left[ f_{X_{n+1}} (\theta_n) \right] + \gamma^2 \mathbb{E}_{\theta_n} \left[ \| f_{X_{n+1}} (\theta_n) \|_2^2 \right] \\
= \| \theta_n - \theta^* \|_2^2 + 2\gamma \mathbb{E}_{\theta_n} \| \epsilon_n \|_2^2 + 2\gamma^2 \mathbb{E}_{\theta_n} \left[ \| f_{X_{n+1}} (\theta_n) \|_2^2 \right] + 2\gamma (\theta_n - \theta^*)^\top \mathbb{E}_{\theta_n} (f_{X_n}) (\theta_n - 1)
\]
So, applying (27), we have
\[
\mathbb{E}_{\theta_n} \| \theta_{n+1} - \theta^* \|_2^2 \leq \| \theta_n - \theta^* \|_2^2 + 2\gamma \mathbb{E}_{\theta_n} \| \epsilon_n \|_2^2 + 2\gamma \Phi (\theta_n - \theta^*)^\top \mathbb{E}_{\theta_n} \left[ f_{X_{n+1}} (\theta_n) \right] + \gamma^2 \mathbb{E}_{\theta_n} \left[ \| f_{X_{n+1}} (\theta_n) \|_2^2 \right] \\
\leq \| \theta_n - \theta^* \|_2^2 + 2\gamma (1 - \beta - d^2 \gamma) \| \Phi (\theta_n - \theta^*) \|_\Psi^2 + \gamma^2 d^2 \left( \| \Phi (\tilde{\theta}^{(m)} - \theta^*) \|_\Psi^2 \right) + \gamma^2 \mathbb{E}_{\theta_n} \| \epsilon_n \|_2^2 .
\]
Summing the above inequality over an epoch and noting that \( \mathbb{E}_{\Phi,\theta_n} \| \theta_{n+1} - \theta^* \|_2^2 \geq 0 \), we have
\[
2\gamma M((1 - \beta) - d^2 \gamma) \| \Phi (\tilde{\theta}^{(m+1)} - \theta^*) \|_\Psi^2 \leq \| \theta_0 - \theta^* \|_2^2 + \gamma^2 M d^2 \left( \| \Phi (\tilde{\theta}^{(m)} - \theta^*) \|_\Psi^2 \right) \\
+ \gamma^2 \sum_{i=(m-1)M}^{mM} \mathbb{E}_{\theta_i} \| \epsilon_i \|_2^2 .
\]
Notice that
\[
(\tilde{\theta}^{(m)} - \theta^*)^\top I (\tilde{\theta}^{(m)} - \theta^*) \leq \frac{1}{\mu} (\tilde{\theta}^{(m)} - \theta^*)^\top \Phi^\top \Phi (\tilde{\theta}^{(m)} - \theta^*)
\]
and hence we obtain the following by setting \( \theta_0 = \tilde{\theta}^{(m)} \):
\[
2\gamma M((1 - \beta) - d^2 \gamma) \| \Phi (\tilde{\theta}^{(m+1)} - \theta^*) \|_\Psi^2 \leq \left( \frac{1}{\mu} + \gamma^2 M d^2 \right) \left( \| \Phi (\tilde{\theta}^{(m)} - \theta^*) \|_\Psi^2 \right) + \gamma^2 \sum_{i=(m-1)M}^{mM} \mathbb{E}_{\theta_i} \| \epsilon_i \|_2^2 .
\]

**Step 4: Combining across epochs.**
Finally, we obtain (26) by unrolling (across epochs) the final recursion in the previous step.

**6 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 for TD(0) with linear function approximators. These include bounds both in high-probability as well as in expectation. From our results, it is observed that in order to obtain the optimal rate of convergence of the order \( O (1 / \sqrt{\gamma}) \), it is necessary to incorporate iterate averaging. This is because, to obtain the optimal rate with the classic step-size choice that is inversely proportional to the number of iterations of TD(0), 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.
Appendix

A Convergence rate of TD(0): Non-averaged case

Throughout this section let $X_n = (s_n, s_{n+1})$ and $f_{X_n}(\theta) := (r(s_n, \pi(s_n)) + \beta \theta_{n-1}^T \phi(s_{n+1}) - \theta_n^T \phi(s_n))$. Recall that $\theta^*$ is the solution of $\mathbb{E}_\psi(f_{X_n}(\theta^*)) = 0$, and let $z_{n+1} = \theta_{n+1} - \theta^*$. Finally, let $\mathcal{F} = \{\mathcal{F}_n := \sigma(\theta_1, \ldots, \theta_n)\}$ denote the filtration generated by the iterates $\theta_n, n \geq 0$.

A.1 Bound in expectation

Theorem 11. Under (A1)-(A6), we have,

$$
\mathbb{E}\|\theta_n - \theta^*\|_2 \leq \sqrt{1 + 9B_0^2 \exp(-(1-\beta)\mu\Gamma_n)} \|\theta_0 - \theta^*\|_2 \tag{A.1}
$$

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

Proof. The update rule (1) can be re-written as follows:

$$
\theta_{n+1} = \theta_n + \gamma_n [\mathbb{E}_\psi(f_{X_n}(\theta_n)) + \epsilon_n + \Delta M_n], \tag{A.2}
$$

where $\epsilon_n := \mathbb{E}(f_{X_n}(\theta_n) | \mathcal{F}_n) - \mathbb{E}_\psi(f_{X_n}(\theta_n))$ is the mixing error term, while $\Delta M_n := f_{X_n}(\theta_n) - \mathbb{E}(f_{X_n}(\theta_n) | \mathcal{F}_n)$ is a martingale sequence.

We notice that

$$
\mathbb{E}_\psi(f_{X_n}(\theta_n)) = \mathbb{E}_\psi(f_{X_n}(\theta_n) - f_{X_n}(\theta^*))
\begin{align*}
&= \mathbb{E}_\psi(\beta \theta_n^T \phi(s_{n+1}) - \beta \theta^{*T} \phi(s_{n+1}) - (\theta_n^T \phi(s_n) - \theta^{*T} \phi(s_n))) \\
&= \mathbb{E}_\psi((\theta_n - \theta^*)^T [\beta \phi(s_{n+1}) - \phi(s_n)]) \\
&= \mathbb{E}_\psi(\phi(s_n)^T [\beta (s_{n+1})^T - \phi(s_n)^T] (\theta_n - \theta^*)) \\
&= -A(\theta_n - \theta^*), \tag{A.3}
\end{align*}
$$

where $A := \Phi^T \Psi(I - \beta P) \Phi$ (here $P = P_\pi$ denotes the one-step transition probability matrix of the underlying Markov chain induced under a stationary policy $\pi$). Plugging (A.3) into (A.2), we obtain the following recursive procedure:

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

where $\Pi_n := \prod_{k=1}^{n} (I - \gamma_k A)$. 


The mixing error can be bounded as follows:

\[
\sum_{n=0}^{\infty} \|\epsilon_n\|_2 \leq \sum_{n=0}^{\infty} \|\mathbb{E}(r(s_n, \pi(s_n))\phi(s_n) \mid \mathcal{F}_n) - \mathbb{E}(r(s_n, \pi(s_n))\phi(s_n))\|_2 \tag{A.4}
\]

\[
+ \sum_{n=0}^{\infty} \|\mathbb{E}(\beta \phi(s_n)\phi(s_{n+1})^T \mid \mathcal{F}_n) - \mathbb{E}(\beta \phi(s_n)\phi(s_{n+1})^T)\|_2 \tag{A.5}
\]

\[
+ \sum_{n=0}^{\infty} \|\mathbb{E}(\phi(s_n)\phi(s_n)^T \mid \mathcal{F}_n) - \mathbb{E}(\phi(s_n)\phi(s_n)^T)\|_2 \tag{A.6}
\]

\[
\leq 3B(s_0). \tag{A.7}
\]

In particular we know from the above that \(\|\epsilon_n\|_2 \leq 3B(s_0)\).

By Jensen’s inequality, we obtain

\[
\mathbb{E}(\|z_n\|_2) \leq (\mathbb{E}(\|z_n\|_2))^\frac{1}{2}
\]

\[
= \left( \mathbb{E} \left( \|\Pi_n z_0\|_2^2 + \sum_{k=1}^{n} \gamma_k \|\Pi_n \Pi_k^{-1} \epsilon_k\|_2^2 + \|\Pi_n \Pi_k^{-1} \Delta M_k\|_2^2 + \left\langle \Pi_n z_0, \sum_{k=1}^{n} \gamma_k \Pi_n \Pi_k^{-1} \Delta M_k \right\rangle \right) \right)^\frac{1}{2}
\]

\[
\leq \sqrt{1 + 9B(s_0)^2} \left( \mathbb{E} \|\Pi_n z_0\|_2^2 + \sum_{k=1}^{n} \gamma_k^2 \mathbb{E} \|\Pi_n \Pi_k^{-1} \epsilon_k\|_2^2 + \sum_{k=1}^{n} \gamma_k^2 \mathbb{E} \|\Pi_n \Pi_k^{-1} \Delta M_k\|_2^2 \right)^\frac{1}{2} \tag{A.8}
\]

where we have used the bound on \(\|\epsilon\|_2\) derived above, together with the fact that \(\Delta M_k\) is a martingale difference. Notice that for any vector \(x\)

\[
x^T \Phi^T \Psi P \Phi x = (x^T \Phi^T, P \Phi x)^T \Psi \leq \|x^T \Phi^T\|_\Psi \|P \Phi x\|_\Psi = \|x^T \Phi^T\|_\Psi^2,
\]

where we have used Lemma 1 in [Tsitsiklis and Van Roy, 1997] for the final equality. So we deduce that \(A - (1 - \beta)\mu I\) is positive definite by (A3), and hence

\[
\|\Pi_n \Pi_k^{-1}\|_2 = \left\| \prod_{j=k+1}^{n} (I - \gamma_j A) \right\|_2 \leq \prod_{j=k+1}^{n} \|(1 - \gamma_j (1 - \beta)\mu) I - \gamma_j (A - (1 - \beta)\mu I)\|_2 \leq \prod_{j=k+1}^{n} \|(1 - \gamma_j) (1 - \beta)\mu I\|_2 \leq \prod_{j=k+1}^{n} \exp(- (1 - \beta)\mu (\Gamma_n - \Gamma_k)), \tag{A.9}
\]

By using the definition of \(f_X(s)\), the martingale difference can be rewritten as \(\Delta M_n := \beta \theta_n^\top [\phi(s_{n+1}) - \mathbb{E}[\phi(s_{n+1}) \mid \mathcal{F}_n]] \phi(s_n)\). \(\Delta M_n\) can now be bounded as follows:

\[
\mathbb{E}[\|\Delta M_n\|_2^2] \leq \mathbb{E} \left( \beta \|\theta_n\|_2^2 \|\phi(s_{n+1}) - \mathbb{E}[\phi(s_{n+1}) \mid \mathcal{F}_n]\|_2^2 \|\phi(s_n)\|_2^2 \right) \leq 2 \beta H, \tag{A.10}
\]

where the last inequality follows from assumption (A4), and the projection step of the algorithm. The claim now follows by plugging (A.9) and (A.10) into (A.8).
A.2 High Probability Bound

**Theorem 12.** Under (A1)-(A6), we have,

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

where

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

**Proof.** Recall that \(z_n := \theta_n - \theta^*\). First, we rewrite \(\|z_n\|_2^2 \leq \mathbb{E}[\|z_n\|_2^2] = \sum_{i=1}^{n} D_i\), as a telescoping sum of martingale differences:

\[
\|z_n\|_2^2 \leq \mathbb{E}[\|z_n\|_2^2] = \sum_{i=1}^{n} g_i - \mathbb{E}[g_i | \mathcal{F}_{i-1}] = \sum_{i=1}^{n} D_i,
\]

where \(D_i := g_i - \mathbb{E}[g_i | \mathcal{F}_{i-1}], g_i := \mathbb{E}[\|z_n\|_2 | \theta_i],\) and \(\mathcal{F}_i\) denotes the sigma algebra generated by the iterates \(\{\theta_1, \ldots, \theta_i\}\).

The proof is given through two lemmas. The first lemma establishes that the functions \(g_i\) are Lipschitz continuous with Lipschitz constants \(L_i\), and makes use of assumption (A4). This is the crucial ingredient to invoke the concentration bound then given in Lemma 14.

**Lemma 13.** Recall that \(X_n = (s_n, s_{n+1})\) and \(f_{X_n}(\theta) := (r(s_n, \pi(s_n)) + \beta \theta_{n-1} - \theta \phi(s_n)\phi(s_n))\). Then, conditioned on \(\mathcal{F}_{i-1}\), the functions \(g_i\) are Lipschitz continuous in \(f_{X_i}(\theta_{i-1})\), the random innovation at time \(i\), with constants

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

**Proof.** Let \(\Theta_j^i(\theta)\) denote the mapping that returns the value of the iterate \(\theta_j\) at instant \(j\), given that \(\theta_i = \theta\).

\[
\Theta_{j+1}^i(\theta) - \Theta_{j+1}^i(\theta') = \Theta_j^i(\theta) - \Theta_j^i(\theta') - \gamma_j \left[f_{X_{j+1}}(\Theta_j^i(\theta)) - f_{X_{j+1}}(\Theta_j^i(\theta'))\right]
\]

\[
= \Theta_j^i(\theta) - \Theta_j^i(\theta') - \gamma_j \left[\phi(s_{j+1})\phi(s_{j+1})^T - \beta \phi(s_{j+1})\phi(s_{j+2})^T\right]((\Theta_j^i(\theta) - \Theta_j^i(\theta'))
\]

\[
= [I - \gamma_j \left[\phi(s_{j+1})\phi(s_{j+1})^T - \beta \phi(s_{j+1})\phi(s_{j+2})^T\right]]((\Theta_j^i(\theta) - \Theta_j^i(\theta'))
\]

The second equality follows from the definition of \(f_j\). Let \(a_{j+1} := [\phi(s_{j+1})\phi(s_{j+1})^T - \beta \phi(s_{j+1})\phi(s_{j+2})^T]\). Then note that

\[
a_{j+1}^Ta_{j+1} = \phi(s_{j+1})\phi(s_{j+1})^T \phi(s_{j+1})\phi(s_{j+1})^T
\]

\[- 2\beta \phi(s_{j+1})\phi(s_{j+2})\phi(s_{j+1})^T + \beta^2 \phi(s_{j+2})\phi(s_{j+1})^T \phi(s_{j+2})^T
\]

\[
= \|\phi(s_{j+1})\|^2 \phi(s_{j+1})\phi(s_{j+1})^T - \beta(2 - \|\phi(s_{j+1})\|^2) \beta \phi(s_{j+2})\phi(s_{j+2})^T,
\]

where the in the first inequality we have used that for two column vectors of equal dimension, \(x\) and \(y\), \((xy)^T = yx^T\), and \((xx)^T = xx^T\).
Now, from the tower property of conditional expectations, it follows that:

\[
\begin{align*}
\mathbb{E} \left[ \left\| \Theta_{n+1}(\theta) - \Theta_{n+1}(\theta') \right\|_2^2 \right] \\
&= \mathbb{E} \left[ \left\| \Theta_{n+1}(\theta) - \Theta_{n+1}(\theta') \right\|_2^2 \mid \Theta_n(\theta), \Theta_n(\theta') \right] \\
&= \mathbb{E} \left[ (\Theta_n^i(\theta) - \Theta_n^i(\theta'))^T \left( I - 2\gamma_{n+1} \mathbb{E} \left[ (a_{n+1} - \frac{\gamma_{n+1} a_{n+1}^T a_{n+1}}{2} \mid \Theta_{n-1}(\theta), \Theta_{n-1}(\theta') \right] \right) (\Theta_n^i(\theta) - \Theta_n^i(\theta')) \right] \\
&= \mathbb{E} \left[ (\Theta_n^i(\theta) - \Theta_n^i(\theta'))^T \left( I - 2\gamma_{n+1} \left( \mathbb{E}_\Psi \left[ (a_{n+1} - \gamma_{n+1} a_{n+1}^T a_{n+1}/2) \mid \Theta_{n-1}(\theta), \Theta_{n-1}(\theta') \right] + \epsilon_{n+1}' \right) \right) (\Theta_n^i(\theta) - \Theta_n^i(\theta')) \right]
\end{align*}
\]

where

\[
\epsilon_{n+1}' := \mathbb{E} \left[ (a_{n+1} - \gamma_{n+1} a_{n+1}^T a_{n+1}/2) \mid \Theta_{n-1}(\theta), \Theta_{n-1}(\theta') \right] - \mathbb{E}_\Psi \left[ (a_{n+1} - \gamma_{n+1} a_{n+1}^T a_{n+1}/2) \right]
\]

By a similar argument to that used in (8), we have from (A6)

\[
\sum_{k=1}^\infty \| \epsilon_k' \|_2^2 \leq \sum_{n=1}^\infty (1 - \gamma_k/2) \| \mathbb{E}(\beta \phi(s_k)\phi(s_k)^T \mid \mathcal{F}_n) - \mathbb{E}_\Psi(\beta \phi(s_k)\phi(s_k)^T) \|_2^2 \\
+ \beta \sum_{n=1}^\infty \| \mathbb{E}(\beta \phi(s_k)\phi(s_k+1)^T \mid \mathcal{F}_n) - \mathbb{E}_\Psi(\beta \phi(s_k)\phi(s_k+1)^T) \|_2^2 \\
+ \beta (2 - \beta) \sum_{n=1}^\infty \| \mathbb{E}(\phi(s_k+1)\phi(s_k+1)^T \mid \mathcal{F}_n) - \mathbb{E}_\Psi(\phi(s_k+1)\phi(s_k+1)^T) \|_2^2 \\
\leq [1 + \beta(3 - \beta)] B(s_0)
\]

and, consequently, \( \| \epsilon_k' \|_2^2 \leq [1 + \beta(3 - \beta)] B(s_0) \) for all \( k \). Now

Setting \( \Delta \) be the diagonal matrix with entries \( \Delta_{i,i} = \Phi_{i,1,d}^T \Phi_{i,1,d} \). Then we find that for any vector \( \theta \):

\[
\theta^T \mathbb{E}_\Psi \left[ a_{j+1} - \frac{\gamma_{j+1} a_{j+1}^T a_{j+1}}{2} \right] \theta \\
= \theta^T \Phi^T \left( I - \beta \Psi P - \frac{\gamma_{j+1}}{2} (\Delta - \beta P^T (2I - \Delta) \Psi P) \right) \Phi \theta \\
= \theta^T \Phi^T \left( I - \beta \Psi \Pi P - \frac{\gamma_{j+1}}{2} (\Delta - \beta P^T \Pi P (2I - \Delta) \Psi \Pi P) \right) \Phi \theta \\
\geq \| \Phi \theta \|_\Psi^2 - \beta \| \Phi \theta \|_\Psi^2 - \frac{\gamma_{j+1}}{2} \theta^T \Phi (\Delta - \beta \Pi^T (2I - \Delta) \Psi \Pi P) \Phi \theta \geq \| \Phi \theta \|_\Psi^2 - \beta \| \Phi \theta \|_\Psi^2 - \frac{\gamma_{j+1}}{2} \| \Phi \theta \|_\Psi^2 \\
\geq \| \Phi \theta \|_\Psi^2 - \beta \| \Phi \theta \|_\Psi^2 - \frac{\gamma_{j+1}}{2} \| \Phi \theta \|_\Psi^2 + \frac{\gamma_{j+1}}{2} \beta (2 - \beta) \| \Pi P \Phi \theta \|_\Psi^2 \\
\geq \mu \left( 1 - \beta - \frac{\gamma_{j+1}}{2} \right) \| \theta \|_2^2,
\]

where (A.13) follows from the fact that \( \theta^T \Phi^T \Psi (I - \Pi) x = 0 \) since \( \Pi \) is a projection, (A.14) by an application of Cauchy-Schwarz inequality and from the non-expansiveness property of \( \Pi \) and \( P \). (A.15) and (A.16) follow from the fact that, by (A4), the matrix \( \Delta - I \) is positive semi-definite. The final inequality (A.17) follows since (A3).
Lemma 14. Under the conditions of Theorem 11, we have

\[ \mathbb{E} \left[ \| \Theta_i^n(\theta) - \Theta_i^n(\theta') \|_2^2 \right] \leq \left( 1 - 2\gamma_n \left( \mu \left( 1 - \beta - \frac{\gamma_n}{2} \right) + [1 + \beta(3 - \beta)] B(s_0) \right) \right) \mathbb{E} \left[ \| \Theta_{n-1}^i(\theta) - \Theta_{n-1}^i(\theta') \|_2^2 \right] \]

Hence we can conclude that:

\[ \mathbb{E} \left[ \| \Theta_i^n(\theta) - \Theta_i^n(\theta') \|_2^2 \right] \leq \prod_{j=i+1}^{n} \left( 1 - 2\gamma_j \left( \mu \left( 1 - \beta - \frac{\gamma_j}{2} \right) + [1 + \beta(3 - \beta)] B(s_0) \right) \right) \| \theta - \theta' \|_2^2 \]

and, finally, we have

\[ \| \mathbb{E} \left[ \| \theta_n - \theta^* \|_2 \mid \theta_{i-1}, f_i = f \right] - \mathbb{E} \left[ \| \theta_n - \theta^* \|_2 \mid \theta_{i-1}, f_i = f' \right] \|_2 \leq \mathbb{E} \left[ \| \Theta_i^n(\theta) - \Theta_i^n(\theta') \|_2 \right] \]

\[ \leq \prod_{j=i+1}^{n} \left( 1 - 2\gamma_j \left( \mu \left( 1 - \beta - \frac{\gamma_j}{2} \right) + [1 + \beta(3 - \beta)] B(s_0) \right) \right)^{\frac{1}{2}} \gamma_i \| f - f' \|_2 = L_i \| f - f' \|_2 . \]

In the following lemma, we invoke a standard martingale concentration bound using the \(L_i\)-Lipschitz property of the \(g_i\) functions and the assumption (A3).

**Lemma 14.** Under the conditions of Theorem 11 we have

\[ P(\| z_n \|_2 - \mathbb{E} \| z_n \|_2 \geq \epsilon) \leq \exp(-\lambda \epsilon) \exp \left( \frac{\alpha \lambda^2}{2} \sum_{i=1}^{n} L_i^2 \right) . \] (A.18)

**Proof.** Note that

\[ P(\| z_n \|_2 - \mathbb{E} \| z_n \|_2 \geq \epsilon) = P\left( \sum_{i=1}^{n} D_i \geq \epsilon \right) \leq \exp(-\lambda \epsilon) \mathbb{E} \left( \exp \left( \lambda \sum_{i=1}^{n} D_i \right) \right) \]

\[ = \exp(-\lambda \epsilon) \mathbb{E} \left( \exp \left( \lambda \sum_{i=1}^{n-1} D_i \right) \mathbb{E} \left( \exp(\lambda D_n) \mid \mathcal{F}_{n-1} \right) \right) . \]

The first equality above follows from (A.11), while the inequality follows from Markov inequality. Now for any bounded random variable \( f \), and \(L\)-Lipschitz function \( g \) we have

\[ \mathbb{E} (\exp(\lambda g(f))) \leq \exp \left( \lambda^2 L^2 / 2 \right) . \]

Note that each \( f_i(\theta_{i-1}) \) is a bounded random variable by (A3), and, conditioned on \( \mathcal{F}_{i-1} \), \( g_i \) is Lipschitz in \( f_i(\theta_{i-1}) \) with constant \( L_i \) (Lemma 13). So we obtain

\[ \mathbb{E} (\exp(\lambda D_n) \mid \mathcal{F}_{n-1}) \leq \exp \left( \frac{\lambda^2 L_n^2}{2} \right) , \]

and so

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

The proof of Theorem 11 follows by optimizing over \( \lambda \) in (A.18).
A.3 Derivation of rates (Proof of Theorem 11)

Theorem 15. Under (A1)-(A6), choosing $\gamma_n = \frac{c}{(c+n)^2}$ we have, for any $\delta > 0$, 
\[ E \| \theta_n - \theta^* \|_2 \leq \frac{K_1(n)}{\sqrt{n+c}} \] when $(1 - \beta)^2 \mu c > 1/2$, 
\[ & \& \mathbb{P} \left( \| \theta_n - \theta^* \|_2 \geq \frac{K_2(n)}{\sqrt{n+c}} \right) \geq 1 - \delta \text{ when } (\mu(1 - \beta)/2 + 3B(s_0))c \in (1, \infty), \] (A.19)

where
\[ K_1(n) := \sqrt{1 + 9B(s_0)^2} \left[ \frac{\| \theta_0 - \theta^* \|_2}{(n+c)^2(1-\beta)^2 \mu c - 1/2} + \frac{(1 - \beta)c^2\beta H}{\sqrt{1 - 2(1 - \beta)^2 \mu c}} \right] \]
and
\[ K_2(n) := \sqrt{1 + 9B(s_0)^2} \left( \frac{(1 - \beta)c \ln(1/\delta)}{\mu(1 - \beta)/2 + 3B(s_0)^2 c - 1} + K_1(n) \right) \]

High probability bound: Note that when $\gamma_n = \frac{(1 - \beta)c}{(c+n)^2}$,
\[ \sum_{i=1}^{n} L_i^2 = \sum_{i=1}^{n} \frac{(1 - \beta)^2 c^2}{(c+i)^2} \prod_{j=i}^{n} \left( 1 - \frac{2(1 - \beta)c}{c+j} \left( \mu(1 - \beta - \frac{1 - \beta}{2(c+j)}) + [1 + \beta(3 - \beta)] B(s_0) \right) \right) \]
\[ \leq \sum_{i=1}^{n} \frac{(1 - \beta)^2 c^2}{(c+i)^2} \prod_{j=i}^{n} \left( 1 - \frac{2(1 - \beta)c}{c+j} \left( \mu(1 - \beta)/2 + 3B(s_0) \right) \right) \]
\[ \leq \sum_{i=1}^{n} \frac{(1 - \beta)^2 c^2}{(c+i)^2} \exp \left( - (\mu(1 - \beta)/2 + 3B(s_0) c) \right) \]
\[ \leq \frac{(1 - \beta)^2 c^2}{(n+c)(\mu(1-\beta)/2+3B(s_0)c)} \sum_{i=1}^{n} (i+c)^{-2 - (\mu(1 - \beta)/2 + 3B(s_0)c)} \]

We now find three regimes for the rate of convergence, based on the choice of $c$:
(i) $\sum_{i=1}^{n} L_i^2 = O \left( (n+c)^{\mu(1-\beta)/2 + 3B(s_0)c} \right)$ when $(\mu(1 - \beta)/2 + 3B(s_0)c) c \in (0,1)$,
(ii) $\sum_{i=1}^{n} L_i^2 = O \left( n^{-1} \ln n \right)$ when $(\mu(1 - \beta)/2 + 3B(s_0)c) c = 1$, and
(iii) $\sum_{i=1}^{n} L_i^2 \leq \frac{(1 - \beta)^2 c^2}{(\mu(1-\beta)/2+3B(s_0)c)}^{-1} (n+c)^{-1}$ when $(\mu(1 - \beta)/2 + 3B(s_0)c) c \in (1, \infty)$.
(We have used comparisons with integrals to bound the summations.) Setting $c$ so that we are in regime (iii), the high probability bound from Theorem 11 gives
\[ \mathbb{P} \left( \| \theta_n - \theta^* \|_2 - E \| \theta_n - \theta^* \|_2 \geq \epsilon \right) \leq \exp \left( - \frac{\epsilon^2 (n+c)}{2K_{\mu,c,\beta}} \right) \] (A.20)

where $K_{\mu,c,\beta} := (1 - \beta)^2 c^2 / ((\mu(1 - \beta)/2 + 3B(s_0)c) c - 1)$.

Expectation bound: Under the same choice of step-size, and supposing that $c$ is chosen so that $(1 - \beta)^2 \mu c > 1/2$ the bound in expectation in Theorem 11 we have:
\[ \sum_{k=1}^{n-1} 2\beta H_k^2 \gamma_{k+1} \exp(-2(1 - \beta)\mu (\Gamma_n - \Gamma_{k+1})) \]
\[ \leq \frac{(1 - \beta)^2 c^2 (2\beta H_k^2)^2}{(n+c)^{2(1-\beta)^2 \mu c}} \sum_{k=1}^{n} (c+k)^{-2(2(1 - \beta)^2 \mu c)} \leq \frac{(1 - \beta)^2 c^2 H^2}{1 - 2(1 - \beta)^2 \mu c} \frac{1}{c+n} \]

where, in the last inequality, we have again 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
\[
E \|\theta_n - \theta^*\|_2 \leq (\sqrt{c} \|\theta_0 - \theta^*\|_2 + (1 - \beta)c2\beta H)(c + n)^{-\frac{1}{2}},
\] (A.21)
and the result in Theorem 16 follows.

**B  Convergence rate of TD(0) with Iterate Averaging**

**Theorem 16.** 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 \(\delta > 0\),

\[
E \left\| \theta_n - \hat{\theta}_T \right\|_2 \leq \frac{K_1^{IA}(n)}{(n + c)^{\alpha/2}} \quad \text{and} \quad L \left( \left\| \theta_n - \hat{\theta}_T \right\|_2 \right) \geq 1 - \delta,
\] (B.1)

where

\[
K_1(n) := \sqrt{1 + 9B(s_0)^2} \left[ \frac{\|\theta_0 - \theta^*\|_2}{(n + c)^{(1-\alpha)/2}} + \frac{(1 - \beta)c2\beta H}{\mu c(1 - \beta)^2 (1 + \frac{2\alpha}{1 + \frac{2\alpha}{1+2\alpha}})} \right]
\]

and

\[
K_2(n) := \sqrt{1 + 9B(s_0)^2} \left[ \frac{3^\alpha + \left[ \mu \left( \frac{1-\beta}{2} + B(s_0) \right) \right]^{\frac{2\alpha}{1 + \frac{2\alpha}{1+2\alpha}}} + \frac{2\alpha}{\alpha}}{\mu \left( \frac{1}{2} + \frac{B(s_0)}{1-\beta} \right) n^{(1-\alpha)/2}} + K_1(n).
\]

**B.1  Proof of Lemma 10**

**Proof.** We perform the calculation:

\[
\sum_{i=1}^{n} L_i^2 = \sum_{i=1}^{n} \left[ \frac{\gamma_i}{n} \left( 1 + \sum_{l=1+1}^{n-1} \prod_{j=i}^{l} \left( 1 - 2\gamma_{j+1} + 1 \mu \left[ \left( (1 - \beta) - \frac{\gamma_{j+1}}{2} \right) + (1 + \beta(3 - \beta))B(s_0) \right] \right) \right) \right]^2
\]

\[
\leq \frac{1}{n^2} \sum_{i=1}^{n} \left[ \gamma_i \left( 1 + \sum_{l=i+1}^{n-1} \exp \left( - \mu \left[ \frac{1 - \beta}{2} + B(s_0) \right] \prod_{j=i}^{l} \left( 1 \right) \right) \right) \right]^2
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} \left[ \frac{1 - \beta}{(c + i)^\alpha} \left( 1 + \sum_{l=i+1}^{n-1} \exp \left( - \mu \left[ \frac{1 - \beta}{2} + B(s_0) \right] \prod_{j=i}^{l} \left( 1 \right) \right) \right) \right]^2
\]

\[
\leq \left( \frac{1 - \beta}{\mu \left[ \frac{1-\beta}{2} + B(s_0) \right] n} \right)^2 \sum_{i=1}^{n} \left( \frac{1}{(c + i)^\alpha} \right) \quad \text{and}
\]

\[
\sum_{i=1}^{n} \left[ \frac{1 - \beta}{2} + B(s_0) \right] c^{\alpha} \left( (c + i)^1 - (c + i)^{1-\alpha} \right) \left( (c + l + 2)^\alpha - (c + l + 1)^\alpha \right)^2
\]

\[
\leq \left( \mu \left[ \frac{1}{2} + B(s_0) \right] n \right)^{-2} \left( \left[ 2\alpha \frac{\mu \left[ \frac{1-\beta}{2} + B(s_0) \right] c^{\alpha} + \frac{2\alpha}{\alpha} \right] n \right)
\]

19
In the second equality, we have substituted $\gamma_i = (1 - \beta) \left( \frac{\gamma_i}{T + \eta} \right)^\alpha$. For the second inequality, we have used an Abel transform (see page 15 of Fathi and Frikha [2013], display (2.2), for details). For the last inequality, we have noted, as in page 15 of Fathi and Frikha [2013], that

$$(A) := \sum_{l=n+1}^{n-1} \exp \left( -\mu \left[ \frac{(1 - \beta)}{2} + B(s_0) \right] c^\alpha (1 - \beta)^2 ((c + l)^{1-\alpha} - (c + i)^{1-\alpha}) / (1 - \alpha) \right)$$

$$\cdot ((c + l + 2)^\alpha - (c + l + 1)^\alpha)$$

$$\leq \frac{1}{1 - \alpha} \exp \left( \mu \left[ \frac{1 - \beta}{2} + B(s_0) \right] c^\alpha (1 - \beta)^2 (c + i)^{1-\alpha} / (1 - \alpha) \right)$$

$$\cdot \int_{(c+i+1)^{1-\alpha}}^{(c+n)^{1-\alpha}} \exp \left( \mu \left[ \frac{(1 - \beta)}{2} + B(s_0) \right] c^\alpha (1 - \beta)^2 l / (1 - \alpha) \right) l^{\frac{2\alpha - 1}{1 - \alpha}} dl.$$

Now, by taking the derivative and setting it to zero, we find that

$$l \mapsto \exp \left( \mu \left[ \frac{(1 - \beta)}{2} + B(s_0) \right] c^\alpha (1 - \beta)^2 l / (1 - \alpha) \right) l^{\frac{2\alpha - 1}{1 - \alpha}}$$

is decreasing on $[2\alpha / \mu \left[ (1 - \beta)/2 + B(s_0) \right] c^\alpha (1 - \beta)^2, \infty)$, and so we deduce that $(A) \leq (c + i + 1)^\alpha / \alpha$ when $c + i \geq 2\alpha / \mu \left[ (1 - \beta)/2 + B(s_0) \right] c^\alpha (1 - \beta)^2$. When $c + i < 2\alpha / \mu c^\alpha (1 - \beta)^2$ we use that the summand is bounded by 1.

**B.2 Proof of Lemma 9**

**Proof.** Substituting $\gamma_n = (1 - \beta) \left( c / (c + n) \right)^{-\alpha}$ for some $\alpha \in \left( \frac{1}{2}, 1 \right)$ gives

$$\mathbb{E} \| \theta_n - \theta \|^2_2 \leq \exp \left( -\mu c^\alpha (1 - \beta)^2 (n + c)^{1-\alpha} \right) \| \theta_0 - \theta^* \|_2$$

$$+ \left[ \sum_{k=1}^{n} (2\beta H)^2 c^2 (1 - \beta)^2 \frac{c}{k + c + 1} \right]^{2\alpha} \exp \left( -\mu c^\alpha (1 - \beta)^2 \left( (n + c)^{1-\alpha} - (k + 1 + c)^{1-\alpha} \right) \right)$$

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

$$\cdot \left[ \| \theta_0 - \theta^* \|_2 + 2\beta H c^\alpha (1 - \beta) \left\{ \int_0^{n} x^{-2\alpha} \exp(\mu c^\alpha (1 - \beta)^2 x^{1-\alpha}) dx \right\}^{\frac{1}{2}} \right]$$

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

$$\cdot \left[ \| \theta_0 - \theta^* \|_2 + 2\beta H c^\alpha (1 - \beta) \left\{ \frac{\mu c^\alpha (1 - \beta)^2}{2 \alpha} \int_0^{\frac{\mu c^\alpha (1 - \beta)^2}{2 \alpha} (1 - \alpha) y^{-2\alpha} - \alpha y^{-(1+\alpha)} \exp(y^{1-\alpha}) dy \right\}^{\frac{1}{2}} \right]$$

$$\leq \exp( -\mu c^{1-\alpha} ) \| \theta_0 - \theta^* \|_2 + 2\beta H c^\alpha (1 - \beta) \left\{ \frac{\mu c^\alpha (1 - \beta)^2}{2 \alpha} \int_0^{\frac{\mu c^\alpha (1 - \beta)^2}{2 \alpha} (1 - \alpha) y^{-2\alpha} - \alpha y^{-(1+\alpha)} \exp(y^{1-\alpha}) dy \right\}^{\frac{1}{2}}$$

$$\leq \exp( -\mu c^{1-\alpha} ) \| \theta_0 - \theta_T \|_2 + 2\beta H c^\alpha (1 - \beta) \left\{ \frac{\mu c^\alpha (1 - \beta)^2}{2 \alpha} \int_0^{\frac{\mu c^\alpha (1 - \beta)^2}{2 \alpha} (1 - \alpha) y^{-2\alpha} - \alpha y^{-(1+\alpha)} \exp(y^{1-\alpha}) dy \right\}^{\frac{1}{2}}$$
So we have
\[ E \| \hat{\theta}_n - \theta^* \|_2 \leq \sum_{n=1}^{\infty} \exp(-\mu c(n+c)^{1-\alpha}) \| \theta_0 - \theta^* \|_2 n^{-1} \]
\[ + 2\beta H c^{\alpha}(1 - \beta) \left( \mu c^{\alpha}(1 - \beta)^2 \right)^{-\frac{\alpha+\delta_0}{\alpha(1-\alpha)}} (n+c)^{-\frac{1}{2}}. \]
\[ \square \]

### C Convergence rate of TD(0) with Centering

Let \( X_n = (s_n, s_{n+1}) \) and \( f_{X_n}(\theta) := \langle r(s_n, \pi(s_n)) + \beta \theta_{n-1} \phi(s_{n+1}) - \theta^*_n \phi(s_n) \rangle \). Each epoch is of length \( M \).

Let \( \bar{\theta}^{(m)} \), be chosen uniformly from the iterates \( \{\theta_{(m-1)M}, \ldots, \theta_{mM-1}\} \), with \( \theta_0 \) chosen arbitrarily. Then, the update rule for the centered TD algorithm is given by: For \( n = mM + 1, \ldots, (m+1)M \) update

\[
\theta_{n+1} = \theta_n + \gamma \left( f_{X_n}(\theta_n) - f_{X_n}(\bar{\theta}^{(m)}) + \frac{1}{M} \sum_{i=(m-1)M}^{mM} f_{X_i}(\bar{\theta}^{(m)}) \right), \tag{C.1}
\]

where \( i_n \) is drawn uniformly at random in \( \{(m-1)M, \ldots, mM-1\} \), and \( \theta_{mM} := \bar{\theta}^{(m)} \).

**Theorem 17.** Let \( \theta^* \) denote the solution of \( F(\theta) = 0 \). Let the epoch length \( M \) of the CTD algorithm (15) be chosen such that \( C_1 \),

\[
C_1 := \left( \frac{1}{2\mu \gamma M((1-\beta) - d^2 \gamma)} + \frac{\gamma d^2}{2((1-\beta) - d^2 \gamma)} \right)
\]

Then, under (A1)-(A4) and (A6), we have,

\[
\| \Phi(\bar{\theta}^{(m)} - \theta^*) \|^2 \leq C_1^m \left( \| \Phi(\bar{\theta}^{(0)} - \theta^*) \|^2 \right) + C_2 \sum_{k=1}^{m-1} C_1^{(m-2)-k} \sum_{i=(k-1)M}^{kM} E_{\theta_i} \| \epsilon_i \|^2 \]

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

In particular, if the Markov chain underlying policy \( \pi \) satisfies the following property:

\[
| P(s_t = s \mid s_0) - \psi(s) | \leq C \rho^{t/M}, \tag{C.2}
\]

then,

\[
\| \Phi(\bar{\theta}^{(m)} - \theta^*) \|^2 \psi \leq C_1^m \left( \| \Phi(\bar{\theta}^{(0)} - \theta^*) \|^2 \right) + CC_2 \max \{ C_1, \rho \}^{(m-1)} \tag{C.3}
\]

**Proof.** The proof proceeds along the following steps:

**Step 1: Rewriting the recursion (15).**

We rewrite the last term in (15) above as follows:

\[
\frac{1}{M} \sum_{i=(m-1)M}^{mM} f_{X_i}(\bar{\theta}^{(m)}) = \mathbb{E}_{\psi}(f_{X_{mM}}(\bar{\theta}^{(m)})) + \mathbb{E}(f_{X_{mM}}(\bar{\theta}^{(m)}) \mid F_{mM}) - \mathbb{E}_{\psi}(f_{X_{mM}}(\bar{\theta}^{(m)}))
\]
\[
+ \frac{1}{M} \sum_{i=(m-1)M}^{mM} f_{X_i}(\bar{\theta}^{(m)}) - \mathbb{E}(f_{X_{mM}}(\bar{\theta}^{(m)}) \mid F_{mM})
\]
\[
+ \mathbb{E}_{\psi}(f_{X_{mM}}(\bar{\theta}^{(m)})) + \epsilon_n.
\]

\[3\text{For any } v \in \mathbb{R}^d, \text{ we take } \| v \psi \| := \sqrt{v^t \psi v}. \]
The equality above follows because $\Delta M_n = 0$.

**Step 2: Bounding the variance using the centering term.**

\[
\mathbb{E}_\psi \left( \left\| f_{X,n}(\theta_n) - f_{X,n}(\overline{\theta}(m)) \right\|^2 \right) \\
\leq \mathbb{E}_\psi \left( \left\| f_{X,n}(\theta_n) - f_{X,n}(\theta^*) \right\|^2 + \left\| f_{X,n}(\overline{\theta}(m)) - f_{X,n}(\theta^*) - \mathbb{E}_\psi (f_{X,n}(\overline{\theta}(m))) \right\|^2 \right) \\
\leq \mathbb{E}_\psi \left( \left\| f_{X,n}(\theta_n) - f_{X,n}(\theta^*) \right\|^2 + \mathbb{E}_\psi \left( \left\| f_{X,n}(\overline{\theta}(m)) - f_{X,n}(\theta^*) \right\|^2 \right) \right),
\]

where we have used that for any random variable $\xi$, $\mathbb{E}(\|x - \mathbb{E}x\|^2) \leq \mathbb{E}(\|x\|^2)$.

For any $\theta$, we have

\[
\mathbb{E}_\psi \left( \left\| f_{X,n}(\theta) - f_{X,n}(\theta^*) \right\|^2 \right) \\
= (\theta - \theta^*)^T \mathbb{E}_\psi \left[ (\beta\phi(s_{n+1}) - \phi(s_n))\phi(s_n)^T (\beta\phi(s_{n+1}) - \phi(s_n))^T \right] (\theta - \theta^*) \\
\leq (\theta - \theta^*)^T (\Phi^T (I - \beta P) \Phi)^T \Phi (\theta - \theta^*) \\
\leq d^2 \|\Phi(\theta - \theta^*)\|_F^2.
\]

In the final inequality, we have used $\|\Phi^T \Psi \phi\|_2 \leq d^2$.

Plugging the above in (C.6), we obtain

\[
\mathbb{E}_\psi \left( \left\| f_{X,n}(\theta_n) - f_{X,n}(\overline{\theta}(m)) \right\|_2 + \mathbb{E}_\psi (f_{X,n}(\overline{\theta}(m))) \right) \\
\leq d^2 \left( \|\Phi(\theta_n - \theta^*)\|_F^2 + \|\Phi(\overline{\theta}(m) - \theta^*)\|_F^2 \right)
\]

**Step 3: Analysis for a particular epoch.**

\[
\mathbb{E}_{\theta_n} \|\theta_{n+1} - \theta^*\|_2^2 \leq \|\theta_n - \theta^*\|_2^2 + 2\gamma (\theta_n - \theta^*)^T \mathbb{E}_{\theta_n} \left[ f_{X,n}(\theta_n) - f_{X,n}(\overline{\theta}(m)) + \mathbb{E}(f_{X,n}(\overline{\theta}(m))) \right] \\
+ \gamma^2 \mathbb{E}_{\theta_n} \left( \left\| f_{X,n}(\theta_n) - f_{X,n}(\overline{\theta}(m)) \right\|^2 + \mathbb{E}(f_{X,n}(\overline{\theta}(m))) \right) + \mathbb{E}_{\theta_n} \|\epsilon_n\|_2^2 \\
= \|\theta_n - \theta^*\|_2^2 + 2\gamma (\theta_n - \theta^*)^T \mathbb{E}(f_{X,n}(\theta_n -1)) \\
+ \gamma^2 \mathbb{E} \left( \left\| f_{X,n}(\theta_n) - f_{X,n}(\overline{\theta}(m)) \right\|^2 + \mathbb{E}(f_{X,n}(\overline{\theta}(m))) \right) + \mathbb{E}_{\theta_n} \|\epsilon_n\|_2^2 \\
\leq \|\theta_n - \theta^*\|_2^2 + 2\gamma \|\theta_n - \theta^*\|^T [\Phi^T (I - \beta P) \Phi] (\theta_n - \theta^*) \\
+ \gamma^2 d^2 \left( \|\Phi(\theta_n - \theta^*)\|_F^2 + \|\Phi(\overline{\theta}(m) - \theta^*)\|_F^2 \right) + \gamma^2 \mathbb{E}_{\theta_n} \|\epsilon_n\|_2^2 \\
\leq \|\theta_n - \theta^*\|_2^2 + 2\gamma ((1 - \beta) - d^2\gamma) \|\Phi(\theta_n - \theta^*)\|_F^2 + \gamma^2 d^2 \left( \|\Phi(\overline{\theta}(m) - \theta^*)\|_F^2 \right) + \gamma^2 \mathbb{E}_{\theta_n} \|\epsilon_n\|_2^2
\]

Summing the above inequality over an epoch and noting the LHS term $\mathbb{E}_\psi, \theta_n, \|\theta_{n+1} - \theta^*\|_2^2$ is non-negative, we obtain

\[
2\gamma M ((1 - \beta) - d^2\gamma) \|\Phi(\overline{\theta}(m+1) - \theta^*)\|_F^2 \leq \|\theta_0 - \theta^*\|_2^2 + \gamma^2 Md^2 \left( \|\Phi(\overline{\theta}(m) - \theta^*)\|_F^2 \right) + \gamma^2 \sum_{i=(m-1)M}^{mM} \mathbb{E}_{\theta_i} \|\epsilon_i\|_2^2.
\]
Notice that \((\bar{\theta}^{(m)} - \theta^*)^\top I(\bar{\theta}^{(m)} - \theta^*) \leq \frac{1}{\mu} (\bar{\theta}^{(m)} - \theta^*)^\top \Phi \Psi (\bar{\theta}^{(m)} - \theta^*)\) and hence we obtain the following by setting \(\theta_0 = \bar{\theta}^{(m)}\):

\[
2\gamma M ((1 - \beta) - d^2\gamma) \|\Phi(\bar{\theta}^{(m+1)} - \theta^*)\|_\Psi^2 \leq \left( \frac{1}{\mu} + \gamma^2Md^2 \right) \|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2 + \gamma^2 \sum_{i=(m-1)M}^{mM} \mathbb{E}_{\theta_i} \|\epsilon_i\|_2^2 \]

\[\iff \|\Phi(\bar{\theta}^{(m+1)} - \theta^*)\|_\Psi^2 \leq C_1 \left( \|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2 \right) + C_2 \sum_{i=(m-1)M}^{mM} \mathbb{E}_{\theta_i} \|\epsilon_i\|_2^2 \]

**Step 4: Combining across epochs.**

Unrolling the above recursion across epochs, we obtain

\[
\|\Phi(\bar{\theta}^{(m)} - \theta^*)\|_\Psi^2 \leq C_1 \left( \|\Phi(\bar{\theta}^{(0)} - \theta^*)\|_\Psi^2 \right) + C_2 \sum_{k=1}^{m-1} C_1^{(m-2)-k} \sum_{i=(k-1)M}^{kM} \mathbb{E}_{\theta_i} \|\epsilon_i\|_2^2 
\]

Hence proved.

**References**

Dimitri P Bertsekas. *Approximate dynamic programming*. 2011.

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

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

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

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

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

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

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

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

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

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

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

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