Finite Sample Analysis for TD(0) with Linear Function Approximation

Gal Dalal∗
gald@tx.technion.ac.il

Balázs Szőrényi∗
szorenyi.balazs@gmail.com

Gugan Thoppe∗
gugan.thoppe@gmail.com

Shie Mannor
shie@ee.technion.ac.il

Abstract

TD(0) is one of the most commonly used algorithms in reinforcement learning. Despite this, there is no existing finite sample analysis for TD(0) with function approximation, even for the linear case. Our work is the first to provide such a result. Works that managed to obtain concentration bounds for online Temporal Difference (TD) methods analyzed modified versions of them, carefully crafted for the analyses to hold. These modifications include projections and step-sizes dependent on unknown problem parameters. Our analysis obviates these artificial alterations by exploiting strong properties of TD(0) and tailor-made stochastic approximation tools.

1 Introduction

Temporal Difference (TD) algorithms lie at the core of Reinforcement Learning (RL), dominated by the celebrated TD(0) algorithm. The term has been coined in [Sutton and Barto, 1998], describing an iterative process of updating an estimate of a value function $V^\pi(s)$ with respect to a given policy $\pi$ based on temporally-successive samples. The classical version of the algorithm uses a tabular representation, i.e., entry-wise storage of the value estimate per each state $s \in S$. However, in many problems the state-space $S$ is too large for such a vanilla approach. The common practice to mitigate this caveat is to approximate the value function using some parameterized family. Often, linear regression is used, i.e., $V^\pi(s) \approx \theta^T \phi(s)$. This allows for an efficient implementation of TD(0) even on large state-spaces and has shown to perform well in a variety of problems Tesauro [1995], Powell [2007]. More recently, TD(0) has become prominent in many state-of-the-art RL solutions when combined with deep neural network architectures, as an integral part of fitted value iteration [Mnih et al., 2015, Silver et al., 2016]. In this work we focus on the former case of linear Function Approximation (FA); nevertheless, we consider this work as a preliminary milestone in route to achieving theoretical guarantees for deep RL approaches.

Two types of convergence rate results exist in literature: with high probability and in expectation. We stress that no results of the first type exist for the actual, commonly used, TD(0) algorithm with linear FA; our work is the first to provide such a result. In fact, it is the first work to give a concentration bound for an unaltered online TD algorithm of any type. To emphasize, TD(0) with linear FA is formulated and used with non-problem-specific step-sizes. Also, it does not require a projection step to keep $\theta$ in a 'nice' set. In contrast, the few recent works that managed to provide concentration bounds for TD(0) analyzed only altered versions of them, carefully crafted for the analyses to hold. These modifications include a projection step and eigenvalue-dependent step-sizes; we expand on this in the coming section. As for the second type of results, i.e., expectation bounds, existing results

∗Equal contribution
either apply only to altered versions of TD(0) as described above, or applies to average of iterates. In this work, we obtain the first expectation bound directly on the iterates for the unaltered TD(0).

1.1 Existing Literature

The first TD(0) convergence result was obtained by Tsitsiklis et al. [1997] for both finite and infinite state-spaces. Following that, a key result by Borkar and Meyn [2000] paved the path to a unified and convenient tool for convergence analyses of Stochastic Approximation (SA), and hence of TD algorithms. This tool is based on the Ordinary Differential Equation (ODE) method. Essentially, that work showed that under the right conditions, the SA trajectory follows the solution of a suitable ODE, often referred to as its limiting ODE; thus, it eventually converges to the solution of the limiting ODE. Several usages of this tool in RL literature can be found in [Sutton et al., 2009a,b, 2015].

As opposed to the case of asymptotic convergence analysis of TD algorithms, very little is known on their finite sample behavior. We now briefly discuss the few existing results on this topic. In Borkar [2008], a concentration bound is given for generic SA algorithms. Recent works [Kamal, 2010, Thoppe and Borkar, 2015] obtain better concentration bounds via tighter analyses. The results in these works are conditioned on the event that the $n_0$-th iterate lies in some a-priori chosen bounded region containing the desired equilibria; this, therefore, is the caveat in applying them to TD(0).

In Korda and Prashanth [2015], concentration bounds for TD(0) with mixing-time consideration have been given. However, unlike in our work, a strong requirement for all their high probability bounds is that the iterates need to lie in some a-priori chosen bounded set; this is ensured there via projections (personal communication). Additionally, their results require the learning rate to be set based on prior knowledge about system dynamics, which, as argued in the paper, is problematic; alternatively, they apply to average of iterates. An additional work by Liu et al. [2015] considered the gradient TD algorithms GTD(0) and GTD2, which were first introduced in Sutton et al. [2009b,a]. That work interpreted the algorithms as gradient methods to some saddle-point optimization problem. This enabled them to obtain concentration bounds on altered versions of these algorithms using results from the convex optimization literature. Despite the alternate approach, in similar fashion to the results above, a projection step that keeps the parameter vectors in a convex set is needed there.

Bounds similar in flavor to ours are also given in Frikha and Menozzi [2012, Fathi and Frikha, 2013]. However, they apply only to a class of SA methods satisfying strong assumptions, which do not hold for TD(0). In particular, neither the uniformly Lipschitz assumption nor its weakened version, the Lyapunov Stability-Domination criteria, hold for TD(0) when formulated in their iid noise setup.

Two additional works Yu and Bertsekas [2009, Lazaric et al., 2010] provide sample complexity bounds on the batch LSTD algorithms. However, in the context of finite sample analysis, these belong to a different class of algorithms. The case of online TD learning has proved to be more practical, at the expense of increased analysis difficulty compared to LSTD methods.

1.2 Our Contribution

Our work is the first to give a bound on the convergence rate of TD(0) in its original, unaltered form. In fact, it is the first to obtain a concentration bound for an unaltered online TD algorithm of any type. Indeed, as discussed earlier, existing convergence rates apply only to online TD algorithms with alterations such as projections and step-sizes dependent on unknown problem parameters; alternatively, they only apply to average of iterates. The key ingredients in our approach to obviate these alterations are i) show that the $n$-th iterate at worst is only $O(n)$ away from the solution $\theta^*$; and ii) based on that, show that after some additional steps all subsequent iterates are $\epsilon$-close to the solution w.h.p. We believe this approach is not limited to TD(0) alone.

Moreover, we provide the first expectation decay rate of the actual TD(0) iterates. It applies for a general family of step-sizes that is not restricted to square-summable sequences, as is assumed in most works.

2 Problem Setup

We consider the problem of policy evaluation for a Markov Decision Process (MDP). A MDP is defined by the 5-tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ [Sutton, 1988], where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of
actions, $P = P(s'|s, a)$ is the transition kernel, $R(s, a, s')$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. In each time-step, the process is in some state $s \in S$, an action $a \in \mathcal{A}$ is taken, the system transitions to a next state $s' \in S$ according to the transition kernel $P$, and an immediate reward $r$ is received according to $R(s, a, s')$. Let policy $\pi : S \rightarrow \mathcal{A}$ be a stationary mapping from states to actions. Assuming the associated Markov chain is ergodic and uni-chain, let $\nu$ be the induced stationary distribution. Moreover, let $V^\pi(s)$ be the value function at state $s$ w.r.t. $\pi$ defined via the Bellman equation $V^\pi(s) = \mathbb{E}_\nu [r + \gamma V^\pi(s')]$. In our policy evaluation setting, the goal is to estimate $V^\pi(s)$ using linear regression, i.e., $V^\pi(s) \approx \phi^\top(s) \phi(s)$, where $\phi(s) \in \mathbb{R}^d$ is a feature vector at state $s$, and $\theta \in \mathbb{R}^d$ is a weight vector.

Let $\{(\phi_n, \phi'_n, r_n)\}_n$ be iid samples of $(\phi, \phi', r)$. Then the TD(0) algorithm has the update rule

$$\theta_{n+1} = \theta_n + \alpha_n [r_n + \gamma \phi^\top_n \theta_n - \phi^\top_n \theta_n] \phi_n,$$

where $\alpha_n$ is the step-size. For analysis, we can rewrite the above as

$$\theta_{n+1} = \theta_n + \alpha_n [h(\theta_n) + M_{n+1}],$$

where $h(\theta) = b - A \theta$ and

$$M_{n+1} = (r_n + \gamma \phi^\top_n \theta_n - \phi^\top_n \theta_n) \phi_n - [b - A \theta_n],$$

with $A = \mathbb{E}_\nu [\phi(\theta - \phi')^\top]$ and $b = \mathbb{E}_\nu [\phi^\top]$. It is known that $A$ is positive definite [Bertsekas 2012] and that (2) converges to $\theta^* := A^{-1} b$ [Borkar 2008]. Note that

$$h(\theta) = -A[\theta - \theta^*].$$

We make the following assumption:

$${\mathcal{A}}_1.$$ All rewards $r(s, a, s')$ and feature vectors $\phi(s)$ are uniformly bounded, i.e., $\|\phi(s)\| \leq 1/2, \forall s \in S$, and $|r(s, a, s')| \leq 1, \forall s, s' \in S, a \in \mathcal{A}$.

### 3 Main Result

Our main result is the following. The $O$ notation hides problem dependent constants and polynomial terms.

**Theorem 1 (TD(0) Concentration Bound).** Let $\lambda \in (0, \min_{i \in [d]} \{\text{real}(\lambda_i(A))\})$, where $\lambda_i(A)$ is the $i$-th eigenvalue of $A$. Let $\alpha_n = (n + 1)^{-1}$. Then for $\epsilon > 0$ and $\delta \in (0, 1)$, there exists a function

$$N(\epsilon, \delta) = O\left(\max\left\{\left[\frac{1}{\epsilon}\right]^{1+\frac{1}{2}}, \left[\frac{1}{\epsilon}\right]^{1+\frac{1}{2}}, \left[\frac{1}{\epsilon}\right]^{2}, \left[\frac{\lambda}{\delta}\right]^{3}\right\}\right)$$

such that

$$\Pr\{\|\theta_n - \theta^*\| \leq \epsilon \ \forall n \geq N(\epsilon, \delta)\} \geq 1 - \delta.$$ 

The proof of Theorem 1 also gives the following result; instead of fixed $\epsilon$, we have a decreasing rate.

**Theorem 2.** Let $\lambda, \alpha_n$ be as in Theorem 1. Fix $\delta \in (0, 1)$. Then there exists some function $N_0(\delta)$ such that for all $n \geq N_0(\delta)$,

$$\Pr\{\|\theta_n - \theta^*\| = O\left(n^{-\min\{1/2, \lambda/(\lambda+1)\}}\right)\} \geq 1 - \delta.$$ 

**Remark 1.** Theorem 1, [Korda and Prashanth 2015] requires the TD(0) step-sizes to satisfy: $\alpha_n = f_n(\lambda)$ for some function $f_n$, where $\lambda$ is as above. Further, Theorem 2 there applies to average of iterates. Also, concentration bounds in these results require projecting the iterates to some bounded set (personal communication). In contrast, our result applies directly to the original TD(0) algorithm and we obviate all the above modifications. However, our result is weaker than Theorem 1 there when $\lambda < 1$.

Our other main result is a bound on the expected decay rate of the TD(0) iterates.
Table 1: Chronological Summary of Analysis Outline

| Stepsize | Martingale Noise Impact | Discretization Error | TD(0) Behavior |
|----------|-------------------------|----------------------|----------------|
| Large    | Large $O(n_0)$ w.h.p.   | Large $O(n_0)$       | Diverging      |
| Moderate | $\epsilon/3$ w.h.p.     | $\epsilon/3$         | Stay in $O(n_0)$ ball w.h.p. |
| Small    |                         |                      | Converging     |

Theorem 3 (Expected Decay Rate for TD(0)). Fix $\sigma \in (0,1)$ and let $\alpha_n = (n+1)^{-\sigma}$. Fix $\lambda \in (0, \lambda_{\min}(A+A^T))$. Then, for $n \geq 0$,

$$E\|\theta_{n+1} - \theta^*\|^2 \leq K_1 e^{-(\lambda/2)(n+2)^{1-\sigma}} + \frac{K_2}{(n+1)\sigma},$$

where $K_1, K_2 \geq 0$ are some constants that depend on both $\lambda$ and $\sigma$; see Theorem 12 for the exact expression.

Remark 2. The exponentially decaying term in Theorem 3 corresponds to the convergence rate of the noiseless TD(0) algorithm, while the inverse polynomial term appears due to the martingale noise $M_n$. The inverse impact of $\sigma$ on these two terms introduces the following tradeoff:

1. For $\sigma$ close to 0, the first term converges faster and corresponds to slowly decaying stepsizes, which, in turn, speed up the noiseless TD(0) convergence.
2. For $\sigma$ close to 1, the second term decays quickly and corresponds to small step-sizes that better mitigate the effect of martingale noise; this originates in the term $\alpha_n M_{n+1}$.

While this insight is folklore, a formal estimate of the tradeoff, to the best of our knowledge, has been obtained here for the first time.

Remark 3. The expectation bound in Theorem 1, [Korda and Prashanth, 2015] again requires the stepsize sequence be scaled as in Remark 1. Theorem 2 obviates this, but it applies to average of iterates. In contrast, our expectation bound applies directly to the TD(0) iterates and does not need any scaling of the above kind. Moreover, our result applies to a broader family of stepsizes; see Remark 4. Our expectation bound when compared to that of Theorem 2, [Korda and Prashanth, 2015] is of the same order (even though theirs is for average of iterates).

Remark 4. In Theorem 3, unlike most works, $\sum_{n \geq 0} \alpha_n^2$ need not be finite. Thus this result is applicable for a wider class of stepsizes; e.g., $1/n^\kappa$ with $\kappa \in (0, 1/2]$. In [Borkar, 2008], on which much of the existing RL literature is based on, the square summability assumption is due to the Gronwall inequality. In contrast, in our work, we use the Variation of Parameters Formula [Lakshmikantham and Deo, 1998] for comparing the SA trajectory to appropriate trajectories of the limiting ODE; it is a stronger tool than Gronwall inequality.

Remark 5. In the proof of Theorem 3, one can see that the requirement on the martingale noise is of the form $E\|M_{n+1}\|^2 \leq C(1 + \|\theta_n\|^2)$ where $C$ is a constant, in correspondence to Remark 4. It is in fact a weaker requirement than the one obtained for TD(0), as is given in Lemma 4.

4 Proof of Theorem 1

This section outlines the analysis conducted for Theorem 1. All proofs are given in Appendix B.

4.1 Outline of Approach

We compare the TD(0) iterates $\{\theta_n\}$ with suitable solutions of its limiting ODE using the Variation of Parameters (VoP) method [Lakshmikantham and Deo, 1998]. As the solutions of the ODE are continuous functions of time, we first define a linear interpolation $\{\bar{\theta}(t)\}$ of $\{\theta_n\}$. Let $t_0 = 0$. For $n \geq 0$, let $t_{n+1} = t_n + \alpha_n$ and let

$$\bar{\theta}(\tau) = \begin{cases} 
\theta_n & \text{if } \tau = t_n, \\
\theta_n + \frac{\tau - t_n}{\alpha_n}[\theta_{n+1} - \theta_n] & \text{if } \tau \in (t_n, t_{n+1}).
\end{cases}$$

(5)
We establish some preliminary results here that will be used throughout this section. Let \( \theta_t \) be as in Theorem 1. From Corollary 3.6, \( \theta_t \) is positive definite, for all \( n \), \( n \geq 1 \) s.t. \( n \geq n_0 \) so that first the TD(0) iterates for \( n \geq n_0 \) stay within a \( O(n) \) distance from \( \theta^* \); then after for some additional time when the stepsizes decay enough the TD(0) iterates start behaving almost like a noiseless version. These three different behaviours are summarized in Table 1 and illustrated in Figure 1.

\[ \theta(t) = h(\theta(t)) = b - A\theta(t) = -A(\theta(t) - \theta^*) . \] (6)

Let \( \theta(t, s, u_0) \), \( t \geq s \), denote the solution to the above ODE starting at \( u_0 \) at time \( t = s \). When the starting point and time are unimportant, we will denote this solution by \( \theta(t) \).

Initially, \( \theta(t) \) could stray away from \( \theta^* \) when the step-sizes may not be small enough to tame the noise. However, we show that \( \|\theta(t, s) - \theta^*\| = O(n) \), i.e., \( \theta_n \) does not stray away from \( \theta^* \) too fast. Later, we show that we can fix some \( n_0 \) so that first the TD(0) iterates for \( n \geq n_0 \) stay within a \( O(n) \) distance from \( \theta^* \); then after for some additional time when the stepsizes decay enough the TD(0) iterates start behaving almost like a noiseless version.

**4.2 Preliminaries**

We establish some preliminary results here that will be used throughout this section. Let \( s \in \mathbb{R} \), and \( u_0 \in \mathbb{R}^d \). Using results from Chapter 6, \[ \text{Hirsch et al. 2012} \], it follows that the solution \( \theta(t, s, u_0) \), \( t \geq s \), of (6) satisfies the relation

\[ \theta(t, s, u_0) = \theta^* + e^{-A(t-s)}(u_0 - \theta^*) . \] (7)

As the matrix \( A \) is positive definite, for \( \theta(t) \equiv \theta(t, s, u_0) \),

\[ \frac{d}{dt}\|\theta(t) - \theta^*\|^2 = -2(\theta(t) - \theta^*)^\top A(\theta(t) - \theta^*) < 0 . \]

Hence

\[ \|\theta(t', s, u_0) - \theta^*\| \leq \|\theta(t, s, u_0) - \theta^*\| , \] (8)

for all \( t' \geq t \geq s \) and \( u_0 \).

Let \( \lambda \) be as in Theorem 1. From Corollary 3.6, \[ \text{Teschl 2012} \], \( \exists K_\lambda \geq 1 \) so that \( \forall t \geq s \)

\[ \|e^{-A(t-s)}\| \leq K_\lambda e^{-\lambda(t-s)} . \] (9)

Separately, as \( t_{n+1} - t_{k+1} = \sum_{t=k+1}^{n} \alpha_t = \sum_{t=k+1}^{n} \frac{1}{t+1} \),

\[ \frac{(k + 1)^\lambda}{(n + 1)^\lambda} \leq e^{-\lambda(t_{n+1} - t_{k+1})} \leq \frac{(k + 2)^\lambda}{(n + 2)^\lambda} . \] (10)

The following result gives a bound on the martingale difference noise as a function of the iterates. We emphasize that this bound is significant in our work and that this strong behavior of TD(0) is usually overlooked in existing literature.
Lemma 4 (Martingale Noise Behavior). For all \( n \geq 0 \),

\[
\|M_{n+1}\| \leq K_m \left[ 1 + \|\theta_n - \theta^*\| \right],
\]

where

\[
K_m := \frac{1}{4} \max \left\{ 2 + [1 + \gamma] A^{-1} \|b\|, 1 + \gamma + 4 \|A\| \right\}.
\]

Remark 6. The noise behavior usually used in the literature (e.g., [Sutton et al., 2009a]) is

\[
\mathbb{E}[\|M_{n+1}\|^2 | \mathcal{F}_n] \leq C(1 + \|\theta_n\|^2),
\]

for some constant \( C \geq 0 \). The result on the noise behavior in Lemma 4 is in fact stronger than that. For easier comparison, we also provide following result (the proof technique is similar to that in Lemma 4). For all \( n \geq 0 \),

\[
\|M_{n+1}\|^2 \leq 3[1 + \gamma + \max(\|A\|, \|b\|)]^2(1 + \|\theta_n\|^2).
\]

The remaining parts of the analysis rely on the comparison of the discrete TD(0) trajectory \( \{\theta_n\} \) to the continuous solution \( \theta(t) \) of the limiting ODE. For this, we first switch from directly treating \( \{\theta_n\} \) to treating their linear interpolation \( \{\bar{\theta}(t)\} \) as defined in [4]. The key idea then is to use the VoP method [Lakshmikantham and Deo, 1998] and express \( \bar{\theta}(t) \) as a perturbation of \( \theta(t) \) due to two factors: the discretization error and the martingale difference noise. This is discussed further in Lemma A in Appendix A.

For the interval \([t_{\ell_1}, t_{\ell_2}]\), let \( E_{[t_{\ell_1}, t_{\ell_2}]}^d := \sum_{k=\ell_1}^{\ell_2-1} \int_{t_k}^{t_{k+1}} e^{-A(t_{k+1}-\tau)} A\bar{\theta}(\tau) - \bar{\theta}(\tau) d\tau \) and \( E_{[t_{\ell_1}, t_{\ell_2}]}^m := \sum_{k=\ell_1}^{\ell_2-1} \int_{t_k}^{t_{k+1}} e^{-A(t_{k+1}-\tau)} d\tau \) \( M_{k+1} \). Corollary 5 below shows that \( \bar{\theta}(t_{\ell_2}) - \theta^* \) differs from \( \theta(t_{\ell_2}, t_{\ell_1}, \bar{\theta}(t_{\ell_1})) - \theta^* \) by \( E_{[t_{\ell_1}, t_{\ell_2}]}^d + E_{[t_{\ell_1}, t_{\ell_2}]}^m \). We highlight that both the paths, \( \bar{\theta}(t) \) and \( \theta(t, t_{\ell_1}, \bar{\theta}(t_{\ell_1})) \), \( t \geq t_{\ell_1} \), start at the same point \( \bar{\theta}(t_{\ell_1}) \) at time \( t_{\ell_1} \). As mentioned above, \( E_{[t_{\ell_1}, t_{\ell_2}]}^d \) and \( E_{[t_{\ell_1}, t_{\ell_2}]}^m \) respectively denote the cumulative discretization error and martingale difference noise over the interval \([t_{\ell_1}, t_{\ell_2}]\).

Corollary 5 (Comparison of SA Trajectory and ODE Solution). For every \( \ell_2 \geq \ell_1 \),

\[
\bar{\theta}(t_{\ell_2}) - \theta^* = \theta(t_{\ell_2}, t_{\ell_1}, \bar{\theta}(t_{\ell_1})) - \theta^* + E_{[t_{\ell_1}, t_{\ell_2}]}^d + E_{[t_{\ell_1}, t_{\ell_2}]}^m.
\]

We shall use this result later in Lemmas 14 and 15 in Appendix B.

4.3 Part I – Initial Possible Divergence

In this section we show that the TD(0) iterates lie in a \( O(n) \)-ball around \( \theta^* \). We emphasize that this is one of the results that enables us to accomplish more than existing literature. Previously, the distance of the initial iterates from \( \theta^* \) was bounded using various assumptions, often justified with an artificial projection step which we are able to avoid.

Let \( R_0 := 1 + \|\theta_0 - \theta^*\| \).

Lemma 6 (Worst-case Iterates Bound). For \( n \geq 0 \),

\[
\|\theta_n - \theta^*\| \leq R_{wc}(n),
\]

where

\[
R_{wc}(n) := [n + 1] C_* R_0
\]

and \( C_* := 1 + \|\theta^*\| \leq 1 + \|A^{-1}\| \|b\| \).

Next, since \( \|M_{n+1}\| \) is linearly bounded by \( \|\theta_n - \theta^*\| \), the following result shows that \( \|M_{n+1}\| \) is \( O(n) \) as well. It follows from Lemmas 4 and 6.

Corollary 7 (Worst-case Noise Bound). For \( n \geq 0 \),

\[
\|M_{n+1}\| \leq K_m [1 + C_* R_0][n + 1].
\]
### 4.4 Part II – Rate of Convergence

Our formal aim here is to obtain an estimate on the probability of the event

\[ \mathcal{E}(n_0, n_1) := \{ \| \theta_n - \theta^* \| \leq \epsilon \ \forall n > n_0 + n_1 \} \]

for sufficiently large \( n_0, n_1 \geq 1 \); how large they ought to be will be elaborated later. We do this by comparing the TD(0) trajectory \( \theta_{n+1} \) with the ODE solution \( \theta(t_{n+1}, t_n, \theta(t_n)) \) \( \forall n \geq n_0 \); for that we use Corollary 5 along with Lemma 6. In this section we show that if \( n_0 \) is sufficiently large, or equivalently the stepsizes \( \{a_n\}_{n \geq n_0} \) are small enough, then after a finite number of iterations from \( n_0 \), the TD(0) iterates are \( \epsilon \)-close to \( \theta^* \) w.h.p. This holds as the small stepsize and sufficiently long waiting time ensure that the ODE solution \( \theta(t_{n+1}, t_n, \theta(t_n)) \) is \( \epsilon \)-close to \( \theta^* \), the discretization error \( E^m_{[n_0, n+1]} \) is small and martingale difference noise \( E^m_{[n_0, n+1]} \) is small w.h.p.

Let \( \delta \in (0, 1) \), and let \( \epsilon \) be such that \( \epsilon > 0 \). Also, for an event \( \mathcal{E} \), let \( \mathcal{E}^c \) denote its complement and let \( \{ \mathcal{E}_1, \mathcal{E}_2 \} \) denote \( \mathcal{E}_1 \cap \mathcal{E}_2 \). We begin with a careful decomposition of \( \mathcal{E}^c(n_0, n_1) \), the complement of the event of interest. The idea is to break it down into an incremental union of events. Each such event has an inductive structure: good up to iterate \( n \) (denoted by \( G_{n_0, n} \) below) and the \((n+1)\)th iterate is bad. The good event \( G_{n_0, n} \) holds when all the iterates up to \( n \) remain in an \( \mathcal{O}(n_0) \) ball around \( \theta^* \). For \( n < n_0 + n_1 \), the bad event means that \( \theta_{n+1} \) is outside the \( \mathcal{O}(n_0) \) ball around \( \theta^* \), while for \( n \geq n_0 + n_1 \), the bad event means that \( \theta_{n+1} \) is outside the \( \epsilon \) ball around \( \theta^* \). Formally, for \( n_1 \geq 1 \), define the events

\[
\mathcal{E}_{n_0, n_1}^{mid} := \bigcup_{n=n_0}^{n_0+n_1-1} \{ G_{n_0, n}, \| \theta_{n+1} - \theta^* \| > 2R_{wc}(n_0) \},
\]

\[
\mathcal{E}_{n_0, n_1}^{after} := \bigcup_{n=n_0+n_1}^{\infty} \{ G_{n_0, n}, \| \theta_{n+1} - \theta^* \| > \min\{\epsilon, 2R_{wc}(n_0)\} \},
\]

and, \( \forall n \geq n_0 \), let

\[
G_{n_0, n} := \left\{ \bigcap_{k=n_0}^{n} \{ \| \theta_k - \theta^* \| \leq 2R_{wc}(n_0) \} \right\}.
\]

Using the above definitions, the decomposition of \( \mathcal{E}^c(n_0, n_1) \) is the following relation.

**Lemma 8 (Decomposition of Event of Interest).** For \( n_0, n_1 \geq 1 \),

\[ \mathcal{E}^c(n_0, n_1) \subseteq \mathcal{E}_{n_0, n_1}^{mid} \cup \mathcal{E}_{n_0, n_1}^{after}. \]

For the following results, define the constants

\[
C_{m2} := \begin{cases} 6K_m^2 \lambda^{2-0.5} & \text{if } \lambda > 0.5 \\ 6K_m \lambda^{2-1} & \text{if } \lambda < 0.5 \end{cases}.
\]

Next we show that on the “good” event \( G_{n_0, n} \), the discretization error is small for all sufficiently large \( n \).

**Lemma 9 (Part II Discretization Error Bound).** For any \( n \geq n_0 \geq K_6\|A\|\|A\|+2K_m \)

\[
\| E^d_{[n_0, n+1]} \| \leq \frac{1}{3} \{ n_0 + 1 \} C_{m} R_{0} = \frac{1}{3} R_{wc}(n_0).
\]

Furthermore, for \( n \geq n_c \geq \left( 1 + \frac{K_6\|A\|\|A\|+2K_m}{\lambda \min\{\epsilon, R_{wc}(n_0)\}} \right) \{ n_0 + 1 \} \) it thus also holds on \( G_{n_0, n} \) that

\[
\| E^d_{[n_c, n+1]} \| \leq \frac{1}{3} \min\{\epsilon, [n_0 + 1] C_{m} R_{0} \} = \frac{1}{3} \min\{\epsilon, R_{wc}(n_0)\}.
\]

The next result gives a bound on the probability that, on the “good” event \( G_{n_0, n} \), the martingale difference noise is small when \( n \) is large. The bound has two forms for the different values of \( \lambda \).
Lemma 10 (Part II Martingale Difference Noise Concentration). Let \( n_0 \geq 1 \) and \( R \geq 0 \). Let \( n \geq n' \geq n_0 \). For \( \lambda > 1/2 \),

\[
\Pr\{G_{n_0,n}, \|E_{n',n+1}^n\| \geq R\} \leq 2d^2 \exp\left(\frac{(n' + 1)^2}{2d^2 \lambda^2 R^2} e^{\lambda C_{n_0}(n)}\right).
\]

For \( \lambda < 1/2 \),

\[
\Pr\{G_{n_0,n}, \|E_{n',n+1}^n\| \geq R\} \leq 2d^2 \exp\left(\frac{[n' + 1]^{1-2\lambda}(n + 1)^{2\lambda} R^2}{2d^2 C_{n_0}^2 \lambda^2 R_w(n)}\right).
\]

Lemmas 8, 9 and 10 are the key ingredients for proving Theorem 1. The detailed proof is given in Appendix B. However, we now outline the underlying idea.

From Lemma 8 by a union bound,

\[
\Pr\{E\mid n_0, n_1\} \leq \Pr\{E\mid n_0, n_1\} \leq \Pr\{E\mid n_0, n_1\} + \Pr\{E\mid n_0, n_1\}.
\]

Next, we use Lemmas 9 and 10 to set \( n_0 \) and \( n_1 \) in the following way to bound the terms on the RHS. The behavior of \( E_{n_0,n} \) is dictated by \( n_0 \), while the behavior of \( E_{n_0,n_1} \) by \( n_1 \). We set \( n_0 \) so that \( E_{n_0,n} \) is less than \( \delta/2 \) by substituting \( R_{n_0}(n_0)/2 \) in \( r \) from Lemma 10, resulting in the condition \( n_0 \leq O(\ln \frac{1}{\delta}) \).

Next, we set \( r = \frac{1}{\lambda} \) for bounding \( \{E\mid n_0,n_1\} \) by \( 2\delta/2 \), resulting in \( n_1 = \hat{O}\left(\left(\frac{1}{\lambda} \ln \frac{1}{\delta}\right)^{1+1/\lambda}\right) \) for \( \lambda > 1/2 \), and \( n_1 = \hat{O}\left(\left(\frac{1}{\lambda} \ln \frac{1}{\delta}\right)^{1+1/\lambda}\right) \) for \( \lambda < 1/2 \).

5 Proof of Theorem 3

The expectation bound is due to an inductive argument and an application of a subtle trick from Kamal [2010]. Building on the approach there, our key steps are: identifying a “nice” Liapunov function \( V \) of the TD(0) method’s limiting ODE; and then using conditional expectation suitably to get rid of the linear noise terms in the relation between \( V(\theta_n) \) and \( V(\theta_{n+1}) \). Induction then leads to the desired result.

Throughout this section only, \( \{\alpha_n\} \) is a stepsize sequence satisfying \( \sum_{n \geq 0} \alpha_n = \infty, \lim_{n \to \infty} \alpha_n = 0 \) and \( \sup_{n \geq 0} \alpha_n \leq 1 \). The proofs in this section are provided in Appendix C.

Recall that all eigenvalues of a symmetric matrix are real. For a symmetric matrix \( X \), let \( \lambda_{\min}(X) \) and \( \lambda_{\max}(X) \) be its minimum and maximum eigenvalues, respectively.

Theorem 11 (Technical Result: Expectation Bound). Fix \( \lambda \in (0, \lambda_{\min}(A + A^\top)) \).

\[
\mathbb{E}\|\theta_{n+1} - \theta^*\|^2 \leq K_P \left[ e^{-\lambda \sum_{k=0}^n \alpha_k} \right] \mathbb{E}\|\theta_0 - \theta^*\|^2 + 4K_m^2 K_p \sum_{i=0}^n \left[ e^{-\lambda \sum_{k=i+1}^n \alpha_k} \right] \alpha_i^2,
\]

where \( K_P, K_m \geq 0 \) are constants as defined in Lemmas 16 and 17 respectively.

The next result provides closed form estimates of the expectation bound given in Theorem 11 for the specific stepsize sequence \( \alpha_n = 1/(n + 1)^{\sigma} \), with \( \sigma \in (0, 1) \).

Theorem 12. Fix \( \sigma \in (0, 1) \) and let \( \alpha_n = 1/(n + 1)^{\sigma} \). Then

\[
\mathbb{E}\|\theta_{n+1} - \theta^*\|^2 \leq \left[ K_P e^{\lambda} \mathbb{E}\|\theta_0 - \theta^*\|^2 e^{-(\lambda/2)(n+2)^{1-\sigma}} + \frac{8K_m^2 K_p e^{\lambda}}{\lambda} \right] e^{-(\lambda/2)(n+2)^{1-\sigma}}
\]

\[
+ \frac{8K_m^2 K_p e^{\lambda/2}}{\lambda} \frac{1}{(n+1)^\sigma}.
\]

where \( K_b = e^{(\lambda/2) \sum_{k=0}^{i_0} \alpha_k} \) with \( i_0 \) denoting a number larger than \( (2\sigma/\lambda)^{1/(1-\sigma)} \).
6 Discussion

In this work we obtained the first concentration bound for an unaltered version of the celebrated TD(0); it is, in fact, the first to show the convergence rate of an unaltered online TD algorithm of any type. Our proof technique is general and can be used to provide convergence rates for additional TD methods. Specifically, using the non-linear analysis presented in [Thoppe and Borkar, 2015], we believe it can be extended to a broader family of function approximators, e.g., neural networks. Furthermore, future work can extend to a more general family learning rates, including the commonly used adaptive ones. Building upon Remark 5, we believe that a stronger expectation bound may hold for TD(0). This may enable obtaining tighter concentration bounds for TD(0) even with generic stepsizes.

References

D. P. Bertsekas. Dynamic Programming and Optimal Control. Vol II. Athena Scientific, fourth edition, 2012.

Vivek S Borkar. Stochastic approximation: a dynamical systems viewpoint. 2008.

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. SIAM Journal on Control and Optimization, 38(2):447–469, 2000.

Max Fathi and Noufel Frikha. Transport-entropy inequalities and deviation estimates for stochastic approximation schemes. Electron. J. Probab., 18:36 pp., 2013. doi: 10.1214/EJP.v18-2586. URL http://dx.doi.org/10.1214/EJP.v18-2586

Noufel Frikha and Stéphane Menozzi. Concentration bounds for stochastic approximations. Electron. Commun. Probab., 17:15 pp., 2012. doi: 10.1214/ECP.v17-1952. URL http://dx.doi.org/10.1214/ECP.v17-1952

Morris W Hirsch, Stephen Smale, and Robert L Devaney. Differential equations, dynamical systems, and an introduction to chaos. Academic press, 2012.

Sameer Kamal. On the convergence, lock-in probability, and sample complexity of stochastic approximation. SIAM Journal on Control and Optimization, 48(8):5178–5192, 2010.

Nathaniel Korda and LA Prashanth. On td (0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In ICML, pages 626–634, 2015.

Vangipuram Lakshmikantham and Sadashiv Deo. Method of variation of parameters for dynamic systems. CRC Press, 1998.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of lstd. In ICML-27th International Conference on Machine Learning, pages 615–622, 2010.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In UAI, pages 504–513. Citeseer, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.

Warren B Powell. Approximate Dynamic Programming: Solving the curses of dimensionality, volume 703. John Wiley & Sons, 2007.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.

Richard S Sutton. Learning to predict by the methods of temporal differences. Machine learning, 3(1):9–44, 1988.

Richard S. Sutton and Andrew G. Barto. Introduction to Reinforcement Learning. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.
Richard S Sutton, Hamid R Maei, and Csaba Szepesvári. A convergent o (n) temporal-difference algorithm for off-policy learning with linear function approximation. In Advances in neural information processing systems, pages 1609–1616, 2009a.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 993–1000. ACM, 2009b.

Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. The Journal of Machine Learning Research, 17:1–29, 2015.

Gerald Tesauro. Temporal difference learning and td-gammon. Communications of the ACM, 38(3): 58–68, 1995.

Gerald Teschl. Ordinary Differential Equations and Dynamical Systems. 2012.

Gugan Thoppe and Vivek S Borkar. A concentration bound for stochastic approximation via alekseev’s formula. arXiv:1506.08657, 2015.

John N Tsitsiklis, Benjamin Van Roy, et al. An analysis of temporal-difference learning with function approximation. IEEE transactions on automatic control, 42(5):674–690, 1997.

Huizhen 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.
A Variation of Parameters Formula

Let \( \theta(t, s, \tilde{\theta}(s)) \), \( t \geq s \), be the solution to (6) starting at \( \tilde{\theta}(s) \) at time \( t = s \). For \( k \geq 0 \), and \( \tau \in [t_k, t_{k+1}) \), let

\[
\zeta_1(\tau) := h(\theta_k) - h(\tilde{\theta}(\tau)) = A[\tilde{\theta}(\tau) - \theta_k]
\]

and

\[
\zeta_2(\tau) := M_{k+1}.
\]

Lemma 13. Let \( i \geq 0 \). For \( t \geq t_i \),

\[
\tilde{\theta}(t) = \theta(t, t_i, \tilde{\theta}(t_i)) + \int_{t_i}^{t} e^{-A(t-\tau)}[\zeta_1(\tau) + \zeta_2(\tau)]d\tau.
\]

Proof. For \( n \geq 0 \) and \( t \in [t_n, t_{n+1}) \), by simple algebra,

\[
\bar{\theta}(t) - \bar{\theta}(t_i) = \frac{t-t_n}{\alpha_n}[(\theta_{n+1} - \theta_n)] + \sum_{k=i}^{n-1}[(\theta_{k+1} - \theta_k)].
\]

Combining this with (2), (11), and (12), we have

\[
\bar{\theta}(t) = \bar{\theta}(t_i) + \int_{t_i}^{t} h(\tilde{\theta}(\tau))d\tau + \int_{t_i}^{t} [\zeta_1(\tau) + \zeta_2(\tau)]d\tau.
\]

Separately, writing (6) in integral form, we have

\[
\theta(t, t_i, \tilde{\theta}(t_i)) = \bar{\theta}(t_i) + \int_{t_i}^{t} h(\tilde{\theta}(\tau))d\tau.
\]

From the above two relations and the VoP formula [Lakshmikantham and Deo, 1998], the desired result follows.

B Supplementary Material for Proof of Theorem 1

Proof of Lemma 2. We have

\[
\|M_{n+1}\| = \|r_n \phi_n + (\gamma \phi'_n - \phi_n)\top \theta_n \phi_n - [b - A \theta_n]\|
\]

\[
= \|r_n \phi_n + (\gamma \phi'_n - \phi_n)\top (\theta_n - \theta^*)\phi_n
\]

\[
+ (\gamma \phi'_n - \phi_n)\top \theta^* \phi_n + A(\theta_n - \theta^*)\|
\]

\[
\leq \frac{1}{2} + \frac{1 + \gamma}{4} \|A^{-1}\| \|b\| + \frac{1 + \gamma + 4\|A\|}{4} \|\theta_n - \theta^*\|,
\]

where the first relation follows from (5), the second holds as \( b = A \theta^* \), while the third follows since \( A_i \) holds and \( \theta^* = A^{-1}b \). The desired result is now easy to see.

Proof of Corollary 3. The result follows by using Lemma 13 from Appendix A with \( i = \ell_1 \), \( t = t_{\ell_2} \), and subtracting \( \theta^* \) from both sides.

Proof of Lemma 6. The proof is by induction. The claim holds trivially for \( n = 0 \). Assume the claim for \( n \). Then from (11),

\[
\|\theta_{n+1} - \theta^*\| \leq \|\theta_n - \theta^*\| + \alpha_n \|\gamma \phi'_n - \phi_n\top \theta^* \phi_n\|
\]

\[
+ \alpha_n \|r_n \phi_n\| + \alpha_n \|\gamma \phi'_n - \phi_n\top [\theta_n - \theta^*] \phi_n\|
\]

Applying the Cauchy-Schwarz inequality, and using \( A_i \) and the fact that \( \gamma \leq 1 \), we have

\[
\|\theta_{n+1} - \theta^*\| \leq \|\theta_n - \theta^*\| + \frac{\alpha_n}{2} C_* + \frac{\alpha_n}{2} \|\theta_n - \theta^*\|.
\]

Now as \( 1 \leq R_0 \), we have

\[
\|\theta_{n+1} - \theta^*\| \leq \left[ 1 + \frac{\alpha_n}{2} \right] \|\theta_n - \theta^*\| + \frac{\alpha_n}{2} C_* R_0.
\]

Using the induction hypothesis and the stepsize choice, the claim for \( n + 1 \) is now easy to see. The desired result thus follows.
Proof of Lemma 8. For any two events $\mathcal{E}_1$ and $\mathcal{E}_2$, note that
\[
\mathcal{E}_1 = [\mathcal{E}_2^c \cap \mathcal{E}_1] \cup [\mathcal{E}_2 \cap \mathcal{E}_1] \subseteq \mathcal{E}_2^c \cup [\mathcal{E}_2 \cap \mathcal{E}_1].
\] (13)

Separately, for any sequence of events $\{\mathcal{E}_k\}$, observe that
\[
\bigcup_{k=1}^{m} \mathcal{E}_k = \left[ \bigcup_{k=1}^{m} \left( \bigcup_{i=1}^{k-1} \mathcal{E}_i \right) \right] \cap \mathcal{E}_k.
\] (14)

where $\bigcup_{i=1}^{j_2} \mathcal{E}_i = \emptyset$ whenever $i_1 > i_2$. Using (13), we have
\[
\mathcal{E}^c(n_0, n_1) \subseteq G^c_{n_0, n_0 + n_1} \cup [G_{n_0, n_0 + n_1} \cap \mathcal{E}^c(n_0, n_1)].
\] (15)

From Lemma 6, $\{\|\theta_{n_0} - \theta^*\| \leq R_{wc}(n_0)\}$ is a certain event. Hence it follows from (14) that
\[
G^c_{n_0, n_0 + n_1} = \mathcal{E}^\text{mid}_{n_0, n_1}.
\] (16)

Similarly, from (14) and the fact that $\epsilon \leq R_0$,
\[
G_{n_0, n_0 + n_1} \cap \mathcal{E}^c(n_0, n_1) \subseteq \mathcal{E}^\text{after}_{n_0, n_1}.
\] (17)

Substituting (16) and (17) in (15) gives
\[
\mathcal{E}^c(n_0, n_1) \subseteq \mathcal{E}^\text{mid}_{n_0, n_1} \cup \mathcal{E}^\text{after}_{n_0, n_1}.
\]

The claimed result follows.

Proof of Lemma 9. For $n \geq n' \geq n_0 \geq 0$, by its definition and the triangle inequality,
\[
\|E_{[n', n+1]}^d\| \leq \sum_{k=n'}^{n} \int_{t_k}^{t_{k+1}} e^{-\lambda(t_{k+1} - \tau)} \|A\| \|\bar{\theta}(\tau) - \theta_k\| d\tau.
\]

Fix a $k \in \{n', \ldots, n\}$ and $\tau \in (t_k, t_{k+1})$. Then using (5), (2), (4), and the fact that $(\tau - t_k) \leq \alpha_k$, we have
\[
\|\bar{\theta}(\tau) - \theta_k\| \leq \alpha_k [\|A\| \|\theta_k - \theta^*\| + \|M_{k+1}\|].
\]

Combining this with Lemma 4, we get
\[
\|\bar{\theta}(\tau) - \theta_k\| \leq \alpha_k [K_m + (\|A\| + K_m) \|\theta_k - \theta^*\|].
\]

As the event $G_{n_0, n}$ holds, and since $\alpha_k \leq \alpha_{n'}$ and $R_{wc}(n_0) \geq 1$, we have
\[
\|\bar{\theta}(\tau) - \theta_k\| \leq 2\|A\| + 2K_m \alpha_{n'}[n_0 + 1]C, R_0.
\]

From the above discussion, (9), the step-size choice, and the facts that
\[
\sum_{k=n'}^{n} \int_{t_k}^{t_{k+1}} e^{-\lambda(t_{k+1} - \tau)} d\tau = \int_{t_{n'}}^{t_n} e^{-\lambda(t_n - \tau)} d\tau \leq \frac{1}{\lambda},
\]
and $\alpha_k \leq \alpha_{n'} \leq \alpha_{n_0}$, we get
\[
\|E_{[n', n+1]}^d\| \leq \frac{K_m\|A\| (\|A\| + 2K_m) [n_0 + 1]C, R_0}{\lambda(n'+1)}.
\]

The desired results now follow by substituting $n'$ first with $n_0$ and then with $n_c$.

Proof of Lemma 10. Let $Q_{k,n} = \int_{t_k}^{t_{k+1}} e^{-A(t_{k+1} - \tau)} d\tau$. Then, for any $n_0 \leq n' \leq n$,
\[
E_{[n', n+1]}^m = \sum_{k=n'}^{n} Q_{k,n} M_{k+1},
\]
a sum of martingale differences. When the event $G_{n_0, n}$ holds, it follows that the indicator $1_{G_{n_0, k}} = 1$ for all $k \in \{n_0, \ldots, n', \ldots, n\}$. Hence, for any $R \geq 0$,
\[
\Pr\{G_{n_0, n}, \|E_{[n', n+1]}^m\| \geq R\} = \Pr\{G_{n_0, n}, \|\sum_{k=n'}^{n} Q_{k,n} M_{k+1} 1_{G_{n_0, k}}\| \geq R\}
\]
As from (9), let
\[ Q_{k,n}^i, \quad i,j = \text{entry of matrix } Q_{k,n} \quad \text{and let } M_{k+1}^j = \text{coordinate of } M_{k+1}. \]

Then using the union bound twice on the above relation, we have
\[
\Pr\{G_{n_0,n}^1, \| E_{n',n+1}^m \| \geq R \} \leq \sum_{i=1}^d \sum_{j=1}^d \Pr\left\{ \left\| \sum_{k=n'}^n Q_{k,n}^i M_{k+1}^j 1_{G_{n_0,k}} \right\| \geq R \right\}.
\]

As \( |Q_{k,n}^i M_{k+1}^j 1_{G_{n_0,k}}| \leq \| Q_{k,n}^i \| \| M_{k+1}^j 1_{G_{n_0,k}} =: \beta_{k,n} \), Azuma-Hoeffding inequality now gives
\[
\Pr\{G_{n_0,n}^1, \| E_{n',n+1}^m \| \geq R \} \leq 2d^2 \exp\left[-\frac{R^2}{2d^3 \sum_{k=n'}^n \beta_{k,n}^2}\right]. \tag{18}
\]

On the event \( G_{n_0,k}; \| \theta_k - \theta^* \| \leq 2R_{\text{wc}}(n_0) \) by definition. Hence from Lemma 4 we have
\[
\| M_{k+1}^j 1_{G_{n_0,k}} \| \leq 3K_m R_{\text{wc}}(n_0). \tag{19}
\]

Also from (9), \( \| Q_{k,n}^i \| \leq K_\lambda e^{-\lambda(t_{n+1}-t_{k+1})} \alpha_k \). Combining the two inequalities, and using (10) along with the fact that \( 1/(k+1) \leq 2/(k+2) \), we get
\[
\beta_{k,n} \leq 3K_m K_\lambda R_{\text{wc}}(n_0) e^{-\lambda(t_{n+1}-t_{k+1})} \alpha_k \leq 6K_m K_\lambda R_{\text{wc}}(n_0) \frac{(k+2)^{\lambda-1}}{(n+2)^\lambda}.
\]

Consider the case \( \lambda > 1/2 \). By treating the sum as a right Riemann sum, we have
\[
\sum_{k=n'}^n (k+2)^{2\lambda-2} \leq (n+3)^{2\lambda-1}/(2\lambda - 1).
\]

As \( (n+3) \leq 2(n+2) \) and \( (n+2) \geq (n+1) \), we have
\[
\sum_{k=n'}^n \beta_{k,n}^2 \leq C_{m2}^2 \frac{R_{\text{wc}}^2(n_0)}{n+1}.
\]

Now consider the case \( \lambda < 1/2 \). Again treating the sum as a right Riemann sum, we have
\[
\sum_{k=n'}^n (k+2)^{2\lambda-2} \leq \frac{1}{(1-2\lambda)[n'+1]^{1-2\lambda}}.
\]

As \( (n+2) \geq (n+1) \), it follows that
\[
\sum_{k=n'}^n \beta_{k,n}^2 \leq C_{m2}^2 \frac{R_{\text{wc}}^2(n_0)}{[n'+1]^{1-2\lambda}(n+1)^{2\lambda}}.
\]

Substituting \( \sum_{k=n'}^n \beta_{k,n}^2 \) bounds in (18), the desired result is easy to see. \( \square \)

### B.1 Conditional Results on the Bad Events

On the first “bad” event \( \varepsilon_{n_0,n_1}^\text{mid} \), the TD(0) iterate \( \theta_n \) for at least one \( n \) between \( n_0 + 1 \) and \( n_0 + n_1 \) leaves the \( 2R_{\text{wc}}(n_0) \) ball around \( \theta^* \). The next lemma shows that this event has low probability.

**Lemma 14 (Bound on Probability of \( \varepsilon_{n_0,n_1}^\text{mid} \)).** Let \( n_0 \geq \max \left\{ \frac{K_\lambda \| A \| (\| A \| + 2K_m \lambda)}{\lambda}, 2 \right\} \) and \( n_1 \geq 1 \).

- If \( \lambda > 1/2 \), then
  \[ \Pr\{\varepsilon_{n_0,n_1}^\text{mid}\} \leq 16d^5 C_{m2}^2 \exp\left[-\frac{n_0}{8d^6 C_{m2}^2}\right]. \]
• If \( \lambda < 1/2 \), then

\[
\Pr\{E_{n_0,n_1}^{\text{mid}}\} \leq 2d^2 \left[ \frac{8d^3C_{m_2}}{\lambda} \right]^{1/2} \exp\left[ -\frac{n_0}{6d^3C_{m_2}} \right] \frac{1}{(n_0 + 1)^{1-2\lambda}}.
\]

**Proof.** From Corollary 5 we have

\[
\|\theta_{n+1} - \theta^*\| \leq \|\theta(t_{n+1}, t_{n_0}, \theta_{n_0}) - \theta^*\| + \|E_{[n_0,n_1]}^d\| + \|E_{[n_0,n_1]}^m\|.
\]

Suppose the event \( G_{n_0,n} \) holds. Then from 8,

\[
\|\theta(t_{n+1}, t_{n_0}, \theta_{n_0}) - \theta^*\| \leq \|\theta_{n_0} - \theta^*\| \leq R_{wc}(n_0).
\]

Also, as \( n_0 \geq \frac{K_6\|A\|\|\lambda\| + 2K_m}{\lambda} \), by Lemma 9

\[
\|E_{[n_0,n_1]}^d\| \leq R_{wc}(n_0)/3.\]

From all of the above, we have

\[
\{G_{n_0,n}, \|\theta_{n+1} - \theta^*\| > 2R_{wc}(n_0)\} \subseteq \{G_{n_0,n}, \|E_{[n_0,n_1]}^m\| > R_{wc}(n_0)/2\}.
\]

From this, we get

\[
E_{n_0,n_1}^{\text{mid}} \subseteq \bigcup_{n=n_0}^{n_0+n_1-1} \left\{G_{n_0,n}, \|E_{[n_0,n_1]}^m\| > \frac{R_{wc}(n_0)}{2}\right\}
\]

\[
\subseteq \bigcup_{n=n_0}^{\infty} \left\{G_{n_0,n}, \|E_{[n_0,n_1]}^m\| > \frac{R_{wc}(n_0)}{2}\right\}.
\]

Consequently,

\[
\Pr\{E_{n_0,n_1}^{\text{mid}}\} \leq \sum_{n=n_0}^{\infty} \Pr\left\{G_{n_0,n}, \|E_{[n_0,n_1]}^m\| > \frac{R_{wc}(n_0)}{2}\right\}.
\]

(20)

Consider the case \( \lambda > 1/2 \). Lemma 10 shows that

\[
\Pr\left\{G_{n_0,n}, \|E_{[n_0,n_1]}^m\| > \frac{R_{wc}(n_0)}{2}\right\} \leq 2d^2 \exp \left[ -\frac{n + 1}{8d^3C_{m_2}^2}\right].
\]

Substituting this in (20) and treating the resulting expression as a right Riemann sum, the desired result is easy to see.

Now consider the case \( \lambda < 1/2 \). From Lemma 10 we get

\[
\Pr\left\{G_{n_0,n}, \|E_{[n_0,n_1]}^m\| > \frac{R_{wc}(n_0)}{2}\right\} \leq 2d^2 \exp \left[ -\frac{(n_0 + 1)^{1-2\lambda}}{8d^3C_{m_2}^2}\right].
\]

Let \( \ell_{n_0} := (n_0 + 1)^{1-2\lambda} / 8d^3C_{m_2}^2 \). Observe that

\[
\sum_{n=n_0}^{\infty} \exp\left[ -\ell_{n_0}(n + 1)^{2\lambda}\right]
\]

\[
\leq \sum_{i=[(n_0+1)^{2\lambda}]}^{\infty} e^{-i\ell_{n_0}} |\{n : [(n + 1)^{2\lambda}] = i\}|
\]

\[
\leq \frac{1}{2\lambda} \sum_{i=[(n_0+1)^{2\lambda}]}^{\infty} e^{-i\ell_{n_0}/2} e^{-i\ell_{n_0}/2} \left( i + 1 \right)^{1-2\lambda} (1 - 2\lambda)
\]

\[
\leq \frac{1}{2\lambda} \left[ \frac{(1 - 2\lambda)}{\ell_{n_0}\lambda} \right]^{1-2\lambda} e^{\frac{1}{2} \ell_{n_0} - \frac{1-2\lambda}{\lambda}} \sum_{i=[(n_0+1)^{2\lambda}]}^{\infty} e^{-i\ell_{n_0}/2}.
\]

(21)

(22)
Also, as follows by treating the sum as a right Riemann sum, (24) follows by substituting the value of (22) holds since, again by calculus,

\[
\max_{i \geq 0} e^{-\frac{1}{\ell_n 0}/2 (i + 1)^{2 - \lambda}} \leq \left( \frac{1 - 2 \lambda}{\ell_n 0} \right)^{1 - \lambda} e^{\frac{1}{2} (\ell_n 0 - \frac{1 - 2 \lambda}{\ell_n 0})}.
\]

(23) follows by treating the sum as a right Riemann sum, (24) follows by substituting the value of \( \ell_0 \) and using the fact that \( n_0^2 \lambda \geq 4 \) and (25) holds since \( 1 - 2 \lambda < 1 \). Substituting (25) in (20), the desired result follows.

On the second “bad” event \( e_{n_0, n_1} \), the TD(0) iterate \( \theta_n \) for at least one \( n > n_0 + n_1 \) lies outside the \( \min\{\epsilon, 2R_{\text{wec}}(n_0)\} \) radius ball around \( \theta^* \). The next result shows that this event also has low probability.

**Lemma 15** (Bound on Probability of \( E_{n_0, n_1} \)). Let \( n_0 \geq \max\{K_{\lambda}6\|A\|\|\|A\|+2K_m\|\chi\}, 2^{1/\lambda} \) and

\[
n_c \geq \left( 1 + \frac{K_{\lambda}6\|A\|\|\|A\|+2K_m\|\chi\}}{\lambda \min(\epsilon, R_{\text{wec}}(n_0))} \right) R_{\text{wec}}(n_0),
\]

Let \( n_1 = n_1(\epsilon, n_c, n_0) \geq (n_c + 1) \left[ \frac{6K_{\lambda}R_{\text{wec}}(n_0)}{\epsilon} \right]^{1/\lambda} - n_0.

If \( \lambda > 1/2 \), then

\[
\Pr\{E_{n_0, n_1}\} \leq 36d^5 C_{m2} \left[ \frac{R_{\text{wec}}(n_0)^2}{\epsilon} \right] \exp\left[ \frac{-(6K_{\lambda})^{1/\lambda}(n_c + 1) \left[ \frac{\epsilon}{R_{\text{wec}}(n_0)} \right]^{2 - 1/\lambda}}{18d^5 C_{m2}^2} \right].
\]

If \( \lambda < 1/2 \), then

\[
\Pr\{E_{n_0, n_1}\} \leq 2d^5 \left[ \frac{18d^5 C_{m2}^2 R_{\text{wec}}(n_0)^2}{\epsilon^2 \lambda} \right] \exp\left[ - \frac{K_{\lambda}^2}{4d^5 C_{m2}^2}(n_c + 1) \right].
\]

**Proof.** Assume the event \( G_{n_0, n} \) holds for some \( n \geq n_c \). Then \n
\[
\|\theta_n - \theta^*\| \leq 2R_{\text{wec}}(n_0).
\]

Hence from (27) and (29), for \( t \geq t_{n_c} \), we have

\[
\|\theta(t, t_{n_c}, \theta_{n_c}) - \theta^*\| \leq K_{\lambda} e^{-\lambda (t - t_{n_c})} 2R_{\text{wec}}(n_0) \leq \frac{\epsilon}{3}.
\]

Now as \( n_1 = (n_c + 1) \left[ \frac{6K_{\lambda}R_{\text{wec}}(n_0)}{\epsilon} \right]^{1/\lambda} - n_0 \), it follows that \( \forall n \geq n_0 + n_1, \)

\[
\|\theta(t_{n+1}, t_{n}, \theta_{n}) - \theta^*\| \leq \frac{\epsilon}{3}.
\]

Also, as \( n_c \geq \left( 1 + \frac{K_{\lambda}6\|A\|\|\|A\|+2K_m\|\chi\}}{\lambda \min(\epsilon, R_{\text{wec}}(n_0))} \right) (n_0 + 1), \) from Lemma 15 we have \( \|E_{\text{wec}}(n_{n-1})\| \leq \epsilon/3 \) for all \( n \geq n_c \). Combining these with Corollary 5 it follows that \( \forall n \geq n_0 + n_1 \),

\[
\{ G_{n_0, n}, \|\theta_{n+1} - \theta^*\| > \min\{\epsilon, 2R_{\text{wec}}(n_0)\} \} \subseteq \{ G_{n_0, n}, \|\theta_{n+1} - \theta^*\| > \epsilon \}
\]
Hence from the definition of $\mathcal{E}^\text{after}_{n_0, n_1}$,
\[
\Pr\{\mathcal{E}^\text{after}_{n_0, n_1}\} \leq \sum_{n = n_0 + 1}^{\infty} \Pr\{G_{n_0, n}, \|E_m^{n, n_1}\| \geq \frac{\epsilon}{3}\}.
\]  

(27)

Consider the case $\lambda > 1/2$. Lemma 10 and the definition of $R_{\text{wc}}(n_0)$ in Theorem 6 shows that
\[
\Pr\{G_{n_0, n}, \|E_m^{n, n_1}\| \geq \frac{\epsilon}{3}\} \leq 2d^2 \exp\left[-\frac{(n_0 + 1)^{1-2\lambda}(n + 1)^2\epsilon^2}{18d^3C_m^2C^2R_0^2}\right].
\]

Using this in (27) and treating the resulting expression as a right Riemann sum, we get
\[
\Pr\{\mathcal{E}^\text{after}_{n_0, n_1}\} \leq 36d^2C_m^2 \left[\frac{R_{\text{wc}}(n_0)}{\epsilon}\right]^2 \exp\left[-\frac{(n_0 + 1)^{1-2\lambda}(n + 1)^2\epsilon^2}{18d^3C_m^2[R_{\text{wc}}(n_0)]^2}\right].
\]
Substituting the given relation between $n_1$ and $n_0$, the desired result is easy to see.

Consider the case $\lambda < 1/2$. From Lemma 10 and the definition of $R_{\text{wc}}(n_0)$ in Theorem 6, we have
\[
\Pr\{G_{n_0, n}, \|E_m^{n, n_1}\| \geq \frac{\epsilon}{3}\} \leq 2d^2 \exp\left[-\frac{(n_0 + 1)^{1-2\lambda}(n + 1)^2\epsilon^2}{18d^3C_m^2[R_{\text{wc}}(n_0)]^2}\right].
\]

Let $k_{n_0} := c^2(n_0 + 1)^{1-2\lambda}/(18d^3C_m^2[R_{\text{wc}}(n_0)]^2)$. Then by the same technique that we use to obtain (23) in the proof for Lemma 14, we have
\[
\sum_{n = n_0 + 1}^{\infty} \exp[-k_{n_0}(n + 1)^{2\lambda}] \
\leq \frac{1}{k_{n_0} \lambda} \left[\frac{1}{2\lambda}\right]^{1-2\lambda} e^{\frac{1}{2}[k_{n_0} - (1-2\lambda)]} e^{-k_{n_0}(n_0 + n_1)^{2\lambda}} \
\leq \frac{1}{k_{n_0} \lambda} \left[\frac{1}{2\lambda}\right]^{1-2\lambda} e^{-k_{n_0}(n_0 + n_1)^{2\lambda}} \
= \left[\frac{18d^3C_m^2[R_{\text{wc}}(n_0)]^2}{c^2\lambda(n_0 + 1)^{1-2\lambda}}\right]^{\frac{1}{2\lambda}} \exp\left[-\frac{c^2(n_0 + 1)^{1-2\lambda}(n + 1)^2\epsilon^2}{144d^6C_m^2[R_{\text{wc}}(n_0)]^2}\right]
\]
where the second inequality is obtained using the facts that $(n_0 + 1)^{2\lambda} \geq n_0^{2\lambda} \geq 4$ and $1 - 2\lambda \leq 1$ and the last equality is obtained by substituting the value of $k_{n_0}$. From this, after substituting the given relation between $n_0$ and $n_1$, the desired result is easy to see.

\textit{Proof of Theorem 7} From Lemma 8 by a union bound,
\[
\Pr\{\mathcal{E}'(n_0, n_1)\} \leq \Pr\{\mathcal{E}_{\text{mid}}^\text{after}\} + \Pr\{\mathcal{E}_{n_0, n_1}\}.
\]

We now show how to set $n_0$ and $n_1$ so that each of the two terms above is less than $\delta/2$.

Consider the case $\lambda > 1/2$. Let
\[
N_0(\delta) = \max\left\{\frac{K_6\|A\|\|A\| + 2K_m}{\lambda}, 2^{\frac{1}{\lambda}}, 8d^3C_m^2 \ln\left[\frac{32d^6C_0^2}{\delta}\right]\right\} = O\left(\ln\frac{1}{\delta}\right),
\]  

(28)

\[
N_c(\epsilon, \delta, n_0) = \max\left\{\left[\left(1 + \frac{K_6\|A\|\|A\| + 2K_m}{\lambda \min \{\epsilon, R_{\text{wc}}(n_0)\}}\right) R_{\text{wc}}(n_0)\right], \right\}
\]
\[
\frac{18d^3C_m^2}{(6K_\lambda)^{1/\lambda}} \left[\frac{R_{\text{wc}}(n_0)}{\epsilon}\right]^{-2 - \frac{1}{\lambda}} \ln\left[72d^6C_m^2 \left[\frac{1}{\delta}\right] \left[\frac{R_{\text{wc}}(n_0)}{\epsilon}\right]^2\right],
\]
so that \( N_c(\epsilon, \delta, N_0(\delta)) = \hat{O}\left( \max\left\{ \frac{1}{\epsilon} \ln \left[ \frac{1}{\delta} \right], \left[ \frac{1}{\epsilon} \right]^{2 - \frac{1}{\lambda}} \left[ \ln \frac{1}{\delta} \right]^{3 - \frac{1}{\lambda}} \right\} \right) \), and let

\[
N_1(\epsilon, n_c, n_0) = (n_c + 1) \left[ \frac{6K\lambda R_{we}(n_0)}{\epsilon} \right]^{1/\lambda} - n_0,
\]

so that

\[
N_1(\epsilon, N_c(\epsilon, \delta, N_0(\delta)), N_0(\delta)) = \hat{O}\left( \max\left\{ \left[ \frac{1}{\epsilon} \right]^{1 + \frac{1}{\lambda}} \left[ \ln \frac{1}{\delta} \right]^{1 + \frac{1}{\lambda}}, \left[ \frac{1}{\epsilon} \right]^2 \left[ \ln \frac{1}{\delta} \right]^3 \right\} \right). \tag{29}
\]

Let \( n_0 \geq N_0(\delta), n_c \geq N_c(\epsilon, \delta, n_0) \) and \( n_1 \geq N_1(\epsilon, n_c, n_0) \). Then from Lemma 14, \( \Pr\{E_{n_0, n_1}^{mid} \} \leq \delta/2 \) and from Lemma 15, \( \Pr\{E_{n_0, n_1}^{after} \} \leq \delta/2 \). Hence \( \Pr\{E(\epsilon, n_0, n_1) \} \leq \delta \). Consequently, \( N(\epsilon, \delta) = N_1(\epsilon, N_c(\epsilon, \delta, N_0(\delta)), N_0(\delta)) \) satisfies the desired properties, which completes the proof for \( \lambda > 1/2 \).

Now consider the case \( \lambda < 1/2 \). The same exact proof can be repeated, with the following \( N_0, N_c \) and \( N_1 \).

\[
N_0(\delta) = \max\left\{ \frac{K_{\alpha}6\||A||\|A\|+2K_m}{\lambda}, 2^{\frac{1}{\lambda}}, \frac{6d^2C_m^2}{2\lambda} \ln\left( \frac{32d^2C_m^2}{\lambda} \right) \right\} = O\left( \ln \frac{1}{\delta} \right), \tag{30}
\]

\[
N_c(\epsilon, \delta, n_0) = \max\left\{ \left[ 1 + \frac{K_{\alpha}6\||A||\|A\|+2K_m}{\Lambda_{\min}(\epsilon, R_{we}(n_0))} \right] R_{we}(n_0) \right\},
\]

so that \( N_c(\epsilon, \delta, N_0(\delta)) = \hat{O}\left( \frac{1}{\epsilon} \ln \frac{1}{\delta} \right) \) and let

\[
N_1(\epsilon, n_c, n_0) = (n_c + 1) \left[ \frac{6K\lambda R_{we}(n_0)}{\epsilon} \right]^{1/\lambda} - n_0, \tag{31}
\]

so that \( N_1(\epsilon, N_c(\epsilon, \delta, N_0(\delta)), N_0(\delta)) = \hat{O}\left( \left[ (1/\epsilon) \ln (1/\delta) \right]^{1+1/\lambda} \right) \). Thus \( N(\epsilon, \delta) = N_1(\epsilon, N_c(\epsilon, \delta, N_0(\delta)), N_0(\delta)) \) satisfies the desired properties for the case \( \lambda < 1/2 \).

For \( \lambda = 1/2 \), the same process can be repeated, resulting in the same \( O \) and \( \hat{O} \) results as in (30) and (31).

\[\square\]

C Supplementary Material for Proof of Theorem 3

Notice that the matrices \( (A^T + A) \) and \( (A^T A + K_m I) \) are symmetric. Further, as \( A \) is positive definite, the above matrices are also positive definite. Hence their minimum and maximum eigenvalues are strictly positive.

**Lemma 16.** For \( n \geq 0 \), let \( \lambda_n := \lambda_{\max}(\Lambda_n) \), where

\[
\Lambda_n := I - \alpha_n(A + A^T) + \alpha_n^2(A^T A + 4K_m^2 I).
\]

Fix \( \lambda \in (0, \lambda_{\min}(A + A^T)) \). Let \( m \) be so that \( \forall k \geq m, \alpha_k \leq \frac{\lambda_{\min}(A + A^T) - \lambda}{\lambda_{\max}(A^T A + 4K_m^2)} \). Then for any \( k, n \) such that \( n \geq k \geq 0 \),

\[
\prod_{i=k}^{n} \lambda_k \leq K_pe^{-\lambda[\sum_{i=k}^{n} \alpha_k]},
\]

where

\[
K_p := \max_{\ell_1 \leq \ell_2 \leq m} \prod_{\ell \in [\ell_1, \ell_2]} e^{\alpha_{\ell} (\mu + \lambda)},
\]

with \( \mu = -\lambda_{\min}(A + A^T) + \lambda_{\max}(A^T A + 4K_m^2 I) \).
Remark 7. Such $m$ exists since $\alpha_k \to 0$ as $k \to \infty$.

Proof. Using Weyl’s inequality, we have

$$\lambda_n \leq \lambda_{\text{max}}(I - \alpha_n (A + A^\top)) + \alpha_n^2 \lambda_{\text{max}}(A^\top A + 4K_m^2 I).$$  \hfill (32)

Since $\lambda_{\text{max}}(I - \alpha_n (A + A^\top)) \leq (1 - \alpha_n \lambda_{\text{min}}(A + A^\top))$, we have

$$\lambda_n \leq e^{[-\alpha_n \lambda_{\text{min}}(A^\top A) + \alpha_n^2 \lambda_{\text{max}}(A^\top A + 4K_m^2 I)]}.$$

For $n < m$, using $\alpha_n \leq 1$ and hence $\alpha_n^2 \leq \alpha_n$, we have the following weak bound:

$$\lambda_n \leq e^{\alpha_n \mu}.$$  \hfill (33)

On the other hand, for $n \geq m$, we have

$$\lambda_n \leq e^{-\alpha_n} e^{-\alpha_n (\lambda_{\text{min}}(A^\top A) - \lambda) - \alpha_n \lambda_{\text{max}}(A^\top A + 4K_m^2 I)} \leq e^{-\alpha_n \lambda_n}.$$  \hfill (34)

To prove the desired result, we consider three cases: $k \leq n \leq m$, $m \leq k \leq n$ and $k \leq m \leq n$. For the last case, using (33) and (34), we have

$$\prod_{k=0}^{n} \lambda_k \leq \prod_{k=0}^{m} \lambda_k e^{-\lambda \sum_{k=0}^{m} \alpha_k} \leq \prod_{k=0}^{m} \lambda_k e^{-\lambda \sum_{k=0}^{n} \alpha_k} \leq K_p e^{-\lambda \sum_{k=0}^{n} \alpha_k},$$

as desired. Similarly, it can be shown that bound holds in other cases as well. The desired result thus follows.

Proof of Theorem 17. Let $V(\theta) = \|\theta - \theta^*\|^2$. Using (2) and (4), we have

$$\theta_{n+1} - \theta^* = (I - \alpha_n A) (\theta_n - \theta^*) + \alpha_n M_{n+1}.$$  

Hence

$$V(\theta_{n+1}) = (\theta_{n+1} - \theta^*)^\top (\theta_{n+1} - \theta^*)$$

$$= [(I - \alpha_n A) (\theta_n - \theta^*) + \alpha_n M_{n+1}]^\top [(I - \alpha_n A) (\theta_n - \theta^*) + \alpha_n M_{n+1}]$$

$$= (\theta_n - \theta^*)^\top (I - \alpha_n (A^\top A) + \alpha_n^2 A^\top A) (\theta_n - \theta^*)$$

$$+ \alpha_n (\theta_n - \theta^*)^\top (I - \alpha_n A)^\top M_{n+1} + \alpha_n M_{n+1}^\top (I - \alpha_n A) (\theta_n - \theta^*) + \alpha_n^2 \| M_{n+1} \|^2.$$  

Taking conditional expectation and using $E[\| M_{n+1} \|^2 | F_n] = 0$, we get

$$E[V(\theta_{n+1}) | F_n] = \left( \theta_n - \theta^* \right)^\top \left[ I - \alpha_n (A^\top A) + \alpha_n^2 A^\top A \right] \left( \theta_n - \theta^* \right) + \alpha_n^2 E[\| M_{n+1} \|^2 | F_n].$$

From Lemma 4 and as $|x| \leq 1 + x^2$, we have $\| M_{n+1} \|^2 \leq 4K_m^2 [1 + \|\theta_n - \theta^*\|^2]$. This immediately shows that $E[\| M_{n+1} \|^2 | F_n] \leq 4K_m^2 [1 + \|\theta_n - \theta^*\|^2]$. Hence

$$E[V(\theta_{n+1}) | F_n] \leq \left( \theta_n - \theta^* \right)^\top \Lambda_n \left( \theta_n - \theta^* \right) + 4K_m^2 \alpha_n^2,$$

where $\Lambda_n = [I - \alpha_n (A^\top A) + \alpha_n^2 (A^\top A + 4K_m^2 I)]$. Since $\Lambda_n$ is a symmetric matrix, all its eigenvalues are real. With $\lambda_n := \lambda_{\text{max}}(\Lambda_n)$, we have

$$E[V(\theta_{n+1}) | F_n] \leq \lambda_n V(\theta_n) + 4K_m^2 \alpha_n^2.$$  

Taking expectation on both sides and letting $w_n = E[V(\theta_n)]$, we have

$$w_{n+1} \leq \lambda_n w_n + 4K_m^2 \alpha_n^2.$$  

Sequentially using the above inequality, we have

$$w_{n+1} \leq \sum_{k=1}^{n} \lambda_k w_0 + 4K_m^2 \sum_{k=1}^{n} \lambda_k \alpha_k^2.$$  

Using Lemma 16 and using the constant $K_p$ defined there, the desired result follows.

\hfill \qed
Proof of Theorem 12. Let \( \{t_n\} \) be as defined in Subsection 4.1. Observe that

\[
\sum_{i=0}^{n} \left[ e^{-((\lambda)/2) \sum_{k=i+1}^{n} \alpha_k} \right] \alpha_i \leq \left( \sup_{i \geq 0} e^{((\lambda)/2) \alpha_i} \right) \sum_{i=0}^{n} \left[ e^{-((\lambda)/2) \sum_{k=i+1}^{n} \alpha_k} \right] \alpha_i = \left( \sup_{i \geq 0} e^{((\lambda)/2) \alpha_i} \right) \sum_{i=0}^{n} \left[ e^{-((\lambda)/2)(t_{n+1} - t_k)} \right] \alpha_i \leq \left( \sup_{i \geq 0} e^{((\lambda)/2) \alpha_i} \right) \int_{0}^{t_{n+1}} e^{-((\lambda)(t_{n+1} - s))} ds \leq \left( \sup_{i \geq 0} e^{((\lambda)/2) \alpha_i} \right) \frac{2}{\lambda} \leq \frac{2e^{(\lambda)/2}}{\lambda}.
\]

where the third relation follows by treating the sum as right Riemann sum, and the last inequality follows since \( \sup_{i \geq 0} \alpha_i \leq 1 \). Hence it follows that

\[
\sum_{i=0}^{n} \left[ e^{-((\lambda)/2) \sum_{k=i+1}^{n} \alpha_k} \right] \alpha_i^2 \leq \left( \sup_{0 \leq i \leq n} \left[ \alpha_i e^{-(\lambda/2) \sum_{k=i+1}^{n} \alpha_k} \right] \right) \sum_{i=0}^{n} \left[ e^{-((\lambda)/2) \sum_{k=i+1}^{n} \alpha_k} \right] \alpha_i \leq \frac{2e^{(\lambda)/2}}{\lambda} \left( \sup_{0 \leq i \leq n} \left[ \alpha_i e^{-(\lambda/2) \sum_{k=i+1}^{n} \alpha_k} \right] \right). \tag{35}
\]

We claim that for all \( n \geq i_0 \),

\[
\sup_{i_0 \leq i \leq n} \left[ \alpha_i e^{-(\lambda/2) \sum_{k=i+1}^{n} \alpha_k} \right] \leq \frac{1}{(n+1)^{\sigma}}. \tag{36}
\]

To establish this, we show that for any \( n \geq i_0 \), \( \alpha_t e^{-(\lambda/2) \sum_{k=i+1}^{n} \alpha_k} \) monotonically increases as \( i \) is varied from \( i_0 \) to \( n \). To prove the latter, it suffices to show that \( \alpha_t e^{-(\lambda/2) \alpha_{i+1}} \leq \alpha_{i+1} \), or equivalently \((i+2)^{\sigma}/(i+1)^{\sigma} \leq e^{\lambda/2(i+2)^{\sigma}} \) for all \( i \geq i_0 \). But the latter is indeed true. Thus \( 35 \) holds. From \( 35 \) and \( 36 \), we then have

\[
\sum_{i=0}^{n} \left[ e^{-\lambda \sum_{k=i+1}^{n} \alpha_k} \right] \alpha_i^2 \leq \frac{2e^{(\lambda)/2}}{\lambda} \left( \sup_{0 \leq i \leq i_0} \left[ \alpha_i e^{-(\lambda/2) \sum_{k=i+1}^{n} \alpha_k} \right] \right) + \left( \sup_{0 \leq i \leq n} \left[ \alpha_i e^{-(\lambda/2) \sum_{k=i+1}^{n} \alpha_k} \right] \right) + \frac{1}{(n+1)^{\sigma}}.
\]

where the first relation holds as \( \sup_{a_0, \ldots, a_n} \leq \sup_{a_0, \ldots, a_i} + \sup_{a_{i+1}, \ldots, a_n} \) for any positive sequence \( \{a_0, \ldots, a_n\} \) with \( 0 \leq i_0 \leq n \), and the last relation follows as \( \alpha_i \leq 1 \) and \( \sup_{0 \leq i \leq i_0} e^{(\lambda/2) \sum_{k=0}^{i_0} \alpha_k} \leq K \). Combining the above inequality with the relation from Theorem 11 we have

\[
\mathbb{E}[\|\theta_{n+1} - \theta^*\|^2] \leq K_p \left[ e^{-\lambda \sum_{k=0}^{n} \alpha_k} \right] + \mathbb{E}[\|\theta_0 - \theta^*\|^2 + \frac{8K^2}{\lambda} K_p e^{\lambda/2} \left[ e^{-((\lambda)/2) \sum_{k=0}^{n} \alpha_k} \right] \frac{1}{(n+1)^{\sigma}}.
\]

Since

\[
\sum_{k=0}^{n} \alpha_k \geq \int_{0}^{n+1} \frac{1}{(x+1)^{\sigma}} dx = (n+2)^{1-\sigma} - 1,
\]

the desired result follows. \hfill \Box