The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Finite Sample Analyses for TD(0) with Function Approximation

Gal Dalal* gald@campus.technion.ac.il
Balázs Szörényi* 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 results. Works that managed to obtain convergence rates for online Temporal Difference (TD) methods analyzed somewhat modified versions of them that include projections and stepsize dependent on unknown problem parameters. Our analysis obviates these artificial alterations by exploiting strong properties of TD(0). We provide convergence rates both in expectation and with high-probability. Both are based on relatively unknown, recently developed stochastic approximation techniques.

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 π 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 $s \in \mathcal{S}$. However, in many problems, the state-space $\mathcal{S}$ 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., $V^\pi(s) \approx \theta^\top \phi(s)$. 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 non-linear RL architectures.

*Equal contribution

Two types of convergence rate results exist in literature: in expectation and with high probability. We stress that no results of either type exist for the actual, commonly used, TD(0) algorithm with linear FA; our work is the first to provide such results. In fact, it is the first work to give a convergence rate for an unaltered online TD algorithm of any type. We emphasize that TD(0) with linear FA is formulated and used with non-problem-specific stepsizes. Also, it does not require a projection step to keep θ in a 'nice' set. In contrast, the few recent works that managed to provide convergence rates for TD(0) analyzed only altered versions of them. These modifications include a projection step and eigenvalue-dependent stepsizes, or they apply only to the average of iterates; we expand on this in the coming section.

Existing Literature

The first TD(0) convergence result was obtained by (Tsitsiklis, Van Roy, and others 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, Maei, and Szepesvári 2009; Sutton et al. 2009; Sutton, Mahmood, and White 2015).

As opposed to the case of asymptotic convergence analysis of TD algorithms, very little is known about 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), convergence rates for TD(0) with mixing-time consideration have been given. We note that even though doubts were recently raised regarding the correctness results there (Narayanan and Szepesvári 2017), we shall treat them as correct for the sake of discussion. The results in (Korda and Prashanth 2015) 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 the average of iterates. Additionally, unlike in our work, a strong requirement for all 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). Results in similar fashion regarding required knowledge of system parameters are also given for TD(0) in (Konda 2002). 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. 2009; Sutton, Maei, and Szepesvári 2009). That work interpreted the algorithms as gradient methods to some saddle-point optimization problem. This enabled them to obtain convergence rates on altered versions of these algorithms using results from the convex optimization literature. Despite the alternate approach, in a 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.

Three additional works (Yu and Bertsekas 2009; Lazaric, Ghavamzadeh, and Munos 2010; Pan, White, and White 2017) 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.

Our Contributions

Our work is the first to give bounds on the convergence rate of TD(0) in its *original, unaltered form*. In fact, it is the first to obtain convergence rate results 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 stepsizes dependent on unknown problem parameters; alternatively, they only apply to average of iterates.

The methodologies for obtaining the expectation and high probability bounds are quite different. The former has a short and elegant proof that follows via induction using a subtle trick from (Kamal 2010). This bound applies to a general family of stepsizes that is not restricted to square-summable sequences, as usually was required by most previous works. This result reveals an explicit interplay between the stepsizes and noise.

As for the key ingredients in proving our high-probability bound, we first show that the n-th iterate at worst is only $O(n)$ away from the solution θ^*. Based on that, we then utilize tailor-made stochastic approximation tools to show that after some additional steps all subsequent iterates are ϵ-close to the solution w.h.p. This novel analysis approach allows us to obviate the common alterations mentioned above. Our key insight regards the role of the driving matrix's smallest eigenvalue λ. The convergence rate is dictated by it when it is below some threshold; for larger values, the rate is dictated by the noise.

We believe these two analysis approaches are not limited to TD(0) alone.

2 Problem Setup

We consider the problem of policy evaluation for a Markov Decision Process (MDP). A MDP is defined by the 5-tuple $(\mathcal{S}, \mathscr{A}, P, \mathcal{R}, \gamma)$ (Sutton 1988), where $\mathcal{S}$ is the set of states, $\mathscr{A}$ is the set of actions, $P = P(s'|s,a)$ is the transition kernel, $\mathcal{R}(s,a,s')$ is the reward function, and $\gamma \in (0,1)$ is the discount factor. In each time-step, the process is in some state $s \in \mathcal{S}$, an action $a \in \mathscr{A}$ is taken, the system transitions to a next state $s' \in \mathcal{S}$ according to the transition kernel P, and an immediate reward r is received according to $\mathcal{R}(s,a,s')$. Let policy $\pi : \mathcal{S} \to \mathscr{A}$ be a stationary mapping from states to actions. Assuming the associated Markov chain is ergodic and uni-chain, let ν be the induced stationary distribution. Moreover, let $V^\pi(s)$ be the value function at state s w.r.t. π 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., $V^\pi(s) \approx \theta^\top \phi(s)$, where $\phi(s) \in \mathbb{R}^d$ is a feature vector at state s, and $\theta \in \mathbb{R}^d$ is a weight vector. For brevity, we omit the notation π and denote $\phi(s),\ \phi(s')$ by $\phi,\ \phi'$.

Let $\{(\phi_n, \phi'_n, r_n)\}_n$ be iid samples of (ϕ, ϕ', r).[1] Then the TD(0) algorithm has the update rule

$$\theta_{n+1} = \theta_n + \alpha_n[r_n + \gamma \phi_n'^\top \theta_n - \phi_n^\top \theta_n]\phi_n, \tag{1}$$

where α_n is the stepsize. For analysis, we can rewrite the above as

$$\theta_{n+1} = \theta_n + \alpha_n[h(\theta_n) + M_{n+1}]\ , \tag{2}$$

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

$$M_{n+1} = \left(r_n + \gamma \phi_n'^\top \theta_n - \phi_n^\top \theta_n\right)\phi_n - [b - A\theta_n]\ , \tag{3}$$

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

$$h(\theta) = -A[\theta - \theta^*]\ . \tag{4}$$

[1]The iid assumption does not hold in practice; however, it is standard when dealing with convergence bounds in reinforcement learning (Liu et al. 2015; Sutton, Maei, and Szepesvári 2009; Sutton et al. 2009). It allows for sophisticated and well-developed techniques from SA theory, and it is not clear how it can be avoided. Indeed, the few papers that obviate this assumption assume other strong properties such as exponentially-fast mixing time (Korda and Prashanth 2015; Tsitsiklis, Van Roy, and others 1997). In practice, drawing samples from the stationary distribution is often simulated by taking the last sample from a long trajectory, even though knowing when to stop the trajectory is again a hard theoretical problem. Additionally, in most recent implementations of TD algorithms, long replay buffers that shuffle samples reduce the correlation between them, making this assumption more realistic.

3 Main Results

Our first main result is a bound on the expected decay rate of the TD(0) iterates. It requires the following assumption.

$\boldsymbol{\mathcal{A}_1}$. For some $K_s > 0$,

$$\mathbb{E}[\|M_{n+1}\|^2|\mathcal{F}_n] \leq K_s[1 + \|\theta_n - \theta^*\|^2].$$

This assumption follows from (3) when, for example, $\{(\phi_n, \phi'_n, r_n)\}_n$ have uniformly bounded second moments. The latter is a common assumption in such results; e.g., (Sutton et al. 2009; Sutton, Maei, and Szepesvári 2009).

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 3.1 (Expected Decay Rate for TD(0))**.** *Fix $\sigma \in (0,1)$ and let $\alpha_n = (n+1)^{-\sigma}$. Fix $\lambda \in (0, \lambda_{\min}(A + A^\top))$. Then, under $\boldsymbol{\mathcal{A}_1}$, for $n \geq 1$,*

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

where $K_1, K_2 \geq 0$ are some constants that depend on both λ and σ; see (11) and (12) for the exact expressions.

Remark 3.2 (Stepsize tradeoff – I)**.** *The exponentially decaying term in Theorem 3.1 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 σ on these two terms introduces the following tradeoff:*

1. *For σ 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 σ close to 1, the second term decays quickly and corresponds to small stepsizes 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 3.3 (Stepsize tradeoff – II)**.** *A practitioner might expect initially large stepsizes to speed up convergence. However, Theorem 3.1 shows that as σ becomes small, the convergence rate starts being dominated by the martingale difference noise; i.e., choosing a larger stepsize will help speed up convergence only up to some threshold.*

Remark 3.4 (Non square-summable stepsizes)**.** *In Theorem 3.1, 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.*

Our second main result is a high-probability bound for a specific stepsize. It requires the following assumption.

$\boldsymbol{\mathcal{A}_2}$. All rewards $r(s,a,s')$ and feature vectors $\phi(s)$ are uniformly bounded, i.e., $\|\phi(s)\| \leq 1/2, \ \forall s \in \mathcal{S}$, and $|r(s,a,s')| \leq 1, \forall s, s' \in \mathcal{S}, a \in \mathscr{A}$.

The $\tilde{O}$ notation hides problem dependent constants and poly-logarithmic terms.

Theorem 3.5 (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, under $\boldsymbol{\mathcal{A}_2}$, for $\epsilon > 0$ and $\delta \in (0,1)$, there exists a function*

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

such that

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

To enable direct comparison with previous works, one can obtain a following weaker implication of Theorem 3.5 by dropping quantifier $\forall$ inside the event. This translates to the following.

Theorem 3.6. *[TD(0) High-Probability Convergence Rate] Let λ and α_n be as in Theorem 3.5. Fix $\delta \in (0,1)$. Then, under $\boldsymbol{\mathcal{A}_2}$, there exists some function $N_0(\delta) = O(\ln(1/\delta))$ such that for all $n \geq N_0(\delta)$,*

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

Proof. Fix some n, and choose $\epsilon = \epsilon(n)$ so that $n = N(\epsilon,\delta)$. Then, on one hand, $1 - \delta \leq \Pr\{\|\theta_n - \theta^*\| \leq \epsilon\}$ due to Theorem 3.5 and, on the other hand, $\epsilon = \tilde{O}\left(n^{-\min\{1/2,\lambda/(\lambda+1)\}}\right)$ by the definition of $N(\epsilon,\delta)$. The claimed result follows. ☐

Remark 3.7 (Eigenvalue dependence)**.** *Theorem 3.6 shows that the rate improves as λ increases from 0 to 1; however, beyond 1 it remains fixed at $1/\sqrt{n}$. As seen in the proof of Theorem 3.5, this is because the rate is dictated by noise when $\lambda > 1$, and by the limiting ODE when $\lambda < 1$.*

Remark 3.8 (Comparison to (Korda and Prashanth 2015))**.** *Recently, doubts were raised in (Narayanan and Szepesvári 2017) regarding the correctness of the results in (Korda and Prashanth 2015). Nevertheless, given current form those results, the following discussion is in order.*

The expectation bound in Theorem 1, (Korda and Prashanth 2015) requires the TD(0) stepsize to satisfy $\alpha_n = f_n(\lambda)$ for some function f_n, where λ is as above. Theorem 2 there obviates this, but it applies to the average of iterates. In contrast, our expectation bound does not need any scaling of the above kind and applies directly to the TD(0) iterates. Moreover, our result applies to a broader family of stepsizes; see Remark 3.4. Our expectation bound when compared to that of Theorem 2, (Korda and Prashanth 2015) is of the same order (even though theirs is for the average of iterates). As for the high-probability concentration bounds in Theorems 1&2, (Korda and Prashanth 2015), they 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.

4 Proof of Theorem 3.1

We begin with an outline of the analysis conducted for Theorem 3.1. Our first key step is to identify a "nice" Liapunov function $V(\theta)$. Then, apply conditional expectation to get rid of the linear noise terms in the relation between $V(\theta_n)$ and $V(\theta_{n+1})$; this subtle trick appeared in (Kamal 2010). Lastly, induction leads to the desired result.

Our first two results hold for a general stepsize sequence that is not necessarily square-summable. Namely, $\{\alpha_n\}$ is a stepsize sequence satisfying $\sum_{n\geq 0}\alpha_n = \infty$, $\lim_{n\to\infty}\alpha_n = 0$ and $\sup_{n\geq 0}\alpha_n \leq 1$.

Notice that the matrices $(A^\top + A)$ and $(A^\top A + K_s I)$ are symmetric, where K_s is the constant from $\boldsymbol{\mathcal{A}_1}$. Further, as A is positive definite, the above matrices are also positive definite. Hence their minimum and maximum eigenvalues are strictly positive. This is used in the proofs in this section.

Lemma 4.1. *For $n \geq 0$, let $\lambda_n := \lambda_{\max}(\Lambda_n)$, where*

$$\Lambda_n := \mathbb{I} - \alpha_n(A + A^\top) + \alpha_n^2(A^\top A + K_s I).$$

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

$$\prod_{i=k}^{n}\lambda_i \leq K_p e^{-\lambda[\sum_{i=k}^{n}\alpha_i]}\ ,$$

where

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

with $\mu = -\lambda_{\min}(A + A^\top) + \lambda_{\max}(A^\top A + K_s I)$.

Note that such m exists since $\alpha_k \to 0$ as $k \to \infty$.

Proof. Using Weyl's inequality, we have

$$\lambda_n \leq \lambda_{\max}(\mathbb{I}-\alpha_n(A+A^\top)) + \alpha_n^2\lambda_{\max}(A^\top A+K_s I). \quad (5)$$

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

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

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

$$\lambda_n \leq e^{\alpha_n\mu}. \quad (6)$$

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

$$\begin{aligned}\lambda_n &\leq e^{-\lambda\alpha_n}e^{-\alpha_n[(\lambda_{\min}(A^\top+A)-\lambda)-\alpha_n\lambda_{\max}(A^\top A+K_s I)]}\\ &\leq e^{-\lambda\alpha_n}. \quad (7)\end{aligned}$$

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 (6) and (7), we have

$$\begin{aligned}\prod_{\ell=k}^{n}\lambda_\ell &\leq \left[\prod_{\ell=k}^{m}\lambda_\ell\right]e^{-\lambda(\sum_{\ell=m+1}^{n}\alpha_\ell)}\\ &= \left[\prod_{\ell=k}^{m}\lambda_\ell\right]e^{\lambda(\sum_{\ell=k}^{m}\alpha_\ell)}e^{-\lambda(\sum_{\ell=k}^{n}\alpha_\ell)}\\ &\leq K_p e^{-\lambda(\sum_{\ell=k}^{n}\alpha_\ell)}\ ,\end{aligned}$$

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

Using Lemma 4.1, we now prove a convergence rate in expectation for general stepsizes.

Theorem 4.2 (Technical Result: Expectation Bound)**.** *Fix $\lambda \in (0, \lambda_{\min}(A + A^\top))$. Then, under $\boldsymbol{\mathcal{A}_1}$,*

$$\begin{aligned}\mathbb{E}\|\theta_{n+1} - \theta^*\|^2 \leq& K_p\left[e^{-\lambda\sum_{k=0}^{n}\alpha_k}\right]\mathbb{E}\|\theta_0 - \theta^*\|^2\\ &+ K_s K_p\sum_{i=0}^{n}\left[e^{-\lambda\sum_{k=i+1}^{n}\alpha_k}\right]\alpha_i^2,\end{aligned}$$

where $K_p, K_s \geq 0$ are constants as defined in Lemmas 4.1 and $\boldsymbol{\mathcal{A}_1}$, respectively.

Proof. 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

$$\begin{aligned}&V(\theta_{n+1})\\ =&(\theta_{n+1} - \theta^*)^\top(\theta_{n+1} - \theta^*)\\ =&[(I - \alpha_n A)(\theta_n - \theta^*) + \alpha_n M_{n+1}]^\top\\ &\times[(I - \alpha_n A)(\theta_n - \theta^*) + \alpha_n M_{n+1}]\\ =&(\theta_n - \theta^*)^\top[I - \alpha_n(A^\top + A) + \alpha_n^2 A^\top A](\theta_n - \theta^*)\\ &+ \alpha_n(\theta_n - \theta^*)^\top(I - \alpha_n A)^\top M_{n+1}\\ &+ \alpha_n M_{n+1}^\top(I - \alpha_n A)(\theta_n - \theta^*) + \alpha_n^2\|M_{n+1}\|^2.\end{aligned}$$

Taking conditional expectation and using $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0$, we get

$$\begin{aligned}&\mathbb{E}[V(\theta_{n+1})|\mathcal{F}_n] = \alpha_n^2\mathbb{E}[\|M_{n+1}\|^2|\mathcal{F}_n]\\ &\quad+ (\theta_n - \theta^*)^\top[I - \alpha_n(A^\top + A) + \alpha_n^2 A^\top A](\theta_n - \theta^*).\end{aligned}$$

Therefore, using $\boldsymbol{\mathcal{A}_1}$,

$$\mathbb{E}[V(\theta_{n+1})|\mathcal{F}_n] \leq (\theta_n - \theta^*)^\top\Lambda_n(\theta_n - \theta^*) + K_s\alpha_n^2,$$

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

$$\mathbb{E}[V(\theta_{n+1})|\mathcal{F}_n] \leq \lambda_n V(\theta_n) + K_s\alpha_n^2.$$

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

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

Sequentially using the above inequality, we have

$$w_{n+1} \leq \left[\prod_{k=0}^{n}\lambda_k\right]w_0 + K_s\sum_{i=0}^{n}\left[\prod_{k=i+1}^{n}\lambda_k\right]\alpha_i^2.$$

Using Lemma 4.1 and using the constant K_p defined there, the desired result follows. $\square$

The next result provides closed form estimates of the expectation bound given in Theorem 4.2 for the specific stepsize sequence $\alpha_n = 1/(n+1)^\sigma$, with $\sigma \in (0, 1)$. Notice this family of stepsizes is more general than other common choices in the literature as it is non-square summable for $\sigma \in (0, 1/2]$. See Remark 3.4 for further details.

Theorem 4.3. *Fix $\sigma \in (0,1)$ and let $\alpha_n = 1/(n+1)^\sigma$. Then, under $\boldsymbol{\mathcal{A}_1}$,*

$$\begin{aligned}\mathbb{E}\|\theta_{n+1}-\theta^*\|^2 \leq& \bigg[K_p e^{\lambda}\mathbb{E}\|\theta_0-\theta^*\|^2 e^{-(\lambda/2)(n+2)^{1-\sigma}} \\ &+ \frac{2K_sK_pK_be^{\lambda}}{\lambda}\bigg] e^{-(\lambda/2)(n+2)^{1-\sigma}} \\ &+ \frac{2K_sK_pe^{\lambda/2}}{\lambda}\frac{1}{(n+1)^\sigma},\end{aligned}$$

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

Proof. Let $t_n = \sum_{i=0}^{n-1}\alpha_i$ for $n \geq 0$. Observe that

$$\begin{aligned}&\sum_{i=0}^{n}\left[e^{-(\lambda/2)\sum_{k=i+1}^{n}\alpha_k}\right]\alpha_i \\ &\leq \left(\sup_{i\geq 0} e^{(\lambda/2)\alpha_i}\right)\sum_{i=0}^{n}\left[e^{-(\lambda/2)\sum_{k=i}^{n}\alpha_k}\right]\alpha_i \\ &= \left(\sup_{i\geq 0} e^{(\lambda/2)\alpha_i}\right)\sum_{i=0}^{n}\left[e^{-(\lambda/2)(t_{n+1}-t_i)}\right]\alpha_i \\ &\leq \left(\sup_{i\geq 0} e^{(\lambda/2)\alpha_i}\right)\int_0^{t_{n+1}} e^{-(\lambda/2)(t_{n+1}-s)}\mathrm{d}s \\ &\leq \left(\sup_{i\geq 0} e^{(\lambda/2)\alpha_i}\right)\frac{2}{\lambda} \\ &\leq \frac{2e^{\lambda/2}}{\lambda},\end{aligned}$$

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

$$\sum_{i=0}^{n}\left[e^{-\lambda\sum_{k=i+1}^{n}\alpha_k}\right]\alpha_i^2 \tag{8}$$

$$\begin{aligned}&\leq \left(\sup_{0\leq i\leq n}\left[\alpha_i e^{-\frac{\lambda}{2}\sum_{k=i+1}^{n}\alpha_k}\right]\right)\sum_{i=0}^{n}\left[e^{-\frac{\lambda}{2}\sum_{k=i+1}^{n}\alpha_k}\right]\alpha_i \\ &\leq \left(\sup_{0\leq i\leq n}\left[\alpha_i e^{-\frac{\lambda}{2}\sum_{k=i+1}^{n}\alpha_k}\right]\right)\frac{2e^{\frac{\lambda}{2}}}{\lambda}.\end{aligned} \tag{9}$$

We claim that for all $n \geq i_0$,

$$\sup_{i_0\leq i\leq n}\left[\alpha_i e^{-(\lambda/2)\sum_{k=i+1}^{n}\alpha_k}\right] \leq \frac{1}{(n+1)^\sigma}. \tag{10}$$

To establish this, we show that for any $n \geq i_0$, $\alpha_i e^{-(\lambda/2)[\sum_{k=i+1}^{n}\alpha_k]}$ monotonically increases as i is varied from i_0 to n. To prove the latter, it suffices to show that $\alpha_i e^{-(\lambda/2)\alpha_{i+1}} \leq \alpha_{i+1}$, 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 (10) holds. From (9) and (10), we then have

$$\begin{aligned}&\sum_{i=0}^{n}\left[e^{-\lambda\sum_{k=i+1}^{n}\alpha_k}\right]\alpha_i^2 \\ &\leq \tfrac{2e^{\lambda/2}}{\lambda}\Bigg[\left(\sup_{0\leq i\leq i_0}\left[\alpha_i e^{-(\lambda/2)\sum_{k=i+1}^{n}\alpha_k}\right]\right) \\ &\qquad + \left(\sup_{i_0\leq i\leq n}\left[\alpha_i e^{-(\lambda/2)\sum_{k=i+1}^{n}\alpha_k}\right]\right)\Bigg] \\ &\leq \tfrac{2e^{\lambda/2}}{\lambda}\left[\left(\sup_{0\leq i\leq i_0}\left[\alpha_i e^{-(\lambda/2)\sum_{k=i+1}^{n}\alpha_k}\right]\right) + \tfrac{1}{(n+1)^\sigma}\right] \\ &\leq \tfrac{2e^{\lambda/2}}{\lambda}\Bigg[e^{-[(\lambda/2)\sum_{k=0}^{n}\alpha_k]}\left(\sup_{0\leq i\leq i_0}\left[\alpha_i e^{(\lambda/2)\sum_{k=0}^{i}\alpha_k}\right]\right) \\ &\qquad + \tfrac{1}{(n+1)^\sigma}\Bigg] \\ &\leq \tfrac{2e^{\lambda/2}}{\lambda}\left[K_b e^{-[(\lambda/2)\sum_{k=0}^{n}\alpha_k]} + \tfrac{1}{(n+1)^\sigma}\right],\end{aligned}$$

where the first relation holds as $\sup\{a_0,\ldots,a_n\} \leq \sup\{a_0,\ldots,a_{i_0}\} + \sup\{a_{i_0},\ldots,a_n\}$ 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 $\sup_{0\leq i\leq i_0} e^{(\lambda/2)\sum_{k=0}^{i}\alpha_k} \leq K_b$. Combining the above inequality with the relation from Theorem 4.2, we have

$$\begin{aligned}\mathbb{E}\|\theta_{n+1}-\theta^*\|^2 \leq& K_p\left[e^{-\lambda\sum\limits_{k=0}^{n}\alpha_k}\right]\mathbb{E}\|\theta_0-\theta^*\|^2 \\ &+ \frac{2K_sK_pe^{\lambda/2}}{\lambda}\left[K_b e^{-\left[(\lambda/2)\sum\limits_{k=0}^{n}\alpha_k\right]} + \frac{1}{(n+1)^\sigma}\right],\end{aligned}$$

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. ☐

To finalize the proof of Theorem 3.1 we employ Theorem 4.3 with the following constants.

$$K_1 = K_pe^{\lambda}\mathbb{E}\|\theta_0-\theta^*\|^2 + \frac{2K_sK_pK_be^{\lambda}}{\lambda}, \tag{11}$$

$$K_2 = \frac{2K_sK_pe^{\lambda/2}}{\lambda}, \tag{12}$$

where K_s is the constant from $\boldsymbol{\mathcal{A}_1}$,

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

with $\mu = -\lambda_{\min}(A+A^\top) + \lambda_{\max}(A^\top A + K_sI)$ and $m =$

$\left\lceil \left(\frac{\lambda_{\max}(A^\top A+K_s I)}{\lambda_{\min}(A+A^\top)-\lambda}\right)^{1/\sigma} \right\rceil$, and

$$K_b = \exp\left[(\lambda/2)\sum_{k=0}^{\lceil (2\sigma/\lambda)^{1/(1-\sigma)}\rceil} \alpha_k\right]$$
$$\leq \exp\left[(\lambda/2)\frac{\left(\lceil (2\sigma/\lambda)^{1/(1-\sigma)}\rceil+1\right)^{1/\sigma}+\sigma}{1-\sigma}+1\right].$$

5 Proof of Theorem 3.5

This section outlines the analysis conducted for Theorem 3.5. Throughout this section we assume $\mathcal{A}_2$. All proofs are given in Appendix B.

Outline of Approach

We begin with defining the limiting ODE for (2):

$$\dot\theta(t) = h(\theta(t)) = b - A\theta(t) = -A(\theta(t)-\theta^*) \ . \quad (13)$$

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 $\theta(t)$.

As the solutions of the ODE are continuous functions of time, we also 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

$$\bar\theta(\tau) = \begin{cases} \theta_n & \text{if } \tau = t_n \ , \\ \theta_n + \frac{\tau - t_n}{\alpha_n}[\theta_{n+1}-\theta_n] & \text{if } \tau \in (t_n, t_{n+1}) \ . \end{cases} \quad (14)$$

Our tool for comparing $\bar\theta(t)$ to $\theta(t)$ is the *Variation of Parameters* (VoP) method (Lakshmikantham and Deo 1998). Initially, $\bar\theta(t)$ could stray away from θ^* when the stepsizes may not be small enough to tame the noise. However, we show that $\|\bar\theta(t_n)-\theta^*\| = O(n)$, i.e., θ_n does not stray away from θ^* 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 an $O(n_0)$ distance from θ^*; 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.

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, (Hirsch, Smale, and Devaney 2012), it follows that the solution $\theta(t,s,u_0)$, $t \geq s$, of (13) satisfies the relation

$$\theta(t,s,u_0) = \theta^* + e^{-A(t-s)}(u_0-\theta^*) \ . \quad (15)$$

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

$$\|\theta(t',s,u_0)-\theta^*\| \leq \|\theta(t,s,u_0)-\theta^*\| \ , \quad (16)$$

for all $t' \geq t \geq s$ and u_0.

Let λ be as in Theorem 3.5. From Corollary 3.6, p71, (Teschl 2012), $\exists K_\lambda \geq 1$ so that $\forall t \geq s$

$$\|e^{-A(t-s)}\| \leq K_\lambda\, e^{-\lambda(t-s)} \ . \quad (17)$$

Separately, as $t_{n+1}-t_{k+1} = \sum_{\ell=k+1}^n \alpha_\ell = \sum_{\ell=k+1}^n \frac{1}{\ell+1}$,

$$\frac{(k+1)^\lambda}{(n+1)^\lambda} \leq e^{-\lambda(t_{n+1}-t_{k+1})} \leq \frac{(k+2)^\lambda}{(n+2)^\lambda} \ . \quad (18)$$

The following result is a consequence of $\mathcal{A}_2$ that gives a bound directly on the martingale difference noise as a function of the iterates. We emphasize that this strong behavior of TD(0) is significant in our work. We also are not aware of other works that utilized it even though $\mathcal{A}_2$ or equivalents are often assumed and widely accepted.

Lemma 5.1 (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 5.2. *The noise behavior usually used in the literature (e.g., (Sutton et al. 2009; Sutton, Maei, and Szepesvári 2009)) is the same as we assumed in $\mathcal{A}_1$ for Theorem 3.1:*

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

for some constant $K_s \geq 0$. However, here we assume the stronger $\mathcal{A}_2$, which, using a similar proof technique to that of Lemma 5.1, implies

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

for all $n \geq 0$.

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 (14). The key idea then is to use the VoP method (Lakshmikantham and Deo 1998) as in Lemma A.1, and express $\bar\theta(t)$ as a perturbation of $\theta(t)$ due to two factors: the discretization error and the martingale difference noise. Our quantification of these two factors is as follows. For the interval $[t_{\ell_1}, t_{\ell_2}]$, let

$$E^{\mathrm{d}}_{[\ell_1,\ell_2]} := \sum_{k=\ell_1}^{\ell_2-1}\int_{t_k}^{t_{k+1}} e^{-A(t_{n+1}-\tau)}A[\bar\theta(\tau)-\theta_k]\mathrm{d}\tau \ ,$$

and

$$E^{\mathrm{m}}_{[\ell_1,\ell_2]} := \sum_{k=\ell_1}^{\ell_2-1}\left[\int_{t_k}^{t_{k+1}} e^{-A(t_{n+1}-\tau)}\mathrm{d}\tau\right]M_{k+1} \ .$$

Corollary 5.3 (Comparison of SA Trajectory and ODE Solution)**.** *For every $\ell_2 \geq \ell_1$,*

$$\bar\theta(t_{\ell_2}) - \theta^* = \theta(t_{\ell_2}, t_{\ell_1}, \bar\theta(t_{\ell_1})) - \theta^* + E^d_{[\ell_1,\ell_2]} + E^m_{[\ell_1,\ell_2]} \ .$$

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_{ℓ_1}. Consequently, by bounding $E^{\mathrm{d}}_{[\ell_1,\ell_2]}$ and $E^{\mathrm{m}}_{[\ell_1,\ell_2]}$ we can estimate the distance of interest.

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

Table 1: Chronological Summary of Analysis Outline

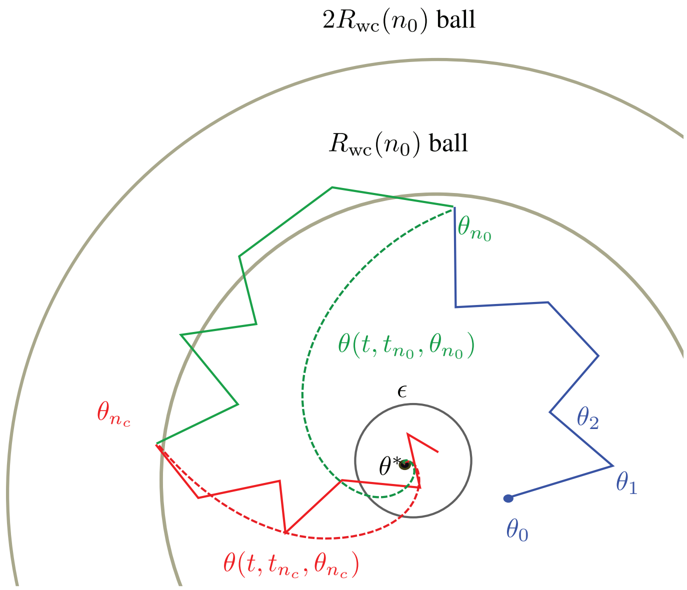

Figure 1: Visualization of the proof outline. The three balls (from large to small) are respectively the $2R_{\text{wc}}(n_0)$ ball, $R_{\text{wc}}(n_0)$ ball, and ϵ ball, where $R_{\text{wc}}(n_0)$ is from Lemma 5.4. The blue curve is the initial, possibly diverging phase of $\bar{\theta}(t)$. The green curve is $\bar{\theta}(t)$ when the stepsizes are moderate in size ($t_{n_0} \leq t \leq t_{n_c}$ in the analysis). Similarly, the red curve is $\bar{\theta}(t)$ when the stepsizes are sufficiently small ($t > t_{n_c}$). The dotted curves are the associated ODE trajectories $\theta(t, t_n, \theta_n)$.

Part I – Initial Possible Divergence

In this section, we show that the TD(0) iterates lie in an $O(n)$-ball around θ^*. We stress that this is one of the results that enable us to accomplish more than existing literature. Previously, the distance of the initial iterates from θ^* 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 5.4 (Worst-case Iterates Bound)**.** *For $n \geq 0$,*

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

where

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

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 5.1 and 5.4.

Corollary 5.5 (Worst-case Noise Bound)**.** *For $n \geq 0$,*

$$\|M_{n+1}\| \leq K_m \left[1 + C_* R_0\right][n+1] \ .$$

Part II – Rate of Convergence

In this section we bound 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; how large they should be will be elaborated later. We do this by comparing the TD(0) trajectory θ_n with the ODE solution $\theta(t_n, t_{n_0}, \bar{\theta}(t_{n_0})) \ \forall n \geq n_0$; for this we will use Corollary 5.3 along with Lemma 5.4. Next, we show that if n_0 is sufficiently large, or equivalently the stepsizes $\{\alpha_n\}_{n \geq n_0}$ are small enough, then after a finite number of iterations from n_0, the TD(0) iterates are $\epsilon-$close to θ^* w.h.p. This holds as the small stepsize and sufficiently long waiting time ensure that the ODE solution $\theta(t_{n+1}, t_{n_0}, \bar{\theta}(t_{n_0}))$ is $\epsilon-$close to θ^*, the discretization error $E^{\text{d}}_{[n_0, n+1]}$ is small, and martingale difference noise $E^{\text{m}}_{[n_0, n+1]}$ is small w.h.p. Having summarized the outline, we now continue with the formal proof.

Let $\delta \in (0,1)$, and let ϵ 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 $O(n_0)$ ball around θ^*. For $n < n_0 + n_1$, the bad event means that θ_{n+1} is outside the $O(n_0)$ ball around θ^*, while for $n \geq n_0 + n_1$, the bad event means that θ_{n+1} is outside the ϵ ball around θ^*. Formally, for $n_1 \geq 1$, define the events

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

$$\mathcal{E}^{\text{after}}_{n_0,n_1} \tag{19}$$

$$:= \bigcup_{n=n_0+n_1}^{\infty} \{G_{n_0,n}, \|\theta_{n+1} - \theta^*\| > \min\{\epsilon, 2R_{\text{wc}}(n_0)\}\} \ , \tag{20}$$

and, $\forall n \geq n_0$, let

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

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

Lemma 5.6 (Decomposition of Event of Interest)**.** *For* $n_0, n_1 \geq 1$,

$$\mathcal{E}^c(n_0, n_1) \subseteq \mathcal{E}^{\text{mid}}_{n_0,n_1} \cup \mathcal{E}^{\text{after}}_{n_0,n_1} \quad .$$

For the following results, define the constants

$$C_{m2} := \begin{cases} \frac{6K_m\, K_\lambda\, 2^{\lambda-0.5}}{\sqrt{2\lambda-1}} & \text{if } \lambda > 0.5 \\ \frac{6K_m\, K_\lambda}{\sqrt{1-2\lambda}} & \text{if } \lambda < 0.5 \end{cases} \quad .$$

Next, we show that on the "good" event $G_{n_0,n}$, the discretization error is small for all sufficiently large n.

Lemma 5.7 (Part II Discretization Error Bound)**.** *For any*

$$n \geq n_0 \geq \tfrac{K_\lambda 6\|A\|(\|A\|+2K_m\,)}{\lambda},$$

$$\|E^d_{[n_0,n+1]}\| \leq \tfrac{1}{3}[n_0+1]C_* R_0 = \tfrac{1}{3}R_{\text{wc}}(n_0).$$

Furthermore, for

$$n \geq n_{\text{c}} \geq \left(1 + \tfrac{K_\lambda 6\|A\|(\|A\|+2K_m\,)C_* R_0}{\lambda \min\{\epsilon, R_{\text{wc}}(n_0)\}}\right)(n_0+1)$$

it thus also holds on $G_{n_0,n}$ that

$$\begin{aligned}\|E^d_{[n_{\text{c}},n+1]}\| &\leq \tfrac{1}{3}\min\{\epsilon, [n_0+1]C_* R_0\} \\ &= \tfrac{1}{3}\min\{\epsilon, R_{\text{wc}}(n_0)\} \quad .\end{aligned}$$

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 λ.

Lemma 5.8 (Part II Martingale Difference Noise Concentration)**.** *Let $n_0 \geq 1$ and $R \geq 0$. Let $n \geq n' \geq n_0$.*

- *For $\lambda > 1/2$,*

$$\begin{aligned}&\Pr\{G_{n_0,n}, \|E^m_{[n',n+1]}\| \geq R\} \\ &\qquad \leq 2d^2 \exp\left[-\frac{(n+1)R^2}{2d^3 C^2_{m2} R^2_{\text{wc}}(n_0)}\right] \quad .\end{aligned}$$

- *For $\lambda < 1/2$,*

$$\begin{aligned}&\Pr\{G_{n_0,n}, \|E^m_{[n',n+1]}\| \geq R\} \\ &\qquad \leq 2d^2 \exp\left[-\frac{[n'+1]^{1-2\lambda}(n+1)^{2\lambda}R^2}{2d^3 C^2_{m2} R^2_{\text{wc}}(n_0)}\right] \quad .\end{aligned}$$

Having Lemma 5.7, we substitute $R = \frac{R_{\text{wc}}(n_0)}{2}$ in Lemma 5.8 and estimate the resulting sum to bound $\mathcal{E}^{\text{mid}}_{n_0,n_1}$.

Lemma 5.9 (Bound on Probability of $\mathcal{E}^{\text{mid}}_{n_0,n_1}$)**.** *Let* $n_0 \geq \max\left\{\frac{K_\lambda 6\|A\|(\|A\|+2K_m\,)}{\lambda}, 2^{\frac{1}{\lambda}}\right\}$ *and* $n_1 \geq 1$.

- *If $\lambda > 1/2$, then*

$$\Pr\{\mathcal{E}^{\text{mid}}_{n_0,n_1}\} \leq 16d^5 C^2_{m2} \exp\left[-\frac{n_0}{8d^3 C^2_{m2}}\right] \quad .$$

- *If $\lambda < 1/2$, then*

$$\Pr\{\mathcal{E}^{\text{mid}}_{n_0,n_1}\} \leq 2d^2 \left[\frac{8d^3 C^2_{m2}}{\lambda}\right]^{\frac{1}{2\lambda}} \frac{\exp[-\frac{n_0}{64d^3 C^2_{m2}}]}{(n_0+1)^{\frac{1-2\lambda}{2\lambda}}} \quad .$$

Lastly, we upper bound $\mathcal{E}^{\text{after}}_{n_0,n_1}$ in the same spirit as $\mathcal{E}^{\text{mid}}_{n_0,n_1}$ in Lemma 5.9, again using Lemmas 5.7 and 5.8, this time with $R = \frac{\epsilon}{3}$. For brevity, the result is given in Lemma B.1 in Appendix B.

We are now ready to put the pieces together for proving Theorem 3.5. For the detailed calculations see end of Appendix B.

Proof of Theorem 3.5. From Lemma 5.6, by a union bound,

$$\Pr\{\mathcal{E}^c(n_0, n_1)\} \leq \Pr\{\mathcal{E}^{\text{mid}}_{n_0,n_1}\} + \Pr\{\mathcal{E}^{\text{after}}_{n_0,n_1}\} \quad .$$

The behavior of $\mathcal{E}^{\text{mid}}_{n_0,n_1}$ is dictated by n_0, while the behavior of $\mathcal{E}^{\text{after}}_{n_0,n_1}$ by n_1. Using Lemma 5.9, we set n_0 so that $\mathcal{E}^{\text{mid}}_{n_0,n_1}$ is less than $\delta/2$, resulting in the condition $n_0 = O\left(\ln \frac{1}{\delta}\right)$. Next, using Lemma B.1, we set n_1 so that $\mathcal{E}^{\text{after}}_{n_0,n_1}$ is less than $\delta/2$, resulting in

$$n_1 = \tilde{O}\left(\left[(1/\epsilon)\ln(1/\delta)\right]^{\max\{1+1/\lambda, 2\}}\right)$$

for $\lambda > 1/2$, and

$$n_1 = \tilde{O}\left(\left[(1/\epsilon)\ln(1/\delta)\right]^{1+1/\lambda}\right)$$

for $\lambda < 1/2$. $\square$

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.

As can be seen from Theorem 3.5, the bound explodes when the matrix A is ill-conditioned. We stress that this is not an artifact of the bound but an inherent property of the algorithm itself. This happens because along the eigenspace corresponding to the zero eigenvalues, the limiting ODE makes no progress and consequently no guarantees for the (noisy) TD(0) method can be given in this eigenspace. As is well known, the ODE will, however, advance in the eigenspace corresponding to the non-zero eigenvalues to a solution which we refer to as the truncated solution. Given this, one might expect that the (noisy) TD(0) method may also converge to this truncated solution. We now provide a short example that suggests that this is in fact not the case. Let

$$A := \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \text{ and } b := \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

Clearly, $\theta^* := [1 \quad 1]^\top$ is a vector satisfying $b = A\theta^*$ and the eigenvalues of A are 1 and 0. Consider the update rule $\theta_{n+1} = \theta_n + \alpha_n[b - A\theta_n + M_{n+1}]$ with

$$M_{n+1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} Z_{n+1}[\theta_n(2) - \theta^*(2)].$$

Here $v(i)$ is the $i-$th coordinate of vector v, and $\{Z_n\}$ are IID Bernoulli $\{-1, +1\}$ random variables. For an initial value θ_0, one can see that the (unperturbed) ODE for the above update rule converges to $[-1 \quad 1]^\top \theta_0(2) + b$; this is not θ^*, but the truncated solution mentioned above. For the same initial point, predicting the behavior of the noisy update is not easy. Rudimentary simulations show the following. In the initial phase (when the stepsizes are large) the noise dictates how the iterates behave. Afterwards, at a certain stage when the stepsizes become sufficiently small, an effective θ_0 is detected, from which the iterates start converging to a new truncated solution, corresponding to this effective θ_0. This new truncated solution is different per each run and is often very different from the truncated solution corresponding the initial iterate θ_0.

Separately, we stress that our proof technique is general and can be used to provide convergence rates for TD with non-linear function approximations, such as neural networks. Specifically, this can be done using the non-linear analysis presented in (Thoppe and Borkar 2015). There, the more general form of Variation of Parameters is used: the so-called Alekseev's formula. However, as mentioned in Section 1, the caveat there is that the n_0-th iterate needs to be in the domain of attraction of the desired asymptotically stable equilibrium point. Nonetheless, we believe that one should be able to extend our present approach to non-linear ODEs with a unique global equilibrium point. For non-linear ODEs with multiple stable points, the following approach can be considered. In the initial phase, the location of the SA iterates is a Markov chain with the state space being the domain of attraction associated with different attractors (Williams and others 2002). Once the stepsizes are sufficiently small, analysis as in our current paper via Alekseev's formula should enable one to obtain expectation and high probability concentration bounds.

Finally, future work can extend to a more general family learning rates, including the commonly used adaptive ones. Building upon Remark 5.2, we believe that a stronger expectation bound may hold for TD(0) with uniformly bounded features and rewards. This may enable obtaining tighter concentration bounds for TD(0) even with generic stepsizes.

References

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

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

Borkar, V. S. 2008. *Stochastic approximation: a dynamical systems viewpoint*.

Fathi, M., and Frikha, N. 2013. Transport-entropy inequalities and deviation estimates for stochastic approximation schemes. *Electron. J. Probab.* 18:36 pp.

Frikha, N., and Menozzi, S. 2012. Concentration bounds for stochastic approximations. *Electron. Commun. Probab.* 17:15 pp.

Hirsch, M. W.; Smale, S.; and Devaney, R. L. 2012. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press.

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

Konda, V. 2002. *Actor-Critic Algorithms*. Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, MIT.

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

Lakshmikantham, V., and Deo, S. 1998. *Method of variation of parameters for dynamic systems*. CRC Press.

Lazaric, A.; Ghavamzadeh, M.; and Munos, R. 2010. Finite-sample analysis of lstd. In *ICML-27th International Conference on Machine Learning*, 615–622.

Liu, B.; Liu, J.; Ghavamzadeh, M.; Mahadevan, S.; and Petrik, M. 2015. Finite-sample analysis of proximal gradient td algorithms. In *UAI*, 504–513. Citeseer.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Narayanan, C., and Szepesvári, C. 2017. Finite time bounds for temporal difference learning with function approximation: Problems with some "state-of-the-art" results. Technical Report.

Pan, Y.; White, A. M.; and White, M. 2017. Accelerated gradient temporal difference learning. In *AAAI*, 2464–2470.

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

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Sutton, R. S., and Barto, A. G. 1998. *Introduction to Reinforcement Learning*. Cambridge, MA, USA: MIT Press, 1st edition.

Sutton, R. S.; Maei, H. R.; Precup, D.; Bhatnagar, S.; Silver, D.; Szepesvári, C.; and Wiewiora, E. 2009. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 993–1000. ACM.

Sutton, R. S.; Maei, H. R.; and Szepesvári, C. 2009. A convergent o (n) temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*, 1609–1616.

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

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

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

Teschl, G. 2012. *Ordinary Differential Equations and Dynamical Systems*.

Thoppe, G., and Borkar, V. S. 2015. A concentration bound for stochastic approximation via alekseev's formula. *arXiv:1506.08657*.

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

Williams, N., et al. 2002. Stability and long run equilibrium in stochastic fictitious play. *Manuscript, Princeton University*.

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

A Variation of Parameters Formula

Let $\theta(t, s, \bar{\theta}(s))$, $t \geq s$, be the solution to (13) starting at $\bar{\theta}(s)$ at time $t = s$. For $k \geq 0$, and $\tau \in [t_k, t_{k+1})$, let

$$\zeta_1(\tau) := h(\theta_k) - h(\bar{\theta}(\tau)) = A[\bar{\theta}(\tau) - \theta_k] \tag{21}$$

and

$$\zeta_2(\tau) := M_{k+1} \ . \tag{22}$$

Lemma A.1. *Let $i \geq 0$. For $t \geq t_i$.*

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

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), (21), and (22), and using the relations $\tau - t_n = \int_{t_n}^{t} \mathrm{d}\tau$ and $\alpha_k = \int_{t_k}^{t_{k+1}} \mathrm{d}\tau$, we have

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

Separately, writing (13) in integral form, we have

$$\theta(t, t_i, \bar{\theta}(t_i)) = \bar{\theta}(t_i) + \int_{t_i}^{t} h(\theta(\tau))\mathrm{d}\tau.$$

From the above two relations and the VoP formula (Lakshmikantham and Deo 1998), the desired result follows. □

B Supplementary Material for Proof of Theorem 3.5

Proof of Lemma 5.1. We have

$$\begin{aligned}
\|M_{n+1}\| &= \|r_n\phi_n + (\gamma\phi'_n - \phi_n)^\top \theta_n \phi_n - [b - A\theta_n]\| \\
&= \|r_n\phi_n + (\gamma\phi'_n - \phi_n)^\top (\theta_n - \theta^*)\phi_n \\
&\quad + (\gamma\phi'_n - \phi_n)^\top \theta^* \phi_n + A(\theta_n - \theta^*)\| \\
&\leq \frac{1}{2} + \frac{[1+\gamma]}{4}\|A^{-1}\| \, \|b\| + \frac{[1 + \gamma + 4\|A\|]}{4}\|\theta_n - \theta^*\|,
\end{aligned}$$

where the first relation follows from (3), the second holds as $b = A\theta^*$, while the third follows since $\boldsymbol{\mathcal{A}_2}$ holds and $\theta^* = A^{-1}b$. The desired result is now easy to see. □

Proof of Corollary 5.3. The result follows by using Lemma A.1 from Appendix A, with $i = \ell_1$, $t = t_{\ell_2}$, and subtracting θ^* from both sides. □

Proof of Lemma 5.4. The proof is by induction. The claim holds trivially for $n = 0$. Assume the claim for n. Then from (1),

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

Applying the Cauchy-Schwarz inequality, and using $\boldsymbol{\mathcal{A}_2}$ 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 5.6. For any two events $\mathcal{E}_1$ and $\mathcal{E}_2$, note that

$$\mathcal{E}_1 = [\mathcal{E}_2^c \cap \mathcal{E}_1] \cup [\mathcal{E}_2 \cap \mathcal{E}_1] \subseteq \mathcal{E}_2^c \cup [\mathcal{E}_2 \cap \mathcal{E}_1] \ . \tag{23}$$

Separately, for any sequence of events $\{\mathcal{E}_k\}$, observe that

$$\bigcup_{k=1}^{m} \mathcal{E}_k = \left[\bigcup_{k=1}^{m} \left(\left[\bigcup_{i=1}^{k-1} \mathcal{E}_i \right]^c \cap \mathcal{E}_k \right) \right] \ , \tag{24}$$

where $\bigcup_{i=i_1}^{i_2} \mathcal{E}_i = \emptyset$ whenever $i_1 > i_2$. Using (23), we have

$$\mathcal{E}^c(n_0, n_1) \subseteq G^c_{n_0, n_0+n_1} \cup [G_{n_0, n_0+n_1} \cap \mathcal{E}^c(n_0, n_1)] \ . \tag{25}$$

From Lemma 5.4, $\{\|\theta_{n_0} - \theta^*\| \leq R_{\text{wc}}(n_0)\}$ is a certain event. Hence it follows from (24) that

$$G^c_{n_0, n_0+n_1} = \mathcal{E}^{\text{mid}}_{n_0, n_1}. \tag{26}$$

Similarly, from (24) and the fact that $\epsilon \leq R_0$,

$$G_{n_0, n_0+n_1} \cap \mathcal{E}^c(n_0, n_1) \subseteq \mathcal{E}^{\text{after}}_{n_0, n_1} \ . \tag{27}$$

Substituting (26) and (27) in (25) 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 5.7. For $n \geq n' \geq n_0 \geq 0$, by its definition and the triangle inequality,

$$\|E^{\text{d}}_{[n', n+1]}\| \leq \sum_{k=n'}^{n} \int_{t_k}^{t_{k+1}} \|e^{-A(t_{n+1}-\tau)}\| \|A\| \|\bar{\theta}(\tau) - \theta_k\| \mathrm{d}\tau.$$

Fix a $k \in \{n', \ldots, n\}$ and $\tau \in [t_k, t_{k+1})$. Then using (14), (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 5.1, 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 $R_{\text{wc}}(n_0) \geq 1$, we have

$$\|\bar{\theta}(\tau) - \theta_k\| \leq 2[\|A\| + 2K_m]\alpha_{n'}[n_0 + 1]C_* R_0 \ .$$

From the above discussion, (17), the stepsize choice, and the facts that

$$\sum_{k=n'}^{n} \int_{t_k}^{t_{k+1}} e^{-\lambda(t_{n+1}-\tau)} \mathrm{d}\tau = \int_{t_{n'}}^{t_{n+1}} e^{-\lambda(t_{n+1}-\tau)} \mathrm{d}\tau \leq \frac{1}{\lambda} \ ,$$

and $\alpha_k \leq \alpha_{n'} \leq \alpha_{n_0}$, we get

$$\|E^{\text{d}}_{[n', n+1]}\| \leq \tfrac{K_\lambda 2\|A\|(\|A\| + 2K_m)(n_0+1)C_* R_0}{\lambda(n'+1)} \ .$$

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

Proof of Lemma 5.8. Let $Q_{k,n} = \int_{t_k}^{t_{k+1}} e^{-A(t_{n+1}-\tau)} \mathrm{d}\tau$. Then, for any $n_0 \leq n' \leq n$,

$$E^m_{[n', n+1]} = \sum_{k=n'}^{n} Q_{k,n} M_{k+1} \ ,$$

a sum of martingale differences. When the event $G_{n_0, n}$ holds, it follows that the indicator $1_{G_{n_0,k}} = 1 \ \forall k \in \{n_0, \ldots, n', \ldots n\}$. Hence, for any $R \geq 0$,

$$\begin{aligned}
\Pr\{G_{n_0,n}, \|E^m_{[n', n+1]}\| \geq R\} &= \Pr\left\{ G_{n_0,n}, \left\| \sum_{k=n'}^{n} Q_{k,n} M_{k+1} 1_{G_{n_0,k}} \right\| \geq R \right\} \\
&\leq \Pr\left\{ \left\| \sum_{k=n'}^{n} Q_{k,n} M_{k+1} 1_{G_{n_0,k}} \right\| \geq R \right\}.
\end{aligned}$$

Let $Q_{k,n}^{ij}$ be the $i,j-$th entry of the matrix $Q_{k,n}$ and let M_{k+1}^{j} be the $j-$th coordinate of M_{k+1}. Then using the union bound twice on the above relation, we have

$$\Pr\{G_{n_0,n}, \|E^{m}_{[n',n+1]}\| \geq R\} \leq \sum_{i=1}^{d}\sum_{j=1}^{d} \Pr\left\{\left|\sum_{k=n'}^{n} Q_{k,n}^{ij} M_{k+1}^{j} 1_{G_{n_0,k}}\right| \geq \frac{R}{d\sqrt{d}}\right\}.$$

As $|Q_{k,n}^{ij} M_{k+1}^{j}| 1_{G_{n_0,k}} \leq \|Q_{k,n}\| \|M_{k+1}\| 1_{G_{n_0,k}} =: \beta_{k,n}$, Azuma-Hoeffding inequality now gives

$$\Pr\{G_{n_0,n}, \|E^{m}_{[n',n+1]}\| \geq R\} \leq 2d^2 \exp\left[-\frac{R^2}{2d^3 \sum_{k=n'}^{n} \beta_{k,n}^2}\right] \quad . \tag{28}$$

On the event $G_{n_0,k}$, $\|\theta_k - \theta^*\| \leq 2R_{\mathrm{wc}}(n_0)$ by definition. Hence from Lemma 5.1, we have

$$\|M_{k+1}\| 1_{G_k} \leq 3K_m\, R_{\mathrm{wc}}(n_0) \quad . \tag{29}$$

Also from (17), $\|Q_{k,n}\| \leq K_\lambda\, e^{-\lambda(t_{n+1}-t_{k+1})}\alpha_k$. Combining the two inequalities, and using (18) along with the fact that $1/(k+1) \leq 2/(k+2)$, we get

$$\begin{aligned} \beta_{k,n} &\leq 3K_m\, K_\lambda\, R_{\mathrm{wc}}(n_0) e^{-\lambda(t_{n+1}-t_{k+1})}\alpha_k \\ &\leq 6K_m\, K_\lambda\, R_{\mathrm{wc}}(n_0) \frac{(k+2)^{\lambda-1}}{(n+2)^{\lambda}} \quad . \end{aligned}$$

Consider the case $\lambda > 1/2$. By treating the sum as a right Riemann sum, we have

$$\sum_{k=n'}^{n} (k+2)^{2\lambda-2} \leq (n+3)^{2\lambda-1}/(2\lambda-1) \quad .$$

As $(n+3) \leq 2(n+2)$ and $(n+2) \geq (n+1)$, we have

$$\sum_{k=n'}^{n} \beta_{k,n}^2 \leq C_{m2}^2 \frac{R_{\mathrm{wc}}^2(n_0)}{n+1} \quad .$$

Now consider the case $\lambda < 1/2$. Again treating the sum as a right Riemann sum, we have

$$\sum_{k=n'}^{n} (k+2)^{2\lambda-2} \leq \frac{1}{(1-2\lambda)[n'+1]^{1-2\lambda}} \quad .$$

As $(n+2) \geq (n+1)$, it follows that

$$\sum_{k=n'}^{n} \beta_{k,n}^2 \leq C_{m2}^2 \frac{R_{\mathrm{wc}}^2(n_0)}{[n'+1]^{1-2\lambda}(n+1)^{2\lambda}} \quad .$$

Substituting $\sum_{k=n_0}^{n} \beta_{k,n}^2$ bounds in (28), the desired result is easy to see. ☐

Conditional Results on the Bad Events

On the first "bad" event $\mathcal{E}^{\mathrm{mid}}_{n_0,n_1}$, the TD(0) iterate θ_n for at least one n between n_0+1 and n_0+n_1 leaves the $2R_{\mathrm{wc}}(n_0)$ ball around θ^*. Lemma 5.9 shows that this event has low probability. Its proof is the following.

Proof of Lemma 5.9. From Corollary 5.3, we have

$$\|\theta_{n+1} - \theta^*\| \leq \|\theta(t_{n+1}, t_{n_0}, \theta_{n_0}) - \theta^*\| + \|E^{\mathrm{d}}_{[n_0,n+1]}\| + \|E^{\mathrm{m}}_{[n_0,n+1]}\| \quad .$$

Suppose the event $G_{n_0,n}$ holds. Then from (16),

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

Also, as $n_0 \geq \frac{K_\lambda 6\|A\|(\|A\|+2K_m)}{\lambda}$, by Lemma 5.7, $\|E^{\mathrm{d}}_{[n_0,n+1]}\| \leq R_{\mathrm{wc}}(n_0)/3$. From all of the above, we have

$$\{G_{n_0,n}, \|\theta_{n+1} - \theta^*\| > 2R_{\mathrm{wc}}(n_0)\} \subseteq \{G_{n_0,n}, \|E^{\mathrm{m}}_{[n_0,n+1]}\| > R_{\mathrm{wc}}(n_0)/2\} \quad .$$

From this, we get

$$\mathcal{E}^{\text{mid}}_{n_0,n_1} \subseteq \bigcup_{n=n_0}^{n_0+n_1-1} \left\{G_{n_0,n}, \|E^{\text{m}}_{[n_0,n+1]}\| > \tfrac{R_{\text{wc}}(n_0)}{2}\right\}$$
$$\subseteq \bigcup_{n=n_0}^{\infty} \left\{G_{n_0,n}, \|E^{\text{m}}_{[n_0,n+1]}\| > \tfrac{R_{\text{wc}}(n_0)}{2}\right\} \quad .$$

Consequently,

$$\Pr\{\mathcal{E}^{\text{mid}}_{n_0,n_1}\} \leq \sum_{n=n_0}^{\infty} \Pr\left\{G_{n_0,n}, \|E^{\text{m}}_{[n_0,n+1]}\| > \tfrac{R_{\text{wc}}(n_0)}{2}\right\}. \tag{30}$$

Consider the case $\lambda > 1/2$. Lemma 5.8 shows that

$$\Pr\left\{G_{n_0,n}, \|E^{\text{m}}_{[n_0,n+1]}\| > \tfrac{R_{\text{wc}}(n_0)}{2}\right\} \leq 2d^2 \exp\left[-\frac{n+1}{8d^3C_{m2}^2}\right] \quad .$$

Substituting this in (30) 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 5.8, we get

$$\Pr\left\{G_{n_0,n}, \|E^{\text{m}}_{[n_0,n+1]}\| > \tfrac{R_{\text{wc}}(n_0)}{2}\right\} \leq 2d^2 \exp\left[-\frac{(n_0+1)^{1-2\lambda}(n+1)^{2\lambda}}{8d^3C_{m2}^2}\right] \quad .$$

Let $\ell_{n_0} := (n_0+1)^{1-2\lambda}/8d^3C_{m2}^2$. Observe that

$$\begin{aligned}
&\sum_{n=n_0}^{\infty} \exp[-\ell_{n_0}(n+1)^{2\lambda}] \\
&\leq \sum_{i=\lfloor (n_0+1)^{2\lambda}\rfloor}^{\infty} e^{-i\ell_{n_0}} |\{n : \lfloor (n+1)^{2\lambda}\rfloor = i\}| \\
&\leq \frac{1}{2\lambda} \sum_{i=\lfloor (n_0+1)^{2\lambda}\rfloor}^{\infty} e^{-i\ell_{n_0}} \ (i+1)^{\frac{1-2\lambda}{2\lambda}} \qquad (31)\\
&\leq \frac{1}{2\lambda} \sum_{i=\lfloor (n_0+1)^{2\lambda}\rfloor}^{\infty} e^{-i\ell_{n_0}/2} e^{-i\ell_{n_0}/2} \ (i+1)^{\frac{1-2\lambda}{2\lambda}} \\
&\leq \frac{1}{2\lambda}\left[\frac{(1-2\lambda)}{\ell_{n_0}\lambda}\right]^{\frac{1-2\lambda}{2\lambda}} e^{\frac{1}{2}[\ell_{n_0} - \frac{1-2\lambda}{\lambda}]} \sum_{i=\lfloor (n_0+1)^{2\lambda}\rfloor}^{\infty} e^{-i\ell_{n_0}/2} \qquad (32)\\
&\leq \frac{1}{\ell_{n_0}\lambda}\left[\frac{(1-2\lambda)}{\ell_{n_0}\lambda}\right]^{\frac{1-2\lambda}{2\lambda}} e^{\frac{1}{2}[\ell_{n_0} - \frac{1-2\lambda}{\lambda}]} e^{-\frac{\ell_{n_0} {n_0}^{2\lambda}}{4}} \qquad (33)\\
&\leq \left[\frac{1-2\lambda}{e}\right]^{\frac{1-2\lambda}{2\lambda}} \left[\frac{8d^3C_{m2}^2}{\lambda}\right]^{\frac{1}{2\lambda}} \frac{\exp[-\frac{n_0}{64d^3C_{m2}^2}]}{(n_0+1)^{\frac{1-2\lambda}{2\lambda}}} \qquad (34)\\
&\leq \left[\frac{8d^3C_{m2}^2}{\lambda}\right]^{\frac{1}{2\lambda}} \frac{\exp[-\frac{n_0}{64d^3C_{m2}^2}]}{(n_0+1)^{\frac{1-2\lambda}{2\lambda}}} \quad . \qquad (35)
\end{aligned}$$

The relation (31) follows, as by calculus,

$$|\{n : \lfloor (n+1)^{2\lambda}\rfloor = i\}| \leq \frac{1}{2\lambda}(i+1)^{\frac{1-2\lambda}{2\lambda}} \quad ,$$

(32) holds since, again by calculus,

$$\max_{i\geq 0} e^{-i\ell_{n_0}/2}(i+1)^{\frac{1-2\lambda}{2\lambda}} \leq \left[\frac{(1-2\lambda)}{\ell_{n_0}\lambda}\right]^{\frac{1-2\lambda}{2\lambda}} e^{\frac{1}{2}[\ell_{n_0} - \frac{1-2\lambda}{\lambda}]} \quad ,$$

(33) follows by treating the sum as a right Riemann sum, (34) follows by substituting the value of ℓ_{n_0} and using the fact that $n_0^{2\lambda} \geq 4$ and (35) holds since $1-2\lambda \leq 1$. Substituting (35) in (30), the desired result follows. □

On the second "bad" event $\mathcal{E}^{\text{after}}_{n_0,n_1}$, the TD(0) iterate θ_n for at least one $n > n_0 + n_1$ lies outside the $\min\{\epsilon, 2R_{\text{wc}}(n_0)\}$ radius ball around θ^*. The next result shows that this event also has low probability.

Lemma B.1 (Bound on Probability of $\mathcal{E}^{\text{after}}_{n_0,n_1}$)**.** *Let* $n_0 \geq \max\left\{\frac{K_\lambda 6\|A\|(\|A\|+2K_m)}{\lambda}, 2^{\frac{1}{\lambda}}\right\}$ *and*

$$n_{\text{c}} \geq \left(1 + \frac{K_\lambda 6\|A\|(\|A\|+2K_m)}{\lambda \min\{\epsilon, R_{\text{wc}}(n_0)\}}\right) R_{\text{wc}}(n_0),$$

Let $n_1 \equiv n_1(\epsilon, n_{\text{c}}, n_0) \geq (n_{\text{c}}+1)\left[\frac{6K_\lambda R_{\text{wc}}(n_0)}{\epsilon}\right]^{1/\lambda} - n_0$.

- *If* $\lambda > 1/2$, *then*

$$\Pr\{\mathcal{E}^{\text{after}}_{n_0,n_1}\} \leq 36 d^5 C_{m2}^2 \left[\frac{R_{\text{wc}}(n_0)}{\epsilon}\right]^2 \exp\left[-\frac{(6K_\lambda)^{1/\lambda}}{18 d^3 C_{m2}^2}(n_{\text{c}}+1)\left[\frac{\epsilon}{R_{\text{wc}}(n_0)}\right]^{2-\frac{1}{\lambda}}\right] .$$

- *If* $\lambda < 1/2$, *then*

$$\Pr\{\mathcal{E}^{\text{after}}_{n_0,n_1}\} \leq 2d^2 \left[\frac{18 d^3 C_{m2}^2 [R_{\text{wc}}(n_0)]^2}{\epsilon^2 \lambda}\right]^{\frac{1}{2\lambda}} \exp\left[-\frac{K_\lambda^2}{4 d^3 C_{m2}^2}(n_{\text{c}}+1)\right] .$$

Proof. Assume the event $G_{n_0,n}$ holds for some $n \geq n_{\text{c}}$. Then

$$\|\theta_{n_{\text{c}}} - \theta^*\| \leq 2R_{\text{wc}}(n_0).$$

Hence from (15) and (17), for $t \geq t_{n_{\text{c}}}$, we have

$$\|\theta(t, t_{n_{\text{c}}}, \theta_{n_{\text{c}}}) - \theta^*\| \leq K_\lambda\, e^{-\lambda(t - t_{n_{\text{c}}})} 2R_{\text{wc}}(n_0) \quad . \tag{36}$$

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

$$\|\theta(t_{n+1}, t_{n_{\text{c}}}, \theta_{n_{\text{c}}}) - \theta^*\| \leq \frac{\epsilon}{3} \quad .$$

Also, as $n_{\text{c}} \geq \left(1 + \frac{K_\lambda 6\|A\|(\|A\|+2K_m)C_* R_0}{\lambda \min\{\epsilon, R_{\text{wc}}(n_0)\}}\right)(n_0+1)$, from Lemma 5.7, we have $\|E^{\text{d}}_{[n_{\text{c}},n+1]}\| \leq \epsilon/3$ for all $n \geq n_{\text{c}}$. Combining these with Corollary 5.3, it follows that $\forall n \geq n_0 + n_1$,

$$\begin{aligned}\{G_{n_0,n}, \|\theta_{n+1} - \theta^*\| > \min\{\epsilon, 2R_{\text{wc}}(n_0)\}\} &\subseteq \{G_{n_0,n}, \|\theta_{n+1} - \theta^*\| > \epsilon\} \\ &\subseteq \{G_{n_0,n}, \|E^{\text{m}}_{[n_{\text{c}},n+1]}\| \geq \tfrac{\epsilon}{3}\} \quad .\end{aligned}$$

Hence from the definition of $\mathcal{E}^{\text{after}}_{n_0,n_1}$,

$$\Pr\{\mathcal{E}^{\text{after}}_{n_0,n_1}\} \leq \sum_{n=n_0+n_1}^{\infty} \Pr\left\{G_{n_0,n}, \|E^{\text{m}}_{[n_{\text{c}},n+1]}\| \geq \tfrac{\epsilon}{3}\right\} \quad . \tag{37}$$

Consider the case $\lambda > 1/2$. Lemma 5.8 and the definition of $R_{\text{wc}}(n_0)$ in Theorem 5.4 shows that

$$\Pr\left\{G_{n_0,n}, \|E^{\text{m}}_{[n_{\text{c}},n+1]}\| \geq \tfrac{\epsilon}{3}\right\} \leq 2d^2 \exp\left[-\frac{(n_0+1)^{-2}(n+1)\epsilon^2}{18 d^3 C_{m2}^2 C_*^2 R_0^2}\right] .$$

Using this in (37) and treating the resulting expression as a right Riemann sum, we get

$$\Pr\{\mathcal{E}^{\text{after}}_{n_0,n_1}\} \leq 36 d^5 C_{m2}^2 \left[\frac{R_{\text{wc}}(n_0)}{\epsilon}\right]^2 \exp\left[-\frac{(n_0+n_1)\epsilon^2}{18 d^3 C_{m2}^2 [R_{\text{wc}}(n_0)]^2}\right] .$$

Substituting the given relation between n_1 and n_{c}, the desired result is easy to see.

Consider the case $\lambda < 1/2$. From Lemma 5.8 and the definition of $R_{\text{wc}}(n_0)$ in Theorem 5.4, we have

$$\Pr\left\{G_{n_0,n}, \|E^{m}_{[n_{\text{c}},n+1]}\| \geq \tfrac{\epsilon}{3}\right\} \leq 2d^2 \exp\left[-\frac{(n_{\text{c}}+1)^{1-2\lambda}(n+1)^{2\lambda}\epsilon^2}{18 d^3 C_{m2}^2 [R_{\text{wc}}(n_0)]^2}\right] .$$

Let $k_{n_{\text{c}}} := \epsilon^2 (n_{\text{c}}+1)^{1-2\lambda} / (18 d^3 C_{m2}^2 [R_{\text{wc}}(n_0)]^2)$.

$$\Pr\left\{G_{n_0,n}, \|E^{m}_{[n_{\text{c}},n+1]}\| \geq \tfrac{\epsilon}{3}\right\} \leq 2d^2 \exp\left[-k_{n_{\text{c}}}(n+1)^{2\lambda}\right] .$$

Then by the same technique that we use to obtain (33) in the proof for Lemma 5.9, we have

$$\begin{aligned}
&\sum_{n=n_0+n_1}^{\infty} \exp[-k_{n_{\rm c}}(n+1)^{2\lambda}] \\
&\leq \frac{1}{k_{n_{\rm c}}\lambda}\left[\frac{(1-2\lambda)}{k_{n_{\rm c}}\lambda}\right]^{\frac{1-2\lambda}{2\lambda}} e^{\frac{1}{2}[k_{n_{\rm c}}-\frac{1-2\lambda}{\lambda}]} e^{-\frac{k_{n_{\rm c}}(n_0+n_1)^{2\lambda}}{4}} \\
&\leq \left[\frac{1}{k_{n_{\rm c}}\lambda}\right]^{\frac{1}{2\lambda}} e^{-\frac{k_{n_{\rm c}}(n_0+n_1)^{2\lambda}}{8}} \\
&= \left[\frac{18d^3C_{m2}^2[R_{\rm wc}(n_0)]^2}{\epsilon^2\lambda(n_{\rm c}+1)^{1-2\lambda}}\right]^{\frac{1}{2\lambda}} \exp\left[-\frac{\epsilon^2(n_{\rm c}+1)^{1-2\lambda}(n_0+n_1)^{2\lambda}}{144d^3C_{m2}^2[R_{\rm wc}(n_0)]^2}\right]
\end{aligned}$$

where the second inequality is obtained using the facts that $(n_0+n_1)^{2\lambda} \geq n_0^{2\lambda} \geq 4$ and $1-2\lambda \leq 1$ and the last equality is obtained by substituting the value of $k_{n_{\rm c}}$. From this, after substituting the given relation between n_c and n_1, the desired result is easy to see. □

Detailed Calculations the Proof of Theorem 3.5

We conclude by providing all detailed calculations for our main result, Theorem 3.5.

From Lemma 5.6, by a union bound,

$$\Pr\{\mathcal{E}^c(n_0,n_1)\} \leq \Pr\{\mathcal{E}^{\rm mid}_{n_0,n_1}\} + \Pr\{\mathcal{E}^{\rm after}_{n_0,n_1}\} \ .$$

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

$$N_0(\delta) = \max\left\{\tfrac{K_\lambda 6\|A\|(\|A\|+2K_m)}{\lambda}, 2^{\frac{1}{\lambda}}, 8d^3C_{m2}^2\ln\left[\tfrac{32d^5C_{m2}^2}{\delta}\right]\right\} = O\left(\ln\tfrac{1}{\delta}\right) \ , \tag{38}$$

$$\begin{aligned}
N_c(\epsilon,\delta,n_0) = \max\Bigg\{ &\left\lceil\left(1+\tfrac{K_\lambda 6\|A\|(\|A\|+2K_m)}{\lambda\min\{\epsilon,R_{\rm wc}(n_0)\}}\right)R_{\rm wc}(n_0)\right\rceil, \\
&\frac{18d^3C_{m2}^2}{(6K_\lambda)^{1/\lambda}}\left[\frac{R_{\rm wc}(n_0)}{\epsilon}\right]^{2-\frac{1}{\lambda}}\ln\left[72d^5C_{m2}^2\left[\frac{1}{\delta}\right]\left[\frac{R_{\rm wc}(n_0)}{\epsilon}\right]^2\right]\Bigg\},
\end{aligned}$$

so that $N_c(\epsilon,\delta,N_0(\delta)) = \tilde{O}\left(\max\left\{\frac{1}{\epsilon}\ln\left[\frac{1}{\delta}\right], \left[\frac{1}{\epsilon}\right]^{2-\frac{1}{\lambda}}\left[\ln\frac{1}{\delta}\right]^{3-\frac{1}{\lambda}}\right\}\right)$, and let

$$N_1(\epsilon,n_{\rm c},n_0) = (n_{\rm c}+1)\left[\frac{6K_\lambda R_{\rm wc}(n_0)}{\epsilon}\right]^{1/\lambda} - n_0,$$

so that

$$N_1(\epsilon,N_c(\epsilon,\delta,N_0(\delta)),N_0(\delta)) = \tilde{O}\left(\max\left\{\left[\frac{1}{\epsilon}\right]^{1+\frac{1}{\lambda}}\left[\ln\frac{1}{\delta}\right]^{1+\frac{1}{\lambda}}, \left[\frac{1}{\epsilon}\right]^2\left[\ln\frac{1}{\delta}\right]^3\right\}\right). \tag{39}$$

Let $n_0 \geq N_0(\delta)$, $n_{\rm c} \geq N_c(\epsilon,\delta,n_0)$ and $n_1 \geq N_1(\epsilon,n_{\rm c},n_0)$. Then from Lemma 5.9, $\Pr\{\mathcal{E}^{\rm mid}_{n_0,n_1}\} \leq \delta/2$ and from Lemma B.1, $\Pr\{\mathcal{E}^{\rm after}_{n_0,n_1}\} \leq \delta/2$. Hence $\Pr\{\mathcal{E}^c(n_0,n_1)\} \leq \delta$. 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.

$$N_0(\delta) = \max\left\{\tfrac{K_\lambda 6\|A\|(\|A\|+2K_m)}{\lambda}, 2^{\frac{1}{\lambda}}, \tfrac{64d^3C_{m2}^2}{2\lambda}\ln\left(\tfrac{32d^5C_{m2}^2}{\delta\lambda}\right)\right\} = O\left(\ln\tfrac{1}{\delta}\right) \ , \tag{40}$$

$$\begin{aligned}
N_c(\epsilon,\delta,n_0) = \max\Bigg\{ &\left\lceil\left(1+\tfrac{K_\lambda 6\|A\|(\|A\|+2K_m)}{\lambda\min\{\epsilon,R_{\rm wc}(n_0)\}}\right)R_{\rm wc}(n_0)\right\rceil, \\
&\frac{4d^3C_{m2}^2}{2\lambda K_\lambda^2}\ln\left(\frac{72d^5C_{m2}^2}{\lambda}\left[\frac{1}{\delta}\right]\frac{[R_{\rm wc}(n_0)]^2}{\epsilon^2}\right)\Bigg\},
\end{aligned}$$

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

$$N_1(\epsilon, n_c, n_0) = (n_c + 1) \left[\frac{6K_\lambda \, R_{wc}(n_0)}{\epsilon}\right]^{1/\lambda} - n_0, \tag{41}$$

so that $N_1(\epsilon, N_c(\epsilon, \delta, N_0(\delta)), N_0(\delta)) = \tilde{O}\left(\left[(1/\epsilon) \ln (1/\delta)\right]^{1+1/\lambda}\right)$. 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 $\tilde{O}$ results as in (40) and (41).