A Tale of Two-Timescale Reinforcement Learning with the Tightest Finite-Time Bound

Gal Dalal,[1] Balázs Szörényi,[2] Gugan Thoppe[3]*
[1]Technion, Israel Institute of Technology, Haifa, Israel; gald@technion.ac.il
[2]Yahoo! Research, New York, NY, USA; szorenyi.balazs@gmail.com
[3]Duke University, Durham, NC, USA; gugan.thoppe@gmail.com

Abstract

Policy evaluation in reinforcement learning is often conducted using two-timescale stochastic approximation, which results in various gradient temporal difference methods such as GTD(0), GTD2, and TDC. Here, we provide convergence rate bounds for this suite of algorithms. Algorithms such as these have two iterates, θ_n and w_n, which are updated using two distinct stepsize sequences, α_n and β_n, respectively. Assuming $\alpha_n = n^{-\alpha}$ and $\beta_n = n^{-\beta}$ with $1 > \alpha > \beta > 0$, we show that, with high probability, the two iterates converge to their respective solutions θ^* and w^* at rates given by $\|\theta_n - \theta^*\| = \tilde{O}(n^{-\alpha/2})$ and $\|w_n - w^*\| = \tilde{O}(n^{-\beta/2})$; here, $\tilde{O}$ hides logarithmic terms. Via comparable lower bounds, we show that these bounds are, in fact, tight. To the best of our knowledge, ours is the first finite-time analysis which achieves these rates. While it was known that the two timescale components decouple asymptotically, our results depict this phenomenon more explicitly by showing that it in fact happens from some finite time onwards. Lastly, compared to existing works, our result applies to a broader family of stepsizes, including non-square summable ones.

1 Introduction

Stochastic Approximation (SA) (Kushner and Yin 1997) is the name given to algorithms useful for finding optimal points or zeros of a function for which only noisy access is available. This makes SA theory vital to machine learning and, specifically, to Reinforcement Learning (RL). Here, we obtain tight convergence rate estimates for the special class of linear two-timescale SA, which involves two interleaved update rules with distinct stepsize sequences. In the context of RL, the analysis here applies to *policy evaluation* schemes with function approximation.

A generic linear two-timescale SA has the form:

$$\theta_{n+1} = \theta_n + \alpha_n[h_1(\theta_n, w_n) + M_{n+1}^{(1)}] \ , \quad (1)$$

$$w_{n+1} = w_n + \beta_n[h_2(\theta_n, w_n) + M_{n+1}^{(2)}] \ , \quad (2)$$

where $\alpha_n, \beta_n \in \mathbb{R}$ are stepsizes and $M_n^{(i)} \in \mathbb{R}^d$ denotes noise. Further, $h_i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ has the form

$$h_i(\theta, w) = v_i - \Gamma_i\theta - W_i w \quad (3)$$

for a vector $v_i \in \mathbb{R}^d$ and matrices $\Gamma_i, W_i \in \mathbb{R}^{d\times d}$.

Within RL, this class of algorithms mainly concerns the suite of gradient Temporal Difference (TD) methods, which was introduced in (Sutton, Maei, and Szepesvári 2009) and has gradually gained increasing attention since then. That work presented a gradient descent variant of TD(0), called GTD(0). As it supports off-policy learning, GTD(0) is advantageous over TD(0). More recently, additional variants were introduced such as GTD2 and TDC (Sutton et al. 2009); while being better than TD(0), these are also faster than GTD(0). The above gradient TD methods have been shown to converge asymptotically in the case of linear and non-linear function approximation (Sutton, Maei, and Szepesvári 2009; Sutton et al. 2009; Bhatnagar et al. 2009). Separately, there are also a few convergence rate results for altered versions of the GTD family (Liu et al. 2015) and sparsely-projected variants (Dalal et al. 2018b). Both works apply projections to keep the iterates in a confined region around the solutions. However, in (Liu et al. 2015), the learning rates are set to a fixed ratio which makes the altered algorithms single-timescale variants of the original ones.

To place our work in the landscape of the existing literature on generic two-timescale SA, we now briefly review a few seminal papers. The first well-known use of the two-timescale idea is the Polyak-Ruppert averaging scheme (Ruppert 1988; Polyak 1990). There, iterate averaging is used to improve the convergence rate of a one-timescale algorithm, which is especially beneficial when the driving matrices have poor conditioning. The general two-timecale SA scheme is formulated in (Borkar 1997); this work provided conditions for convergence. Since then, relatively little work has been published on the topic; the main results obtained so far include weak convergence and asymptotic convergence rates (Gerencsér 1997; Konda and Tsitsiklis 2004; Mokkadem and Pelletier 2006), and stability (Lakshminarayanan and Bhatnagar 2017).

We now discuss two specific papers from the above list that are the closest to our work. Denote by θ^* and w^* the respective solutions of (1) and (2); i.e., $h_1(\theta^*, w^*) =$

*Research supported by NSF grants DEB-1840223 and DMS 17-13012.

$h_2(\theta^*, w^*) = 0$. In (Konda and Tsitsiklis 2004), it was shown that both, $(\theta_n - \theta^*)/\sqrt{\alpha_n}$ and $(w_n - w^*)/\sqrt{\beta_n}$, are asymptotically normal. This result surprisingly tells us that eventually the two components do not influence each other's convergence rates. However, one of the assumptions there is that the noise sequence is independent of its past values, and their variance-covariance matrices are constant across the iterations. This make their results inapplicable to the RL methods of our interest. In (Mokkadem and Pelletier 2006), a similar weak convergence result has been derived in the context of nonlinear SA under the assumptions that the stepsizes are square summable. This result also explicitly establishes asymptotic independence (see (5) there) between the two components. A separate result in this last work is that of almost-sure asymptotic convergence rate. The issue with this last result is that it cannot be used to obtain explicit form for the constants. In fact, by its very nature, the constants involved depend on the sample paths.

In this work, we revisit the convergence rate question for two-timescale RL methods with a focus on finite-time behaviour. In order to highlight the merits of this work over existing literature, we first classify common types of convergence results. The first class is of asymptotic convergence, which is beneficial for the rudimentary verification that an algorithm converges after an infinite amount of time. The second class is asymptotic convergence rates; these are stronger in the sense of telling us that an algorithm would asymptotically converge at a certain rate, but again they have little practical implications; even given exact knowledge of all parameters of the problem, with these results one cannot numerically compute a bound on the distance from the solution with a corresponding numerical probability value. The third class, to which the results in this work belong, are finite time bounds. These contain explicit constants — both controllable such as stepsize parameters and uncontrollable such as eigenvalues — as well as finite-time rates, thereby revealing intriguing dependencies among such parameters that crucially affect convergence rates (e.g., $1/q_i$; see (Dalal, Thoppe, and Szörényi 2019)[Table 3]). Moreover, the constants are trajectory-independent and thus can be of help in obtaining stopping time theorems. We consider this a significant step forward in obtaining practical results that would enable to assuredly adapt algorithm parameters so as to maximize their efficiency.

Our Contributions In (Dalal et al. 2018b), the first finite time bound for the GTD family was proved. Here, we significantly strengthen it and, in fact, obtain a tight rate. Specifically, our key result (Theorem 3) is that the iterates θ'_n and w'_n, obtained by sparsely projecting θ_n and w_n, respectively, satisfy $\|\theta'_n - \theta^*\| = \tilde{O}(n^{-\alpha/2})$ and $\|w'_n - w^*\| = \tilde{O}(n^{-\beta/2})$ with high probability. Here, $\tilde{O}$ hides logarithmic terms and α and β originate in the stepsize choice $\alpha_n = n^{-\alpha}$ and $\beta_n = n^{-\beta}$ with $1 > \alpha > \beta > 0$. We establish the tightness of this upper bound by deriving a matching lower bound.

We emphasize that we have explicit formulas for the constants hidden in these order notations and also bounds on the iteration index from where these rates apply. In particular, our bound shows how the convergence rate of a given GTD method depends on the parameters of the MDP itself; e.g., the eigenvalues of the driving matrix.

As in (Dalal et al. 2018b) which dealt with single-timescale algorithms, the bounds in this work are applicable for both square-summable and non-square-summable stepsizes. This was indeed also the case in (Konda and Tsitsiklis 2004); however, as pointed earlier, the noise assumptions there are significantly stronger than ours.

The sparse projection scheme used here is novel but is similar in spirit to the one used in (Dalal et al. 2018b). There, the iterates were only projected when the iteration indices were powers of 2, whereas here we project whenever the iteration index is of the form $k^k = 2^{k \log_2 k}$, $k \geq 0$. The motivation for using projections is to keep the iterates bounded. However, projections also modify the original algorithm by introducing non-linearity. This highly complicates the analysis. Evidently, the literature almost doesn't contain analyses of projected algorithms at all. Moreover, projections are often empirically found to be unnecessary. The advantages of using a sparse projection scheme is that we effectively almost never project and, more importantly, it makes the analysis oblivious to its non-linearity.

An additional novelty of this paper is its proof technique. At its heart lie two induction tricks–one inspired from (Thoppe and Borkar 2019) and the other, being rather non-standard, from (Mokkadem and Pelletier 2006). The first induction is on the iteration index n; together with projections it enables us to show that both θ'_n and w'_n iterates are $O(1)$, i.e., bounded, with high probability. On each sample path where the iterates are bounded, we then use the second induction to show that the convergence rate of the w'_n iterates can be improved from $\tilde{O}(n^{-\beta/2}\mathbb{1}[\ell \neq 0] + n^{-(\alpha-\beta)\ell})$ to $\tilde{O}(n^{-\beta/2}\mathbb{1}[\ell \neq 0] + n^{-(\alpha-\beta)(\ell+1)})$ for all suitable ℓ. In particular, we use this to show that the bound on the behaviour of w'_n iterates can be incrementally improved from $O(1)$, established above, to the desired $\tilde{O}(n^{-\beta/2})$. Finally, we use this latter result to show that $\|\theta'_n - \theta^*\| = \tilde{O}(n^{-\alpha/2})$.

We end this section by describing the key insights that our main result in Theorem 3 provides.

Decoupling after Finite Time: Even though both θ'_n and w'_n influence each other, our result shows that, from some finite time onwards, their convergence rates do not depend on β and α, respectively. While from the results in (Konda and Tsitsiklis 2004) and (Mokkadem and Pelletier 2006), one would expect the two-timescale components to indeed decouple asymptotically, our result shows that this in fact happens from some finite time that can conceptually be numerically evaluated. All of this is in sharp contrast to the former state-of-the-art finite-time result given in (Dalal et al. 2018b) which showed that the convergence rate is $\tilde{O}(n^{-\min\{\alpha-\beta, \beta/2\}})$.

One vs Two-Timescale: A natural question for an RL practitioner is whether to run the algorithm given in (1) and (2) in the one-timescale mode, i.e., with α_n/β_n being constant, or in the two-timescale mode, i.e., with $\alpha_n/\beta_n \to 0$. Judging solely on the convergence rate order – based on this work and on single-timescale results from, e.g., (Liu et al.

2015), the answer[1] is to pick the single timescale mode with $\alpha_n = \beta_n \approx 1/n$. This then brings forth an imperative question for future work: "what indeed are the provable benefits of two-timescale RL methods?" A comparison to recent gradient descent literature suggests that this question can be better answered via iteration complexity, i.e., the the number of iterations required to hit some $\epsilon-$ball around the solution. In particular, we believe the eigenvalues of the driving matrices — hiding in the constants — can have dramatic influence on the actual rate. A predominant recent example is how the heavy-ball method, which is similar in nature to a two-timescale algorithm, has an $O(\sqrt{\kappa}\ln(1/\epsilon))$ iteration complexity as compared to the usual stochastic gradient descent which has $O(\kappa\ln(1/\epsilon))$ (Loizou and Richtárik 2017); here, κ is the condition number. Thus, we believe that finite-time analyses of two-timescale methods are crucial for understanding their potential merits over one-single variants.

2 Main Result

We state our main convergence rate result here. It applies to the iterates θ'_n and w'_n which are obtained by sparsely-projecting θ_n and w_n from (1) and (2). We begin by stating our assumptions and defining the projection operator.

$\boldsymbol{\mathcal{A}_1}$ (Matrix Assumptions). W_2 and $X_1 = \Gamma_1 - W_1 W_2^{-1}\Gamma_2$ are positive definite (not necessarily symmetric).

$\boldsymbol{\mathcal{A}_2}$ (Stepsize Assumption). $\alpha_n = (n+1)^{-\alpha}$ and $\beta_n = (n+1)^{-\beta}$, where $1 > \alpha > \beta > 0$.

Definition 1 (Noise Condition). *$\{M_n^{(1)}\}$ and $\{M_n^{(2)}\}$ are said to be (θ_n, w_n)-*dominated martingale differences with parameters m_1 and m_2, *if they are martingale difference sequences w.r.t. the family of $\sigma-$fields $\{\mathcal{F}_n\}$, where $\mathcal{F}_n = \sigma(\theta_0, w_0, M_1^{(1)}, M_1^{(2)}, \ldots, M_n^{(1)}, M_n^{(2)})$, and*

$$\begin{aligned}\left\|M_{n+1}^{(1)}\right\| &\leq m_1(1 + \|\theta_n\| + \|w_n\|)\\ \left\|M_{n+1}^{(2)}\right\| &\leq m_2(1 + \|\theta_n\| + \|w_n\|)\end{aligned}$$

for all $n \geq 0$.

Definition 2 (Sparse Projection). *For $R > 0$, let $\Pi_R(x) = \min\{1, R/\|x\|\}\cdot x$ be the projection into the ball with radius R around the origin. The* sparse projection *operator*

$$\Pi_{n,R} = \begin{cases}\Pi_R, & \text{if } n = k^k - 1 \text{ for some } k \in \mathbb{Z}_{>0},\\ I, & \text{otherwise}.\end{cases} \quad (4)$$

We call it sparse as it projects only on specific indices that are exponentially far apart.

Pick an arbitrary $p > 1$. Fix some constants $R^{\theta}_{\text{proj}} > 0$ and $R^{w}_{\text{proj}} > 0$ for the radius of the projection balls. Further, let

$$\theta^* = X_1^{-1}b_1, \quad w^* = W_2^{-1}(v_2 - \Gamma_2\theta^*)$$

with $b_1 = v_1 - W_1W_2^{-1}v_2$. Using (Borkar 2009) and (Lakshminarayanan and Bhatnagar 2017), it can be shown that $(\theta_n, w_n) \to (\theta^*, w^*)$ a.s.

[1]The $\alpha_n = \beta_n = 1/n$ case above would bring the condition number of the driving matrices into picture (Dalal et al. 2018a). To overcome this, one could use Polyak-Ruppert iterate averaging for two-timescale SA (Mokkadem and Pelletier 2006).

Constant	Definition
N_3	$\min\{n \geq N_3' : n = k^k - 1 \text{ for some integer } k\}$
N_3'	$\max\{N_6, K_{22,a}, K_{22,b}, K_{3,w}, K_{3,\theta}, e^{1/\beta}, (2/\beta)^{2/\beta}\}$
N_6	$\max\{N_7, N_8\}$
N_7	$\max\{K_{15,\alpha}, K_{15,\beta}, K_{20,\alpha}(0), K_{21,\beta}, K_9, (p-1)^{-1/(p-1)}\}$
N_8	$\max\{K_{20,\alpha}(\beta/2), k_\beta(\beta/2), e^{1/\beta}\delta^{1/p}/(4d^2)^{1/p}, K_{30,a}, K_{30,b}, K_{35,a}, K_{35,b}\} + 1$

Table 1: A summary of all n_0 lower bounds

Theorem 3 (Main Result). *Assume $\boldsymbol{\mathcal{A}_1}$ and $\boldsymbol{\mathcal{A}_2}$. Let $\theta'_0, w'_0 \in \mathbb{R}^d$ be arbitrary. Consider the update rules*

$$\theta'_{n+1} = \Pi_{n+1,R^{\theta}_{proj}}\Big(\theta'_n + \alpha_n[h_1(\theta'_n, w'_n) + M_{n+1}^{(1')}]\Big), \quad (5)$$

$$w'_{n+1} = \Pi_{n+1,R^{w}_{proj}}\Big(w'_n + \beta_n[h_2(\theta'_n, w'_n) + M_{n+1}^{(2')}]\Big), \quad (6)$$

where $\{M_n^{(1')}\}$ and $\{M_n^{(2')}\}$ are (θ'_n, w'_n)-dominated martingale differences with parameters m_1 and m_2 (see Def. 1). Then, with probability larger than $1-\delta$, for all $n \geq N_3$

$$\|\theta'_n - \theta^*\| \leq C_{3,\theta}\frac{\sqrt{\ln(4d^2(n+1)^p/\delta)}}{(n+1)^{\alpha/2}} \quad (7)$$

$$\|w'_n - w^*\| \leq C_{3,w}\frac{\sqrt{\ln(4d^2(n+1)^p/\delta)}}{(n+1)^{\beta/2}}. \quad (8)$$

Refer to Table 1 and (Dalal, Thoppe, and Szörényi 2019)[Table 3] for the constants.

Comments on Main Result

1. Our analysis goes through even if $\theta_n \in \mathbb{R}^{d_1}, w_n \in \mathbb{R}^{d_2}$ with $d_1 \neq d_2$. For brevity, we work with $d_1 = d_2 = d$.
2. The constants in the above result equal infinity when $\alpha = \beta$. This is because the algorithm then ceases to be two-timescale, thereby making our analysis invalid.

2.1 Tightness

Here, we accompany our upper bound by a lower bound. This bound is asymptotic and holds for unprojected algorithms. Nonetheless, a coupling argument as in the proof of Theorem 3 can be used to obtain a similar bound for projected ones. We thus establish the tightness (up to logarithmic terms) of the result in Theorem 3.

Proposition 4 (Lower Bound). *Assume $\boldsymbol{\mathcal{A}_1}$ and $\boldsymbol{\mathcal{A}_2}$. Consider* (1) *and* (2) *with $\{M_n^{(1)}\}$ and $\{M_n^{(2)}\}$ being (θ_n, w_n)-dominated martingale differences (see Def. 1). Then, there exists an algorithm for which*

$$\|\theta_n - \theta^*\| = \Omega_p(n^{-\alpha/2}) \quad \text{and} \quad \|w_n - w^*\| = \Omega_p(n^{-\beta/2}),$$

where $X_n = \Omega_p(\gamma_n)$ means that for any $\epsilon > 0$, there are constants c and K so that $\mathbb{P}\{|X_n|/\gamma_n < c\} \leq \epsilon$, $\forall n \geq K$.

Proof. See (Dalal, Thoppe, and Szörényi 2019)[Appendix A]. □

3 Applications to Reinforcement Learning

Here, we apply our results on the general linear two-timescale setup to the specific RL use case. Namely, we apply Theorem 3 to derive the tightest existing finite sample bound for the GTD family. This section relies on a similar procedure as in Section 5, (Dalal et al. 2018b). Nonetheless, we reiterate it here for completeness.

3.1 Background

A Markov Decision Processes (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (Sutton 1988), where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the transition kernel, R is the reward function, and γ the discount factor. A policy $\pi : \mathcal{S} \to \mathcal{A}$ is a stationary mapping from states to actions and $V^\pi(s) = \mathbb{E}^\pi[\sum_{n=0}^{\infty} \gamma^n r_n | s_0 = s]$ is the value function at state s w.r.t π.

As mentioned above, our results apply to GTD, which is a suite of policy evaluation algorithms. These algorithms are used to estimate the value function $V^\pi(s)$ with respect to a given π 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 parameter vector. For brevity, we omit the notation π and denote $\phi(s_n),\ \phi(s'_n)$ by $\phi_n,\ \phi'_n$. Finally, let $\delta_n = r_n + \gamma\theta_n^\top \phi'_n - \theta_n^\top \phi_n$, $A = \mathbb{E}[\phi(\phi - \gamma\phi')^\top]$, $C = \mathbb{E}[\phi\phi^\top]$, and $b = \mathbb{E}[r\phi]$, where the expectations are w.r.t. the stationary distribution of the induced chain [2] .

We assume all rewards $r(s)$ and feature vectors $\phi(s)$ are bounded: $|r(s)| \leq 1, \|\phi(s)\| \leq 1\ \forall s \in S$. Also, it is assumed that the feature matrix Φ is full rank, so A and C are full rank. This assumption is standard (Maei et al. 2010; Sutton, Maei, and Szepesvári 2009). Therefore, due to its structure, A is also positive definite (Bertsekas 2012). Moreover, by construction, C is positive semi-definite; thus, by the full-rank assumption, it is actually positive definite.

3.2 The GTD(0) Algorithm

First introduced in (Sutton, Maei, and Szepesvári 2009), GTD(0) is designed to minimize the objective function $J^{\mathrm{NEU}}(\theta) = \frac{1}{2}(b - A\theta)^\top (b - A\theta)$. Its update rule is

$$\begin{aligned}\theta_{n+1} =& \theta_n + \alpha_n (\phi_n - \gamma\phi'_n)\, \phi_n^\top w_n,\\ w_{n+1} =& w_n + \beta_n r_n \phi_n + \phi_n[\gamma\phi'_n - \phi_n]^\top \theta_n.\end{aligned}$$

It thus takes the form of (1) and (2) with $h_1(\theta, w) = A^\top w$, $h_2(\theta, w) = b - A\theta - w$, $M_{n+1}^{(1)} = (\phi_n - \gamma\phi'_n)\, \phi_n^\top w_n - A^\top w_n$, $M_{n+1}^{(2)} = r_n\phi_n + \phi_n[\gamma\phi'_n - \phi_n]^\top \theta_n - (b - A\theta_n)$. That is, in case of GTD(0), the relevant matrices in the update rules are $\Gamma_1 = 0$, $W_1 = -A^\top$, $v_1 = 0$, and $\Gamma_2 = A$, $W_2 = I$, $v_2 = b$. Additionally, $X_1 = \Gamma_1 - W_1 W_2^{-1} \Gamma_2 = A^\top A$. By our assumption above, both W_2 and X_1 are symmetric positive definite matrices, and thus the real parts of their eigenvalues are also positive. Also, $\|M_{n+1}^{(1)}\| \leq (1 + \gamma + \|A\|)\|w_n\|$, $\|M_{n+1}^{(2)}\| \leq 1 + \|b\| + (1 + \gamma + \|A\|)\|\theta_n\|$. Hence, the noise condition in Defn. 1 is satisfied with constants $m_1 = (1 + \gamma + \|A\|)$ and $m_2 = 1 + \max(\|b\|, \gamma + \|A\|)$.

We can now apply Theorem 3 to get the following result.

Corollary 5. *Consider the Sparsely Projected variant of GTD(0) as in* (5) *and* (6)*. Then, for* $\alpha_n = 1/(n+1)^\alpha$, $\beta_n = 1/(n+1)^\beta$, *with probability larger than* $1 - \delta$, *for all* $n \geq N_3$, *we have*

$$\|\theta'_n - \theta^*\| \leq C_{3,\theta} \frac{\sqrt{\ln(4d^2(n+1)^p/\delta)}}{(n+1)^{\alpha/2}} \tag{9}$$

$$\|w'_n - w^*\| \leq C_{3,w} \frac{\sqrt{\ln(4d^2(n+1)^p/\delta)}}{(n+1)^{\beta/2}}. \tag{10}$$

For GTD2 and TDC (Sutton et al. 2009), the above result can be similarly reproduced. The detailed derivation and relevant constants are provided in (Dalal, Thoppe, and Szörényi 2019)[Appendix K].

4 Outline of Proof of the Main Result

Here, we first state an intermediary result in Thereom 6 and using that we sketch a proof of Theorem 3. The full proof is in (Dalal, Thoppe, and Szörényi 2019)[Appendix C].

Assume $\boldsymbol{\mathcal{A}_1}$ and $\boldsymbol{\mathcal{A}_2}$. Consider (1) and (2) with $\{M_n^{(1)}\}$ and $\{M_n^{(2)}\}$ being (θ_n, w_n)-dominated martingale differences with parameters m_1 and m_2 (see Def. 1). Let $\mathcal{G}'_{n_0}$ be the event given by

$$\mathcal{G}'_{n_0} = \{\|\theta_{n_0} - \theta^*\| \leq R^\theta_{\text{proj}}, \|w_{n_0} - w^*\| \leq R^w_{\text{proj}}\}$$

and let $\nu(n; \gamma) = (n+1)^{-\gamma/2}\sqrt{\ln(4d^2(n+1)^p/\delta)}$.

Theorem 6. *Let* $\delta \in (0, 1)$. *Suppose that* $n_0 \geq N_6$ *and that the event* $\mathcal{G}'_{n_0}$ *holds. Then, with probability larger than* $1 - \delta$,

$$\|\theta_n - \theta^*\| \leq A_{5,n_0}\, \nu(n, \alpha) \tag{11}$$

$$\|w_n - w^*\| \leq A_{4,n_0}\, \nu(n, \beta) \tag{12}$$

for all $n \geq n_0$.

Sketch of Proof for Theorem 3. Our idea is to use a coupling argument to show that the projected iterates, given in (5) and (6), and the unprojected iterates, given in (1) and (2), are identically distributed from some time on. This then allows us to use Theorem 6 to conclude Theorem 3.

The key steps in our argument are as follows.

1. First we note that, for the projected algorithm, the event $\mathcal{G}'_{n_0}$ holds whenever n_0 is of the form $k^k - 1$.
2. Further, recalling (4), we observe that, for any $k \geq 0$, between projection steps $k^k - 1$ and $(k+1)^{k+1} - 1$, the projected iterates $\{\theta'_n, w'_n\}$ behave exactly as the unprojected iterates $\{\theta_n, w_n\}$ that are initiated at $(\theta'_{k^k-1}, w'_{k^k-1})$.
3. It then follows from Theorem 6 that if k is large enough so that $n_0 = k^k - 1 \geq N_6$, then (11) and (12) apply to $\{\theta'_n, w'_n\}$ for $k^k - 1 \leq n < (k+1)^{k+1} - 1$.
4. In fact, if k is enlarged a bit more so that $n_0 = k^k - 1 \geq N_3 \geq N_6$, then not only does the above claim hold, it is also true that the RHSs in (11) and (12) are less than R^θ_{proj} and R^w_{proj}, respectively, for $n \geq (k+1)^{k+1} - 1$.

[2] Here, the samples $\{(\phi_n, \phi'_n)\}$ are drawn iid. This assumption is standard when dealing with convergence bounds in RL (Liu et al. 2015; Sutton, Maei, and Szepesvári 2009; Sutton et al. 2009).

5. In turn, the latter implies that the projected iterates and unprojected iterates, starting from (θ'_n, w'_n), behave exactly the same $\forall n \geq k^k - 1$. Consequently, (11) and (12) hold for the projected iterates $\forall n \geq N_3$. Substituting $n_0 = N_3$ then establishes Theorem 3.

See (Dalal, Thoppe, and Szörényi 2019)[Appendix C] for the actual proof. □

Next, we discuss the proof of Theorem 6; note that this result only concerns the unprojected iterates. First, we introduce some further notations.

Fix any $p > 1$ and let $\mathcal{U}(n_0)$ be the event given by

$$\mathcal{U}(n_0) := \bigcap_{n \geq n_0} \big\{\|\theta_n - \theta^*\| \leq C_R^\theta R_{\text{proj}}^\theta, \|L_{n+1}^{(\theta)}\| \leq \epsilon_n^{(\theta)}, \|w_n - w^*\| \leq C_R^w R_{\text{proj}}^w, \|L_{n+1}^{(w)}\| \leq \epsilon_n^{(w)}\big\}, \quad (13)$$

where

$$\epsilon_n^{(\theta)} = \sqrt{d^3 L_\theta C_{14,\theta}}\, \nu(n, \alpha), \quad (14)$$

$$\epsilon_n^{(w)} = \sqrt{d^3 L_w C_{14,w}}\, \nu(n, \beta). \quad (15)$$

Further, let $L_{n+1}^{(\theta)}$ and $L_{n+1}^{(w)}$ be appropriate aggregates of the martingale noise terms given by

$$L_{n+1}^{(w)} = \sum_{k=n_0}^{n} \left[\prod_{j=k+1}^{n} [I - \beta_j W_2]\right] \beta_k M_{k+1}^{(2)}, \quad (16)$$

$$L_{n+1}^{(\theta)} = \sum_{k=n_0}^{n} \left[\prod_{j=k+1}^{n} [I - \alpha_j X_1]\right] \alpha_k \times \left[-W_1 W_2^{-1} M_{k+1}^{(2)} + M_{k+1}^{(1)}\right]. \quad (17)$$

For the definition of the constants above, see (Dalal, Thoppe, and Szörényi 2019)[Table 3].

As a first step in proving Theorem 6, we show that the co-occurrence of the events $\mathcal{G}'_{n_0}$ and $\mathcal{U}(n_0)$ has small probability if n_0 is large enough. The proof, inspired from (Thoppe and Borkar 2019), uses induction on the iteration index n. Specifically, we show that if, at time n, the iterates are bounded and the aggregate noise is well-behaved (respectively bounded by $\epsilon_n^{(\theta)}$ and $\epsilon_n^{(w)}$), then the iterates continue to remain bounded at time $n+1$ as well w.h.p.

Theorem 7. *Let $\delta \in (0,1)$ and $n_0 \geq N_7$. Then,*

$$\mathbb{P}\{\mathcal{U}^c(n_0) | \mathcal{G}'_{n_0}\} \leq \delta.$$

Next, we show that, on the event $\mathcal{U}(n_0)$, the convergence rates of $\{\theta_n\}$ and $\{w_n\}$ are $\tilde{O}(n^{-\alpha/2})$ and $\tilde{O}(n^{-\beta/2})$, respectively. The proof proceeds as follows. By refining an induction trick from (Mokkadem and Pelletier 2006), we first show that the convergence rate estimate for the $\{w_n\}$ iterates can be improved from $O(1)$ to $\tilde{O}(n^{-\beta/2})$. Using this, we then show that $\|\theta_n - \theta^*\| = \tilde{O}(n^{-\alpha/2})$. We emphasize that these results are deterministic.

Theorem 8. *Let $n_0 \geq N_8$. Then,*

$$\mathcal{U}(n_0) \subseteq \{\|w_n - w^*\| \leq A_{4,n_0}\, \nu(n;\beta), \forall n \geq n_0\} \quad (18)$$

and

$$\mathcal{U}(n_0) \subseteq \{\|\theta_n - \theta^*\| \leq A_{5,n_0}\, \nu(n;\alpha), \forall n \geq n_0\}. \quad (19)$$

Proof of Theorem 6. Theorems 7 and 8 together establish Theorem 6. □

The next two subsections highlight the key steps in the proofs of these last two results.

4.1 Proof of Theorem 7

Let $C_R^\theta = 3$ and $C_R^w = 3/2 + (e^{q_2}/q_2 \|\Gamma_2\| C_{16,w}) C_R^\theta \frac{R_{\text{proj}}^\theta}{R_{\text{proj}}^w}$. Further, let $\mathcal{G}_n, \mathcal{L}_n$, and $\mathcal{A}_n$ be the events given by

$$\mathcal{G}_n = \bigcap_{k=n_0}^{n} \{\|\theta_k - \theta^*\| \leq C_R^\theta R_{\text{proj}}^\theta, \|w_k - w^*\| \leq C_R^w R_{\text{proj}}^w\}, \quad (20)$$

$$\mathcal{L}_n = \bigcap_{k=n_0}^{n} \left\{\|L_{k+1}^{(\theta)}\| \leq \epsilon_k^{(\theta)}, \|L_{k+1}^{(w)}\| \leq \epsilon_k^{(w)}\right\}, \quad (21)$$

and $\mathcal{A}_n = \mathcal{G}_n \cap \mathcal{L}_n$. Using (13), note that $\mathcal{U}(n_0) = \lim_{n\to\infty} \mathcal{A}_n = \bigcap_{n \geq n_0} \mathcal{A}_n$. Lastly, define

$$\mathcal{Z}_n = \{\|\theta_n - \theta^*\| \leq C_R^\theta R_{\text{proj}}^\theta, \|w_n - w^*\| \leq C_R^w R_{\text{proj}}^w, \|L_{n+1}^{(\theta)}\| \leq \epsilon_n^{(\theta)}, \|L_{n+1}^{(w)}\| \leq \epsilon_n^{(w)}\}. \quad (22)$$

Proof of Theorem 7. By adopting ideas from (Thoppe and Borkar 2019), we first decompose the event $\mathcal{G}'_{n_0} \cap \mathcal{U}^c_{n_0}$. From (149) - (161) in (Dalal, Thoppe, and Szörényi 2019), we have

$$\mathcal{G}'_{n_0} \cap \mathcal{U}^c(n_0) = (\mathcal{G}'_{n_0} \cap \mathcal{Z}^c_{n_0}) \cup (\mathcal{G}'_{n_0} \cap \mathcal{A}_{n_0} \cap \mathcal{Z}^c_{n_0+1}) \cup (\mathcal{G}'_{n_0} \cap \mathcal{A}_{n_0+1} \cap \mathcal{Z}^c_{n_0+2}) \cup \ldots, \quad (23)$$

$$\mathcal{G}'_{n_0} \cap \mathcal{Z}^c_{n_0} \subseteq \mathcal{G}_{n_0} \cap \big(\{\|L^{(\theta)}_{n_0+1}\| > \epsilon^{(\theta)}_{n_0}\} \cup \{\|L^{(w)}_{n_0+1}\| > \epsilon^{(w)}_{n_0}\}\big), \quad (24)$$

and

$$\mathcal{G}'_{n_0} \cap \mathcal{A}_n \cap \mathcal{Z}^c_{n+1} \quad (25)$$

$$\subseteq \mathcal{G}'_{n_0} \cap \mathcal{A}_n \cap \Big[\{\|\theta_{n+1} - \theta^*\| > C_R^\theta R_{\text{proj}}^\theta\} \cup \{\|w_{n+1} - w^*\| > C_R^w R_{\text{proj}}^w\}\Big] \quad (26)$$

$$\cup \Big(\mathcal{G}_{n+1} \cap \big[\{\|L^{(\theta)}_{n+2}\| > \epsilon^{(\theta)}_{n+1}\} \cup \{\|L^{(w)}_{n+2}\| > \epsilon^{(w)}_{n+1}\}\big]\Big). \quad (27)$$

With regards to (27), we also have the following fact.

Lemma 9. *Let $n \geq n_0 \geq \max\{K_{15,\alpha}, K_{15,\beta}, K_{20,\alpha}(0), K_{21,\beta}, K_9\}$. Then,*

$$\mathcal{G}'_{n_0} \cap \mathcal{A}_n \cap \Big[\{\|\theta_{n+1} - \theta^*\| > C_R^\theta R^\theta_{proj}\} \cup \{\|w_{n+1} - w^*\| > C_R^w R^w_{proj}\}\Big] = \emptyset.$$

Proof. See (Dalal, Thoppe, and Szörényi 2019)[Appendix F]. □

Therefore, it follows that for $n \geq n_0$

$$\mathcal{G}'_{n_0} \cap \mathcal{A}_n \cap \mathcal{Z}^c_{n+1} \subseteq \mathcal{G}_{n+1} \cap \left[\{\|L^{(\theta)}_{n+2}\| > \epsilon^{(\theta)}_{n+1}\} \cup \{\|L^{(w)}_{n+2}\| > \epsilon^{(w)}_{n+1}\}\right]. \quad (28)$$

Equations (23), (24) and (28) together imply

$$\mathcal{G}'_{n_0} \cap \mathcal{U}^c(n_0) \subseteq \bigcup_{n\geq n_0} \Big(\mathcal{G}_n \cap \big[\{|L^{(\theta)}_{n+1}\| > \epsilon^{(\theta)}_n\} \cup \{\|L^{(w)}_{n+1}\| > \epsilon^{(w)}_n\}\big]\Big). \quad (29)$$

The usefulness of this decomposition lies in the fact that each term in the union contains the event $\mathcal{G}_n$ which ensures that the iterates are bounded. This, along with our noise assumption in Definition 1, implies that the Martingale differences are in turn bounded and the Azuma-Hoeffding inequality can now be invoked (see (Dalal, Thoppe, and Szörényi 2019)[Lemma 29]). Applying this on (29) after using the union bound gives

$$\mathbb{P}\{\mathcal{U}^c(n_0) \,|\, \mathcal{G}'_{n_0}\} = \mathbb{P}\{\mathcal{U}^c(n_0) \cap \mathcal{G}'_{n_0} \,|\, \mathcal{G}'_{n_0}\} \quad (30)$$

$$\leq \sum_{n\geq n_0} \mathbb{P}\Big(\mathcal{G}_n \cap \{\|L^{(w)}_{n+1}\| > \epsilon^{(w)}_n\} \,|\, \mathcal{G}'_{n_0}\Big) \quad (31)$$

$$+ \sum_{n\geq n_0} \mathbb{P}\Big(\mathcal{G}_n \cap \{\|L^{(w)}_{n+1}\| > \epsilon^{(w)}_n\} \,|\, \mathcal{G}'_{n_0}\Big) \quad (32)$$

$$\leq \sum_{n\geq n_0} 2d^2 \exp\left(-\tfrac{(\epsilon^{(\theta)}_n)^2}{d^3 L_\theta a_{n+1}}\right) + \sum_{n\geq n_0} 2d^2 \exp\left(-\tfrac{(\epsilon^{(w)}_n)^2}{d^3 L_w b_{n+1}}\right). \quad (33)$$

Additionally, due to Lemma 14 in (Dalal, Thoppe, and Szörényi 2019),

$$\begin{aligned} a_{n+1} &\leq C_{14,\theta}(n+1)^{-\alpha}, \\ b_{n+1} &\leq C_{14,w}(n+1)^{-\beta}. \end{aligned} \quad (34)$$

Substituting (34) and (14) in (33) gives

$$\mathbb{P}\left\{\mathcal{G}'_{n_0} \cap \mathcal{U}^c(n_0) \,\middle|\, \mathcal{G}'_{n_0}\right\} \leq \sum_{n\geq n_0} \frac{\delta}{(n+1)^p} \leq \delta \frac{n_0^{-(p-1)}}{p-1}. \quad (35)$$

Now, since

$$n_0 \geq (p-1)^{-1/(p-1)}, \quad (36)$$

it eventually follows that (35) $\leq \delta$, as desired. □

4.2 Proof of Theorem 8

For a sequence $u \in \mathbb{R}^\infty_+$, let

$$\mathcal{W}_n(u) := \{\|w_k - w^*\| \leq u_k \ \forall n_0 \leq k \leq n\}. \quad (37)$$

Definition 10. *We say that* $u \in \mathbb{R}^\infty_+$ *is* α-moderate from k_0 onwards *if*

$$\frac{u_k}{u_{k+1}} \leq \frac{\alpha_{k+1}}{\alpha_k} \frac{\beta_k}{\beta_{k+1}} e^{q_1/2\ \alpha_{k+1}}, \quad \forall k \geq k_0.$$

Definition 11. *We say that* $u \in \mathbb{R}^\infty_+$ *is* β-moderate from k_0 onwards *if*

$$\frac{u_k}{u_{k+1}} \leq \frac{\alpha_{k+1}}{\alpha_k} \frac{\beta_k}{\beta_{k+1}} e^{q_2/2\ \beta_{k+2}}, \quad \forall k \geq k_0.$$

We consider these definitions to be part of the novelty of this work. They characterize a sequence via the ratio of its consecutive terms. Ratios in a decaying sequence (such as the ones used in this paper) satisfying Defs. 10 or 11 will converge to 1. Examples of sequences satisfying these definitions are constant sequences and those that decay at an inverse polynomial rate. On the other hand, sequences that decay exponentially fast do not satisfy these conditions. These definitions play a crucial role in enabling our induction; i.e., they help us show that the estimates on the rate of convergence of $\|w_n - w^*\|$ can be incrementally improved. One quick way to see this is via (43) given later; it shows that if the bound on $\|w_n - w^*\|$ was u_n, then it can be improved via induction to $O(\epsilon_n) + O\left(\frac{\alpha_n}{\beta_n} u_n\right)$. These definitions are motivated by Definitions 1 and 2 in (Mokkadem and Pelletier 2006). However, there they are expressed as a certain asymptotic behavior, while ours provide the exact sequence, including constants, and thereby enable finite time analysis.

For $\ell \geq 0$, let $\mathcal{E}(n_0;\ell) := \bigcap_{n\geq n_0}\{\|w_n - w^*\| \leq u_n(\ell)\}$, where

$$u_n(\ell) := \left[A_{1,n_0} \sum_{i=0}^{\ell-1} A_2^i\right] \epsilon^{(w)}_n + \left[A_3 A_2^\ell\right] \left[\frac{\alpha_n}{\beta_n}\right]^\ell; \quad (38)$$

all the constants are given in (Dalal, Thoppe, and Szörényi 2019)[Table 3].

Proof of Theorem 8. Our proof idea inspired by (Mokkadem and Pelletier 2006) is as follows. We use induction to show that whenever $\mathcal{U}(n_0)$ holds, the rate of convergence of w_n is bounded by (38) for all $\ell \leq \ell^*$, where the latter is as in (39). Notice that there are two terms in (38) that depend on n, one is ϵ_n and the other is α_n/β_n. As ℓ increases, $(\alpha_n/\beta_n)^\ell$ decays faster. Thus, eventually, for $\ell = \ell^*$, the convergence rate of w_n would be dictated by ϵ_n, thereby giving us our desired result.

Formally, we begin with proving the following claim.

Claim: Let

$$\ell^* = \left\lceil \frac{\beta}{2(\alpha-\beta)} \right\rceil; \quad (39)$$

i.e., let ℓ^* be the smallest integer ℓ such that $(\alpha-\beta)\ell \geq \beta/2$. Then, for $0 \leq \ell \leq \ell^*$,

$$\mathcal{U}(n_0) \subseteq \mathcal{E}(n_0;\ell). \quad (40)$$

Induction Base: By definition, $\mathcal{U}(n_0) \subseteq \mathcal{E}(n_0, 0)$.

Induction Hypothesis: Suppose (40) holds for some ℓ such that $0 \leq \ell < \ell^*$.

Induction Step: For the ℓ defined in the hypothesis above, we have $(\alpha-\beta)\ell < \beta/2$. Making use of this, we now show that $\mathcal{U}(n_0) \subseteq \mathcal{E}(n_0, \ell+1)$.

From the induction hypothesis, on $\mathcal{U}(n_0)$, for $n \geq n_0 - 1$,

$$\|w_{n+1} - w^*\| \leq u_{n+1}(\ell). \quad (41)$$

A useful result for improving this bound is the following.

Lemma 12. *Let $n_0 \in \mathbb{N}$. Let $u \in \mathbb{R}_+^\infty$ be a monotonically decreasing sequence that is both α-moderate and β-moderate from $n_0 - 1$ onwards. Let $n \geq n_0 - 1$. Suppose that the event $\mathcal{W}_n(u)$ holds, $\|L_n^{(\theta)}\|\mathbb{1}[n \geq n_0 + 1] \leq \epsilon_{n-1}^{(\theta)}$, and $\|L_n^{(w)}\|\mathbb{1}[n \geq n_0 + 1] \leq \epsilon_{n-1}^{(w)}$. If $n \geq n_0$, then*

$$\|\theta_n - \theta^*\| \leq C_{32,b}\frac{\alpha_{n-1}}{\beta_{n-1}}u_{n-1} + \epsilon_{n-1}^{(\theta)} \tag{42}$$
$$+ C_{32,a}\left[\|\theta_{n_0} - \theta^*\| + \frac{\alpha_{n_0}}{\beta_{n_0}}\|w_{n_0} - w^*\|\right] e^{-q_1\sum_{j=n_0+1}^{n-1}\alpha_j}.$$

Additionally, if $n_0 \geq \max\{K_{30,a}, K_{30,b}, K_{35,a}, K_{35,b}, K_{20,\alpha}(\beta/2)\} + 1$ and $n \geq n_0 - 1$, then

$$\|w_{n+1} - w^*\| \leq A_{1,n_0}\epsilon_{n+1}^{(w)} + A_2\frac{\alpha_{n+1}}{\beta_{n+1}}u_{n+1}. \tag{43}$$

All the constants are as in (Dalal, Thoppe, and Szörényi 2019)[Table 3].

Proof. See (Dalal, Thoppe, and Szörényi 2019)[Appendix H]. □

We now verify the conditions necessary to apply this result. After substituting the value of $\epsilon_n^{(w)}$ from (14), and those of α_n, β_n into (38), and then pulling out p from (14) to the constants, observe that $u_n(\ell)$ is of the form

$$u_n(\ell) = B_1(n+1)^{-\beta/2}\sqrt{\ln[B_2(n+1)]} + B_3(n+1)^{-(\alpha-\beta)\ell} \tag{44}$$

for some suitable constants B_1, B_2 and B_3. Clearly, B_1 and B_3 are strictly positive, while $B_2 = (4d^2/\delta)^{1/p} \geq 1$. Lemma 34 in (Dalal, Thoppe, and Szörényi 2019) then shows $\{u_n(\ell)\}$ is α-moderate, β-moderate, and monotonically decreasing from $n_0 - 1$ onwards.

Additionally, notice that due to (41) the event $\mathcal{W}_n(u)$ holds for $u = \{u_n(\ell)\}$, while on $\mathcal{U}(n_0)$ the events $\{\|L_n^{(\theta)}\|\mathbb{1}[n \geq n_0 + 1] \leq \epsilon_{n-1}^{(\theta)}\}$ and $\{\|L_n^{(w)}\|\mathbb{1}[n \geq n_0 + 1] \leq \epsilon_{n-1}^{(w)}\}$ hold. Since $n_0 \geq N_8 \geq \max\{K_{30,a}, K_{30,b}, K_{35,a}, K_{35,b}, K_{20,\alpha}(\beta/2)\} + 1$, we can now employ Lemma 12 with $\{u_n\} = \{u_n(\ell)\}$ and obtain that, on the event $\mathcal{U}(n_0)$,

$$\|w_{n+1} - w^*\| \leq A_{1,n_0}\epsilon_{n+1}^{(w)} + A_2\frac{\alpha_{n+1}}{\beta_{n+1}}u_{n+1}(\ell).$$

By substituting the value of $u_{n+1}(\ell)$ from (38) and making use of the fact that $\alpha_n/\beta_n \leq 1$, we get

$$A_{1,n_0}\epsilon_{n+1}^{(w)} + A_2\frac{\alpha_{n+1}}{\beta_{n+1}}u_{n+1}(\ell) \leq u_{n+1}(\ell+1).$$

This completes the proof of the induction step.

When $\ell = \ell^*$, it now follows that $\mathcal{U}(n_0) \subseteq \mathcal{E}(n_0; \ell^*)$. That is, when the event $\mathcal{U}(n_0)$ holds,

$$\|w_{n+1} - w^*\| \leq u_{n+1}(\ell^*), \quad \forall n \geq n_0 - 1.$$

We now bound $u_n(\ell^*)$. Since $\lceil \frac{\beta}{2(\alpha-\beta)} \rceil \geq \frac{\beta}{2(\alpha-\beta)}$, we have $(\alpha_n/\beta_n)^{\lceil \frac{\beta}{2(\alpha-\beta)} \rceil} \leq (n+1)^{-\beta/2}$. Substituting the value of $\epsilon_n^{(w)}$ and using the above relation along with the fact that $4 \geq e$ which implies $\sqrt{\ln(4d^2(n+1)^p/\delta)} \geq 1$, we have

$$u_n(\ell^*) \leq \left[A_{1,n_0}\sum_{i=0}^{\lceil \frac{\beta}{2(\alpha-\beta)} \rceil - 1} A_2^i\sqrt{d^3 L_w C_{14,w}} + A_3 A_2^{\lceil \frac{\beta}{2(\alpha-\beta)} \rceil}\right]\nu(n;\beta). \tag{45}$$

Consequently, for $n \geq n_0 - 1$,

$$\|w_{n+1} - w^*\| \leq u_{n+1}(\ell^*) \leq A_{4,n_0}\,\nu(n+1;\beta) \tag{46}$$

which establishes (18).

We now prove (19). On the event $\mathcal{U}(n_0)$, we can apply (42) from Lemma 12 with $\{u_n\} = \{u_n(\ell^*)\}$ and use the fact that $\alpha_{n_0}/\beta_{n_0} \leq 1$, as well as bound $\|\theta_{n_0} - \theta^*\|$ and $\|w_{n_0} - w^*\|$ using $\mathcal{U}(n_0)$, to get

$$\|\theta_n - \theta^*\| \leq C_{32,b}\frac{\alpha_{n-1}}{\beta_{n-1}}u_{n-1}(\ell^*) + C_{32,a}\left[C_R^\theta R_{\text{proj}}^\theta + C_R^w R_{\text{proj}}^w\right] e^{-q_1\sum_{j=n_0+1}^{n-1}\alpha_j} + \epsilon_{n-1}^{(\theta)}.$$

Now, Lemma 35 in (Dalal, Thoppe, and Szörényi 2019) and the fact that $q_1 \geq q_{\min}$ imply (in Lemma 35 we require $n \geq n_0$ but here we use it from $n_0 - 1$, which is justified since $n_0 \geq K_{35,b} + 1$), on $\mathcal{U}(n_0)$,

$$\|\theta_n - \theta^*\| \leq C_{32,b}\frac{\alpha_{n-1}}{\beta_{n-1}}u_{n-1}(\ell^*) + \left[C_{32,a}\left[C_R^\theta R_{\text{proj}}^\theta + C_R^w R_{\text{proj}}^w\right]/\epsilon_{n_0-1}^{(\theta)} + 1\right]\epsilon_{n-1}^{(\theta)}.$$

Consequently, using (14), (46) and the facts that $\alpha_{n-1}/\beta_{n-1} = n^{-(\alpha-\beta)}$ and $\alpha/2 = \alpha - \alpha/2 \leq \alpha - \beta/2$, we have that, on $\mathcal{U}(n_0)$,

$$\|\theta_n - \theta^*\| \leq C_{32,b}[A_{4,n_0}\nu(n-1,\alpha)] + \left[C_{32,a}\left[C_R^\theta R_{\text{proj}}^\theta + C_R^w R_{\text{proj}}^w\right]/\epsilon_{n_0-1}^{(\theta)} + 1\right]\epsilon_{n-1}^{(\theta)}.$$

Since $\nu(n-1,\alpha) \leq 2\nu(n,\alpha)$, the theorem follows. □

5 Discussion

Two-timescale SA lies at the foundation of RL in the shape of several popular evaluation and control methods. This work introduces the tightest finite sample analysis for the GTD algorithm suite. We provide it as a general methodology that applies to all linear two-timescale SA algorithms.

Extending our methodology to the case of GTD algorithms with non-linear function-approximation, in similar fashion to (Bhatnagar et al. 2009), would be a natural future direction to consider. Such a result could be of high interest due to the attractiveness of neural networks. Finite time analysis of non-linear SA would also be of use in better understanding actor-critic RL algorithms. An additional direction for future research could be finite sample analysis of distributed SA algorithms of the kind discussed in (Mathkar and Borkar 2016).

Lastly, it would also be interesting to see how adaptive stepsizes can help improve sample complexity in all the above scenarios.

References

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

Bhatnagar, S.; Precup, D.; Silver, D.; Sutton, R. S.; Maei, H. R.; and Szepesvári, C. 2009. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*, 1204–1212.

Borkar, V. S. 1997. Stochastic approximation with two time scales. *Systems & Control Letters* 29(5):291–294.

Borkar, V. S. 2009. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Dalal, G.; Szorenyi, B.; Thoppe, G.; and Mannor, S. 2018a. Finite sample analyses for td(0) with function approximation. In *AAAI*.

Dalal, G.; Thoppe, G.; Szörényi, B.; and Mannor, S. 2018b. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In Bubeck, S.; Perchet, V.; and Rigollet, P., eds., *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, 1199–1233. PMLR.

Dalal, G.; Thoppe, G.; and Szörényi, B. 2019. A tale of two-timescale reinforcement learning with the tightest finite-time bound. *arXiv preprint arXiv:2937714*.

Gerencsér, L. 1997. Rate of convergence of moments of spall's spsa method. In *Control Conference (ECC), 1997 European*, 2192–2197. IEEE.

Konda, V. R., and Tsitsiklis, J. N. 2004. Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability* 14(2):796–819.

Kushner, H. J., and Yin, G. G. 1997. *Stochastic Approximation Algorithms and Applications*.

Lakshminarayanan, C., and Bhatnagar, S. 2017. A stability criterion for two timescale stochastic approximation schemes. *Automatica* 79:108–114.

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.

Loizou, N., and Richtárik, P. 2017. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*.

Maei, H. R.; Szepesvári, C.; Bhatnagar, S.; and Sutton, R. S. 2010. Toward off-policy learning control with function approximation. In *ICML*, 719–726.

Mathkar, A. S., and Borkar, V. S. 2016. Nonlinear gossip. *SIAM Journal on Control and Optimization* 54(3):1535–1557.

Mokkadem, A., and Pelletier, M. 2006. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability* 16(3):1671–1702.

Polyak, B. T. 1990. New stochastic approximation type procedures. *Automat. i Telemekh* 7(98-107):2.

Ruppert, D. 1988. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

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

Thoppe, G., and Borkar, V. 2019. A concentration bound for stochastic approximation via alekseev's formula. *Stochastic Systems* 9(1):1–26.