Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation

Gandharv Patil
McGill University, Mila

L.A. Prashanth
IIT Madras

Dheeraj Nagaraj
Google Research

Doina Precup
McGill University, Mila

Abstract

We study the finite-time behaviour of the popular temporal difference (TD) learning algorithm, when combined with tail-averaging. We derive finite time bounds on the parameter error of the tail-averaged TD iterate under a step-size choice that does not require information about the eigenvalues of the matrix underlying the projected TD fixed point. Our analysis shows that tail-averaged TD converges at the optimal $O\left(1/t\right)$ rate, both in expectation and with high probability. Our bounds exhibit a sharper rate of decay for the initial error (bias), which is an improvement over averaging all iterates. We also propose and analyse a variant of TD that incorporates regularisation, and show that this variant fares favourably in problems with ill-conditioned features.

1 Introduction

Temporal difference (TD) [21] learning is an efficient and easy to implement stochastic approximation algorithm used for evaluating the long-term performance of a decision policy. The algorithm predicts the value function using a single sample path obtained by simulating the Markov decision process (MDP) with a given policy. Analysis of TD algorithms is challenging, and researchers have devoted significant effort in studying its asymptotic properties [23, 15, 19, 10]. In recent years, there has been an interest in characterising the finite-time behaviour of TD, and several papers [6, 17, 2, 12, 3] have tackled this problem under various assumptions.

For t iterations/updates, most existing works either provide a $O\left(\frac{1}{t^\alpha}\right)$ (with universal step-size) [6, 2] or a $O\left(\frac{1}{t}\right)$ (with constant step-size) [17, 2, 12] convergence rate to the TD-fixed point $\theta^\star$ defined as $\theta^\star \triangleq A^{-1}b$, where A and b are quantities which depend on the MDP and the policy (see Section 2 for the notational information). To obtain a $O\left(\frac{1}{t}\right)$ rate with a constant step-size, [17, 2] assume that the minimum eigenvalue of the matrix A is known apriori. However, in a typical RL setting, such eigenvalue information is not available. Estimating the matrix A and its lowest eigenvalue accurately might require a large number of additional samples, which makes the algorithm more complicated. Therefore, obtaining a $O\left(\frac{1}{t}\right)$ rate for TD with a *universal* step-size is an important open problem.

In this paper, we provide a solution to this problem by establishing a $O\left(\frac{1}{t}\right)$ bound on the convergence rate for a variant of TD that incorporates tail-averaging, and uses a constant "universal" step-size. In [2, 17, 12] the authors study an alternate version called iterate averaging which was introduced independently by Polyak and Juditsky [16] and Ruppert [18] for general stochastic-approximation algorithms. A shortcoming of iterate averaging is that the initialisation error (i.e., distance between θ_0 and $\theta^\star$) is forgotten at a slower rate than the non-averaged case, and in practical implementations, one usually performs averaging after a sufficient number of iterations have been performed. This type of delayed averaging, called 'tail-averaging', has been explored in the context of ordinary least squares by Jain et al. [11].

Inspired by the analysis of TD learning, we propose a variant of TD that incorporates regularisation, wherein we introduce a parameter λ and solve for the regularised TD fixed point given by $\theta^\star_{\mathsf{reg}} = (A + \lambda \mathbf{I})^{-1} b$. The update rule for this algorithm is similar to vanilla TD except that it involves an additional factor with λ. Through our analysis we observe that using regularisation can be helpful in obtaining better non-asymptotic bounds for many problems, where the discount factor is close to 1.

Concretely, the contributions of this paper are as follows: First, we establish a $O(1/t)$ finite time bounds on the convergence rate of tail-averaged TD and tail-averaged TD with regularisation. Similar to [2, 6], the analysis assumes that the data is sampled in an i.i.d. fashion from a fixed distribution. The resulting bounds are valid under a universal step-size and hold in expectation as well as high probability. We also show that Markov sampling can be handled with simple mixing arguments. The salient features of the bounds for each variant are as follows:

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Tail averaged TD: In this variant, the step-size is a function of the discount factor and a bound on the norm of the state features. The expectation bound provides a $O(1/t)$ convergence rate for tail-averaged TD iterate, while the high-probability bound establishes an exponential concentration of tail-averaged TD around the projected TD fixed point.

Tail-averaged TD with regularisation: For this variant, the step-size is a function of the discount factor, regularisation parameter λ, and a bound on the norm of the state features. Although this variant converges to the regularised TD fixed-point $\theta^\star_{\mathsf{reg}}$, we show that the worse-case bound on the difference between TD fixed point $\theta^\star$ and $\theta^\star_{\mathsf{reg}}$ is $O(\lambda)$ in the ℓ_2 norm. Moreover, our analysis makes a case for using the regularised TD algorithm for problems with ill-conditioned features.

Next, we show that under mixing assumptions, we can extend our results to Markov sampling instead of i.i.d. sampling. These error bounds contain an extra $\tilde{O}(\tau_{\mathsf{mix}})$, where τ_{mix} is the underlying Markov chain's mixing time. This is no better than making the samples appear approximately i.i.d. by considering one out of every $\tilde{O}(\tau_{\mathsf{mix}})$ samples, and then dropping the rest. In fact, as per Nagaraj et al. [13, Theorem 2], even with the discount factor $\beta = 0$, it is information-theoretically impossible to do any better without further assumptions on the nature of the linear approximation. Recently Agarwal et al. [1] showed that for linear MDPs, one can use reverse experience replay with function approximation to obtain finite time bounds which are independent of the mixing time constant. We leave the study of TD with different experience replay strategies as an interesting future direction, and for the sake of completeness, present the bounds for Markov sampling in Remark 8, and provide a proof sketch in Section 6.

In Table 1 we compare our expectation bounds with existing bounds in the literature. In addition, we also derive high-probability bounds for tail-averaged TD with/without regularisation, and we provide a summary of these bounds in a tabular form in Table 2.

Related work. Over the past few years, there has been significant interest in understanding the finite-time behaviour of TD learning. Several researchers have proposed interesting frameworks establishing bounds on TD's convergence rate under different assumptions. In [20, 25, 7, 24, 9] the authors analyse the finite time behaviour of TD using Lyapunov drift-conditions and establish finite time bounds that hold under expectation. The advantage of this framework is that it can be used directly for analysing TD with Markov noise. However, to provide an $O(1/t)$ bound, these analyses use a step-size which depends on the eigenvalue of A. For *eg.*, in [20, Theorem 7], we have $\epsilon = O(\frac{\log T}{\gamma_{\max} T})$ where $\gamma_{\max}$ is essentially the smallest eigenvalue of A. Similar conditions can also be found in [7, Eq. (88)], [24, Proposition 2], and [9, Eq. (18)].

Table 1: Summary of the bounds in expectation of the form $\mathbb{E}[\|\theta_{\mathrm{Alg},t} - \theta^\star\|_2^2]$, where $\theta^\star$ is the TD fixed point, and $\theta_{\mathrm{Alg},t}$ is the parameter picked by an algorithm after t iterations of TD.

Reference	Algorithm	Step-size	Rate
Bhandari et al. [2]	Last iterate Averaged iterate	c/t[1] $\frac{1}{\sqrt{t}}$	$O(1/t)$ $O(1/\sqrt{t})$
Dalal et al. [6]	Last iterate	$1/t^\alpha$	$O(1/t^\alpha)$
Lakshminarayanan and Szepesvari [12]	Constant step-size with averaging	c	$O(1/t)$
Prashanth et al. [17]	Last iterate Averaged iterate	$c/n, c \propto 1/\mu$ $c/t^\alpha, c > 0$	$O(1/t)$ $O(1/t^\alpha)$
Our work	Tail-averaged TD Regularised TD[2]	$c > 0$ $c > 0$	$O(1/t)$ $O(1/t)$

[1]Step-size requires information about eigenvalue of the feature covariance matrix Σ.
[2]The convergence here is to the regularised TD solution.

Table 2: Summary of the high-probability bounds of the form $\mathbb{P}\left[\|\theta_{\mathrm{Alg},t} - \theta^\star\|_2^2 \leq h(t)\right]$, where $\theta^\star$ is the TD fixed point, $\theta_{\mathrm{Alg},t}$ is the parameter picked by an algorithm after t iterations of TD, and $h(t)$ is a function of t that depends on Alg.

Reference	Algorithm	Step-size	$h(t)$
Dalal et al. [6]	Last iterate	$1/t^\alpha$	$O(1/t^\alpha)$
Prashanth et al. [17]	Last iterate Averaged iterate	$c/n, c \propto 1/\mu$ $c/t^\alpha, c > 0$	$O(1/t)$ $O(1/t^\alpha)$
Our work	Tail-averaged TD Regularised TD[2]	$c > 0$ $c > 0$	$O(1/t)$ $O(1/t)$

[2]The convergence here is to the regularised TD solution.

The analysis presented in this work is closely related to bounds established in [2, 17, 12], where the authors provide an $O(1/t)$ bound in expectation on the mean square error of the parameters. Our bounds match the overall order of these bounds under comparable assumptions. The principal advantage with our bounds is that they hold for a 'universal' step-size choice, while the aforementioned references required the knowledge of μ. Another advantage with our bounds, owing to tail averaging, is that the initial error is forgotten exponentially fast, while the corresponding term in the aforementioned references exhibit a power law decay. In another related work, for a universal step-size the authors in [6] provide a $O(1/t^\alpha)$ bound in expectation, where $\alpha \in (0, 1)$, while we obtain a $O(1/t)$ bound under similar assumptions. Finally, high-probability bounds for TD have been derived in [6, 17]. In comparison to these works, the high-probability bound that we derive is easy to interpret and exhibits better concentration properties. The related Q-learning algorithm and modifications have also been considered in the finite-time regime with linear function approximation (cf. [5, 4] and the references therein). However, these results too require the knowledge of the condition number to set the step-size.

The rest of the paper is organised as follows: In Section 2, we present the main model of TD with function approximation used for our analysis. In Section 3, we describe the tail-averaged TD algorithm, and also present the finite time bounds for this algorithm. In Section 4, we combine tail-averaging with regularisation in a TD algorithm, and provide finite time bounds for this algorithm. In Section 5, we present a sketch of the proofs of our main results, and the detailed proofs are available in [14, Section 6]. In Section 6, we discuss the extension of our results to address the case of Markov sampling. Finally, in Section 7, we provide the concluding remarks.

2 TD with linear function approximation

Consider an MDP $\langle \mathcal{S}, \mathcal{A}, P, r, \beta \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s'|s,a)$ is the probability of transitioning to the state s' from the state s on choosing action a, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the per step reward, and $\beta \in (0,1]$ is the discount factor. We assume that the state and action spaces are both finite. A stationary randomised policy π maps every state s to a distribution over actions. For a given policy π, we define the value function V^π as follows:

$$V^\pi(s) = \mathbb{E}^{\pi,P}\left[\sum_{t=0}^{\infty} \beta^t r(s_t, a_t) | S_0 = s\right], \tag{1}$$

where the action a_t in state s_t is chosen using policy π, i.e., $a_t \sim \pi(s_t)$. The value function V^π obeys the Bellman equation $\mathcal{T}^\pi V^\pi = V^\pi$, where the Bellman operator $\mathcal{T}^\pi$ is defined by $(\mathcal{T}^\pi V)(s) \triangleq \mathbb{E}^{\pi,P}\left[r(s,a) + \beta V(s')\right]$, where the action a is chosen using π, i.e., $a \sim \pi(s)$ and the next state s' is drawn from $P(\cdot|s)$.

2.1 Value function approximation

Most practical applications have high-dimensional state-spaces making exact computation of the value function infeasible. One solution to overcome this problem is to use a parametric approximation of the value function. In this work, we consider the linear function approximation architecture [22], where the value function $V^\pi(s)$, for any $s \in \mathcal{S}$, is approximated as follows:

$$V^\pi(s) \approx \widetilde{V}(s;\theta) := \phi(s)^\top \theta. \tag{2}$$

In the above, $\phi(s) \in \mathbb{R}^d$ is a fixed feature vector for state s, and $\theta \in \mathbb{R}^d$ is a parameter vector that is shared across states. When the state space is a finite set, say $\mathcal{S} = \{1, 2, \ldots, n\}$, the n-vector $\widetilde{V}(\theta)$ with components $\widetilde{V}(s;\theta)$ can be expressed as follows:

$$\widetilde{V}(\theta) = \underbrace{\begin{bmatrix} \phi_1(1) & \phi_2(1) & \ldots & \phi_d(1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(n) & \phi_2(n) & \ldots & \phi_d(n) \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}}_{\theta}, \tag{3}$$

where $\Phi \in \mathbb{R}^{n \times d}$, and $\theta \in \mathbb{R}^d$.

The objective is to learn the best parameter for approximating V^π within the following linear space:

$$\mathcal{B} := \{\Phi\theta \mid \theta \in \mathbb{R}^d\}. \tag{4}$$

Naturally, with a linear function approximation, it is not possible to find the fixed point $V^\pi = \mathcal{T}^\pi V^\pi$. Instead, one can approximate V^π within $\mathcal{B}$ by solving a projected system of equations. The system of equations, which is also referred to as the projected Bellman equation, is given by

$$\Phi\theta^\star = \Pi \mathcal{T}^\pi(\Phi\theta^\star), \tag{5}$$

where Π is the orthogonal projection operator onto the set $\mathcal{B}$ using a weighted ℓ_2-norm. More precisely, let $D = \text{diag}(\rho(1), \ldots, \rho(n)) \in \mathbb{R}^{n \times n}$ denote a diagonal matrix, whose elements are given by the stationary distribution ρ of the Markov chain underlying the policy π. We assume that the stationary distribution exists (see Assumption 1). Let $\|V\|_D = \sqrt{V^\top D V}$ denote the weighted norm of a n-vector V, and assume that the matrix Φ has full column rank. Then, the operator Π projects orthogonally onto the $\mathcal{B}$ using the $\|\cdot\|_D$ norm, and it can be shown that $\Pi = \Phi(\Phi^\top D \Phi)^{-1}\Phi^\top D$.

Next, the projected TD fixed point θ^* for (5) is given by:

$$A\theta^\star = b, \text{ where } A \triangleq \Phi^\top D(\mathbf{I} - \beta P)\Phi, \quad b \triangleq \Phi^\top D\mathcal{R}, \tag{6}$$

and $\mathcal{R} = \sum_{a \in \mathcal{A}} \pi(s,a) r(s,a)$.

2.2 Temporal Difference (TD) Learning

Temporal difference (TD) [22] algorithms are a class of stochastic approximation methods used for solving the projected linear system given in (5). These algorithms start with a initial guess for the θ_0, and at every time-step t and update them using samples from the Markov chain induced by a policy π.

The update rule is given as follows:

$$\theta_t = \theta_{t-1} + \gamma f_t(\theta_{t-1}), \text{ where}$$
$$f_t(\theta) \triangleq (r_t + \beta\theta^\top \phi(s'_t) - \theta^\top \phi(s_t))\phi(s_t). \tag{7}$$

In the above, γ is the step-size parameter.

An alternate version of the algorithm (which we consider for deriving the high probability bounds) uses the projection Γ as follows:

$$\theta_t = \Gamma(\theta_{t-1} + \gamma f_t(\theta_{t-1})). \tag{8}$$

In (8), operator Γ projects the iterate θ_t onto the nearest point in a closed ball $\mathcal{C} \in \mathbb{R}^d$ with a radius H, which is large enough to include $\theta^\star$.

An interesting result by [23] tells us that for any $\theta \in \mathbb{R}^d$, the function

$f(\theta) \triangleq (r(s,a) + \beta\theta^\top\phi(s') - \theta^\top\phi^\top(s))\phi(s)$ has a well defined steady-state expectation given by

$$\mathbb{E}^{\rho,P}[f(\theta)] = \sum_{s,s'\in\mathcal{S},a\in\mathcal{A}} \rho(s)\pi(s,a)\Big((r(s,a) + P(s'|s,a)\beta\theta^\top\phi(s') - \theta^\top\phi^\top(s))\phi(s)\Big), \tag{9}$$

We can rearrange (9) as $\sum_{s,s'\in\mathcal{S},a\in\mathcal{A}} P(s'|s,a)(r(s,s') + \beta\theta^\top\phi(s'_t)) = (\mathcal{T}^\pi\Phi\theta)(s)$, and use [23, Lemma 8] to get the following:

$$\mathbb{E}^{\rho,P}[f(\theta)] = \Phi^\top D(\mathcal{T}^\pi(\Phi\theta) - \Phi\theta) \tag{10}$$

$$= -A\theta + b, \tag{11}$$

where A and b are as defined in (6). We can then characterise the mean behaviour of TD algorithm using the following update rule:

$$\theta_t = \theta_{t-1} + \gamma\Big(\Phi^\top D(\mathcal{T}^\pi(\Phi\theta_{t-1}) - \Phi\theta_{t-1})\Big) = \theta_{t-1} + \gamma\mathbb{E}^{\rho,P}[f(\theta_{t-1})]. \tag{12}$$

The characterisation of TD's behaviour in (12) is of particular importance as it forms the basis of our analysis.

3 Tail-averaged TD

3.1 Basic algorithm

Tail averaging or suffix averaging refers to returning the average of the final few iterates of the optimisation process, to improve its variance properties. Specifically, for any t, the tail-averaged iterate $\theta_{k+1,N}$ is the average of $\{\theta_{k+1},\dots,\theta_t\}$, computed as follows:

$$\theta_{k+1,N} = \frac{1}{N}\sum_{i=k+1}^{k+N}\theta_i, \tag{13}$$

where $N = t - k$.

An alternative to tail-averaging is the Polyak-Ruppert averaging, where one takes an average of all the iterates. This approach has the best asymptotic convergence rate, as shown by [16]. However, from a non-asymptotic analysis viewpoint, it is usually observed that the initial error (the rate at which the initial point is forgotten) is forgotten slower with iterate averaging compared to the non-averaged case, see [8]. Tail averaging retains the advantages of iterate averaging while ensuring that the initial error is forgotten exponentially fast – a conclusion that can be inferred from the finite time bounds that we derive for the TD algorithm.

Algorithm 1 presents the pseudocode of the tail-averaged TD algorithm.

Algorithm 1: Tail-averaged TD(0)

Input : Initial parameter θ_0, step-size γ, initial state distribution ζ_0, tail-average index k.

Sample an initial state $s_0 \sim \zeta_0$;
for $t = 0, 1, \dots$ **do**
  Choose an action $a_t \sim \pi(s_t)$;
  Observe r_t, and next state s'_t;
  Update parameters: $\theta_t = \theta_{t-1} + \gamma f(\theta_{t-1})$;
  Average the final N iterates:
  $\theta_{k+1,N} = \frac{1}{N}\sum_{i=k+1}^{k+N}\theta_i$, where $N = t - k$.
end

3.2 Finite time bounds

Before presenting our results, we list the assumptions under which we conduct our analysis.

Assumption 1. The Markov chain underlying the policy π is irreducible.

Assumption 2. The samples $\{s_t, r_t, s'_t\}_{t\in\mathbb{N}}$ are independently and identically drawn from the following distribution given as $\rho(s)P(s'|s)$ where ρ is the stationary distribution induced by the policy π and P is the MDP's transition probability matrix.

Assumption 3. For all $s \in \mathcal{S}$, $\|\phi(s)\|_2 \le \Phi_{\max} < \infty$.

Assumption 4. For all $s \in \mathcal{S}$, and $a \in \mathcal{A}$, $|r(s,a)| \le R_{\max} < \infty$.

Assumption 5. The matrix Φ has full column rank.

Assumption 6. The set $\mathcal{C} \triangleq \{\theta \in \mathbb{R}^d | \|\theta\|_2 \le H\}$ used for projection through Γ satisfies $H > \frac{\|\bar{b}\|_2}{\mu}$.

We now discuss the assumptions listed above. Assumption 1 ensures the existence of the stationary distribution for the Markov chain underlying policy π, since the underlying state and action spaces are assumed to be finite. We study the non-asymptotic behaviour of the tail-averaged TD algorithm under the i.i.d observation model specified in Assumption 2, and later show that we can extend our results to handle Markov sampling. Next, Assumptions 3 and 4 are boundedness requirements on the underlying features and rewards, and are common in the finite time analysis of the TD algorithm, see [2, 17]. Assumption 5 requires the columns of the feature matrix Φ to be linearly independent, which ensures the uniqueness of the TD solution θ^*. It also ensures that the minimum eigenvalue, say μ' of $B = \mathbb{E}^{\rho,P}[\Phi\Phi^\top]$ is strictly positive, which implies that the minimum eigenvalue μ of the matrix A defined in (6) is strictly positive. Assumption 6 is required for the high-probability bounds, while the bounds in expectation do not require projection.

The first result we state below is a bound in expectation on the parameter error $\|\theta_{k+1,N} - \theta^\star\|_2^2$.

Theorem 1 (Bound in expectation). *Suppose Assumptions 1 to 5 hold. Choose a step-size γ satisfying*

$$\gamma \leq \gamma_{\max} = \frac{1-\beta}{(1+\beta)^2\Phi_{\max}^2}, \tag{14}$$

where β is the discount factor and $\Phi_{\max}$ is a bound on the features (see Assumption 3).

Then the expected error of the tail-averaged iterate $\theta_{k+1,N}$ formed using Algorithm 1 satisfies

$$\mathbb{E}\left[\|\theta_{k+1,N} - \theta^\star\|_2^2\right] \leq \frac{10e^{(-k\gamma(1-\beta)\mu')}}{\gamma^2(1-\beta)^2\mu'^2N^2}\mathbb{E}\left[\|\theta_0 - \theta^\star\|_2^2\right] + \frac{10\sigma^2}{(1-\beta)^2\mu'^2N}, \tag{15}$$

where $N = t - k$, θ_0 is the initial point, $\sigma = (R_{\max} + (1+\beta)\Phi_{\max}^2\|\theta^\star\|_2)$, with $\theta^\star$ denoting the TD fixed point specified in (6), and μ' is the minimum eigenvalue of $B = \mathbb{E}^{\rho,P}[\Phi\Phi^\top]$.

Proof. See Section 5.1 for the proof sketch and [14, Section 6] for detailed proof. □

A few remarks are in order.

Remark 1. It is apparent that the bound presented above scales inversely with the square of $(1-\beta)\mu'$. More importantly, the bound presented above is for a step-size choice that does not require information about the eigenvalues of matrices A or B. To the best of our knowledge, this is the first bound of $O(1/t)$ for a 'universal' step-size for tail-averaged TD. Previous results, such as those by [2, 17] provide a comparable bound, albeit for a diminishing step-size of the form c/k, where setting c requires knowledge of μ. On the other hand, [6, 17] provide a $O(1/t^\alpha)$ bound for larger step-sizes of the form c/t^α, where c is a universal constant and $\alpha < 1$

Remark 2. The first term on the RHS of (15) relates to the rate at which the initial parameter θ_0 is forgotten, while the second term arises from a martingale difference noise term associated with the i.i.d. sampling model. Setting $k = t/2$, we observe that the first term is forgotten at an exponential rate, while the noise term is $O(1/t)$.

Remark 3. In [12], the authors consider iterate averaging in the linear stochastic approximation setting. Comparing their Theorem 1 to the result we have presented above, we note that the first term on the RHS of (15) exhibits an exponential decay, while the corresponding decay is of order $O(1/t)$ in [12]. The second term in their result as well as in (15) is of order $O(1/t)$. While the second dominates the rate, the first term, which relates to the rate at which the initial parameter is forgotten, decays much faster with tail averaging. Intuitively, it makes sense to average after sufficient iterations have passed instead of averaging from the beginning, and our bounds confirm this viewpoint.

Remark 4. A closely related result under comparable assumptions is Theorem 2 of [2]. This result provides two bounds corresponding to constant and diminishing step-sizes, respectively, while assuming the knowledge of μ. The bound there corresponding to the constant stepsize for the last iterate of TD is the sum of an exponentially decaying 'initial error' term and a constant offset with the noise variance. The second bound in the aforementioned work is $O(1/t)$ for both initial error and noise terms. The bound we derived in (15) combines the best of these two bounds through tail averaging, i.e., an exponentially decaying initial error and a $O(1/t)$ noise term. As an aside, our bound is for the projection-free variant of TD, while the bounds in [2] requires projection, with an assumption similar to Assumption 6 specified.

Remark 5. Another closely related result is Theorem 4.4 of [17], where the authors analyse TD with linear function approximation, with input data from a batch of samples. One can easily extend their analysis to cover our i.i.d. sampling model. As in the remark above, while the overall rate is $O(1/t)$ in their result as well as (15), the initial error in our bound is forgotten much faster. A similar observation also holds w.r.t. the bound in the recent work [3], but the authors do not state their bound explicitly.

Remark 6. It is possible to extend our analysis to cover the Markov noise observation model, as specified in Section 8 of [2]. In this model, we assume that the underlying Markov chain is fast mixing. For finite Markov chains, irreducibility and aperiodicity is sufficient to establish this (see Assumption 1). The fast mixing assumption allows us to translate the i.i.d. sample bounds to Markov sample bounds. We provide the details of such an extension in Remark 8 and Section 6.

Next, we turn to provide a bound that holds with high probability for the parameter error $\|\theta_{k+1,N} - \theta^\star\|_2^2$ of the projected TD algorithm. For this result, we require the TD update parameter to stay within a bounded region that houses θ^*, which is formalised in Assumption 6.

Theorem 2 (High-probability bound). *Suppose Assumptions 1 to 6 hold. Choose the step-size such that $\gamma \leq \gamma_{\max}$, where $\gamma_{\max}$ is defined in (14). Then, for any $\delta \in (0,1]$, we have the following bound for the projected tail-averaged iterate $\theta_{k+1,N}$:*

$$P\left(\|\theta_{k+1,N} - \theta^\star\|_2 \leq \frac{2\sigma}{(1-\beta)\mu'\sqrt{N}}\sqrt{\log\left(\frac{1}{\delta}\right)} + \frac{4e^{(-k\gamma(1-\beta)\mu')}}{\gamma(1-\beta)\mu' N}\mathbb{E}\left[\|\theta_0 - \theta^\star\|_2\right]\right.$$

$$+\frac{4\sigma}{(1-\beta)\mu'\sqrt{N}}\Bigg) \geq 1-\delta, \tag{16}$$

where $N, \sigma, \mu, \theta_0, \theta^\star$ are as specified in Theorem 1.

Proof. See Section 5.2 for the proof sketch and [14, Section 6] for detailed proof. □

Remark 7. High-probability bounds for TD algorithm have been derived earlier in [17, 6]. In comparison to Theorem 4.2 of [17], we note that our bound is an improvement since the sampling error (the first and third terms in $\mathcal{K}(n)$ defined above) decays at a much faster rate for tail-averaged TD. Next, unlike [6], we note that our bound requires projection. However, it does exhibit a $O(1/t)$ rate. The result by [6] (Theorem 3.6) is of the form $O(1/t^\lambda)$ where λ is related to the μ of A, and hence cannot be guaranteed to be of order $O(1/t)$.

Remark 8. Consider the case when Assumption 2 does not hold, but we sample (s_t, r_t, s'_t) from a trajectory corresponding the policy π. We assume exponential ergodicity for the total variation distance as used in [2, 20], with mixing time τ_{mix}. For this case, we consider a variant of TD which uses one sample in every $\tilde{O}(\tau_{\mathsf{mix}})$ consecutive samples for the update iteration. The guarantees for the resulting TD algorithm with tail averaging with N data points in the trajectory correspond to the guarantees for $\theta_{k+1,N'}$ in Theorem 2 where $N' = \tilde{O}\left(\frac{N}{\tau_{\mathsf{mix}}}\right)$, and $(1-\delta)$ is replaced by $(1-2\delta)$. This gives an error of the order $\tilde{O}\left(\sqrt{\frac{\tau_{\mathsf{mix}}}{N}}\right)$, which is similar to the bounds in [2, Theorem 3]. These follow from standard mixing arguments, and we refer to Section 6 for further details. As an aside, we remark that without further assumptions on the linear approximation, it is information theoretically impossible to get a better bound (cf. Theorem 2 in [13]).

4 Regularised TD Learning

In this section, we present the regularised TD algorithm. From the results in Theorems 1 and 2 one can observe that although tail-averaged TD achieves a $O\left(\frac{1}{t}\right)$ rate of convergence, the bounds depend inversely on $(1-\beta)\mu'$, where μ' is the minimum eigenvalue of $B = \mathbb{E}^{\rho,P}[\Phi\Phi^\top]$. In the following results we will show that the non-asymptotic bounds for regularised TD scale inversely with μ (minimum eigenvalue of matrix A). Such dependence may be preferable over vanilla TD, as there are problem instances where $(1-\beta)\mu' \ll \mu$. To make this intuition more concrete, consider the following problem instance.

Example 1. Consider a two state MDP with the transition dynamics as depicted in Figure 1, for a given policy, say π. The one-dimensional state features are given as follows: $\phi(1) = 1$, and $\phi(2) = \frac{1}{2}$. For the case of $p = \frac{1}{2}$, we have

$$A = (1/2)\left(\phi(1)^2 + \phi(2)^2\right) - \beta/4(\phi(1)^2 + \phi(2)^2$$

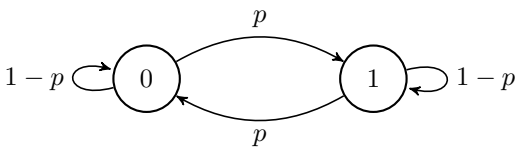

Figure 1: A two state Markov chain

$$+\phi(1)\phi(2) + \phi(2)\phi(1))$$
$$= \frac{5}{8} - \frac{9\beta}{16}.$$

Further, $B = \frac{5}{8}$. Thus, for any $\beta \in [0,1]$, we have

$$(1-\beta)B \leq A.$$

Further, as β approaches 1, $(1-\beta)B \to 0$, while $A \to \frac{9}{16}$. Since the convergence rate of tail-averaged TD depends inversely on $(1-\beta)\mu'$ (see Theorems 1 and 2), there is a concrete case for an algorithm whose convergence rate depends on μ instead of $(1-\beta)\mu'$. The regularised TD variant that we present next achieves this objective.

4.1 Basic algorithm

Instead of the TD solution (6), we solve the following regularised problem for a given regularisation parameter $\lambda > 0$:

$$\theta^\star_{\mathsf{reg}} = (A + \lambda \mathbf{I})^{-1} b, \tag{17}$$

The update for the regularised version of TD is as follows:

$$\hat{\theta}_t = (\mathbf{I} - \gamma\lambda)\hat{\theta}_{t-1} + \gamma(r_t + \beta\hat{\theta}_{t-1}^\top\phi(s'_t) - \hat{\theta}_{t-1}^\top\phi(s_t))\phi(s_t). \tag{18}$$

Similarly, the projected regularised TD update (which we consider for deriving the high probability bounds) uses the projection Γ as follows:

$$\hat{\theta}_t = \Gamma((\mathbf{I} - \gamma\lambda)\hat{\theta}_{t-1} + \gamma(r_t + \beta\hat{\theta}_{t-1}^\top\phi(s'_t) - \hat{\theta}_{t-1}^\top\phi(s_t))\phi(s_t)). \tag{19}$$

In (19), operator Γ projects the iterate θ_t onto the nearest point in a closed ball $\mathcal{C} \in \mathbb{R}^d$ with a radius H which is large enough to include $\theta^\star_{\mathsf{reg}}$.

Using arguments similar to vanilla TD, it is easy to see that the iterate $\hat{\theta}_t$ converges to (17) under Assumptions 1 to 5, and a standard stochastic approximation condition on the step-size.

The overall flow of the regularised TD algorithm would be similar to Algorithm 1, except that the iterate is updated according to (18), and an additional regularisation parameter is involved.

4.2 Finite time bounds

Using a technique similar to that used in establishing the bound for tail-averaged TD in Theorem 1, we arrive at the following bound in expectation for regularised TD.

Theorem 3 (Bound in expectation). *Suppose Assumptions 1 to 5 hold. Choose a step-size γ satisfying*

$$\gamma \leq \gamma_{\max} = \frac{\lambda}{\lambda^2 + 2\lambda(1+\beta)\Phi_{\max}^2 + (1+\beta)^2\Phi_{\max}^4}. \tag{20}$$

Then the expected error of the tail-averaged regularised TD iterate $\hat{\theta}_{k+1,N}$ satisfies

$$\mathbb{E}\left[\left\|\hat{\theta}_{k+1,N} - \theta^\star_{\text{reg}}\right\|_2^2\right] \leq \frac{10e^{(-k\gamma(\mu+\lambda))}}{\gamma^2(\mu+\lambda)^2N^2}\mathbb{E}\left[\left\|\hat{\theta}_0 - \theta^\star_{\text{reg}}\right\|_2^2\right] + \frac{10\sigma^2}{(\mu+\lambda)^2N}, \tag{21}$$

where $N = t - k$, $\sigma = \left(R_{\max} + (1+\beta)\Phi_{\max}^2\left\|\theta^\star_{\text{reg}}\right\|_2\right)$ and μ is the minimum eigenvalue of the matrix A defined in (6).

Proof. See Section 5.3 for the proof sketch, and [14, Section 6] for the detailed proof □

Note that the result above bounds the distance to the regularised TD solution. In the next result we will show that the distance between regularised TD iterate and vanilla projected TD fixed point is of $O(\lambda)$.

Corollary 1. Under conditions of Theorem 3, we have

$$\mathbb{E}\left[\left\|\hat{\theta}_{k+1,N} - \theta^\star\right\|_2^2\right] \leq \frac{20e^{(-k\gamma(\mu+\lambda))}}{\gamma^2(\mu+\lambda)^2N^2}\mathbb{E}\left[\left\|\hat{\theta}_0 - \theta^\star_{\text{reg}}\right\|_2^2\right] + \frac{20\sigma^2}{(\mu+\lambda)^2N} + \frac{2\lambda^2\Phi_{\max}^2R_{\max}^2}{\mu(\mu+\lambda)}. \tag{22}$$

With a suitable choice of λ, the following result shows that regularised TD obtains a $O(1/t)$ rate (e.g., with $k = t/2$), and the bound scales inversely with the eigenvalue μ of the matrix A. From the discussion earlier, recall that there are problem instances where $\mu \gg (1-\beta)\mu'$, and the bound for tail-averaged TD sans regularisation depended inversely on $(1-\beta)\mu'$.

Corollary 2. Under conditions of Theorem 3, and with $\lambda = \frac{1}{\sqrt{N}}$, we obtain

$$\mathbb{E}\left[\left\|\hat{\theta}_{k+1,N} - \theta^\star\right\|_2^2\right] \leq \frac{K}{\mu^2N^3}\mathbb{E}\left[\left\|\hat{\theta}_0 - \theta^\star_{\text{reg}}\right\|_2^2\right] + \frac{20\sigma^2}{\mu^2N} + \frac{2\Phi_{\max}^2R_{\max}^2}{\mu^2N}, \tag{23}$$

where $K = 20(1 + (1+\beta)\Phi_{\max}^2\sqrt{N})^4 e^{\frac{(-k\mu)}{(1+\beta)^2\Phi_{\max}^4\sqrt{N}}}$

A few remarks are in order.

Remark 9. The choice of step-size in the bound of Theorem 3 is universal, i.e., does not require the knowledge of μ, and the rate of convergence is $O(1/t)$, if we set $k = t/2$, or any constant multiple of t. Although the regularised TD iterate converges to (17), which is different from the vanilla TD fixed point, in Corollary 1 we show that the distance between the regularised and vanilla TD solutions is $O(\lambda)$. This implies that for a small value of λ, the regularised TD solution is a good proxy for the vanilla TD, and one can use regularised TD iterate can be used in place of vanilla TD iterate, to obtain a good approximation to the TD fixed point.

Remark 10. The initial and sampling errors in (21) are as in the tail-averaged TD (see Theorem 1), i.e., initial error is forgotten at an exponential rate, while the sampling error is $O(1/t)$.

Remark 11. In [17], the authors analyse the iterate-average variant of TD, and derive a $O(1/t^\alpha)$ bound for a step-size $\Theta(1/k^\alpha)$, where $1/2 < \alpha < 1$. Further, their step-size choice is universal as is the case of tail-averaged TD. Our bound for regularised TD exhibits a better rate than [17].

Next, we present a high-probability bound for regularised TD in the spirit of Theorem 2.

Theorem 4 (High-probability bound). *Suppose Assumptions 1 to 6 hold. Choose the step-size such that $\gamma \leq \gamma_{\max}$, where $\gamma_{\max}$ is defined in (20). Then, for any $\delta \in (0,1]$, we have the following bound for the projected tail-averaged regularised TD iterate $\hat{\theta}_{k+1,N}$:*

$$P\left(\left\|\hat{\theta}_{k+1,N} - \theta^\star_{\text{reg}}\right\|_2 \leq \frac{2\sigma}{(\mu+\lambda)\sqrt{N}}\sqrt{\log\left(\frac{1}{\delta}\right)} + \frac{4e^{(-k\gamma(\mu+\lambda))}}{\gamma(\mu+\lambda)N}\mathbb{E}\left[\left\|\hat{\theta}_0 - \theta^\star_{\text{reg}}\right\|_2\right] + \frac{4\sigma}{(\mu+\lambda)\sqrt{N}}\right) \geq 1 - \delta,$$

where $N, \sigma, \mu, \hat{\theta}_0, \theta^\star_{\text{reg}}$ are as specified in Theorem 3.

Proof. See Section 5.4 for a proof sketch, and [14, Section 6] for the detailed proof. □

As discussed in Remark 8 and Section 6 for tail-averaged TD sans regularisation, it is straightforward to extend the results in Theorems 3 and 4 to cover the case of Markov sampling.

5 Proof Ideas

5.1 Proof of Theorem 1 *(Sketch)*

Proof. We present the framework for obtaining the results obtained in the paper; the framework has been introduced in the work of [17, 2, 6]. Towards that end, we begin by introducing some notation. First, we define the centered

error $z_t \triangleq \theta_t - \theta^\star$. Using the TD update (7), the centred error can be seen to satisfy the following recursive relation:

$$z_t = (\mathbf{I} - \gamma a_t) z_{t-1} + \gamma f_t(\theta^\star), \tag{24}$$

where $f(\cdot)$ defined as in (7), and
$a_t \triangleq \phi(s_t)\phi(s_t)^\top - \beta\phi(s_t)\phi(s_t')^\top$.

The centred error is decomposed into a bias and variance term as follows:

$$\begin{aligned}\mathbb{E}\left[\|z_t\|_2^2\right] &= 2\mathbb{E}\left[\left\|C^{t:1} z_0\right]\right\|_2^2\Big] \\ &\quad + 2\gamma^2 \sum_{k=0}^{t} \mathbb{E}\left[\left\|\gamma \sum_{k=0}^{t} C^{t:k+1} f_k(\theta^\star)\right\|_2^2\right] \\ &= 2z_t^{\mathsf{bias}} + 2\gamma^2 z_t^{\mathsf{variance}},\end{aligned} \tag{25}$$

where

$$C^{i:j} = \begin{cases} (\mathbf{I} - \gamma a_i)(\mathbf{I} - \gamma a_{i-1}) \ldots (\mathbf{I} - \gamma a_j), & \text{if } i >= j, \\ \mathbf{I}, & \text{otherwise.} \end{cases}$$

The bias term is then bounded as follows:

$$z_t^{\mathsf{bias}} \leq \exp(-\gamma(1-\beta)\mu' t)\mathbb{E}[\|z_0\|_2^2].$$

On the other hand, the variance term is bounded above by

$$z_t^{\mathsf{variance}} \leq \frac{\sigma^2}{\gamma(1-\beta)\mu'}.$$

The centered error corresponding to the tail-averaged iterate $\theta_{k+1,N}$ is given by $z_{k+1,N} = \frac{1}{N}\sum_{i=k+1}^{k+N} z_i$. The analysis proceeds by bounding the expectation of the norm $\mathbb{E}[\|z_{k+1,N}\|_2^2]$ using the following decomposition:

$$\begin{aligned}\mathbb{E}\left[\|z_{k+1,N}\|_2^2\right] \leq \frac{1}{N^2}\Bigg(&\sum_{i=k+1}^{k+N} \mathbb{E}\left[\|z_i\|_2^2\right] \\ &+ 2\sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N} \mathbb{E}\left[z_i^\top z_j\right]\Bigg).\end{aligned}$$

Using the definitions of z_t^{bias} and z_t^{variance}, we simplify the RHS above as follows:

$$\begin{aligned}\mathbb{E}\left[\|z_{k+1,N}\|_2^2\right] \leq &\underbrace{\frac{2}{N^2}\left(1 + \frac{4}{\gamma(1-\beta)\mu'}\right)\sum_{i=k+1}^{k+N} z_i^{\mathsf{bias}}}_{z_{k+1,N}^{\mathsf{bias}}} \\ &+ \underbrace{\frac{2}{N^2}\left(1 + \frac{4}{\gamma(1-\beta)\mu'}\right)\gamma^2 \sum_{i=k+1}^{k+N} z_i^{\mathsf{variance}}}_{z_{k+1,N}^{\mathsf{variance}}},\end{aligned} \tag{26}$$

where z_i^{bias} and z_i^{variance} are defined in (25).

The main result follows by substituting the bounds on z_i^{bias} and z_i^{variance} followed by some algebraic manipulations. □

5.2 Proof of Theorem 2 *(Sketch)*

Proof. To obtain the high-probability bound, we use the proof technique by Prashanth et al. [17], where we consider separately the deviation of the centred error from its mean, *i.e.*, $\|z_{k+1,N}\|_2^2 - \mathbb{E}\left[\|z_{k+1,N}\|_2^2\right]$. We decompose this quantity as a sum of martingale differences, establish a Lipschitz property followed by a sub-Gaussian concentration bound to infer

$$\begin{aligned}&P\Big(\|z_{k+1,N}\|_2 - \mathbb{E}\left[\|z_{k+1,N}\|_2\right] > \epsilon\Big) \\ &\leq \exp\left(-\frac{\epsilon^2}{(R_{\mathsf{max}} + (1+\beta)H\Phi_{\mathsf{max}}^2)^2 \sum_{i=k+1}^{k+N} L_i^2}\right),\end{aligned} \tag{27}$$

where $L_i \triangleq \frac{\gamma}{N}\sum_{j=i+1}^{i+N}\left(1 - \frac{\gamma(1-\beta)\mu'}{2}\right)^{j-i+1}$.
Next, under the choice of step-size γ specified in the theorem statement, we establish that

$$\sum_{i=k+1}^{k+N} L_i^2 \leq \frac{4}{N(1-\beta)\mu'^2}.$$

The main claim follows by (i) substituting the bound obtained above in (27); (ii) using the bound on $\mathbb{E}\left[\|z_{k+1,N}\|_2\right]$ specified in Theorem 1; and (iii) converting the tail bound resulting from (i) and (ii) into a high-probability bound. □

5.3 Proof of Theorem 3 (*Sketch*)

The template for the proof of Theorem 3 is more or less similar to Theorem 1. The main difference in the proof technique is the following lemma that helps us capture the effect of the interplay of the step-size and regularisation parameters on the constants and decay rates in Theorem 3's result.

Lemma 1. *With $\gamma \leq \gamma_{\mathsf{max}}$ as given in (20), the following bound holds*

$$\left\|\left[\mathbf{I} - \gamma(A + \lambda\mathbf{I})\right]^\top \left[\mathbf{I} - \gamma(A + \lambda\mathbf{I})\right]\right\|_2 \leq 1 - \gamma(\mu + \lambda),$$

and

$$\|(\mathbf{I} - \gamma A)\|_2 \leq 1 - \frac{\gamma(\mu+\lambda)}{2}.$$

The main consequence of the above result is that the bounds in Theorem 3 directly depend on the minimum eigenvalue of A as opposed to the minimum eigenvalue of B in Theorem 1.

5.4 Proof of Theorem 4 (*Sketch*)

Similar to Theorem 3, high-probability bounds for regularised TD depend on the μ. The main lemma that helps establish this result is as follows:

Lemma 2. *Let* $a_j \triangleq [\phi(s_j)\phi(s_j)^\top - \beta\phi(s_j)\phi(s_j')^\top]$ *and* $\gamma \leq \gamma_{\mathsf{max}}$*, where* γ_{max} *is set as per* (20)*. Then for any* $\mathcal{F}_{i-1}$ *measurable* $\hat{\theta} \in \mathbb{R}^d$ *we have*

$$\mathbb{E}\left[\hat{\theta}^\top(\mathbf{I}-\gamma(\lambda\mathbf{I}+a_j))^\top(\mathbf{I}-\gamma(\lambda\mathbf{I}+a_j))\hat{\theta}|\mathcal{F}_{j-1}\right] \leq (1-\gamma(\mu+\lambda))\left\|\hat{\theta}\right\|_2^2$$

and,

$$\mathbb{E}\left[\left\|(\mathbf{I}-\gamma(\lambda\mathbf{I}+a_j))\hat{\theta}\right\|_2|\mathcal{F}_{j-1}\right] \leq \left(1-\frac{\gamma(\mu+\lambda)}{2}\right)\left\|\hat{\theta}\right\|_2.$$

The rest of the proof then follows steps similar to that of Theorem 2.

6 Bounds for Markov Sampling

In this section we will analyse the performance of tail-averaged TD when $(s_t)_{t\in\mathbb{N}}$ are drawn from a single stationary trajectory of the Markov chain with policy π. To derive our results we assume that the Markov chain is exponentially egrodic, which holds true for any finite Markov chain which is irreducible. Let ρ denote the stationary distribution of the Markov chain under policy π.

Assumption 7. With $s_1 \sim \rho$, there exist constants C and τ_{mix} such that for every $t, \tau \in \mathbb{N}$

$$D(\tau) := \sup_{s\in\mathcal{S}} \mathsf{TV}(s_{t+\tau}|s_t = s, \rho) \leq C\exp(-\tfrac{\tau}{\tau_{\mathsf{mix}}}),$$

where TV denotes the total variation distance between probability measures.

This is a standard assumption in the literature [2, 20]. We now adapt Lemma 3 from [13] to our present setting.

Lemma 3 (Adaptation of Lemma 3 in [13])**.** *For any* $K \in \mathbb{N}$*, define the random variable*

$$S_{K,n} := ((s_1, s_2), (s_{K+1}, s_{K+2}), (s_{2K+1}, s_{2K+2}), \ldots, (s_{nK+1}, s_{nK+2})).$$

Let P^π *denote the transition kernel for the Markov chain under policy* π*. By* $\rho^{(2)}$ *denote the joint distribution of* (s_1, s_2)*. Under Assumption 7, we have*

$$\mathsf{TV}(S_{K,n}, (\rho^{(2)})^{\otimes n}) \leq nD(K-1) \leq nC\exp(-\tfrac{K-1}{\tau_{\mathsf{mix}}}).$$

Proof. Let $R_{K,n} = (r_1, r_{K+1}, \ldots, r_{nK+1})$ be the random rewards corresponding to $S_{K,n}$ and consider i.i.d random variables $\tilde{S}_{K,n} = ((\tilde{s}_1, \tilde{s}_2), (\tilde{s}_{K+1}, \tilde{s}_{K+2}), (\tilde{s}_{2K+1}, \tilde{s}_{2K+2}), \ldots, (\tilde{s}_{nK+1}, \tilde{s}_{nK+2})) \sim (\rho^{(2)})^{\otimes n}$ along with the corresponding rewards $\tilde{R}_{K,n}$. We can define these random variables on a common probability space such that

$$\mathbb{P}((S_{K,n}, R_{K,n}) \neq (\tilde{S}_{K,n}, \tilde{R}_{K,n})) \leq nD(K-1) \leq nC\exp(-\tfrac{K-1}{\tau_{\mathsf{mix}}}). \quad (28)$$

□

Since the samples (s_t, r_t, s_{t+1}) now belong to a trajectory (as opposed being i.i.d), we will modify Algorithm 1 in the following ways to account for the mixing.

We fix $K \in \mathbb{N}$.

Modification 1: Run Algorithm 1 with data $S_{K,n}, R_{K,n}$ - i.e, we input $(s_{tK+1}, r_{tK+1}, s_{tK+2})$ at step t.

Modification 2: Run Algorithm 1 with data $\tilde{S}_{K,n}, \tilde{R}_{K,n}$.

Note that the Modification 2 is exactly same as running the algorithm under Assumption 2 for n steps and therefore the results of Theorem 2 apply to this case if we replace N with n. By the results in Lemma 3, we conclude that the trajectories (θ_t) generated by modification 1 and $(\tilde{\theta}_t)$ generated by modification 2 can be coupled such that

$$\mathbb{P}\left[(\theta_t)_{t=1}^{n+1} \neq (\tilde{\theta}_t)_{t=1}^{n+1}\right] \leq nD(K-1).$$

This is based on the fact that whenever the algorithm is fed with the same input, we obtain the same output. Setting $K = \tau_{\mathsf{mix}}\log(\frac{Cn}{\delta})$, we conclude that under Assumption 7, we have

$$\mathbb{P}\left[(\theta_t)_{t=1}^{n+1} \neq (\tilde{\theta}_t)_{t=1}^{n+1}\right] \leq \delta. \quad (29)$$

Therefore, we conclude the bounds in Remark 8.

7 Conclusions

We presented a finite time analysis of tail-averaged TD algorithm. We obtained $O\left(\frac{1}{t}\right)$ bounds, both in expectation as well as high-probability, for a step-size choice that is 'universal', and this is an improvement over previously known results. Additionally, we proposed and analysed a variant of TD that incorporated regularisation. This algorithm in conjunction with tail averaging was shown to be useful over vanilla tail-averaged TD on problem instances, where the feature matrix is ill-conditioned.

References

[1] N. Agarwal, S. Chaudhuri, P. Jain, D. M. Nagaraj, and P. Netrapalli. Online Target Q-learning with Reverse Experience Replay: Efficiently finding the Optimal Policy for Linear MDPs. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[2] J. Bhandari, D. Russo, and R. Singal. A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation. In *Conference On Learning Theory (COLT)*, volume 75 of *Proceedings*

of Machine Learning Research, pages 1691–1692. PMLR, 2018.

[3] S. Chen, A. Devraj, A. Busic, and S. Meyn. Explicit Mean-Square Error Bounds for Monte-Carlo and Linear Stochastic Approximation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 4173–4183. PMLR, 2020.

[4] Z. Chen, J. P. Clarke, and S. T. Maguluri. Target Network and Truncation Overcome The Deadly triad in Q-Learning. *arXiv preprint arXiv:2203.02628*, 2022.

[5] Z. Chen, S. Zhang, T. T. Doan, J. P. Clarke, and S. T. Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. In *Automatica*, volume 146, 2022.

[6] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor. Finite Sample Analyses for TD(0) With Function Approximation. In *AAAI Conference on Artificial Intelligence, (AAAI)*, pages 6144–6160. AAAI Press, 2018.

[7] A. Durmus, É. Moulines, A. Naumov, S. Samsonov, and H.-T. Wai. On the Stability of Random Matrix Product with Markovian Noise: Application to Linear Stochastic Approximation and TD Learning. In *COLT*, 2021.

[8] M. Fathi and N. Frikha. Transport-entropy inequalities and deviation estimates for stochastic approximation schemes. In *Electronic Journal of Probability*, volume 18, pages 1–36. Institute of Mathematical Statistics and Bernoulli Society, 2013.

[9] B. Hu and U. A. Syed. Characterizing the Exact Behaviors of Temporal Difference Learning Algorithms Using Markov Jump Linear System Theory. In *Advances in neural information processing systems (NeurIPS)*, 2019.

[10] T. S. Jaakkola, M. I. Jordan, and S. P. Singh. On the Convergence of Stochastic Iterative Dynamic Programming Algorithms. In *Neural Computation*, volume 6, pages 1185–1201, 1994.

[11] P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. In *Journal of Machine Learning Research*, volume 18, 2018.

[12] C. Lakshminarayanan and C. Szepesvari. Linear Stochastic Approximation: How Far Does Constant Step-Size and Iterate Averaging Go? In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pages 1347–1355. PMLR, Apr 2018.

[13] D. Nagaraj, X. Wu, G. Bresler, P. Jain, and P. Netrapalli. Least squares regression with Markovian data: Fundamental limits and algorithms. In *Advances in neural information processing systems (NeurIPS)*, volume 33, pages 16666–16676, 2020.

[14] G. Patil, L. Prashanth, D. Nagaraj, and D. Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation, 2022. URL https://arxiv.org/abs/2210.05918.

[15] F. J. Pineda. Mean-Field Theory For Batched-TD(l). In *Neural Computation*, volume 9, pages 1403–1419, 1997.

[16] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. In *SIAM Journal on Control and Optimization*, volume 30, pages 838–855. SIAM, 1992.

[17] L. A. Prashanth, N. Korda, and R. Munos. Concentration bounds for temporal difference learning with linear function approximation: the case of batch data and uniform sampling. In *Machine Learning*, volume 110, pages 559–618, 2021.

[18] D. Ruppert. Stochastic approximation. In *Handbook of Sequential Analysis*, pages 503–529, 1991.

[19] R. Schapire and M. K. Warmuth. On the Worst-Case Analysis of Temporal-Difference Learning Algorithms. In *Machine Learning*, volume 22, pages 95–121, 2004.

[20] R. Srikant and L. Ying. Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning. In *Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 2803–2830. PMLR, Jun 2019.

[21] R. S. Sutton. Learning to Predict by the Methods of Temporal Differences. In *Machine Learning*, volume 3, pages 9–44, 1988.

[22] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018. ISBN 0-262-19398-1.

[23] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. In *IEEE Transactions on Automatic Control*, volume 42, pages 674–690, 1997.

[24] G. Wang and G. B. Giannakis. Finite-Time Error Bounds for Biased Stochastic Approximation with Applications to Q-Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 3015–3024. PMLR, 2020.

[25] S. Zhang, Z. Zhang, and S. T. Maguluri. Finite Sample Analysis of Average-Reward TD Learning and Q-

Learning. In *Advances in neural information processing systems (NeurIPS)*, 2021.