Finite time analysis of temporal difference learning with linear function approximation: the tail averaged case

Gandharv Patil [1] **Prashanth L.A.** [2] **Doina Precup** [1]

Abstract

In this paper, we study the finite-time behaviour of temporal difference (TD) learning algorithms when combined with tail-averaging, and present instance dependent bounds on the parameter error of the tail-averaged TD iterate. Our error bounds hold in expectation as well as with high probability, exhibit a sharper rate of decay for the initial error (bias), and are comparable with existing bounds in the literature.

1. Introduction

Temporal difference (TD) (Sutton, 1988) learning is an efficient and easy to implement stochastic approximation algorithm used for evaluating the long-term performance of a decision policy, as quantified by the value function. The algorithm predicts the value function using a single sample path obtained by simulating the Markov decision process with a given policy.

Analysis of TD algorithms is challenging, and researchers have devoted significant effort in studying its asymptotic properties (Tsitsiklis & Van Roy, 1997; Pineda, 1997; Schapire & Warmuth, 2004; Jaakkola et al., 1994). In recent years, there has been an interest in characterising the finite-time behaviour of TD, and several papers (Dalal et al., 2018; Prashanth et al., 2021; Bhandari et al., 2018; Lakshminarayanan & Szepesvari, 2018; Chen et al., 2020) have tackled this problem under various assumptions.

The goal of this paper is to study the finite-time performance of TD with tail-averaging and a constant step-size. A closely related alternative is iterate averaging, proposed independently in (Polyak & Juditsky, 1992; Ruppert, 1991) for general stochastic approximation schemes. A shortcoming of iterate averaging is that the initial error, i.e., rate at which the initial point of the algorithm, is forgotten at a slower rate than the non-averaged case. In practical implementations, one usually performs averaging after a sufficient number of iterations have been performed, and we refer to such a scheme as 'tail-averaging'. Such an approach has been explored in the context of least-squares problems in (Jain et al., 2018).

We consider TD with tail-averaging under i.i.d. sampling: a model previously considered in (Bhandari et al., 2018; Dalal et al., 2018). Concretely, our contributions are as follows: First, we provide a finite-time bound in expectation that establishes a $O(1/t)$ rate of convergence for tail-averaged TD, where t is the number of update iterations. Second, we provide a high-probability bound that establishes exponential concentration of tail-averaged TD around the fixed point of the projected Bellman equation.

Related work. Several recent works contribute finite-time bounds for TD with linear function approximation. In (Bhandari et al., 2018; Prashanth et al., 2021; Lakshminarayanan & Szepesvari, 2018), 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. A particular advantage with our bounds 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, the authors in (Dalal et al., 2018) provide a $O(1/t^{\alpha})$ bound in expectation, where α is a parameter that is restricted to be within $(0,1)$. This bound is for a universal step-size choice, while we require the knowledge of a minimum eigenvalue of the matrix that is used to define the projected TD fixed point. Such information is assumed to be available in (Prashanth et al., 2021; Bhandari et al., 2018; Chen et al., 2020; Lakshminarayanan & Szepesvari, 2018). Finally, high-probability bounds for TD have been derived in (Dalal et al., 2018; Prashanth et al., 2021). In comparison to these works, the high-probability bound that we derive is easily interpretable, and exhibits better concentration properties.

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 present our main results and compare them with previously known re-

[1]McGill University, Montreal, QC, Canada [2]Indian Institute of Technology Madras, Chennai, Tamil Nadu, India. Correspondence to: Gandharv Patil <gandharv.patil@mail.mcgill.ca>.

Workshop on Reinforcement Learning Theory, 38th International Conference on Machine Learning, 2021.

sults in the literature. Finally, in Section 4, 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. Further, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the per step reward, and $\beta \in (0,1]$ is the discount factor. A stationary randomized 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:

$$V^\pi(s_0) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \beta^t r(s_t, a_t) | S_0 = s_0 \right], \tag{1}$$

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

$$(\mathcal{T}^\pi V)(s) = \mathbb{E}^{\pi, P} \Big[r(s,a) + \beta V(s') \Big], \tag{2}$$

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 (Sutton & Barto, 1998), where the value function $V^\pi(s)$, for any $s \in \mathcal{S}$, is approximated as follows:

$$V^\pi(s) \approx \widetilde{V}(s;\theta) := \phi(s)^\top \theta. \tag{3}$$

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 $\mathcal{S} = \{1, 2, \ldots, n\}$, $\widetilde{V}(\cdot;\theta) \in \mathbb{R}^d$ can be expressed compactly as:

$$\widetilde{V}(\theta) = \underbrace{\begin{bmatrix} \phi_1(1) & \phi_2(2) & \ldots & \phi_d(1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(n) & \phi_2(n) & \ldots & \phi_d(n) \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}}_{\theta}, \tag{4}$$

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:

$$\mathcal{B} := \{ \Phi\theta \mid \theta \in \mathbb{R}^d \} \tag{5}$$

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.

Before describing the projected Bellman equation we will define a weighted Euclidean norm on $\mathbb{R}^d$. For a symmetric positive definite matrix A, define the inner product $\langle x, y \rangle_A = x^\top A y$ and the associated norm $\|x\|_A = \sqrt{x^\top A x}$.

Now, the projected Bellman equation is given as:

$$\Phi\theta^\star = \Pi \mathcal{T}^\pi (\Phi\theta^\star). \tag{6}$$

In (6), Π is the orthogonal projection onto $\mathcal{B}$. Assuming the matrix Φ has full column rank, it can be shown that $\Pi = \Phi(\Phi^\top D \Phi)^{-1} \Phi^\top D$, where $D = \text{diag}(\rho(1), \ldots, \rho(n)) \in \mathbb{R}^{n \times n}$ denote the 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 (A1)).

Next, the approximate TD solution θ^* for (6) is given by:

$$A\theta^\star = b, \text{ or, equivalently } \theta^\star = A^{-1} b, \tag{7}$$

where

$$A := \Phi^\top D (\mathbf{I} - \beta P) \Phi, \quad b := \Phi^\top D \mathcal{R}, \tag{8}$$

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

2.2. Temporal Difference (TD) Learning

Temporal difference (TD) (Sutton & Barto, 1998) algorithms are a class of stochastic approximation methods used for solving the projected linear system given in (6). 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 for the parameters is given as follows:

$$\theta_t = \theta_{t-1} + \gamma f_t(\theta_{t-1}), \tag{9}$$

where $f_t(\theta_{t-1}) \triangleq (r_t + \beta \theta_{t-1}^\top \phi(s_t') - \theta_{t-1}^\top \phi(s_t)) \phi(s_t)$, and 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:

$$\theta_t = \Gamma(\theta_{t-1} + \gamma f_t(\theta_{t-1})). \tag{10}$$

In (10), 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 from Tsitsiklis & Van Roy (1997) tells us that for any $\theta \in \mathbb{R}^d$, the function $f(\theta)$ has a well defined

steady-state expectation given by: .

$$\mathbb{E}^{\rho,P}[f(\theta)] = \sum_{s,s'\in\mathcal{S}} \rho(s)\Big((P(s'|s)(r(s,s') + \beta\theta^\top\phi(s')) - \theta^\top\phi^\top(s))\phi(s)\Big) \quad (11)$$

$$= -A\theta + b, \quad (12)$$

where A and b are as defined in (8).

Using the fact that $\sum_{s,s'\in\mathcal{S}} P(s'|s)(r(s,s') + \beta\theta^\top\phi(s'_t)) = \mathcal{T}^\pi\Phi\theta$, we can rearrange (11) as follows:

$$\mathbb{E}^{\rho,P}[f(\theta)] = \Phi^\top D(\mathcal{T}^\pi\Phi\theta - \Phi\theta). \quad (13)$$

As a result, the mean behaviour of TD algorithm can be characterised using the following update iteration:

$$\theta_t = \theta_{t-1} + \gamma\Big(\Phi^\top D(\mathcal{T}^\pi(\Phi\theta_{t-1}) - \Phi\theta_{t-1})\Big) \quad (14)$$

$$= \theta_{t-1} + \gamma\mathbb{E}^{\rho,P}[f(\theta_{t-1})]. \quad (15)$$

The aforementioned characterisation of TD's behaviour is of particular importance as it forms the basis of our analysis.

2.2.1. Tail-averaged TD

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 $\{\theta_k, \ldots, \theta_t\}$, computed as follows:

$$\theta_{k+1,N} = \frac{1}{N}\sum_{i=k+1}^{k+N}\theta_i, \quad (16)$$

where $N = t - k$.

The tail-averaged TD algorithm is as follows:

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 $t = 0, 1, \ldots$ **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. Main Results

Before presenting our results, we would like to enlist the key assumptions under which we conduct our analysis.

(A1). The Markov chain underlying the policy π is irreducible.

This assumption ensures the existence of the stationary distribution for the Markov chain underlying policy π.

We study the non-asymptotic behaviour of the tail-averaged TD algorithm under the i.i.d observation model, where the sequence of states observed by the algorithm are drawn from the stationary distribution of the underlying Markov chain in an i.i.d. fashion, i.e., satisfying the following assumption:

(A2). The samples in the tuple $\{s_t, r_t, s'_t\}_{t\in\mathbb{N}}$ are independently and identically drawn from the Markov chain's stationary distribution given as follows:

$$\rho(s)P(s'|s), \quad (17)$$

where at time t the state s_t is drawn from the stationary distribution ρ, and the next state s' is drawn from $P(\cdot|s_t)$

(A3). For all $s \in \mathcal{S}$, $\|\phi(s)\|_2 \leq \Phi_{\max} < \infty$.

(A4). For all $s \in \mathcal{S}$, and $a \in \mathcal{A}$, $|r(s,a)| \leq R_{\max} < \infty$.

(A5). The matrix Φ has full column rank.

(A6). For the projected TD update in (10), the set $\mathcal{C} \triangleq \{\theta \in \mathbb{R}^d | \|\theta\|_2 \leq H\}$ used for projection through Γ satisfies $H > \frac{\|\bar{b}\|_2}{\mu}$.

Assumption (A3) and Assumption (A4) are boundedness requirements on the underlying features and rewards, and are common in the finite time analysis of TD algorithm, see (Bhandari et al., 2018; Prashanth et al., 2021). (A5) requires the columns of the feature matrix Φ to be linearly independent, in turn ensuring the uniqueness of the TD solution θ^*. Moreover, this assumption ensures that the minimum eigenvalue μ of the matrix A defined in (8) is strictly positive.

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

Theorem 1 (Bound in expectation). *Assume (A1)–(A5). For any step size $\gamma \leq \gamma_{max} = \frac{\mu}{(1+\beta)^2\Phi_{max}^4}$, the expected error of the tail-averaged iterate $\theta_{k+1,N}$ is bounded as follows:*

$$\mathbb{E}[\left\|\theta_{k+1,N} - \theta^\star\right\|_2^2] \leq \left(1 + \frac{4}{\gamma\mu}\right)\left(\frac{2e^{(-k\gamma\mu)}}{\gamma\mu N^2}\mathbb{E}[\|\theta_0 - \theta^\star\|_2^2] + \frac{2\sigma^2}{\mu N}\right), \quad (18)$$

where $N = t - k$, and $\sigma = (R_{max} + (1+\beta)\Phi_{max}^2\|\theta^\star\|_2)$.

Proof. See Appendix C.4. □

A few remarks are in order.

Remark 1. The first term on the RHS of (18) 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 2. In (Lakshminarayanan & Szepesvari, 2018), the authors consider iterate averaging in 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 (18) exhibits an exponential decay, while the corresponding decay is of order $O(1/t)$ in (Lakshminarayanan & Szepesvari, 2018). The second term in their result as well as in (18) 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 view.

Remark 3. In (Dalal et al., 2018), the authors derive a bound in expectation for the last iterate of TD with a universal step-size choice that is diminishing, say $\gamma_t = 1/t^\alpha$ for $\alpha \in (0,1)$. Under this stepsize, the authors in (Dalal et al., 2018) obtain an order $O(1/t^\alpha)$, where $\alpha < 1$. In comparison, our step-size choice is constant, and more importantly, requires the knowledge of the minimum eigenvalue μ of the matrix A. While we obtain a $O(1/t)$ bound, it assumes more information than (Dalal et al., 2018). However, some knowledge of the underlying eigenvalues, either the min or the max, is assumed to be available in previous works such as (Lakshminarayanan & Szepesvari, 2018; Bhandari et al., 2018).

Remark 4. A closely related result under comparable assumptions is Theorem 2 of (Bhandari et al., 2018). This result provides two bounds corresponding to constant and diminishing stepsizes, 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 (18) 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 (Bhandari et al., 2018) requires projection, with an assumption similar to (A6) specified below.

Remark 5. Another closely related result is Theorem 4.4 of (Prashanth et al., 2021), where the authors analyse TD with linear function approximation, with input data from a batch of samples. The analysis there can be easily extended 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 (18), the initial error in our bound is forgotten much faster. A similar observation also holds w.r.t. to the bound in the recent work (Chen et al., 2020), 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 (Bhandari et al., 2018). In this model, we assume that the underlying Markov chain is uniformly ergodic, which intuitively translates to a fast mixing rate. Since we focus on a finite irreducible Markov chain (see (A1)), it is enough to an aperiodicity requirement to ensure uniform ergodicity. Using the fast mixing assumption, one can obtain a bound for the Markov noise model by incorporating an additional term that controls the error until the Markov chain has mixed. The rest of the bound would follow from the analysis of the i.i.d. noise case. We omit the details.

Next, we turn to providing a bound that holds with high probability the parameter error $\left\|\theta_{k+1,N} - \theta^\star\right\|_2^2$ 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 formalized in the assumption below.

Theorem 2 (High-probability bound). *Assume (A1)–(A6). For any step size $\gamma \leq \gamma_{max} = \frac{\mu}{(1+\beta)^2\Phi_{max}^4}$, and any $\delta \in (0,1)$, the parameter error of the tail-averaged iterate $\theta_{k+1,N}$ as per (10) satisfies the following bound with $N = t - k$:*

$$P\left(\left\|\theta_{k+1,N} - \theta^\star\right\|_2^2 \leq \mathcal{K}(n)\right) \geq 1 - \delta, \text{ where}$$

$$\mathcal{K}(n) = \frac{\sigma^2 e^{(-k\gamma\mu)}}{\mu N^2} + \left[1 + \frac{4}{\gamma\mu}\right]\left[\frac{2e^{(-k\gamma\mu)}}{\gamma\mu N^2}\mathbb{E}[\|\theta_0 - \theta^\star\|_2^2] + \frac{2\sigma^2}{\mu N}\right],$$

with σ as defined in Theorem 1.

Proof. See Appendix D. □

Remark 7. High-probability bounds for TD algorithm have been derived earlier in (Prashanth et al., 2021; Dalal et al., 2018). In comparison to Theorem 4.2 of (Prashanth et al., 2021), 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 (Dalal et al., 2018), we note that our bound requires projection, and the knowledge of μ, however it does exhibit a $O(1/t)$ rate. The result by (Dalal et al., 2018) (Theorem 3.6) is of the form $O(1/t^\lambda)(\frac{1}{t^\lambda})$ where λ is related to the μ, and hence cannot be guaranteed to be of order $O(1/t)$.

4. Conclusions

In this short paper, we have presented the finite sample analysis of tail-averaged TD algorithm. Our bounds are easy to

interpret, and improve the previously known results. Additionally, our results highlight the trade-off between sample complexity and domain knowledge, wherein we are able to obtain a sharp characterisation of TD learning's behaviour by assuming the knowledge of the problem specific information. Finally, analysing TD without these problem specific assumptions is an interesting future research direction.

References

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1691–1692. PMLR, 2018. URL http://proceedings.mlr.press/v75/bhandari18a.html.

Chen, S., Devraj, A., Busic, A., and Meyn, S. Explicit mean-square error bounds for monte-carlo and linear stochastic approximation. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4173–4183. PMLR, 26–28 Aug 2020. URL http://proceedings.mlr.press/v108/chen20e.html.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analyses for TD(0) with function approximation. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 6144–6160. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16392.

Jaakkola, T. S., Jordan, M. I., and Singh, S. P. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Comput.*, 6(6):1185–1201, 1994. doi: 10.1162/neco.1994.6.6.1185. URL https://doi.org/10.1162/neco.1994.6.6.1185.

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

Lakshminarayanan, C. and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1347–1355. PMLR, 09–11 Apr 2018. URL http://proceedings.mlr.press/v84/lakshminarayanan18a.html.

Pineda, F. J. Mean-field theory for batched-td(l). *Neural Comput.*, 9(7):1403–1419, 1997. doi: 10.1162/neco.1997.9.7.1403. URL https://doi.org/10.1162/neco.1997.9.7.1403.

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

Prashanth, L. A., Korda, N., and Munos, R. Concentration bounds for temporal difference learning with linear function approximation: the case of batch data and uniform sampling. *Mach. Learn.*, 110(3):559–618, 2021. doi: 10.1007/s10994-020-05912-5. URL https://doi.org/10.1007/s10994-020-05912-5.

Ruppert, D. Stochastic approximation. *Handbook of Sequential Analysis*, pp. 503–529, 1991.

Schapire, R. and Warmuth, M. K. On the worst-case analysis of temporal-difference learning algorithms. *Machine Learning*, 22:95–121, 2004.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3:9–44, 1988. doi: 10.1007/BF00115009. URL https://doi.org/10.1007/BF00115009.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1998. ISBN 0-262-19398-1. URL http://www.cs.ualberta.ca/%7Esutton/book/ebook/the-book.html.

Tsitsiklis, J. N. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. URL http://www.stanford.edu/~bvr/psfiles/td.pdf.

A. Preliminaries

Let $\mathcal{F}_t$ denote the sigma-field generated by $\theta_0 \dots \theta_t \quad t \geq 0$, and let:

$$f_t(\theta) \triangleq (r_t + \beta \theta^\top \phi(s'_t) - \theta^\top \phi(s_t))\phi(s_t). \tag{19}$$

Moreover, recall for a finite set of states, the state feature can be represented as a $(|\mathcal{S}| \times d)$-matrix, where $\Phi \triangleq (\phi_1(s_1), \dots, \phi_d(s_n))$, and the next state feature can be represented as $\Phi' \triangleq (\phi_1(s'_1), \dots, \phi_d(s'_n))$, similarly $R_t = (r_1, \dots, r_n)$ can be represented as a $|\mathcal{S}| \times 1$ vector.

According to the characterisation of TD's steady-state behaviour in (15), it's final solution can be written as follows:

$$\theta = A^{-1}b, \tag{20}$$

where $A = \frac{1}{|\mathcal{S}|}(\Phi^\top \Phi - \beta \Phi^\top \Phi')$, and $b = \frac{1}{|\mathcal{S}|}\Phi R$

We can also use (15) to rewrite the TD update in eq. (9) as:

$$\theta_t = \theta_{t-1} + \gamma f_t(\theta_{t-1}), \tag{21}$$

$$= \theta_{t-1} + \gamma\bigg(-A\theta_{t-1} + b + \Delta M_t \bigg), \tag{22}$$

where $\Delta M_t = f_t(\theta_{t-1}) - \mathbb{E}^{\rho,P}[f_t(\theta_{t-1})|\mathcal{F}_{t-1}]$ is a martingale sequence with $f(\cdot)$ defined as in (19)

B. Bias-variance decomposition of the non-asymptotic error

We will first define the centered update rule as: $z_t = \theta_t - \theta^\star$, and rewrite eq. (9) in terms of the centered error z_t Let us first rewrite the TD update rule:

$$z_t = \theta_{t-1} - \theta^\star - \gamma\bigg(r_t + \beta \theta_{t-1}^\top \phi(s'_t) - \theta_{t-1}^\top \phi(s_t) \bigg)\phi(s_t), \tag{23}$$

$$= \bigg(\mathbf{I} - \gamma(\phi(s_t) - \beta\phi(s'_t))\phi(s_t)^\top \bigg) z_{t-1} - \gamma \Delta M_t, \tag{24}$$

$$= C_t z_{t-1} + \gamma \Delta M_t \tag{25}$$

where $C_t = (\mathbf{I} - \gamma(\phi(s_t) - \beta\phi(s'_t)) \cdot \phi(s_t)^\top)$.

Unrolling eq. (25) until the initial θ, we get:

$$z_t = \prod_{k=0}^{t} C_k z_0 + \gamma \sum_{k=1}^{t} \bigg(\prod_{j=k+1}^{t} C_j \bigg) \Delta M_k, \tag{26}$$

$$z_t = z_t^{\text{bias}} + z_t^{\text{variance}}, \tag{27}$$

$$\text{where, } z_t^{\text{bias}} \triangleq \prod_{k=0}^{t} C_k z_0, \text{ and, } z_t^{\text{variance}} \triangleq \gamma \sum_{k=1}^{t} \bigg(\prod_{j=k+1}^{t} C_j \bigg) \Delta M_k]$$

We are interested finding an analytic expression for the non-asymptotic error given by: $\mathbb{E}[\|z_t\|_2]$.

Now, from the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we get:

$$\mathbb{E}\,\|z_t\|^2 \leq 2\mathbb{E}\left\|z_t^{\text{bias}}\right\|^2 + 2\mathbb{E}\left\|z_t^{\text{variance}}\right\|^2. \tag{28}$$

Therefore, a bound on $\|z_t\|_2$ can obtained by by bounding individual terms in eq. (28).

B.1. Bounding the bias error

For bounding the bias error, we will consider the first term in eq. (28).

We first present the following intermediate results:

Lemma 1. *With $m = \phi^\top(s_t)\phi(s_t) - \beta\phi(s_t)^\top\phi(s_t')$, we have the following*

$$\mathbf{I} - \gamma\Big(\mathbb{E}[m] + \mathbb{E}[m]^\top - \gamma\mathbb{E}[m^\top]\mathbb{E}[m]\Big) \leq 1 - \gamma(2\mu - \gamma(1+\beta)^2\Phi_{max}^4). \tag{29}$$

Proof. To prove this lemma, we will separately bound the two terms inside the parenthesis of the LHS of eq. (29).

Bounding Term 1:

$$\mathbb{E}[m] + \mathbb{E}[m]^\top \overset{a}{=} \frac{1}{T}(2\Phi^\top\Phi - \beta(\Phi^\top\Phi' + \Phi'^\top\Phi)), \tag{30}$$

where (a) is because $\sum_{i=1}^{|\mathcal{S}|}\phi(s_i)\phi(s_i)^\top = \Phi^\top\Phi$

Now,

$$\begin{aligned}\lambda_{\min}(2\Phi^\top\Phi - \beta(\Phi^\top\Phi' + \Phi'^\top\Phi)) &= \lambda_{\min}((\Phi^\top\Phi - \beta(\Phi^\top\Phi')) + (\Phi^\top\Phi - \beta\Phi'^\top\Phi)),\\ &= \lambda_{\min}(T(A + A^\top)),\\ &\overset{a}{\geq} 2T\mu,\end{aligned} \tag{31}$$

where (a) follows from assumption (A5).

Bounding Term 2:

$$\mathbb{E}[m^\top]\mathbb{E}[m] \overset{a}{=} \frac{1}{T}\Big(\Phi^\top\mathrm{Tr}(\Phi^\top\Phi)\Phi - \beta(\Phi^\top\mathrm{Tr}(\Phi^\top\Phi)\Phi' + \Phi'^\top\mathrm{Tr}(\Phi^\top\Phi)\Phi) + \beta^2\Phi'^\top\mathrm{Tr}(\Phi^\top\Phi)\Phi')\Big), \tag{32}$$

where (a) is because $\sum_{i=1}^{|\mathcal{S}|}\phi(s_i)\phi(s_i)^\top = \Phi^\top\Phi$

Let $\Delta = \mathrm{Tr}(\Phi^\top\Phi)$, therefore,

$$\begin{aligned}&\Phi^\top\Delta\Phi' + \Phi'^\top\Delta\Phi \overset{a}{\leq} 2\Phi_{\max}^4, &(33)\\ &\implies -2\Phi_{\max}^4 \leq \Phi^\top\Delta\Phi' + \Phi'^\top\Delta\Phi, &(34)\\ &\implies \Phi^\top\Delta\Phi - \beta(\Phi^\top\Delta\Phi' + \Phi'^\top\Delta\Phi) + \beta^2\Phi'^\top\Delta\Phi' \leq (1+2\beta+\beta^2)\Phi_{\max}^4, &(35)\\ &= (1+\beta)^2\Phi_{\max}^4, &(36)\end{aligned}$$

where (a) follows from the boundedness of features in assumption (A3).

Finally, combining eqs. (31) and (36), we get

$$\mathbf{I} - \gamma\Big(\mathbb{E}[m] + \mathbb{E}[m]^\top\Big) + \gamma^2\mathbb{E}[m^\top]\mathbb{E}[m] \leq 1 - \gamma(2\mu - \gamma(1+\beta)^2\Phi_{\max}^4). \tag{37}$$

□

Lemma 2. *With $\gamma \leq \gamma_{max} = \frac{\mu}{(1+\beta)^2\Phi_{max}^4}$, the following bound holds*

$$\mathbb{E}\Big[\Big(\mathbf{I} - \gamma(\phi^\top(s_t)\phi(s_t) - \beta\phi(s_t)^\top\phi(s_t'))\Big)^\top\Big(\mathbf{I} - \gamma(\phi^\top(s_t)\phi(s_t) - \beta\phi(s_t)^\top\phi(s_t'))\Big)\Big] \leq 1 - \gamma\mu.$$

Proof. For ease of exposition, let $m = \phi^\top(s_t)\phi(s_t) - \beta\phi(s_t)^\top\phi(s'_t)$, therefore, we get

$$\mathbb{E}\Big[\Big(\mathbf{I}-\gamma m\Big)^\top \cdot \Big(\mathbf{I}-\gamma m\Big)\Big] = \mathbb{E}\Big[\Big(\mathbf{I}-\gamma m-\gamma m^\top+\gamma^2 m^\top m\Big)\Big], \tag{38}$$

$$= \mathbf{I}-\gamma\Big(\mathbb{E}[m]+\mathbb{E}[m]^\top\Big)+\gamma^2\mathbb{E}[m^\top]\mathbb{E}[m], \tag{39}$$

$$\overset{a}{\leq} 1-\gamma(2\mu-\gamma(1+\beta)^2\Phi^4_{\max}), \tag{40}$$

$$\overset{b}{\leq} 1-\mu\gamma, \tag{41}$$

where (a) follows from lemma 1, and c follows as per definition of γ. Note that $1-\gamma\mu \geq 0 \implies (1-\gamma\mu)^{\frac{1}{2}} \leq 1-\frac{\gamma\mu}{2}$. Therefore,

$$\mathbb{E}[(\mathbf{I}-\gamma m)] \leq 1-\frac{\gamma\mu}{2}. \tag{42}$$

□

Lemma 3. *For any step size $\gamma \leq \gamma_{max}$, the bias or the initial error of the TD update is upper bounded as*

$$\mathbb{E}[\|z_t^{bias}\|_2^2] \leq \exp(-\gamma\mu t)\mathbb{E}[\|z_0^{bias}\|_2^2]. \tag{43}$$

Proof. For ease of exposition, let $m = \phi^\top(s_t)\phi(s_t) - \beta\phi(s_t)^\top\phi(s'_t)$, therefore, we get

$$\mathbb{E}[\|z_t^{\text{bias}}\|_2^2] = \mathbb{E}\Big[z_{t-1}^\top(\mathbf{I}-m)^\top(\mathbf{I}-k)z_{t-1}\Big], \tag{44}$$

$$= \mathbb{E}\Big[z_{t-1}^\top\mathbb{E}\Big[(\mathbf{I}-m)^\top(\mathbf{I}-m)\Big|\mathcal{F}_{t-1}\Big]z_{t-1}\Big], \tag{45}$$

$$\overset{a}{\leq} (1-\mu\gamma)\mathbb{E}[\|z_{t-1}^{\text{bias}}\|_2^2], \tag{46}$$

$$\overset{b}{=} (1-\gamma\mu)^t\mathbb{E}[\|z_0^{\text{bias}}\|_2^2], \tag{47}$$

$$\leq \exp(-t\mu\gamma)\mathbb{E}[\|z_0^{\text{bias}}\|_2^2], \tag{48}$$

where a follows from lemma 2, and b is the recursive application of the bound in eq. (46). □

B.2. Bounding the variance error

Lemma 4. *With $\prod_{j=i+1}^t C_j \triangleq B_{i+1:t}$ such that $B_{i+1:t} = (\mathbf{I}-m_{i+1})(\mathbf{I}-m_{i+2})\ldots(\mathbf{I}-m_t)$, for $\forall \mathbf{x} \in \mathbb{R}^d$, where $m_t = \phi^\top(s_t)\phi(s_t) - \beta\phi(s_t)^\top\phi(s'_t)$, we have:*

$$\mathbb{E}[\gamma\sum_{i=0}^t \|B_{i+1:t}\mathbf{x}\|_2^2] \leq \gamma\sum_{i=0}^t(1-\gamma\mu)^i\|\mathbf{x}\|_2^2. \tag{49}$$

Proof.

$$\mathbb{E}[\gamma\sum_{i=0}^{t}\|B_{i+1:t}\mathbf{x}\|_2^2] = \mathbb{E}\left[\gamma\sum_{i=0}^{t}\left(B_{i+1:t}\mathbf{x}\right)^\top\left(B_{i+1:t}\mathbf{x}\right)\right], \tag{50}$$

$$= \gamma\sum_{i=0}^{t}\mathbb{E}\left[\left(\mathbf{x}^\top B_{i+1:t}^\top B_{i+1:t}\mathbf{x}\right)\right], \tag{51}$$

$$\overset{a}{=} \gamma\sum_{i=0}^{t}\mathbb{E}\left[\mathbb{E}\left[\left(\mathbf{x}^\top B_{i+1:t}^\top B_{i+1:t}\mathbf{x}\right)\middle|\mathcal{F}_{i-1}\right]\right], \tag{52}$$

$$\overset{b}{=} \gamma\sum_{i=0}^{t}\mathbb{E}\left[\left(\mathbf{x}^\top B_{i+2:t}^\top\mathbb{E}\left[(\mathbf{I}-m_{i+1})^\top(\mathbf{I}-m_{i+1})|\mathcal{F}_{i-1}\right]B_{i+2:t}\mathbf{x}\right], \tag{53}$$

$$\overset{c}{\leq} \gamma\sum_{i=0}^{t}(1-\gamma\mu)\mathbb{E}\left[\left(\mathbf{x}^\top B_{i+2:t}^\top B_{i+1:t}\mathbf{x}\right)\middle|\mathcal{F}_{i-1}\right], \tag{54}$$

$$\overset{d}{=} \gamma\sum_{i=0}^{t}(1-\gamma\mu)^i\|\mathbf{x}\|_2^2, \tag{55}$$

$$\overset{e}{\leq} \frac{\|\mathbf{x}\|_2^2}{\mu}, \tag{56}$$

Where (a) follows from tower-rule of conditional expectation, (b) follows from the definition of $B_{i+1:t}$, (c) follows from lemma 2, (d) follows from unrolling the recursion, and (e) follows from the fact that $\sum_{i=1}^{t}(1-\gamma\mu)^i \leq \frac{1}{\gamma\mu}$.

□

Lemma 5. *With* $\mathbb{E}[\|\Delta M_t\|_2^2|\mathcal{F}_{t-1}] \leq \sigma$, *where* $\sigma = (R_{max}+(1+\beta)\Phi_{max}^2\|\theta^\star\|_2^2)$, *we have*

$$\mathbb{E}[\left\|z_t^{variance}\right\|^2] \leq \frac{\sigma^2}{\mu}. \tag{57}$$

Proof. Recall that from eq. (27) we have $z_t^{\text{variance}} \triangleq \gamma\sum_{i=1}^{t}\left(\prod_{j=i+1}^{t}C_j\right)\Delta M_i$

$$\mathbb{E}[\left\|z_t^{\text{variance}}\right\|_2^2] = \gamma\mathbb{E}\left[\sum_{i=1}^{t}\left(\left(\prod_{j=i+1}^{t}C_j\right)\Delta M_i\right)^\top\left(\left(\prod_{j=i+1}^{t}C_j\right)\Delta M_i\right)\right], \tag{58}$$

$$\overset{a}{\leq} \frac{1}{\mu}\mathbb{E}[\|\Delta M_t\|_2^2|\mathcal{F}_{t-1}], \tag{59}$$

$$\overset{b}{\leq} \frac{\sigma^2}{\mu}, \tag{60}$$

where (a) follows from lemma 4, and (b) follows from assumption (A4). □

C. Non-Asymptotic analysis of TD with tail-averaging

Tail averaging or suffix averaging refers to returning final few iterates of the optimisation process to improve its variance properties. In particular, if a parameter is updated for t iterations, we can average the centered updates z N-times starting from s iterations as follows:

$$z_{k+1,N} = \frac{1}{N}\sum_{i=k+1}^{k+N}z_i, \tag{61}$$

$$\overset{a}{=} \frac{1}{N}\sum_{i=k+1}^{k+N}\left(z_i^{\text{bias}}+z_i^{\text{variance}}\right), \tag{62}$$

$$= z_{k+1,N}^{\text{bias}}+z_{k+1,N}^{\text{variance}}, \tag{63}$$

where (a) is from eq. (27), and $N = t-k$.

C.1. Bias-Variance decomposition of the tail-averaged iterate

Lemma 6. *Define* $\prod_{j=i+1}^{t} M_j \triangleq B_{i+1:t}$ *such that* $B_{i+1:t} = (\mathbf{I} - m_{i+1})(\mathbf{I} - m_{i+2})\dots(\mathbf{I} - m_t)$, *where* $m_t = \phi^\top(s_t)\phi(s_t) - \beta\phi(s_t)^\top\phi(s_t')$. *Then for* $\forall j > i$ *we have:*

$$\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[z_i^\top z_j] \leq \frac{2}{\gamma\mu}\sum_{i=0}^{n}\mathbb{E}[\|z_i\|^2]. \tag{64}$$

Proof.

$$\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[z_i^\top z_j] = \sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[z_i^\top(B_{j:i+1}z_i + \sum_{k=i+1}^{j}B_{j:k+1}f_j(\theta^\star))], \tag{65}$$

$$\overset{a}{=} \sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[z_i^\top B_{j:i+1}z_i], \tag{66}$$

$$\overset{b}{=} \sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[z_i^\top(\prod_{l=j}^{i+1}M_l)z_i], \tag{67}$$

$$\overset{c}{=} \sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[z_i^\top(\mathbb{E}[(\mathbf{I}-m_j|\mathcal{F}_{j-1}](\prod_{l=j-1}^{i+1}M_l))z_i], \tag{68}$$

$$\overset{d}{\leq} \sum_{i=0}^{n-1}\sum_{j=i+1}^{n}(1-\frac{\gamma\mu}{2})\mathbb{E}[z_i^\top(\prod_{l=j-1}^{i+1}M_l)z_i], \tag{69}$$

$$\overset{e}{\leq} \sum_{i=0}^{n-1}\sum_{j=i+1}^{n}(1-\frac{\gamma\mu}{2})^{i-j}\mathbb{E}[\|z_i\|^2], \tag{70}$$

$$\leq \sum_{i=0}^{n-1}\mathbb{E}[\|z_i\|^2]\sum_{j=i+1}^{\infty}(1-\frac{\gamma\mu}{2})^{j-i}, \tag{71}$$

$$\overset{f}{\leq} \frac{2}{\gamma\mu}\sum_{i=0}^{n}\mathbb{E}[\|z_i\|^2], \tag{72}$$

where (a) follows from the fact that $\mathbb{E}[\Delta M_t] = 0$, (b) follows from the definition, (c) follows from the tower rule of conditional expectation, (d) follows lemma 2, (e) follows from recursive application of (d), and (f) follows from the summation of the geometric series.

□

C.2. Bounding the bias error

Lemma 7. *For any step size* $\gamma \leq \gamma_{max}$, *the bias or the initial error of tail-averaged TD update is upper bounded as*

$$\mathbb{E}[\|z_t^{bias}\|^2] \leq \frac{1}{\gamma\mu N^2}(1-\gamma\mu)^{k+1}(1+\frac{4}{\gamma\mu})\mathbb{E}[\|z_0^{bias}\|^2]. \tag{73}$$

Proof. Before deriving the bound, let us first bound the per-step contraction properties of the TD update in case of the bias

error ($\gamma\sum_{k=1}^{t}\left(\prod_{j=k+1}^{t}C_j\right)\Delta M_k = 0$). For ease of exposition, let $m = \phi^\top(s_t)\phi(s_t) - \beta\phi(s_t)^\top\phi(s_t')$, therefore, we get

$$\mathbb{E}[\|z_t\|^2] = \mathbb{E}\left[z_{t-1}^\top(\mathbf{I}-m_{t-1})^\top(\mathbf{I}-m_{t-1})z_{i-1}\right], \tag{74}$$

$$= \mathbb{E}\left[z_{t-1}^\top\mathbb{E}\left[(\mathbf{I}-m_{t-1})^\top(\mathbf{I}-m_{t-1})\middle|\mathcal{F}_{t-1}\right]z_{t-1}\right], \tag{75}$$

$$\overset{a}{\leq} (1-\mu\gamma)\mathbb{E}[\left\|z_{t-1}^{\text{bias}}\right\|^2], \tag{76}$$

$$\overset{b}{\leq} (1-\mu\gamma)^t\mathbb{E}[\left\|z_0^{\text{bias}}\right\|^2], \tag{77}$$

where a follows from lemma 2, b is the recursive application of the bound in eq. (76)

Coming back to the main quantity of interest,

$$\mathbb{E}[\left\|z_{k+1,N}^{\text{bias}}\right\|^2] = \frac{1}{N^2}\mathbb{E}[\left\|\sum_{i=k+1}^{k+N} z_i^\top z_j\right\|^2], \tag{78}$$

$$\overset{a}{=} \frac{1}{N^2}\left(\sum_{i=k+1}^{k+N}\mathbb{E}[\|z_i\|^2] + 2\sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N}\mathbb{E}[z_i^\top z_j]\right), \tag{79}$$

$$\overset{b}{\leq} \frac{1}{N^2}(1+\frac{4}{\gamma\mu})\sum_{i=k+1}^{k+N}[\left\|z_i^{\text{bias}}\right\|^2], \tag{80}$$

$$\leq \frac{1}{N^2}(1+\frac{4}{\gamma\mu})\sum_{i=k+1}^{\infty}[\left\|z_i^{\text{bias}}\right\|^2], \tag{81}$$

$$\overset{c}{\leq} \frac{1}{N^2}(1+\frac{4}{\gamma\mu})\sum_{i=k+1}^{\infty}(1-\gamma\mu)^i\mathbb{E}[\left\|z_0^{\text{bias}}\right\|^2], \tag{82}$$

$$\overset{d}{=} \frac{1}{\gamma\mu N^2}(1-\gamma\mu)^{k+1}(1+\frac{4}{\gamma\mu})\mathbb{E}[\left\|z_0^{\text{bias}}\right\|^2], \tag{83}$$

where (a) follows by separating the diagonal and "cross" terms in the summation, (b) follows from Lemma 6, (c) follows from eq. (77) and ,(d) follows from the summation of the geometric series. □

C.3. Bounding the variance error

Lemma 8. *For any step size $\gamma \leq \gamma_{max}$, with $\mathbb{E}[\|\Delta M_t\|_2^2|\mathcal{F}_{t-1}] \leq \sigma$, where $\sigma^2 = (R_{max} + (1+\beta)\Phi_{max}^2\|\theta^\star\|_2)$, and setting $z^{bias} = 0$, variance of the tail averaged TD update is bounded as:*

$$\mathbb{E}[\left\|z_{s+1,N}^{variance}\right\|^2] \leq \left(1+\frac{4}{\gamma\mu}\right)\frac{\sigma^2}{\mu N}. \tag{84}$$

Proof.

$$\mathbb{E}[\|z_{k+1,N}^{\text{variance}}\|^2] = \frac{1}{N^2}\mathbb{E}[\left\|\sum_{i=k+1}^{k+N} z_i^\top z_j\right\|^2], \tag{85}$$

$$\overset{a}{=} \frac{1}{N^2}\left(\sum_{i=k+1}^{k+N}\mathbb{E}[\|z_i\|^2] + 2\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[z_i^\top z_j]\right), \tag{86}$$

$$\overset{b}{\leq} \frac{1}{N^2}(1+\frac{4}{\gamma\mu})\sum_{i=k+1}^{k+N}\mathbb{E}[\|z_i\|^2], \tag{87}$$

$$\overset{c}{\leq} \frac{1}{N^2}(1+\frac{4}{\gamma\mu})\sum_{i=k+1}^{k+N}\frac{\sigma^2}{\mu}, \tag{88}$$

$$= \left(1+\frac{4}{\gamma\mu}\right)\frac{\sigma^2}{\mu N}, \tag{89}$$

where (a) follows by separating the diagonal and "cross" terms, (b) follows from argument similar to lemma 6, (c) follows from lemma 5 □

C.4. Final Result

Theorem 3. *For any step size $\gamma \leq \gamma_{max}$, we have*

$$\mathbb{E}[\|\theta_t - \theta^\star\|_2^2] \leq \left(1+\frac{4}{\gamma\mu}\right)\left(\frac{2\exp(-k\gamma\mu)}{\gamma\mu N^2}\mathbb{E}[\|\theta_0-\theta^\star\|_2^2] + \frac{2\sigma^2}{\mu N}\right). \tag{90}$$

Proof.

$$\mathbb{E}[\left\|z_{k+1,N}\right\|_2^2] \overset{a}{\leq} 2\mathbb{E}[\left\|z_{k+1,N}^{\text{bias}}\right\|_2^2] + 2\mathbb{E}[\left\|z_{k+1,N}^{\text{variance}}\right\|_2^2], \tag{91}$$

$$\overset{b}{=} \left(1+\frac{4}{\gamma\mu}\right)\left(\frac{2}{\gamma\mu N^2}(1-\gamma\mu)^{k+1}\mathbb{E}[\left\|z_0^{\text{bias}}\right\|_2^2] + \frac{2\sigma^2}{\mu N}\right), \tag{92}$$

$$\overset{c}{\leq} \left(1+\frac{4}{\gamma\mu}\right)\left(\frac{2\exp(-k\gamma\mu)}{\gamma\mu N^2}\mathbb{E}[\left\|z_0^{\text{bias}}\right\|_2^2] + \frac{2\sigma^2}{\mu N}\right), \tag{93}$$

where (a) follows from the inequality $(a+b)^2 \leq 2a^2+2b^2$, and eq. (63), (b) follows from the Lemma 7 and 8. □

D. High probability bound for non-asymptotic error in TD

Proposition 1. *Under assumption (A3) to (A4), for all $\varepsilon \geq 0$, and $t \geq 1$,*

$$P(\left\|z_{k+1,N}\right\|_2 - \mathbb{E}[\left\|z_{k+1,N}\right\|_2] \leq \exp\left(-\frac{\varepsilon^2}{(R_{max}+(1+\beta)H\phi_{max}^2)^2\sum_{i=1}^t L_i^2}\right), \tag{94}$$

where $L_i \triangleq \frac{\gamma}{N}(\sum_{l=k+1}^{k+N}(1-\gamma\mu)^l)$

Proof. **D.1. Step 1**

$$\|z_t\|_2 - \mathbb{E}[\|z_t\|_2] = \sum_{i=1}^t \mathbb{E}[\|z_t\|_2|\mathcal{F}_i] - \mathbb{E}[\mathbb{E}[\|z_t\|_2|\mathcal{F}_i]|\mathcal{F}_{i-1}], \tag{95}$$

$$= \sum_{i=1}^t g_i - \mathbb{E}[g_i|\mathcal{F}_{i-1}], \tag{96}$$

$$= \sum_{i=1}^t D_i. \tag{97}$$

D.2. Step 2

We need to prove that that functions g_i are Lipschitz continuous in the random innovation at time k with new constant L_k. Towards that end, define $\Theta^i_{t,k+1}(\theta)$ to be the value of the tail-averaged iterate at time t that evolves according to eq. (9) beginning from θ at time i. Therefore,

$$\bar{\Theta}^i_{k+1,N} = \frac{(i-k)\bar{\theta}}{N} + \frac{1}{N}\sum_{j=i}^{k+N}\Theta^i_j(\theta). \tag{98}$$

Let f and f' denote two possible values of the random innovation at time i. Now, setting $\theta = \theta_{i-1} + \gamma f$, and $\theta' = \theta_{i-1} + \gamma f'$ we get:

$$\mathbb{E}[\left\|\bar{\Theta}^i_{k+1}(\bar{\theta}_{i-1},\theta) - \bar{\Theta}^i_{k+1}(\bar{\theta}_{i-1},\theta')\right\|_2] = \frac{1}{N}\mathbb{E}\left[\left\|\sum_{j=k+1}^{s+N}(\Theta^i_j(\theta) - \Theta^i_j(\theta'))\right\|_2\right]. \tag{99}$$

We will now bound the term $\Theta^i_j(\theta) - \Theta^i_j(\theta')$ inside the summation of (99), note that since the projection Γ is non-expansive, we have the following:

$$\mathbb{E}\left[\left\|\Theta^i_j(\theta) - \Theta^i_j(\theta')\right\|_2 |\mathcal{F}_{j-1}\right] \leq \mathbb{E}\left[\left\|\Theta^i_{j-1}(\theta) - \Theta^i_{j-1}(\theta') - \gamma[f_i(\Theta^i_{j-1}(\theta)) - f_i(\Theta^i_{j-1}(\theta'))]\right\|_2 |\mathcal{F}_{i-1}\right]. \tag{100}$$

Expanding random innovation terms, we have

$$\begin{aligned}\Theta^i_{j-1}(\theta) - \Theta^i_j(\theta') - \gamma[f_i(\Theta^i_{j-1}(\theta)) - f_i(\Theta^i_{j-1}(\theta')] &= \Theta^i_{j-1}(\theta) - \Theta^i_{j-1}(\theta') \\ &\quad - \gamma[\phi(s_{i_j})\phi(s_{i_j})^\top - \beta\phi(s_{i_j})\phi(s'_{i_j})^\top][(\Theta^j_{i-1}(\theta)) - (\Theta^j_{i-1}(\theta')], \quad (101)\\ &= [\mathbf{I} - \gamma a_j](\Theta^i_{j-1}(\theta) - \Theta^i_{j-1}(\theta')), \quad (102)\end{aligned}$$

where $a_j \triangleq [\phi(s_{i_j})\phi(s_{i_j})^\top - \beta\phi(s_{i_j})\phi(s'_{i_j})^\top]$

Substituting (102) in (99), and using lemma 1 and 2, we get the following:

$$\mathbb{E}[\left\|\bar{\Theta}^i_{k+1}(\bar{\theta}_{i-1},\theta) - \bar{\Theta}^i_{k+1}(\bar{\theta}_{i-1},\theta')\right\|_2] \leq \frac{1}{N}\left(\sum_{j=k+1}^{s+N}(1-\gamma\mu)^{\frac{j}{2}}\right)\left\|\theta - \theta'\right\|_2. \tag{103}$$

Now, let us substitute f and f' as values of random innovations at time i, then we know that, $\theta = \theta_{i-1} + \gamma f$, and $\theta' = \theta_{i-1} + \gamma f'$, therefore,

$$\begin{aligned}\left|\mathbb{E}[\|\theta - \theta^\star\|_2 |\theta_i = \theta] - \mathbb{E}[\|\theta - \theta^\star\|_2 |\theta_i = \theta']\right| &\overset{a}{\leq} \mathbb{E}[\left\|\Theta^i_t(\theta) - \Theta^i_t(\theta')\right\|_2] \quad (104)\\ &\leq L_i\left\|f - f'\right\|_2, \quad (105)\end{aligned}$$

where (a) follows from Jensen's inequality.

Comparing eq. (103) and eq. (105), it is clear that that g_i is L_i- Lipschtiz in the random innovation at time i, which further implies that D_i is Lipschitz. Moreover, the Lipschitz constant L_i is equal to $\frac{1}{N}\sum_{j=s+1}^{s+N}(1-\gamma\mu)^{\frac{j}{2}}$

D.3. Step 3

Next, we derive a standard martingale concentration bound for the iterate z_t such that $t = \{s+1, s+2, \dots s+N\}$. Note that for any $\lambda > 0$,

$$P(\|z_t\|_2 - \mathbb{E}[\|z_t\|_2] \geq \varepsilon) = P\Big(\sum_{i=1}^{t} D_i \geq \varepsilon\Big), \tag{106}$$

$$\overset{a}{\leq} \exp(-\lambda\varepsilon)\mathbb{E}\Big[\exp(\lambda \sum_{i=1}^{t} D_i)\Big], \tag{107}$$

$$\overset{b}{=} \exp(-\lambda\varepsilon)\mathbb{E}\Big[\exp(\lambda \sum_{i=1}^{t-1} D_i)\mathbb{E}\Big[\exp(\lambda D_t)|\mathcal{F}_{t-1}\Big]\Big], \tag{108}$$

$$\tag{109}$$

where (a) follows from Markov inequality and (b) follows from eq. (95).

Now, let Z be a zero-mean random variables satisfying $|Z| \leq B$ w.p 1, and g be a L- Lipschitz. Letting Z' denote an independent copy of Z and ε a Rademacher random variable, we have,

$$\mathbb{E}[\exp(\lambda g(Z))] = \mathbb{E}[\exp(\lambda(g(Z) - \mathbb{E}[g(Z')])], \tag{110}$$

$$\overset{a}{\leq} \mathbb{E}[\exp(\lambda(g(Z) - g(Z'))], \tag{111}$$

$$\overset{b}{=} \mathbb{E}[\exp(\varepsilon\lambda(g(Z) - g(Z'))], \tag{112}$$

$$\overset{c}{\leq} \mathbb{E}\Big[\exp\Big(\frac{\lambda^2(g(Z) - g(Z'))^2}{2}\Big)\Big], \tag{113}$$

$$\overset{d}{\leq} \mathbb{E}\Big[\exp\Big(\frac{\lambda^2 L^2(Z - Z')^2}{2}\Big)\Big], \tag{114}$$

$$\overset{e}{\leq} \exp\Big(\frac{\lambda^2 B^2 L^2}{2}\Big), \tag{115}$$

where (a) follows from Jensen's inequality, (b) follows from the fact that $g(Z) - g(Z')$ is the same as $\varepsilon(g(Z) - g(Z'))$, (d) follows from since g is Lipschitz and (e) follows since Z is bounded.

Next, from assumption (A4), and the projection step of the algorithm we have that $f_i(\theta_{i-1}) < (R_{\max} + (1+\beta)H\phi^2_{\max})$ is a bounded random variable, and conditioned on $\mathcal{F}_{i-1}$, D_i is Lipshitz in $f_i(\theta_{i-1})$ with constant L_i, therefore we have

$$\mathbb{E}[\exp(\lambda D_t)|\mathcal{F}_{t-1}] \leq \exp\Big(\frac{\lambda^2(R_{\max} + (1+\beta)H\phi^2_{\max})^2 L_t^2}{2}\Big). \tag{116}$$

And hence,

$$P(\|z_t\|_2 - \mathbb{E}[\|z_t\|_2] \geq \varepsilon) \leq \exp(-\lambda\varepsilon)\exp\Big(\frac{\lambda^2(R_{\max} + (1+\beta)H\phi^2_{\max})^2 L_t^2}{2}\Big). \tag{117}$$

Finally optimising over λ gives us

$$P(\big\|z_{k+1,N}\big\|_2 - \mathbb{E}[\big\|z_{k+1,N}\big\|_2]) \leq \exp\Big(-\frac{\varepsilon^2}{(R_{\max} + (1+\beta)H\phi^2_{\max})^2 \sum_{i=1}^{t} L_i^2}\Big). \tag{118}$$

□

D.4. Bounding the Lipschitz constant

Lemma 9. *For the tail-averaged TD, the Lipschtiz constant L_i in proposition 1 is upper bounded as:*

$$L_i^2 \leq \frac{\gamma}{\mu N^2}\exp(-k\gamma\mu). \tag{119}$$

Proof.

$$L_i^2 = \frac{\gamma^2}{N^2}\left(\sum_{k+1}^{k+N}(1-\gamma\mu)^{\frac{j}{2}}\right)^2, \tag{120}$$

$$\leq \frac{\gamma^2}{N^2}\left(\sum_{k+1}^{\infty}(1-\gamma\mu)^{\frac{j}{2}}\right)^2, \tag{121}$$

$$\overset{a}{=} \frac{\gamma^2}{\gamma\mu N^2}(1-\gamma\mu)^{k+1}, \tag{122}$$

$$\overset{b}{\leq} \frac{\gamma}{\mu N^2}\exp(-k\gamma\mu), \tag{123}$$

where (a) is due to summing the geometric series and (b) follows form the elementary inequality $(1+x)^y = \exp(y\log(1+x)) \leq \exp(xy)$ □

Theorem 4.

$$P\left(\left\|z_{k+1,N}\right\|_2^2 \leq \frac{\sigma^2\exp(-k\gamma\mu)}{\mu N^2} + \left(1+\frac{4}{\gamma\mu}\right)\left(\frac{2\exp(-k\gamma\mu)}{\gamma\mu N^2}\mathbb{E}[\left\|z_0^{bias}\right\|^2] + \frac{2\sigma^2}{\mu N}\right)\right) \geq 1-\delta. \tag{124}$$

$$\left\|z_{k+1,N}\right\|_2^2 - \mathbb{E}\left\|z_{k+1,N}\right\|_2^2 \leq \frac{\sigma^2\exp(-k\gamma\mu)}{\mu N} \text{ with probability } 1-\delta \tag{125}$$

$$\left\|z_{k+1,N}\right\|_2^2 \leq \frac{\sigma^2\exp(-k\gamma\mu)}{\mu N^2} + \left(1+\frac{4}{\gamma\mu}\right)\left(\frac{2\exp(-k\gamma\mu)}{\gamma\mu N^2}\mathbb{E}[\left\|z_0^{\text{bias}}\right\|^2] + \frac{2\sigma^2}{\mu N}\right) \tag{126}$$

Theorem 5 (High-probability bound). *Under assumptions (A2)–(A6), for any step size $\gamma \leq \gamma_{max} = \frac{\mu}{(1+\beta)^2\Phi_{max}^4}$, any number δ, and with $\sigma = (R_{max} + (1+\beta)\Phi_{max}^2\|\theta^\star\|_2)$, the worst-case error of the tail-averaged iterate $\theta_{k+1,N}$ as per (10) is bounded as follows:*

$$P\left(\left\|\theta_{k+1,N} - \theta^\star\right\|_2^2 \leq \frac{\sigma^2 e^{(-k\gamma\mu)}}{\mu N^2} + \left(1+\frac{4}{\gamma\mu}\right)\left(\frac{2e^{(-k\gamma\mu)}}{\gamma\mu N^2}\mathbb{E}[\|\theta_0 - \theta^\star\|_2^2] + \frac{2\sigma^2}{\mu N}\right)\right) \geq 1-\delta, \tag{127}$$

where $N = t-k$

Proof.

$$P(\left\|z_{k+1,N}\right\|_2 - \mathbb{E}[\left\|z_{k+1,N}\right\|_2]) \overset{a}{\leq} \exp\left(-\frac{\varepsilon^2}{(R_{\max} + (1+\beta)H\phi_{\max}^2)^2\sum_{i=1}^t L_i^2}\right), \tag{128}$$

$$\overset{b}{\leq} \exp\left(-\frac{\mu N^2\varepsilon^2}{\sigma^2\exp(-k\gamma\mu)}\right), \tag{129}$$

where (a) follows from Proposition 1, and (b) follows from Lemma 9.

Now let

$$\exp\left(-\frac{\mu N^2\varepsilon^2}{\sigma^4\exp(-k\gamma\mu)}\right) = \delta, \tag{130}$$

$$\frac{\mu N^2\varepsilon^2}{\sigma^2\exp(-k\gamma\mu)} = \log(\frac{1}{\delta}), \tag{131}$$

The final bound follows by substituting the value for ε from (131) in (129), and using the result from theorem 1. □