Almost Sure Convergence of Differential Temporal Difference Learning for Average Reward Markov Decision Processes

Ethan Blaser
University of Virginia
blaser@email.virginia.edu

Jiuqi Wang
University of Virginia
jiuqi@email.virginia.edu

Shangtong Zhang
University of Virginia
shangtong@virginia.edu

Abstract

The average reward is a fundamental performance metric in reinforcement learning (RL) focusing on the long-run performance of an agent. Differential temporal difference (TD) learning algorithms are a major advance for average reward RL as they provide an efficient online method to learn the value functions associated with the average reward in both on-policy and off-policy settings. However, existing convergence guarantees require a local clock in learning rates tied to state visit counts, which practitioners do not use and does not extend beyond tabular settings. We address this limitation by proving the almost sure convergence of on-policy n-step differential TD for any n using standard diminishing learning rates without a local clock. We then derive three sufficient conditions under which off-policy n-step differential TD also converges without a local clock. These results strengthen the theoretical foundations of differential TD and bring its convergence analysis closer to practical implementations.

1 INTRODUCTION

The average reward is an important performance metric in Reinforcement Learning (RL, Sutton and Barto (2018)). Compared with the commonly used discounted total rewards performance metric, the average reward setting more heavily emphasizes the long-term behavior of the RL agent, making it particularly suitable for applications like network resource allocation (Marbach et al., 1998; Bakhshi et al., 2021; Yang et al., 2024), robotics (Kober et al., 2013), and scheduling (Ghavamzadeh and Mahadevan, 2007).

Differential temporal difference (TD) (Wan et al., 2021b) learning is one of the most important recent advances for average reward RL. Differential TD is designed to estimate the corresponding value function for the average reward performance metric and can be used in both on-policy and off-policy settings. However, the convergence analysis of differential TD remains less satisfactory. In Wan et al. (2021b), almost sure convergence is proved only when the stepsizes depend on a local clock. Specifically, they require the learning rates of the form $\left\{\alpha_{\nu(t,S_t)}\right\}$, where $\{\alpha_t\}$ is a sequence of deterministic, nonnegative, and diminishing scalars and a local clock (i.e., a counter) $\nu(t,s)$, which counts the number of visits to a state s up to timestep t. In other words, at time t the stepsize depends not only on t, but also on the number of past visits to the current state S_t.

We argue that this local clock based learning rate is unsatisfactory for at least three reasons. First, to our best knowledge, practitioners do not actually use the local clock in their learning rates, including Wan et al. (2021b) in their experiments. The local clock seems to be primarily a theoretically motivated technique (Borkar, 2009). Although recent work demonstrates that it can occasionally be required for convergence in certain settings (Chen, 2025), its adoption in practical implementations remains rare. Second, the local clock cannot be used in many function approximation settings, especially those considered in Sutton and Barto (2018), where the agent only has access to the feature of the current state, denoted as $\phi(S_t)$, not the state S_t itself. With only $\phi(S_t)$, it is not clear how to count the visits to S_t since the feature function ϕ is usually not a one-to-one mapping. This means the local clock technique is only viable in the tabular setting. Third, although convergence analyses of discounted TD (Sutton, 1988) also require the local clock in learning rates (Jaakkola et al., 1993; Tsitsiklis, 1994), later works removed this requirement (Tsitsiklis and Roy, 1996;

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Liu et al., 2025a). Therefore, there is a theoretical gap in the literature for average reward RL, and gives rise to the central question this work aims to answer:

Can we establish the convergence of differential TD without using a local clock in the learning rates?

This question seems trivial at first glance. After all, local clocks can be avoided in the discounted setting, so one might expect the same argument to carry over to the average reward setting. However, as we will now explain, extending that analysis to the average-reward case introduces several fundamental obstacles.

The convergence of the discounted TD with a local clock rests on the global asymptotic stability (G.A.S.) of the following ODE[1]

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = Av(t), \tag{1}$$

where $v(t) \in \mathbb{R}^{|\mathcal{S}|}$ can be viewed as the estimation of the value function and $A \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ corresponds to the discounted TD algorithm, with $|\mathcal{S}|$ being the number of states. Essentially, (1) is G.A.S. because the A matrix corresponding to discounted TD with a local clock is negative definite (n.d.)[2]. When the local clock is removed from the learning rates, the corresponding ODE becomes

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = DAv(t), \tag{2}$$

where $D \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is a diagonal matrix whose entries are the stationary state distribution. The change from (1) to (2) is intuitive. With a local clock, the total magnitude of updates applied to each state is forced to be the same, regardless of how frequently that state is visited. Without a local clock, the magnitude of the updates naturally depends on visitation frequency, which appears as the multiplier D in (2). For instance, when a state s is visited for the first time, the learning rate is always α_1 with a local clock, whereas without it the learning rate may be α_{100} if s is first visited at time $t = 100$. Nevertheless, when A is n.d., it is straightforward to show that DA is also n.d., implying that (2) is G.A.S., and thus that discounted TD converges even without a local clock.

However, as we shall show soon, the corresponding A matrix for differential TD with the local clock is only Hurwitz and not necessarily n.d. When A is Hurwitz, whether DA is also Hurwitz is a long-standing open problem in the linear algebra community, called the D-stability problem (Johnson, 1974b; Giorgi and Zuccotti, 2015). Progress on the D-stability problem has been limited in the past decade (Kushel, 2019; Kushel and Pavani, 2023; Tong and Su, 2024). As a result, verifying whether (2) is G.A.S. for differential TD without a local clock is substantially more challenging than it initially appears.

Nevertheless, this paper makes three contributions. First, we establish the almost sure convergence of on-policy n-step differential TD for any n without the local clock. Second, we give three different sufficient conditions for the almost sure convergence of off-policy n-step differential TD without the local clock. Admittedly, our characterization in the off-policy case is incomplete and we correspondingly present our third contribution: we outline a few challenges and open problems in this area.

2 BACKGROUND

In this work, all vectors are column. The ℓ_2 norm in $\mathbb{R}^d$ is denoted by $\|\cdot\|$. The identity matrix is denoted by I, and we use e to denote the all-one vector. For a matrix $A \in \mathbb{R}^{n\times n}$, we denote its spectral radius by $\lambda_{\max}(A) \doteq \max\{|\lambda| : \lambda \in \sigma(A)\}$, with $\sigma(A)$ as the set of eigenvalues of A. We say that a matrix A is (strictly) *positive stable* if $\forall\lambda \in \sigma(A)$ $\mathrm{Re}\,\lambda \geq 0$ ($\mathrm{Re}\,\lambda > 0$). It is easy to see that A is strictly positive stable if and only if $-A$ is Hurwitz.[3] If a matrix A has only nonnegative (positive) entries, we write $A \geq 0$, ($A > 0$). If a matrix A is positive definite, we write $A \succ 0$. Given any vector x, $\sum x$ denotes the sum of all elements in x. We use $A_{i,j}$ to refer to the (i,j)-th entry of A.

Definition 2.1. An $M-$matrix is a matrix of the form $\gamma I - P$ where $P \in \mathbb{R}^{n\times n}$, $P \geq 0$, and $\gamma \geq \lambda_{\max}(P)$.

In RL, we consider a Markov Decision Process (MDP; Bellman (1957); Puterman (2014)) with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a reward function $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, a transition function $p : \mathcal{S}\times\mathcal{S}\times\mathcal{A} \to [0,1]$, an initial distribution $p_0 : \mathcal{S} \to [0,1]$. At time step 0, an initial state S_0 is sampled from p_0. At time t, given the state S_t, the agent samples an action $A_t \sim \pi(\cdot|S_t)$, where $\pi : \mathcal{A}\times\mathcal{S} \to [0,1]$ is the policy being followed by the agent. A reward $R_{t+1} \doteq r(S_t, A_t)$ is then emitted and the agent proceeds to a successor state $S_{t+1} \sim p(\cdot|S_t, A_t)$.

We assume the Markov chain $\{S_t\}$ induced by

[1] The ODE (1) is G.A.S. if and only if the A matrix is Hurwitz (Theorem 4.5 from Khalil (2002)). A matrix A is Hurwitz if the real part of any of its eigenvalues is strictly negative.

[2] A matrix A, not necessarily symmetric, is n.d. if for any $y \neq 0$, it holds that $y^\top Ay < 0$. A n.d. matrix must be Hurwitz. But a Hurwitz matrix does not need to be n.d. For example, $\begin{bmatrix} -1 & 10 \\ 0 & -1 \end{bmatrix}$ is Hurwitz but not n.d.

[3] "Hurwitz" is often used in the control community while "positive stable" is often used in the linear algebra community

the policy π is ergodic and thus adopts a unique stationary distribution d_π. We define $D_\pi = \mathrm{diag}(d_\pi)$. The average reward (a.k.a. gain, Puterman (2014)) is defined as $\bar{J}_\pi \doteq \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[R_t\right]$. Consequently, the differential value function (a.k.a. bias, Puterman (2014)) is defined as $v_\pi(s) \doteq \lim_{T\to\infty} \frac{1}{T}\sum_{\tau=1}^{T} \mathbb{E}\left[\sum_{i=1}^{\tau}(R_{t+i} - \bar{J}_\pi) \mid S_t = s\right]$. The corresponding Bellman equation (a.k.a. Poisson's equation) is then

$$v = r_\pi - \bar{J}_\pi e + P_\pi v, \tag{3}$$

where $v \in \mathbb{R}^{|\mathcal{S}|}$ is the free variable, $r_\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the reward vector induced by the policy π, i.e., $r_\pi(s) \doteq \sum_a \pi(a|s) r(s,a)$, and $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is the transition matrix induced by the policy π, i.e., $P_\pi(s,s') \doteq \sum_a \pi(a|s)p(s'|s,a)$. It is known (Puterman, 2014) that all solutions to (3) form a set $\mathcal{V}_* \doteq \{v_\pi + ce \mid c \in \mathbb{R}\}$. The policy evaluation problem in average reward MDPs is to estimate v_π, perhaps up to a constant offset ce.

In the off-policy setting, an agent aims to evaluate a target policy π but follows a behavior policy μ. We define the importance sampling ratio $\rho(s,a) \doteq \frac{\pi(a|s)}{\mu(a|s)}$ and $\rho_t \doteq \rho(S_t, A_t)$.

3 DIFFERENTIAL TEMPORAL DIFFERENCE LEARNING

Differential TD is designed to estimate v_π in an online manner. The differential TD algorithm proposed by Wan et al. (2021b) only considers a one-step lookahead. Inspired by the success of n-step TD in the discounted setting (Sutton and Barto, 2018), we first extend the 1-step differential TD to the n-step case. As we shall see soon, this extension is vital to the analysis in the off-policy setting. Here we only present off-policy n-step differential TD as the on-policy version is just a special case with $\mu = \pi$. Suppose a trajectory $\{S_0, A_0, R_1, S_1, \dots\}$ is generated by following a behavior policy μ as $A_t \sim \mu(\cdot \mid S_t)$. Since the n-step return for S_t is only observable after reaching S_{t+n}, the iterates $\{v_t \in \mathbb{R}^{|\mathcal{S}|}\}, \{J_t \in \mathbb{R}\}$ are updated at time $t+n$ as

$$\delta_t = R_{t+1:t+n} - nJ_{t+n-1} + v_{t+n-1}(S_{t+n}) - v_{t+n-1}(S_t),$$

$$\begin{aligned} J_{t+n} &= J_{t+n-1} + \tfrac{\eta}{n}\alpha_{t+n-1}\rho_{t:t+n-1}\delta_t, \\ v_{t+n}(S_t) &= v_{t+n-1}(S_t) + \alpha_{t+n-1}\rho_{t:t+n-1}\delta_t, \end{aligned} \tag{4}$$

where $\rho_{t:t+n-1} \doteq \prod_{k=t}^{t+n-1}\rho_k$ and $R_{t+1:t+n} \doteq \sum_{k=1}^{n} R_{t+k}$ are shorthands.

To our knowledge, this is the first time that n-step differential TD is formalized, and the complete derivation is presented in Appendix B. When $n = 1$, it recovers the 1-step differential TD in Wan et al. (2021b). However, in the convergence analysis in Wan et al. (2021b), they replace the learning rate $\{\alpha_t\}$ with $\{\alpha_{\nu(t,S_t)}\}$. We recall that $\nu(t,s)$ counts the number of visits to the state s until time t and is referred to as the local clock. In this work, we shall conduct our analysis of (4) directly without altering the learning rates.

Inspired by Wan et al. (2021b), to facilitate our analysis, we first rewrite (4) to eliminate the iterates $\{J_t\}$. Define $\Sigma_t \doteq \sum_s v_t(s)$. Since the $n-$step return for S_t is only available after time $t+n$, we adopt the standard convention that no updates occur before the first $n-$step return is observed, so J_t and v_t are constant for $t < n$. Making use of the fact that v_{t+n} and v_{t+n-1} differ from each other only for the S_t-indexed entry, we obtain

$$\begin{aligned} J_{t+n} - J_{n-1} &= \textstyle\sum_{i=0}^{t} \frac{\eta}{n}\alpha_{i+n-1}\rho_{i:i+n-1}\delta_i, \\ &= \textstyle\sum_{i=0}^{t} \frac{\eta}{n}(v_{i+n}(S_i) - v_{i+n-1}(S_i)), \\ &= \textstyle\sum_{i=0}^{t} \frac{\eta}{n}\sum_s (v_{i+n}(s) - v_{i+n-1}(s)), \\ &= \tfrac{\eta}{n}\,(\Sigma_{t+n} - \Sigma_{n-1}). \end{aligned}$$

We can then rewrite δ_t from (4) as

$$\begin{aligned} \delta_t = R_{t+1:t+n} - n\big(J_{n-1} + \tfrac{\eta}{n}(\Sigma_{t+n-1} - \Sigma_{n-1})\big) \\ + v_{t+n-1}(S_{t+n}) - v_{t+n-1}(S_t). \end{aligned}$$

Then, (4) can be expressed more compactly as

$$\begin{aligned} v_{t+n}(S_t) = v_{t+n-1}(S_t) + \alpha_{t+n-1}\rho_{t:t+n-1}\big(\tilde{R}_{t+1:t+n} \\ - \eta\Sigma_{t+n-1} + v_{t+n-1}(S_{t+n}) - v_{t+n-1}(S_t)\big), \end{aligned} \tag{5}$$

where $\tilde{R}_{t+1:t+n} \doteq \sum_{k=1}^{n}(R_{t+k} - J_{n-1} + \frac{\eta}{n}\Sigma_{n-1})$. We assume initialization with $J_{n-1} \doteq 0$ and $v_{n-1} \doteq 0$ for simplifying presentation so that $\tilde{R}_{t+1:t+n} = R_{t+1:t+n}$. For nonzero initialization of J_0 and v_0, we only need to conduct the same analysis in a new MDP with a shifted reward function $\tilde{r}(s,a) \mapsto r(s,a) - J_{n-1} + \frac{\eta}{n}\Sigma_{n-1}$ (Wan et al., 2021b).

4 CONVERGENCE OF DIFFERENTIAL TEMPORAL DIFFERENCE LEARNING

We make the following standard assumptions.

Assumption 4.1 (Ergodicity and coverage)**.** The Markov chains induced by the behavior policy μ and target policy π are finite, irreducible, and aperiodic. The behavior policy μ covers π i.e. $\forall\, s \in \mathcal{S},\ \forall\, a \in \mathcal{A}$: $\pi(a \mid s) > 0 \implies \mu(a \mid s) > 0$.

The ergodicity assumption is standard for analyzing RL algorithms (Bertsekas and Tsitsiklis, 1996). Furthermore, the coverage assumption is the same as in Sutton and Barto (2018).

From Assumption 4.1, the Markov chains induced by the behavior policy and target policy each adopt a unique stationary distribution, which we denote respectively as d_π and d_μ. Because the Markov chains are irreducible, $d_\pi > 0$ and $d_\mu > 0$ (Puterman, 2014).

Assumption 4.2. The learning rates $\{\alpha_t\}$ are positive, decreasing, and satisfy

$$\textstyle\sum_{t=0}^{\infty} \alpha_t = \infty, \lim_{t\to\infty} \alpha_t = 0, \text{ and } \frac{\alpha_t - \alpha_{t+1}}{\alpha_t} = \mathcal{O}(\alpha_t).$$

This is the standard set of assumptions for learning rates in stochastic approximation (Borkar, 2009). We emphasize that this definition of the learning rates is far more widely used compared to the state visitation-dependent learning rates found in Wan et al. (2021b). For example, Assumption 4.2 is satisfied by any learning rate of the form $\alpha_t = \frac{C_1}{(n+C_2)^\beta}$ where C_1 and C_2 are constants and $\beta \in (0.5, 1]$.

Our proof of convergence will utilize some results from the stochastic approximation (SA) community. Thus, we begin by writing the update (5) as a canonical stochastic approximation update by first defining an augmented Markov chain $\{Y_t\}$ evolving in a finite state space $\mathcal{Y}$ as,

$$Y_{t+1} = (S_t, A_t, S_{t+1}, A_{t+1}, \ldots, A_{t+n-1}, S_{t+n}). \quad (6)$$

From Assumption 4.1, it is clear that $\{Y_t\}$ is also irreducible and aperiodic, and we denote its stationary distribution as $d_\mathcal{Y}$. Then, we can define the operator $H : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{Y} \to \mathbb{R}^{|\mathcal{S}|}$ as,

$$\begin{aligned} H(v, y)[s] \doteq \rho_{0:n-1}(y)\big(\textstyle\sum_{k=0}^{n-1} r(s_k, a_k) - \eta \sum v \\ + v(s_n) - v(s_0)\big)\mathbb{I}\{s = s_0\}. \end{aligned} \quad (7)$$

where we have $y \doteq (s_0, a_0, \ldots s_n)$ and $\rho_{0:n-1}(y) \doteq \prod_{k=0}^{n-1} \rho(s_k, a_k)$. Then we can write (5) as a canonical SA algorithm according to

$$v_{t+1} = v_t + \alpha_t H(v_t, Y_{t+1}). \quad (8)$$

Note that the SA iteration index t differs from the environment time step in (5). One SA update corresponds to an n-step block of experience, so SA step t corresponds to environment time $t + n - 1$, since the tuple Y_{t+1} becomes available only after observing up to S_{t+n}.

One prominent method for analyzing the asymptotic behavior of $\{v_t\}$ is to regard $\{v_t\}$ as Euler's discretization of the ODE

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = h(v(t)) \quad (9)$$

where the expected operator $h(v) \doteq \mathbb{E}_{y\sim d_\mathcal{Y}}[H(v, y)]$. Using this method, the asymptotic behavior of the discrete and stochastic updates $\{v_t\}$ can be characterized by the continuous and deterministic trajectories of the ODE (9), if the stability of the iterates can be established. The Borkar-Meyn theorem (Borkar and Meyn, 2000) establishes the desired stability given the ODE@∞ is G.A.S., which is defined as

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = h_\infty(v(t)),$$

where $h_\infty \doteq \lim_{c\to\infty} \frac{h(cv)}{c}$. Although the original work of Borkar and Meyn (2000) only allows for $\{Y_t\}$ to be i.i.d, recently Liu et al. (2025a) generalized the Borkar-Meyn theorem to Markovian noise $\{Y_t\}$ under equally mild assumptions, an important extension which we will leverage here since our $\{Y_t\}$ in (6) is a Markov chain.

We therefore proceed by studying the expected operator for (5):

$$\begin{aligned} h(v)[s] &= \mathbb{E}_{y\sim d_\mathcal{Y}}[H(v, y)[s]] \\ &= \mathbb{E}_\mu\Big[\rho_{0:n-1}(y)\big(\textstyle\sum_{k=0}^{n-1} r(s_k, a_k) - \eta \sum v \\ &\qquad + v(s_n) - v(s)\big)\mathbb{I}\{s = s_0\}\Big] \\ &= d_\mu(s)\mathbb{E}_\mu\Big[\rho_{0:n-1}(y)\Big(\textstyle\sum_{k=0}^{n-1} r(s_k, a_k) \\ &\qquad - \eta \sum v + v(s_n) - v(s)\Big)\Big| s_0 = s\Big], \end{aligned}$$

where we have used $\mathbb{E}_\mu$ to abbreviate the expectation over the trajectory generated by $a_k \sim \mu(\cdot|s_k)$ and $s_{k+1} \sim p(\cdot|s_k, a_k)$. Isolating the reward-sum term, we define the expected n-step reward by

$$\begin{aligned} r^{(n)}(s) &\doteq \mathbb{E}_\mu\Big[\rho_{0:n-1} \textstyle\sum_{k=0}^{n-1} r(s_k, a_k)\Big| s = s_0\Big] \\ &= \mathbb{E}_\pi\Big[\textstyle\sum_{k=0}^{n-1} r(s_k, a_k)\Big| s = s_0\Big] \\ &= \textstyle\sum_{k=0}^{n-1} \big(P_\pi^k r_\pi\big)(s). \end{aligned}$$

Therefore, the expected operator for (5) is

$$h(v) = D_\mu(r^{(n)} - \eta e e^\top v + P_\pi^n v - v), \quad (10)$$

where D_μ is a diagonal matrix whose entries are d_μ. The corresponding ODE@∞ is

$$\begin{aligned} \frac{\mathrm{d}v(t)}{\mathrm{d}t} &= h_\infty(v(t)) \\ &= D_\mu\big(P_\pi^n - I - \eta e e^\top\big)v(t) \\ &= -Av(t), \end{aligned} \quad (11)$$

where $A \doteq D_\mu\big(I - P_\pi^n + \eta e e^\top\big)$.

We now outline the structure of our proof. The first milestone is to prove that the ODE (11) is G.A.S. It is then trivial to see that (9) is also G.A.S.. We will

use v_∞ to denote the G.A.S. equilibrium of (9) and we have $h(v_\infty) = 0$. This means that $r^{(n)} - \eta e e^\top v_\infty + P_\pi^n v_\infty - v_\infty = 0$. The analysis in Appendix B.2.1 of Wan et al. (2021b), which we omit to avoid redundancy, then immediately confirms that $v_\infty \in \mathcal{V}_*$. The second milestone is to invoke a result from Liu et al. (2025a) (stated as Lemma A.7 in the Appendix) to prove that the iterates $\{v_t\}$ generated in (5) converge to v_∞ almost surely. We now proceed to carry out this proof strategy in detail.

It is known that a necessary and sufficient condition for (11) to be G.A.S. is that A is strictly positive stable (Theorem 4.5 from Khalil (2002)). Wan et al. (2021b) essentially prove that the matrix $I - P_\pi^n + \eta e e^\top$ is strictly positive stable. However, this does not mean that the A matrix is strictly positive stable. This is an instance of the D-stability problem. As discussed in Section 6, this is a very challenging problem in the linear algebra community. Nevertheless, to prove A is strictly positive stable, we will utilize the results from Bierkens and Ran (2014), which we present as Lemma 4.3, that establish conditions under which M-matrices (see Definition 2.1) are strictly positive stable under rank one perturbations.

Lemma 4.3. *(Theorem 2.7 from Bierkens and Ran (2014)). Let $B \in \mathbb{R}^{n\times n}$ and $v, w \in \mathbb{R}^n$. Then $B + vw^\top$ is strictly positive stable if:*

1. *$B = \lambda_{\max}(K)I - K$ is a singular M-matrix where $K \in \mathbb{R}^{n\times n}$.*
2. *0 is a geometrically simple eigenvalue of B with left and right eigenvectors $z_l \neq 0$ and $z_r \neq 0$. (i.e. $z_l^\top B = 0$ and $Bz_r = 0$).*
3. *$(z_l^\top v)(w^\top z_r) \neq 0$*

*and **either** of the following conditions hold:*

4. *$Bv = 0$, or $w^\top B = 0$.*
5. *$v, w > 0$ and $2K_{i,j} \geq v_i w_j \, \forall \, i, j$ (where v_i, respectively, w_j denote the i-th entry of v and the j-th entry of w)*

To utilize Lemma 4.3 to prove the strict positive stability of A from (11), we begin by decomposing A into the form of $B + vw^\top$ with,

$$\begin{aligned} A &= D_\mu\big(I - P_\pi^n + \eta\, e\, e^\top\big) \\ &= I - I + D_\mu(I - P_\pi^n) + \eta\, d_\mu\, e^\top \\ &= I - (I + D_\mu(P_\pi^n - I)) + \eta d_\mu e^\top \\ &\doteq B + \eta d_\mu e^\top, \end{aligned} \tag{12}$$

where $B \doteq I - (I + D_\mu(P_\pi^n - I))$, and we recall that η is a positive constant.

Without any additional assumptions, we can verify the first three conditions of Lemma 4.3 in the following Lemma.

Lemma 4.4. *Let Assumption 4.1 hold. Then, $B \doteq I - (I + D_\mu(P_\pi^n - I))$, $v \doteq \eta d_\mu$, and $w \doteq e$ satisfy conditions 1-3 of Lemma 4.3.*

Proof. First, we verify Condition 1. If we define $K \doteq I + D_\mu(P_\pi^n - I)$, we have $B = I - K$. Therefore, it is sufficient to prove that $\lambda_{\max}(K) = 1$. Since $D_\mu = \mathrm{diag}(d_\mu)$ with $d_\mu > 0$ and P_π^n is non-negative, it is easy to see that

$$K \doteq I + D_\mu(P_\pi^n - I) = (I - D_\mu) + D_\mu P_\pi^n \tag{13}$$

is non-negative. Additionally, computing the row sums of K, we see that it is row-stochastic:

$$Ke = (I - D_\mu)e + D_\mu P_\pi^n e = e - D_\mu e + D_\mu e = e,$$

where the second equality holds because the transition matrix P_π^n is row-stochastic. Then, we are guaranteed that $\lambda_{\max}(K) = 1$ (Theorem 8.1.22 from Horn and Johnson (2012)).

To verify Condition 2, we demonstrate that B has 0 as an algebraically (and therefore geometrically) simple eigenvalue with left and right eigenvectors $z_l, z_r \neq 0$ (i.e. $z_l^\top B = 0$ and $Bz_r = 0$). We have

$$\ker B = \ker (I - K) = \{z_r : Kz_r = z_r\}. \tag{14}$$

Additionally, since $B^\top = (I - K)^\top$ we have

$$\ker B^\top = \ker (I - K)^\top = \big\{z_l : z_l^\top K = z_l^\top\big\}.$$

Since P_π is irreducible and aperiodic by Assumption 4.1, P_π^n is irreducible for every $n \geq 1$ (Puterman, 2014). Multiplying a positive diagonal matrix and adding another positive diagonal matrix leaves the zero pattern unchanged so $K = I - D_\mu + D_\mu P_\pi^n$ is also irreducible. Therefore, the Perron-Frobenius theorem (Theorem 8.4.4 in Horn and Johnson (2012)) guarantees that one is an algebraically simple eigenvalue of K. This implies that $\ker B$ and $\ker B^\top$ are one-dimensional, and thus zero is a geometrically simple eigenvalue of B. We identify the one-dimensional left and right kernels of B as,

$$\ker B = \mathrm{span}(e), \quad \ker B^\top = \mathrm{span}(d_\pi / d_\mu),$$

where d_π / d_μ represents element-wise division. For the right kernel, it holds trivially from (14) and the fact that K is row-stochastic. For the left kernel, with $B^\top = I - K^\top = D_\mu - P_\pi^{n\top} D_\mu$ we have,

$$\begin{aligned} B^\top\left(\tfrac{d_\pi}{d_\mu}\right) &= D_\mu\left(\frac{d_\pi}{d_\mu}\right) - P_\pi^{n\top} D_\mu\left(\frac{d_\pi}{d_\mu}\right) \\ &= d_\pi - P_\pi^{n\top} d_\pi \\ &= 0. \end{aligned}$$

Clearly $z_l = d_\pi/d_\mu$ and $z_r = e$ are both nonzero, so Condition 2 is satisfied.

To verify Condition 3, we note that all components of e and $\frac{d_\pi}{d_\mu}$ are strictly positive, so any non-zero vector $z_l^\top \in \text{span}(\frac{d_\pi}{d_\mu})$ and $z_r \in \text{span}(e)$ will have uniform sign. Using the fact that $v = \eta d_\mu$ and $w = e$ are strictly positive, it is easy to see that

$$(z_l^\top v)(w^\top z_r) = \eta(z_l^\top d_\mu)(e^\top z_r) \neq 0.$$

□

Although we have verified Conditions 1-3 of Lemma 4.3 using only Assumption 4.1, we still need either Condition 4 or 5 to establish that A is strictly positive stable. We therefore split the analysis into two regimes. In the on-policy case with $\mu = \pi$ (Section 4.1), we are able to directly satisfy Condition 4. In the off-policy case (Section 4.2), additional restrictions are needed, and we provide three sufficient conditions.

4.1 On-Policy Case

In the on-policy case, we consider the following assumption.

Assumption 4.5 (On-policy)**.** The behavior policy followed by the agent is the target policy, i.e., $\pi(a|s) = \mu(a|s)\,\forall s \in \mathcal{S}, a \in \mathcal{A}$.

To prove the strict positive stability of A in the on-policy case, since Conditions 1-3 are already in place from Lemma 4.4, it remains only to verify Condition 4. Theorem 4.6 does so, thereby establishing the strict positive stability of A. Corollary 4.7 then gives the almost-sure convergence of Differential TD.

Theorem 4.6. *Let Assumptions 4.1 and 4.5 hold. Then, $A = B + \eta d_\mu e^\top$ is strictly positive stable for any $n \geq 1$ and any $\eta > 0$.*

Proof. Recall from (12), we have expressed A in the form of $B + vw^\top$ where $B \doteq I - (I + D_\mu(P_\pi^n - I))$, $v = \eta d_\mu$ and $w = e$. Lemma 4.3 states that A is strictly positive-stable once Conditions 1-3 together with either Condition 4 or 5, are satisfied. In Lemma 4.4 we verify the first three conditions of Lemma 4.3 with this choice of B, v, w. In the on-policy case (Assumption 4.5), we have $\mu = \pi$, which further gives

$$B = I - (I + D_\pi(P_\pi^n - I)), \quad v = \eta d_\pi, \quad w = e.$$

To demonstrate Condition 4 holds in the on-policy setting, we show $w^\top B = 0$ with,

$$w^\top B = e^\top D_\pi(I - P_\pi^n) = d_\pi^\top(I - P_\pi^n) = 0.$$

□

Corollary 4.7. *Let Assumptions 4.1, 4.2 and 4.5 hold. Then the iterates $\{v_t\}$ in (5) satisfy: $\lim_{t\to\infty} v_t = v_\infty$, a.s., where $v_\infty \in \mathcal{V}_*$.*

Proof. To prove the almost sure convergence of the differential TD iterates in (5) to fixed point v_∞ we will utilize Corollary 8 from (Liu et al., 2025a) which we present as Lemma A.7. We proceed by verifying the requisite Assumptions A.1-A.6. Starting with Assumption A.5, in Theorem 4.6 we prove that A defined in (11) is strictly positive stable under Assumptions 4.1, 4.2, 4.5. Therefore, the ODE@∞ in (11), is G.A.S. (Theorem 4.5 from Khalil (2002)).

Verifying the remaining assumptions is straightforward. Note that our Assumptions 4.1 and 4.2 are sufficient to directly satisfy Assumptions A.1, A.2 and A.6. We refer the reader to Remarks 1-3 of (Liu et al., 2025a) for a discussion on how these are trivially satisfied for ergodic and finite $\{Y_t\}$. We then verify Assumption A.3 in Lemma C.1. It is easy to verify that $H(x, y)$ is Lipschitz, which we present for completeness in Lemma C.2 that satisfies A.4. Then, Lemma A.7 guarantees that $\{v_t\}$ converges to the invariant set of the ODE, which is a singleton we denote as v_∞. □

4.2 Off-Policy Case

In the off-policy setting, we consider three additional assumptions.

We will first prove that there exists some η_0 for which for $\eta \in (0, \eta_0]$, A is strictly positively stable using an extension of Lemma 4.3, Lemma 4.8.

Lemma 4.8 (Lemma 2.11 from Bierkens and Ran (2014))**.** *Let $B \in \mathbb{R}^{n\times n}, v, w \in \mathbb{R}^n, v, w \geq 0$ satisfy Conditions 1-3 of Lemma 4.3. Additionally, let 0 be an algebraically simple eigenvalue of B. Define a matrix-valued curve $\Gamma(t) : t \to B + tvw^\top, t \in \mathbb{R}$. There exists a $t_0 > 0$ such that $\Gamma(t)$ is strictly positive stable for $t \in (0, t_0]$.*

Lemma 4.9. *Let Assumption 4.1 hold. Then, there exits a $\eta_0 > 0$ such that $A = B + \eta d_\mu e^\top$ is strictly positive stable for $\eta \in (0, \eta_0]$.*

Proof. Recall from (12) that $A = B + \eta d_\mu e^\top$ where $B \doteq I - (I + D_\mu(P_\pi^n - I))$. By Lemma 4.4, B, d_μ, and e satisfy Conditions 1-3 of Lemma 4.3. [4] Therefore, if we set $v = d_\mu$, $w = e$ and $t = \eta$, then $A = B + tvw^\top$ and Lemma 4.8 guarantees the existence of some $t_0 > 0$ (hence η_0) for which $\Gamma(t) = B + tvw^\top$ is strictly positive stable on $(0, t_0]$. It remains only to check that 0 is algebraically simple for B.

[4] In Lemma 4.4, we prove this for $v = \eta d_\mu$ instead of $v = d_\mu$. However, since η is a positive constant, its easy to see that the argument still holds.

To show this, we need to show that 0 is a simple root of the characteristic polynomial of B. We use $\chi_M(\lambda) = \det(M - \lambda I)$ to denote the characteristic polynomial of a matrix M. Recall the definition of K from (13). Then, the characteristic polynomial of $B = I - K$ is,

$$\begin{aligned}\chi_B(\lambda) &= \det\left((I - K) - \lambda I\right) \\ &= (-1)^{|\mathcal{S}|} \det\left(K - (1 - \lambda) I\right) \\ &= (-1)^{|\mathcal{S}|} \chi_K(1 - \lambda).\end{aligned}$$

We proved in Lemma 4.6 that 1 is an algebraically simple eigenvalue of K. In other words, $\chi_K(\kappa)$ has a simple root at $\kappa = 1$. Then the change of variable $\kappa \mapsto 1 - \lambda$ implies $\lambda = 0$ is a simple root of χ_B. Thus 0 is an algebraically simple eigenvalue of B.

Then, Lemma 4.8 proves that there exists some $\eta_0 > 0$ for which $A = B + \eta d_\mu e^\top$ is strictly positive stable on $(0, \eta_0]$. □

Having established that A is strictly positive–stable, the almost-sure convergence of (5) to v_∞ follows immediately by the same argument used in Corollary 4.7. In that proof, every assumption except A.5 was checked without invoking Assumption 4.5, and A.5 itself is a direct consequence of the strict positive stability of A. Therefore, we omit the proof of the corollary to avoid redundancy.

Corollary 4.10. *Let Assumptions 4.1, 4.2 hold. Then there exists some positive constant $\eta_0 > 0$ such that for $\eta \in (0, \eta_0]$ the iterates $\{v_t\}$ in (5) satisfy $\lim_{t\to\infty} v_t = v_\infty$, a.s., where $v_\infty \in \mathcal{V}_*$.*

The main limitation of this result is that, while it guarantees some $\eta_0 > 0$, it does not offer a closed form for its value. To address this, we impose the additional assumption that P_π^n becomes strictly positive under sufficiently large n. Under this condition, we can characterize η_0.

Assumption 4.11. P_π^n is strictly positive.

Such an n is guaranteed to exist by the irreducibility of P_π from Assumption 4.1 (Levin and Peres, 2017).

Theorem 4.12. *Let Assumption 4.1 and 4.11 hold. Then, $A = B + \eta d_\mu e^\top$ is strictly positive stable for $\eta \in (0, \eta_0]$ where $\eta_0 \doteq 2 \min_{i,j} P_\pi^n(i, j)$.*

Proof. To prove that A is strictly positive stable with the addition of Assumption 4.11, we will once again utilize Lemma 4.3. Lemma 4.4 shows that B defined in (12), $v = \eta d_\mu$, and $w = e$ satisfy the first three conditions of Lemma 4.3. In addition to Conditions 1-3, we will also prove that Condition 5 holds, which is sufficient to guarantee the strict positive stability of A. Because P_π^n is strictly positive, we have

$$\begin{aligned}K_{ij} &= (1 - d_\mu(i))\mathbb{I}\{i = j\} + d_\mu(i) P_\pi^n(i, j) \\ &\geq d_\mu(i)\, p_{\min} \quad \forall\, i, j.\end{aligned}$$

where we define $p_{\min} \doteq \min_{i,j} P_\pi^n(i, j) > 0$. For any $\eta \in (0, \eta_0]$, we therefore have

$$2K_{ij} \geq 2 d_\mu(i) p_{\min} \geq \eta d_\mu(i) = v_i w_j,\ \forall\, i, j,$$

so the entry-wise inequality in Condition 5 of Lemma 4.3 holds, and the theorem follows. □

Having established that A is strictly positive–stable, the almost-sure convergence of (5) to v_∞ follows immediately by the same argument used in Corollary 4.7.

Corollary 4.13. *Let Assumptions 4.1, 4.2, and 4.11 hold. Then for $\eta \in (0, \eta_0]$, where $\eta_0 \doteq 2 \min_{i,j} P_\pi^n(i, j)$, the iterates $\{v_t\}$ in (5) satisfy: $\lim_{t\to\infty} v_t = v_\infty$, a.s., where $v_\infty \in \mathcal{V}_*$.*

Having presented two sufficient conditions on η and n, we now present the third sufficient condition on P_π. Namely, under the assumption that P_π is doubly stochastic, we are able to establish the strict positive stability of A for any $\eta \geq 0$ and $n > 0$.

Assumption 4.14. P_π^n is doubly stochastic (i.e. $P_\pi^n e = e$ and $e^\top P_\pi^n = e^\top$).

Admittedly, doubly stochastic matrices are a small portion of the transition matrices considered in RL. They do arise, however, in simple random walks on k-regular, undirected graphs, such as cycles and complete graphs (Levin and Peres, 2017). Furthermore, doubly stochastic matrices are also a popular mathematical model (Section 8.7 Horn and Johnson (2012)).

Theorem 4.15. *Let Assumptions 4.1, 4.2, and 4.14 hold. Then A is strictly positive-stable for every n and $\eta > 0$.*

Proof. By the Lyapunov theorem (Theorem 4.6 in Khalil (2002)), A is positive-stable if and only if there exists a symmetric positive-definite matrix M such that $A^\top M + MA \succ 0$. In the off-policy case we take $M \doteq D_\mu^{-1}$. Then with the definition of A from (11), we have

$$A^\top M + MA = (I - P_\pi^{n\top} + \eta\, ee^\top) + (I - P_\pi^n + \eta\, ee^\top).$$

We now show that $(I - P_\pi^n + \eta ee^\top)$ is positive-definite. When P_π^n is doubly stochastic (so $\|P_\pi^n\| = 1$), for any

nonzero $v \in \mathbb{R}^{|\mathcal{S}|}$,

$$\begin{aligned} v^\top \left(I - P_\pi^n + \eta\, ee^\top\right) v &= v^\top v - v^\top P_\pi^n v + \eta (v^\top e)^2 \\ &\geq \|v\|^2 - \|P_\pi^n\| \|v\|^2 + \eta (v^\top e)^2 \\ &= (1 - \|P_\pi^n\|)\|v\|^2 + \eta (v^\top e)^2 \\ &= \eta (v^\top e)^2 > 0 \end{aligned}$$

where the first inequality holds by Cauchy-Schwarz, and the second equality holds because $\|P_\pi^n\| = 1$. □

Corollary 4.16. *Let Assumptions 4.1, 4.2, and 4.14 hold. Then the iterates* $\{v_t\}$ *in* (5) *satisfy:* $\lim_{t\to\infty} v_t = v_\infty$, *a.s., where* $v_\infty \in \mathcal{V}_*$.

5 CHALLENGES AND OPEN PROBLEMS

We note that our off-policy convergence guarantee in Corollary 4.13 rests on the conservative bound $\eta \leq \eta_0 = 2 \min_{i,j} P_\pi^n(i,j)$, which requires P_π^n to be strictly positive (Assumption 4.11). In this section, we demonstrate that empirically, this estimate for the upper bound of η is quite pessimistic. We consider a 5×5 gridworld and set $n = 3$. Notably, an agent cannot reach every state in three steps. So $\eta_0 = \min_{i,j} P_\pi^3(i,j) = 0$ in this environment and Assumption 4.11 is violated. However, as Figure 1 shows, the algorithm still converges for a wide range of η values.

This empirical result seems to suggest that the convergence can be obtained for any η. Furthermore, Wan et al. (2021b) also prove the convergence for any η with local-clock-based learning rates. Then we might expect that some future work may be able to prove the convergence for any η without the local clock as well. However, we must be cautious here. Anehila and Ran (2022) exhibit M-matrix counterexamples satisfying Conditions 1–3 for which $B + tvw^\top$ fails to remain strictly positive-stable once t exceeds some finite threshold. This implies our Lemma 4.9 *may* not hold for large η. Crucially, the linear algebra community still lacks a tight upper bound on admissible η and has no known necessary and sufficient characterization of triples (B, v, w) that ensure stability under rank-one perturbation (Anehila and Ran, 2022). Closing that theoretical gap would immediately yield sharper convergence guarantees here, and thus represents an important direction for future work.

6 RELATED WORK

Average Reward RL. Several temporal difference methods have been proposed for Markov decision processes with an average-reward objective. The best known is the average reward TD algorithm of Tsitsiklis and Roy (1999), whose convergence guarantees were first analyzed in the linear function approximation case, and further extended to the tabular setting by Blaser and Zhang (2026). The differential TD algorithm we analyze here belongs to the same family but estimates the average reward with the full temporal-difference error instead of only using the reward sample (Wan et al., 2021b). Additional TD-based algorithms for policy evaluation and control in the average-reward setting include Konda and Tsitsiklis (1999); Abounadi et al. (2001); Yang et al. (2016); Wan et al. (2021a); Zhang and Ross (2021); Zhang et al. (2021a,b); He et al. (2023); Saxena et al. (2023); Xie et al. (2025).

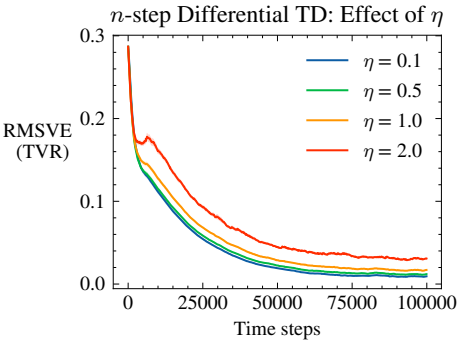

Figure 1: Off-policy convergence of n–step differential TD in a 5×5 gridworld ($n = 3$) for various η. Although $\eta_0 = 0$ here, the algorithm is stable across η. We use a variant of root mean-squared value error from Tsitsiklis and Roy (1999), denoted as 'RMSVE (TVR)', which measures the distance of the estimated values to the nearest solution that satisfies the Bellman equation (3). The trials are averaged over 30 seeds with shaded regions as 1 standard error. The experimental details and results for other n values appear in Appendix C.

Convergence of RL Algorithms. The investigation of the almost sure convergence of RL algorithms is an active area of research. Most prior work relies on the ODE based approach (Benveniste et al., 1990; Kushner and Yin, 2003; Borkar, 2009; Liu et al., 2025a), where the corresponding ODE is relatively easy to analyze (Tsitsiklis and Roy, 1996; Konda and Tsitsiklis, 1999; Sutton et al., 2008, 2009; Zhang et al., 2020a,b; Maei, 2011; Zhang and Whiteson, 2022). By contrast, the ODE studied in this work is highly nontrivial to analyze, and and we still do not have a complete characterization of it. Other notable works involving nontrivial ODEs include Meyn (2024); Wang and Zhang (2024). In addition to the ODE based approach, the Robbins-Siegmund theorem (Robbins and Siegmund, 1971) and its variant (Liu et al., 2025b) are gaining increasing attention for establishing almost sure con-

vergence (Bertsekas and Tsitsiklis, 1996; Qian et al., 2024; Qian and Zhang, 2025; Liu et al., 2025b), and has recently been formally verified (Zhang, 2025). Beyond (asymptotic) almost sure convergence, the L^2 convergence rates of RL algorithms are also widely studied. Notable works include Mahadevan et al. (2014); Liu et al. (2015); Wang et al. (2017); Srikant and Ying (2019); Zou et al. (2019); Wu et al. (2020); Zhang et al. (2022); Xie et al. (2025); Liu et al. (2025c).

Matrix Stability Under Perturbations and D-stability. The stability question in our paper lies within the broader D-stability problem which asks whether a given real matrix $A \in \mathbb{R}^{n\times n}$ remains strictly positive stable under left multiplication by any positive diagonal matrix $D \succ 0$. Despite Johnson's necessary and sufficient criteria in low dimensions (Johnson, 1974b,a), the general case ($n > 4$) remains open, see Hershkowitz (1992) and Kushel (2019) for comprehensive surveys. Our analysis is based on the results of Bierkens and Ran (2014) who investigated the D-stability of M matrices under nonnegative rank-one perturbations. Their work extends a broader area of research investigating the eigenvalues and Jordan structure of rank-one perturbations of matrices (Moro and Dopico, 2003; Savchenko, 2004; Ding and Zhou, 2007; Mehl et al., 2011; Ran and Wojtylak, 2012; Fourie et al., 2013; Mehl et al., 2014; Ran and Wojtylak, 2021).

7 CONCLUSION

Learning rates that use a local clock have played an essential role in the theoretical analysis of differential TD (Wan et al., 2021b), yet they remain largely unused by practitioners. This work bridges that divide by applying $D-$stability and rank-one perturbation theory from the linear algebra community, to provide novel almost sure convergence results of differential TD. To our knowledge, this is the first use of $D-$stability and rank-one perturbation techniques in RL. We expect this approach to enable further theoretical advances in RL, such as convergence proofs for differential Q-learning (Wan et al., 2021b) and for RVI Q-learning (Abounadi et al., 2001) without relying on a local clock.

Acknowledgements

EB acknowledges support from the NSF Graduate Research Fellowship under award 1842490. This work is supported in part by the US National Science Foundation under the awards III-2128019, SLES-2331904, and CAREER-2442098, the Commonwealth Cyber Initiative's Central Virginia Node under the award VV-1Q26-001, a Cisco Faculty Research Award, and an NVIDIA academic grant program award.

References

Abounadi, J., Bertsekas, D., and Borkar, V. S. (2001). Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*.

Anehila, B. and Ran, A. (2022). A note on a conjecture concerning rank one perturbations of singular m-matrices. *Quaestiones Mathematicae*.

Bakhshi, B., Mangues-Bafalluy, J., and Baranda, J. (2021). R-learning-based admission control for service federation in multi-domain 5g networks. In *IEEE Global Communications Conference*.

Bellman, R. (1957). A markovian decision process. *Journal of Mathematics and Mechanics*.

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific Belmont, MA.

Bierkens, J. and Ran, A. (2014). A singular M-matrix perturbed by a nonnegative rank one matrix has positive principal minors; is it D-stable? *Linear Algebra and Its Applications*.

Blaser, E. and Zhang, S. (2026). Asymptotic and finite sample analysis of nonexpansive stochastic approximations with markovian noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*. Springer.

Borkar, V. S. and Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*.

Chen, Z. (2025). Non-asymptotic guarantees for average-reward Q-learning with adaptive stepsizes. *ArXiv Preprint*.

Ding, J. and Zhou, A. (2007). Eigenvalues of rank-one updated matrices with some applications. *Applied Mathematics Letters*.

Fourie, J., Groenewald, G., van Rensburg, D. J., and Ran, A. (2013). Rank one perturbations of h-positive real matrices. *Linear Algebra and its Applications*.

Ghavamzadeh, M. and Mahadevan, S. (2007). Hierarchical average reward reinforcement learning. *Journal of Machine Learning Research*.

Giorgi, G. and Zuccotti, C. (2015). An overview on d-stable matrices. *Department of Economics and Management DEM Working Paper Series*.

He, J., Che, F., Wan, Y., and Mahmood, A. R. (2023). Loosely consistent emphatic temporal-difference learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Hershkowitz, D. (1992). Recent directions in matrix stability. *Linear Algebra and its Applications*.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis (2nd Edition)*. Cambridge university press.

Jaakkola, T., Jordan, M., and Singh, S. (1993). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*.

Johnson, C. R. (1974a). D-stability and real and complex quadratic forms. *Linear Algebra and its Applications*.

Johnson, C. R. (1974b). Sufficient conditions for d-stability. *Journal of Economic Theory*.

Khalil, H. K. (2002). *Nonlinear Systems*. Prentice Hall.

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *International Journal of Robotics Research*.

Konda, V. R. and Tsitsiklis, J. N. (1999). Actor-critic algorithms. In *Advances in Neural Information Processing Systems*.

Kushel, O. Y. (2019). Unifying matrix stability concepts with a view to applications. *SIAM Review*.

Kushel, O. Y. and Pavani, R. (2023). Novel versions of d-stability in matrices provide new insights into ode dynamics. *Mediterranean Journal of Mathematics*.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*. American Mathematical Soc.

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., and Petrik, M. (2015). Finite-sample analysis of proximal gradient TD algorithms. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Liu, S., Chen, S., and Zhang, S. (2025a). The ODE method for stochastic approximation and reinforcement learning with markovian noise. *Journal of Machine Learning Research*.

Liu, X., Xie, Z., and Zhang, S. (2025b). Extensions of robbins-siegmund theorem with applications in reinforcement learning. *ArXiv Preprint*.

Liu, X., Xie, Z., and Zhang, S. (2025c). Linear Q-learning does not diverge in L^2: Convergence rates to a bounded set. In *Proceedings of the International Conference on Machine Learning*.

Maei, H. R. (2011). *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta.

Mahadevan, S., Liu, B., Thomas, P. S., Dabney, W., Giguere, S., Jacek, N., Gemp, I., and Liu, J. (2014). Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces. *ArXiv Preprint*.

Marbach, P., Mihatsch, O., and Tsitsiklis, J. N. (1998). Call admission control and routing in integrated services networks using reinforcement learning. In *Proceedings of the IEEE Conference on Decision and Control*.

Mehl, C., Mehrmann, V., Ran, A. C., and Rodman, L. (2011). Eigenvalue perturbation theory of classes of structured matrices under generic structured rank one perturbations. *Linear Algebra and Its Applications*.

Mehl, C., Mehrmann, V., Ran, A. C., and Rodman, L. (2014). Eigenvalue perturbation theory of symplectic, orthogonal, and unitary matrices under generic structured rank one perturbations. *BIT Numerical Mathematics*.

Meyn, S. (2024). The projected bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*.

Moro, J. and Dopico, F. M. (2003). Low rank perturbation of jordan structure. *SIAM Journal on Matrix Analysis and Applications*.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qian, X., Xie, Z., Liu, X., and Zhang, S. (2024). Almost sure convergence rates and concentration of stochastic approximation and reinforcement learning with markovian noise. *ArXiv Preprint*.

Qian, X. and Zhang, S. (2025). Revisiting a design choice in gradient temporal difference learning. In *Proceedings of the International Conference on Learning Representations*.

Ran, A. C. and Wojtylak, M. (2012). Eigenvalues of rank one perturbations of unstructured matrices. *Linear Algebra and its Applications*.

Ran, A. C. and Wojtylak, M. (2021). Global properties of eigenvalues of parametric rank one perturbations for unstructured and structured matrices. *Complex Analysis and Operator Theory*.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. *Optimizing Methods in Statistics*.

Savchenko, S. V. (2004). On the change in the spectral properties of a matrix under perturbations of sufficiently low rank. *Functional Analysis and Its Applications*.

Saxena, N., Khastagir, S., Kolathaya, S., and Bhatnagar, S. (2023). Off-policy average reward actor-critic with deterministic policy search. In *Proceedings of the International Conference on Machine Learning*.

Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation andtd learning. In *Proceedings of the Conference on Learning Theory*.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction (2nd Edition)*. MIT press.

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

Sutton, R. S., Szepesvári, C., and Maei, H. R. (2008). A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*.

Tong, Y. and Su, S. W. (2024). Sufficient d-stability conditions for non-square matrices. *ArXiv Preprint*.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*.

Tsitsiklis, J. N. and Roy, B. V. (1996). Analysis of temporal-diffference learning with function approximation. In *IEEE Transactions on Automatic Control*.

Tsitsiklis, J. N. and Roy, B. V. (1999). Average cost temporal-difference learning. *Automatica*.

Wan, Y., Naik, A., and Sutton, R. (2021a). Average-reward learning and planning with options. In *Advances in Neural Information Processing Systems*.

Wan, Y., Naik, A., and Sutton, R. S. (2021b). Learning and planning in average-reward markov decision processes. In *Proceedings of the International Conference on Machine Learning*.

Wang, J. and Zhang, S. (2024). Almost sure convergence of linear temporal difference learning with arbitrary features. *ArXiv Preprint*.

Wang, Y., Chen, W., Liu, Y., Ma, Z., and Liu, T. (2017). Finite sample analysis of the GTD policy evaluation algorithms in markov setting. In *Advances in Neural Information Processing Systems*.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. (2020). A finite-time analysis of two time-scale actor-critic methods. In *Advances in Neural Information Processing Systems*.

Xie, Z., Liu, X., Chandra, R., and Zhang, S. (2025). Finite sample analysis of linear temporal difference learning with arbitrary features. In *Advances in Neural Information Processing Systems*.

Yang, K., Yang, J., and Shen, C. (2024). Average reward reinforcement learning for wireless radio resource management. In *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*.

Yang, S., Gao, Y., An, B., Wang, H., and Chen, X. (2016). Efficient average reward reinforcement learning using constant shifting values. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, S. (2025). Towards formalizing reinforcement learning theory. *ArXiv Preprint*.

Zhang, S., Liu, B., and Whiteson, S. (2020a). GradientDICE: Rethinking generalized offline estimation of stationary values. In *Proceedings of the International Conference on Machine Learning*.

Zhang, S., Liu, B., Yao, H., and Whiteson, S. (2020b). Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proceedings of the International Conference on Machine Learning*.

Zhang, S., Tachet, R., and Laroche, R. (2022). Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *Journal of Machine Learning Research*.

Zhang, S., Wan, Y., Sutton, R. S., and Whiteson, S. (2021a). Average-reward off-policy policy evaluation with function approximation. In *Proceedings of the International Conference on Machine Learning*.

Zhang, S. and Whiteson, S. (2022). Truncated emphatic temporal difference methods for prediction and control. *Journal of Machine Learning Research*.

Zhang, S., Yao, H., and Whiteson, S. (2021b). Breaking the deadly triad with a target network. In *Proceedings of the International Conference on Machine Learning*.

Zhang, Y. and Ross, K. W. (2021). On-policy deep reinforcement learning for the average-reward criterion. In *Proceedings of the International Conference on Machine Learning*.

Zou, S., Xu, T., and Liang, Y. (2019). Finite-sample analysis for SARSA with linear function approximation. In *Advances in Neural Information Processing Systems*.

Checklist

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. No, but we merely are analyzing an algorithm that exists, not proposing a new one.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, we provide a link in Appendix D.

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. Yes

(b) Complete proofs of all theoretical results. Yes

(c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, link provided in Appendix D.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes. See Appendix D

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes. See Appendix D

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). No, but all experiments ran on a standard laptop.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. Not Applicable

(b) The license information of the assets, if applicable. Not Applicable

(c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable

(d) Information about consent from data providers/curators. Not Applicable

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. Not Applicable

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Appendix

A MATHEMATICAL BACKGROUND

A.1 Main Results from Liu et al. (2025a)

First we will restate the main results from Liu et al. (2025a) concerning the convergence of SA iterates of the form (8) for completeness.

Assumption A.1. The Markov chain $\{Y_n\}$ has a unique invariant probability measure (i.e. stationary distribution), denoted by $d_{\mathcal{Y}}$.

Assumption A.2. The learning rates $\{\alpha_n\}$ are positive, decreasing and satisfy

$$\sum_{i=0}^{\infty} \alpha_i = \infty, \lim_{n\to\infty} \alpha_n = 0, \text{ and } \frac{\alpha_n - \alpha_{n+1}}{\alpha_n} = \mathcal{O}(\alpha_n).$$

Assumption A.3. Let $H_c(x,y) \doteq \frac{1}{c}H(cx,y)$. There exists a measurable function $H_\infty(x,y) \doteq \lim_{c\to\infty} H_c(x,y)$, a scalar function $\kappa\colon \mathbb{R}\to\mathbb{R}$ (independent of x,y), and a measurable function $b(x,y)$ such that for any x,y:

$$\begin{aligned} H_c(x,y) - H_\infty(x,y) &= \kappa(c)b(x,y), \\ \lim_{c\to\infty} \kappa(c) &= 0. \end{aligned} \tag{15}$$

Moreover, there exists a measurable function $L_b(y)$ such that for all x, x', y,

$$\big\|b(x,y) - b(x',y)\big\| \le L_b(y)\,\|x - x'\|, \tag{16}$$

and the expectation,

$$L_b \doteq \mathbb{E}_{y\sim d_y}\big[L_b(y)\big]$$

is well-defined and finite.

Assumption A.4. There exists a measurable function $L(y)$ such that for any x, x', y,

$$\big\|H(x,y) - H(x',y)\big\| \le L(y)\,\|x - x'\|, \tag{17}$$

$$\big\|H_\infty(x,y) - H_\infty(x',y)\big\| \le L(y)\,\|x - x'\|. \tag{18}$$

Moreover, the following expectations are well-defined and finite for every x:

$$h(x) \doteq \mathbb{E}_{y\sim d_y}\big[H(x,y)\big], \tag{19}$$

$$h_\infty(x) \doteq \mathbb{E}_{y\sim d_y}\big[H_\infty(x,y)\big], \tag{20}$$

$$L \doteq \mathbb{E}_{y\sim d_y}\big[L(y)\big]. \tag{21}$$

Assumption A.5. As $c\to\infty$, $h_c(x)$ converges to $h_\infty(x)$ uniformly on x on any compact subsets of $\mathbb{R}^d$. The ODE,

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = h_\infty(x(t))$$

has 0 as its G.A.S equilibrium.

Assumption A.6. Let g denote any of the following functions:

$$y \mapsto H(x,y) \quad (\forall x),$$
$$y \mapsto L_b(y),$$
$$y \mapsto L(y).$$

Then for any initial condition Y_1, it holds that

$$\lim_{n\to\infty} \alpha_n \sum_{i=1}^{n} \Big(g(Y_i) - \mathbb{E}_{y\sim d_y}[g(y)] \Big) = 0 \quad \text{a.s.}$$

Lemma A.7. *Let Assumptions A.1 - A.6 hold. Then the iterates $\{x_n\}$ generated by (8) converge almost surely to a (sample-path-dependent) bounded invariant set of the ODE*

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = h(x(t)).$$

B DERIVATION OF N-STEP DIFFERENTIAL TD

We begin with the Bellman equation for differential TD in matrix-vector form (3). Recall that $P_\pi \in [0,1]^{|\mathcal{S}|\times|\mathcal{S}|}$ denotes the stochastic matrix of the Markov chain induced by the target policy π, $r_\pi \in \mathbb{R}^{|\mathcal{S}|}$ represents the expected rewards under π, and $\bar{J}_\pi$ is the average reward. The one-step Bellman equation of v_π is

$$v_\pi = r_\pi - \bar{J}\pi e + P_\pi v_\pi.$$

We can keep unrolling it for n steps and get

$$\begin{aligned} v_\pi &= r_\pi - \bar{J}_\pi e + P_\pi\big(r_\pi - \bar{J}_\pi e + P_\pi v_\pi\big) \\ &= r_\pi - 2\bar{J}_\pi e + P_\pi r_\pi + P_\pi^2 v_\pi \\ &\;\;\vdots \\ &= -n\bar{J}_\pi e + r_\pi + P_\pi r_\pi + P_\pi^2 r_\pi + \cdots + P_\pi^{n-1} r_\pi + P_\pi^n v_\pi \end{aligned} \tag{22}$$

Hence, by (22), we have for all $s \in \mathcal{S}$,

$$\begin{aligned} v_\pi(s) =& \mathbb{E}\left[\sum_{i=0}^{n-1} \big(r(S_i,A_i) - \bar{J}_\pi \big) + v_\pi(S_n) \middle| S_0 = s, A_i \sim \pi(\cdot \mid S_i), S_{i+1} \sim p(\cdot \mid S_i, A_i) \right] \\ =& v_\pi(s) + \mathbb{E}\left[\sum_{i=0}^{n-1} \big(r(S_i,A_i) - \bar{J}_\pi \big) + v_\pi(S_n) - v_\pi(S_0) \middle| S_0 = s, A_i \sim \pi(\cdot \mid S_i), S_{i+1} \sim p(\cdot \mid S_i, A_i) \right] \\ =& v_\pi(s) + \mathbb{E}\left[\rho_{0:n-1} \left(\sum_{i=0}^{n-1} \big(r(S_i,A_i) - \bar{J}_\pi \big) + v_\pi(S_n) - v_\pi(S_0) \right) \middle| S_0 = s, A_i \sim \mu(\cdot \mid S_i), S_{i+1} \sim p(\cdot \mid S_i, A_i) \right]. \end{aligned}$$

Therefore, we have the n-step bootstrapped differential TD update

$$v_{t+n}(S_t) = v_{t+n-1}(S_t) + \alpha_{t+n-1}\rho_{t:t+n-1}(R_{t+1:t+n} - nJ_{t+n-1} + v_{t+n-1}(S_{t+n}) - v_{t+n-1}(S_t))$$

where J_t is the average reward estimate to be defined shortly.

Let $d_\mu \in [0,1]^{|\mathcal{S}|}$ denote the stationary distribution induced by the behavior policy μ. Rearranging (22), we get

$$\begin{aligned} n\bar{J}_\pi e &= r_\pi + P_\pi r_\pi + P_\pi^2 r_\pi + \cdots + P_\pi^{n-1} r_\pi + P_\pi^n v_\pi - v_\pi \\ \bar{J}_\pi e &= \frac{1}{n}\left(r_\pi + P_\pi r_\pi + P_\pi^2 r_\pi + \cdots + P_\pi^{n-1} r_\pi + P_\pi^n v_\pi - v_\pi \right) \\ d_\mu^\top \bar{J}_\pi e &= \frac{d_\mu^\top}{n}\left(r_\pi + P_\pi r_\pi + \cdots + P_\pi^{n-1} r_\pi + P_\pi^n v_\pi - v_\pi \right) \\ \bar{J}_\pi &= \frac{d_\mu^\top}{n}\left(r_\pi + P_\pi r_\pi + \cdots + P_\pi^{n-1} r_\pi + P_\pi^n v_\pi - v_\pi \right) \end{aligned} \tag{23}$$

Therefore, by (23), we have

$$\begin{aligned}\bar{J}_\pi =&\frac{1}{n}\mathbb{E}\left[\sum_{i=0}^{n-1} r(S_i,A_i)+v_\pi(S_n)-v_\pi(S_0)\middle| S_0\sim d_\pi, A_i\sim\pi(\cdot\mid S_i), S_{i+1}\sim p(\cdot\mid S_i,A_i)\right]\\ =&\bar{J}_\pi+\frac{1}{n}\mathbb{E}\left[\sum_{i=0}^{n-1}\big(r(S_i,A_i)-\bar{J}_\pi\big)+v_\pi(S_n)-v_\pi(S_0)\middle| S_0\sim d_\pi, A_i\sim\pi(\cdot\mid S_i), S_{i+1}\sim p(\cdot\mid S_i,A_i)\right]\\ =&\bar{J}_\pi+\frac{1}{n}\mathbb{E}\left[\rho_{0:n-1}\left(\sum_{i=0}^{n-1}\big(r(S_i,A_i)-\bar{J}_\pi\big)+v_\pi(S_n)-v_\pi(S_0)\right)\middle| S_0\sim d_\mu, A_i\sim\mu(\cdot\mid S_i), S_{i+1}\sim p(\cdot\mid S_i,A_i)\right].\end{aligned}$$

As a result, we update the average reward estimate as

$$J_{t+n}=J_{t+n-1}+\frac{\eta}{n}\alpha_{t+n-1}\rho_{t:t+n-1}(R_{t+1:t+n}-nJ_{t+n-1}+v_{t+n-1}(S_{t+n})-v_{t+n-1}(S_t)),$$

where η is a positive multiplicative constant to allow for a different update rate relative to v.

C TECHNICAL LEMMAS

Lemma C.1. *The function $H(v,y)$ defined in (7) satisfies Assumption A.3.*

Proof. Recall that for a scalar $c>0$, the scaled operator is defined as $H_c(x,y)\doteq\frac{1}{c}H(cx,y)$. Substituting the definition of H from (7), we obtain

$$H_c(v,y)[s]=\rho_{0:n-1}(y)\left(\frac{1}{c}\sum_{k=0}^{n-1}r(s_k,a_k)-\eta\sum v+v(s_n)-v(s_0)\right)\mathbb{I}\{s=s_0\}.$$

This implies that the limit of the operator $H_\infty(v,y)=\lim_{c\to\infty}H_c(v,y)$ is given by

$$H_\infty(v,y)[s]\doteq\rho_{0:n-1}(y)\Big(-\eta\sum v+v(s_n)-v(s_0)\Big)\mathbb{I}\{s=s_0\},\tag{24}$$

and the difference $H_c(v,y)-H_\infty(v,y)$ simplifies to

$$H_c(v,y)[s]-H_\infty(v,y)[s]=\rho_{0:n-1}(y)\left(\frac{1}{c}\sum_{k=0}^{n-1}r(s_k,a_k)\right)\mathbb{I}\{s=s_0\}.$$

Therefore, Assumption A.3 (15) is satisfied by setting $\kappa(c)\doteq\frac{1}{c}$, which vanishes as $c\to\infty$, and defining

$$b(x,y)[s]\doteq\rho_{0:n-1}(y)\left(\sum_{k=0}^{n-1}r(s_k,a_k)\right)\mathbb{I}\{s=s_0\}.$$

Importantly, $b(x,y)$ is independent of x, so for all x,x',y we have $\|b(x,y)-b(x',y)\|=0$. Thus, the Lipschitz condition in (16) from Assumption A.3 is trivially satisfied with $L_b(y)=0$, and the expectation $\mathbb{E}[L_b(y)]$ is finite.

□

Lemma C.2. *The function $H(v,y)$ defined in (7) satisfies Assumption A.4.*

Proof. We first verify that H is Lipschitz continuous in the $\infty-$norm, i.e. (17). Fix any transition $y=(s_0,a_0,s_1,\dots s_n)$ and any state s. Using the indicator ($\mathbb{I}\{s=s_0\}$), we have

$$\begin{aligned}H(v)[s]-H(w)[s]=\rho(y)\Big[&\left(\sum_{k=0}^{n-1}r(s_k,a_k)-\eta\sum v+v(s_n)-v(s_0)\right)\\ &-\left(\sum_{k=0}^{n-1}r(s_k,a_k)-\eta\sum w+w(s_n)-w(s_0)\right)\Big]\mathbb{I}\{s=s_0\}\\ =&\begin{cases}\rho(y)\big[-\eta\sum(v-w)+(v(s_n)-w(s_n))-(v(s_0)-w(s_0))\big], & s=s_0,\\ 0, & s\neq s_0.\end{cases}\end{aligned}$$

Hence

$$\big|H(v)[s] - H(w)[s]\big| \le \begin{cases} \rho(y)\big[\eta|\mathcal{S}|\|v-w\|_\infty + 2\,\|v-w\|_\infty\big], & s = s_0, \\ \|v-w\|_\infty, & s \neq s_0. \end{cases}$$

The only remaining task is to upper-bound $\rho(y)$ defined in (7). Under our standard "coverage" assumption (Assumption 4.1), whenever $\pi(a|s) > 0$, we also have $\mu(a|s) > 0$, and because both $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite, there is a uniform lower bound

$$\mu_{\min} = \min_{s,a:\,\pi(a|s)>0} \mu(a \mid s) > 0.$$

This implies that

$$\rho(y) = \prod_{k=0}^{n-1} \frac{\pi(a_k \mid s_k)}{\mu(a_k \mid s_k)} \le \left(\frac{1}{\mu_{\min}}\right)^n \doteq \bar{\rho} < \infty.$$

Therefore,

$$\begin{aligned} \|H(v,y) - H(w,y)\|_\infty &= \max_{s\in\mathcal{S}} |H(v,y)[s] - H(w,y)[s]| \\ &\le L\|v-w\|_\infty, \end{aligned}$$

where $L \doteq \max\{1, \bar{\rho}(\eta|\mathcal{S}|+2)\}$.

Verifying the Lipschitz continuity of H_∞, starting from (24) from the proof of Lemma C.1 to avoid repetition, we have,

$$\begin{aligned} H_\infty(v)[s] - H_\infty(w)[s] &= \rho(y)\Big[\Big(-\eta\sum v + v(s_n) - v(s_0)\Big) - \Big(-\eta\sum w + w(s_n) - w(s_0)\Big)\Big]\mathbb{I}\{s = s_0\} \\ &= \begin{cases} \rho(y)\big[-\eta\sum(v-w) + (v(s_n) - w(s_n)) - (v(s_0) - w(s_0))\big], & s = s_0, \\ 0, & s \neq s_0. \end{cases} \\ &= H(v)[s] - H(w)[s] \end{aligned}$$

Therefore, the Lipschitz continuity of H_∞ holds for L as well, proving (18) holds.

From equation (10), its easy to see that h is finite, verifying (19). To verify (20), its also easy to see that

$$\begin{aligned} \mathbb{E}_{y\sim d_\mathcal{Y}}\big[H_\infty(v,y)[s]\big] &= \mathbb{E}_\mu\Big[\rho_{0:n-1}(y)\,\big(-\eta\textstyle\sum_p v(p) + v(s_n) - v(s_0)\big)I\{s = s_0\}\Big] \\ &= d_\mu(s)\,\mathbb{E}_\pi\Big[-\eta\, e^\top v + v(s_n) - v(s) \;\Big|\; s_0 = s\Big] \\ &= d_\mu(s)\Big[-\eta\, e^\top v + (P_\pi^n v)(s) - v(s)\Big] \\ &= \big[D_\mu\big(P_\pi^n - I - \eta\, ee^\top\big)v\big](s). \end{aligned}$$

In vector form,

$$\mathbb{E}_{y\sim d_\mathcal{Y}}\big[H_\infty(v,y)\big] = D_\mu\big(P_\pi^n - I - \eta\, ee^\top\big)v,$$

which is clearly finite. Finally, since the Lipschitz constant L is independent of y, (21) holds trivially. □

D EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

D.1 Experimental Details

Our experiments were carried out in a simple, continuing gridworld. At each timestep t the agent occupies one cell in an 5×5 grid, can move to any of the four orthogonal neighbors (subject to walls at the borders where an illegal move keeps the agent in the same state), and receives a unit reward each time it reaches the designated

"goal" state in the bottom right corner. Rather than terminating, after reaching the goal state, the agent is put back to the start corner (top left) on the very next step with probability 1, making this a continuing task.

We evaluated an off-policy, multi-step differential TD learner for a fixed ϵ-greedy target policy with $\epsilon = 0.1$. The behavior policy is random. The value function and reward estimate were both initialized to zero. Throughout training, the agent took actions following the random behavior policy, observed the deterministic next state and reward, and performed its n-step TD updates with Off-Policy Differential TD to estimate the value function associated with the target policy. The experiment was run for $100,000$ steps, and repeated with 30 random seeds. For Figure 1, the η values were reported for $\eta = \{0.1, 0.5, 1, 2\}$, while keeping $n = 3$. In Figure 2, $\eta = 0.1$ and $n = \{1, 2, 3, 4\}$. In both experiments, we used a constant learning rate $\alpha = 0.01$. The source code can be found at `https://github.com/blaserethan/Differential-TD`.

We now elaborate on the evaluation metric. The convergence was assessed using a variant of root-mean-squared value error from (Tsitsiklis and Roy, 1999) which we denote as 'RMSVE (TVR)', which is also used in (Wan et al., 2021b). As noted in Section 2, the solutions to the differential Bellman equation (3) form a set $\mathcal{V}_* = \{v_\pi + ce\}$. Which point in this set an algorithm converges to depends on initializations and the design choices of the algorithm. Therefore, computing the value error with respect to v_π does not say much about convergence. To remedy this, (Tsitsiklis and Roy, 1999) proposed computing the error with respect to the nearest valid solution to the Bellman equations. The metric is defined as

$$\text{RMSVE(TVR)}(v, v_\pi) \doteq \inf_c \|v - (v_\pi + ce)\|_{d_\pi}.$$

Algorithmically, this amounts to computing the offset of the learned value function, subtracting it, and then computing the RMSVE with respect to v_π. As demonstrated in Section C.4 of Wan et al. (2021b) this v_π can be analytically computed using the Bellman equations with the additional constraint that $d_\pi^\top v_\pi = 0$ (effectively centering the value function).

D.2 Additional Experiment on the effect of n

Here, we present additional experiments, demonstrating that the conclusions from Section 5 also hold for various n values. Despite the fact that $\eta_0 = 0$, we empirically observe the convergence of differential TD with $\eta = 0.1$ under several choices of n. In this experiment, we used the environment described in Section D.1.

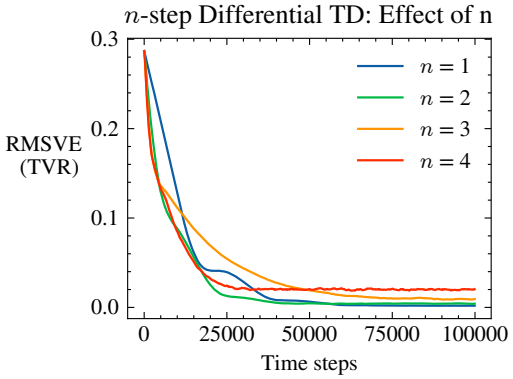

Figure 2: Off-policy convergence of n–step differential TD in a 5×5 gridworld for various n with fixed $\eta = 0.1$. See Section D.1 for the complete experiment description. Despite $\eta_0 = 0$, we still observe the convergence of differential TD with across several n values.