On the Divergence of Differential Temporal Difference Learning without Local Clocks

David Antrobius
Department of Computer Science
University of Virginia
davidantrobius@email.virginia.edu

Shangtong Zhang
Department of Computer Science
University of Virginia
shangtong@virginia.edu

Abstract

Learning rate is a critical component of reinforcement learning (RL). This work uses global and local clocks to distinguish two types of learning rates. The former is of the standard form α_t that depends only on the time step t (i.e., a global clock). The latter is of the form $\alpha_{\nu(S_t,t)}$, where $\nu(s,t)$ counts the number of visits to state s until time t (i.e., a local clock). In discounted RL, an RL algorithm that is convergent with a local clock is always also convergent with a global clock, and vice versa. We are not aware of any counterexample. The key contribution of this work is to show that this nice correspondence breaks down in average-reward RL. Specifically, we construct a counterexample showing that although differential temporal difference (TD) learning is convergent with a local clock, it can diverge with a global clock. We empirically observe this divergence in our experiments and also prove that the associated ODE of differential TD with a global clock is unstable. This counterexample closes the open problem in Wan et al. [2021], Blaser et al. [2026].

1 Introduction

Tabular methods form a fundamental class of algorithms in reinforcement learning (RL, Sutton and Barto [2018]). Looking back at the history of tabular RL, we can see a clear pattern that early theoretical analysis of tabular RL methods often uses a local clock in the learning rate. Here, by a local clock in a learning rate, we mean a learning rate of the form $\alpha_{\nu(S_t,t)}$, where $\nu(s,t)$ counts the number of visits to state s until time t. This is in contrast to a learning rate with a global clock, which is of the form α_t that depends only on the time step t. The global clock refers to the time step t in the learning rate, while the local clock refers to the state-dependent visit count $\nu(s,t)$ in the learning rate. For example, classical convergence proofs for Q-learning usually use local clocks [Watkins, 1989, Watkins and Dayan, 1992, Jaakkola et al., 1993, Tsitsiklis, 1994, Bertsekas and Tsitsiklis, 1996].

We argue that the use of local clocks in theoretical analysis is mostly because of technical convenience. Chapter 7 of Borkar [2009] gives a thorough exposition of the role of local clocks in the convergence analysis of stochastic approximation algorithms. Intuitively, local clocks allow more balanced updates across states, essentially converting the Markovian samples in RL into i.i.d. samples, which significantly simplifies the analysis. However, a local clock is not really a practitioner's first choice. For example, the entire textbook Sutton and Barto [2018] does not mention or use any local clock in the learning rates. A global clock is preferred by most practitioners.

This theorist-practitioner gap is not a problem in the discounted RL setting. Although early theoretical analysis of discounted RL methods often uses local clocks, recent works have shown that the same algorithms are also convergent with global clocks. For example, the convergence of Q-learning with a global clock is also established by Lee and He [2020], Chen et al. [2024], Liu et al. [2025a]. To our

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

knowledge, there is no known counterexample showing that an algorithm that is convergent with a local clock can diverge with a global clock in discounted RL.

In average-reward RL, early theoretical analysis such as RVI Q-learning also uses local clocks in the learning rates [Abounadi et al., 2001]. The seminal work Wan et al. [2021] develops a new set of temporal difference algorithms for average-reward RL that removes the need for a reference state in RVI Q-learning. The (off-policy) policy evaluation algorithm in Wan et al. [2021] is called differential TD. And the control algorithm in Wan et al. [2021] is called differential Q-learning. In the main text of Wan et al. [2021], both algorithms are presented with a global clock in the learning rates. However, in the convergence analysis in the appendix, both algorithms are shown to be convergent only with a local clock in the learning rates. So the open problem is, do these algorithms also converge with a global clock in the learning rates? Blaser et al. [2026] partially answer this question by showing that when η is sufficiently small, differential TD is convergent even with a global clock. Here η is a hyperparameter in differential TD that we will introduce later. However, there is still a gap because Wan et al. [2021] show that with a local clock, differential TD is convergent for any $\eta > 0$. Does differential TD also converge with a global clock for any $\eta > 0$? This is an open problem from Blaser et al. [2026].

One may optimistically conjecture that the answer to the above question is yes. After all in the discounted RL setting, we have seen that algorithms that are convergent with a local clock are also convergent with a global clock. We are not aware of any counterexample in the literature. The key contribution of this work is to show that this nice correspondence breaks down in average-reward RL. Particularly, we construct a counterexample showing that differential TD can diverge with a global clock. Empirically, we observe this divergence in our experiments. Theoretically, we prove that the ODE associated with differential TD with a global clock can be unstable. This negative result is surprising and interesting because it shows that the choice of learning rates can have a much more significant impact on the convergence of RL algorithms in the average-reward setting than in the discounted setting.

Our contribution is beyond the counterexample itself. Specifically, we provide a detailed analysis of the stability of the ODE associated with differential TD with a global clock. Beyond the sufficient small-η condition of Blaser et al. [2026], we characterize the first loss of positive stability and the structure of the full stability region for that ODE. Our counterexample shows that this region can be disconnected, with stability lost and then recovered as η increases.

2 Background

We consider an infinite horizon Markov Decision Process (MDP, Bellman [1957]) defined by a tuple $(\mathcal{S}, \mathcal{A}, p, p_0, r)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the transition function, $p_0 : \mathcal{S} \to [0, 1]$ is the initial state distribution, and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. At time step 0, an initial state S_0 is drawn from p_0. At each time step $t \geq 0$, an action A_t is chosen according to a policy $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$ as $A_t \sim \pi(\cdot|S_t)$. Then a reward $R_{t+1} = r(S_t, A_t)$ is emitted and the next state S_{t+1} is drawn from $p(\cdot|S_t, A_t)$.

In the discounted RL setting, we consider a discount factor $\gamma \in [0, 1)$. The goal of policy evaluation is to estimate the value function for a given policy π, defined as $v_\pi(s) \doteq \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i+1} | S_t = s\right]$. In the average-reward RL setting, we consider the average reward $J_\pi \doteq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} R_{t+1}\right]$ and are interested in the differential value function of a policy π, defined as $\bar{v}_\pi(s) \doteq \mathbb{E}[\sum_{i=0}^{\infty}(R_{t+i+1} - J_\pi)|S_t = s]$. We assume that the Markov chain induced by the target policy is irreducible, so J_π is independent of the initial state.

Temporal difference (TD) learning is the most fundamental class of algorithms for policy evaluation in RL. In this work, we consider off-policy TD, where the data is generated by a behavior policy μ (i.e., $A_t \sim \mu(\cdot|S_t)$) while we want to estimate the value function for a different target policy π. Let $\{\alpha_n\}_{n\geq 0}$ denote the deterministic base learning-rate sequence. In the discounted setting, the TD update is given by

$$\begin{aligned} \delta_t &= R_{t+1} + \gamma v_t(S_{t+1}) - v_t(S_t), \qquad (1) \\ v_{t+1}(s) &= \begin{cases} v_t(s) + \alpha_t \rho_t \delta_t, & \text{if } s = S_t, \\ v_t(s), & \text{otherwise} \end{cases}, \end{aligned}$$

where $\rho_t = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$ is the importance sampling ratio. The vector $v_t \in \mathbb{R}^{|\mathcal{S}|}$ estimates the discounted value function v_π. Here, α_t is the learning rate, e.g., $\alpha_t = \frac{1}{t+1}$ and we refer to (1) as *discounted TD with a global clock*. Let $\nu(s,t) \doteq \sum_{i=0}^{t} \mathbb{I}_{\{S_i=s\}}$ be the number of times state s has been visited by time t with $\mathbb{I}$ being the indicator function. If we replace α_t with $\alpha_{\nu(S_t,t)}$ in (1), we get *discounted TD with a local clock*. Both clocks use asynchronous value updates: at each time step, only the component corresponding to the visited state is updated. Discounted TD converges with either clock under the corresponding chain, coverage, and learning-rate conditions stated in Section 3 and Appendix C.

Wan et al. [2021] extend off-policy TD to the average-reward setting and propose the following update for differential TD:

$$\begin{aligned} \delta_t &= R_{t+1} - \hat{J}_t + v_t(S_{t+1}) - v_t(S_t), \\ \hat{J}_{t+1} &= \hat{J}_t + \alpha_t \eta \rho_t \delta_t, \\ v_{t+1}(s) &= \begin{cases} v_t(s) + \alpha_t \rho_t \delta_t, & \text{if } s = S_t, \\ v_t(s), & \text{otherwise} \end{cases}, \end{aligned} \tag{2}$$

where $\eta > 0$ is a hyperparameter and $\hat{J}_t$ is an estimate of the average reward J_π. The vector v_t estimates the differential value function $\bar{v}_\pi$ up to an additive constant. The learning rate for $\hat{J}_t$ is η times that for $v_t(S_t)$. We refer to (2) as *differential TD with a global clock*. Similarly, by replacing both appearances of α_t in (2) with $\alpha_{\nu(S_t,t)}$, we get *differential TD with a local clock*. Wan et al. [2021] prove the almost sure convergence of differential TD with a local clock for any $\eta > 0$ in their appendix under the assumptions summarized in Appendix C, although their main text only presents differential TD with a global clock and does not mention the local clock at all. Blaser et al. [2026] show that when η is sufficiently small, differential TD with a global clock is also convergent under their stated assumptions; Appendix C states sufficient conditions used there. It is thus an open problem whether differential TD with a global clock is convergent for an η that is not small enough. Surprisingly, we will show in this paper that the answer is negative: the divergence of the algorithm is empirically observed and the instability of the associated ODE is proved.

We now introduce some definitions for matrix analysis.

Definition 2.1. A matrix A is called an M-matrix if A can be expressed as $A = sI - B$, where $s > 0$, $B \geq 0$ (i.e., all entries of B are nonnegative), and $s \geq \rho(B)$ with $\rho(B)$ being the spectral radius of B.

Definition 2.2. A matrix A is called positive stable if all eigenvalues of A have strictly positive real parts. A matrix A is Hurwitz if $-A$ is positive stable.

3 How Is Differential TD Different from Discounted TD?

We now provide some necessary background to illustrate the fundamental difference between how discounted and differential TD interact with global and local clocks. This is largely due to Blaser et al. [2026] and we provide details here only for completeness and to set the stage for our counterexample in the next section. For the discussion below, we assume that the target and behavior chains are irreducible and that $\pi(a|s) > 0$ implies $\mu(a|s) > 0$ for every state-action pair.

The ODE method is a powerful tool for establishing the almost sure convergence of RL algorithms. Take discounted TD with a local clock as an example. To apply the ODE method, we study the ODE@∞ associated with discounted TD with a local clock, which is given by

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -(I - \gamma P_\pi)x(t),$$

where I is the identity matrix and P_π is the transition matrix under policy π, i.e., $P_\pi(s,s') = \sum_a \pi(a|s)p(s'|s,a)$. The standard ODE method (e.g., Borkar [2009], Borkar et al. [2025], Liu et al. [2025a]) needs to verify that the above ODE is globally asymptotically stable (GAS) and 0 is its unique equilibrium point. Fortunately, it is the case here because $I - \gamma P_\pi$ is a nonsingular M-matrix. It is well-known that a nonsingular M-matrix is positive stable [Plemmons, 1977]. The ODE@∞[1]

[1] Here, ODE@∞ denotes the ODE at infinity technique used in the Borkar-Meyn stability framework [Borkar and Meyn, 2000].

associated with discounted TD with a global clock is instead

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = - D_\mu(I - \gamma P_\pi)x(t),$$

where D_μ is a diagonal matrix with $D_\mu(s,s) = d_\mu(s)$ and d_μ is the stationary distribution of the Markov chain induced by μ. Suppose the Markov chain induced by μ is irreducible, then $d_\mu(s) > 0$ for all s. So D_μ is also positive stable. It's also known that the left multiplication of a nonsingular M-matrix by a positive stable diagonal matrix is still positive stable [Fiedler and Pták, 1962, Plemmons, 1977], which implies that $D_\mu(I - \gamma P_\pi)$ is also positive stable and thus the ODE@∞ associated with discounted TD with a global clock is also GAS. The difference between the two ODEs is intuitive. With a local clock, all the states are updated at the same "magnitude", so the actual sampling distribution does not matter. But with a global clock, the states are updated at different magnitudes according to how frequently they are visited. So there is a D_μ term to capture the sampling distribution.

For differential TD with a local clock, the corresponding ODE@∞ is

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -(I - P_\pi + \eta e e^\top)x(t),$$

where e is the all-one vector. The detailed derivation can be found in Wan et al. [2021]. Although $(I - P_\pi + \eta e e^\top)$ is in general not an M-matrix, it is still positive stable for all $\eta > 0$. So the ODE is still GAS. The ODE@∞ associated with differential TD with a global clock is instead

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -D_\mu(I - P_\pi + \eta e e^\top)x(t).$$

See Blaser et al. [2026] for the detailed derivation. Unfortunately, the left multiplication of a positive stable matrix by a positive stable diagonal matrix does not necessarily preserve the positive stability. So the ODE@∞ associated with differential TD with a global clock is not automatically GAS for all $\eta > 0$. To summarize, the fundamental difference is that **discounted TD has an M-matrix but differential TD has only a positive stable matrix**.

Blaser et al. [2026] approach this challenge via a rank-one perturbation perspective. Let

$$A_\eta \doteq D_\mu(I - P_\pi + \eta e e^\top) = D_\mu(I - P_\pi) + \eta d_\mu e^\top.$$

Then A_η can be viewed as a rank-one perturbation of a singular M-matrix because $D_\mu(I - P_\pi)$ is a singular M-matrix (see Blaser et al. [2026] for a simple proof) and $\eta d_\mu e^\top$ is a rank-one matrix. With results from the rank-one perturbation community [Bierkens and Ran, 2014], Blaser et al. [2026] show that when η is sufficiently small, A_η is positive stable and thus differential TD with a global clock is convergent under the assumptions stated in Appendix C.

4 Counterexamples

In this section, we provide a full characterization about when A_η is positive stable, based on which we construct concrete counterexamples showing that A_η is not positive stable for some $\eta > 0$. The following assumption is standard and is used in all the rest of the paper.

Assumption 4.1. The target transition matrix P_π is irreducible and $d_\mu(s) > 0$ for all s.

We follow the rank-one perturbation perspective in Blaser et al. [2026] and first study

$$L \doteq D_\mu(I - P_\pi).$$

We use $\mathrm{spec}(L) \doteq \{z \in \mathbb{C} : \det(zI - L) = 0\}$ to denote the spectrum of L, i.e., the set of its eigenvalues. Our first observation is that every eigenvalue of L has nonnegative real part and every nonzero eigenvalue has strictly positive real part.

Lemma 4.2. *For any $\lambda \in \mathrm{spec}(L)$, $\mathrm{Re}(\lambda) \geq 0$. If $\lambda \neq 0$, then $\mathrm{Re}(\lambda) > 0$.*

Proof. Let $\lambda \in \mathrm{spec}(L)$ and let $x \neq 0$ satisfy $Lx = \lambda x$. Choose k such that $|x_k| = \max_j |x_j|$. The k-th row of $Lx = \lambda x$ gives $d_\mu(k)\left(x_k - \sum_j P_\pi(k,j)x_j\right) = \lambda x_k$, so $\sum_j P_\pi(k,j)x_j = \left(1 - \frac{\lambda}{d_\mu(k)}\right)x_k$. Since P_π is row-stochastic,

$$\left|1 - \tfrac{\lambda}{d_\mu(k)}\right| |x_k| \leq \textstyle\sum_j P_\pi(k,j)|x_j| \leq |x_k|.$$

Thus $|1-\lambda/d_\mu(k)| \le 1$. Writing $\lambda = a+ib$ yields $\left(1-\frac{a}{d_\mu(k)}\right)^2 + \left(\frac{b}{d_\mu(k)}\right)^2 \le 1$. By rearranging terms, we have $2d_\mu(k)\,\mathrm{Re}(\lambda) \ge |\lambda|^2$, which completes the proof. □

Lemma 4.2 clarifies that *0 is the only eigenvalue of L on the imaginary axis and all the other eigenvalues of L are in the open right half plane*. Now, define a function $g : \mathbb{C} \setminus \mathrm{spec}(L) \to \mathbb{C}$ as

$$g(z) \doteq e^\top (zI - L)^{-1} d_\mu.$$

It is easy to see that $g(z)$ is well-defined for every $z \notin \mathrm{spec}(L)$. Particularly, let i be the imaginary unit. Then for every $\omega > 0$, $i\omega \notin \mathrm{spec}(L)$ thanks to Lemma 4.2. So $g(i\omega)$ is well-defined. This allows us to define

$$\eta_* \doteq \inf\{\eta > 0 \mid \exists \omega > 0 \text{ such that } 1 - \eta g(i\omega) = 0\}.$$

with the convention that $\inf \emptyset = +\infty$. This η_* is the threshold we want.

Theorem 4.3 (Maximal Stability Threshold)**.** *For every $\eta \in (0, \eta_*)$, A_η is positive stable. Furthermore, the above statement does not hold for any $\eta'_* > \eta_*$, i.e., for every $\eta'_* > \eta_*$, there exists some $\eta_c \in [\eta_*, \eta'_*)$ such that A_{η_c} is not positive stable.*

Proof. We view A_η as a rank-one perturbation of L and study how the eigenvalues of A_η move as η increases from 0. Since eigenvalues depend continuously on the entries of the matrix, they also depend continuously on η. By Lemma 4.9 of Blaser et al. [2026], there exists $\eta_0 > 0$ such that A_η is positive stable for every $\eta \in (0, \eta_0]$.[2] In other words, all the eigenvalues of A_η are in the open right half plane for all sufficiently small $\eta > 0$. Thus for A_η to lose positive stability when η increases, at least one eigenvalue must cross the imaginary axis for some $\eta > 0$.

First, we will show that it cannot cross the imaginary axis through the origin. This is because A_η has no zero eigenvalue for any $\eta > 0$. To see this, let d_π be the stationary distribution of P_π and define $\ell^\top \doteq d_\pi^\top D_\mu^{-1}$. If $A_\eta x = 0$ for some $x \neq 0$ and $\eta > 0$, then

$$0 = \ell^\top A_\eta x = \ell^\top D_\mu (I - P_\pi)x + \eta \ell^\top d_\mu e^\top x = \eta e^\top x.$$

Then $0 = A_\eta x = Lx + \eta d_\mu e^\top x = Lx$. So x is in the kernel of L. Since D_μ is full rank, the kernel of L is the same as the kernel of $I - P_\pi$. But by the Perron-Frobenius theorem, the kernel of $I - P_\pi$ is one-dimensional and is spanned by the all-one vector e. So there must exist $c \neq 0$ such that $x = ce$, yielding $0 = e^\top x = ce^\top e \neq 0$, which is a contradiction.

Now we characterize every possible non-origin imaginary axis crossing. Notice that A_η is a real matrix, so nonreal eigenvalues of A_η come in conjugate pairs. It is thus sufficient to characterize every possible crossing at $i\omega$ with $\omega > 0$. We thus study when $i\omega$ is an eigenvalue of A_η for some $\omega > 0$ and $\eta > 0$, i.e., when $\det(i\omega I - A_\eta) = 0$. By the rank-one determinant identity $\det\big(M + uv^\top\big) = \det(M) + v^\top \,\mathrm{adj}(M) u$ with $\mathrm{adj}(M)$ being the adjugate of M, we have

$$0 = \det(i\omega I - A_\eta) = \det\big(i\omega I - L - \eta d_\mu e^\top\big) = \det(i\omega I - L) - \eta e^\top \,\mathrm{adj}(i\omega I - L) d_\mu.$$

Since $i\omega \notin \mathrm{spec}(L)$, we have $\mathrm{adj}(i\omega I - L) = \det(i\omega I - L)(i\omega I - L)^{-1}$, so

$$0 = \det(i\omega I - L) - \eta e^\top \,\mathrm{adj}(i\omega I - L) d_\mu = \det(i\omega I - L)(1 - \eta g(i\omega)).$$

Again since $i\omega \notin \mathrm{spec}(L)$, we have $\det(i\omega I - L) \neq 0$. So the above is equivalent to $1 - \eta g(i\omega) = 0$. By definition, η_* is the infimum of such η, which completes the proof. □

Theorem 4.3 gives the largest η_* such that the claim "A_η is positive stable for every $\eta \in (0, \eta_*)$" holds. In the following, we show that the stability region can be more complicated than just $(0, \eta_*)$.

Theorem 4.4 (Exact Stability Region)**.** *The stability region*

$$\Pi \doteq \{\eta > 0 \mid A_\eta \text{ is positive stable}\}$$

is always a finite union of open intervals, possibly including unbounded intervals.

[2] Lemma 4.9 of Blaser et al. [2026] proves this using Lemma 2.11 of Bierkens and Ran [2014]. For completeness, Appendix A gives a self-contained perturbation proof for our setting.

Proof. To characterize Π, we will use the Routh-Hurwitz criterion [Gantmacher, 1980]. To this end, we consider the characteristic polynomial of $-A_\eta$ defined as

$$q_\eta(z) \doteq \det(zI + A_\eta) = z^m + a_1(\eta)z^{m-1} + \cdots + a_m(\eta),$$

where $m = |\mathcal{S}|$ is the number of states and $\{a_j(\eta)\}_{j=1}^m$ are the coefficients of the characteristic polynomial. The roots of $q_\eta(z) = \det(zI + A_\eta)$ are exactly $z = -\lambda$, where $\lambda \in \operatorname{spec}(A_\eta)$. Therefore A_η is positive stable if and only if every root of q_η has negative real part, i.e., if and only if q_η is a Hurwitz polynomial. By the Routh-Hurwitz criterion, a real monic polynomial is Hurwitz if and only if all of its Hurwitz determinants are positive. Using the convention that $a_0(\eta) \doteq 1$ and $a_j(\eta) \doteq 0$ whenever $j < 0$ or $j > m$, for $k = 1, \ldots, m$, the k-th Hurwitz determinant of the polynomial q_η is defined as

$$\Delta_k(\eta) \doteq \det\left([a_{2j-i}(\eta)]_{i,j=1}^k\right).$$

Thus A_η is positive stable if and only if $\Delta_k(\eta) > 0$ for every $k = 1, \ldots, m$. Since $A_\eta = L + \eta d_\mu e^\top$, we have

$$q_\eta(z) = \det\big(zI + L + \eta d_\mu e^\top\big) = \det(zI + L) + \eta e^\top \operatorname{adj}(zI + L) d_\mu.$$

Hence $q_\eta(z) = q_0(z) + \eta q_1(z)$, where q_0 and q_1 are polynomials in z. Consequently, each coefficient $a_j(\eta)$ must be linear in η. Since $\Delta_k(\eta)$ is the determinant of a $k \times k$ matrix whose entries are linear in η, it is a polynomial in η of degree at most k. If some $\Delta_k(\eta)$ is zero for all $\eta > 0$, then no η can satisfy $\Delta_k(\eta) > 0$, so the stability region is empty. Otherwise, the finitely many real roots of the nonzero polynomials $\Delta_1, \ldots, \Delta_m$ partition $(0, \infty)$ into finitely many intervals, and the sign of each Δ_k is constant on each such interval. Selecting exactly those intervals on which all determinants are positive gives Π. Thus the stability region is a finite union of open intervals, possibly empty and possibly including unbounded intervals. □

We note that Theorems 4.3 & 4.4 do not rule out the possibility that Π is $(0, \infty)$. Rather, they provide a characterization of what Π may look like. Building on these insights, we will now construct a concrete counterexample showing that for some $\eta > 0$, A_η is not positive stable. The construction combines a cyclic target-policy structure through three groups of states with unequal behavior-state visitation frequencies. Its symmetry isolates a three-dimensional subspace in which the rank-one term $\eta d_\mu e^\top$ can destabilize the dynamics, allowing an exact calculation of where stability is lost and recovered.

Example 4.5 (Counterexample)**.** *Suppose $m > 22$ is an integer. Then there exists a positive diagonal matrix $D_\mu \in \mathbb{R}^{(m+2)\times(m+2)}$ and an irreducible and aperiodic row-stochastic matrix P_π such that A_η is positive stable if and only if*

$$\eta \in (0, \varepsilon) \cup (3\varepsilon, \infty), \qquad \varepsilon \doteq \frac{m - 22}{m^2 + m - 2}. \tag{3}$$

The smallest instance in this family is obtained at $m = 23$, yielding $\varepsilon = 1/550$ and a 25-state system.

Proof. Consider $m + 2$ states $a_1, \ldots, a_m, b, c$. Set $\varepsilon \doteq (m - 22)/(m^2 + m - 2)$. Construct a vector $d_\mu \in \mathbb{R}^{m+2}$ as $d_\mu[a_i] = \varepsilon$ for $i = 1, \ldots, m$, $d_\mu[b] = \varepsilon$, and $d_\mu[c] = 1 - (m + 1)\varepsilon$, and let $D_\mu \doteq \operatorname{diag}(d_\mu)$. Since $m > 22$, we have $\varepsilon > 0$. Moreover,

$$d_\mu[c] = \frac{22m + 20}{m^2 + m - 2} > 0, \qquad d_\mu[c] - \varepsilon = \frac{21m + 42}{m^2 + m - 2} > 0.$$

Also, $\sum_s d_\mu[s] = m\varepsilon + \varepsilon + d_\mu[c] = 1$. Thus d_μ is a strictly positive probability vector, and $\varepsilon \le d_\mu[s]$ for every state s.

Define the row-stochastic matrix Q by

$$Q[a_i, b] = 1 \ (i = 1, \ldots, m), \qquad Q[b, c] = 1, \qquad Q[c, a_i] = \frac{1}{m} \ (i = 1, \ldots, m),$$

with all other entries zero. Define $P_\pi \doteq I + \varepsilon D_\mu^{-1}(Q - I)$. Then, for each state s,

$$P_\pi[s, \cdot] = \left(1 - \tfrac{\varepsilon}{d_\mu[s]}\right) e_s^\top + \tfrac{\varepsilon}{d_\mu[s]} Q[s, \cdot],$$

where e_s is the standard basis vector at state s, not the all-one vector e. Therefore every row of P_π is a convex combination of $e_s^\top$ and $Q[s,\cdot]$, so P_π is row-stochastic. It is irreducible because all transitions of Q remain available, and it is aperiodic because $P_\pi[c,c] = 1 - \varepsilon/d_\mu[c] > 0$.

For $\eta > 0$, write $t \doteq \eta/\varepsilon$. Then

$$A_\eta = D_\mu(I - P_\pi + \eta e e^\top) = \varepsilon(I - Q + t d_\mu e^\top),$$

since $D_\mu(I - P_\pi) = \varepsilon(I - Q)$ and $D_\mu \eta e e^\top = \eta d_\mu e^\top = \varepsilon t d_\mu e^\top$. Since $\varepsilon > 0$, multiplication by ε does not change positive stability. So it is sufficient to analyze $B_t \doteq I - Q + t d_\mu e^\top$.

We next decompose the ambient vector space. Work over $\mathbb{C}$ and order the coordinates as $(a_1, \ldots, a_m, b, c)$. Let

$$\mathcal{H} \doteq \{x \in \mathbb{C}^{m+2} : x[b] = x[c] = 0,\ \sum_{i=1}^m x[a_i] = 0\}.$$

If $x \in \mathcal{H}$, then $Qx = 0$ and $e^\top x = 0$, so $B_t x = x$. Since $\dim(\mathcal{H}) = m - 1$, the restriction $B_t|_{\mathcal{H}}$ is the identity map on $\mathcal{H}$, so this block contributes $m - 1$ eigenvalues equal to 1. Now consider the symmetric subspace

$$\mathcal{U} \doteq \{x \in \mathbb{C}^{m+2} : x[a_1] = \cdots = x[a_m]\}.$$

Define the linear isomorphism

$$\Phi : \mathbb{C}^3 \to \mathcal{U}, \qquad \Phi(u, v, w) = (u, \ldots, u, v, w).$$

Let $y = (u, v, w)^\top$. Then $Q\Phi y = \Phi(Cy), \qquad e^\top \Phi y = r^\top y$, where

$$C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad r^\top = \begin{pmatrix} m & 1 & 1 \end{pmatrix}.$$

Moreover,

$$d_\mu e^\top \Phi y = \Phi(\bar{d} r^\top y), \qquad \bar{d} = \begin{pmatrix} \varepsilon \\ \varepsilon \\ 1 - (m+1)\varepsilon \end{pmatrix}.$$

Hence $(Q - t d_\mu e^\top)\Phi y = \Phi\big((C - t\bar{d} r^\top) y\big)$. In particular, $\mathcal{U}$ is invariant under $Q - t d_\mu e^\top$, and therefore also under $B_t = I - Q + t d_\mu e^\top$. The matrix of the restricted map $(Q - t d_\mu e^\top)|_{\mathcal{U}}$ in the coordinates given by Φ is $M_t \doteq C - t\bar{d} r^\top$. Consequently, the restriction of B_t to $\mathcal{U}$ is represented by $I - M_t$. Finally, $\mathbb{C}^{m+2} = \mathcal{H} \oplus \mathcal{U}$. Indeed, for any $x \in \mathbb{C}^{m+2}$, let $\bar{u} \doteq m^{-1} \sum_{i=1}^m x[a_i]$. Then

$$x = \underbrace{(x[a_1] - \bar{u}, \ldots, x[a_m] - \bar{u}, 0, 0)}_{\in \mathcal{H}} + \underbrace{(\bar{u}, \ldots, \bar{u}, x[b], x[c])}_{\in \mathcal{U}},$$

and $\mathcal{H} \cap \mathcal{U} = \{0\}$. Since both subspaces are B_t-invariant, the matrix of B_t in a basis adapted to this direct sum is block diagonal:

$$B_t \sim \begin{pmatrix} I_{m-1} & 0 \\ 0 & I - M_t \end{pmatrix}.$$

Thus the remaining three eigenvalues of B_t are exactly the eigenvalues of $I - M_t$. Now compute the characteristic polynomial of M_t. Let $N = zI - C$. Then $\det N = z^3 - 1$ and

$$\operatorname{adj}(N) = \begin{pmatrix} z^2 & z & 1 \\ 1 & z^2 & z \\ z & 1 & z^2 \end{pmatrix}.$$

By the matrix determinant lemma,

$$\det(zI - M_t) = \det\big(N + t\bar{d} r^\top\big) = \det N + t r^\top \operatorname{adj}(N) \bar{d}.$$

$$r^\top \operatorname{adj}(N) \bar{d} = z^2 + z + \big[m - \varepsilon(m^2 + m - 2)\big] = z^2 + z + 22.$$

Therefore $\det(zI - M_t) = z^3 - 1 + t(z^2 + z + 22)$. The Hurwitz polynomial of $I - M_t$ is

$$q_t(s) = \det(sI + I - M_t) = s^3 + (t+3)s^2 + (3t+3)s + 24t.$$

For a cubic $s^3 + as^2 + bs + c$, Hurwitz stability is equivalent to $a > 0$, $b > 0$, $c > 0$, and $ab > c$. Here $a, b, c > 0$ for $t > 0$, and the remaining condition is

$$(t+3)(3t+3) > 24t \quad \Longleftrightarrow \quad 3(t-1)(t-3) > 0.$$

Thus the three-dimensional block is positive stable if and only if $t \in (0,1) \cup (3,\infty)$. Since the other $m-1$ eigenvalues of B_t are equal to 1, we conclude that

$$A_\eta \text{ is positive stable} \quad \Longleftrightarrow \quad \eta \in (0,\varepsilon) \cup (3\varepsilon,\infty).$$

□

Lemma 4.9 of Blaser et al. [2026] guarantees positive stability for sufficiently small η. Theorem 4.3 characterizes the maximal initial interval on which positive stability holds. In Example 4.5, its endpoint is $\eta_* = \varepsilon$, while the full stability region is $(0,\varepsilon) \cup (3\varepsilon,\infty)$. Thus, stability can return after it is lost, and η_* is not a threshold beyond which every η is unstable. Lemma 4.9 of Blaser et al. [2026] does not specify a numerical value of η_0 (only the existence), so we cannot further make more detailed statements about the numerical gap between η_0 and η_*.

We believe the spirit of this counterexample extends to other settings. Its three types of states, a_i, b, and c, have a simple structure. Under the target policy, the process usually stays at c, with rare excursions $c \to a_i \to b \to c$. Here c can be thought of as a normal state of the system, a_i can be thought of as the i-th possible rare failure of the system and b can be thought of as a common recovery for every failure. Under the behavior policy, c is visited far more often than each a_i or b.

5 Experiments

We first provide a numerical validation of Example 4.5. Figure 1 shows the nontrivial eigenvalues of A_η/ε as η increases from 0 to 4ε. The conjugate pair starts in the open right half plane, crosses the imaginary axis at $\eta = \varepsilon$, remains in the open left half plane for $\eta \in (\varepsilon, 3\varepsilon)$, and crosses back into the open right half plane at $\eta = 3\varepsilon$.

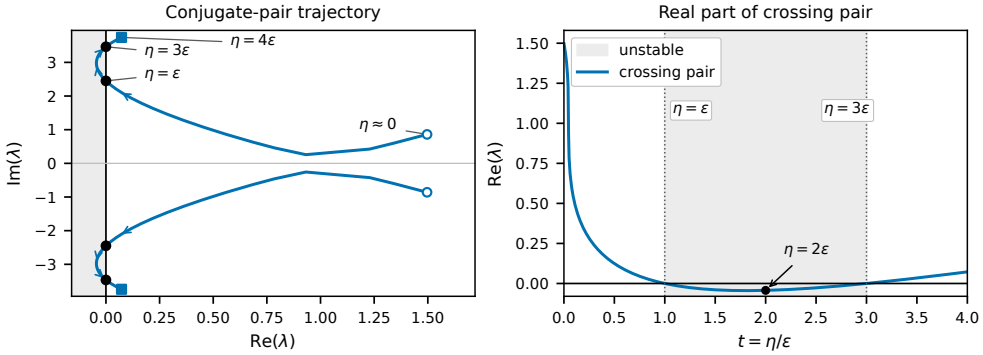

Figure 1: Phase portrait of the nontrivial eigenvalues of A_η/ε for the counterexample in Example 4.5. Left: trajectory of the conjugate pair in the complex plane. Right: real part of this pair as a function of $t = \eta/\varepsilon$, with the unstable interval $t \in (1,3)$ shaded.

We then construct a concrete MDP instance based on Example 4.5 and run the global-clock and local-clock differential TD algorithms on it with the same η. We consider a finite MDP with $m+2$ states and two actions $\{a_0, a_1\}$, where $m > 22$ is an integer. The target policy π always selects action a_0, and the transition matrix of a_0 is defined to be P_π in the proof of Example 4.5. So the transition matrix of the target policy π coincides with P_π. The behavior policy μ selects action a_0 with probability $\kappa \in (0,1)$ and action a_1 otherwise, where κ is chosen so that $\kappa \le \min_{i,j:P_\pi[i,j]>0} \frac{d_\mu[j]}{P_\pi[i,j]}$, where both P_π and d_μ are defined in the proof of Example 4.5. Define the transition matrix of a_1 as $R \doteq (ed_\mu^\top - \kappa P_\pi)/(1-\kappa)$. The matrix R is a row stochastic matrix by the choice of κ and the fact that $ed_\mu^\top$ and P_π are row stochastic. The transition matrix of the behavior policy is then

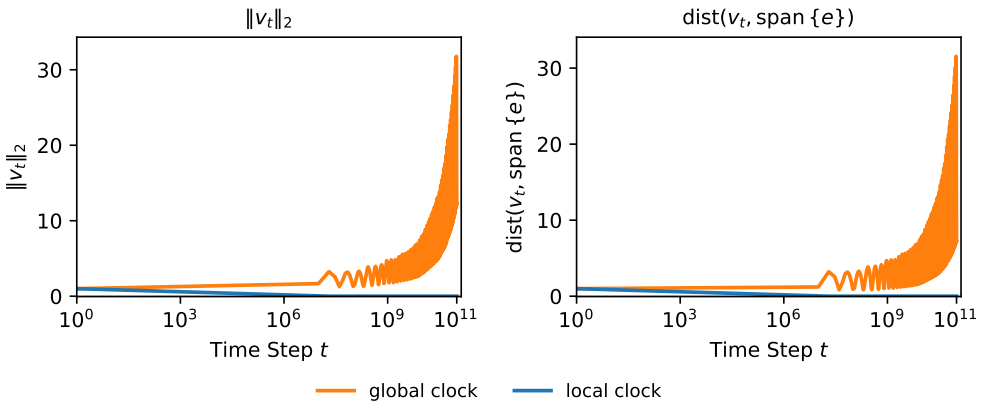

Figure 2: Differential TD with $\eta = 2\varepsilon$. Left: $\|v_t\|_2$. Right: distance from v_t to $\text{span}\{e\}$. The global-clock trajectory is oscillatory because the unstable mode is the conjugate pair shown in Figure 1. The curve here is only a typical run with one seed, because we are interested in the sample path behavior. We provide 9 more seeds in Appendix B, which show the same phenomenon.

$P_\mu = \kappa P_\pi + (1-\kappa)R = ed_\mu^\top$. So d_μ coincides with the stationary distribution of the behavior policy. The reward is always 0. For the reported run, the initial state is fixed to be a_1. We use the smallest instance in this family, $m = 23$, and choose $\eta = 2\varepsilon = \frac{1}{275}$, which lies in the unstable interval $(\varepsilon, 3\varepsilon)$ for the global clock. We then run the algorithm (2) with both the global clock and the local clock in the learning rates, using the base sequence $\alpha_n = 0.45/(10^4 + n)^{0.6}$. The global-clock run uses $n = t+1$, while the local-clock run uses $n = \nu(S_t, t)$. We choose v_0 as the real part of an eigenvector corresponding to the eigenvalue of A_η with the smallest real part, normalized so that $\|v_0\|_2 = 1$, and set $\hat{J}_0 = \eta e^\top v_0$. With this particular choice of $\hat{J}_0$, the results in Wan et al. [2021] show that differential TD with a local clock should converge to 0 for any $\eta > 0$. We run the algorithm for 10^{11} time steps and plot both $\|v_t\|_2$ and $\text{dist}(v_t, \text{span}\{e\})$. Since the reward is always 0, the latter is the distance from v_t to $\{\bar{v}_\pi + ce : c \in \mathbb{R}\}$. The result in Figure 2 shows that the local-clock recursion contracts to 0 while the global-clock recursion grows, matching the exact stability calculation above. Additional experiments in Appendix B.1 vary η and show consistent algorithmic behavior with the stability region in (3).

6 Related Work

The closest to our work is Chen [2025], which also shows that an average-reward algorithm converges with a local clock but can diverge with a global clock. However, there are several differences between this work and Chen [2025]. First, Chen [2025] considers only the base learning-rate sequence $\alpha_n = \mathcal{O}(n^{-1})$. (This rate describes the base sequence. With a local clock, the index is the corresponding state-action visit count.) By using a local clock in this form of learning rate, Chen [2025] implicitly approximates the state-action visitation distribution and thus implicitly applies importance sampling. Their insights and analyses do not apply to learning rates that do not have this form, such as $\alpha_n = \mathcal{O}(n^{-\beta})$ for some $\beta \in (1/2, 1)$. By contrast, our analysis applies to any learning rate that satisfies the standard Robbins-Monro conditions, including those with $\beta \in (1/2, 1)$. Second, Chen [2025] focuses on running standard discounted Q-learning with $\gamma = 1$ in the average-reward setting directly. With the local clock implicitly applying importance sampling, their algorithm converges to the *set* of optimal action-value functions, i.e., the q value estimates themselves are not guaranteed to converge to a unique limit. So Chen [2025] has to introduce a new centered version for convergence to a unique limit. By contrast, the differential TD algorithm we analyze is designed for the average-reward setting and implicitly applies the reference state technique from RVI Q-learning, which directly guarantees convergence to a unique limit under the local-clock convergence assumptions in Appendix C. Nevertheless, one may ask, can the insights and techniques of Chen [2025] be applied towards understanding the open problems in Wan et al. [2021], Blaser et al. [2026]? Our evaluation is negative. At most, Chen [2025] can provide an alternative convergence proof for differential TD with a local clock and base learning-rate sequence $\alpha_n = \mathcal{O}(n^{-1})$. But

combining the results from Blaser et al. [2026] and this work, it is easy to see that the effect of the global clock interacts with the hyperparameter η. But Chen [2025] does not have the mechanism to make an evaluation based on η.

This work studies the divergence of differential TD through the lens of the instability of the associated ODE. It is important to note that there are works showing that even if the ODE is not GAS, an algorithm can still converge in a weaker sense, e.g., convergence to a set of solutions [Xie et al., 2025, Wang and Zhang, 2026], convergence to a bounded set [Meyn, 2024, Liu et al., 2025c,b], or convergence to a sample path dependent limit [Blaser and Zhang, 2026]. However, as demonstrated by our experiments, the instability of the ODE we encounter does not appear benign. That being said, proving that differential TD with a global clock can diverge almost surely is still an open problem.

This work also sheds light on the deadly triad problem [Sutton and Barto, 2018, Zhang, 2022] in RL, which arises from the interaction of function approximation, off-policy learning, and bootstrapping. There have been numerous studies for breaking the deadly triad in the discounted setting [Sutton et al., 2008, 2009, 2016, Zhang and Whiteson, 2022, Zhang et al., 2021b, 2020, Qian and Zhang, 2025]. Some of these approaches extend to the average-reward setting, e.g., Zhang et al. [2021a] extend Sutton et al. [2008, 2009] to the average-reward setting. But since a global clock is commonly used with function approximation, future works on breaking the deadly triad in the average-reward setting should carefully investigate the interaction between the global clock and the average-reward estimation rate hyperparameter η. Similarly, policy gradient and actor critic methods [Sutton et al., 1999, Konda and Tsitsiklis, 1999] are well studied in the discounted setting with a global clock [Agarwal et al., 2020, Mei et al., 2020, Zhang et al., 2022]. But future works extending the analysis to the average-reward setting should also carefully consider the interaction between the global clock and the average-reward estimation rate hyperparameter η.

7 Conclusion

This work shows that the choice of local and global clocks can affect the stability of average-reward RL algorithms in an arguably surprising way: for some $\eta > 0$, the ODE associated with differential TD can be stable with a local clock but unstable with a global clock, a phenomenon that has never been reported in the discounted setting. Our result shows that, in average-reward RL, replacing local clocks by the practically standard global clock is not a benign implementation detail. A theorem proved with local clocks may be misleading for the algorithm people actually implement. A possible direction for future work is to extend the analysis to understand the behavior of differential Q-learning with a global clock and its dependence on the hyperparameter η. We imagine that such an analysis would be very challenging given the nonlinear nature of the differential Q-learning updates. However, we argue that our analysis for differential TD can serve as a foundation because the ODE associated with differential Q-learning can be regarded as a switched system (cf. Lee and He [2020]), where each subsystem's stability can be analyzed similarly to the differential TD case.

This work also makes a contribution to the rank-one perturbation theory [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]. Particularly, Bierkens and Ran [2014] study the positive stability of rank-one perturbation to a singular M-matrix and establish a set of sufficient conditions. One important result is regarding the positive stability of $B_\eta \doteq L + \eta u v^\top$, where L is a singular M-matrix with 0 being an algebraically simple eigenvalue and u and v are elementwise nonnegative. Lemma 2.11 of Bierkens and Ran [2014] shows that B_η is positive stable for sufficiently small $\eta > 0$. The positive stability of B_η for sufficiently large η is left as Conjecture 2.17 in Bierkens and Ran [2014]. Later on, Anehila and Ran [2022] show by a counterexample that Conjecture 2.17 of Bierkens and Ran [2014] is false. Our counterexample further shows that even if we impose additional conditions such as u being strictly positive and v being the all-one vector, the positive stability of B_η does not necessarily hold for all $\eta > 0$. Those counterexamples together testify to the hardness of characterizing the positive stability of rank-one perturbation to a singular M-matrix. It remains an open problem whether there is some nontrivial sufficient condition that can guarantee the positive stability of B_η for sufficiently large η. Notably, as noted by Bierkens and Ran [2014], Anehila and Ran [2022], such counterexamples are hard to construct and random sampling is unlikely to find them. Our experience is the same. We have tried to find counterexamples by random sampling and failed. And we have to construct the counterexample manually based on the theoretical insights in Theorems 4.3 & 4.4.

Acknowledgments and Disclosure of Funding

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 awards VV-1Q26-001, VV-1Q27-007, and VV-1Q27005, 4VA at UVA, a Cisco Faculty Research Award, and a Google Research Award.

References

Jinane Abounadi, Dimitri Bertsekas, and Vivek S Borkar. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 2001.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Proceedings of the Conference on Learning Theory*, 2020.

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

Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, 1957.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific Belmont, MA, 1996.

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

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

Ethan Blaser, Jiuqi Wang, and Shangtong Zhang. Almost sure convergence of differential temporal difference learning for average reward markov decision processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2026.

Vivek Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *The Annals of Applied Probability*, 2025.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009.

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

Zaiwei Chen. Non-asymptotic guarantees for average-reward Q-learning with adaptive stepsizes. In *Advances in Neural Information Processing Systems*, 2025.

Zaiwei Chen, Siva T Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A lyapunov theory for finite-sample guarantees of markovian stochastic approximation. *Operations Research*, 2024.

Jiu Ding and Aihui Zhou. Eigenvalues of rank-one updated matrices with some applications. *Applied Mathematics Letters*, 2007.

Miroslav Fiedler and Vlastimil Pták. On matrices with non-positive off-diagonal elements and positive principal minors. *Czechoslovak Mathematical Journal*, 1962.

JH Fourie, GJ Groenewald, DB Janse van Rensburg, and ACM Ran. Rank one perturbations of h-positive real matrices. *Linear Algebra and its Applications*, 2013.

F.R. Gantmacher. *The Theory of Matrices*. Chelsea Publishing Company, 1980.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, 1993.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, 1999.

Donghwan Lee and Niao He. A unified switching system perspective and convergence analysis of Q-Learning algorithms. In *Advances in Neural Information Processing Systems*, 2020.

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

Xinyu Liu, Zixuan Xie, and Shangtong Zhang. Extensions of robbins-siegmund theorem with applications in reinforcement learning. *arXiv Preprint*, 2025b.

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

Christian Mehl, Volker Mehrmann, André CM Ran, and Leiba Rodman. Eigenvalue perturbation theory of classes of structured matrices under generic structured rank one perturbations. *Linear Algebra and Its Applications*, 2011.

Christian Mehl, Volker Mehrmann, André CM Ran, and Leiba Rodman. Eigenvalue perturbation theory of symplectic, orthogonal, and unitary matrices under generic structured rank one perturbations. *BIT Numerical Mathematics*, 2014.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the International Conference on Machine Learning*, 2020.

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

Julio Moro and Froilán M Dopico. Low rank perturbation of jordan structure. *SIAM Journal on Matrix Analysis and Applications*, 2003.

Robert J. Plemmons. M-matrix characterizations.i—nonsingular m-matrices. *Linear Algebra and its Applications*, 1977.

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

André CM Ran and Michał Wojtylak. Eigenvalues of rank one perturbations of unstructured matrices. *Linear Algebra and its Applications*, 2012.

André CM Ran and Michał Wojtylak. Global properties of eigenvalues of parametric rank one perturbations for unstructured and structured matrices. *Complex Analysis and Operator Theory*, 2021.

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

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 1999.

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

Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the International Conference on Machine Learning*, 2009.

Richard S. Sutton, Ashique Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 2016.

John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 1994.

Yi Wan, Abhishek Naik, and Richard S. Sutton. Learning and planning in average-reward markov decision processes. In *Proceedings of the International Conference on Machine Learning*, 2021.

Jiuqi Wang and Shangtong Zhang. Almost sure convergence of linear temporal difference learning with arbitrary features. *Journal of Machine Learning Research*, 2026.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine Learning*, 1992.

Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, 1989.

Zixuan Xie, Xinyu Liu, Rohan Chandra, and Shangtong Zhang. Finite sample analysis of linear temporal difference learning with arbitrary features. In *Advances in Neural Information Processing Systems*, 2025.

Shangtong Zhang. *Breaking the deadly triad in reinforcement learning*. PhD thesis, University of Oxford, 2022.

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

Shangtong Zhang, Bo Liu, and Shimon Whiteson. GradientDICE: Rethinking generalized offline estimation of stationary values. In *Proceedings of the International Conference on Machine Learning*, 2020.

Shangtong Zhang, Yi Wan, Richard S. Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2021a.

Shangtong Zhang, Hengshuai Yao, and Shimon Whiteson. Breaking the deadly triad with a target network. In *Proceedings of the International Conference on Machine Learning*, 2021b.

Shangtong Zhang, Remi Tachet des Combes, and Romain Laroche. Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *Journal of Machine Learning Research*, 2022.

A Small-η positive stability

Lemma A.1 (Small-η positive stability)**.** *Suppose that P_π is irreducible and $d_\mu(s) > 0$ for every state s. Then there exists $\eta_0 > 0$ such that*

$$A_\eta = D_\mu(I - P_\pi + \eta e e^\top)$$

is positive stable for every $\eta \in (0, \eta_0]$.

Proof. Recall that $L \doteq D_\mu(I - P_\pi)$, so $A_\eta = L + \eta d_\mu e^\top$. Lemma 4.2 shows that every nonzero eigenvalue of L has strictly positive real part. It remains to understand the zero eigenvalue of L. Since D_μ is invertible, $\ker(L) = \ker(I - P_\pi)$. Because P_π is irreducible and row-stochastic, $\ker(I - P_\pi) = \operatorname{span}\{e\}$ by the Perron-Frobenius theorem.

Let d_π be the stationary distribution of P_π and define

$$\ell^\top \doteq d_\pi^\top D_\mu^{-1}.$$

Then

$$\ell^\top L = d_\pi^\top (I - P_\pi) = 0, \qquad \ell^\top e = d_\pi^\top D_\mu^{-1} e > 0.$$

The zero eigenvalue of L is algebraically simple. If we suppose otherwise, we know that since its eigenspace is one-dimensional (because $\ker(L) = \operatorname{span}\{e\}$ is one-dimensional), there would exist a generalized eigenvector y such that $Ly = e$. Multiplying by $\ell^\top$ yields

$$0 = \ell^\top L y = \ell^\top e > 0,$$

which is a contradiction.

We now compute how this simple zero eigenvalue moves under the perturbation. Define

$$F(x, \lambda, \eta) \doteq \begin{pmatrix} A_\eta x - \lambda x \\ \ell^\top x - \ell^\top e \end{pmatrix}.$$

At $(x, \lambda, \eta) = (e, 0, 0)$, we have $F(e, 0, 0) = 0$. The derivative of F with respect to (x, λ) at this point maps (y, θ) to

$$\begin{pmatrix} Ly - \theta e \\ \ell^\top y \end{pmatrix}.$$

This map is injective. If $Ly - \theta e = 0$ and $\ell^\top y = 0$, then multiplying the first equation by $\ell^\top$ gives $-\theta \ell^\top e = 0$, so $\theta = 0$. Then $Ly = 0$, hence $y = ce$, and $\ell^\top y = 0$ gives $c = 0$. Since the domain and codomain have the same finite dimension, the map is invertible.

By the finite-dimensional implicit function theorem, applied after identifying complex vectors with real vectors, there are differentiable functions $x(\eta)$ and $\lambda(\eta)$ near 0 such that

$$A_\eta x(\eta) = \lambda(\eta) x(\eta), \qquad x(0) = e, \qquad \lambda(0) = 0.$$

Differentiating at $\eta = 0$ yields

$$d_\mu e^\top e + L x'(0) = \lambda'(0) e.$$

Multiplying by $\ell^\top$ gives

$$\lambda'(0) = \frac{\ell^\top d_\mu \, e^\top e}{\ell^\top e} = \frac{|\mathcal{S}|}{d_\pi^\top D_\mu^{-1} e} > 0,$$

because $\ell^\top d_\mu = d_\pi^\top D_\mu^{-1} d_\mu = d_\pi^\top e = 1$ and $e^\top e = |\mathcal{S}|$. Therefore the zero eigenvalue of L moves into the open right half plane for all sufficiently small positive η.

Here, all other eigenvalues of L already have strictly positive real part by Lemma 4.2, and they remain in the open right half plane for sufficiently small $\eta > 0$ by continuity of the roots of the characteristic polynomial. Hence there exists $\eta_0 > 0$ such that A_η is positive stable for every $\eta \in (0, \eta_0]$. □

B Additional Online TD Runs

Figures 3 and 4 repeat the differential TD experiment from Figure 2 for nine additional random seeds.

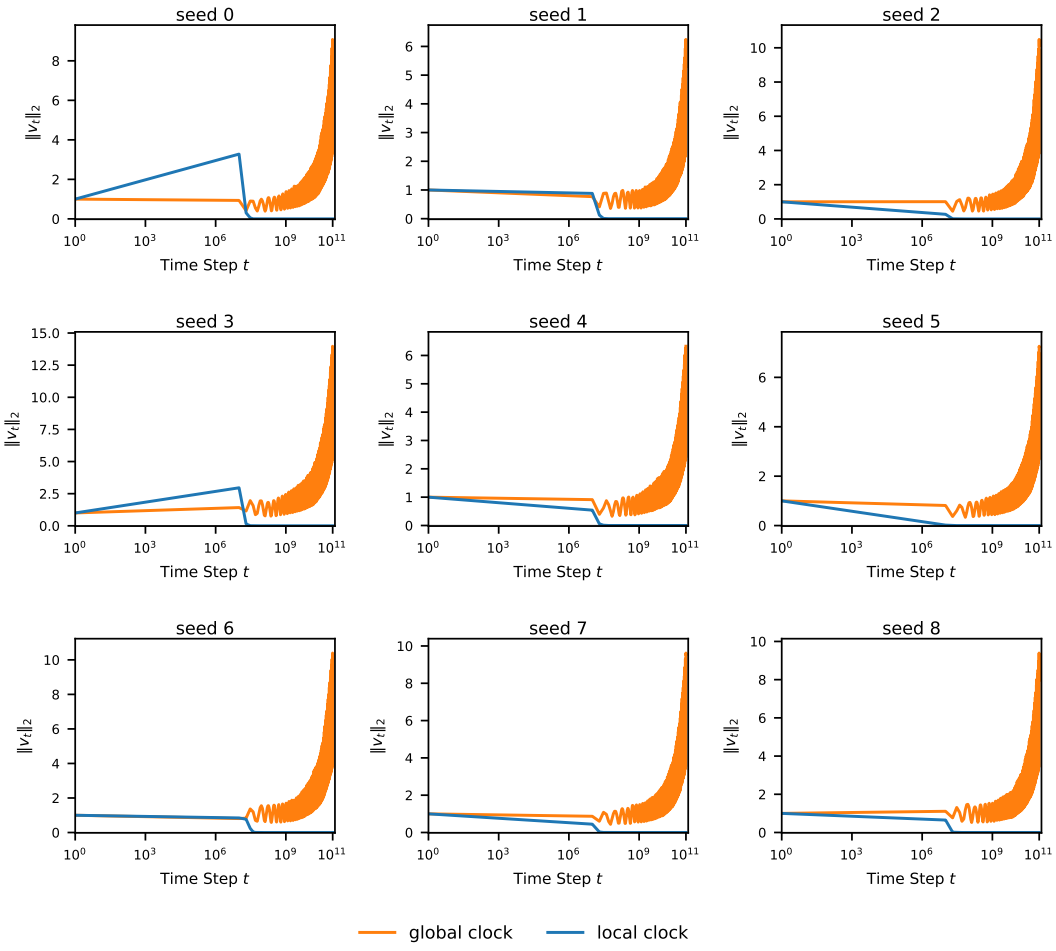

Figure 3: Additional differential TD runs with $\eta = 2\varepsilon$, showing $\|v_t\|_2$.

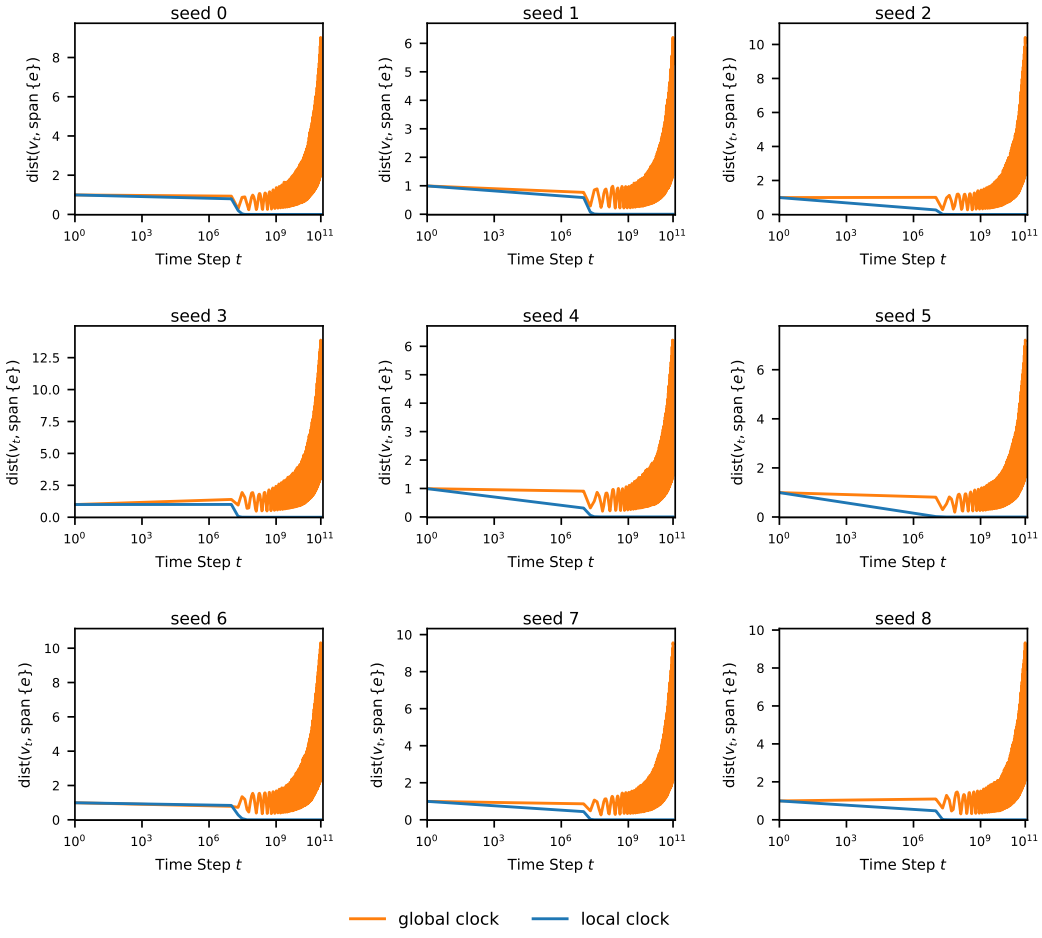

Figure 4: Additional differential TD runs with $\eta = 2\varepsilon$, showing the distance from v_t to $\mathrm{span}\{e\}$.

B.1 Threshold sweep

We also run a threshold sweep for Example 4.5 on both sides of $\eta = \varepsilon$ and $\eta = 3\varepsilon$, with $m = 23$ and $\varepsilon = 1/550$. We use $\kappa = \varepsilon$, the initial state a_1, the base sequence $\alpha_n = 0.45/(10^4 + n)^{0.6}$, and $T = 10^9$ time steps. For each η, we choose v_0 as the real part of an eigenvector corresponding to the eigenvalue of A_η with the smallest real part, normalized so that $\|v_0\|_2 = 1$, and set $\hat{J}_0 = \eta e^\top v_0$. All four settings use the same behavior trajectory (seed 9). Since the critical mode is a complex conjugate pair and the norm oscillates, Table 1 reports the change over the second half of the run. The endpoint changes are consistent with the stable–unstable–stable regions. At the exact boundaries, the critical eigenvalues of the associated ODE are purely imaginary, so exponential divergence is not predicted.

Table 1: Threshold sweep for global-clock differential TD. Here, v_t^{G} denotes the global-clock iterate and $\mathrm{dist}_t^{\mathrm{G}} = \mathrm{dist}(v_t^{\mathrm{G}}, \mathrm{span}\{e\})$ is its Euclidean distance to the differential-value solution set.

η/ε	ODE region	$\max_{\lambda \in \mathrm{spec}(-A_\eta)} \mathrm{Re}(\lambda)$	$\|v_{T/2}^{\mathrm{G}}\|_2 \to \|v_T^{\mathrm{G}}\|_2$	$\mathrm{dist}_{T/2}^{\mathrm{G}} \to \mathrm{dist}_T^{\mathrm{G}}$
0.85	stable	-4.40×10^{-5}	$2.500 \to 1.575$	$2.191 \to 1.482$
1.15	unstable	$+3.16 \times 10^{-5}$	$2.647 \to 3.645$	$2.607 \to 3.369$
2.70	unstable	$+3.16 \times 10^{-5}$	$3.375 \to 3.803$	$3.318 \to 3.790$
3.30	stable	-3.61×10^{-5}	$2.035 \to 0.980$	$1.981 \to 0.693$

C Convergence assumptions

This appendix states the conditions used when invoking stochastic convergence. The matrix results in Section 4 use Assumption 4.1 and do not require a learning-rate sequence.

Local-clock differential TD. Theorem B.3 of Wan et al. [2021] invokes Assumptions B.3, B.4, and B.6–B.8:

1. The Markov chain induced by the target policy is unichain. The irreducibility assumption used here is sufficient.
2. Coverage holds: $\pi(a \mid s) > 0$ implies $\mu(a \mid s) > 0$ for every state-action pair.
3. The deterministic base learning-rate sequence satisfies
$$\alpha_n > 0, \qquad \sum_{n=0}^{\infty} \alpha_n = \infty, \qquad \sum_{n=0}^{\infty} \alpha_n^2 < \infty.$$
4. For every $x \in (0, 1)$,
$$\sup_n \frac{\alpha_{\lfloor xn \rfloor}}{\alpha_n} < \infty, \qquad \frac{\sum_{j=0}^{\lfloor yn \rfloor} \alpha_j}{\sum_{j=0}^{n} \alpha_j} \longrightarrow 1 \quad \text{uniformly for } y \in [x, 1].$$
5. There exists $\Delta > 0$ such that every state satisfies
$$\liminf_{n\to\infty} \frac{\nu(s, n)}{n + 1} \geq \Delta \quad \text{almost surely.}$$
Furthermore, for every $u > 0$, define
$$N(n, u) = \min \left\{ m \geq n : \sum_{i=n+1}^{m} \alpha_i \geq u \right\}.$$
For every pair of states s, s', the limit
$$\lim_{n\to\infty} \frac{\sum_{i=\nu(s,n)}^{\nu(s,N(n,u))} \alpha_i}{\sum_{i=\nu(s',n)}^{\nu(s',N(n,u))} \alpha_i}$$
exists almost surely.

Under these conditions, local-clock differential TD converges almost surely for every $\eta > 0$.

Global-clock differential TD. We adopt the following sufficient conditions for stochastic convergence:

1. The finite target and behavior chains are irreducible and aperiodic.
2. Coverage holds: $\pi(a \mid s) > 0$ implies $\mu(a \mid s) > 0$ for every state-action pair.
3. The deterministic base learning-rate sequence is positive and nonincreasing, with
$$\alpha_n \to 0, \qquad \sum_{n=0}^{\infty} \alpha_n = \infty, \qquad \sum_{n=0}^{\infty} \alpha_n^2 < \infty, \qquad \frac{\alpha_n - \alpha_{n+1}}{\alpha_n} = O(\alpha_n).$$
4. A_η is positive stable. Lemma 4.9 of Blaser et al. [2026] supplies this condition for $\eta \in (0, \eta_0]$ for some $\eta_0 > 0$.

These are sufficient conditions adopted here; the square-summability condition is additional to the learning-rate conditions printed in Assumption 4.2 of Blaser et al. [2026]. For the finite MDP considered here, ergodic behavior-policy sampling and coverage give a finite ergodic transition-sample chain and bounded affine update coefficients. These satisfy Assumption 6′ of Liu et al. [2025a]. Together with the stated learning-rate conditions and positive stability, Corollary 8 of Liu et al. [2025a] gives almost-sure convergence to the unique equilibrium of the reduced affine mean ODE.

Discounted TD. The corresponding chain, coverage, and learning-rate conditions above also suffice for discounted tabular TD with the respective clock. The associated matrices $I - \gamma P_\pi$ and $D_\mu(I - \gamma P_\pi)$ are positive stable for $\gamma \in [0, 1)$, as discussed in Section 3.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as a counterexample for differential TD in the average-reward setting. The formal claims are then proved in the theoretical sections and illustrated in the experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion states that the result is a constructed counterexample for differential TD rather than a broad empirical characterization across algorithms or environments. It also clarifies that the experiments visualize the exact spectral mechanism rather than serving as a benchmark comparison.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical results state the relevant assumptions on the finite-state MDP, policies, transition matrices, and stepsizes. The proofs are given in the main paper, including the spectral and Routh–Hurwitz calculations used for the counterexample.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experiments are based on the explicit construction in Example 4.5, with the MDP, target and behavior policies, choice of η, and clock variants described in the paper. The supplemental script reproduces the plotted eigenvalue portrait and online TD curves.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The submitted supplementary material includes the self-contained Python script used to generate the experimental figures. No external dataset is required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The MDP construction and the value $\eta = 2\varepsilon$ are specified in the experiments section. The remaining numerical settings, including stepsize schedule, number of steps, sampling interval, seed, and initialization, are recorded in the supplemental script.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because the experiments are illustrative and the main instability claim follows from the exact Routh–Hurwitz calculation. The main online TD figure uses a fixed seed to visualize the theoretically predicted separation, and Appendix B reports additional seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments section states that the figures are generated by a CPU-only Python script, with no GPU or external dataset required. The largest reported runs use the 25-state instance and 10^{11} behavior-chain steps per seed. Memory is dominated by small dense transition matrices and cached trajectory summaries.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and computational, uses no private or sensitive data, and does not involve human subjects. The authors have no known deviations from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work is foundational reinforcement learning theory and does not introduce a deployed system, dataset, or model with a direct societal use case. No specific positive or negative societal impact is identified beyond general scientific understanding of algorithmic stability.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk of misuse. The only planned release is code for reproducing the mathematical figures and simulations.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use existing datasets, pretrained models, or benchmark assets. The experiments are generated from the finite MDP constructed in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The only new asset is the supplementary code used to reproduce the figures. Its parameters and construction are documented in the script and correspond to the experiment description in the paper.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, data annotation, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core method, theorem development, experimental protocol, or originality of the research. Any writing or formatting assistance does not require declaration under the stated policy.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.