Journal of Machine Learning Research 27 (2026) 1-36 Submitted 9/24; Revised 12/25; Published 3/26

Almost Sure Convergence of Linear Temporal Difference Learning with Arbitrary Features

Jiuqi Wang JIUQI@EMAIL.VIRGINIA.EDU
Department of Computer Science
University of Virginia
85 Engineer's Way, Charlottesville, VA, 22903

Shangtong Zhang SHANGTONG@VIRGINIA.EDU
Department of Computer Science, University of Virginia
University of Virginia
85 Engineer's Way, Charlottesville, VA, 22903

Editor: Laurent Orseau

Abstract

Temporal difference (TD) learning with linear function approximation (linear TD) is a classic and powerful prediction algorithm in reinforcement learning. While it is well-understood that linear TD converges almost surely to a unique point, this convergence traditionally requires the assumption that the features used by the approximator are linearly independent. However, this linear independence assumption does not hold in many practical scenarios. This work is the first to establish the almost sure convergence of linear TD without requiring linearly independent features. We prove that the weight iterates of linear TD converge to a bounded set, and that the value estimates derived from the weights in that set are the same almost everywhere. We also establish a notion of local stability of the weight iterates. Importantly, we do not impose assumptions tailored to feature dependence and do not modify the linear TD algorithm. Key to our analysis is a novel characterization of bounded invariant sets of the mean ODE of linear TD.

Keywords: temporal difference learning, linear function approximation, reinforcement learning, almost sure convergence, bounded invariant sets

1. Introduction

Function approximation is crucial in reinforcement learning (RL) algorithms when the problem involves an intractable discrete or continuous state space (Sutton and Barto, 2018). The idea is to encode the states into finite-dimensional real-valued vectors called features. A parameterized function of the features, with learnable weights, is then used to approximate the desired function. For instance, linear function approximation computes the approximated value by taking the dot product of the feature and the weight. Most existing convergence results with linear function approximation assume the features are linearly independent (Tsitsiklis and Roy, 1996; Konda and Tsitsiklis, 1999; Sutton et al., 2008, 2009; Maei, 2011; Yu, 2015; Sutton et al., 2016; Lee and He, 2019; Nachum et al., 2019; Zou et al., 2019; Carvalho et al., 2020; Zhang et al., 2020a,b, 2021a,b; Zhang and Whiteson, 2022; Zhang et al., 2023; Qian and Zhang, 2025). However, the linear independence assumption is not desired for at least four reasons. First, many well-known empirical successes of RL with linear function approximation (Liang et al., 2016; Azagirre et al., 2024) do not have linearly independent features, leaving a gap between theory and practice. Second, in the continual learning setting (Ring, 1994; Khetarpal et al., 2022; Abel et al., 2023), the observations received by an agent are usually served one after another. There is no way to verify whether the features used in the observations are linearly independent. Third, the features are sometimes constructed via neural networks (Chung et al., 2019). Usually, it is impossible to guarantee that those neural network-based features are linearly independent. Fourth, sometimes the features are gradients of another neural network, e.g., in the compatible feature framework for actor-critic algorithms (see Sutton et al. (1999); Konda and Tsitsiklis (1999); Zhang et al. (2020b) for details). One cannot guarantee those features are linearly independent either. As a result, although the linear independence assumption greatly simplifies the theoretical analysis, it is unrealistically restrictive. Dayan (1992) and Tsitsiklis and Roy (1996, 1999) also identify the removal of the linear independence assumption of the features as a future research direction.

This work makes progress towards closing this gap using linear temporal difference (TD) learning (Sutton, 1988) as an example, since linear TD is arguably one of the most fundamental RL algorithms. In particular, this work is the first to establish the almost sure convergence of linear TD without requiring linearly independent features. The main contributions of this work are

1. characterization of TD fixed points under arbitrary features and proof that all such fixed points produce the same value estimate almost everywhere;
2. mean ODE analysis with Jordan normal form demonstrating every ODE trajectory converges to an initial-condition-dependent point;
3. characterization of bounded invariant sets of the mean ODE in the absence of global asymptotic stability;
4. establishment of a notion of local stability of the weight iterates.

Importantly, our proof does not introduce assumptions tailored to feature dependence and keep linear TD in its original form. Building upon our work, Xie et al. (2025) conduct finite sample analysis of linear TD under arbitrary features to characterize its convergence rate, further completing the understanding of linear TD in its canonical form.

2. Background

Notations. A square complex matrix (not necessarily symmetric) $M \in \mathbb{C}^{n\times n}$ is said to be positive definite if, for any non-zero vector $x \in \mathbb{C}^n$, it holds that $\mathrm{Re}(x^H M x) > 0$, where x^H denotes the conjugate transpose of x and $\mathrm{Re}(\cdot)$ denotes the real part. A matrix M is negative definite if $-M$ is positive definite. Likewise, a matrix M is positive semi-definite if $\mathrm{Re}(x^H M x) \geq 0$ for any non-zero $x \in \mathbb{C}^n$. A matrix M is negative semi-definite if $-M$ is positive semi-definite. Given a vector $x \in \mathbb{C}^n$, we define the ℓ_2 norm $\|x\| \doteq \sqrt{x^H x}$. The ℓ_2 norm $\|\cdot\|$ also induces a matrix norm. Given a matrix $M \in \mathbb{C}^{m\times n}$, the induced matrix norm is defined as $\|M\| \doteq \sup_{x\in\mathbb{C}^n, x\neq 0} \frac{\|Mx\|}{\|x\|}$. We now restrict ourselves to real vectors and matrices. A real symmetric positive definite matrix $D \in \mathbb{R}^n$ induces a vector norm $\|\cdot\|_D$, where $\|x\|_D \doteq \sqrt{x^\top D x}$ for $x \in \mathbb{R}^n$. We overload $\|\cdot\|_D$ to also denote the induced matrix norm.

We consider a Markov Reward Process (MRP[1], Bellman (1957); Puterman (2014)) with a state space $\mathcal{S} \subseteq \mathbb{R}^K$ for some $K \in \mathbb{N}$. The MRP employs an initial distribution $p_0 : \mathcal{B}(\mathcal{S}) \to [0,1]$, where $\mathcal{B}$ denotes the Borel algebra. The dynamics of the MRP are characterized by a transition kernel $p : \mathcal{B}(\mathcal{S}) \times \mathcal{S} \to [0,1]$[2]. The MRP also adopts a measurable and bounded reward function $r : \mathcal{S} \to \mathbb{R}$. At time step t, suppose the agent is at state S_t. It transitions to the next state $S_{t+1} \sim p(\cdot|S_t)$ following the transition kernel. At the same time, the environment emits a reward $R_{t+1} \doteq r(S_t)$ to the agent for transitioning out of S_t. Note that (p, p_0) defines a Markov chain $\{S_t\}$. We further define a p-induced operator $P_\pi : \mathcal{F} \to \mathcal{F}$, where $\mathcal{F} \doteq \{f : \mathcal{S} \to \mathbb{R}^n \mid f \text{ is measurable and integrable}\}$, as $(P_\pi f)(s) \doteq \int_\mathcal{S} f(s')p(\mathrm{d}s' \mid s)$.

The state-value function $v_\pi : \mathcal{S} \to \mathbb{R}$ is defined as $v_\pi(s) \doteq \mathbb{E}\big[\sum_{t=0}^\infty \gamma^t R_{t+1} | S_0 = s\big]$, where $\gamma \in [0,1)$ is a discount factor. The state-value function maps each state to the expected cumulative reward the agent gets starting from that state. We consider a linear function approximation of v_π with a feature mapping $x : \mathcal{S} \to \mathbb{R}^d$ and a weight vector $w \in \mathbb{R}^d$. The function x maps each state $s \in \mathcal{S}$ to a d-dimensional real vector $x(s)$. One simply takes $x(s)^\top w$ to compute the approximated state value for s. The goal is thus to adjust w such that $x(s)^\top w \approx v_\pi(s)$ for all $s \in \mathcal{S}$. Linear TD updates the weight w iteratively as

$$w_{t+1} = w_t + \alpha_t \Big(R_{t+1} + \gamma x(S_{t+1})^\top w_t - x(S_t)^\top w_t \Big) x(S_t). \qquad \text{(Linear TD)}$$

where α_t is the learning rate and we recall that $\{S_t\}$ is a Markov chain evolving as $S_{t+1} \sim p(\cdot|S_t)$ and S_0 is sampled from an arbitrary initial distribution. The following assumptions are commonly made in analyzing linear TD (Tsitsiklis and Roy, 1996).

Assumption 2.1 *The learning rates $\{\alpha_t\}$ is a decreasing sequence of positive real numbers such that $\sum_{t=0}^\infty \alpha_t = \infty$, $\sum_{t=0}^\infty \alpha_t^2 < \infty$, and $\lim_{t\to\infty} \left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right) < \infty$.*

Assumption 2.2 *The Markov chain $\{S_t\}$ admits a well-defined and unique stationary distribution $\mu : \mathcal{B}(\mathcal{S}) \to [0,1]$, such that $\mu(Z) = \int_\mathcal{S} p(Z \mid s)\mu(\mathrm{d}s)$ for all $Z \in \mathcal{B}(\mathcal{S})$ and $\mu(U) > 0$ for all non-empty open sets $U \subseteq \mathcal{S}$.*

The sequence $\left\{\frac{1}{(t+1)^k}\right\}$ where $k \in (0.5, 1]$ satisfies Assumption 2.1 as a valid example. In light of Assumption 2.2, we define an inner product $\langle \cdot, \cdot \rangle_\mu$ induced by μ as $\langle f, g \rangle_\mu \doteq \int_\mathcal{S} \langle f(s), g(s) \rangle \mu(\mathrm{d}s)$, where $f : \mathcal{S} \to \mathbb{R}^n$ and $g : \mathcal{S} \to \mathbb{R}^n$ are measurable functions. The inner product $\langle \cdot, \cdot \rangle_\mu$ further induces a semi-norm $\|f\|_\mu^2 \doteq \langle f, f \rangle_\mu$.

Assumption 2.3 *The feature function and reward function are bounded, i.e., $\sup_{s\in\mathcal{S}} \|x(s)\| < \infty, \sup_{s\in\mathcal{S}} |r(s)| < \infty$.*

We use $L_p(\mathcal{S}, \mu)$ to denote the set of functions defined on $\mathcal{S}$ that are p-integrable with respect to μ. A direct consequence of Assumption 2.3 is that $x \in L_p(\mathcal{S}, \mu)$ for $p \in [1, \infty)$ because $\left(\int_\mathcal{S} \|x(s)\|^p \mu(\mathrm{d}s)\right)^{1/p} \leq \left((\sup_{s\in\mathcal{S}} \|x(s)\|)^p \int_\mathcal{S} \mu(\mathrm{d}s)\right)^{1/p} = \sup_{s\in\mathcal{S}} \|x(s)\| < \infty$.

[1] In policy evaluation, the policy π is fixed and induces an MRP from a Markov Decision Process (MDP).
[2] We define the kernel this way to be compatible with the conventional notation $p(\cdot|s)$

Assumption 2.4 *Given any $f : \mathcal{S} \to \mathbb{R}^n$ and $f = 0$ a.e. with respect to μ, it holds that $P_\pi f = 0$ a.e. with respect to μ as well.*

In the rest of the paper, we use "a.e." to denote "a.e. with respect to μ" for simplifying notations. Notably, Assumptions 2.2 - 2.4 trivially hold when $\mathcal{S}$ is finite and $\{S_t\}$ is irreducible.

We can represent the feature mapping x as $x(s) = \begin{bmatrix} x_1(s) & x_2(s) & \cdots & x_d(s) \end{bmatrix}^\top$, where each $x_i : \mathcal{S} \to \mathbb{R}$ is a basis function for $i = 1, 2, \dots, d$. Since it is defined on a general state space, we define the linear dependence/independence of the features in the a.e. sense. We say the set of basis functions $\{x_1, x_2, \dots, x_d\}$ are linearly dependent a.e. if there exists $c \in \mathbb{R}^d \setminus \{0\}$, such that $c_1 x_1 + c_2 x_2 + \cdots + c_d x_d = 0$ a.e. Likewise, we say the set of basis functions $\{x_1, x_2, \dots, x_d\}$ are linearly independent a.e. if $c_1 x_1 + c_2 x_2 + \cdots + c_d x_d = 0$ a.e. holds if and only if $c = 0$. In the rest of the text, we omit a.e. from linear dependence/independence whenever it does not confuse.

Intuitively, when $\{S_t\}$ reaches the stationary distribution and the learning rate α is "small", linear TD behaves like a deterministic algorithm as $w_{t+1} = w_t + \alpha_t(Aw_t + b)$, where

$$
\begin{aligned}
A &\doteq \mathbb{E}_{S \sim \mu, S' \sim p(\cdot|S)} \left[x(S) \left(\gamma x(S')^\top - x(S)^\top \right) \right] \\
&= \textstyle\int_{\mathcal{S}} x(s) \left(\gamma (P_\pi x)(s)^\top - x(s)^\top \right) \mu(\mathrm{d}s) \in \mathbb{R}^{d \times d} && (1) \\
b &\doteq \mathbb{E}_{S \sim \mu}[x(S) r(S)] = \textstyle\int_{\mathcal{S}} x(s) r(s) \mu(\mathrm{d}s) \in \mathbb{R}^d. && (2)
\end{aligned}
$$

We can then further relate the stochastic and discrete iterative updates (Linear TD) with deterministic and continuous trajectories of the following ordinary differential equation (ODE)

$$\tfrac{\mathrm{d}w(t)}{\mathrm{d}t} = Aw(t) + b, \tag{3}$$

which is known as the ODE method in stochastic approximation (Benveniste et al., 1990; Kushner and Yin, 2003; Borkar, 2009; Borkar et al., 2025; Liu et al., 2025a).

In the finite case, it is also well known that P_π is nonexpansive and A is negative semi-definite (Tsitsiklis and Roy, 1996) (if $\{x_i\}$ are linearly independent, then A would be negative definite). These are also true in the continuous state space we consider.

Lemma 1 *Let Assumption 2.2 hold. For any $v \in L_2(\mathcal{S}, \mu)$, it holds that $\|P_\pi v\|_\mu \leq \|v\|_\mu$.*

See Appendix B.1 for the proof.

Lemma 2 *Let Assumptions 2.2 & 2.3 hold. Then A is negative semi-definite.*

See Appendix B.2 for the proof.

3. TD Fixed Points

As previously discussed, linear TD is approximately a stochastic discretization of ODE (3). The fixed points of linear TD, commonly referred to as TD fixed points (Tsitsiklis and Roy, 1996; Sutton and Barto, 2018), are linked to the equilibria of this ODE. We hence consider the linear system

$$Aw + b = 0, \tag{4}$$

where A is defined in (1) and b is defined in (2). With linearly independent features, (4) adopts a unique solution — the unique TD fixed point. Without assuming linear independence, matrix A is merely negative semi-definite (Lemma 2), so (4) can potentially adopt infinitely many solutions. In light of this, we refer to all solutions to the linear system as TD fixed points. Namely, we define $\mathcal{W}_* \doteq \{w_* | Aw_* + b = 0\}$ and refer to $\mathcal{W}_*$ as *the set of TD fixed points*. A few questions arise naturally from this definition.

(Q1) Is $\mathcal{W}_*$ always non-empty?

(Q2) If $\mathcal{W}_*$ contains multiple weights, do those weights give the same value estimate?

(Q3) Do the iterates $\{w_t\}$ generated by (Linear TD) converge to $\mathcal{W}_*$?

We shall give affirmative answers to all the questions above in the rest of the paper. We use

$$v_w(s) \doteq x(s)^\top w \tag{5}$$

to denote the value estimate for a state s given a weight w. Below, we answer (Q1) and (Q2) affirmatively as a warm-up. The affirmative answer to (Q3) is much more involved and is deferred to the next section.

Lemma 3 *Let Assumptions 2.2 & 2.3 hold. Then,* $w^\top Aw = 0 \iff v_w = 0$ *a.e.*

The proof is in Appendix C.1.

Lemma 4 *Let Assumptions 2.2, 2.3, & 2.4 hold. Then* $\mathcal{W}_*$ *is non-empty.*

The proof is in Appendix C.2. This answers (Q1) affirmatively.

Lemma 5 *Let Assumptions 2.2 & 2.3 hold. Given any* $w \in \mathcal{W}_*$ *and any* $w' \in \mathbb{R}^d$*, it holds that* $v_w = v_{w'}$ *a.e.* $\iff w' \in \mathcal{W}_*$.

The proof is in Appendix C.3. This answers (Q2) affirmatively.

4. ODE Solutions

In this section, we study the mean ODE (3) related to linear TD and establish its convergence. We use $w(t; w_0)$ to denote the solution to (3) with the initial condition $w(0; w_0) = w_0$. Recall that matrix A would be negative definite if the features are linearly independent. Then, it would follow from standard dynamical system results (Khalil, 2002) that $\lim_{t\to\infty} w(t; w_0) = -A^{-1}b$. In other words, regardless of the initial condition w_0, a solution always converges to the globally asymptotically stable equilibrium $-A^{-1}b$. Without assuming linear independence, matrix A is merely negative semi-definite (Lemma 2). It is then impractical to expect all solutions to converge to the same point. However, can we still expect each $w(t; w_0)$ to converge to a w_0-dependent limit? The answer is affirmative. To proceed, we first perform a standard change of variable. For any $w_0 \in \mathbb{R}^d$ and any $w_* \in \mathcal{W}_*$, we have $\frac{\mathrm{d}(w(t;w_0)-w_*)}{\mathrm{d}t} = \frac{\mathrm{d}w(t;w_0)}{\mathrm{d}t} = Aw(t; w_0) + b - (Aw_* + b) = A(w(t; w_0) - w_*)$. This indicates that $w(t; w_0) - w_*$ is a solution to the shifted ODE

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = Az(t). \tag{6}$$

starting from $w_0 - w_*$. Therefore, to study the original ODE (3), it is sufficient to study the shifted ODE (6). We use $z(t; z_0)$ to denote a solution to (6) with the initial condition $z(0; z_0) = z_0$. We then have

$$z(t; w_0 - w_*) = w(t; w_0) - w_* \tag{7}$$

When it does not confuse, we write $z(t; z_0)$ as $z(t)$ for simplicity. Analogous to the definition of $\mathcal{W}_*$, we define $\mathcal{Z}_* \doteq \left\{ z \middle| Az = 0 \right\}$.

Corollary 6 *Let Assumptions 2.2 & 2.3 hold. Then* $v_z = 0$ *a.e.* $\iff z \in \mathcal{Z}_*$.

Proof We recall that Lemma 5 holds for any reward function r. By letting $r = 0$, we obtain $b = 0$ and $\mathcal{W}_* = \mathcal{Z}_*$. Therefore, given any $z_* \in \mathcal{Z}_*$ and any $z \in \mathbb{R}^d$, it holds that $v_{z_*} = v_z$ a.e. $\iff z \in \mathcal{Z}_*$. Furthermore, since $z_* \in \mathcal{Z}_*$, it holds that $Az_* = 0$, which implies $z_*^\top A z_* = 0$. Hence, by Lemma 3, it holds that $v_{z_*} = 0$ a.e. As a result, we have $v_z = 0$ *a.e.* $\iff z \in \mathcal{Z}_*$, which completes the proof. ■

4.1 Value Convergence

Here, we prove the value estimate of the mean ODE (3) converges for almost all states.

Theorem 7 *Let Assumptions 2.2, 2.3, & 2.4 hold. For any* $w_0 \in \mathbb{R}^d$ *and any* $w_* \in \mathcal{W}_*$, *there exists* $\mathcal{S}^+ \subseteq \mathcal{S}$ *with* $\mu(\mathcal{S}^+) = 1$, *such that* $\forall s \in \mathcal{S}^+$, $\lim_{t\to\infty} v_{w(t;w_0)}(s) = v_{w_*}(s)$.

Proof Fix an arbitrary $z_* \in \mathcal{Z}_*$. We define $U(z) \doteq \frac{1}{2}\|z - z_*\|^2$. Then, for any $z_0 \in \mathbb{R}^d$, we have

$$\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} = (z(t) - z_*)^\top \frac{\mathrm{d}z(t)}{\mathrm{d}t} = (z(t) - z_*)^\top A z(t) \tag{8}$$

$$= (z(t) - z_*)^\top A (z(t) - z_*) \leq 0. \qquad \text{(Lemma 2)}$$

We now claim

$$\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} = 0 \iff z(t) \in \mathcal{Z}_*. \tag{9}$$

To see $\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} = 0 \impliedby z(t) \in \mathcal{Z}_*$, we note $Az(t) = 0$ by definition when $z(t) \in \mathcal{Z}_*$. Hence, $\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} = (z(t) - z_*)^\top A z(t) = 0$. To see $\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} = 0 \implies z(t) \in \mathcal{Z}_*$, we note that $v_{z(t)-z_*} = 0$ a.e. when $\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} = 0$ according to Lemma 3. That implies $z(t) - z_* \in \mathcal{Z}_*$ by Corollary 6, which further implies $z(t) \in \mathcal{Z}_*$ by the definition of $\mathcal{Z}_*$.

The result in (9) suggests that U is almost a Lyapunov function except that $\frac{\mathrm{d}U(z(t))}{\mathrm{d}t}$ has multiple zeros. The standard Lyapunov stability theorem (e.g., Theorem 4.1 of Khalil (2002)) thus does not apply.

Theorem 4.1 *(LaSalle's theorem, Theorem 4.4 of Khalil (2002)) Let* $\Omega \subset \mathbb{R}^d$ *be a compact set that is positively invariant*[3] *with respect to* (6). *Let* $U : \mathbb{R}^d \to \mathbb{R}$ *be a continuously*

[3] A set Z is a positively invariant set of (6) if $z(0) \in Z \implies \forall t \in [0, \infty), z(t) \in Z$. Here, $z(t)$ denotes a solution to (6) on $[0, \infty)$.

differentiable function such that $\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} \leq 0$ *whenever* $z(t) \in \Omega$*. Let* E *be the set of all points in* Ω *satisfying* $\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} = 0$ *whenever* $z(t) \in E$*. Let* M *be the largest invariant set*[4] *in* E*. Then every solution* $z(t)$ *with* $z_0 \in \Omega$ *satisfies*[5] $\lim_{t\to\infty} d(z(t), M) = 0$.

Given any $z_0 \in \mathbb{R}^d$, to apply LaSalle's theorem, we define $\Omega \doteq \{z \mid \|z - z_*\| \leq \|z_0 - z_*\|\}$. This set Ω is compact because it defines a closed ball centered on z_* in $\mathbb{R}^d$, thus is closed and bounded. We first show that Ω is positively invariant. For any $z_0' \in \Omega$, it holds that for any $t \geq 0$, $\|z(t; z_0') - z_*\|^2 \leq \|z(0; z_0') - z_*\|^2 = \|z_0' - z_*\|^2 \leq \|z_0 - z_*\|^2$, where the first inequality holds because $\frac{\mathrm{d}\left\|z(t;z_0') - z_*\right\|^2}{\mathrm{d}t} \leq 0$ by (8) and the second inequality holds because $z_0' \in \Omega$. Then, we conclude that $z(t; z_0') \in \Omega$ for any $t \geq 0$, implying that Ω is positively invariant. We use our previously defined U as the U for Theorem 4.1. Then $\frac{\mathrm{d}U(z(t))}{\mathrm{d}t} \leq 0$ holds by (8).

In light of (9), set E defined in Theorem 4.1 is then $E = \mathcal{Z}_* \cap \Omega$. We now show that E itself is an invariant set, so that the set M defined in Theorem 4.1 is just E. Let $z(t; z_0)$ be a solution to (6) in $(-\infty, \infty)$ with $z_0 \in E$. Define $\mathcal{S}_{z_0}^+ \doteq \{s \mid v_{z_0}(s) = 0, s \in \mathcal{S}\}$. Since $z_0 \in \mathcal{Z}_*$, it holds that $\mu(\mathcal{S}_{z_0}^+) = 1$ by Corollary 6. Then, for any $t \in (-\infty, \infty)$ and $s' \in \mathcal{S}_{z_0}^+$, we have

$$
\begin{aligned}
z(t) =& z_0 + \textstyle\int_0^t Az(\tau)\,\mathrm{d}\tau\,, && (10)\\
x(s')^\top z(t) =& x(s')^\top z_0 + \textstyle\int_0^t \int_{\mathcal{S}} x(s')^\top x(s)\big(\gamma(P_\pi x)(s)^\top - x(s)^\top\big)\mu(\mathrm{d}s)z(\tau)\,\mathrm{d}\tau \\
=& \textstyle\int_0^t \int_{\mathcal{S}} x(s')^\top x(s)\big(\gamma(P_\pi x)(s)^\top - x(s)^\top\big)\mu(\mathrm{d}s)z(\tau)\,\mathrm{d}\tau\,. && (s' \in \mathcal{S}_{z_0}^+)
\end{aligned}
$$

Then, recall the definition of v_w in (5), we have

$$
\begin{aligned}
v_{z(t)}(s') &= \textstyle\int_0^t \int_{\mathcal{S}} v_{x(s')}(s)\big(\gamma(P_\pi v_{z(\tau)})(s) - v_{z(\tau)}(s)\big)\mu(\mathrm{d}s)\,\mathrm{d}\tau \\
&= \textstyle\int_0^t \left\langle v_{x(s')}, \gamma P_\pi v_{z(\tau)} - v_{z(\tau)}\right\rangle_\mu \mathrm{d}\tau = \int_t^0 \left\langle v_{x(s')}, v_{z(\tau)} - \gamma P_\pi v_{z(\tau)}\right\rangle_\mu \mathrm{d}\tau\,.
\end{aligned}
$$

Therefore, when $t \geq 0$, we get

$$
\begin{aligned}
v_{z(t)}(s')^2 =& \Big(\textstyle\int_0^t \left\langle v_{x(s')}, \gamma P_\pi v_{z(\tau)} - v_{z(\tau)}\right\rangle_\mu \mathrm{d}\tau\Big)^2 \\
\leq& t \textstyle\int_0^t \left\langle v_{x(s')}, \gamma P_\pi v_{z(\tau)} - v_{z(\tau)}\right\rangle_\mu^2 \mathrm{d}\tau && \text{(Cauchy-Schwarz inequality)} \\
\leq& t\left\|v_{x(s')}\right\|_\mu^2 \textstyle\int_0^t \left\|\gamma P_\pi v_{z(\tau)} - v_{z(\tau)}\right\|_\mu^2 \mathrm{d}\tau\,. && \text{(Cauchy–Schwarz inequality)}
\end{aligned}
$$

It then holds that

$$
\begin{aligned}
\textstyle\int_{\mathcal{S}} v_{z(t)}(s')^2\mu(\mathrm{d}s') \leq& t \textstyle\int_{\mathcal{S}} \left\|v_{x(s')}\right\|_\mu^2 \mu(\mathrm{d}s') \int_0^t \left\|\gamma P_\pi v_{z(\tau)} - v_{z(\tau)}\right\|_\mu^2 \mathrm{d}\tau \\
\leq& tC \textstyle\int_0^t \left\|\gamma P_\pi v_{z(\tau)} - v_{z(\tau)}\right\|_\mu^2 \mathrm{d}\tau && (C \doteq \sup_{s'\in\mathcal{S}} \left\|v_{x(s')}\right\|_\mu^2 < \infty) \\
\leq& tC \textstyle\int_0^t \left(\gamma^2\left\|P_\pi v_{z(\tau)}\right\|_\mu^2 + 2\gamma\left\|P_\pi v_{z(\tau)}\right\|_\mu \left\|v_{z(\tau)}\right\|_\mu + \left\|v_{z(\tau)}\right\|_\mu^2\right) \mathrm{d}\tau \\
\leq& tC(\gamma+1)^2 \textstyle\int_0^t \left\|v_{z(\tau)}\right\|_\mu^2 \mathrm{d}\tau\,. && \text{(Lemma 1)}
\end{aligned}
$$

[4]A set Z is an invariant set of (6) if $z(0) \in Z \implies \forall t \in (-\infty, \infty), z(t) \in Z$. Here, $z(t)$ denotes a solution to (6) on $(-\infty, \infty)$.

[5]The distance between a point z and a set Ω is defined as $d(z, \Omega) \doteq \inf_{z'\in\Omega} \|z - z'\|$.

We thus have $\left\|v_{z(t)}\right\|_\mu^2 \leq tC(\gamma+1)^2 \int_0^t \left\|v_{z(\tau)}\right\|_\mu^2 \mathrm{d}\tau$ for $t \geq 0$. By Gronwall's inequality (see, e.g., Theorem A.1 of Liu et al. (2025a)), we get $\left\|v_{z(t)}\right\|_\mu^2 \leq 0 \exp\left(t^2 C(\gamma+1)^2\right) = 0$ for $t \geq 0$. Similarly, when $t \leq 0$, we have

$$\begin{aligned}
v_{z(t)}(s')^2 =& \left(\int_t^0 \left\langle v_{x(s')}, v_{z(\tau)} - \gamma P_\pi v_{z(\tau)} \right\rangle_\mu \mathrm{d}\tau\right)^2 \\
\leq& -t \int_t^0 \left\langle v_{x(s')}, v_{z(\tau)} - \gamma P_\pi v_{z(\tau)} \right\rangle_\mu^2 \mathrm{d}\tau && \text{(Cauchy-Schwarz inequality)} \\
\leq& -t \left\|v_{x(s')}\right\|_\mu^2 \int_t^0 \left\|v_{z(\tau)} - \gamma P_\pi v_{z(\tau)}\right\|_\mu^2 \mathrm{d}\tau\,. && \text{(Cauchy-Schwarz inequality)}
\end{aligned}$$

Following the same steps as above, we arrive at $\left\|v_{z(t)}\right\|_\mu^2 \leq -tC(\gamma+1)^2 \int_t^0 \left\|v_{z(\tau)}\right\|_\mu^2 \mathrm{d}\tau$ for $t \leq 0$. By a reverse time version of Gronwall's inequality (see, e.g., Theorem A.2 of Liu et al. (2025a)), it holds that $\left\|v_{z(t)}\right\|_\mu^2 \leq 0 \exp\left(t^2 C(\gamma+1)^2\right) = 0$ for $t \leq 0$. The fact $\left\|v_{z(t)}\right\|_\mu^2 = 0$ for $t \in (-\infty, \infty)$ implies that $v_{z(t)} = 0$ a.e. for $t \in (-\infty, \infty)$. According to Corollary 6, we have $z(t) \in \mathcal{Z}_*$ for $t \in (-\infty, \infty)$. This means $Az(\tau) = 0$ in (10), implying that $z(t) = z_0 \in E$ for all $t \in (-\infty, \infty)$. We have now proved that E is an invariant set of (6). Theorem 4.1 then implies that for any z_0,

$$\lim\nolimits_{t\to\infty} d(z(t; z_0), E) = 0. \tag{11}$$

We now convert the convergence of $z(t; z_0)$ back to $w(t; w_0)$. To this end, fix any $z_* \in E$. Define $\mathcal{S}_{z_*}^+ \doteq \left\{s \mid x(s)^\top z_* = 0, s \in \mathcal{S}\right\}$. Since $z_* \in E$, it holds that $\mu(\mathcal{S}_{z_*}^+) = 1$ by Corollary 6. For any $s \in \mathcal{S}_{z_*}^+$, it holds that $\inf_{z\in E} |\langle x(s), z(t; z_0) - z\rangle| = \inf_{z\in E} |\langle x(s), z(t; z_0)\rangle - \langle x(s), z\rangle|$ $\geq |\langle x(s), z(t; z_0)\rangle| - |\langle x(s), z_*\rangle|$. Therefore, we have

$$\begin{aligned}
\lim\nolimits_{t\to\infty} |\langle x(s), z(t; z_0)\rangle| &\leq \lim\nolimits_{t\to\infty} \inf\nolimits_{z\in E} |\langle x(s), z(t; z_0) - z\rangle| + |\langle x(s), z_*\rangle| \\
&\leq \|x(s)\| \lim\nolimits_{t\to\infty} \inf\nolimits_{z\in E} \|z(t; z_0) - z\| + |\langle x(s), z_*\rangle| \\
&= \|x(s)\| \lim_{t\to\infty} d(z(t; z_0), E) + |\langle x(s), z_*\rangle| = |\langle x(s), z_*\rangle| = 0. \\
& && \text{(By (11))}
\end{aligned}$$

It then follows immediately that

$$\lim\nolimits_{t\to\infty} \langle x(s), z(t; z_0)\rangle = 0. \tag{12}$$

Since (12) holds for any z_0 and corresponding trajectory $z(t; z_0)$, it also holds for the trajectory $w(t; w_0) - w_*$ that starts from $w_0 - w_*$, i.e.,

$$\begin{aligned}
&\lim\nolimits_{t\to\infty} x(s)^\top (w(t; w_0) - w_*) = 0 \\
&\qquad \lim\nolimits_{t\to\infty} v_{w(t;w_0)}(s) = v_{w_*}(s),
\end{aligned}$$

which completes the proof. ■

4.2 Weight Convergence

The value convergence in Theorem 7 immediately implies that any solution $w(t; w_0)$ would eventually converge to the set $\mathcal{W}_*$ as time progresses. But is it possible that $w(t; w_0)$ keeps oscillating within $\mathcal{W}_*$ or in neighbors of $\mathcal{W}_*$ without ever converging to any single point? In this section, we rule out this possibility and prove that any solution will always converge to some fixed point, formalized in the next theorem.

Theorem 8 *Let Assumptions 2.2, 2.3, & 2.4 hold. For any $w_0 \in \mathbb{R}^d$, there exists a constant $w_\infty(w_0) \in \mathcal{W}_*$, such that $\lim_{t\to\infty} w(t; w_0) = w_\infty(w_0)$.*

We shall perform a finer analysis of the mean ODE (6) and propose several helper lemmas to prove this theorem. Firstly, it is well-known that $z(t; z_0)$ has a closed-form solution (Khalil, 2002) as

$$z(t; z_0) = \exp(At)z_0. \tag{13}$$

The standard approach to work with matrix exponential[6] is to consider the Jordan normal form (Horn and Johnson, 2012). Namely, we decompose $A = PJP^{-1}$. Here, P is the invertible matrix in Jordan decomposition and J is the Jordan matrix of A. We use $\lambda_1, \lambda_2, \dots, \lambda_k$ to denote the k distinct eigenvalues of A. We use $m_1, \dots, m_k$ to denote the algebraic multiplicity of each eigenvalue. Likewise, we use $g_1, \dots, g_k$ to denote the geometric multiplicity of each eigenvalue. It always holds that $1 \le g_i \le m_i$ for $i = 1, 2, \dots, k$. Each distinct eigenvalue λ_i has exactly g_i corresponding Jordan blocks. The dimensions of each Jordan block are inconsequential for our analysis. Therefore, to simplify our notation, we use $\rho_{i,j}$ to denote the dimension of the j-th Jordan block of λ_i. Notably, $m_i = \sum_{j=1}^{g_i} \rho_{i,j}$. Then, the Jordan matrix can be expressed as $J = \bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} B_{i,j}$, where $B_{i,j} \in \mathbb{C}^{\rho_{i,j}\times\rho_{i,j}}$ is the j-th Jordan block corresponding to the eigenvalue λ_i and $\bigoplus$ denotes the direct matrix sum[7]. The matrix exponential can then be computed as

$$\begin{aligned}\exp(At) &= \textstyle\sum_{n=0}^\infty \frac{1}{n!}(PJP^{-1})^n t^n = P\exp(Jt)P^{-1} \\ &= P\exp\Big(\textstyle\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} B_{i,j}t\Big)P^{-1} = P\left[\textstyle\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} \exp(B_{i,j}t)\right]P^{-1}.\end{aligned} \tag{14}$$

As an example of the Jordan blocks, suppose $\rho_{1,1} = 3$, then $B_{1,1} = \begin{bmatrix}\lambda_1 & 1 & 0\\ 0 & \lambda_1 & 1\\ 0 & 0 & \lambda_1\end{bmatrix}$. We may also express each Jordan block as $B_{i,j} = \lambda_i I_{\rho_{i,j}} + N_{i,j}$, where $N_{i,j}$ is a nilpotent matrix[8]. Continuing with our example, $N_{1,1} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 0\end{bmatrix}$. We denote the index of $N_{i,j}$ as $y_{i,j}$. Using the property of a nilpotent matrix and matrix exponential, we can compute $\exp(B_{i,j}t)$ as

$$\exp(B_{i,j}t) = \exp\big(\lambda_i I_{\rho_{i,j}} t + N_{i,j}t\big) = \exp(\lambda_i t)\exp(N_{i,j}t)$$

[6] For a square matrix X, its exponential is $\exp(X) \doteq \sum_{n=0}^\infty \frac{1}{n!}X^n$.

[7] Given block matrices A and B, $A \bigoplus B = \begin{bmatrix}A & \\ & B\end{bmatrix}$.

[8] A nilpotent matrix (Section 0.9.13 of Horn and Johnson (2012)) is a square matrix N, such that $N^k = 0$ for some positive integer k. The smallest k is called the index of N.

$$
\begin{aligned}
&= \exp(\lambda_i t) \textstyle\sum_{n=0}^{\infty} \frac{1}{n!} t^n N_{i,j}^n && \text{(power series)}\\
&= \exp(\lambda_i t) \textstyle\sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n && (N_{i,j}^n = 0, \forall n \geq y_{i,j})\\
&= \exp(\mathrm{Re}(\lambda_i) t) \exp(\mathtt{i}\, \mathrm{Im}(\lambda_i) t) \textstyle\sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n\\
&= \exp(\mathrm{Re}(\lambda_i) t) \left(\cos(\mathrm{Im}(\lambda_i) t) + \mathtt{i} \sin(\mathrm{Im}(\lambda_i))\right) \textstyle\sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n.\\
& && \text{(Euler's formula)}
\end{aligned}
$$

Here, $\mathrm{Im}(\cdot)$ denotes the imaginary part. The fact that matrix A is negative semi-definite (Lemma 2) implies $\mathrm{Re}(\lambda_i) \leq 0$.[9] We now branch into cases for $\mathrm{Re}(\lambda) < 0$ and $\mathrm{Re}(\lambda) = 0$.

Lemma 9 *Let Assumptions 2.2 & 2.3 hold. If* $\mathrm{Re}(\lambda_i) < 0$*, then it holds that* $\lim_{t\to\infty} \exp(B_{i,j} t) = 0 \quad \forall j \in \{1, \dots, g_i\}$.

The proof is in Appendix D.1. This is intuitive because all terms other than $\exp(\mathrm{Re}(\lambda_i)t)$ are at most polynomial and are dominated by the exponential. To analyze the case of $\mathrm{Re}(\lambda_i) = 0$, we make the following two observations.

Lemma 10 *Let Assumptions 2.2 & 2.3 hold.* $\forall i, j$*, it holds that* $\sup_{t\in[0,\infty)} \|\exp(B_{i,j} t)\| < \infty$.

The proof is in Appendix D.2. Intuitively, the boundedness holds because $w(t; w_0)$ is bounded for any $w_0 \in \mathbb{R}^d$ (Lemma D.1). Furthermore, it follows from the value convergence that

Corollary 11 *Let Assumpitons 2.2, 2.3, & 2.4 hold. Then* $\forall z_0 \in \mathbb{R}^d$, $\lim_{t\to\infty} \frac{\mathrm{d}z(t;z_0)}{\mathrm{d}t} = 0$.

The proof is in Appendix D.3. We are now ready to discuss the case $\mathrm{Re}(\lambda_i) = 0$.

Lemma 12 *Let Assumptions 2.2, 2.3, and 2.4 hold. If* $\mathrm{Re}(\lambda_i) = 0$*, then it holds that for any* $j \in \{1, \dots, g_i\}$ *and any* $t \geq 0$, $\exp(B_{i,j} t) = I_{\rho_{i,j}}$.

The proof is in Appendix D.4. Intuitively, if $\mathrm{Re}(\lambda_i) = 0$, we must have $y_{i,j} = 1$. Otherwise, $\exp(B_{i,j} t)$ cannot be bounded, leading to a contradiction with Lemma 10. Furthermore, it must hold that $\mathrm{Im}(\lambda_i) = 0$. Otherwise, the derivative will not diminish, leading to a contradiction with Corollary 11.

Combining Lemmas 9 & 12, we have $\lim_{t\to\infty} \exp(B_{i,j} t) = I_{\rho_{i,j}} \mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}$. Here, $\mathbb{I}\{\cdot\}$ is the indicator function. Therefore, $\lim_{t\to\infty} \exp(At)$ exists, and we define

$$
A_\infty \doteq \lim_{t\to\infty} \exp(At) = P\Big(\textstyle\bigoplus_{i=1}^{k} \bigoplus_{j=1}^{g_i} I_{\rho_{i,j}} \mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}\Big) P^{-1}. \tag{15}
$$

Plugging this back to (13) confirms that for any trajectory $z(t; z_0)$ of the ODE (6),

$$
\lim_{t\to\infty} z(t; z_0) = A_\infty z_0. \tag{16}
$$

Proof [of Theorem 8] Now, we have everything to prove the main statement. We recall that $w(t; w_0) - w_*$ is a trajectory of (6) starting from $w_0 - w_*$. Therefore, we have from (16) that $\lim_{t\to\infty} w(t; w_0) - w_* = A_\infty(w_0 - w_*)$. In other words, $\lim_{t\to\infty} w(t; w_0) = w_\infty(w_0) \doteq A_\infty(w_0 - w_*) + w_*$. Referring to Theorem 7, there exists $\mathcal{S}^+ \subseteq \mathcal{S}$, with $\mu(\mathcal{S}^+) = 1$, such that $\lim_{t\to\infty} x(s)^\top w(t; w_0) = x(s)^\top w_\infty(w_0) = v_{w_*}(s)$ for all $s \in \mathcal{S}^+$ and $w_0 \in \mathbb{R}^d$. We get $v_{w_\infty(w_0)} = v_{w_*}$ a.e. for any $w_0 \in \mathbb{R}^d$. Therefore, we have $w_\infty(w_0) \in \mathcal{W}_*$ for all $w_0 \in \mathbb{R}^d$ by Lemma 5, which completes the proof. ■

[9] Let A be a negative semi-definite matrix and u be an eigenvector of A having eigenvalue λ, it holds that $\mathrm{Re}(u^H A u) = \mathrm{Re}(\lambda \|u\|^2) = \mathrm{Re}(\lambda)\|u\|^2 \leq 0$. Since $\|u\|^2 > 0$, we have $\mathrm{Re}(\lambda) \leq 0$.

4.3 Bounded Invariant Sets

In the ODE methods for stochastic approximation, if the mean ODE of the stochastic approximation algorithm (cf. ODE (3) for linear TD (Linear TD)) is not globally asymptotically stable, usually one can only expect that the iterates of the stochastic approximation converge to a bounded invariant set of the ODE. In light of this, we now study the bounded invariant sets of ODE (3). We first study the bounded solutions to the ODE on $(-\infty, +\infty)$.

Theorem 13 *Let Assumptions 2.2, 2.3, & 2.4 hold. Let $w(t)$ be a bounded solution to ODE (3) on $(-\infty, +\infty)$, i.e., $\sup_{t\in(-\infty,+\infty)} \|w(t)\| < \infty$. It then holds that $w(t)$ is constant and is in $\mathcal{W}_*$, i.e., there exists some $w_* \in \mathcal{W}_*$ such that $w(t) = w_*$ holds for any $t \in (-\infty, +\infty)$.*

Proof Let $z(t; z_0)$ be any bounded solution to (6) on $(-\infty, \infty)$. By (14) and (15), it holds that, for any $t \in (-\infty, \infty)$,

$$\begin{aligned}
\exp(At)A_\infty =& P\Big(\textstyle\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} \exp(B_{i,j}t)\Big)\Big(\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} I_{\rho_{i,j}}\mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}\Big)P^{-1} \\
=& P\Big(\textstyle\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} I_{\rho_{i,j}}\mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}\Big)\Big(\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} \exp(B_{i,j}t)\Big)P^{-1} \\
=& A_\infty \exp(At).
\end{aligned}$$

For any $t \in (-\infty, \infty)$ and $t' > 0$, we then have

$$\begin{aligned}
\|z(t; z_0) - A_\infty z_0\| =& \|\exp(At)z_0 - A_\infty z_0\| \\
=& \|\exp(A(t'+t) - At')z_0 - \exp(At')\exp(-At')A_\infty z_0\| \\
=& \|\exp(A(t'+t) - At')z_0 - \exp(At')A_\infty \exp(-At')z_0\| \\
\leq& \big\|\exp\big(A(t'+t)\big) - \exp\big(At'\big)A_\infty\big\|\big\|\exp\big(-At'\big)z_0\big\| \\
=& \|\exp(A(t'+t)) - \exp(At')A_\infty\|\|z(-t'; z_0)\| \\
\leq& \big\|\exp\big(A(t'+t)\big) - \exp\big(At'\big)A_\infty\big\| \sup_{t''\in(-\infty,\infty)} \big\|z(t''; z_0)\big\|.
\end{aligned} \tag{17}$$

We note that $\lim_{t'\to\infty} \exp(A(t'+t)) - \exp(At')A_\infty = A_\infty - A_\infty A_\infty$. In light of (15), it holds that

$$\begin{aligned}
A_\infty A_\infty &= P\Big(\textstyle\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} I_{\rho_{i,j}}\mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}\Big)P^{-1}P\Big(\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} I_{\rho_{i,j}}\mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}\Big)P^{-1} \\
&= P\Big(\textstyle\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} I_{\rho_{i,j}}\mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}\Big)\Big(\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} I_{\rho_{i,j}}\mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}\Big)P^{-1} \\
&= P\Big(\textstyle\bigoplus_{i=1}^k \bigoplus_{j=1}^{g_i} I_{\rho_{i,j}}\mathbb{I}\{\mathrm{Re}(\lambda_i) = 0\}\Big)P^{-1} = A_\infty.
\end{aligned}$$

Hence, we obtain $\lim_{t'\to\infty} \exp(A(t'+t)) - \exp(At')A_\infty = 0$. Taking $t' \to \infty$ on both sides of (17) then yields $\|z(t; z_0) - A_\infty z_0\| \leq 0 \cdot \sup_{t''\in(-\infty,\infty)} \|z(t''; z_0)\| = 0$, where, to obtain the last equality, we have used the boundedness of $z(t; z_0)$. This concludes that for any $t \in (-\infty, \infty)$, we have $z(t; z_0) = A_\infty z_0$. Thus, $z(t; z_0)$ is constant.

We fix any $w_* \in \mathcal{W}_*$ and recall (7). Suppose $w(t; w_0)$ is a bounded solution to (3) for $t \in (-\infty, \infty)$. Then, $z(t; w_0 - w_*)$ is a bounded solution to (6) for $t \in (-\infty, \infty)$. We have shown that $z(t; w_0 - w_*)$ is constant whenever it is a bounded solution. So, $w(t; w_0)$ is also constant. By Theorem 8, it holds that $\lim_{t\to\infty} w(t; w_0) = w_\infty(w_0) \in \mathcal{W}_*$ for any $w_0 \in \mathbb{R}^d$. Consequently, it must hold that $w(t; w_0) \in \mathcal{W}_*$ for all $t \in (-\infty, \infty)$ because it is a constant solution. ■

Theorem 13 leads to the following characterization of a bounded invariant set.

Corollary 14 *Let Assumptions 2.2, 2.3, & 2.4 hold. If $\mathcal{W}$ is a bounded invariant set of ODE* (3)*, then $\mathcal{W} \subseteq \mathcal{W}_*$.*

The proof is in Appendix D.5.

5. Convergence of Linear TD

Having fully characterized the mean ODE (3), we are now ready to connect the linear TD update (4) with the mean ODE. To this end, we consider the joint process $Y_t = (S_t, S_{t+1})$, which is also a Markov chain. Recall that $S_t \in \mathbb{R}^K$. We then regard Y_t as a vector in $\mathbb{R}^{2K}$. As a result, the Markov chain $\{Y_t\}$ evolves in a subset of $\mathbb{R}^{2K}$, denoted as $\mathcal{Y}$. Since $\{Y_t\}$ is essentially the same chain as $\{S_t\}$ but viewed differently, Assumption 2.2 implies that $\{Y_t\}$ adopts a unique stationary distribution, referred as η. We additionally refer to the transition kernel of $\{Y_t\}$ as $P_\mathcal{Y}$. Given a measurable and integrable function $f : \mathcal{Y} \to \mathbb{R}^d$, we define $P_\mathcal{Y} f$ as $(P_\mathcal{Y} f)(y) \doteq \int_\mathcal{Y} f(y') P_\mathcal{Y}(\mathrm{d}y' \mid y)$. With this joint process, the linear TD update (Linear TD) can be written as $w_{t+1} = w_t + \alpha_t H(w_t, Y_{t+1})$, where $H(w, y) = \big(r(s) + \gamma w^\top x(s') - w^\top x(s)\big)x(s)$. Here, we have used shorthand $y \doteq (s, s')$. The expected update is then $h(w) = \mathbb{E}_{y\sim\eta}[H(w, y)] = Aw + b$. We now make a few standard assumptions on the behavior of $\{Y_t\}$.

Assumption 5.1 (Poisson Equation) *There exists a function $\nu : \mathcal{Y} \to \mathbb{R}^d$, such that $\nu_w(y) - (P_\mathcal{Y}\nu_w)(y) = H(w, y) - h(w)$ holds for all w, y. Furthermore, there exists a constant $C_{5.1}$ such that $\forall w, y$, $\|\nu_w(y)\| \leq C_{5.1}(1 + \|w\|)(1 + \|y\|)$; $\|(P_\mathcal{Y}\nu_w)(y) - (P_\mathcal{Y}\nu_{w'})(y)\| \leq C_{5.1}\|w - w'\|(1 + \|y\|)$.*

Assumption 5.2 (Law of Iterated Logarithm) *For any w, there exists a sample-path-dependent finite constant (i.e., a random variable that is finite a.s.) $\zeta_{5.2}$, such that*
$\|\sum_{t=1}^n (H(w, Y_t) - h(w))\| \leq \zeta_{5.2}\sqrt{n \log\log n}$ *a.s.*

Remark 15 *Both Assumptions 5.1 & 5.2 are concerned about the behavior of the Markov chain $\{Y_t\}$. Assumption 5.1 is the standard way to handle Markovian dependence in the update noise $H(w, Y_t) - h(w)$, while Assumption 5.2 says the cumulative fluctuation of that update noise grows sufficiently slowly almost surely to be absorbed by the diminishing step sizes. Furthermore, they are weak in that, if $\mathcal{S}$ is finite and $\{S_t\}$ is irreducible and aperiodic, they hold automatically (in fact they hold for any function $H : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}^k$ and the corresponding expectation $h : \mathbb{R}^k \to \mathbb{R}^k$, not just the ones corresponding to linear TD). In particular, see Section 8.2.3 of Puterman (2014) for how Assumption 5.1 is verified for finite irreducible chains. See Remark 3 of Liu et al. (2025a) for how Assumption 5.2 is verified for finite irreducible and aperiodic chains. For a generic state space $\mathcal{S}$, there are multiple sufficient conditions to imply both. For example, if $\{Y_t\}$ is positive Harris*[10] *and satisfies the*

[10] See pages 235 and 204 of Meyn and Tweedie (2012) for the definitions of positive and Harris chains, respectively.

Lyapunov drift condition (V4)[11] *with a constant function* $V(y) = 1\,\forall y$,[12] *Assumption 5.1 can be satisfied via the boundedness of the fundamental kernel through Theorem 2.3 of Glynn and Meyn (1996). Assumption 5.2 can be satisfied via V-uniform ergodicity*[13] *(which is implied by (V4) through Theorem 16.0.1 (iv) of Meyn and Tweedie (2012)) through Theorem 17.0.1 (iii) and (iv) of Meyn and Tweedie (2012). Such implications are standard and have been extensively studied in the existing literature. But detailing them necessitates introducing quite a few additional definitions, which is not our contribution and deviates from the goal of this paper. So, we present the current forms of Assumption 5.1 & 5.2 directly. They are also widely used in existing works. For example, Assumption 5.1 is the most important assumption in Benveniste et al. (1990). Assumption 5.2 is also used in Liu et al. (2025a).*

We additionally make the following assumption regarding the state space.

Assumption 5.3 $\mathcal{S}$ *is compact.*

Remark 16 *We note that Assumption 5.3 is not needed to characterize the TD fixed point or the mean ODE in Sections 3 & 4. It is also unnecessary for linear TD itself. We employ Assumption 5.3 as a convenient sufficient condition to satisfy the assumptions made in Benveniste et al. (1990) to work with infinite state spaces. That being said, we do envision that Assumption 5.3 can be relaxed to some non-compact state space with some additional assumptions, as long as the assumptions in Benveniste et al. (1990) are satisfied. This relaxation is standard and too technical to be included in this paper, but it is worth noting that it is possible.*

The next theorem demonstrates the stability of linear TD.

Theorem 17 *Let Assumptions 2.1, 2.2, 2.3, 2.4, 5.1, & 5.3 hold. Then, the iterates* $\{w_t\}$ *generated by* (Linear TD) *is stable, i.e.,* $\sup_t \|w_t\| < \infty \quad a.s.$

The proof is in Appendix E.1. It is based on Theorem 17(a) of Benveniste et al. (1990), where we use $U(w) \doteq \|w - w_*\|^2 + \|w_*\|^2$ as an energy function. Here, w_* is any fixed point in $\mathcal{W}_*$. In the canonical analysis with linearly independent features, $\|w - w_*\|^2$ is commonly used as the Lyapunov function. Without assuming linear independence, we additionally add $\|w_*\|^2$ such that $U(w) \geq \frac{1}{2}\|w\|^2$ always holds. Apparently, this $U(w)$ is not a Lyapunov function, but it is sufficient to prove stability per Benveniste et al. (1990). In particular, this $U(w)$ satisfies Conditions (i) and (ii) on page 239 of Benveniste et al. (1990).

Having established the stability, the convergence of $\{w_t\}$ to a bounded invariant set is now expected from standard stochastic approximation results (Benveniste et al., 1990; Kushner and Yin, 2003; Borkar, 2009; Liu et al., 2025a). Here, we use Corollary 1 of Liu et al. (2025a) to establish the desired convergence, which is essentially a simplified version of Theorem 1 in Chapter 5 of Kushner and Yin (2003).

Theorem 18 *Suppose Assumptions 2.1, 2.2, 2.3, 2.4, 5.1, 5.2, & 5.3 hold. Let* $w_* \in \mathcal{W}_*$. *Then, for any* $w_0 \in \mathbb{R}^d$, *there exists* $\mathcal{S}^+ \subseteq \mathcal{S}$ *with* $\mu(\mathcal{S}^+) = 1$, *such that the iterates* $\{w_t\}$ *generated by* (Linear TD) *satisfy* $\forall s \in \mathcal{S}^+$, $\lim_{t\to\infty} x(s)^\top w_t = x(s)^\top w_* \quad a.s.$

[11] See page 386 of Meyn and Tweedie (2012) for the definition of the (V4) condition.
[12] This condition again trivially holds for finite irreducible and aperiodic chains.
[13] See page 387 of Meyn and Tweedie (2012) for the definition of V-uniform ergodicity.

The proof is in E.2. We conclude this section with an open problem. Theorem 18 is in parallel with Theorem 7 in that both are concerned with the weight convergence to a set. In light of Theorem 8, a natural question arises: can we prove that $\{w_t\}$ converges to a single (possibly sample path dependent) point? Unfortunately, we do not have a definite answer now. The best we know is that from Theorem 18, any convergent subsequence of $\{w_t\}$ will converge to some sample path dependent point in $\mathcal{W}_*$. We further show that $\{w_t\}$ exhibits some local stability along a convergent subsequence in the following sense.

Corollary 19 *Let Assumptions 2.1, 2.2, 2.3, 2.4, 5.1, 5.2, & 5.3 hold. Then, there exists at least one convergent subsequence of $\{w_t\}$, denoted as $\{w_{t_k}\}$, such that for any $T < \infty$, it holds that $\lim_{k\to\infty} \max_{t_k \le j \le m(t_k, T)} \|w_j - w_*\| = 0$, where $w_* \in \mathcal{W}_*$ is the limit of $\{w_{t_k}\}$.*

Here, $m(t, T)$ is defined as $m(t, T) \doteq \max\{n | \sum_{i=t}^{n} \alpha_i \le T\}$. Intuitively, the magnitude of the updates of linear TD in (Linear TD) is controlled by the learning rate α_i. Then, $\{t, t+1, \dots, m(t, T)\}$ denotes a period during which the total magnitude of updates is no more than T. The proof is in E.3. Corollary 19 essentially confirms that $\{w_t\}$ will visit (arbitrarily small) neighbors of w_* infinitely many times. The number of update steps during each visit (i.e., $m(t_k, T) - t_k$) diverges to ∞.

6. Finite State Space

We now consider a special case where S is finite, which is also considered in many previous works (Sutton, 1988; Dayan, 1992; Tsitsiklis and Roy, 1996). This special case allows us to make a more detailed comparison with previous approaches and provide finer characterization of the set of TD fixed points $\mathcal{W}_*$.

When $\mathcal{S}$ is finite, we can represent the transition dynamics as a stochastic matrix $P \in [0,1]^{|\mathcal{S}|\times|\mathcal{S}|}$, where $P(i,j) = p(s_j \mid s_i)$. Similarly, we overload r to denote the vector representation of the reward function when it does not confuse. We now define the Bellman operator $\mathcal{T} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as $\mathcal{T}v \doteq r + \gamma P v$, given $v \in \mathbb{R}^{|\mathcal{S}|}$. We shall also represent the feature function x compactly as a matrix $X \in \mathbb{R}^{|\mathcal{S}|\times d}$, where $x(s)$ is the s-th row of X. Remarkably, Assumptions 2.2, 2.3, 2.4, 5.1, 5.2, & 5.3 trivially hold for finite $\mathcal{S}$ when $\{S_t\}$ is irreducible and aperiodic. We also define $D \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ as the diagonal matrix whose diagonal terms are the stationary distribution of the induced Markov chain $\{S_t\}$. Under Assumption 2.2, D is symmetric and positive definite.

We now review the canonical convergence analysis of linear TD (cf. Tsitsiklis and Roy (1996)) under linearly independent features (i.e., X has a full column rank). Given any vector $v \in \mathbb{R}^{|\mathcal{S}|}$, define the projection of v onto the column space of X as

$$\Pi v \doteq \arg\min_{v_0 \in \{Xw | w \in \mathbb{R}^d\}} \|v_0 - v\|_D^2 = X \arg\min_{w\in\mathbb{R}^d} \|Xw - v\|_D^2. \tag{18}$$

Suppose the features are linearly independent, i.e., X has linearly independent columns, Tsitsiklis and Roy (1996) prove that $\Pi = X\left(X^\top D X\right)^{-1} X^\top D$. The fact that $X^\top D X$ is nonsingular follows directly from the linear independence of the features. The composition $\Pi\mathcal{T}$ of the projection operator and the Bellman operator is a contraction mapping with respect to $\|\cdot\|_D$ (Tsitsiklis and Roy, 1996) and thus adopts a unique fixed point v_*, such that

$$\Pi\mathcal{T} v_* = v_* \tag{19}$$

according to the Banach fixed-point theorem. Because of feature linear independence, there exists a unique w_*, such that $Xw_* = v_*$. This w_* has the following two remarkable properties. First, w_* is the unique zero of the mean squared projected Bellman error (MSPBE) (Sutton and Barto, 2018). Here, MSPBE is defined as $\text{MSPBE}(w) \doteq \|\Pi(\mathcal{T}Xw - Xw)\|_D^2$. Second, w_* is the unique solution to linear system (4), where $A = X^\top D(\gamma P - I)X$ and $b = X^\top Dr$ (Bertsekas and Tsitsiklis, 1996). Notably, A and b here are defined in the same way as (1) and (2). The uniqueness of these two properties follows from the fact that A is negative definite when features are linearly independent (Bertsekas and Tsitsiklis, 1996; Tsitsiklis and Roy, 1996). Under Assumptions 2.1, 2.2, and the feature linear independence assumption, Tsitsiklis and Roy (1996) prove that the iterates $\{w_t\}$ generated by (Linear TD) satisfy $\lim_{t\to\infty} w_t = w_*$ a.s. As a result, the weight w_* is the unique TD fixed point. Importantly, the negative definiteness of A ensures that the ODE (3) is globally asymptotically stable, which is key to the convergence proof of Tsitsiklis and Roy (1996). In the above convergence analysis, the linear independence of the features is vital in the following three aspects:

(i) it ensures that the TD fixed point is unique;

(ii) it ensures that ODE (3) is well-behaved;

(iii) it ensures that TD update (Linear TD) can be properly related to ODE (3).

Consequently, removing the feature linear independence assumption would at least entail three corresponding challenges. We addressed each of them in the previous sections. Namely, in Section 3, we analyzed TD fixed points with arbitrary features. In Section 4, we analyzed ODE trajectories with arbitrary features. In Section 5, we established the convergence of linear TD with arbitrary features.

Next, we show how we can relate $\mathcal{W}_*$ with the MSPBE. If X does not have full column rank, the arg min in (18) may return a set of weights instead of a unique one. In light of this, we redefine Π to always select the weight with the smallest norm. Namely, we redefine Π as

$$\Pi v \doteq X \arg\min_{w \in \arg\min_w \|Xw - v\|_D^2} \|w\|. \tag{20}$$

In the rest of the section, we always use this more general definition of Π. It turns out that this new definition of Π also enjoys a closed-form expression.

Lemma 20 *Let Assumption 2.2 hold. Let $(\cdot)^\dagger$ denote the pseudo-inverse. Then, we have $\Pi = X(D^{1/2}X)^\dagger D^{1/2}$.*

The proof is provided in Appendix F.1, where we have used standard results from least squares, see, e.g., Gallier and Quaintance (2019). Moreover, as expected, this new definition of Π preserves the desired contraction property.

Lemma 21 *Let Assumption 2.2 hold. Then, $\Pi\mathcal{T}$ is a contraction operator w.r.t. $\|\cdot\|_D$.*

The proof is provided in Appendix F.2. Banach's fixed point theorem ensures that $\Pi\mathcal{T}$ adopts a unique fixed point, referred to as v_*, such that

$$\Pi\mathcal{T} v_* = v_*. \tag{21}$$

Here, we have overloaded the definition of v_* in (19) and in the rest of the section we shall use the overloaded definition of v_*. The definition of Π in (20) then immediately ensures that there exists at least one w_*, such that

$$Xw_* = v_*, \tag{22}$$

implying $\Pi\mathcal{T}Xw_* = Xw_*$.

The next lemma mirrors Lemma 5 in the case of finite state space. The only difference is that we now have exact equivalence instead of almost everywhere equivalence for the approximated value functions.

Lemma 22 *Let Assumption 2.2 hold. Then, for any $w, w' \in \mathcal{W}_*$, we have $Xw = Xw'$.*

The proof is provided in Appendix F.3. Having redefined Π and established the equivalence between value estimates for TD fixed points, we are able to relate $\mathcal{W}_*$ to the MSPBE with the following theorem.

Theorem 23 *Let Assumption 2.2 hold. Then, $\forall w \in \mathbb{R}^d$, $Aw + b = 0 \iff \Pi\mathcal{T}Xw = Xw$.*

The proof is provided in Appendix F.4. The above equivalence implies that the w_* defined in (22) must satisfy $Aw_* + b = 0$, i.e., $w_* \in \mathcal{W}_*$. It also confirms that all weights in $\mathcal{W}_*$ minimize MSPBE.

7. Related Work

The seminal work of Sutton (1988) formalizes the idea of temporal difference learning. The linear TD update (Linear TD) in this paper is referred to as TD(0) in Sutton (1988), which is a special case of the more general TD algorithm with eligibility trace, referred to as TD(λ) in Sutton (1988). Sutton (1988) proves the convergence of linear TD(0) in expectation. Extending Sutton's work, Dayan (1992) proves the convergence of linear TD(λ) in expectation for a general λ and the almost sure convergence of tabular TD(0). Later, Dayan and Sejnowski (1994) further show the almost sure convergence of linear TD(λ). One should note that the works of Sutton (1988); Dayan and Sejnowski (1994) require the observations to be linearly independent, i.e., the feature matrix X has full row rank; in most cases, this is an even stronger assumption than having linearly independent columns, because full row rank is essentially equivalent to a tabular representation. Furthermore, the version of TD(λ) Sutton (1988); Dayan (1992); Dayan and Sejnowski (1994) consider is "semi-offline", where the weight update occurs after each sequence of observations rather than at every step. Tsitsiklis and Roy (1996) provide the first proof of almost sure convergence of linear TD(λ), assuming linearly independent features. The linear TD considered in this paper is exactly the same as Tsitsiklis and Roy (1996) with $\lambda = 0$. Tadic (2001) proves the almost sure convergence of linear TD with weaker assumptions than Tsitsiklis and Roy (1996) but still requires linearly independent features. Tadic (2001) argues (without concrete proof) that without the linear independence assumption, the projected iterates $\{\Gamma w_t\}$ converges almost surely, where $\{w_t\}$ are generated by (Linear TD) and Γ projects a vector into the row (column) space of $X^\top DX$ as $\Gamma(w) = \arg\min_{w' \in \{X^\top DXz | z \in \mathbb{R}^d\}} \|w' - w\|^2 = X^\top DX(X^\top DX)^\dagger w$. Without assuming linearly independent features, the convergence of $\{\Gamma w_t\}$ does not necessarily imply the convergence of $\{Xw_t\}$. More recently, Brandfonbrener and Bruna (2020) sought to extend the convergence guarantee of linear TD to nonlinear function approximators. They study the associated ODE of nonlinear function approximators and show that, under certain homogeneous assumptions of the function approximator, the $\liminf$ of the norm of the approximated value function is finite. Under stricter assumptions between the gradient of the function approximator and the reversibility[14] of the transition dynamics, the function approximator converges to the true value function. Cai et al. (2024) provide a finite-sample analysis of TD with a two-layer overparameterized neural network. Neural TD, their proposed TD algorithm, requires a projection operator that confines the network's weights to a ball centered on the initial weights and uses weight averaging. Under those modifications, Cai et al. (2024) establish the value convergence rate of Neural TD in expectation, including an error term that only diminishes when the width of the network goes to infinity. Despite working only with linear function approximation, we do not modify the original linear TD algorithm and still provide almost sure convergence. We envision that our results will shed light on an almost sure convergence of Neural TD, which we leave for future work.

The assumptions made about the features are not exclusive to the asymptotic analysis of linear TD. They are also prevalent in the study of finite-sample guarantees. Recent works on characterizing the convergence rate of linear TD include Bhandari et al. (2018); Lakshminarayanan and Szepesvári (2018); Srikant and Ying (2019); Chen et al. (2025); Mitra (2025). All their analyses rely on the linear independence of the features. As a matter of fact, almost all previous analysis of RL algorithm with linear function approximation assume feature linear independence, see, e.g., Sutton et al. (2008, 2009); Maei (2011); Hackman (2013); Liu et al. (2015); Yu (2015, 2016); Zou et al. (2019); Yang et al. (2019); Zhang et al. (2020a,b); Xu et al. (2020a,b); Wu et al. (2020); Chen et al. (2022); Yang et al. (2021); Qiu et al. (2021); Zhang et al. (2021a,b); Xu et al. (2021); Zhang et al. (2022); Zhang and Whiteson (2022); Zhang et al. (2023); Chen et al. (2023); Dal Fabbro et al. (2024); Wang et al. (2024); Ganesh et al. (2025); Liu et al. (2025a); Qian and Zhang (2025); Maity and Mitra (2025); Peng et al. (2025); Liu et al. (2025c). The only exception we are aware of is Xie et al. (2025), who build on our results and conduct a finite-sample analysis of linear TD under arbitrary features to establish L^2 convergence rates for linear TD in both discounted and average-reward settings.

8. Conclusion

This work contributes to RL with arbitrary features using linear TD as an example, where the commonly used linear independence assumption on features is lifted. The insight and techniques in this work can be easily used to analyze other linear RL algorithms, e.g., linear SARSA (Rummery and Niranjan, 1994; Zou et al., 2019; Zhang et al., 2023), gradient TD methods (Sutton et al., 2008, 2009; Maei, 2011; Zhang et al., 2021a; Qian and Zhang, 2025), emphatic TD methods (Yu, 2015; Sutton et al., 2016; Zhang and Whiteson, 2022), density ratio learning methods (Nachum et al., 2019; Zhang et al., 2020a), TD with target networks (Lee and He, 2019; Carvalho et al., 2020; Zhang et al., 2021b), linear Q-learning (Meyn, 2024; Liu et al., 2025c,b), and actor-critic methods with compatible features (Sutton et al., 1999; Konda and Tsitsiklis, 1999; Zhang et al., 2020b), as well as their in-context

[14] Reversibility measures how symmetric the matrix $D(I - \gamma P)$ is.

learning version (Wang et al., 2025a,b). The mode of convergence may also be extended to high probability convergence with exponential tails (and thus L^p convergence) following the techniques in Chen et al. (2025); Qian et al. (2024). This work is also closely related to overparameterized neural networks, where the linearization of the neural network at the initial weights naturally results in features that are not necessarily linearly independent (Cai et al., 2024). Recently, the convergence of linear TD with linearly independent features has been formally verified in Lean (Zhang, 2025). We envision that the convergence of linear TD with arbitrary features can also be formally verified in Lean based on the results in this paper.

That said, the major open question is whether we can prove that the linear TD weight iterates converge to a possibly sample-path-dependent TD fixed point. We are optimistic about this question because recent work Blaser and Zhang (2026) show that tabular average-reward TD can converge to a sample-path-dependent TD fixed point almost surely. We are, however, pessimistic about solving this problem with the ODE technique. Instead, recent works on stochastic Krasnoselskii-Mann iterations based on a fox-and-hare model (Cominetti et al., 2014; Bravo et al., 2019; Bravo and Cominetti, 2024; Blaser and Zhang, 2026) may be a promising direction to solve this problem.

Acknowledgments

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, and a Cisco Faculty Research Award.

Appendix A. Mathematical Background

We first provide the definition of the Moore-Penrose pseudo-inverse for completeness.

Definition A.1 *(Definition 23.1 of Gallier and Quaintance (2019)) Given any nonzero $m \times n$ matrix A of rank r, if $A = V\Sigma U^\top$ is a singular value decomposition of A such that $\Sigma = \begin{bmatrix} \Lambda & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{bmatrix}$, where $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}$ is an $r \times r$ diagonal matrix consisting of the nonzero singular values of A, then if we let $\Sigma^\dagger$ be the $n \times m$ matrix $\Sigma^\dagger = \begin{bmatrix} \Lambda^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{bmatrix}$, where $\Lambda^{-1} = \begin{bmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_r \end{bmatrix}$, the pseudo-inverse of A is defined as $A^\dagger = U\Sigma^\dagger V^\top$. If $A = 0_{m,n}$ is the zero matrix, the pseudo-inverse of A is defined as $A^\dagger = 0_{n,m}$.*

Theorem A.1 *(Theorems 23.1 & 23.2 of Gallier and Quaintance (2019)) For any matrix A and any vector b, consider the least square problem $\min_x \|Ax - b\|$. Then, $x^\dagger \doteq A^\dagger b$ is a minimizer. Furthermore, any other possible minimizer has a strictly larger norm than $x^\dagger$, i.e., $x^\dagger$ is the unique minimizer of the problem $\min\{\|x\| \| \|Ax - b\| = \min_y \|Ay - b\|\}$.*

Lemma A.1 *For any matrix $A \in \mathbb{R}^{m\times n}$ of rank r, $\left\|AA^\dagger\right\| \leq 1$.*

Proof This follows immediately from the fact that $AA^\dagger$ is an orthogonal projection onto the range of A (see, e.g., Proposition 23.4 of Gallier and Quaintance (2019)) and the well-known fact that the operator norm of an orthogonal projection is 1. We also provide a short proof for completeness. Let $A = V\Sigma U^\top$ be the singular value decomposition. We then have $\left\|AA^\dagger\right\| = \left\|V\Sigma U^\top(U\Sigma^\dagger V^\top)\right\| = \left\|VQV^\top\right\|$, where $Q = \begin{bmatrix} I_r & 0_{r,m-r} \\ 0_{m-r,r} & 0_{m-r,m-r}\end{bmatrix}$. It then follows easily that $\left\|AA^\dagger\right\| \leq \|V\|\|Q\|\left\|V^\top\right\| \leq 1$. ■

We then list supporting theorems from Khalil (2002) about nonlinear systems.

Theorem A.2 *(Theorem 4.2 of Khalil (2002)) Let $z = z_*$ be an equilibrium point for (6). Let $U : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function such that*

1. *$U(z_*) = 0$ and $U(z) > 0, \forall z \neq z_*$;*
2. *$\|z\| \to \infty \implies U(z) \to \infty$;*
3. *$\frac{\mathrm{d}U(z)}{\mathrm{d}t} < 0, \forall z \neq z_*$,*

then $z = z_*$ is globally asymptotically stable.

A.1 Stochastic Approximation

We then present some general results in stochastic approximation from Benveniste et al. (1990) and Liu et al. (2025a). Consider the iterates $\{w_t\}$ in $\mathbb{R}^d$ generated by

$$w_{t+1} = w_t + \alpha_t H(w_t, Y_t), \tag{23}$$

where $\{Y_t\}$ is a Markov chain evolving in $\mathbb{R}^k$ and the function H maps from $\mathbb{R}^d \times \mathbb{R}^k$ to $\mathbb{R}^d$. We use $P_{\mathcal{Y}} : \mathcal{B}(\mathbb{R}^k) \times \mathbb{R}^k \to [0,1]$ to denote the transition kernel of $\{Y_t\}$.

We first present a result from Benveniste et al. (1990) concerning the stability of $\{w_t\}$. We note that Benveniste et al. (1990) considers a time-inhomogeneous Markov chain, but our $\{Y_t\}$ is time-homogeneous. Therefore, some assumptions in Benveniste et al. (1990) trivially hold in our setting. For simplicity, we list only the nontrivial assumptions here.

Assumption A.1 *$\{\alpha_t\}$ is a decreasing sequence of positive real numbers such that*

$$\textstyle\sum_{t=0}^\infty \alpha_t = \infty, \quad \sum_{t=0}^\infty \alpha_t^2 < \infty, \quad \lim_{t\to\infty}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right) < \infty.$$

Assumption A.2 *The Markov chain $\{Y_t\}$ admits a well-defined and unique stationary distribution $\eta : \mathcal{B}(\mathbb{R}^k) \to [0,1]$, such that $\eta(Z) = \int_{\mathbb{R}^k} P_{\mathcal{Y}}(Z \mid y)\eta(\mathrm{d}y)$ for all $Z \in \mathcal{B}(\mathbb{R}^k)$ and $\eta(U) > 0$ for all non-empty open sets $U \subseteq \mathbb{R}^k$.*

Assumption A.3 *There exists a constant $K_{A.3} \in \mathbb{R}$, such that for all $w \in \mathbb{R}^d$ and $y \in \mathbb{R}^k$, we have $\|H(w,y)\| \leq K_{A.3}(1 + \|w\|)(1 + \|y\|)$.*

For any function $f(w,y)$ on $\mathbb{R}^d \times \mathbb{R}^k$, we shall denote the partial mapping $y \to f(w,y)$ by f_w. We shall also denote the function $y \to \int_{\mathbb{R}^k} f(w,y')P_{\mathcal{Y}}(\mathrm{d}y' \mid y)$ as $P_{\mathcal{Y}} f_w$.

Assumption A.4 *There exists a function $g : \mathbb{R}^d \to \mathbb{R}^d$, and for each $w \in \mathbb{R}^d$ a function $\nu_w : \mathbb{R}^k \to \mathbb{R}^d$, such that*

(i) g is locally Lipschitz on $\mathbb{R}^d$;

(ii) $\nu_w - P_{\mathcal{Y}}\nu_w = H_w - g(w)$ for all $w \in \mathbb{R}^d$;

(iii) there exists a constant $K_{A.4} \in \mathbb{R}$, such that for all $w \in \mathbb{R}^d$ and $y \in \mathbb{R}^k$, $\|\nu_w(y)\| \leq K_{A.4}(1 + \|w\|)(1 + \|y\|)$;

(iv) there exists a constant $K'_{A.4} \in \mathbb{R}$, such that for all $w, w' \in \mathbb{R}^d$ and $y \in \mathbb{R}^k$, $\|(P_{\mathcal{Y}}\nu_w)(y) - (P_{\mathcal{Y}}\nu_{w'})(y)\| \leq K'_{A.4}\|w - w'\|(1 + \|y\|)$.

Assumption A.5 *For all $a \in \mathbb{R}^d$, $q > 0$, and $n \geq 0$, there exists $K_{A.5}(q) < \infty$, such that $\mathbb{E}[1 + \|Y_{n+1}\|^q \mid w_0 = a, Y_0 = y] \leq K_{A.5}(q)(1 + \|y\|^q)$.*

Define $h(w) \doteq \mathbb{E}_{Y\sim\eta}[H(w, Y)]$. Then, we have the following theorem in Benveniste et al. (1990).

Theorem A.3 *(Theorem 17(a) of Benveniste et al. (1990)) Let Assumptions A.1 and A.3 - A.5 hold. Assume there exists a function $U : \mathbb{R}^d \to [0, \infty)$ of class C^2 with bounded second derivatives[15] and a constant $c > 0$, such that for all $w \in \mathbb{R}^d$,*

i. $\langle \nabla_w U(w), h(w)\rangle \leq 0$, and

ii. $U(w) \geq c\|w\|^2$,

then for all $w_0 \in \mathbb{R}^d$, the iterates $\{w_t\}$ generated by (23) *is stable, i.e., $\sup_t \|w_t\| < \infty$ a.s.*

We now present results from Liu et al. (2025a) concerning the convergence of $\{w_t\}$.

Assumption A.6 *There exists a function $L : \mathcal{Y} \to \mathbb{R}$ such that for any w, w', y, $\|H(w, y) - H(w', y)\| \leq L(y)\|w - w'\|$. Moreover, the following expectations are well-defined and finite for any $w \in \mathbb{R}^d$: $h(w) \doteq \mathbb{E}_{Y\sim\eta}[H(w, Y)], L \doteq \mathbb{E}_{Y\sim\eta}[L(Y)]$.*

Assumption A.7 *For all $w_0 \in \mathbb{R}^d$, the iterates $\{w_t\}$ generated by* (23) *is stable, i.e., $\sup_t \|w_t\| < \infty$ a.s.*

Theorem A.4 *(Corollary 8 of Liu et al. (2025a)) Let Assumptions A.1, A.2, A.6 and A.7 hold. Then the iterates $\{w_t\}$ generated by* (23) *converge almost surely to a (possibly sample path dependent) bounded invariant set of the ODE $\frac{\mathrm{d}w(t)}{\mathrm{d}t} = h(w(t))$.*

We note that the original form of Corollary 8 of Liu et al. (2025a) involves some additional assumptions regarding $H_\infty(w, y) \doteq \lim_{c\to\infty} \frac{H(cw,y)}{c}$ and $h_\infty(w) \doteq \mathbb{E}_{Y\sim\eta}[H_\infty(w, Y)]$, as well as assumptions regarding the ODE $\frac{\mathrm{d}w(t)}{\mathrm{d}t} = h_\infty(w(t))$. Those assumptions are related to the ODE@∞ technique to establish the stability of the iterates $\{w_t\}$. We refer the reader to Borkar (2009); Borkar et al. (2025); Liu et al. (2025a) for more details about this technique. After establishing stability, the convergence part in Corollary 8 of Liu et al. (2025a) follows

[15] A function of class C^2 is a function whose second derivative is continuous in its domain.

a standard approach based on the Arzela-Ascoli theorem, akin to Theorem 1 of Chapter 5 of Kushner and Yin (2003) and does not rely on those additional assumptions. In this paper, we instead assume stability directly (Assumption A.7). As a result, we no longer need those ODE@∞ related assumptions and thus omit them. In this work, instead of using the ODE@∞ technique, we will use the Lyapunov method in Benveniste et al. (1990) to establish stability.

Appendix B. Proofs in Section 2

B.1 Proof of Lemma 1

Proof By definition, we have

$$\begin{aligned}\|P_\pi v\|_\mu^2 &= \int_{s\in\mathcal{S}} ((P_\pi v)(s))^2 \mu(\mathrm{d}s) = \int_{s\in\mathcal{S}} \left(\int_{s'\in\mathcal{S}} v(s')p(\mathrm{d}s' \,|s)\right)^2 \mu(\mathrm{d}s) \\ &\leq \int_{s\in\mathcal{S}} \int_{s'\in\mathcal{S}} v(s')^2 p(\mathrm{d}s' \,|s)\mu(\mathrm{d}s) && \text{(Jensen's inequality)} \\ &= \int_{s'\in\mathcal{S}} v(s')^2 \int_{s\in\mathcal{S}} p(\mathrm{d}s' \,|s)\mu(\mathrm{d}s) = \int_{s'\in\mathcal{S}} v(s')^2 \mu(\mathrm{d}s') = \|v\|_\mu^2.\end{aligned}$$

■

B.2 Proof of Lemma 2

Proof We prove the negative semi-definiteness of A by showing $\langle w, Aw\rangle \leq 0$ for any $w \in \mathbb{R}^d \setminus \{0\}$. Firstly, we define $v_w(s) \doteq x(s)^\top w$ for all $s \in \mathcal{S}$. In addition, $v_w \in L_2(\mathcal{S}, \mu)$ by Assumption 2.3. Then, we have

$$\begin{aligned}\langle w, Aw\rangle =& \langle v_w, \gamma P_\pi v_w - v_w\rangle_\mu && \text{(By (1))} \\ =& \langle v_w, \gamma P_\pi v_w\rangle_\mu - \|v_w\|_\mu^2 \leq \gamma \|v_w\|_\mu \|P_\pi v_w\|_\mu - \|v_w\|_\mu^2 \\ \leq& (\gamma - 1)\|v_w\|_\mu^2 \leq 0. && \text{(Lemma 1)}\end{aligned}$$

■

Appendix C. Proofs in Section 3

C.1 Proof of Lemma 3

Proof We first note that $v_w \in L_2(\mathcal{S}, \mu)$ because $\|x\|$ is bounded under Assumption 2.3. Then, we have $w^\top Aw = \langle v_w, \gamma P_\pi v_w - v_w\rangle_\mu$. We begin with the direction $w^\top Aw = 0 \impliedby v_w = 0$ a.e. When $v_w = 0$ a.e., we have $w^\top Aw = \langle v_w, \gamma P_\pi v_w - v_w\rangle_\mu = 0$ trivially holds. We then prove $w^\top Aw = 0 \implies v_w = 0$ a.e. via contradiction. First, we have $\langle v_w, \gamma P_\pi v_w - v_w\rangle_\mu = 0$ because $w^\top Aw = 0$. Now, suppose $\|v_w\|_\mu > 0$. We have

$$\begin{aligned}\langle v_w, \gamma P_\pi v_w - v_w\rangle_\mu =& \langle v_w, \gamma P_\pi v_w\rangle_\mu - \|v_w\|_\mu^2 \leq \gamma \|v_w\|_\mu \|P_\pi v_w\|_\mu - \|v_w\|_\mu^2 \\ \leq& (\gamma - 1)\|v_w\|_\mu^2 && \text{(Lemma 1)}\end{aligned}$$

$$<0. \qquad (\|v_w\|_\mu > 0)$$

This generates a contradiction and completes the proof. Thus, we must have $\|v_w\|_\mu = 0$, which holds if and only if $v_w = 0$ a.e. ■

C.2 Proof of Lemma 4

Proof Recall that we can represent the vector-valued feature mapping x as $x(s) = \begin{bmatrix} x_1(s) & x_2(s) & \cdots & x_d(s) \end{bmatrix}^\top$, where each $x_i : \mathcal{S} \to \mathbb{R}$ is a basis function for $i = 1, 2, \ldots, d$. Without further assumptions on x, the basis functions can be linearly dependent. Without loss of generality, suppose the first m basis functions form the largest linearly independent collection of x. When $m = 0$, it implies $x = 0$ a.e., and we end up with a degenerate case, where $A = 0$ and $b = 0$. In this case, $\mathcal{W}_* = \mathbb{R}^d$ and clearly is non-empty. In the rest of the proof, we analyze the case where $m \geq 1$. We denote $\phi(s) \doteq \begin{bmatrix} x_1(s) & x_2(s) & \cdots & x_m(s) \end{bmatrix}^\top \in \mathbb{R}^m$. Then, for each x_k, where $k \in \{m+1, \ldots, d\}$, there exists a vector of coefficients $c \doteq \begin{bmatrix} c_1 & c_2 \cdots & c_m \end{bmatrix}$, such that $x_k = \sum_{i=1}^m c_i x_i$ a.e. Therefore, we can define $\hat{x}(s) \doteq \begin{bmatrix} \phi(s) \\ C\phi(s) \end{bmatrix}$, where $C \in \mathbb{R}^{(d-m)\times m}$ is a coefficient matrix, such that $\hat{x} = x$ a.e. Define $\hat{A} \doteq \int_\mathcal{S} \hat{x}(s)\big(\gamma(P_\pi \hat{x})(s)^\top - \hat{x}(s)^\top\big)\mu(\mathrm{d}s)$ and $\hat{b} \doteq \int_\mathcal{S} \hat{x}(s) r(s)\mu(\mathrm{d}s)$. We have

$$
\begin{aligned}
\hat{A} - A &= \textstyle\int_\mathcal{S} (\hat{x}(s) - x(s) + x(s))\big(\gamma(P_\pi \hat{x})(s)^\top - \hat{x}(s)^\top\big)\mu(\mathrm{d}s) - A \\
&= \textstyle\int_\mathcal{S} (\hat{x}(s) - x(s))\big(\gamma(P_\pi \hat{x})(s)^\top - \hat{x}(s)^\top\big)\mu(\mathrm{d}s) \\
&\quad + x(s)\Big(\gamma(P_\pi \hat{x})(s)^\top - \hat{x}(s)^\top\Big)\mu(\mathrm{d}s) - A \\
&= \textstyle\int_\mathcal{S} \gamma x(s)(P_\pi \hat{x})(s)^\top \mu(\mathrm{d}s) - \int_\mathcal{S} x(s)\hat{x}(s)^\top \mu(\mathrm{d}s) - A && (\hat{x} = x \text{ a.e.}) \\
&= \textstyle\int_\mathcal{S} \gamma x(s)(P_\pi \hat{x})(s)^\top \mu(\mathrm{d}s) - \int_\mathcal{S} x(s)\hat{x}(s)^\top \mu(\mathrm{d}s) \\
&\quad - \big(\textstyle\int_\mathcal{S} \gamma x(s)(P_\pi x)(s)^\top \mu(\mathrm{d}s) - \int_\mathcal{S} x(s)x(s)^\top \mu(\mathrm{d}s)\big) \\
&= \textstyle\int_\mathcal{S} \gamma x(s)(P_\pi(\hat{x} - x))(s)^\top \mu(\mathrm{d}s) - \int_\mathcal{S} x(s)(\hat{x} - x)(s)^\top \mu(\mathrm{d}s) \\
&= \textstyle\int_\mathcal{S} \gamma x(s)(P_\pi(\hat{x} - x))(s)^\top \mu(\mathrm{d}s) && (\hat{x} - x = 0 \text{ a.e.})
\end{aligned}
$$

By Assumption 2.4, it holds that $P_\pi(\hat{x} - x) = 0$ a.e. Consequently, we have $\hat{A} - A = 0$. Similarly, we have $\hat{b} - b = \int_\mathcal{S} (\hat{x} - x)(s) r(s)\mu(\mathrm{d}s) = 0$. Thus, it holds that $\hat{A} = A$ and $\hat{b} = b$.

We now prove that $\mathcal{W}_*$ is non-empty. Let $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$, where $w_1 \in \mathbb{R}^m$ and $w_2 \in \mathbb{R}^{d-m}$. We have $Aw + b = \int_\mathcal{S} \begin{bmatrix} \phi(s) \\ C\phi(s) \end{bmatrix} \begin{bmatrix} \gamma(P_\pi \phi)(s) - \phi(s) \\ C(\gamma(P_\pi \phi)(s) - \phi(s)) \end{bmatrix}^\top \mu(\mathrm{d}s) \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \int_\mathcal{S} \begin{bmatrix} \phi(s) \\ C\phi(s) \end{bmatrix} r_\pi(s)\mu(\mathrm{d}s)$. Define $\delta(s) \doteq \gamma(P_\pi \phi)(s) - \phi(s)$. We then have

$$
\begin{aligned}
Aw + b &= \begin{bmatrix} \int_\mathcal{S} \phi(s)\delta(s)^\top \mu(\mathrm{d}s) & \int_\mathcal{S} \phi(s)\delta(s)^\top \mu(\mathrm{d}s) C^\top \\ C\int_\mathcal{S} \phi(s)\delta(s)^\top \mu(\mathrm{d}s) & C\int_\mathcal{S} \phi(s)\delta(s)^\top \mu(\mathrm{d}s) C^\top \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} \int_\mathcal{S} \phi(s) r_\pi(s)\mu(\mathrm{d}s) \\ C\int_\mathcal{S} \phi(s) r_\pi(s)\mu(\mathrm{d}s) \end{bmatrix} \\
&= \begin{bmatrix} A' & A'C^\top \\ CA' & CA'C^\top \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} b' \\ Cb' \end{bmatrix},
\end{aligned}
$$

where we use shorthands $A' \doteq \int_{\mathcal{S}} \phi(s)\delta(s)^\top \mu(\mathrm{d}s)$ and $b' \doteq \int_{\mathcal{S}} \phi(s) r_\pi(s)\mu(\mathrm{d}s)$. Thus, it leaves us with two linear equations

$$\begin{cases} A'w_1 + A'C^\top w_2 + b' & = 0 \\ CA'w_1 + CA'C^\top w_2 + Cb' & = 0 \end{cases} \tag{24}$$

which we can solve to obtain some w_1, w_2, such that $A\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + b = 0$. Let $\ker(M)$ denote the kernel space of a matrix M. We choose $w_2 \in \ker(A'C^\top)$, such that (24) reduces to

$$\begin{cases} A'w_1 & = -b' \\ CA'w_1 & = -Cb'. \end{cases} \tag{25}$$

We then observe that $w_1 = -A'^{-1}b'$ is a valid solution to the first equation of (25) provided A' has full rank. Furthermore, this w_1 would also satisfy the second equation of (25), thus solving the system of linear equations. We now prove that A' is indeed invertible by showing it is negative definite. The proof mirrors the proof of Lemma 2. We first define $v'_w(s) \doteq \phi(s)^\top w$ for all $s \in \mathcal{S}$ and any $w \in \mathbb{R}^m \setminus \{0\}$, which is square-integrable under Assumption 2.3. Recall that the basis functions $x_1, x_2, \ldots, x_m$ that constitute ϕ are linearly independent. Hence, it is impossible that $v'_w = 0$ a.e. We then have

$$\begin{aligned} \langle w, A'w \rangle =& \langle v'_w, \gamma P_\pi v'_w - v'_w \rangle_\mu = \langle v'_w, \gamma P_\pi v'_w \rangle_\mu - \|v'_w\|_\mu^2 \leq \gamma \|v'_w\|_\mu \|P_\pi v'_w\|_\mu - \|v'_w\|_\mu^2 \\ \leq& (\gamma - 1)\|v'_w\|_\mu^2 < 0. \end{aligned} \tag{Lemma 1}$$

The matrix A' is negative definite and thus invertible. As a result, we have constructed a solution set $\left\{ w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \middle| w_1 = -A'^{-1}b'; w_2 \in \ker(A'C^\top) \right\} \subseteq \mathcal{W}_*$ and proved $\mathcal{W}_*$ is non-empty when $m \geq 1$. ■

C.3 Proof of Lemma 5

Proof We first recall that $v_{w-w'} \doteq x(s)^\top (w - w')$. We begin with $w' \in \mathcal{W}_* \implies v_w = v_{w'}$ a.e. Suppose that $w' \in \mathcal{W}_*$. We have

$$\begin{aligned} Aw + b - (Aw' + b) =& 0 \\ A(w - w') =& 0 \\ (w - w')^\top A(w - w') =& 0. \end{aligned}$$

By Lemma 3, we have $v_{w-w'} = 0$ a.e., implying $v_w = v_{w'}$ a.e.

We now show $v_w = v_{w'}$ a.e. $\implies w' \in \mathcal{W}_*$. Suppose $v_w = v_{w'}$ a.e. holds. We then have $v_{w-w'} = 0$ a.e. Define $|v_{w-w'}|(s) \doteq \left|x(s)^\top (w - w')\right|$. Then, it holds trivially that $|v_{w-w'}| = 0$ a.e. We have

$$\|A(w - w')\| = \left\| \int_{\mathcal{S}} x(s) \big(\gamma (P_\pi x)(s)^\top - x(s)^\top\big) \mu(\mathrm{d}s)(w - w') \right\|$$

$$
\begin{aligned}
=&\left\| \int_{\mathcal{S}} x(s)(\gamma(P_\pi v_{w-w'})(s) - v_{w-w'}(s))\mu(\mathrm{d}s) \right\| \\
\leq& \int_{\mathcal{S}} \| x(s)(\gamma(P_\pi v_{w-w'})(s) - v_{w-w'}(s)) \| \mu(\mathrm{d}s) && \text{(Jensen's inequality)} \\
=& \int_{\mathcal{S}} \| x(s) \| |(\gamma(P_\pi v_{w-w'})(s) - v_{w-w'}(s))| \mu(\mathrm{d}s) \\
\leq& \int_{\mathcal{S}} \| x(s) \| \gamma |(P_\pi v_{w-w'})(s)| \mu(\mathrm{d}s) + \int_{\mathcal{S}} \| x(s) \| |v_{w-w'}|(s) \mu(\mathrm{d}s) \\
\leq& C_x \int_{\mathcal{S}} \gamma |(P_\pi v_{w-w'})(s)| \mu(\mathrm{d}s) + C_x \int_{\mathcal{S}} |v_{w-w'}|(s) \mu(\mathrm{d}s) \\
& && (C_x \doteq \sup_{s \in \mathcal{S}} \| x(s) \| < \infty \text{ by Assumption 2.3}) \\
=& C_x \int_{\mathcal{S}} \gamma |(P_\pi v_{w-w'})(s)| \mu(\mathrm{d}s) && (|v_{w-w'}| = 0 \text{ a.e. w.r.t. } \mu\ ) \\
=& 0. && \text{(Assumption 2.4)}
\end{aligned}
$$

The result suggests $A(w - w') = 0$, which implies $Aw = Aw'$. We therefore have $Aw' + b = Aw + b = 0$ and have proved that $w' \in \mathcal{W}_*$. ■

Appendix D. Proofs in Section 4

Lemma D.1 *Under Assumptions 2.2 and 2.3, it holds that* $\sup_{t \in [0, \infty)} \| w(t; w_0) \| < \infty$ *for all* $w_0 \in \mathbb{R}^d$.

Proof Fix an arbitrary $w_* \in \mathcal{W}_*$. We first show that $\frac{\mathrm{d} \| w(t; w_0) - w_* \|^2}{\mathrm{d}t} \leq 0$ for any $w_0 \in \mathbb{R}^d$.

$$
\begin{aligned}
\frac{\mathrm{d} \| w(t; w_0) - w_* \|^2}{\mathrm{d}t} &= 2(w(t; w_0) - w_*)^\top (Aw(t; w_0) + b) \\
&= 2(w(t; w_0) - w_*)^\top (Aw(t; w_0) + b - (Aw_* + b)) && (Aw_* + b = 0) \\
&= 2(w(t; w_0) - w_*)^\top A(w(t; w_0) - w_*) \leq 0. && \text{(Lemma 2)}
\end{aligned}
$$

Thus, $\| w(t; w_0) - w_* \|$ is monotonically decreasing given any $w_0 \in \mathbb{R}^d$. Hence, $\| w(t; w_0) - w_* \| \leq \| w_0 - w_* \|$ for $t \geq 0$, and by triangle inequality we get $\sup_{t \in [0, \infty)} \| w(t; w_0) \| \leq \| w_* \| + \| w_0 - w_* \|$, which completes the proof. ■

D.1 Proof of Lemma 9

Proof The absolute homogeneity of a matrix norm implies that

$$
\begin{aligned}
&\lim_{t \to \infty} \left\| \exp(\mathrm{Re}(\lambda_i) t)(\cos(\mathrm{Im}(\lambda_i) t) + \mathtt{i} \sin(\mathrm{Im}(\lambda_i) t)) \textstyle\sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n \right\| \\
=& \lim_{t \to \infty} \exp(\mathrm{Re}(\lambda_i) t) |\cos(\mathrm{Im}(\lambda_i) t) + \mathtt{i} \sin(\mathrm{Im}(\lambda_i) t)| \left\| \textstyle\sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n \right\| \\
=& \lim_{t \to \infty} \exp(\mathrm{Re}(\lambda_i) t) \left\| \textstyle\sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n \right\|.
\end{aligned}
$$

It appears that $\left\| \sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n \right\| = \mathcal{O}(t^{y_{i,j}-1})$. When $\mathrm{Re}(\lambda_i) < 0$, we have $\lim_{t \to \infty} \exp(\mathrm{Re}(\lambda_i) t) \left\| \sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n \right\| = 0$ because $\exp(\mathrm{Re}(\lambda_i) t)$ decays exponentially, whereas $\left\| \sum_{n=0}^{y_{i,j}-1} \frac{1}{n!} t^n N_{i,j}^n \right\|$ exhibits polynomial growth. ■

D.2 Proof of Lemma 10

Proof We shall prove this claim by propagating the boundedness of $w(t;w_0)$ through different levels, starting from $z(t;z_0)$. For any $z_0 \in \mathbb{R}^d, w_* \in \mathcal{W}_*$, pick $w_0 = z_0 + w_*$. From (7), we obtain

$$\begin{aligned}\sup_{t\in[0,\infty)} \|z(t;z_0)\| &= \sup_{t\in[0,\infty)} \|w(t;z_0+w_*) - w_*\| \\ &\leq \sup_{t\in[0,\infty)} \|w(t;z_0+w_*)\| + \|w_*\| < \infty. \qquad \text{(Lemma D.1)}\end{aligned}$$

Then, we prove that $\exp(At)$ is bounded. Let e_n be the n-th column of I_d, the identity matrix in $\mathbb{R}^{d\times d}$. By the boundedness of $z(t;z_0)$ and (13), we have for all $n \in \{1,2,\ldots,d\}$, $\sup_{t\in[0,\infty)} \|\exp(At)e_n\| = \sup_{t\in[0,\infty)} \|z(t;e_n)\| < \infty$. By the definition of the induced norm and the invariance of ℓ_2 norm under matrix transpose, we obtain $\|\exp(At)\| \leq \sqrt{d}\max_n \|\exp(At)e_n\| = \sqrt{d}\max_n \|z(t;e_n)\|$. Since n is finite, we have $\sup_{t\in[0,\infty)} \|\exp(At)I_d\| \leq \sup_{t\in[0,\infty)} \sqrt{d}\max_n \|z(t;e_n)\| < \infty$. It then follows from the Jordan decomposition (14) that $\sup_{t\in[0,\infty)} \|\exp(Jt)\| = \sup_{t\in[0,\infty)} \left\|P^{-1}\exp(At)P\right\| \leq \sup_{t\in[0,\infty)} \|\exp(At)\|\left\|P^{-1}\right\|\|P\| < \infty$. Since $\exp(Jt)$ is a block diagonal matrix, we have $\max_{i,j} \|\exp(B_{i,j}t)\| = \|\exp(Jt)\| < \infty$, which completes the proof. ■

D.3 Proof of Corollary 11

Proof For any $z_0 \in \mathbb{R}^d, w_* \in \mathcal{W}_*$, pick $w_0 = z_0 + w_*$. From (7), we obtain $z(t;z_0) = w(t;z_0+w_*) - w_*$. By Theorem 7, we have $\lim_{t\to\infty} v_{w(t;z_0+w_*)} = v_{w_*}$ a.e. Thus, we have $\lim_{t\to\infty} v_{z(t;z_0)} = \lim_{t\to\infty} v_{w(t;z_0+w_*)} - v_{w_*} = 0$ a.e. Then, according to Corollary 6, it holds that $\lim_{t\to\infty} d(z(t;z_0),\mathcal{Z}_*) = 0$. Since $\frac{\mathrm{d}z(t;z_0)}{\mathrm{d}t} = Az(t;z_0)$, we have $\lim_{t\to\infty} \frac{\mathrm{d}z(t;z_0)}{\mathrm{d}t} = 0$ by the definition of $\mathcal{Z}_*$ and the fact that $z(t;z_0)$ converges to $\mathcal{Z}_*$. ■

D.4 Proof of Lemma 12

Proof Given $\mathrm{Re}(\lambda_i) = 0$, we have
$\exp(B_{i,j}t) = (\cos(\mathrm{Im}(\lambda_i)t) + \mathtt{i}\sin(\mathrm{Im}(\lambda_i)t)) \sum_{n=0}^{y_{i,j}-1} \frac{1}{n!}t^n N_{i,j}^n$. We first prove $y_{i,j} = 1$ by contradiction. Suppose $y_{i,j} > 1$. If $\mathrm{Im}(\lambda_i) = 0$, we then have

$$\lim_{t\to\infty} \|\exp(B_{i,j}t)\| = \lim_{t\to\infty} \left\|\sum_{n=0}^{y_{i,j}-1} \frac{1}{n!}t^n N_{i,j}^n\right\| = \infty,$$

yielding a contradiction with Lemma 10. If $\mathrm{Im}(\lambda_i) \neq 0$, then $\exp(B_{i,j}t)$ oscillates with t. Consider the sequence $\left\{t_k : t_k = \frac{2k\pi}{|\mathrm{Im}(\lambda_i)|}, k \geq 0\right\}$. We have

$$\lim_{k\to\infty} \|\exp(B_{i,j}t_k)\| = \lim_{k\to\infty} \left\|\sum_{n=0}^{y_{i,j}-1} \frac{1}{n!}t_k^n N_{i,j}^n\right\| = \infty,$$

again leading to a contradiction with Lemma 10. We then conclude that $y_{i,j} = 1$ must hold, and we are left with $\exp(B_{i,j}t) = (\cos(\mathrm{Im}(\lambda_i)t) + \mathtt{i}\sin(\mathrm{Im}(\lambda_i)t))I_{\rho_{i,j}}$. We now prove $\mathrm{Im}(\lambda_i) = 0$ by contradiction. Assume that $\mathrm{Im}(\lambda_i) \neq 0$ holds. Then,

$$\frac{\mathrm{d}\exp(B_{i,j}t)}{\mathrm{d}t} = \mathrm{Im}(\lambda_i)(\mathtt{i}\cos(\mathrm{Im}(\lambda_i)t) - \sin(\mathrm{Im}(\lambda_i)t))I_{\rho_{i,j}} \qquad (26)$$

and $\frac{\mathrm{d}\exp(B_{i,j}t)}{\mathrm{d}t}$ would oscillate and never converge. On the other hand, we have $\frac{\mathrm{d}z(t;z_0)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\exp(At)z_0 = P\frac{\mathrm{d}}{\mathrm{d}t}\exp(Jt)P^{-1}z_0$. We have showed in Corollary 11 that $\lim_{t\to\infty}\frac{\mathrm{d}z(t;z_0)}{\mathrm{d}t} = 0$ for all $z_0 \in \mathbb{R}^d$. Hence, by setting $z_0 = e_1, e_2, \dots, e_d$, where e_n is the n-th column of I_d, it holds that

$$\begin{aligned} \lim\nolimits_{t\to\infty} P\tfrac{\mathrm{d}}{\mathrm{d}t}\exp(Jt)P^{-1}I_d &= 0 \\ P^{-1}\lim\nolimits_{t\to\infty} P\tfrac{\mathrm{d}}{\mathrm{d}t}\exp(Jt)P^{-1}P &= 0 \\ \lim\nolimits_{t\to\infty}\tfrac{\mathrm{d}}{\mathrm{d}t}\exp(Jt) &= 0, \end{aligned}$$

implying that $\lim_{t\to\infty}\frac{\mathrm{d}}{\mathrm{d}t}\exp(B_{i,j}t) = 0$, yielding a contradiction with (26). We then conclude that $\mathrm{Im}(\lambda_i) = 0$. We are finally left with $\forall t \geq 0$, $\exp(B_{i,j}t) = I_{\rho_{i,j}}$, which completes the proof. ■

D.5 Proof of Corollary 14

Proof Let $\mathcal{W}$ be a bounded invariant set of ODE (3). By definition, for every $w_0 \in \mathcal{W}$, the solution $w(t; w_0)$ to ODE (3) on the domain $(-\infty, \infty)$ would remain in $\mathcal{W}$. Furthermore, since $\mathcal{W}$ is also bounded, it implies that $w(t; w_0)$ is a bounded solution on $(-\infty, \infty)$. By Theorem 13, it holds that $w(t; w_0)$ is constant and in $\mathcal{W}_*$. Thus, it implies that $w_0 \in \mathcal{W}_*$, which completes the proof. ■

Appendix E. Proofs in Section 5

We recall for analyzing (Linear TD), we have $H(w, y) \doteq \big(r(s) + \gamma w^\top x(s') - w^\top x(s)\big)x(s)$, $h(w) \doteq \mathbb{E}_{y\sim\eta}[H(w, y)]$. Some elementary properties then follow.

Lemma E.1 *Let Assumption 2.3 hold. Then there exists some $K_{E.1} > 0$, such that for any $w, w' \in \mathbb{R}^d$ and $y \in \mathcal{Y}$, it holds that $\|H(w, y) - H(w', y)\| \leq K_{E.1}\|w - w'\|$.*

Proof Given $w, w' \in \mathbb{R}^d$, we have

$$\begin{aligned} \|H(w,y) - H(w',y)\| =& \big\|\big(\gamma(w-w')^\top x(s') - (w-w')^\top x(s)\big)x(s)\big\| \qquad (Y \doteq \begin{bmatrix} s & s' \end{bmatrix}^\top) \\ =& |\gamma\langle w-w', x(s')\rangle - \langle w-w', x(s)\rangle|\|x(s)\| \\ \leq& C_x(|\gamma\langle w-w', x(s')\rangle| + |\langle w-w', x(s)\rangle|) \\ & \qquad (C_x \doteq \sup\nolimits_{s\in\mathcal{S}}\|x(s)\| < \infty) \\ \leq& C_x(\gamma C_x\|w-w'\| + C_x\|w-w'\|) \\ =& (1+\gamma)C_x^2\|w-w'\|. \end{aligned}$$

Setting $K_{\mathrm{E.1}} = (1+\gamma)C_x^2$, the lemma is proved. ■

Lemma E.2 *Let Assumptions 2.2 and 2.3 hold. Then there exists some $K_{E.2} > 0$, such that for any $w, w' \in \mathbb{R}^d$, it holds that $\|h(w) - h(w')\| \leq K_{E.2}\|w - w'\|$.*

Proof

$$
\begin{aligned}
\|h(w)-h(w')\| =&\|\mathbb{E}_{Y\sim\eta}[H(w,Y)-H(w',Y)]\| \\
\leq&\mathbb{E}_{Y\sim\eta}[\|H(w,Y)-H(w',Y)\|] \leq \mathbb{E}_{Y\sim\eta}[K_{\text{E.1}}\|w-w'\|] = K_{\text{E.1}}\|w-w'\|.
\end{aligned}
$$

Hence, the inequality holds by setting $K_{\text{E.2}} = K_{\text{E.1}}$. ■

Lemma E.3 *Let Assumption 2.3 hold. Then there exists a constant $K_{E.3} > 0$, such that for all $y \in \mathcal{Y}$, it holds that $\|H(w,y)\| \leq K_{E.3}(1+\|w\|)$.*

Proof We have by Lemma E.1 $\|H(w,y)-H(w',y)\| \leq K_{\text{E.1}}\|w-w'\|$. Then, fixing an arbitrary $\tilde{w} \in \mathbb{R}^d$ with $\|\tilde{w}\| = 1$, we have

$$
\begin{aligned}
\|H(w,y)-H(\tilde{w},y)\| \leq& K_{\text{E.1}}\|w-\tilde{w}\| \\
\|H(w,y)\|-\|H(\tilde{w},y)\| \leq& K_{\text{E.1}}\|w-\tilde{w}\| \\
\|H(w,y)\| \leq& K_{\text{E.1}}\|w-\tilde{w}\| + \big\|\big(r(s)+\gamma\tilde{w}^\top x(s')-\tilde{w}^\top x(s)\big)x(s)\big\| \\
\leq& K_{\text{E.1}}(\|w\|+\|\tilde{w}\|) + C_x\big|r(s)+\gamma\tilde{w}^\top x(s')-\tilde{w}^\top x(s)\big| \\
& \qquad (C_x \doteq \sup_{s\in\mathcal{S}}\|x(s)\| < \infty) \\
\leq& K_{\text{E.1}}(\|w\|+1)C_x\big(|r(s)|+\gamma\big|\tilde{w}^\top x(s')\big|+\big|\tilde{w}^\top x(s)\big|\big) \\
\leq& K_{\text{E.1}}(\|w\|+1)C_x(C_r+\gamma\|\tilde{w}\|\|x(s')\|+\|\tilde{w}\|\|x(s)\|) \\
& \qquad (C_r \doteq \sup_{s\in\mathcal{S}}|r(s)| < \infty) \\
\leq& K_{\text{E.1}}(\|w\|+1)C_x(C_r+\gamma C_x+C_x) \\
=& K_{\text{E.1}}\big(C_xC_r+\gamma C_x^2+C_x^2\big)(\|w\|+1).
\end{aligned}
$$

Hence, setting $K_{\text{E.3}} \doteq K_{\text{E.1}}\big(C_xC_r+\gamma C_x^2+C_x^2\big)$, we get $\|H(w,y)\| \leq K_{\text{E.3}}(1+\|w\|)$, which completes the proof. ■

E.1 Proof of Theorem 17

Proof We shall show that Assumptions A.1 and A.3 – A.5 hold. Invoking Theorem A.3 will then complete the proof. First, Assumption A.1 is satisfied by Assumption 2.1. Assumption A.3 holds by Lemma E.3. We then proceed to show Assumption A.4 can be satisfied. Choosing $g = h$, where we recall that $h(w) \doteq \int_{\mathcal{Y}} H(w,y)\eta(\mathrm{d}y)$, Assumption A.4 (i) holds by Lemma E.2. Assumption 5.1 automatically satisfies (*ii*), (*iii*), and (*iv*) of Assumption A.4. Lastly, we prove Assumption A.5 holds. Since $\mathcal{S}$ is compact by Assumption 5.3, then $\mathcal{Y}$ is also compact. Thus, $\sup_{y\in\mathcal{Y}}\|y\|$ is finite. We denote this quantity as C_y. Then, for all $a \in \mathbb{R}^d$, $q > 0$ and $n \geq 0$, it holds that $\mathbb{E}[1+\|Y_{n+1}\|^q \mid w_0 = a, Y_0 = y] \leq 1 + C_y^q$. Thus, by setting $K_{\text{A.5}}(q) \doteq (1+C_y^q)$, Assumption A.5 is satisfied.

Having verified the assumptions, we now show the existence of the function U required by Theorem A.3. Fixing an arbitrary $w_* \in \mathcal{W}_*$, we define U as $U(w) \doteq \|w-w_*\|^2 + \|w_*\|^2$. The second derivative of U is $2I_d$. Therefore, U is of class C^2 with bounded second derivatives. We now show that it satisfies Condition (i) of Theorem A.3. Taking the gradient, we have

$$
\langle \nabla_w U(w), h(w)\rangle = \langle 2(w-w_*), Aw+b\rangle = 2\,\langle (w-w_*), Aw+b-(Aw_*+b)\rangle
$$

$$= 2\langle w - w_*, A(w - w_*)\rangle \le 0. \qquad \text{(Lemma 2)}$$

Condition (i) holds for our selected U. Next, we prove that Condition (ii) also holds by showing that $U(w) \ge \frac{1}{2}\|w\|^2$ for all $w \in \mathbb{R}^d$. We have

$$\begin{aligned} U(w) - \tfrac{1}{2}\|w\|^2 &= \|w - w_*\|^2 + \|w_*\|^2 - \tfrac{1}{2}\|w\|^2 = \tfrac{1}{2}\|w\|^2 - 2\langle w, w_*\rangle + 2\|w_*\|^2 \\ &= \tfrac{1}{2}\Big(\|w\|^2 - 4\langle w, w_*\rangle + 4\|w_*\|^2\Big) = \tfrac{1}{2}\|w - 2w_*\|^2 \ge 0. \end{aligned}$$

Hence, Condition (ii) holds by setting $c = \frac{1}{2}$. The proof is completed. ■

E.2 Proof of Theorem 18

Proof We apply Theorem A.4 to show that $\{w_t\}$ converges almost surely to a possibly sample path dependent bounded invariant set of ODE (3). To be able to do so, we need to verify Assumptions A.1, A.2, A.6 and A.7 hold. Assumptions A.1 and A.2 are satisfied by Assumptions 2.1 and 2.2, respectively. Regarding Assumption A.6, we have

$$\begin{aligned} \|H(w,y) - H(w',y)\| &= \|(\gamma\langle x(s'), w - w'\rangle - \langle x(s), w - w'\rangle)x(s)\| \\ &= |\langle \gamma x(s') - x(s), w - w'\rangle|\|x(s)\| \\ &\le C_x|\langle \gamma x(s') - x(s), w - w'\rangle| \qquad (C_x \doteq \sup\nolimits_{s\in\mathcal{S}}\|x(s)\| < \infty) \\ &\le C_x\|w - w'\|\|\gamma x(s') - x(s)\| \\ &\le C_x\|w - w'\|(\gamma\|x(s')\| + \|x(s)\|) \le C_x\|w - w'\|(C_x + \gamma C_x) \\ &= (1+\gamma)C_x^2\|w - w'\|. \end{aligned}$$

Thus, simply setting $L(y) = (1+\gamma)C_x^2$ satisfies Assumption A.6. Lastly, Assumption A.7 is satisfied by Theorem 17. As a result, we have $\{w_t\}$ converging almost surely to a bounded invariant set of ODE (3). In light of Corollary 14, any bounded invariant set is a subset of $\mathcal{W}_*$. Hence, $\{w_t\}$ converges almost surely to $\mathcal{W}_*$. Lemma 5 then completes the proof of Theorem 18. ■

E.3 Proof of Corollary 19

Proof For any sample path $\{w_0, w_1, \dots\}$, let $\bar{w} : \mathbb{R} \to \mathbb{R}^d$ be the piece-wise constant interpolation of $\{w_t\}$, i.e.,

$$\bar{w}(t) \doteq \begin{cases} w_{m(0,t)} & t > 0 \\ 0 & t \le 0. \end{cases}$$

Intuitively, on the positive real axis, $\bar{w}(t)$ is the piece-wise constant interpolation of $\{w_t\}$, with each piece having a length $\{\alpha_t\}$. We then define a helper function $\tau : \mathbb{N} \to \mathbb{R}$ as $\tau(k) \doteq \sum_{i=0}^{k} \alpha_k$. Next, we define a sequence of functions $\left\{f_i : \mathbb{R} \to \mathbb{R}^d\right\}$, where $f_i(t) \doteq \bar{w}(\tau(i) + t)$. The boundedness of $\{w_t\}$ follows from Theorem 17, implying that $\bar{w}(t)$ is also bounded, which in turn ensures that $\{f_i\}$ is a sequence of bounded functions. Under Assumptions A.1, A.2, A.6, A.7 (these assumptions have been verified in the proof of Theorem 18), and 5.2, Lemma 32 and the conclusion right above Lemma 34 of Liu et al. (2025a) proves there exists at least one convergent subsequence $\{f_{i_k}\} \subseteq \{f_i\}$, such that $\lim_{k\to\infty} f_{i_k}(t) = \lim_{k\to\infty} \bar{w}(\tau(i_k)+t) = \hat{w}(t)$, where $\hat{w}(t)$ is a bounded solution to ODE (3) on $(-\infty,\infty)$. Additionally, Theorem 13 guarantees that $\hat{w}(t)$ is a constant solution and in $\mathcal{W}_*$. Hence, we can conclude that for all $T<\infty$, we have $\lim_{k\to\infty} \bar{w}(\tau(i_k)+T) = w_*$, where $w_* \in \mathcal{W}_*$. Since $\bar{w}(t)$ is merely a piece-wise constant interpolation of the weight sequence $\{w_t\}$, the statement of this corollary immediately holds. ■

Appendix F. Proofs in Section 6

F.1 Proof of Lemma 20

Proof We first note that $\|Xw-v\|_D^2 = \left\|D^{1/2}(Xw-v)\right\|^2$. Then, by Theorem A.1, it holds that $X \arg\min_{w\in\arg\min_w \|Xw-v\|_D^2} \|w\| = X \arg\min_{w\in\arg\min_w \left\|D^{1/2}(Xw-v)\right\|^2} \|w\| = X(D^{1/2}X)^\dagger D^{1/2}v$. Hence, we have $\Pi = X(D^{1/2}X)^\dagger D^{1/2}$ by (20). ■

F.2 Proof of Lemma 21

Proof Lemma 4 of Tsitsiklis and Roy (1996) proves that $\mathcal{T}$ is a contraction mapping w.r.t. $\|\cdot\|_D$ under Assumption 2.2. Next, we show that Π is nonexpansive w.r.t. $\|\cdot\|_D$. Define $Z \doteq D^{1/2}X$. Then, by Lemma A.1, we have $\|\Pi v\|_D = \left\|X(D^{1/2}X)^\dagger D^{1/2}v\right\|_D = \left\|ZZ^\dagger D^{1/2}v\right\| \leq \left\|ZZ^\dagger\right\|\left\|D^{1/2}v\right\| \leq \left\|D^{1/2}v\right\| = \|v\|_D$. It then follows immediately that $\mathcal{T}\Pi$ is a contraction w.r.t. $\|\cdot\|_D$. ■

F.3 Proof of Lemma 22

Proof Our proof relies on the fact that $D(\gamma P - I)$ is negative definite (Sutton et al., 2016). For $w, w' \in \mathcal{W}_*$, we have

$$
\begin{aligned}
Aw + b - (Aw' + b) &= 0 \\
A(w-w') &= 0 \\
X^\top D(\gamma P - I)X(w-w') &= 0 \\
(w-w')^\top X^\top D(\gamma P - I)X(w-w') &= 0 \\
X(w-w') &= 0 \qquad (D(\gamma P - I) \text{ negative definite}) \\
Xw &= Xw'.
\end{aligned}
$$

■

F.4 Proof of Theorem 23

Proof We begin with the direction $\Pi\mathcal{T}Xw = Xw \implies Aw + b = 0$. Define $\Omega \doteq \{w \mid \Pi\mathcal{T}Xw = Xw\}$. Suppose $w_* \in \Omega$, we have

$$\begin{aligned}
\Pi\mathcal{T}Xw_* &= Xw_* \\
X(D^{1/2}X)^\dagger D^{1/2}\mathcal{T}Xw_* &= Xw_* \\
D^{1/2}X(D^{1/2}X)^\dagger D^{1/2}\mathcal{T}Xw_* &= D^{1/2}Xw_* \\
D^{1/2}X(D^{1/2}X)^\dagger D^{1/2}\mathcal{T}Xw_* &= D^{1/2}X(D^{1/2}X)^\dagger D^{1/2}Xw_* \quad (AA^\dagger A = A) \\
D^{1/2}X(D^{1/2}X)^\dagger D^{1/2}(\mathcal{T}Xw_* - Xw_*) &= 0 \\
(D^{1/2}X)^\top D^{1/2}X(D^{1/2}X)^\dagger D^{1/2}(\mathcal{T}Xw_* - Xw_*) &= 0 \\
(D^{1/2}X)^\top D^{1/2}(\mathcal{T}Xw_* - Xw_*) &= 0 \qquad (A^\top AA^\dagger = A^\top) \\
X^\top D(r_\pi + \gamma PXw_* - Xw_*) &= 0 \\
X^\top D(\gamma P - I)Xw_* + X^\top Dr_\pi &= 0 \\
Aw_* + b &= 0.
\end{aligned}$$

We now have $\Pi\mathcal{T}Xw = Xw \implies Aw + b = 0$. Next, we proceed to proving the other direction, i.e., $\Pi\mathcal{T}Xw = Xw \impliedby Aw + b = 0$. In view of (22), there exists at least one w_* such that $w_* \in \Omega$. The proof in the direction of $\implies$ then confirms that $w_* \in \mathcal{W}_*$. Let w be any weight in $\mathcal{W}_*$. Then, Lemma 22 implies that $Xw = Xw_* = v_*$. In view of (21), this means $w \in \Omega$. So, we have now proved that $w \in \mathcal{W}_* \implies w \in \Omega$, which completes the proof.
■

References

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Xabi Azagirre, Akshay Balwally, Guillaume Candeli, Nicholas Chamandy, Benjamin Han, Alona King, Hyungjun Lee, Martin Loncaric, Sébastien Martin, Vijay Narasiman, Zhiwei (tony) Qin, Baptiste Richard, Sara Smoot, Sean Taylor, Garrett van Ryzin, Di Wu, Fei Yu, and Alex Zamoshchin. A better match for drivers and riders: Reinforcement learning at lyft. *INFORMS Journal on Applied Analytics*, 2024.

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

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer, 1990.

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

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Proceedings of the Conference on Learning Theory*, 2018.

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.

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.

David Brandfonbrener and Joan Bruna. Geometric insights into the convergence of nonlinear td learning. In *Proceedings of the International Conference on Learning Representations*, 2020.

Mario Bravo and Roberto Cominetti. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 2024.

Mario Bravo, Roberto Cominetti, and Matías Pavez-Signé. Rates of convergence for inexact krasnosel'skii–mann iterations in banach spaces. *Mathematical Programming*, 2019.

Qi Cai, Zhuoran Yang, Jason D. Lee, and Zhaoran Wang. Neural temporal difference and Q learning provably converge to global optima. *Mathematics of Operations Research*, 2024.

Diogo Carvalho, Francisco S. Melo, and Pedro Santos. A new convergent variant of Q-learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2020.

Xuyang Chen, Jingliang Duan, Yingbin Liang, and Lin Zhao. Global convergence of two-timescale actor-critic for solving linear quadratic regulator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor-critic with linear function approximation. *IEEE Control Systems Letters*, 2022.

Zaiwei Chen, Siva Theja Maguluri, and Martin Zubeldia. Concentration of contractive stochastic approximation: Additive and multiplicative noise. *The Annals of Applied Probability*, 2025.

Wesley Chung, Somjit Nath, Ajin Joseph, and Martha White. Two-timescale networks for nonlinear value function approximation. In *Proceedings of the International Conference on Learning Representations*, 2019.

Roberto Cominetti, José A Soto, and José Vaisman. On the rate of convergence of krasnosel'skii-mann iterations and their connection with sums of bernoullis. *Israel Journal of Mathematics*, 2014.

Nicolò Dal Fabbro, Arman Adibi, Aritra Mitra, and George J. Pappas. Finite-time analysis of asynchronous multi-agent td learning. In *Proceedings of the American Control Conference*, 2024.

Peter Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 1992.

Peter Dayan and Terrence J. Sejnowski. TD(λ) converges with probability 1. *Machine Learning*, 1994.

Jean Gallier and Jocelyn Quaintance. *Algebra, Topology, Differential Calculus, and Optimization Theory For Computer Science and Engineering*. 2019.

Swetha Ganesh, Jiayu Chen, Washim Uddin Mondal, and Vaneet Aggarwal. Order-optimal global convergence for actor-critic with general policy and neural critic parametrization. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2025.

Peter W Glynn and Sean P Meyn. A liapounov bound for solutions of the poisson equation. *The Annals of Probability*, 1996.

Leah M Hackman. *Faster Gradient-TD Algorithms*. Master's thesis, University of Alberta, 2013.

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

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 2002.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 2022.

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

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003.

Chandrashekar Lakshminarayanan and Csaba Szepesvári. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2018.

Donghwan Lee and Niao He. Target-based temporal-difference learning. In *Proceedings of the International Conference on Machine Learning*, 2019.

Yitao Liang, Marlos C. Machado, Erik Talvitie, and Michael Bowling. State of the art control of atari games using shallow reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents & Multiagent Systems*, 2016.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient TD algorithms. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2015.

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.

Hamid Reza Maei. *Gradient temporal-difference learning algorithms.* PhD thesis, University of Alberta, 2011.

Sreejeet Maity and Aritra Mitra. Adversarially-robust TD learning with markovian data: Finite-time rates and fundamental limits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2025.

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

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

Aritra Mitra. A simple finite-time analysis of td learning with linear function approximation. *IEEE Transactions on Automatic Control*, 2025.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, 2019.

Yang Peng, Kaicheng Jin, Liangyu Zhang, and Zhihua Zhang. A finite sample analysis of distributional TD learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2025.

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

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

Xiaochi Qian, Zixuan Xie, Xinyu Liu, and Shangtong Zhang. Almost sure convergence rates and concentration of stochastic approximation and reinforcement learning with markovian noise. *ArXiv Preprint*, 2024.

Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2021.

Mark Bishop Ring. *Continual learning in reinforcement environments.* PhD thesis, The University of Texas at Austin, 1994.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

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

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

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

Richard S. Sutton, David A. McAllester, Satinder P. 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 $O(n)$ algorithm for off-policy temporal-difference 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.

Vladislav Tadic. On the convergence of temporal-difference learning with linear function approximation. *Machine Learning*, 2001.

John N. Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *IEEE Transactions on Automatic Control*, 1996.

John N. Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 1999.

Jiuqi Wang, Ethan Blaser, Hadi Daneshmand, and Shangtong Zhang. Transformers can learn temporal difference methods for in-context reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2025a.

Jiuqi Wang, Rohan Chandra, and Shangtong Zhang. Towards provable emergence of in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025b.

Yue Wang, Yi Zhou, and Shaofeng Zou. Finite-time error bounds for greedy-GQ. *Machine Learning*, 2024.

Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. In *Advances in Neural Information Processing Systems*, 2020.

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.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. In *Advances in Neural Information Processing Systems*, 2020a.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *ArXiv Preprint*, 2020b.

Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. In *Proceedings of the International Conference on Machine Learning*, 2021.

Long Yang, Gang Zheng, Yu Zhang, Qian Zheng, Pengfei Li, and Gang Pan. On convergence of gradient expected Sarsa(λ). In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In *Advances in Neural Information Processing Systems*, 2019.

Huizhen Yu. On convergence of emphatic temporal-difference learning. In *Proceedings of the Conference on Learning Theory*, 2015.

Huizhen Yu. Weak convergence properties of constrained emphatic temporal-difference learning with constant and slowly diminishing stepsize. *Journal of Machine Learning Research*, 2016.

Shangtong Zhang. Towards formalizing reinforcement learning theory. *ArXiv Preprint*, 2025.

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*, 2020a.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2020b.

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.

Shangtong Zhang, Remi Tachet Des Combes, and Romain Laroche. On the convergence of SARSA with linear function approximation. In *Proceedings of the International Conference on Machine Learning*, 2023.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. In *Advances in Neural Information Processing Systems*, 2019.