Gradient Flow Approximations in Temporal Difference Learning (Extended Version)

Amirreza Neshaei Moghaddam and Bahman Gharesifard

***Abstract*— We consider the continuous-time temporal difference (TD) learning dynamics with nonlinear value function approximations, where there is a slim understanding of the convergence properties in irreversible regimes. Motivated by Krener's linearization idea ala Lie-brackets, we obtain conditions on the approximating value function and irreversibility coeffients under which the TD dynamics behaves close to a gradient flow. We show that our conditions lead to a set of partial differential equations, and study the existence of solutions using the algebraic invertibility of differential operators. Whenever a solution exits, using a perturbation analysis, we provide a stability result for nonlinear TD dynamics. As a by-product, we state the implications of the results for the classical case of linear approximations, where our conditions are algebraic, and easily verifiable.**

I. Introduction

This paper is concerned with the nonlinear temporal difference learning dynamics [1]. We consider a Markov Reward Process (MRP) with a finite number of states s selected from a set S of cardinality n. We denote the probability that the next state is s_{next} by $P(s_{\text{next}}|s)$, with $r(s, s_{\text{next}})$ being the expected reward. Even though this assumption can be relaxed, to focus on the main idea of the paper, throughout we assume that that the underlying Markov process is aperiodic and irreducible, with a stationary distribution denoted by μ.

The classical setting under study here is an infinite-horizon stochastic reward: denoting by r_t the random variable representing the reward at a given time t, our goal is to find the true value function

$$V^*(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right],$$

starting from a given initial state $s_0 = s$, where $\gamma \in (0,1)$ is the discount factor.

Even though one could consider obtaining the value function using Bellman's equation or value iteration [10], in practice, the underlying probability transition matrix is not known; indeed, we often only have access to a sequence of actions/rewards $(s_0, r_1, s_1, r_2, \cdots)$. The main idea behind Temporal Difference (TD) learning method introduced in [9] is as follows: suppose that at state s_t we have an estimate of the true value function V^* denoted by $V(s_t)$. Note that we have obtained a reward r_{t+1} at state s_t. Now, given that the next state is s_{t+1}, by the Bellman's principle, we ask if our estimated value function is consistent with $r_{t+1} + \gamma V(s_{t+1})$. The term "temporal difference" hence naturally refers to

$$r_{t+1} + \gamma V(s_{t+1}) - V(s_t).$$

The temporal difference learning dynamics TD(0) is therefore constructed by updating our estimate of the value function to

$$V(s_t) + \alpha_t \left(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\right),$$

where α_t is an appropriately chosen step-size. In the tabular setting (no function parameterization), by slightly abusing the notation, we denote by V_t our estimate of the true value function at iteration t, and along a sample path, write the TD(0) update as

$$V_{t+1}(s) = V_t(s) + \alpha_t \left(r(s, s_{\text{next}}) + \gamma V_t(s_{\text{next}}) - V_t(s)\right).$$

It is well known that when the system and the learning rate α_t satisfy standard stochastic approximation conditions, the averaged dynamics is a discretized and sampled version of the expected continuous dynamics given by

$$\dot{V}(s) = \mu(s)(R(s) + \gamma \mathbb{E}(V(s_{\text{next}})) - V(s)), \tag{1}$$

where

$$R(s) = \mathbb{E}[r(s, s_{\text{next}})],$$

and the expectation is taken with respect to the underlying Markov process P. Using this fact, and following the notations in [1], we rewrite (1) in vector form as

$$\dot{V} = D_\mu(R + \gamma PV - V), \tag{2}$$

where D_μ is a diagonal matrix with diagonal being the entries of μ. It is worth pointing out that V here is a function of μ, even though we ignore explicitly indexing it. Note that by Bellman's equation, R satisfies

$$R = V^* - \gamma PV^*,$$

and hence we can write (2) as

$$\dot{V} = -A_\mu(V - V^*), \tag{3}$$

where

$$A_\mu = D_\mu(I - \gamma P).$$

Note that while A_μ is only symmetric when the MRP is reversible, it is always positive definite in the sense that $x^\top A_\mu x > 0$ for nonzero x due to $\frac{1}{2}(A_\mu + A_\mu^\top)$ being symmetric positive definite [9].

This setup where we directly update the value function estimate vector V is called the tabular setting. For this specific setting, we have that under the aforementioned conditions, the dynamics (3) always converges to V^* [9]. However, in most realistic settings, the cardinality of S is too large to employ a tabular approach, and hence, we naturally parameterize V by $\theta \in \mathbb{R}^m$ where $m << n$, and denote it by $V(\theta) \in \mathbb{R}^n$. We now focus on ODE associated with the TD(0) update in terms of the underlying parameter, i.e.,

$$\dot{\theta}(t) = -\nabla V(\theta(t))^\top A_\mu (V(\theta(t)) - V^*). \tag{4}$$

In the classical setting, $V(\theta)$ is often assumed to be linear, in that it is assumed that there are linearly independent features functions ϕ_i, where $i \in \{1, \ldots, m\}$ such that

$$V_\theta^{\text{linear}}(s) = \sum_{i=1}^{m} \theta_i \phi_i(s);$$

denoting by Φ the matrix with columns ϕ_i, this equation can be written as

$$V_\theta^{\text{linear}} = \Phi\theta. \tag{5}$$

It is well-known that under linear function approximation, when Φ is a full rank feature matrix, TD(0) converges to a unique fixed point [12]. The proof utilizes the fact that in the linear case, letting $\theta^* = (\Phi^\top A_\mu \Phi)^{-1} \Phi^\top A_\mu V^*$, the dynamics (4) become:

$$\dot{\theta}(t) = -\Phi^\top A_\mu \Phi(\theta - \theta^*), \tag{6}$$

and the positive definiteness of A_μ results in global convergence to θ^*.

In recent applications, neural networks have been employed to approximate $V(\theta)$, challenging the linear assumption. The nonlinear TD(0) dynamics is explored in [11]; notably, it's demonstrated that for a reversible underlying Markov chain, the TD(0) dynamics acts as a gradient flow [8]. Yet, the TD(0) dynamics generally does not constitute a gradient flow, even in linear scenarios[1]. In a recent work [1], interesting findings are presented for the case where the value function is assumed as a homogeneous function. It is important to note that even in this scenario, despite some preliminary results, the convergence characteristics remain incompletely understood.

II. TD(0) and Transformations

Our investigation begins by exploring the extent to which the TD(0) dynamics deviates from being a gradient flow. Gradient flows are naturally defined with respect to a metric and hence it becomes pertinent to inquire about the specific metric when considering whether a particular flow qualifies as a gradient flow, or an approximate to one. It is possible to directly explore this perspective, by asking for the existence of a metric that turns the nonlinear TD(0) into a gradient flow. We instead explore a different perspective with more relaxed conditions stemming from Krener's idea of feedback linearization [6].

To wit, let us rewrite (4) as

$$\dot{\theta}(t) = -\mathscr{S}(\theta(t)) + \mathscr{A}(\theta(t)), \tag{7}$$

where

$$\mathscr{S}(\theta(t)) := \nabla V(\theta(t))^\top (\frac{1}{2}(A_\mu + A_\mu^\top))(V(\theta(t)) - V^*)$$
$$\mathscr{A}(\theta(t)) := -\nabla V(\theta(t))^\top (\frac{1}{2}(A_\mu - A_\mu^\top))(V(\theta(t)) - V^*).$$

Clearly, $\mathscr{S}$ is a gradient, and in particular, for $\theta \in \mathbb{R}^m$

$$\mathscr{S}(\theta) = \nabla \frac{1}{2} \|V(\theta) - V^*\|^2_{\frac{1}{2}(A_\mu + A_\mu^\top)},$$

due to $\frac{1}{2}(A_\mu + A_\mu^\top)$ being a symmetric positive definite matrix. We naturally consider the case where the added term $\mathscr{A}(\theta(t))$ to this gradient is parameterized by a constant $h \geq 0$ by letting

$$h\Omega^\mu := \frac{1}{2}(A_\mu - A_\mu^\top),$$

where Ω^μ is skew-symmetric. We also let

$$S^\mu := \frac{1}{2}(A_\mu + A_\mu^\top).$$

In this sense, from this point onwards, the matrix with respect to which (4) is studied is $S^\mu + h\Omega^\mu$. As a result, (7) can be rewritten as

$$\dot{\theta}(t) = -\mathscr{S}(\theta(t)) + h\mathscr{A}_\Omega(\theta(t)), \tag{8}$$

where

$$\mathscr{S}(\theta(t)) := \nabla V(\theta(t))^\top S^\mu (V(\theta(t)) - V^*)$$
$$\mathscr{A}_\Omega(\theta(t)) := -\nabla V(\theta(t))^\top \Omega^\mu (V(\theta(t)) - V^*).$$

As previously mentioned, S^μ is positive definite, and hence non-singular. We denote the Lie bracket of two given vector field X and Y on $\mathbb{R}^m$ by $[X, Y]$. With this notation in place, we have the following result.

Proposition 2.1: Consider (7) and suppose that there exists a transformation $\mathbf{T}^h : \mathbb{R}^m \to \mathbb{R}^m$ that satisfied

$$[\mathscr{S}(x), \mathbf{T}^h(x)] = \mathscr{A}(x), \tag{9}$$

for all $x \in \mathbb{R}^m$ and is naturally of order h, i.e.,

$$\mathbf{T}^h(x) = h\widetilde{\mathbf{T}}(x),$$

for $\widetilde{\mathbf{T}} \sim \mathcal{O}(1)$ in terms of h. Then

$$z(t) = \theta(t) + \mathbf{T}^h(\theta(t)) \tag{10}$$

satisfies

$$\dot{z}(t) = -\mathscr{S}(z(t)) + h^2 \mathbf{R}^h(z(t)), \tag{11}$$

where

$$\begin{aligned} &\mathbf{R}^h(z) \\ =&\frac{\partial \widetilde{\mathbf{T}}}{\partial \theta}\left((I + \mathbf{T}^h)^{-1}(z)\right) \mathscr{A}_\Omega\left((I + \mathbf{T}^h)^{-1}(z)\right) \\ &+ \left(\widetilde{\mathbf{T}}\left((I + \mathbf{T}^h)^{-1}(z)\right)\right)^\top \nabla^2 \mathscr{S}\left((I + \mathbf{T}^h)^{-1}(z)\right) \\ &\quad \times \left(\widetilde{\mathbf{T}}\left((I + \mathbf{T}^h)^{-1}(z)\right)\right) \\ &- \frac{1}{2}\left(\widetilde{\mathbf{T}}\left((I + \mathbf{T}^h)^{-1}(z)\right)\right)^\top \nabla^2 \mathscr{S}(z) \\ &\quad \times \left(\widetilde{\mathbf{T}}\left((I + \mathbf{T}^h)^{-1}(z)\right)\right) \\ &+ \mathcal{O}(h). \end{aligned} \tag{12}$$

[1] It is noteworthy to mention [7], where "gradient splitting" is introduced and applied to achieve precise convergence rates for TD(0). This concept, though, is limited to linear approximations.

The proof follows from a more general statement presented below.

Lemma 2.2: Consider the dynamics:

$$\dot{x}(t) = -\nabla f(x(t)) + F^h(x(t)), \tag{13}$$

where $x(t) \in \mathbb{R}^m$ for all $t \geq 0$, $f : \mathbb{R}^m \to \mathbb{R}$ and $F^h : \mathbb{R}^m \to \mathbb{R}^m$ are smooth vector fields, and $F^h(\cdot)$ is in $\mathcal{O}(h)$ with $h > 0$, i.e.,

$$F^h(x) = h\widetilde{F}(x),$$

for $\widetilde{F} \sim \mathcal{O}(1)$ in terms of h. Suppose that there exists a transformation $\mathbf{T}^h : \mathbb{R}^m \to \mathbb{R}^m$ such that

$$F^h(x) = \left[\nabla f(x), \mathbf{T}^h(x)\right], \tag{14}$$

for all $x \in \mathbb{R}^m$ and is naturally of order h, i.e.,

$$\mathbf{T}^h(x) = h\widetilde{\mathbf{T}}(x),$$

with $\widetilde{\mathbf{T}} \sim \mathcal{O}(1)$ in terms of h. Then

$$\dot{z}(t) = -\nabla f(z(t)) + h^2\mathbf{R}^h(z),$$

where

$$z = x + \mathbf{T}^h(x),$$

and $\mathbf{R}^h(z)$ is of order $\mathcal{O}(1)$ in h and is given by

$$\begin{aligned}
&\mathbf{R}^h(z)\\
=&\frac{\partial\widetilde{\mathbf{T}}}{\partial x}\left((I+\mathbf{T}^h)^{-1}(z)\right)\widetilde{F}\left((I+\mathbf{T}^h)^{-1}(z)\right)\\
&+\left(\widetilde{\mathbf{T}}\left((I+\mathbf{T}^h)^{-1}(z)\right)\right)^\top\nabla^3 f\left((I+\mathbf{T}^h)^{-1}(z)\right)\\
&\quad\times\left(\widetilde{\mathbf{T}}\left((I+\mathbf{T}^h)^{-1}(z)\right)\right)\\
&-\frac{1}{2}\left(\widetilde{\mathbf{T}}\left((I+\mathbf{T}^h)^{-1}(z)\right)\right)^\top\nabla^3 f(z)\\
&\quad\times\left(\widetilde{\mathbf{T}}\left((I+\mathbf{T}^h)^{-1}(z)\right)\right)\\
&+\mathcal{O}(h).
\end{aligned}$$

Proof: We have that

$$\begin{aligned}
\dot{z} =&\dot{x} + \frac{\partial\mathbf{T}^h}{\partial x}\dot{x}\\
=&(I_n + \frac{\partial\mathbf{T}^h}{\partial x}(x))(-\nabla f(x) + F^h(x))\\
=&-\nabla f(z - \mathbf{T}^h(x)) - \frac{\partial\mathbf{T}^h}{\partial x}(x)\nabla f(x)\\
&+F^h(x) + \frac{\partial\mathbf{T}^h}{\partial x}(x)F^h(x),
\end{aligned}\tag{15}$$

where $\frac{\partial\mathbf{T}^h}{\partial x}$ is the Jacobian of $\mathbf{T}^h$. Approximating now

$$\begin{aligned}
\nabla f(z - \mathbf{T}^h(x)) =&\nabla f(z) - \nabla^2 f(z)\mathbf{T}^h(x)\\
&+\frac{1}{2}\mathbf{T}^h(x)^\top\nabla^3 f(z)\mathbf{T}^h(x) + \mathcal{O}(h^3),
\end{aligned}$$

and for h small enough,

$$\begin{aligned}
\nabla^2 f(z) &= \nabla^2 f(x + \mathbf{T}^h(x))\\
&= \nabla^2 f(x) + \mathbf{T}^h(x)^\top\nabla^3 f(x) + \mathcal{O}(h^2).
\end{aligned}$$

Thus, we can rewrite this as

$$\begin{aligned}
&\nabla f(z - \mathbf{T}^h(x))\\
=&\nabla f(z) - \left(\nabla^2 f(x) + \mathbf{T}^h(x)^\top\nabla^3 f(x) + \mathcal{O}(h^2)\right)\mathbf{T}^h(x)\\
&+\frac{1}{2}\mathbf{T}^h(x)^\top\nabla^3 f(z)\mathbf{T}^h(x) + \mathcal{O}(h^3)\\
\overset{\text{(i)}}{=}&\nabla f(z) - \nabla^2 f(x)\mathbf{T}^h(x) - \mathbf{T}^h(x)^\top\nabla^3 f(x)\mathbf{T}^h(x)\\
&+\frac{1}{2}\mathbf{T}^h(x)^\top\nabla^3 f(z)\mathbf{T}^h(x) + \mathcal{O}(h^3),
\end{aligned}\tag{16}$$

where (i) follows from $\mathbf{T}^h(x) \sim \mathcal{O}(h)$ and therefore $\mathcal{O}(h^2)\mathbf{T}^h(x) \sim \mathcal{O}(h^3)$. Substituting (16) in (15),

$$\begin{aligned}
\dot{z} =&-\nabla f(z) + F^h(x) - \left(\frac{\partial\mathbf{T}^h}{\partial x}\nabla f(x) - \nabla^2 f(x)\mathbf{T}^h(x)\right)\\
&+\frac{\partial\mathbf{T}^h}{\partial x}(x)F^h(x) + \mathbf{T}^h(x)^\top\nabla^3 f(x)\mathbf{T}^h(x)\\
&-\frac{1}{2}\mathbf{T}^h(x)^\top\nabla^3 f(z)\mathbf{T}^h(x) + \mathcal{O}(h^3)\\
\overset{\text{(i)}}{=}&-\nabla f(z) + h^2\left(\frac{\partial\widetilde{\mathbf{T}}}{\partial x}(x)\widetilde{F}(x) + \widetilde{\mathbf{T}}(x)^\top\nabla^3 f(x)\widetilde{\mathbf{T}}(x)\right.\\
&\left.-\frac{1}{2}\widetilde{\mathbf{T}}(x)^\top\nabla^3 f(z)\widetilde{\mathbf{T}}(x) + \mathcal{O}(h)\right),
\end{aligned}$$

where (i) follows from (14). The proof concludes after replacing x with $(I + \mathbf{T}^h)^{-1}(z)$. ■

It is worth displaying the term denoted by $\nabla^2\mathscr{S}$ in (12) in components. Note that

$$\begin{aligned}
\mathscr{S}_i &= \sum_{k,\ell=1}^{m}\frac{\partial V_k}{\partial\theta_i}S^\mu_{k\ell}(V_\ell - V^*_\ell)\\
\frac{\partial\mathscr{S}_i}{\partial\theta_j} &= \sum_{k,\ell=1}^{m}\frac{\partial^2 V_k}{\partial\theta_i\partial\theta_j}S^\mu_{k\ell}(V_\ell - V^*_\ell) + \frac{\partial V_k}{\partial\theta_i}S^\mu_{k\ell}\frac{\partial V_\ell}{\partial\theta_j},
\end{aligned}$$

and hence

$$\begin{aligned}
\frac{\partial^2\mathscr{S}_i}{\partial\theta_j\partial\theta_p} =& \sum_{k,\ell=1}^{m}\frac{\partial^3 V_k}{\partial\theta_p\partial\theta_i\partial\theta_j}S^\mu_{k\ell}(V_\ell - V^*_\ell) + \frac{\partial^2 V_k}{\partial\theta_i\partial\theta_j}S^\mu_{k\ell}\frac{\partial V_\ell}{\partial\theta_p}\\
&+\frac{\partial^2 V_k}{\partial\theta_i\partial\theta_p}S^\mu_{k\ell}\frac{\partial V_\ell}{\partial\theta_j} + \frac{\partial V_k}{\partial\theta_i}S^\mu_{k\ell}\frac{\partial^2 V_\ell}{\partial\theta_j\partial\theta_p}.
\end{aligned}$$

In particular, this term is zero when the approximation to the value function is linear.

Note that in the regime where h is small, existence of such a transformation would bring the dynamics (7) closer to a gradient flow. We will explain shortly why this is useful when one studies convergence properties of TD(0), but before we do that, let us dwell on the existence of such a transformation. Even though the main focus of this paper is on nonlinear temporal difference learning, it is fruitful to start with the linear case (5).

III. Linear Case

We start with a result which characterizes the existence of the transformation (10) for the case where the TD(0) approximation is linear.

Proposition 3.1: Consider the TD(0) dynamics with the the value function approximated linearly as in (5). Suppose that the matrices

$$\begin{bmatrix}\Phi^\top S^\mu \Phi & 0\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}\Phi^\top S^\mu \Phi & h\Phi^\top \Omega^\mu \Phi\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix} \tag{17}$$

are similar. Then there exits a transformation matrix $\mathbf{T}^h$ such that the dynamics (4) in coordinates given by (10) reads as

$$\dot{z}(t) = -\mathscr{S}(z(t)) + h^2 \mathbf{R}^h(z(t)), \tag{18}$$

where $\mathbf{R}^h(z)$ can be computed using (12).

Proof: As before, let us denote the linear approximation of the value function by $V_\theta^{\text{linear}} = \Phi\theta$, where Φ is of full column rank. We seek for a transformation of the form

$$\mathbf{T}^h(\theta) = K_1^h \theta + K_2^h, \tag{19}$$

where $K_1^h \in \mathbb{R}^{m\times m}$ and $K_2^h \in \mathbb{R}^m$, which satisfies (9). Expanding the latter equation with this choice of $\mathbf{T}^h$, we conclude that K_1^h and K_2^h need to satisfy

$$\begin{aligned} K_1^h \Phi^\top S^\mu(\Phi\theta - V^*) - \Phi^\top S^\mu \Phi(K_1^h\theta + K_2^h)& \\ = -h\Phi^\top \Omega(\Phi\theta - V^*).& \end{aligned}$$

As a result,

$$\Phi^\top S^\mu \Phi K_1^h - K_1^h \Phi^\top S^\mu \Phi = h\Phi^\top \Omega \Phi \tag{20}$$

$$-K_1^h \Phi^\top S^\mu V^* - \Phi^\top S^\mu \Phi K_2^h = h\Phi^\top \Omega V^*. \tag{21}$$

Given that S^μ is non-singular, and Φ is of full column rank, one can find K_2^h uniquely from the second equation, given that there exists K_1^h satisfying the first one. Using Lemma 5.1, a solution K_1^h to the first equation exists if and only if the matrices

$$\begin{bmatrix}\Phi^\top S^\mu \Phi & 0\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}\Phi^\top S^\mu \Phi & h\Phi^\top \Omega^\mu \Phi\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix}$$

are similar, which is assumed in the statement of the proposition. This shows that (19) satisfies the conditions of Proposition 2.1. The result then follows by substituting this transformation and using the linearity of the approximating value function to compute the remainder term. ■

Although this condition tends to hold for most cases, there are few instances of it failing. To have a concrete toy example at hand, note that for $\Phi = (S^\mu)^{-1/2}$, the conditions of Proposition 3.1 fail as the left-hand side of (20) will vanish, while the right-hand side may not. Moreover, we provide the following sufficient condition:

Lemma 3.2: A sufficient condition for the similarity mentioned in Proposition 3.1 to hold is the case that $\Phi^\top S^\mu \Phi$ has distinct eigenvalues.

Proof: Since $\Phi^\top S^\mu \Phi \in \mathbb{R}^{m\times m}$ is symmetric, there exists orthogonal matrix P such that

$$P^\top \Phi^\top S^\mu \Phi P = \begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_m\end{bmatrix} =: D.$$

Therefore, let

$$T_1 := \begin{bmatrix}P^\top & 0\\ 0 & P^\top\end{bmatrix},$$

and observe that due to the orthogonality of P,

$$T_1^{-1} = \begin{bmatrix}P & 0\\ 0 & P\end{bmatrix} = T_1^\top.$$

Now note that the matrices

$$\begin{bmatrix}\Phi^\top S^\mu \Phi & 0\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}\Phi^\top S^\mu \Phi & h\Phi^\top \Omega^\mu \Phi\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix}$$

from (17) are similar if and only if the matrices

$$\begin{aligned} &T_1 \begin{bmatrix}\Phi^\top S^\mu \Phi & 0\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix} T_1^{-1} \\ &= \begin{bmatrix}P^\top & 0\\ 0 & P^\top\end{bmatrix} \begin{bmatrix}\Phi^\top S^\mu \Phi & 0\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix} \begin{bmatrix}P & 0\\ 0 & P\end{bmatrix} \\ &= \begin{bmatrix}D & 0\\ 0 & D\end{bmatrix} \quad \text{and} \\ &T_1 \begin{bmatrix}\Phi^\top S^\mu \Phi & h\Phi^\top \Omega^\mu \Phi\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix} T_1^{-1} \\ &= \begin{bmatrix}P^\top & 0\\ 0 & P^\top\end{bmatrix} \begin{bmatrix}\Phi^\top S^\mu \Phi & h\Phi^\top \Omega^\mu \Phi\\ 0 & \Phi^\top S^\mu \Phi\end{bmatrix} \begin{bmatrix}P & 0\\ 0 & P\end{bmatrix} \\ &= \begin{bmatrix}D & hP^\top \Phi^\top \Omega \Phi P\\ 0 & D\end{bmatrix} \end{aligned}$$

are similar. Now let us denote the skew-symmetric matrix $hP^\top \Phi^\top \Omega \Phi P$ by W. Moreover, we define

$$T_2 := \begin{bmatrix}I & X\\ 0 & I\end{bmatrix},$$

for some $X \in \mathbb{R}^{m\times m}$ which will be computed later. It is easy to show

$$T_2^{-1} = \begin{bmatrix}I & -X\\ 0 & I\end{bmatrix}.$$

Now note that

$$\begin{aligned} T_2 \begin{bmatrix}D & W\\ 0 & D\end{bmatrix} T_2^{-1} &= \begin{bmatrix}I & X\\ 0 & I\end{bmatrix} \begin{bmatrix}D & W\\ 0 & D\end{bmatrix} \begin{bmatrix}I & -X\\ 0 & I\end{bmatrix} \\ &= \begin{bmatrix}D & W + XD\\ 0 & D\end{bmatrix} \begin{bmatrix}I & -X\\ 0 & I\end{bmatrix} \\ &= \begin{bmatrix}D & W - (DX - XD)\\ 0 & D\end{bmatrix}. \end{aligned}$$

Thus, if we show there exists X that can satisfy

$$DX - XD = W, \tag{22}$$

we have shown the similarity of $\left[\begin{smallmatrix}D & 0\\ 0 & D\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}D & W\\ 0 & D\end{smallmatrix}\right]$, and hence, finished the proof. In order to do so, observe that due to D being a diagonal matrix, (22) becomes

$$(DX - XD)_{ij} = (\lambda_i - \lambda_j)X_{ij} = W_{ij}, \tag{23}$$

for any $i, j \in \{1, 2, \ldots, m\}$. As a result of the skew-symmetry of W, the equation (23) is readily satisfied for $i = j$. Furthermore, for $i \neq j$, as D is assumed to have distinct eigenvalues, i.e. $\lambda_i \neq \lambda_j$ for $i \neq j$, we can choose

$$X_{ij} = \frac{W_{ij}}{\lambda_i - \lambda_j},$$

which satisfies (22), concluding the proof. ■

As we have mentioned before, the convergence of TD(0) dynamics in the linear case can be obtained directly without the need for approximations by gradient dynamics ala Proposition 3.1; this is in particular the case in light of the recent results in [7], where it is shown that in the linear case, TD(0) can be cast as a gradient splitting. This being said, our results focus more on the trajectories of the dynamics and how close to a gradient flow it can be shown to be (compared to just focusing on convergence to some point). Furthermore, investigating the convergence properties of TD(0) for the linear case provides a useful preparation for our later results in nonlinear settings. We have the following result.

Theorem 3.3: Consider the TD(0) dynamics with linear approximation (5), and suppose that (17) holds. Let z^* be the equilibrium of (18) with $h = 0$, and $Z = \{z \in \mathbb{R}^m \mid \|z - z^*\| < r\}$, where $r > 0$. Suppose that

$$h^2\|\mathbf{R}^h(z)\| \leq \delta < \beta r\sqrt{\frac{\lambda_{\min}^5(\Phi^\top S^\mu\Phi)}{\lambda_{\max}^3(\Phi^\top S^\mu\Phi)}}, \tag{24}$$

for some adjustable $\beta \in (0,1)$ and all $z \in Z$. Then after a finite time

$$T \leq \max\left\{0, \frac{\lambda_{\max}(\Phi^\top S^\mu\Phi)}{(1-\beta)\lambda_{\min}^2(\Phi^\top S^\mu\Phi)}\ln\frac{r\beta\lambda_{\min}^2(\Phi^\top S^\mu\Phi)}{\delta\lambda_{\max}(\Phi^\top S^\mu\Phi)}\right\},$$

the trajectory of (18) from z_0 with $\|z_0 - z^*\| \leq r\sqrt{\frac{\lambda_{\min}(\Phi^\top S^\mu\Phi)}{\lambda_{\max}(\Phi^\top S^\mu\Phi)}}$ satisfies

$$\|z(t) - z^*\| \leq \frac{\delta}{\beta}\sqrt{\frac{\lambda_{\max}^3(\Phi^\top S^\mu\Phi)}{\lambda_{\min}^5(\Phi^\top S^\mu\Phi)}},$$

for all $t \geq T$.

Proof: Since (17) holds, there exists a transformation $\mathbf{T}^h$ such that the dynamics (4) in coordinates given by(10) results in (18). When $h = 0$, the dynamics (18) reads

$$\dot{z}(t) = -\Phi^\top S^\mu\Phi(z - z^*).$$

Clearly, $\mathcal{V}(z) = \frac{1}{2}\|\Phi(z - z^*)\|_{S^\mu}^2$ is a Lyapunov function which, since Φ is of full rank and S^μ is PD, decays exponentially along the trajectories, because

$$\dot{\mathcal{V}}(z) \leq -\lambda_{\min}((\Phi^\top S^\mu\Phi)^2)\|z - z^*\|^2.$$

This, along with the assumption that Φ is of full rank, proves that (18) with $h = 0$ is exponentially stable. Note that

$$\frac{1}{2}\lambda_{\min}(\Phi^\top S^\mu\Phi)\|z-z^*\|^2 \leq \mathcal{V}(z) \leq \frac{1}{2}\lambda_{\max}(\Phi^\top S^\mu\Phi)\|z - z^*\|^2,$$

and

$$\|\frac{\partial\mathcal{V}}{\partial z}\| \leq \lambda_{\max}(\Phi^\top S^\mu\Phi)\|z - z^*\|.$$

We now consider the perturbed dyanmics (18). Suppose that (24) holds on the set Z, and that (18) is initialized at $z_0 \in Z$ with $\|z_0\| \leq r\sqrt{\frac{\lambda_{\min}(\Phi^\top S^\mu\Phi)}{\lambda_{\max}(\Phi^\top S^\mu\Phi)}}$. Gathering all these observations, the non-vanishing perturbation result in [5, Lemma 9.2] readily yields the proof. ■

It is natural to ask what happens if we apply the same non-vanishing perturbation analysis directly to (7). Indeed, the previous result applies verbatim to this case, with the exception that condition (24) is replaced with

$$h\|\Phi^\top\Omega^\mu\Phi(\theta(t) - \theta^*)\| \leq \delta < \beta r\sqrt{\frac{\lambda_{\min}^5(\Phi^\top S^\mu\Phi)}{\lambda_{\max}^3(\Phi^\top S^\mu\Phi)}}.$$

Given that $\mathbf{R}^h(z)$ is of order $\mathcal{O}(1)$ in h, the transformation that has allowed us to bring the dynamics closer to a gradient flow yields a larger range of h for which we can have some guarantees on the behavior of trajectories. We provide a more thorough investigation of this in the next section.

A. Impact of the Transformation on the Size of h Required by the Non-vanishing Perturbation Analysis

In this part, we investigate how the existence of a transformation that can satisfy the condition in Proposition 3.1 can alleviate the requirement on the size of h based on the non-vanishing perturbation analysis. First of all, recall that we pick

$$\mathbf{T}^h(\theta) = K_1^h\theta + K_2^h,$$

and let $K_1 := \frac{1}{h}K_1^h$ so that K_1 has to satisfy

$$\Phi^\top S^\mu\Phi K_1 - K_1\Phi^\top S^\mu\Phi = \Phi^\top\Omega\Phi,$$

which is independent of h. Moreover, note that due to this transformation, we have

$$\begin{aligned} z(t) &= \theta(t) + \mathbf{T}^h(\theta(t)) = (I + hK_1)\theta(t) + K_2^h \\ z^* &= (I + hK_1)\theta^* + K_2^h, \end{aligned}$$

where θ^* and z^* are the equilibria of the original system and the post-transformation system respectively. As a result,

$$\begin{aligned} \|z(t) - z^*\| &= \|(I + hK_1)(\theta(t) - \theta^*)\| \\ &\leq (\|I\| + h\|K_1\|)\|\theta(t) - \theta^*\| \\ &= (1 + h\|K_1\|)\|\theta(t) - \theta^*\|, \end{aligned} \tag{25}$$

and

$$\begin{aligned} \|z(t) - z^*\| &= \|(I + hK_1)(\theta(t) - \theta^*)\| \\ &\geq (\sigma_{\min}(I) - h\|K_1\|)\|\theta(t) - \theta^*\| \\ &= (1 - h\|K_1\|)\|\theta(t) - \theta^*\|. \end{aligned} \tag{26}$$

Now we move on to apply the analysis both on the original system (with θ) and the system after the transformation (with z). Note that the non-perturbed parts of the systems are the same, so we reuse the arguments made in 3.3 for the two cases as follows:

1) For the original system, let

$$\Theta = \{\theta \in \mathbb{R}^m \mid \|\theta - \theta^*\| < r\}.$$

Thus, it needs to hold for all $\theta \in \Theta$ that

$$h\|\Phi^\top\Omega^\mu\Phi(\theta - \theta^*)\| \leq \delta < \beta r\sqrt{\frac{\lambda_{\min}^5(\Phi^\top S^\mu\Phi)}{\lambda_{\max}^3(\Phi^\top S^\mu\Phi)}},$$

which, due to the definition of Θ, implies

$$h\|\Phi^\top \Omega^\mu \Phi\| r \le \delta < \beta r \sqrt{\frac{\lambda_{\min}^5(\Phi^\top S^\mu \Phi)}{\lambda_{\max}^3(\Phi^\top S^\mu \Phi)}}. \tag{27}$$

Now applying the analysis as before, we get that after a finite time τ_1, the trajectories from $\theta(0)$ with

$$\|\theta(0) - \theta^*\| \le r \sqrt{\frac{\lambda_{\min}(\Phi^\top S^\mu \Phi)}{\lambda_{\max}(\Phi^\top S^\mu \Phi)}} \tag{28}$$

satisfy

$$\|\theta(t) - \theta^*\| \le \frac{\delta}{\beta} \sqrt{\frac{\lambda_{\max}^3(\Phi^\top S^\mu \Phi)}{\lambda_{\min}^5(\Phi^\top S^\mu \Phi)}}, \tag{29}$$

for all $t \ge \tau_1$.

2) For the transformed system, let

$$Z = \{z \in \mathbb{R}^m \mid \|z - z^*\| < r'\}.$$

Thus, after calculating from (12) that for the linear case,

$$\begin{aligned}\mathbf{R}^h(z) &= \Phi^\top \Omega^\mu \Phi(\theta - \theta^*)\\ &= \Phi^\top \Omega^\mu \Phi(I + hK_1)^{-1}(z - z^*),\end{aligned}$$

we note that it needs to hold for all $z \in Z$ that

$$\begin{aligned}&h^2\|\Phi^\top \Omega^\mu \Phi(I + hK_1)^{-1}(z - z^*)\|\\ &\le \delta' < \beta r' \sqrt{\frac{\lambda_{\min}^5(\Phi^\top S^\mu \Phi)}{\lambda_{\max}^3(\Phi^\top S^\mu \Phi)}}.\end{aligned} \tag{30}$$

Therefore, since for all $z \in Z$,

$$\begin{aligned}&h^2\|\Phi^\top \Omega^\mu \Phi(I + hK_1)^{-1}(z - z^*)\|\\ &\le h^2\|\Phi^\top \Omega^\mu \Phi\|\|(I + hK_1)^{-1}\|\|(z - z^*)\|\\ &\le h^2\|\Phi^\top \Omega^\mu \Phi\| \frac{r'}{\sigma_{\min}(I + hK_1)}\\ &\le h^2\|\Phi^\top \Omega^\mu \Phi\| \frac{r'}{1 - h\|K_1\|},\end{aligned}$$

it suffices to satisfy that

$$h^2\|\Phi^\top \Omega^\mu \Phi\| \frac{r'}{1 - h\|K_1\|} \le \delta' < \beta r' \sqrt{\frac{\lambda_{\min}^5(\Phi^\top S^\mu \Phi)}{\lambda_{\max}^3(\Phi^\top S^\mu \Phi)}}. \tag{31}$$

Now applying the analysis as before, we get that after a finite time τ_2, the trajectories from $z(0)$ with

$$\|z(0) - z^*\| \le r' \sqrt{\frac{\lambda_{\min}(\Phi^\top S^\mu \Phi)}{\lambda_{\max}(\Phi^\top S^\mu \Phi)}} \tag{32}$$

satisfy

$$\|z(t) - z^*\| \le \frac{\delta'}{\beta} \sqrt{\frac{\lambda_{\max}^3(\Phi^\top S^\mu \Phi)}{\lambda_{\min}^5(\Phi^\top S^\mu \Phi)}}, \tag{33}$$

for all $t \ge \tau_2$. Now if we pick

$$r' = r(1 + h\|K_1\|) \quad \text{and} \quad \delta' = \delta(1 - h\|K_1\|),$$

we have that due to (25), the $z(0)$ starting points in (32) include (but are not limited to) all the respective $\theta(0)$ starting points of (28). Moreover, due to the specific choice of δ' and on account of (26), having $z(t)$ converge to the neighborhood (33) immediately implies convergence of the respective $\theta(t)$ to (29).

In conclusion, using the transformed system, we can guarantee the convergence of the trajectories starting from a wider set of initial points than (28) to a narrower set of endpoints around the equilibrium than (29). The difference between the two cases would be the demanded conditions of the two cases on the perturbation, which for the first case, as shown in (27), is

$$h\|\Phi^\top \Omega^\mu \Phi\| r \le \delta \Rightarrow h \le \frac{\delta}{r\|\Phi^\top \Omega^\mu \Phi\|}, \tag{34}$$

whereas for the second case, as mentioned in (31) and following our choice of δ' and r', we only need

$$\begin{aligned}&h^2\|\Phi^\top \Omega^\mu \Phi\| \frac{r'}{1 - h\|K_1\|} \le \delta' \Rightarrow\\ &h^2\|\Phi^\top \Omega^\mu \Phi\| \frac{r(1 + h\|K_1\|)}{1 - h\|K_1\|} \le \delta(1 - h\|K_1\|) \Rightarrow\\ &h^2 \frac{1 + h\|K_1\|}{(1 - h\|K_1\|)^2} \le \frac{\delta}{r\|\Phi^\top \Omega^\mu \Phi\|}.\end{aligned} \tag{35}$$

So for instance, when the right-hand side $\frac{\delta}{r\|\Phi^\top \Omega^\mu \Phi\|}$ is desired around 0.1 and K_1, which is independent of h, is of unit 2-norm (which is a reasonable value in examples), condition (34) requires $h \le 0.1$ whereas condition (35) only needs $h \lessapprox 0.222$ which allows much more room for perturbation which, in the case of TD, can be interpreted as the irreversibility of the original MRP.

IV. Nonlinear Case

We now arrive at the main part of this paper, where we deal with the nonlinear TD(0) dynamics. Our objective is to find a transformation $\mathbf{T}^h$ such that

$$\mathscr{A}(\theta) = \frac{\partial \mathbf{T}^h}{\partial \theta}(\theta)\mathscr{S}(\theta) - \frac{\partial \mathscr{S}}{\partial \theta}(\theta)\mathbf{T}^h(\theta), \tag{36}$$

where

$$\begin{aligned}\mathscr{S}(\theta) &= \nabla V(\theta)^\top S^\mu (V(\theta) - V^*)\\ \mathscr{A}(\theta) &= -\nabla V(\theta)^\top h\Omega^\mu (V(\theta) - V^*).\end{aligned}$$

Whenever such a transformation exists, we can invoke Proposition 2.1, which implies that for small values of h, the nonlinear TD(0) dynamics approximate, in a sense that will be made precise, a gradient flow. First, note that $V : \mathbb{R}^m \to \mathbb{R}^n$, in that for a given θ, $V(\theta)$ is an n-dimensional vector where each component is the approximation of the value function of the corresponding state $s \in S$. Using this and following through the calculations, it is clear that $\mathscr{S}$ maps $\mathbb{R}^m$ to $\mathbb{R}^m$. Keeping this in mind, we write (36) in coordinates as

$$\mathscr{A}_i(\theta) = \sum_k \frac{\partial \mathbf{T}_i^h}{\partial \theta_k}(\theta)\mathscr{S}_k(\theta) - \frac{\partial \mathscr{S}_i}{\partial \theta_k}(\theta)\mathbf{T}_k^h(\theta). \tag{37}$$

Equation (37) provides us with a partial differential equations on $\mathbf{T}$. We now discuss the existence of solutions to such equation, conveniently following Gromov's algebraic approach (see [3, Chapter 2], and Subsection V-A), simplified and adopted to the discussions here. To wit, let us define the differential operator L with

$$L(\mathbf{T}) = -(\nabla\mathscr{S})\mathbf{T} + \sum_{\ell=1}^{m} \mathscr{S}_\ell \frac{\partial \mathbf{T}}{\partial \theta_\ell},$$

and note that (37) can be written as

$$L(\mathbf{T}^h) = \mathscr{A}. \tag{38}$$

We show later in Proposition 5.2 that this system of partial differential equations (37) does not *generically* have a solution as it corresponds to the determined case ($q = m$) of Proposition 5.2. It is important to note that this does not mean that (37) never has a solution; indeed, the next example shows that in some scenarios a solution can be found.

Example 4.1: Consider the example with $m = n = 2$, and the dynamics

$$A_\mu = \begin{bmatrix} 2 & 1.1 \\ 0.9 & 1 \end{bmatrix},$$

and hence,

$$S^\mu = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad h\Omega^\mu = 0.1 \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

where you can take $h = 0.1$ and $\Omega^\mu = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$. Moreover, let

$$V(\theta) = \begin{bmatrix} \theta_1^2 + 2\theta_1\theta_2 + \theta_2^2 + \theta_1 \\ \theta_1^2 + 2\theta_1\theta_2 + \theta_2^2 + \theta_2 \end{bmatrix},$$

and thus,

$$\nabla V(\theta) = \begin{bmatrix} 2(\theta_1 + \theta_2) + 1 & 2(\theta_1 + \theta_2) \\ 2(\theta_1 + \theta_2) & 2(\theta_1 + \theta_2) + 1 \end{bmatrix}.$$

Taking $V^* = 0$ for simplicity, we have

$$\begin{aligned} \mathscr{S}(\theta) &= \nabla V(\theta)^\top \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} V(\theta) \\ &= \begin{bmatrix} 10(\theta_1 + \theta_2)^3 + 9\theta_1^2 + 16\theta_1\theta_2 + 7\theta_2^2 + 2\theta_1 + \theta_2 \\ 10(\theta_1 + \theta_2)^3 + 8\theta_1^2 + 14\theta_1\theta_2 + 6\theta_2^2 + \theta_1 + \theta_2 \end{bmatrix}, \end{aligned}$$

and

$$\begin{aligned} \mathscr{A}(\theta) &= -0.1\ \nabla V(\theta)^\top \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} V(\theta) \\ &= 0.1 \begin{bmatrix} \theta_1^2 - 2\theta_1\theta_2 - 3\theta_2^2 - \theta_2 \\ 3\theta_1^2 + 2\theta_1\theta_2 - \theta_2^2 + \theta_1 \end{bmatrix}. \end{aligned}$$

Now the transformation

$$\mathbf{T}^h(\theta) = 0.1 \begin{bmatrix} -(\theta_1 - \theta_2) \\ \theta_1 - \theta_2 \end{bmatrix} = 0.1 \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \theta$$

satisfies (37) as

$$\begin{aligned} &[\mathscr{S}(\theta), \mathbf{T}^h(\theta)] \\ =&\frac{\partial \mathbf{T}^h}{\partial \theta}(\theta)\mathscr{S} - \nabla\mathscr{S}(\theta)\mathbf{T}^h(\theta) \\ =&0.1 \begin{bmatrix} -\theta_1^2 - 2\theta_1\theta_2 - \theta_2^2 - \theta_1 \\ \theta_1^2 + 2\theta_1\theta_2 + \theta_2^2 + \theta_1 \end{bmatrix} \\ &+ 0.1 \begin{bmatrix} 2\theta_1^2 - 2\theta_2^2 + \theta_1 - \theta_2 \\ 2\theta_1^2 - 2\theta_2^2 \end{bmatrix} \\ =&0.1 \begin{bmatrix} \theta_1^2 - 2\theta_1\theta_2 - 3\theta_2^2 - \theta_2 \\ 3\theta_1^2 + 2\theta_1\theta_2 - \theta_2^2 + \theta_1 \end{bmatrix} \\ =&\mathscr{A}(\theta). \end{aligned}$$

We present the trajectories of the original and transformed systems for initial conditions $\theta(0) = \left[\begin{smallmatrix} -1 \\ 3 \end{smallmatrix}\right]$ and $z(0) = \theta(0) + \mathbf{T}^h(\theta(0))$ respectively in Figure 1. This verifies that the trajectories of the transformed system can be shown to be closer to the gradient flow compared to the original system.

At last, we state a generalization of Theorem 3.3.

Theorem 4.2: Consider the nonlinear TD(0) dynamics (7), and suppose that V is such that (38) has a solution. Suppose that z^* is a uniformly asymptotically stable equilibrium of (11) with $h = 0$, and $Z = \{z \in \mathbb{R}^m \mid \|z - z^*\| < r\}$, where $r > 0$, and suppose that there exists class $\mathcal{K}$ functions α_1, α_2, α_3, and α_4 such that

$$\alpha_1(\|z - z^*\|) \le \mathcal{E}(z) \le \alpha_2(\|z - z^*\|), \tag{39}$$

$$\alpha_3(\|z - z^*\|) \le \|\nabla\mathcal{E}(z)\|^2 \le \alpha_4(\|z - z^*\|), \tag{40}$$

for $z \in Z$, where $\mathcal{E}(z) = \frac{1}{2}\|V(z) - V^*\|^2_{S^\mu}$. If we have that

$$h^2\|\mathbf{R}^h(z)\| \le \delta < \frac{\beta\alpha_3(\alpha_2^{-1}(\alpha_1(r)))}{\sqrt{\alpha_4(r)}}, \tag{41}$$

for some adjustable $\beta \in (0, 1)$ and all $z \in Z$, then after a finite time T, the trajectory of (11) from z_0 with $\|z_0 - z^*\| \le \alpha_2^{-1}(\alpha_1(r))$ satisfies

$$\|z(t) - z^*\| \le \alpha_1^{-1}\left(\alpha_2\left(\alpha_3^{-1}\left(\frac{\delta\sqrt{\alpha_4(r)}}{\beta}\right)\right)\right) =: \rho(\delta).$$

The proof of this follows immediately from [5, Lemma 9.3].

Remark 4.3: Note that assumption (39) is reasonable and easy to verify. Assumption (40), however, is more involved and along with (41) provides conditions related to the matrix $\nabla V(z)\nabla V(z)^\top$. This matrix shapes the dynamics of TD and in the case of neural networks under particular assumptions it is known as the neural tangent kernel [4] (there is a similar condition on this matrix in [1, Theorem 3]). Indeed,

$$\begin{aligned} \|\nabla\mathcal{E}(z)\|^2 &= \|\mathscr{S}(z)\|^2 \\ &\le 2\lambda_{\max}(S^\mu)\mathcal{E}(z)\|\nabla V(z)\nabla V(z)^\top\| \\ &\le 2\lambda_{\max}(S^\mu)\alpha_2(\|z - z^*\|)\|\nabla V(z)\nabla V(z)^\top\|. \end{aligned}$$

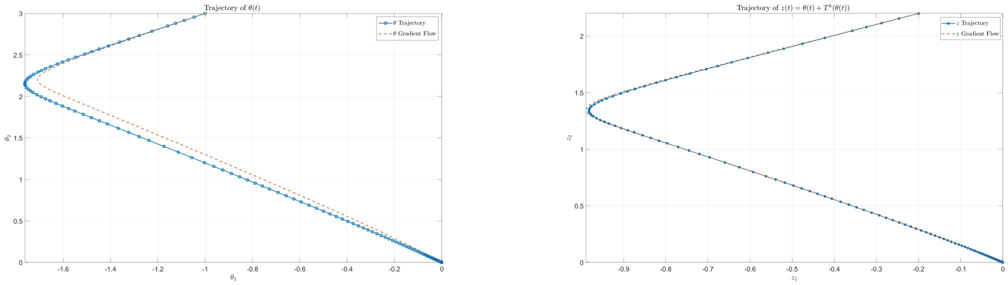

Fig. 1. Sample trajectory of the original and transformed systems compared to the gradient flow (Example 4.1).

V. Summary & Conclusion

We investigated the convergence properties of continuous-time temporal difference (TD) learning dynamics, focusing on both linear and nonlinear cases. Our analysis used Krener's linearization techniques and explored conditions under which the TD dynamics can be brought closer to a gradient flow. For the linear case, we already know that when the feature matrix is full rank, TD(0) converges to a unique fixed point. However, we introduced the concept of a transformation that can bring the dynamics closer to a gradient flow, providing insights into the transient behavior of the TD(0) dynamics as well. Our results also include conditions for the existence of such transformations and their impact on the stability and convergence of the TD(0) dynamics. In the nonlinear case, we demonstrated via Gromov's result that while a universal solution to the associated partial differential equations does not always exist, solutions can be found for certain systems. We provided an example illustrating the transformation of nonlinear TD(0) dynamics and how it can be brought closer to a gradient flow.

The findings offer a deeper understanding of the stability and convergence properties of TD learning and also investigate their transient behavior. Future work may explore further extensions to other reinforcement learning algorithms and their convergence behaviors under different conditions.

References

[1] D. Brandfonbrener and J. Bruna. Geometric insights into the convergence of nonlinear TD learning. In *International Conference on Learning Representations*, 2020.

[2] F. Gerrish and A. J. B. Ward. Sylvester's matrix equation and Roth's removal rule. *The Mathematical Gazette*, 82(495):423–430, 1998.

[3] M. Gromov. *Partial Differential Relations*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer Berlin Heidelberg, 2013.

[4] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

[5] H. K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2002.

[6] A. J. Krener. *Feedback Linearization*, pages 66–98. Springer New York, New York, NY, 1999.

[7] R. Liu and A. Olshevsky. Temporal difference learning as gradient splitting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6905–6913, 2021.

[8] Y. Ollivier. Approximate temporal difference learning is a gradient descent for reversible policies. *arXiv preprint arXiv:1805.00869*, 2018.

[9] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[10] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

[11] H. Tian, I. C. Paschalidis, and A. Olshevsky. On the performance of temporal difference learning with neural networks. In *International Conference on Learning Representations*, 2023.

[12] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

Appendix

We recall the following result from [2].

Lemma 5.1: The equation

$$AX - XB = C,$$

where $A \in \mathbb{R}^{r\times r}, B \in \mathbb{R}^{s\times s}$, and $X, C \in \mathbb{R}^{r\times s}$ has a solution X if and only if the matrices

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \text{ and } \begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$$

are similar, i.e., there exists some matrix $T \in \mathbb{R}^{(r+s)\times(r+s)}$ such that

$$\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} = T \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} T^{-1}.$$

A. Gromov

For completeness and clarity of notations, we suitably adopt a few notions and results from [3, Chapter 2]. Suppose that Y and G are C^∞-smooth vector bundles of dimensions q and m over V, a subspace of dimension n, taken to be $\mathbb{R}^n$ in most parts of what follows. Consider a linear differential operator $L : Y^{(1)} \to G$, where $Y^{(1)}$ is the space of 1-jets of germs of smooth sections of Y, given in local coordinates $\theta_1, \ldots, \theta_n$ on V as

$$L(y) = Ay + \sum_{i=1}^{n} B_i \frac{\partial y}{\partial \theta_i}, \tag{42}$$

where $A : \mathbb{R}^n \to \mathbb{R}^{m\times q}$ and $B_i : \mathbb{R}^n \to \mathbb{R}^{m\times q}$ are smooth functions. Let now g be a smooth section of G and consider the linear partial differential equation

$$L(y) = g. \tag{43}$$

We say that (43) is *underdetermined* if $q > m$, *determined* when $q = m$, and *overdetermined* when $q < m$. For a given smooth section g, Gromov investigates the existence of a *right inverse* M for operator L such that $(L \circ M)g = g$. A more suitable notion is that of a *universal right inverse*, where one seeks for an operator $M = M(L, g)$ such that $L(M(L, g)) = g$ *for all* smooth sections g. In this sense, universal property ensures that the operator L "generically" has a solution. This being said, we note that lack of existence of a universal right inverse does not mean that a right inverse does not exist for a particular choice of g. The following result can be found in [3, Page 153-156].

Proposition 5.2 (Universal right inverse): Consider the linear differential operator L given in (42). Then

1) If $q = m$, (43) does not have a right universal inverse.
2) If $q > n(m + 1) + m$, then the partial differential equation (43) has a universal inverse; in particular, (43) reduces to the system of algebraic equations

$$\begin{aligned} By &= 0 \\ A^* y &= g, \end{aligned} \tag{44}$$

where

$$B = \begin{bmatrix} B_1 \\ \vdots \\ B_n \end{bmatrix} \quad \text{and} \quad A^* = A - \sum_{i=1}^{n} \frac{\partial B_i}{\partial \theta_i}. \tag{45}$$

The condition in the second part is only sufficient, and invertibility may be possible even with smaller q.