DEMYSTIFYING APPROXIMATE RL WITH ϵ-GREEDY EXPLORATION: A DIFFERENTIAL INCLUSION VIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Q-learning and SARSA(0) with ϵ-greedy exploration are leading reinforcement learning methods, and their tabular forms converge to the optimal Q-function under reasonable conditions. However, with function approximation, they exhibit unexpected behaviors, such as i.) policy oscillation and chattering, and ii.) convergence to different attractors (possibly even the worst policy) on different runs, apart from the textbook instability. Accordingly, a theory to explain these phenomena has been a long-standing open problem, even for basic linear function approximation (Sutton, 1999). Our work uses differential inclusion theory to provide the first framework for resolving this problem. We further illustrate via numerical examples how this framework helps explain these algorithms' asymptotic behaviors.

1 INTRODUCTION

Tabular versions of value-based Reinforcement Learning (RL) algorithms such as Q-learning and SARSA are known to converge to the optimal Q-function under reasonable conditions (Singh et al., 2000; Jaakkola et al., 1993; Tsitsiklis, 1994). However, the story of their approximate variants (those using *function approximation*) with *ϵ-greedy exploration* has been inconclusive. On the one hand, these variants, e.g., the Deep Q-Network (DQN) (Mnih et al., 2015), have shown significant empirical successes. On the other, there is also growing evidence of undesirable behaviors such as policy oscillation, i.e., indefinitely cycling between multiple policies, or convergence to a sub-optimal or even the worst possible policy (Gordon, 1996; 2000; De Farias & Van Roy, 2000; Bertsekas, 2011; Young & Sutton, 2020). Accordingly, a mathematical framework to explain such behaviors and, in turn, to identify conditions for some minimal reliability has been a long-standing open problem, even in the basic linear function approximation case (Sutton, 1999, Problem 1).

By reliability of an approximate value-based method, we mean some basic notions like i.) stability, ii.) convergence to the optimal policy when the optimal Q-value function lies in the approximating function class, or iii.) convergence to a policy with a better Q-value function than the initial policy. Tabular variants of Q-learning and SARSA are reliable in all these three viewpoints under the reasonable conditions that guarantee their convergence (see references above). Likewise, there are sufficient, albeit restrictive conditions (Melo et al., 2008; Chen et al., 2019; Carvalho et al., 2020; Lee & He, 2020; Xu & Gu, 2020) under which approximate Q-learning and SARSA with a fixed behavior policy are reliable, at least as per the first two notions above.

In contrast, we claim that the reliability of approximate value-based RL methods with ϵ-greedy exploration cannot be taken for granted. To illustrate why this is the case, let us consider Figure 1, showing different trajectories obtained by running a variant of DQN. This variant continues to use experience replay and a target network as in (Mnih et al., 2015), but uses a *linear* function approximation instead of a neural network. We remove the limitations inherited due to the latter by ensuring that the optimal Q-value function is one of the columns of the feature matrix and, hence, is in the approximating function class (in the figure the optimal Q-function's weight parameter is displayed as a blue dot). The expected behavior here would be convergence to the optimal policy on each run of the experiment, as in the fixed behavior policy case. Surprisingly, we observe either convergence to a (sub-optimal) policy parameter (red trajectory) or oscillation between two sub-optimal policies (black trajectory, tail end).

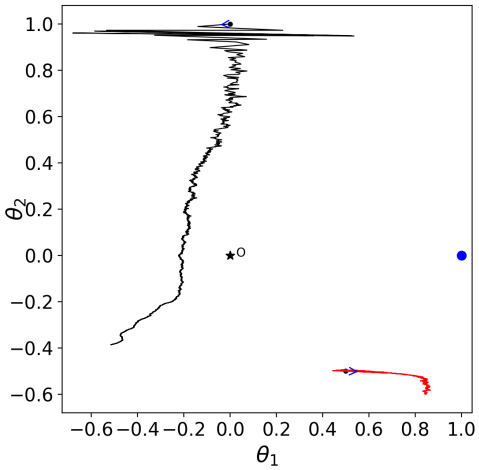

Figure 1: Example trajectories of iterates of DQN with a linear 2-dimensional Q-value approximation, on a 2-state 2-action MDP. The algorithm samples transitions from a replay buffer and carries out updates involving delayed targets (linear weights in this case). The blue dot on the right at (1,0) represents the weight parameter θ that *perfectly represents* the true optimal Q-value function, to which convergence would ideally be expected. Instead, the black trajectory proceeds downward and in its tail exhibits policy oscillation between two greedy regions. The red trajectory continues rightward, with both trajectories moving away from the perfectly approximated Q-function.

The above discussion along with those in (Young & Sutton, 2020) show that approximate value-based methods with ϵ-greedy exploration exhibit several pathological behaviors beyond the *textbook instability*, and have some serious question marks about their usefulness in practice. A theoretical framework to explain these phenomena, thus, is extremely important, and has been an open problem for nearly three decades now. While analysis based on Stochastic Approximation (SA) and Ordinary Differential Equations (ODEs) has tremendously helped in the simpler value prediction problem, the same idea has had limited success in RL control. This is due to its heavy dependence on smoothness of iterate dynamics (Benaïm, 1999, Section 4.1), (Borkar, 2009, Section 2.1), (Benveniste et al., 2012, Section 1.5). This is why all existing results in RL control require stringent smoothness properties on the update rule (Perkins & Precup, 2002; Melo et al., 2008; Chen et al., 2019; Zou et al., 2019; Maei et al., 2010; Bhatnagar & Lakshmanan, 2016; Avrachenkov et al., 2021).

To elaborate on why SA based analyses fail, note that RL schemes can be viewed as update rules in $\mathbb{R}^d$ having the form

$$\theta_{n+1} = \theta_n + \alpha_n[h(\theta_n) + M_{n+1}], \qquad n \geq 0, \tag{1}$$

where h is some $\mathbb{R}^d$-valued driving function, α_n is the stepsize, and M_{n+1} is noise. When h is *Lipschitz continuous*, as in the policy evaluation case, the ODE method can be used to show that the limiting dynamics of (1) is completely governed by the differential equation $\dot{\theta}(t) = h(\theta(t))$ (Benaïm, 1999; Borkar, 2009). However, with approximate value-based methods with ϵ-greedy exploration, the resulting h is *discontinuous*, rendering this standard approach inapplicable.

Key Contributions: The main highlights of this work can be summarized as follows.

1. *Analysis Framework*: Our work uses Differential Inclusion (DI) theory (Aubin & Cellina, 2012) to develop the first framework for analyzing value-based RL methods with discontinuous behavior policy changes. A DI is a generalization of a differential equation that helps circumvent the discontinuity issue by jointly considering all possible local update directions as a set.
2. *Explanation of Asymptotic Behaviors*: Using the above approach, we provide the first detailed explanation for the asymptotic behaviors of linear[1] Q-learning and SARSA(0) with ϵ-greedy ex-

[1]Henceforth, we will use linear Q-learning (resp. linear SARSA(0)) to refer to Q-learning (resp. SARSA(0)) with linear function approximation.

ploration. These variants do not involve experience replay and the target network; instead have oracle access to sample independently[2] from suitable limiting distributions. Our main result is that their asymptotic behaviors are completely governed by their respective limiting DIs. These limiting objects thus provide the first pathway to systematically identify and explain the whole range of limiting phenomena that an approximate value-based RL method with greedy exploration can exhibit, answering the question posed by Sutton (1999).

3. *Illustrations*: In Section 3, we discuss numerical examples to demonstrate how aspects of an algorithm's limiting DI govern whether it possesses unique or multiple attractors, and whether it is susceptible to chattering in policy or parameter space. We are also able to explain policy oscillation via a hitherto unseen 'sliding mode' phenomenon. Separately, we demonstrate that distinct algorithms (Q-learning and SARSA(0)) for the same RL problem can have qualitatively different limiting DIs and, accordingly, different limiting behaviors.

Related Work: Several works identify and report a variety of complex behaviors for approximate algorithms. These include the classic example of instability for linear Q-learning (Baird, 1995) and the phenomenon of chattering in linear SARSA(0) (Gordon, 1996) and its approach to a bounded region (Gordon, 2000). Apart from these, De Farias & Van Roy (2000) show how approximate value iteration may not possess any fixed points, while Bertsekas (2011) argues that approximate policy iteration schemes may generally be prone to policy oscillations, chattering, and convergence to poor solutions. More recently, Young & Sutton (2020) and Schaul et al. (2022) show experimentally how a range of approximate value-based RL methods involving greedification exhibit pathological behaviors such as policy oscillation, multiple fixed points, and consistent convergence to the 'worst' policy. Our framework, along with illustrations in Section 3, helps rigorously explain many of these phenomena.

On the theoretical front, while global convergence results exist for Q-learning and SARSA for the tabular (no approximation) setting, analysis with function approximation has been much more challenging. The difficulty in RL control primarily stems from the changing nature of policies (i.e., sampling distributions) which renders the iterations nonlinear (even with linear function approximation). One prominent stream of work focuses on extensions of linear SA theory to analyze nonlinear SA updates such as the ones in Q-learning with linear function approximation (Melo et al., 2008; Carvalho et al., 2020; Chen et al., 2019) and nonlinear (neural) function approximation (Fan et al., 2020; Xu & Gu, 2020). However, all these works *hold the behavior policy fixed* to mitigate the effect of changing sampling distributions. Another such notable work is that of Lee & He (2020), which uses ideas from switched system theory to analyze nonlinear ODEs. This is perhaps similar in spirit to our endeavor as switched systems are also useful for analyzing discontinuous dynamics.

Another set of analyses apply SA techniques to study SARSA(0) with incrementally changing policies (Perkins & Precup, 2002; Melo et al., 2008; Zhang et al., 2022; Zou et al., 2019). However, a restrictive assumption used here is that the policy improvement operator is Lipschitz-continuous, e.g., softmax, which ensures the limiting ODE is 'very smooth.' Note that ϵ-greedy policies do not fall in this class, which we are able to handle with the DI framework.

Finally, we remark that a few complicated variants of approximate Q-learning have already been analyzed using DI based approaches (Maei et al., 2010; Bhatnagar & Lakshmanan, 2016; Avrachenkov et al., 2021). However, the need for using a DI there is fundamentally different to the one in ours (the use of ϵ-greedy exploration). Maei et al. (2010) and Avrachenkov et al. (2021) look at variants of approximate Q-learning with a *fixed behavior policy* that represent *full gradient descent* with respect to suitably defined objective functions. A DI arises there due to discontinuities in the sub-gradient of the max of Q-function estimates. In contrast, Bhatnagar & Lakshmanan (2016) proposes a two-timescale variant of approximate Q-learning where the max is replaced by a gradient search in the space of *smoothly continuous policies*. A DI is natural there since the gradient search can lead to one of multiple local equilibria. These differences make their DIs very different to the ones we study; hence, these analyses do not carry over to the case with discontinuous behavior policy changes.

[2]Recall the main purpose for using replay buffers was to break correlations between samples.

2 SETUP AND MAIN RESULT

Let $\Delta(U)$ denote the set of probability measures on a set U. At the heart of RL, we have a Markov Decision Process (MDP) represented by the tuple $(\mathcal{S}, \mathcal{A}, \gamma, \mathbb{P}, r)$, where $\mathcal{S}$ denotes a finite state space, $\mathcal{A}$ represents a finite action space equipped with a total order, γ is the discount factor, and $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ are deterministic functions such that $\mathbb{P}(s,a)(s') \equiv \mathbb{P}(s'|s,a)$ specifies the probability of moving from state s to s' under action a, while $r(s,a,s')$ is the one-step reward obtained in this transition. One of the main goals in RL control is to find $Q_* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, the optimal Q-value function, associated with this MDP.

In practice, however, it is often the case that the state and action spaces are large. In such situations, to reduce the search space dimension, it is common to try and find an *approximation* to Q_*, instead of Q_* itself. In this work, we focus on linear function approximation. That is, we presume we are given a feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$ and the goal is to find a $\theta_* \in \mathbb{R}^d$ such that $Q_* \approx \Phi\theta_*$.

Two algorithms to find such a θ_* are linear Q-learning and SARSA(0) with ϵ-greedy exploration. To enable a joint investigation, we first propose a generic learning scheme that includes the above two methods as special cases. Let $\phi^T(s,a)$, where ^{T} implies transpose, denote the (s,a)-th row of Φ. Further, assume that Φ and r satisfy the following condition.

$\mathcal{B}_1$. Φ has full column rank. Further, there exist constants $K_r, K_\phi \geq 0$ such that $\|\phi(s,a)\| \leq K_\phi$ and $|r(s,a,s')| \leq K_r$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$.

Next, let $\epsilon \in (0,1]$ and $\pi_n^\epsilon : \mathcal{S} \to \Delta(\mathcal{A})$ the ϵ-greedy policy at time $n \geq 0$. That is, if θ_n is the estimate of our generic algorithm at time n, then let

$$\pi_n^\epsilon(a'|s) = \begin{cases} 1 - \epsilon(1 - 1/|\mathcal{A}|) & \text{if } a' = \arg\max_a \phi^T(s,a)\theta_n, \\ \epsilon/|\mathcal{A}| & \text{otherwise.} \end{cases} \tag{2}$$

We presume that $\arg\max$ breaks ties between actions that have the same value of $\phi^T(s,a)\theta_n$ using the total order on $\mathcal{A}$. Since the state and action spaces are finite, the number of ϵ-greedy policies is also finite. Suppose these policies satisfy the following condition.

$\mathcal{B}_2$. The Markov chains induced by each of the finitely many ϵ-greedy policies are ergodic and, hence, have their own unique stationary distributions.

Finally, let $\epsilon' \in [0,1]$. Then, given some initial estimate $\theta_0 \in \mathbb{R}^d$, the exact update rule of our generic algorithm at time $n \geq 0$ is

$$\begin{aligned} \theta_{n+1} &= \theta_n + \alpha_n \delta_n \phi(s_n, a_n) \\ \delta_n &= r(s_n, a_n, s_n') + \gamma \phi^T(s_n', a_n')\theta_n - \phi^T(s_n, a_n)\theta_n, \end{aligned} \tag{3}$$

where α_n is the stepsize, s_n is the current state at time n which we presume is independently[3] sampled from the stationary distribution corresponding to the Markov chain[4] induced by π_n^ϵ, $a_n \sim \pi_n^\epsilon(\cdot|s_n)$, while $s_n' \sim \mathbb{P}(\cdot|s_n, a_n)$ and $a_n' \sim \pi_n^{\epsilon'}(\cdot|s_n')$. Clearly, (3) with $\epsilon' = 0$ and $\epsilon' = \epsilon$ corresponds to the linear Q-learning and SARSA(0) algorithms with ϵ-greedy exploration, respectively. In particular, the $\max$ operator for Q-learning is implicitly specified via the manner in which action a_n' is sampled (from an ϵ'-greedy policy).

With regards to the stepsize sequence $(\alpha_n)_{n \geq 0}$, we presume that it satisfies the standard Robbins-Monro condition, i.e.,

$\mathcal{B}_3$. $\sum_{n \geq 0} \alpha_n = \infty$ and $\sum_{n \geq 0} \alpha_n^2 < \infty$.

We need a few more notations before we can state our main result. For $\mathbf{a} \equiv (\mathbf{a}(s))_s \in \mathcal{A}^{\mathcal{S}}$, let the associated *greedy region* $\mathcal{P}_{\mathbf{a}} := \{\theta \in \mathbb{R}^d : \forall s \in \mathcal{S}, \mathbf{a}(s) = \arg\max_a \phi^T(s,a)\theta\}$, where $\arg\max$ breaks ties using the total order. In other words, at each $\theta \in \mathcal{P}_{\mathbf{a}}$, $\mathbf{a}$ is the greedy policy

[3] This assumption ensures an idealized version of the experience replay.

[4] At any time $t \geq 0$, this Markov chain moves from state s to s' with probability $\sum_a \pi_n^\epsilon(a|s)\mathbb{P}(s'|s,a)$.

corresponding to the Q-value funtion $\Phi\theta$. Indeed, for some $\mathbf{a}$, it is possible that $\mathcal{P}_{\mathbf{a}} = \emptyset$. Nonetheless, $\{\mathcal{P}_{\mathbf{a}} : \mathbf{a} \in \mathcal{A}^{\mathcal{S}}\}$ partitions $\mathbb{R}^d$, i.e., for any $\theta \in \mathbb{R}^d$, there exists a unique $\mathbf{a}$ such that $\theta \in \mathcal{P}_{\mathbf{a}}$. Separately, let $\pi^{\epsilon}_{\mathbf{a}}$ (resp. $\pi^{\epsilon'}_{\mathbf{a}}$) be the ϵ-greedy (resp. ϵ'-greedy) policy associated with the policy $\mathbf{a}$. It is easy to see that $\pi^{\epsilon}_n = \pi^{\epsilon}_{\mathbf{a}}$ and $\pi^{\epsilon'}_n = \pi^{\epsilon'}_{\mathbf{a}}$ whenever $\theta_n \in \mathcal{P}_{\mathbf{a}}$.

Next, let $d^{\epsilon}_{\mathbf{a}}$ denote the stationary distribution associated with the Markov chain induced by $\pi^{\epsilon}_{\mathbf{a}}$. Further, let

$$b_{\mathbf{a}} = \mathbb{E}[\phi(s_n, a_n) r(s_n, a_n, s'_n)] = \Phi^T D^{\epsilon}_{\mathbf{a}} \mathbf{r} \tag{4}$$

and

$$A_{\mathbf{a}} = \mathbb{E}[\phi(s_n, a_n)\phi^T(s_n, a_n) - \gamma\phi(s_n, a_n)\phi^T(s'_n, a'_n)] = \Phi^T D^{\epsilon}_{\mathbf{a}}(\mathbb{I} - \gamma P^{\epsilon'}_{\mathbf{a}})\Phi, \tag{5}$$

where $D^{\epsilon}_{\mathbf{a}}$ is the diagonal matrix of size $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|$ whose (s, a)-th diagonal entry is $d^{\epsilon}_{\mathbf{a}}(s)\pi^{\epsilon}_{\mathbf{a}}(a|s)$, $\mathbf{r}$ is the $|\mathcal{S}||\mathcal{A}|$-dimensional vector whose (s, a)-th coordinate is $\mathbf{r}(s, a) = \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s, a) r(s, a, s')$, while $P^{\epsilon'}_{\mathbf{a}}$ is the matrix of size $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|$ such that $P^{\epsilon'}_{\mathbf{a}}((s, a), (s', a')) = \mathbb{P}(s'|s, a)\pi^{\epsilon'}_{\mathbf{a}}(a'|s')$. Note that $b_{\mathbf{a}}$ and $A_{\mathbf{a}}$ are generalizations of similar quantities that appear in the analysis of TD(0) with linear function approximation (e.g., (Sutton & Barto, 2018, (9.11))). Assume that the matrices $A_{\mathbf{a}}, \mathbf{a} \in \mathcal{A}^{\mathcal{S}}$, satisfy the following condition.

$\mathcal{B}_4$. For all $\mathbf{a} \in \mathcal{A}^{\mathcal{S}}$, $A_{\mathbf{a}}$ is positive definite, i.e., $\theta^T A_{\mathbf{a}} \theta > 0 \, \forall \theta \in \mathbb{R}^d$ ($A_{\mathbf{a}}$ need not be symmetric).

Finally, for $\theta \in \mathbb{R}^d$, let $f : \mathbb{R}^d \to \mathbb{R}^d$ and the set-valued $h : \mathbb{R}^d \to 2^{\mathbb{R}^d}$ be given by

$$f(\theta) := \sum_{\mathbf{a} \in \mathcal{A}^{\mathcal{S}}} (b_{\mathbf{a}} - A_{\mathbf{a}}\theta) \mathbb{1}[\theta \in \mathcal{P}_{\mathbf{a}}] \qquad \text{and} \qquad h(\theta) = \bigcap_{\delta > 0} \overline{\text{co}}(f(B(\theta, \delta))), \tag{6}$$

where $\mathbb{1}$ denotes the indicator function, $\overline{\text{co}}$ is the convex closure, and $B(\theta, \delta)$ is the open ball of radius δ centered at θ. As we show later in the proof outline for Theorem 2.1, $f(\theta)$ denotes our generic algorithm's expected update direction at time n if $\theta_n = \theta$, while $h(\theta)$ is the collection of all possible local update directions in the vicinity of θ. Note that, while $f(\theta)$ is *discontinuous*, $h(\theta)$ is *continuous in a set-valued sense*.

Our main result basically states that the limiting behavior of (3) is completely governed by the DI

$$\dot{\theta}(t) \in h(\theta(t)); \tag{7}$$

the convex closure in (6) ensures the above DI admits (absolutely) continuous solutions. Note that, unlike a differential equation, (7) uses $\in$ instead of $=$. Further, the RHS now can be a set of potentially several $\mathbb{R}^d$-valued vectors instead of a singleton. Importantly, unlike in ODEs, solutions of a DI need not be unique. In fact, this is one of the primary sources for the pathological behaviors in linear Q-learning and SARSA(0) with ϵ-greedy exploration.

In relation to (7), we will say a set $\Gamma \subset \mathbb{R}^d$ is *invariant* if, for every $x \in \Gamma$, there is *some* solution trajectory $(\theta(t))_{t \in (-\infty, \infty)}$ of (7) with $\theta(0) = x$ that lies entirely in Γ. An invariant set Γ is additionally said to be *internally chain transitive* if it is compact and, for $x, y \in \Gamma$, $\epsilon > 0$, and $T > 0$, there exist $m \geq 1$ and points $x_0 = x, x_1, \ldots, x_{m-1}, x_m = y$ in Γ such that one of the solution trajectories of (7) initiated at x_i meets the ϵ-neighborhood of x_{i+1} for $0 \leq i < m$ after a time that is equal or larger than T. Such characterizations are useful to restrict the possible choices for limiting sets I.

Theorem 2.1 (**Main Theorem**: SRI-DI connection)**.** *Suppose $\mathcal{B}_1, \ldots, \mathcal{B}_4$ hold. Then, almost surely, the iterates of* (3) *are stable, i.e.,* $\sup_n \|\theta_n\| < \infty$, *and converge to a (possibly sample path dependent) closed connected internally chain transitive invariant set of the DI in* (7).

Proof Outline. The crucial element in the proof is Theorem A.1, originally proved in (Borkar, 2009; Ramaswamy & Bhatnagar, 2017), that discusses the stability and convergence of discontinuous update rules such as (3) and their connections to DIs. Leaving the details to Appendix A, we show below how the expected update direction of our generic algorithm in (3) is $f(\theta_n)$; the connection to the limiting DI in (7) would then be clear.

For $n \geq 0$, let $\mathcal{F}_n := \sigma(\theta_0, s_0, a_0, s'_0, \ldots, s_{n-1}, a_{n-1}, s'_{n-1}, a'_{n-1})$, $y_n = \mathbb{E}[\delta_n \phi(s_n, a_n)|\mathcal{F}_n]$, and $M_{n+1} = \delta_n \phi(s_n, a_n) - y_n$. Separately, let $M_0 = 0$. Since θ_n is measurable with respect to $\mathcal{F}_n$,

$$f(\theta_n) = \sum_{\mathbf{a} \in \mathcal{A}^{\mathcal{S}}} (b_{\mathbf{a}} - A_{\mathbf{a}}\theta_n) \, \mathbb{1}[\theta_n \in \mathcal{P}_{\mathbf{a}}] \in h(\theta_n). \tag{8}$$

Accordingly, the update rule in (3) can now be rewritten as $\theta_{n+1} = \theta_n + \alpha_n[f(\theta_n) + M_{n+1}], n \geq 0$, which is of the desired form. □

Remark 2.2. *Theorem 2.1 represents our main contribution of the differential inclusion framework used to rigorously understand the behavior of RL algorithms with ϵ-greedy exploration. The almost sure convergence shows that the limiting DI in* (7) *captures all possible asymptotic behaviors of our generic scheme given in* (3)*.*

Remark 2.3. *If the only internally chain transitive invariant sets for the DI in* (7) *are isolated equilibrium points, then (θ_n) almost surely converges to a (possibly sample path dependent) equilibrium point. In that case, our generic scheme given in* (3) *will not exhibit parameter chattering.*

Remark 2.4. *Our formal proof for stability builds upon the result in (Ramaswamy & Bhatnagar, 2017). As in (Gordon, 2000), the positive definiteness of the $A_{\mathbf{a}}$ matrices (Assumption $\mathcal{B}_4$) is what enables this result. For SARSA(0), this assumption follows from $\mathcal{B}_2$; see (Gordon, 2000) for details.*

3 Illustration via Numerical Examples

We illustrate in this section how our main result (Theorem 2.1) can be used to explain the limiting dynamics of the Q-learning and SARSA(0) algorithms with linear function approximation and ϵ-greedy exploration. We study the algorithms applied to a range of 2-state 2-action MDPs, whose rewards, probability transition matrices and state-action features have been generated via random search (each MDP's details, such as probability transition function, reward function, discount factor, and the algorithm's exploration parameter ϵ, are provided in the Appendix. In each case, the MDP and features induce 2 greedy policies, for simplicity. We shall see that even with 2-dimensional linear function approximation, the algorithms' trajectories exhibit complex and non-trivial phenomena; we show how our results explain their causes via analysis of the associated DIs.

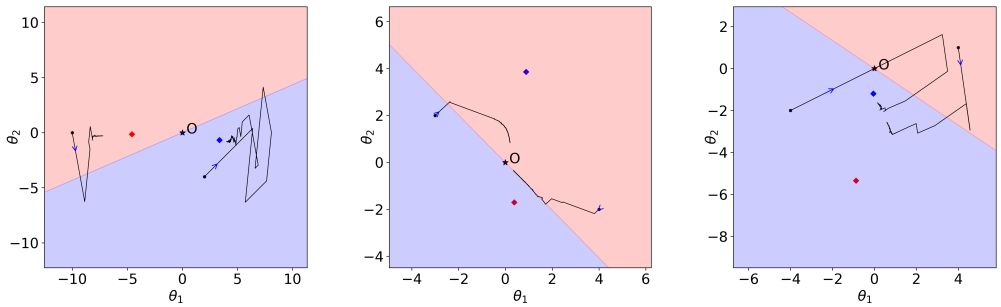

Figure 2: Trajectories under Q-learning. (Left) An MDP where function approximation in $\mathbb{R}^2$ gives rise to 2 greedy regions, both of which are self-consistent. The shaded areas represent the greedy regions, with the (self-consistent) landmarks for each partition's dynamics marked with diamonds of the respective colors. Two sample trajectories of Q-learning, with a $\Theta(1/n)$ stepsize schedule, are overlaid on the partition diagram with their starting iterates marked with black dots. The trajectories converge to any one of the self-consistent landmarks (behavior). The differential inclusion here has only the self-consistent landmarks as its asymptotically stable attractors (cause). (Center) An MDP where function approximation in $\mathbb{R}^2$ gives rise to 2 greedy regions, neither of which is self-consistent. Two sample trajectories of the iterates of Q-learning, with a $\Theta(1/n)$ stepsize schedule, are overlaid on the partition diagram with their starting iterates marked with black dots. The iterates converge to a unique point on the partition boundary (behavior) The differential inclusion here has a single 'sliding mode' attractor on the boundary (cause). (Right) An MDP where function approximation in $\mathbb{R}^2$ gives rise to 2 greedy regions, exactly one of which which is self-consistent. Two sample trajectories of the iterates of Q-learning, with a $\Theta(1/n)$ stepsize schedule, are overlaid on the partition diagram with their starting iterates marked with black dots. The iterates always converge to the unique self-consistent point in one of the partitions (behavior). The differential inclusion here has a unique attractor (cause).

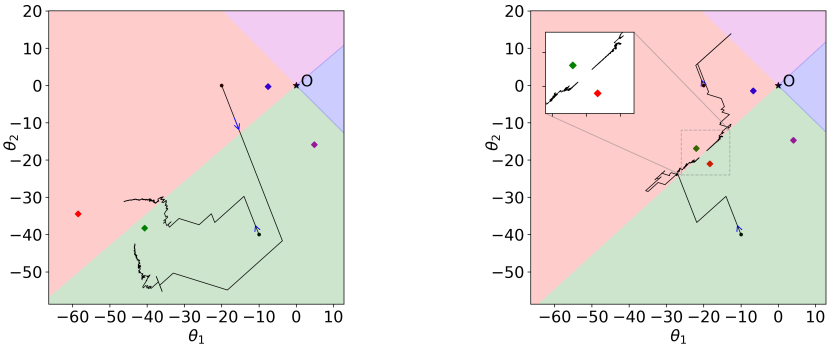

Figure 3: Sample trajectories of the iterates of Q-learning and SARSA(0), with a $\Theta(1/n)$ stepsize schedule, overlaid on their greedy regions. (Left) In the landmark structure induced by Q-learning, there is no 'sliding mode' point on the boundary of the green and red partitions, thus the trajectories move towards the red landmark in this case. (Right) In the landmark structure induced by SARSA(0), there is a 'sliding mode' equilibrium point on the boundary of the green and red regions towards which both trajectories move (see inset).

A. Two self-consistent greedy regions – Multiple attractors. The parameter space $\mathbb{R}^2$ is partitioned into 2 greedy regions (see definition of $\mathcal{P}_{\mathbf{a}}$ in Section 2) $\mathcal{P}_{\mathbf{a}_1}$ and $\mathcal{P}_{\mathbf{a}_2}$ via greedy policies $\mathbf{a}_1$ and $\mathbf{a}_2$, respectively.

Let us call the point $A_{\mathbf{a}_i}^{-1} b_{\mathbf{a}_i}$ a *landmark* for the partition $\mathcal{P}_{\mathbf{a}_i}$, $i = 1, 2$, where the $A_{\mathbf{a}}$ and $b_{\mathbf{a}}$ are as in (4),(5) (all matrices our examples will be positive-definite, hence invertible). The landmark $A_{\mathbf{a}_i}^{-1} b_{\mathbf{a}_i}$ is the (unique) point to which the ODE $\dot{\theta}(t) = b_{\mathbf{a}_i} - A_{\mathbf{a}_i}\theta$ would converge. Put another way, it is the point to which the Q-learning algorithm would converge were it to always sample actions a_n in (3) according to the ϵ-greedy policy based on $\mathbf{a}_i$.

The greedy regions in this example are both *self-consistent*. By this, we mean that each contains its own landmark: $A_{\mathbf{a}_i}^{-1} b_{\mathbf{a}_i} \in \mathcal{P}_{\mathbf{a}_i}$ for $i = 1, 2$. The landmarks, along with their respective color-shaded greedy regions, are shown as diamonds in Fig. 2a. Note that since each greedy region must be a cone and there are only 2 regions, each must be a halfspace with the origin 0 on its boundary.

Fig. 2a shows two sample paths of the iterates θ_n of Q-learning (update (3) and $\epsilon' = 0$), with $\epsilon = 0.1$ greedy exploration, $\gamma = 0.75$ discount factor, decaying stepsizes $\alpha_n = \Theta(1/n)$ and an iteration count of $20,000$, against the backdrop of the partition diagram. The starting positions, marked with black dots, were chosen arbitrarily. Note that segments of a trajectory are often parallel to each other, because the number of directions to move from an iterate is finite in a finite MDP.

Though the trajectories may traverse both greedy regions, they eventually approach one of the two landmarks (this observation remains true after several repeated trials). How can this be explained via our theoretical framework?

To answer this, let us look at the DI $\dot{\theta}(t) \in h(\theta(t))$ induced by Q-learning for this MDP, obtained via (6). Specifically, $h(\theta)$ equals the singleton set $\{b_{\mathbf{a}_1} - A_{\mathbf{a}_1}\theta\}$ for θ lying in the interior of the blue region (say $\mathcal{P}_{\mathbf{a}_1}$), the singleton set $\{b_{\mathbf{a}_2} - A_{\mathbf{a}_2}\theta\}$ in the interior of the red region (say $\mathcal{P}_{\mathbf{a}_2}$), and the connected segment $\text{co}\left(\{b_{\mathbf{a}_1} - A_{\mathbf{a}_1}\theta, b_{\mathbf{a}_2} - A_{\mathbf{a}_2}\theta\}\right)$ whenever θ is on the (1-dimensional) boundary of the two regions. In this DI, each point in the interior of a greedy region suffers a drift 'towards' the region's landmark. Points on the boundary, however, can be subjected to *any* drift in the convex hull of the drifts induced by each partition's affine dynamics at that point.

Theorem 2.1 says that with probability 1, the iterates must converge to an invariant and internally chain transitive set of the DI. Recall that an invariant set I means that $\forall \theta \in I$ there is some trajectory $\tilde{\theta}(t)$, $-\infty < t < \infty$ lying *entirely* in I and passing through θ at time 0. Roughly, for I to be internally chain transitive, for any pair of points $\theta_{\text{start}}, \theta_{\text{end}} \in I$, a closeness tolerance η and a travel time T, we must be able to find a solution of the DI and intermediate waypoints in I such that the

solution starts close to θ_{start}, ends close to θ_{end}, and reaches close to each successive waypoint after time T.

Examining this DI leads to the conclusion that it possesses only 3 invariant, internally chain transitive *singleton* sets – the 2 limiting dynamics points, plus a point θ° on the boundary whose convex hull contains the 0 vector[5] – for the following reasons. These 3 singletons are certainly invariant and internally chain transitive sets, because for each there is a solution of the DI that always stays at that point for all $-\infty < t < \infty$ by virtue of zero drift. It also follows that there can be no other invariant set for the DI, because (a) any solution of the DI that, at any time, is in the interior of a region continues within that region towards its respective landmark, and (b) any solution that, at any time, is on the boundary but not at θ° must suffer a drift that eventually pushes it into some greedy region's interior or along the boundary to θ°. We do not give a formal proof that these are the only invariant, internally chain transitive sets of the DI; the reader is directed to standard references on DIs Aubin & Cellina (2012)

Thus, according to our result, the stochastic iterates of Q-learning must eventually approach one of these 3 isolated points. We additionally remark that the point θ° on the boundary is an unstable equilibrium point, because any neighborhood around it contains points (in the interiors of $\mathcal{P}_{\mathbf{a}_1}$ and $\mathcal{P}_{\mathbf{a}_2}$) from where the DI's solutions will escape towards the landmarks. If the martingale difference noise (M_n) is sufficiently 'rich', i.e., it does not put all its probability mass in a subspace of $\mathbb{R}^2$, then random chance will cause the iterates to eventually escape this unstable equilibrium point. This condition will hold if all state-actions' features span $\mathbb{R}^2$, which they indeed do.

Our analysis also brings out the fact that asymptotically (as $n \to \infty$), there are no policy oscillations (iterates jumping regions infinitely often) nor parameter oscillations (iterates that 'bounce between' two regions at a positive distance) that can occur, because the iterates must eventually stabilize around one of the two landmarks in the interior of the greedy regions.

B. No self-consistent greedy regions – Sliding mode attractor. We consider an MDP and feature set that induce 2 greedy regions, each of which has its landmark in the *other* region. Trajectories of Q-learning, with the same parameters and stepsizes as before, appear to move eventually towards the boundary between the two regions (Fig. 2b).

Reasoning about the DI in this case, and its solutions, paints a qualitatively different picture than that of the previous example. As before, for each point θ in the interior of any greedy region, there is a singleton, nonzero drift $h(\theta) = \{b - A\theta\}$, based on the b, A of the region containing θ. However, this drift is directed toward the *other* region. It follows that any trajectory of the DI initialized at a region's interior point θ will reach the boundary in finite time, which rules out interior points from belonging to any internally chain transitive sets.

The only possible invariant, internally chain transitive sets, then, are to be found on the 1-dimensional boundary between the 2 regions. A moment's thought convinces us that the only such set is the singleton comprising the unique point $\theta^\star$ on the boundary for which both dynamics' drifts point in opposite directions (equivalently, 0 is in their convex hull). This point is strikingly different than the point θ° of the previous example in that is a globally *stable* equilibrium point – indeed, any deviation from this point pushes the DI's trajectory *towards* it rather than *away from* it. Moreover, observe that any trajectory of the DI started from an interior point of a region first reaches the boundary in finite time, due to a positive 'velocity' towards the boundary. Thereafter, it 'slides' along the boundary to reach $\theta^\star$, i.e., $\theta(t)$ always belongs to the boundary $\forall t \geq t_0$ for some time t_0. Such a trajectory is termed a *sliding mode* and has been extensively studied in control theory (see e.g., Utkin (2013)).

To our knowledge, the existence of sliding modes and associated equilibria has not been previously demonstrated for approximate RL algorithms. Also, our main result (Theorem 2.1) guarantees that asymptotically, the stochastic iterates θ_n will not oscillate or chatter, although policy oscillations are possible (the iterates may cross the boundary infinitely often on their way to $\theta^\star$).

C. Only one self-consistent greedy region – Unique landmark attractor. We consider an MDP that, with linear Q-function approximation in $\mathbb{R}^2$, induces two greedy regions with exactly *one* of

[5]The coordinates of this (2-dimensional) point can be analytically obtained by solving 2 equations expressing that: (1) θ° lies on the boundary of the greedy regions, (2) the vectors $b_{\mathbf{a}_{\text{red}}} - A_{\mathbf{a}_{\text{red}}}\theta^\circ$ and $b_{\mathbf{a}_{\text{blue}}} - A_{\mathbf{a}_{\text{blue}}}\theta^\circ$ point in opposite directions.

them being self-consistent (Fig. 2c). The DI for this setting has the property that points in the red region will experience a (vector) drift, towards the boundary, which is bounded away from 0 in norm. As a result, any DI trajectory that starts in the red region will, in finite time, hit the boundary. The driving (set-valued) function $h(\theta)$ for each boundary point can be shown to have all its constituent drift vectors 'pointing into' the blue region, so any solution from the boundary must instantaneously leave it. Continuing from here, the trajectory moves towards the blue landmark, which can be shown to be the sole invariant, internally chain transitive set of the DI. It is thus clear, from Theorem 2.1, that all the original stochastic iterates of the algorithm must eventually approach this point. This is a situation in which no policy or parameter chattering occurs. Also, as observed by Young and Sutton Young & Sutton (2020), it is possible that this (single) landmark point to which every trajectory converges may represent a very poor quality approximation for the optimal value function.

Figs. 3a and 3b show that even with the same MDP and features (and hence greedy regions), Q-Learning and SARSA(0) can induce different landmark structures, resulting in different trajectory behavior. We discuss other examples such as the chattering phenomenon observed in Gordon (1996); Young & Sutton (2020) in the Appendix.

4 Conclusion and Future Directions

We introduced the novel toolset of SRIs and DIs to help understand the algorithmic dynamics of approximate RL algorithms. We may have merely scratched the surface with regard to wielding the DI approach to its fullest potential. Our framework paves the way for a more comprehensive understanding of existing approximate-RL algorithms as well as the design of new schemes with 'well-behaved' DIs, hopefully resulting in value beyond just analysis to the realm of synthesis.

On a somber note, the insights we have uncovered about algorithms using arguably the simplest possible (linear) function approximation casts doubt on their utility in more complicated, nonlinear approximation architectures. It is plausible that there are many more failure modes in these settings than the range observed in the linear case, as also seen by Young & Sutton (2020) in neural approximation. Even with linear function approximation, we have shown in this paper that (a) one could converge to multiple attractors with disparate qualities, (b) one could oscillate between multiple policies' greedy regions, making it difficult to devise simple stopping criteria (c) even if one converges always to a unique point within a greedy partition, the induced greedy policy could be arbitrarily bad – these spell doom for the practitioner unless carefully addressed and investigated. Separately, we acknowledge that our theory addresses only linear function approximation and more work is needed to extend it to nonlinear approximation architectures.

Yet another promising direction for future research, which is not taken up in this paper, is that of the quality of the limiting iterates (assuming there is some reasonable form of convergence). Again, Young & Sutton (2020, 'Worst-case example') and Bertsekas (2011) point out that approximate policy iteration schemes could very well converge to the 'worst possible outcome', which, perhaps, with suitable design considerations based on the DI connection, can be avoided in new algorithms to come. For instance, a whole new class of algorithms can arise by assigning arbitrary values of ϵ, ϵ' in the generic update (3) with favourable properties. On a slightly different but related note, our work also reinforces the fact that merely ensuring stability of an incremental RL algorithm's iterates is by no means sufficient to guarantee good performance – discontinuous policy updates often swamp out gains from stability by inducing complex and varied convergence modes.

References

J-P Aubin and Arrigo Cellina. *Differential inclusions: set-valued maps and viability theory*, volume 264. Springer Science & Business Media, 2012.

Konstantin E Avrachenkov, Vivek S Borkar, Hars P Dolhare, and Kishor Patil. Full gradient dqn reinforcement learning: A provably convergent scheme. In *Modern Trends in Controlled Stochastic Processes:*, pp. 192–220. Springer, 2021.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Elsevier, 1995.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pp. 1–68. Springer, 1999.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions, part ii: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

Dimitri P Bertsekas. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335, 2011.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

Shalabh Bhatnagar and K Mohan Babu. New algorithms of the Q-learning type. *Automatica*, 44(4): 1111–1119, 2008.

Shalabh Bhatnagar and K Lakshmanan. Multiscale q-learning with linear function approximation. *Discrete Event Dynamic Systems*, 26(3):477–509, 2016.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Diogo Carvalho, Francisco S Melo, and Pedro Santos. A new convergent variant of q-learning with linear function approximation. *Advances in Neural Information Processing Systems*, 33:19412–19421, 2020.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *arXiv preprint arXiv:1905.11425*, 2019.

Jorge Cortes. Discontinuous dynamical systems. *IEEE Control systems magazine*, 28(3):36–73, 2008.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Daniela Pucci De Farias and Benjamin Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization theory and Applications*, 105(3):589–608, 2000.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Geoffrey J Gordon. Chattering in sarsa (lambda)-a cmu learning lab internal report. 1996.

Geoffrey J Gordon. Reinforcement learning with function approximation converges to a region. *Advances in neural information processing systems*, 13, 2000.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

A Lasota and Aaron Strauss. Asymptotic behavior for differential equations which cannot be locally linearized. *Journal of Differential Equations*, 10(1):152–172, 1971.

Donghwan Lee and Niao He. A unified switching system perspective and convergence analysis of q-learning algorithms. *Advances in Neural Information Processing Systems*, 33, 2020.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *UAI*, pp. 504–513. Citeseer, 2015.

Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S Sutton. Toward off-policy learning control with function approximation. In *ICML*, 2010.

Peter Marbach and John N Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.

Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pp. 664–671, 2008.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Alexander P Molchanov and Ye S Pyatnitskiy. Criteria of asymptotic stability of differential and difference inclusions encountered in control theory. *Systems & Control Letters*, 13(1):59–64, 1989.

Theodore Perkins and Doina Precup. A convergent form of approximate policy iteration. *Advances in neural information processing systems*, 15, 2002.

Theodore J Perkins and Mark D Pendrith. On the existence of fixed points for q-learning and sarsa in partially observable domains. In *ICML*, pp. 490–497, 2002.

Arunselvan Ramaswamy and Shalabh Bhatnagar. A generalization of the borkar-meyn theorem for stochastic recursive inclusions. *Mathematics of Operations Research*, 42(3):648–661, 2017.

Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. Citeseer, 1994.

Tom Schaul, André Barreto, John Quan, and Georg Ostrovski. The phenomenon of policy churn. *arXiv preprint arXiv:2206.00730*, 2022.

Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38(3):287–308, 2000.

Georgi V Smirnov. *Introduction to the theory of differential inclusions*, volume 41. American Mathematical Society, 2022.

Richard S Sutton. Open theoretical questions in reinforcement learning. In *European Conference on Computational Learning Theory*, pp. 11–17. Springer, 1999.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16 (3):185–202, 1994.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.

Vadim I Utkin. *Sliding modes in control and optimization*. Springer Science & Business Media, 2013.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. In *International Conference on Machine Learning*, pp. 10555–10565. PMLR, 2020.

Vinayaka G Yaji and Shalabh Bhatnagar. Stochastic recursive inclusions with non-additive iterate-dependent markov noise. *Stochastics*, 90(3):330–363, 2018.

Kenny Young and Richard S Sutton. Understanding the pathologies of approximate policy evaluation when combined with greedification in reinforcement learning. *arXiv preprint arXiv:2010.15268*, 2020.

Shangtong Zhang, Remi Tachet, and Romain Laroche. On the chattering of sarsa with linear function approximation. *arXiv preprint arXiv:2202.06828*, 2022.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. *Advances in Neural Information Processing Systems*, 32, 2019.