Why Target Networks Stabilise Temporal Difference Methods

Mattie Fellows [*] [1] Matthew J.A. Smith [*] [1] Shimon Whiteson [1]

Abstract

Integral to many recent successes in deep reinforcement learning has been a class of temporal difference methods that use infrequently updated *target values* for policy evaluation in a Markov Decision Process. At the same time, a complete theoretical explanation for the effectiveness of target networks remains elusive. In this work, we provide an analysis of this popular class of algorithms, to finally answer the question: “why do target networks stabilise TD learning”? To do so, we formalise the notion of a *partially fitted policy evaluation* method, which describes the use of target networks and bridges the gap between fitted methods and semigradient temporal difference algorithms. Using this framework we are able to uniquely characterise the so-called *deadly triad*–the use of TD updates with (nonlinear) function approximation and off-policy data–which often leads to nonconvergent algorithms. This insight leads us to conclude that the use of target networks can mitigate the effects of poor conditioning in the Jacobian of the TD update. Furthermore, we show that under mild regularity conditions and a well tuned target network update frequency, convergence can be guaranteed even in the extremely challenging off-policy sampling and nonlinear function approximation setting.

1. Introduction

Since their introduction in deep Q-networks (DQN) a decade ago (Mnih et al., 2013; 2015), target networks have become a common feature of state-of-the-art deep reinforcement learning algorithms (Lillicrap et al., 2016; Haarnoja et al., 2017; 2018; Fujimoto et al., 2018). Theoretical analysis of target networks has been limited and there has been no satisfactory explanation for their empirical success in stabilising policy evaluation algorithms. Whilst recent analysis has characterised the convergence properties of policy evaluation using target networks (Lee & He, 2019; Fan et al., 2020; Zhang et al., 2021), existing approaches focus on asymptotic results, and usually make simplifying assumptions that neither hold in practice nor account for the true behaviour of target network-based updates. Our work finds that the use of target networks can guarantee that deep RL algorithms will not diverge, even in regimes where traditional RL algorithms fail. Additionally, we establish the first finite-time performance bounds for target networks and general function approximation—without strong simplifying assumptions. Moreover, we prove our key stability assumption can always be satisfied by augmenting our updates with simple ℓ_2 regularisation that does not change the TD fixed points. In doing so, we finally provide theoretical justification for the empirical success that has been observed in challenging, off-policy tasks.

To achieve this, we analyse the use of infrequently updated target value functions by characterising them as a family of methods that we refer to as *partially fitted policy evaluation* (PFPE). This variant bridges the gap between fitted policy evaluation (FPE) (Le et al., 2019)—which iteratively fit the Bellman backups onto the class of representable function approximators —and classic temporal difference (TD) algorithms (Sutton, 1988) by limiting the fitting phase to a fixed number of steps, precisely reflecting the periodically updated target network algorithms as used in practice.

To characterise the performance of PFPE, we express our algorithm–which has traditionally been viewed through the lens of two-timescale analysis–using a single update applied only to the target network parameters. We show that the stability of the algorithm is determined by analysing the eigenvalues of the Jacobian of this update. This formulation allows us to characterise both the limiting (asymptotic) and finite-time (non-asymptotic) convergence properties of PFPE. Furthermore, it suggests, counterintuitively, that target networks are actually the object being optimised rather than merely a means to stabilise conventional TD updates. This insight leads us to empirically investigate a novel target parameter update scheme that uses a momentum-style update (Polyak, 1964), setting the stage for future research of practical target-based algorithms.

Our bounds on the finite-time performance of PFPE apply to off-policy, nonlinear and partially fitted methods, which

[*] Equal contribution [1] Department of Computer Science, University of Oxford,Oxford, United Kingdom. Correspondence to: Mattie Fellows <matthew.fellows@cs.ox.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

have never been investigated previously. We develop key insights into the usefulness of target networks, which we find do not improve asymptotic performance when decaying step sizes are used. Instead, target networks improve the conditioning of TD and fitted methods when the step size *does not tend to zero*, as is often implemented in practice. Under non-decaying stepsizes, our Jacobian analysis shows how PFPE reconditions the TD Jacobian allowing us to prove convergence in regimes where classic TD methods are unstable, thereby breaking the so-called deadly triad that has plagued TD methods (Sutton & Barto, 2018). Furthermore, our results do not depend on unwieldy assumptions or modifications of algorithms used in practice, such as projection, bounded state spaces, linear function approximation, or iterate averaging, as is done in previous analysis. In addition to our theoretical results, we experimentally evaluate our bounds on a toy domain, indicating that they are tight under relevant hyperparameter regimes. Taken together, our results lead to novel insight as to how exactly target networks affect optimisation, and when and why they are effective, leading to actionable results that can be used to further future research.

2. Preliminaries

Proofs for all theorems, propositions and corollaries can be found in Appendix B

We denote the set of all probability distributions on a set $\mathcal{X}$ as $\mathcal{P}(\mathcal{X})$. We use $\|\cdot\|$ to denote the ℓ_2-norm. For a matrix M, we denote the set of eigenvalues as $\lambda(M)$ with the set of maximum normed eigenvalues as $\lambda_{\max}(M) \coloneqq \arg\sup_{\lambda' \in \lambda(M)} |\lambda'|$ and $\lambda_{\min}(M) \coloneqq \arg\inf_{\lambda' \in \lambda(M)} |\lambda'|$. The ℓ_2-norm (spectral norm) for matrix M is $\|M\| = \sqrt{\lambda_{\max}(M^\top M)}$. Given a function $f : \mathcal{X} \to \mathbb{R}$ and a distribution $\mu \in \mathcal{P}(\mathcal{X})$, we denote the L_2-norm as: $\|f\|_\mu \coloneqq \sqrt{\mathbb{E}_{x \sim \mu}[f(x)^2]}$.

2.1. Reinforcement Learning

We consider the infinite horizon discounted RL setting. The agent interacts with an environment, formalised as a Markov Decision Process (MDP): $\mathcal{M} \coloneqq \langle \mathcal{S}, \mathcal{A}, P, P_0, R, \gamma \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, initial state distribution $P_0 \in \mathcal{P}(\mathcal{S})$, bounded stochastic reward kernel $R : \mathcal{S} \times \mathcal{A} \to \mathcal{P}([-r_{max}, r_{max}])$ where $r_{max} \in \mathbb{R} < \infty$ and scalar discount factor $\gamma \in [0, 1)$. An agent in state $s \in \mathcal{S}$ taking action $a \in \mathcal{A}$ observes a reward $r \sim R(s, a)$. The agent's behaviour is determined by a policy that maps a state to a distribution over actions: $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ and the agent transitions to a new state $s' \sim P(s, a)$. We denote the joint distribution of s', a', r conditioned on s, a for policy π as $P^\pi_{sar}(s, a)$. We seek to optimise (in the control case), or estimate (in the policy evaluation case) the expected discounted sum of future rewards starting from a given state $s \in \mathcal{S}$. This quantity is given by the state value function, $V^\pi(s) = \mathbb{E}_{a \sim \pi(s)}[Q^\pi(s, a)]$, with $Q^\pi : \mathcal{S} \times \mathcal{A} \to [-r_{max}/(1-\gamma), r_{max}/(1-\gamma)]$, the action value function, given recursively through the Bellman equation: $Q^\pi(s, a) = \mathcal{T}^\pi[Q^\pi](s, a)$, where the Bellman operator $\mathcal{T}^\pi$ projects functions forwards by one step through the dynamics of the MDP:

$$\mathcal{T}^\pi[Q^\pi](s, a) \coloneqq \mathbb{E}_{s', a', r \sim P^\pi_{sar}(s,a)}\left[r + \gamma Q^\pi(s', a')\right].$$

$\mathcal{T}^\pi$ is a γ-contractive mapping and thus has a fixed point, which corresponds to the true value of π (Puterman, 2014). When estimating MDP values, we employ a value function approximation $Q_\omega : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ parametrised by $\omega \in \Omega \subseteq \mathbb{R}^n$.

Many RL algorithms employ TD learning for policy evaluation, which combines bootstrapping, state samples and sampled rewards to estimate the expectation in the Bellman operator (Sutton, 1988). In their simplest form, TD methods update the function approximation parameters according to:

$$\omega_{i+1} = \\ \omega_i + \alpha_i \left(r + \gamma Q_{\omega_i}(s', a') - Q_{\omega_i}(s, a)\right) \nabla_\omega Q_{\omega_i}(s, a),$$

where $s \sim d, a \sim \mu(s), s', a', r \sim P^\pi_{sar}(s, a)$, $d \in \mathcal{P}(\mathcal{S})$ is a sampling distribution, and μ is a sampling policy that may be different from the target policy π. For simplicity of notation and to accommodate the introduction of target networks in Section 3, we define the tuple $\varsigma \coloneqq (s, a, r, s', a')$ with distribution P_ς and the TD-error vector as:

$$\delta(\omega, \omega', \varsigma) \coloneqq \left(r + \gamma Q_{\omega'}(s', a') - Q_\omega(s, a)\right) \nabla_\omega Q_\omega(s, a),$$

allowing us to write the TD parameter update as:

$$\omega_{i+1} = \omega_i + \alpha_i \delta(\omega_i, \omega_i, \varsigma).$$

We make the following i.i.d. assumption for clarity of exposition, but discuss other sampling regimes in Appendix D:

Assumption 1. *Each $s \sim d$ is drawn i.i.d..*

Typically, d is the steady-state distribution of an ergodic Markov chain. We denote the expected TD-error vector as: $\delta(\omega, \omega') \coloneqq \mathbb{E}_{\varsigma \sim P_\varsigma}[\delta(\omega, \omega', \varsigma)]$ and define the set of TD fixed points as:

$$\omega^\star \in \Omega^\star \coloneqq \{\omega | \delta(\omega, \omega) = 0\}.$$

If a TD algorithm converges, it converges to a TD fixed point. Convergence of TD methods can only be guaranteed for linear function approximators when sampling on-policy in an ergodic MDP, that is the agent sampling and target distributions are the same. We investigate the phenomenon further as part of our asymptotic analysis in Section 4.1.

3. Partially Fitted Policy Evaluation

Unfortunately, real-world applications of RL often demand the expressiveness of nonlinear function approximators like neural networks and/or the ability to use data that has been collected off-policy, i.e., by following a policy μ that differs from the target policy π for policy evaluation.

3.1. Fitted v Partially Fitted Policy Evaluation

Fitted methods improve on the sample efficiency and stability of TD methods by explicitly incorporating the limitations of the function approximation class through the use of a projection operator (Tsitsiklis & Van Roy, 1997). These methods generally perform some variant of the iterate $Q_{\bar{\omega}_{l+1}} = \Pi^{d^\pi} \mathcal{T}^\pi Q_{\bar{\omega}_l}$ where Π^d is the projection operator $\Pi^d Q = \arg\min_{Q'} \|Q' - Q\|_{d,\mu}$. These updates are known as fitted policy evaluation (PFE).

The projection step is needed to accommodate the fact that values generally cannot be exactly represented with function approximation. To obtain a practical way of carrying out the PFE updates, a separate set of *target parameters* can be introduced $\bar{\omega}_l \in \Omega$ that parameterise the TD target and are updated every k timesteps:

$$\omega_{kl+i+1} = \omega_{kl+i} + \alpha_{kl+i}\delta(\omega_{kl+i}, \bar{\omega}_l, \varsigma), \tag{1}$$

$$\bar{\omega}_{l+1} = \omega_{k(l+1)}, \tag{2}$$

The function approximator update in Equation (1) carries out k iterations of stochastic gradient descent (SGD) on the loss:

$$\mathcal{L}(\omega; \bar{\omega}_l) := \|Q_\omega - \mathcal{T}^\pi[Q_{\bar{\omega}_l}]\|_{d,\mu},$$

before updating the target parameters. In the limit as $k \to \infty$, assuming convergence of SGD to a global minimum, fully fitted policy evaluation occurs by finding $\omega_\infty \in \arg\inf_{\omega\in\Omega} \mathcal{L}(\omega, \bar{\omega}_l)$.

In practice k is finite and only partial policy evaluation occurs before updating the target parameters, a setting we call partially fitted policy evaluation (PFPE). Without loss of generality, we assume that $\bar{\omega}_0$ is deterministic with $\|\bar{\omega}_0\| < \infty$ and $\alpha_i = \alpha_l$ for all $kl \leq i < k(l+1)$, that is stepsizes only change after updating target parameters. As the target parameters are updated to the approximator parameters every k timesteps in Equation (2), it suffices to consider the target parameter update in isolation when analysing PFPE. Our goal is thus to analyse a single update for the target parameters in the canonical form:

$$\bar{\omega}_{l+1} = g^k(\bar{\omega}_l, \mathcal{D}, \alpha_l), \quad \mathcal{D} \sim P_{\mathcal{D}}, \tag{3}$$

where $\mathcal{D} := \{\varsigma_i\}_{i=1}^k$ is a set of k samples from the environment with distribution $P_{\mathcal{D}}$ and $g^k(\bar{\omega}_l, \mathcal{D}_l, \alpha_l)$ reduces the k nested updates from Equation (1) into a single update for the target parameters.

3.2. Jacobian Analysis

In our analysis, we show that the stability of the expected PFPE update $g^k(\bar{\omega}_l, \alpha_l) := \mathbb{E}_{\mathcal{D}\sim P_{\mathcal{D}}}\left[g^k(\bar{\omega}_l, \mathcal{D}, \alpha_l)\right]$ is determined by the conditioning of three Jacobians. We denote the Hessian of the loss as: $H(\omega; \bar{\omega}_l) := \nabla_\omega^2 \mathcal{L}(\omega; \bar{\omega}_l)$, the Jacobian of the TD-error vector as: $J_\delta(\omega; \bar{\omega}_l) := \nabla_{\omega'}\delta(\omega, \omega')|_{\omega'=\bar{\omega}_l}$ and define the TD Jacobian as: $J_{\text{TD}}(\bar{\omega}_l) := \nabla_\omega \delta(\omega, \omega)|_{\omega=\bar{\omega}_l}$. Observe that $J_{\text{TD}}(\bar{\omega}_l) = J_\delta(\bar{\omega}_l, \bar{\omega}_l) - H(\bar{\omega}_l; \bar{\omega}_l)$. Without loss of generality, we assume that the Hessian matrix is diagonalisable because, if it is not, an arbitrarily small perturbation can make its eigenvalues distinct and therefore diagonalisable. So that these matrices exist, we require that the expected PFPE update is differentiable almost everywhere, a condition that is guaranteed by a Lipschitz assumption. We also require that the variance of the updates is bounded, motivating the following regularity assumption:

Assumption 2 (Function Approximator Regularity)**.** *We assume that $\delta(\omega, \omega', \varsigma)$ is Lipschitz in ω, ω' with constant L: $\|\delta(\omega_1, \omega_1', \varsigma) - \delta(\omega_2, \omega_2', \varsigma)\| \leq L(\|\omega_1 - \omega_2\| + \|\omega_1' - \omega_2'\|)$ and Ω is convex, $\mathbb{V}_{\varsigma\sim P_\varsigma}[\delta(\omega, \omega, \varsigma)] := \mathbb{E}_{\varsigma\sim P_\varsigma}[\|\delta(\omega, \omega, \varsigma) - \delta(\omega, \omega)\|^2] \leq \sigma_\delta^2$ for some $\sigma_\delta^2 < \infty$.*

The bounded variance assumption can easily be achieved for unbounded function approximators by truncating the TD error vector, much like the commonly used gradient clipping in gradient descent. We now introduce the path-mean Jacobians, which are the principal element of our analysis:

$$\bar{H}(\omega, \omega^\star; \bar{\omega}_l) := -\int_0^1 \nabla_{\omega'}\delta(\omega' = \omega - t(\omega - \omega^\star), \bar{\omega}_l)dt,$$

$$\bar{J}_\delta(\omega, \omega^\star; \bar{\omega}_l) := \int_0^1 \nabla_{\omega'}\delta(\bar{\omega}_l, \omega' = \omega - t(\omega - \omega^\star))dt,$$

$$\bar{J}_{\text{TD}}(\omega, \omega^\star) := \int_0^1 \nabla_{\omega'}\delta(\omega', \omega')|_{\omega'=\omega-t(\omega-\omega^\star)}dt$$

Intuitively, a path-mean Jacobian is the average of all of the Jacobians along the line joining ω to $\omega^\star$. The convexity assumption in Assumption 2 ensures that the line integral joining any two points in Ω always exists. The Lipschitz assumption in Assumption 2 is only required for Section 4 and can be weakened to any condition that ensures the path-mean Jacobians exist for the remainder of the paper.

Our analysis in Section 4 proves that stability of TD and PFPE under decaying stepsizes is determined solely by the negative definiteness of the TD path-mean Jacobian $\bar{J}_{\text{TD}}(\omega, \omega^\star)$. In Section 5, we show for a non-diminishing stepsize regime that through suitable regularisation (which does not affect the TD fixed point), PFPE's stability can be determined *only* by α_l and k, for which stable values exists. As $\bar{H}(\omega, \omega^\star; \bar{\omega}_l)$ is the path-mean Hessian of the loss, convergence can be guaranteed under the same mild assumptions required to prove convergence of a stochastic gradient descent algorithm to minimise $\mathcal{L}(\omega;\bar{\omega}_l)$. This implies that PFPE can converge under regimes where TD will not as $\bar{J}_{\text{TD}}(\omega,\omega^\star)$ is positive definite.

3.3. Analysis of PFE

We now showcase the power of our Jacobian analysis by writing the PFE updates exactly in terms of $(\bar{\omega}_0-\omega^\star)$:

Theorem 1. *Under Assumption 2, the sequence of PFE updates $\bar{\omega}^\star_{l+1}\in\arg\inf_\omega\mathcal{L}(\omega,\bar{\omega}^\star_l)$ satisfy:*

$$\bar{\omega}^\star_l-\omega^\star \\ =\prod_{i=0}^{l-1}\left(\bar{H}(\bar{\omega}^\star_{i+1},\omega^\star;\bar{\omega}^\star_i)^{-1}\bar{J}_\delta(\bar{\omega}^\star_i,\omega^\star;\omega^\star)\right)(\bar{\omega}_0-\omega^\star).$$

We can use Theorem 1 to determine the stabllity of FPE updates. If $\sup_{\omega,\omega'\in\Omega}\left\|\bar{H}(\omega',\omega^\star;\omega)^{-1}\bar{J}_\delta(\omega,\omega^\star;\omega^\star)\right\|<1$ then the FPE updates are a contraction mapping and will converge to a fixed point under the Banach fixed-point theorem. We discuss the convergence of FPE under varying regularisation schemes in Section 5.1.

4. Asymptotic Analysis

We now study the behaviour of Equation (3) in the limit of $l\to\infty$. We introduce the standard Robbins-Munro condition for the decaying stepsizes that is a necessary condition to ensure convergence to a fixed point:

Assumption 3 (Robbins-Munro)**.** *Each α_l is a positive scalar with $\sum_{l=0}^\infty\alpha_l=\infty$ and $\sum_{l=0}^\infty\alpha_l^2<\infty$.*

Now we introduce a core necessary assumption to prove stability of PFPE with diminishing stepsizes:

Assumption 4 (TD Stability)**.** *There exists a region $\mathcal{X}_{TD}(\omega^\star)$ containing a fixed point $\omega^\star$ such that $\bar{J}_{TD}(\omega,\omega^\star)$ has strictly negative eigenvalues for all $\omega\in\mathcal{X}_{TD}(\omega^\star)$.*

The key insight from Assumption 4 is that the stability of PFPE under diminishing stepsizes is determined only by the eigenvalues of the single step path-mean Jacobian $\bar{J}_{\text{TD}}(\omega,\omega^\star)$, regardless of the value of k or α_l. Indeed, stochastic approximation can be shown to be provably divergent if this condition cannot be satisfied (Pemantle, 1990). From this perspective, if TD diverges then so will PFPE under *diminishing stepsizes*, hence the asymptotic stability of PFPE is independent of k and α_l, and, unlike updating under a two-timescale regime, introducing target parameters that are updated periodically every k timesteps does not improve asymptotic convergence properties under this analysis. Once Assumption 4 has been established, there are several approaches to prove convergence of the PFPE update under varying sampling conditions and projection assumptions. We follow the proof of (Vidyasagar, 2022), but discuss approaches that generalise our assumptions in Appendix D

Theorem 2. *Let Assumptions 1 to 4 hold. If there exists some fixed point $\omega^\star$ with region of contraction $\mathcal{X}_{TD}(\omega^\star)$ and timestep t such that $\bar{\omega}_l\in\mathcal{X}_{TD}(\omega^\star)$ for all $l\geq t$ the the sequence of target parameter updates in Equation* (2) *converge almost surely to $\omega^\star$.*

4.1. The Deadly Triad

We have established that it is not possible to prove convergence of PFPE under diminishing stepsizes if Assumption 4 does not hold. We now discuss how adherence to Assumption 4 formalises a phenomenon known as the *deadly triad* (Sutton & Barto, 2018) where it has been established that TD cannot be proved to converge when using function approximators in the off-policy setting. To control for the effect of nonlinear function approximation, we first investigate linear function approximators of the form $Q_\omega(s,a)=\phi(s,a)^\top\omega$ where $\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^n$ is a feature vector. Define the one-step lookahead distribution as: $P^\mu:=\mathbb{E}_{s\sim d,a\sim\mu(s)}\left[P(s,a)\right]$. Introducing the shorthand:

$$\Phi:=\mathbb{E}_{s\sim d,a\sim\mu(s)}[\phi(s,a)\phi(s,a)^\top],$$
$$\Phi':=\mathbb{E}_{s\sim d,a\sim\mu(s)}[\mathbb{E}_{s'\sim P^\mu,a'\sim\pi(s')}[\phi(s',a')]\phi(s,a)^\top],$$

we can derive the TD Jacobian as:

$$\bar{J}_{\text{TD}}(\omega,\omega^\star)=\gamma\Phi'-\Phi.$$

We now examine why the conditioning of $\bar{J}_{\text{TD}}(\omega,\omega^\star)$ explains this phenomenon.

Linear Function Approximation For linear function approximators, we show in Appendix A.1 that $\gamma\|Q_\omega\|_{P^\mu,\pi}<\|Q_\omega\|_{d,\mu}$ for all ω is a sufficient condition for $\gamma\Phi'-\Phi$ to have negative eigenvalues, thereby satisfying Assumption 4. This implies that the function approximator class remains non-expansive under the one-step lookahead distribution P^μ, thereby preventing the function approximator diverging as the Markov chain is traversed. This condition has been introduced previously in the fitted Q-iteration literature (Wang et al., 2020; 2021) as a “low distribution shift” assumption.

In the on-policy setting in an ergodic MDP, we can prove that there exists a stationary distribution d^π induced by following the target policy π, that is $\mu=\pi$. Moreover it is assumed that samples come from d^π; hence by the definition of ergodicity, the one-step lookahead distribution is the stationary distribution: $P^\pi=d^\pi$. It thus follows that $\gamma\|Q_\omega\|_{P^\pi,\pi}=\gamma\|Q_\omega\|_{d^\pi,\pi}<\|Q_\omega\|_{d^\pi,\pi}$ and hence Assumption 4 holds automatically for on-policy TD in an ergodic MDP, thereby establishing the convergence properties as a special case via Theorem 2.

For off-policy data, it is not possible to prove that $\gamma\|Q_\omega\|_{P^\mu,\pi} < \|Q_\omega\|_{d,\mu}$ holds without further assumptions on the sampling policy and MDP. In general, it is not possible to show that $\bar{J}_{\text{TD}}(\omega,\omega^\star)$ is negative definite in the off-policy case as the distribution shift may be too high: there exist counterexample MDPs where off-policy algorithms such as Q-learning provably diverge under linear function approximation (Williams & Baird, 1993; Baird, 1995a).

Nonlinear Function Approximation Even in an on-policy regime, we cannot prove convergence of TD when nonlinear function approximators such as neural networks are used. In these cases, the path-mean Jacobian may not have a closed form solution. However, it can be bounded by the following norm (see Appendix A.2):

$$\begin{aligned}&\sup \lambda\left(\bar{J}_{\text{TD}}(\omega,\omega^\star)\right)\\ &\quad\le \sup_\omega \sup \lambda\left(\mathbb{E}\left[(\mathcal{T}^\pi[Q_\omega]-Q_\omega)\nabla^2_\omega Q_\omega\right]\right.\\ &\qquad\qquad \left.+\mathbb{E}\left[\left(\gamma\mathbb{E}'[\nabla_\omega Q'_\omega]-\nabla_\omega Q_\omega\right)\nabla_\omega Q_\omega^\top\right]\right).\end{aligned}$$

Even making the same assumption as in Section 4.1 of sampling on-policy in an ergodic MDP to show that

$$\begin{aligned}&\omega^\top\mathbb{E}\left[\left(\gamma\mathbb{E}'[\nabla_\omega Q'_\omega]-\nabla_\omega Q_\omega\right)\nabla_\omega Q_\omega^\top\right]\omega\\ &\qquad\le(\gamma-1)\omega^\top\mathbb{E}\left[\nabla_\omega Q_\omega\nabla_\omega Q_\omega^\top\right]\omega<0,\end{aligned}$$

we cannot prove the negative definiteness of $\bar{J}_{\text{TD}}(\omega,\omega^\star)$ required to satisfy Assumption 4. This is because the matrix $\mathbb{E}\left[(\mathcal{T}^\pi[Q_\omega]-Q_\omega)\nabla^2_\omega Q_\omega\right]$ can be arbitrarily positive definite depending on the MDP and choice of function approximator. Indeed, there exist counterexample MDPs with provably divergent nonlinear function approximators when sampling on-policy (Tsitsiklis & Van Roy, 1997).

5. Non-asymptotic Analysis

Our asymptotic analysis in Section 4 shows that increasing k or adjusting α_l for PFPE does not affect the asymptotic strong convergence properties of the TD algorithm, implying that target networks do not stabilise TD if stepsizes tend to zero. We showed that the underlying reason for this was the deadly triad, which we formalised as adherence to Assumption 4. We now replace Assumption 4, that is $\bar{J}_{\text{TD}}(\omega,\omega^\star)$ is negative definite, with the assumption that FPE is stable:

Assumption 5 (FPE Stability)**.** *There exists a region $\mathcal{X}_{FPE}(\omega^\star)$ containing a fixed point $\omega^\star$ such that* $\sup_{\omega,\omega'\in\mathcal{X}_{FPE}(\omega^\star)}\left\|\bar{H}(\omega',\omega^\star;\omega)^{-1}\bar{J}_\delta(\omega,\omega^\star;\omega^\star)\right\|<1.$

5.1. Stabilising FPE

We now prove that Assumption 5 can always be satisfied using regularisation schemes that do not affect the TD fixed points. We introduce the following regularised TD vector:

$$\delta_{\text{Reg}}(\omega,\omega')=\delta(\omega,\omega')+\rho(\omega,\omega'), \tag{4}$$

where $\rho(\omega,\omega')$ is a regularisation term such that $\rho(\omega,\omega)=0$, thereby not changing the TD fixed point or TD update. As an example, $\rho(\omega',\omega')$ can contain powers of regularisation terms $M_{\text{Reg}}(\omega-\omega')$ in addition to combinations of $\delta(\omega',\omega)$ and $\delta(\omega,\omega')$ terms, where $\delta(\omega',\omega)$ is a TD vector with target and Q network parameters swapped. A simpler choice to ensure Assumption 5 holds is the $\rho(\omega,\omega')=-\eta\omega\mathbb{I}(\omega\neq\omega')-\mu(\omega-\omega')$, where $\mathbb{I}(\omega\neq\omega')$ is the identity I for any $\omega\neq\omega'$ and η and μ control the degree of regularisation. This scheme is equivalent to adding ℓ_2-regularisation to the loss $\mathcal{L}(\omega;\omega')$ whilst still ensuring that $\delta(\omega,\omega')$ is differentiable almost everywhere. We emphasise that $\delta_{\text{Reg}}(\bar{\omega}_l,\bar{\omega}_l)=\delta(\bar{\omega}_l,\bar{\omega}_l)$, leaving the TD update unchanged. In contrast, unless $\omega^\star$ is known a priori, introducing regularisation that modifies the TD update—as is done in (Zhang et al., 2021)—will affect the TD fixed points. We now prove that FPE can be stabilised using the regularised update by tuning η,μ:

Proposition 1. *Using the regularised TD vector:*

$$\delta_{Reg}(\omega,\omega')=\delta(\omega,\omega')-\eta\omega\mathbb{I}(\omega\neq\omega')-\mu(\omega-\omega')$$

the path-mean Jacobians are:

$$\begin{aligned}\bar{H}_{Reg}(\omega,\omega^\star;\bar{\omega}_l)&=\bar{H}(\omega,\omega^\star;\bar{\omega}_l)+(\mu+\eta)I,\\ \bar{J}_{\delta,Reg}(\omega,\omega^\star;\bar{\omega}_l)&=\bar{J}_\delta(\omega,\omega^\star;\bar{\omega}_l)+\mu I,\end{aligned}$$

Assumption 5 is satisfied if:

$$\sup_{\omega,\omega'\in\mathcal{X}_{FPE}(\omega^\star)}\left\|\bar{H}_{Reg}(\omega',\omega^\star;\omega)^{-1}\bar{J}_{\delta,Reg}(\omega,\omega^\star;\omega^\star)\right\|<1. \tag{5}$$

There exists a finite η,μ such that Equation (5) *holds.*

5.2. Convergence Analysis

By carrying out a non-asymptotic analysis, we now investigate how the deadly triad can be broken by PFPE using Equation (4) when stepsizes *do not tend to zero*. This leads to a formal understanding of how target parameters stabilise TD under stepsize regimes that are actually used in practice when classic TD methods fail. The foundation of our analysis is a condition function that can be used to determine the stability of the updates:

Definition 1 (Condition Function)**.** *For a subset $\mathcal{X}(\omega^\star)\subseteq\Omega$ with corresponding fixed point $\omega^\star\in\mathcal{X}(\omega^\star)$ such that $\omega_i\in\mathcal{X}(\omega^\star)$ for all $i\geq 0$, let*

$$\lambda_H^\star := \sup_{\omega,\omega',\omega''} \operatorname*{arg\,sup}_{\lambda'\in\lambda(\bar{H}(\omega,\omega';\omega''))} \left|1-\alpha_l\lambda'\right|,$$

$$\left\|\bar{J}_{FPE}^\star\right\| := \sup_{\omega,\omega'\in\mathcal{X}(\omega^\star)} \left\|\bar{H}(\omega',\omega^\star;\omega)^{-1}\bar{J}_\delta(\omega,\omega^\star;\omega')\right\|,$$

$$\left\|\bar{J}_{TD}^\star\right\| := \sup_{\omega\in\mathcal{X}(\omega^\star)} \|I+\alpha\bar{J}_{TD}(\omega,\omega^\star)\|,$$

and define the condition function as:

$$\mathcal{C}(\alpha_l,k) := \left|1-\alpha_l\lambda_H^\star\right|^{k-1}\left\|\bar{J}_{TD}^\star\right\| + \left(1+\left|1-\alpha_l\lambda_H^\star\right|^{k-1}\right)\left\|\bar{J}_{FPE}^\star\right\|. \quad (6)$$

The condition function depends on the maximal eigenvectors of the Jacobians introduced in Section 3.2, and so can still be used to analyse general nonlinear function approximators for which the path-mean Jacobians have no analytic solution. Using the condition function, we decompose the error at a given timestep into the effect of the expected update plus the error induced by variance of the update:

Theorem 3. *Define*

$$\sigma_k := \left(1-\left|1-\alpha_l\lambda_H^\star\right|^k\right)\frac{\sigma_\delta}{\lambda_H^\star},$$

Let Assumptions 1 and 2 hold, then:

$$\mathbb{E}\left[\|\bar{\omega}_{l+1}-\omega^\star\|\right] \le \mathcal{C}(\alpha_l,k)\mathbb{E}\left[\|\bar{\omega}_l-\omega^\star\|\right]+\alpha_l\sigma_k. \quad (7)$$

The effect of the expected update (the first term in Equation (7)) is bounded by the condition function, which depends both on data conditioning but critically, on both k and α_l as well and must diminish with increasing l to ensure convergence. Using this decomposition, we see convergence is guaranteed if the following assumption holds:

Assumption 6 (Contraction Region)**.** *We assume that $\mathcal{C}(\alpha,k)\le c<1$ over $\mathcal{X}_{FPE}(\omega^\star)$.*

allowing us to prove convergence of PFPE for stepsizes that don't tend to zero provided that updates remain in a region of contraction:

Corollary 3.1. *Let Assumptions 1, 2, 5 and 6 hold. For a fixed stepsize $\alpha_l=\alpha>0$,*

$$\mathbb{E}\left[\|\bar{\omega}_l-\omega^\star\|\right] \le \frac{\alpha\sigma_k}{1-c} + \exp(-l(1-c))\left(\|\bar{\omega}_0-\omega^\star\|-\frac{\sigma_k}{1-c}\right).$$

Corollary 3.1 is a key result of this work. Our result demonstrates geometric decay of errors in l, to a ball of fixed radius $\frac{\alpha\sigma_k}{1-c}$. This is analogous to related work in stochastic gradient descent (Bottou et al., 2018), and matches the intuition that, without decaying stepsize, variance in the updates means that convergence to a fixed point does not occur. Note that the radius of the ball which we converge to can be made arbitrarily small by decreasing α.

This supports the use of a hybrid approach, wherein a fixed step size is used until iterates are no longer improving and then reducing step size and repeating to decrease the radius of the ball of convergence whilst maintaining k as small as possible. In the remainder of this section, we explore the properties of the condition function to ensure the existence of a region of contraction satisfying Assumption 6.

5.3. Properties of PFPE Condition Function

We now investigate key properties of Equation (6) to understand how target parameters can lead to convergence when classic TD methods fail. If $\bar{J}_{\text{TD}}(\omega,\omega^\star)$ is positive definite, TD is provably divergent, however our analysis reveals that there are values of k and α_l for which PFPE does converge.

Property 1: Lower bound $\left\|\bar{J}_{\text{FPE}}^\star\right\| \le \mathcal{C}(\alpha_l,k)$.

We first investigate the conditions for which our choice of function approximators can never be used to prove convergence. Our condition function implies that we cannot prove convergence for any $\lambda_H^\star\le 0$ or $\lambda_H^\star\ge\frac{2}{\alpha_l}$ as repeated applications of $|1-\alpha_l\lambda_H^\star|^2$ do not reduce the effect the ill-conditioning of $\bar{J}_{\text{TD}}(\omega,\omega^\star)$. We formalise this in the following regularity assumption:

Assumption 7 (Eigenvalue Regularity Assumption)**.** *Given a region $\mathcal{X}\subseteq\Omega$, for all $\omega,\omega'\in\mathcal{X}$ there exists $0<\lambda_1^{\min}$ and $\lambda_1^{\max}<\infty$ such that $\lambda^{\min}\le\lambda(\nabla_\omega^2\mathcal{L}(\omega;\omega'))\le\lambda^{\max}$.*

We now propose two simple fixes to avoid this issue. Recall from Section 3.2 that $\lambda_H^\star$ is an eigenvalue of the Hessian of a loss. If $\lambda_H^\star$ was negative, this would imply that the Hessian is not positive semidefinite for all ω in the region of interest; hence we cannot prove convergence of stochastic gradient descent on the loss $\mathcal{L}(\omega;\bar{\omega}_l)$, let alone the full PFPE algorithm. To remedy this problem, the eigenvalues of the matrix can be increased using the regularisation introduced in Equation (4) without affecting the TD fixed point. However, if $\lambda_H^\star\ge\frac{2}{\alpha_l}$, then the conditioning of the Hessian matrix is ill-suited to the chosen step-size, and an easy remedy is to decrease α_l. Our bound shows that the condition function is lower bounded by $\|J_{\text{FPE}}^\star\|$, and so if Assumption 5 does not hold, then convergence of PFPE is not provable.

Property 2: Monotonicity For $|1-\alpha_l\lambda_H^\star|<1$, $\mathcal{C}(\alpha_l,k)\le\mathcal{C}(\alpha_l,k')$ for $k\le k'$.

The monotonicity property ensures that $|1-\alpha_l\lambda_H^\star|<1$ defines the interval of Hessian eigenvalues for which there is a regime in which we can increase k in order to ensure PFPE updates are a contraction mapping. This suggests that a key role of the target network is to help mitigate the effects of the ill-conditioning of the TD Jacobian when using fixed step sizes. We now investigate how decreasing stepsizes and increasing the number of PFPE steps affect the conditioning of PFPE, which validates this hypothesis.

Property 3: Limits For any $k < \infty$, $\lim_{\alpha_l \to 0} \mathcal{C}(\alpha_l, k) = \left\|\bar{J}^\star_{\text{TD}}\right\| + 2\left\|\bar{J}^\star_{\text{FPE}}\right\|$. For any $0 < \alpha_l < \frac{2}{\lambda^\star_H}$, $\lim_{k\to\infty} \mathcal{C}(\alpha_l, k) = \left\|\bar{J}^\star_{\text{FPE}}\right\|$.

The first limit illustrates the effects of a diminishing stepsize sequence, confirming our bound is consistent with the results of the previous section that increasing k does not improve the convergence properties of PFPE if stepsizes tend to zero and PFPE only stabilises TD for $0 < \alpha_l$. By taking the limit $k \to \infty$, we compliment our monotonicity result, obtaining a bound for how much we can improve on the stability of TD by increasing k. As expected, in the limit of $k \to \infty$, the condition function tends to $\|J^\star_{\text{FPE}}\|$. Through this insight, we interpret PFPE as mixing FPE and TD updates according the coefficient $|1 - \alpha_l \lambda^\star_H|^{k-1}$: for $k = 1$, PFPE uses only TD updates and in the limit $k \to \infty$, PFPE recovers the FPE update.

5.4. Breaking the Deadly Triad

We now combine all properties presented in this section into our main result, proving that through suitable regularisation and choice of α_l and k, PFPE breaks TD's deadly triad described in Section 4.1:

Theorem 4. *Let Assumption 7 hold over $\mathcal{X}_{FPE}(\omega^\star)$ from Definition 1. For any $\frac{1}{\alpha_l} > \frac{\lambda_1^{\min} + \lambda_1^{\max}}{2}$ such that $\alpha_l > 0$, any*

$$k > 1 + \frac{\log(1 - \|\bar{J}^\star_{FPE}\|) - \log(\|\bar{J}^\star_{TD}\| + \|\bar{J}^\star_{FPE}\|)}{\log(1 - \alpha\lambda^{min})},$$

ensures that $\mathcal{X}_{FPE}(\omega^\star)$ is a region of contraction satisfying Assumption 6.

Theorem 4 demonstrates that appropriate values of α_l and k can be found by treating them as hyperparameters, decreasing α_l and increasing k until the algorithm is stable, reducing the conditions needed to prove convergence of PFPE to those of proving convergence of stochastic gradient descent on the loss $\mathcal{L}(\omega; \bar{\omega}_l)$. The key insight of Theorem 4 is that even when TD is unstable due to $1 < \|I + \alpha_l \bar{J}_{\text{TD}}(\bar{\omega}_l, \omega^\star)\|$, there exists a finite k such that $\mathcal{C}(\alpha_l, k) < 1$ and hence PFPE is stable. We illustrate this phenomenon with a sketch in Figure 1, demonstrating that increasing k ensures PFPE is provably convergent in regimes where TD cannot be proved to converge.

The key insight of our analysis is that, unlike in TD where stability can only be proved if the matrix $\bar{J}_\delta(\omega, \omega^\star; \omega^\star) - \bar{H}(\omega, \omega^\star; \omega)$ is negative definite, with suitable regularisation, the stability of PFPE can be determined solely by tuning α_l and k, regardless of the MDP, sampling regime, or function approximator, thereby breaking the deadly triad. The choice of α_l and k thus becomes a trade-off between maintaining a fast rate of convergence and reducing the residual variance $(\alpha_l \sigma_k)^2$ in Equation (7).

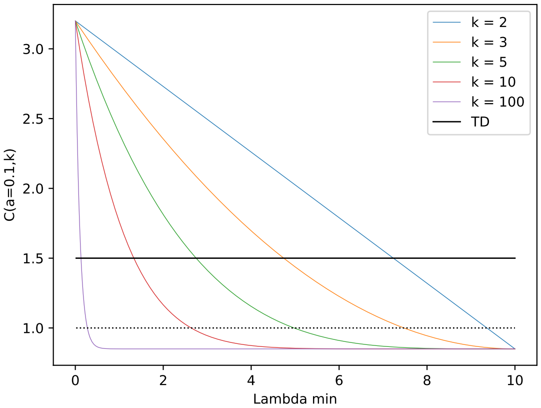

Figure 1: We plot $\mathcal{C}(\alpha = 0.1, k)$ for $\|\bar{J}^\star_{\text{FPE}}\| = 0.85$ and $\|\bar{J}^\star_{\text{TD}}\| \leq 1.5$ with increasing k as a function of $\lambda_{\min}$.

6. Related Work

Our work furthers the analysis of TD, FPE, and target-network based methods. In this section we provide a brief overview of previous investigations of these algorithms.

Fitted Policy Evaluation FPE is a relatively well understood class of RL algorithms from a theoretical perspective. Nedić & Bertsekas (2003) analyse the convergence of the Least-Squares Policy Evaluation (LSPE) of Bertsekas & Ioffe (1996) in an on-policy, linear function approximation setting. Analysis of LSPE shows that learning with constant step size leads to theoretical and empirical gains compared to TD and LSPE with decaying step sizes (Bertsekas et al., 2004), which mirrors our conclusions in Section 5.4.

In the context of fitted methods applied to off-policy and control problems, Munos & Szepesvári (2008) prove generalisation properties of Fitted Q Iteration (Ernst et al., 2005) for general function classes under assumptions of low projection error and limited data distribution shift. Le et al. (2019) coin the term FPE, and formalise the algorithm for general function approximators, with theoretical results under similar assumptions to Munos & Szepesvári (2008).

Theory of TD Previous results concerning convergence rates of classic TD methods largely argue that the Bellman operator is a contraction, and thus most focus on linear function approximation. Tsitsiklis & Van Roy (1997) first proved convergence of linear, on-policy TD, arguing that the projected Bellman operator in this setting is a contraction. This corresponds to a special case of Assumption 4. Dalal et al. (2017) give the first finite time bounds for linear TD(0), under an i.i.d. data model similar to the one that we use here. Bhandari et al. (2018) provide bounds for linear TD in both the i.i.d. data setting and a correlated data setting, through analogy with SGD. Srikant & Ying (2019) approach the problem from the perspective of Ordinary Differential Equations (ODE) analysis, bounding the divergence of a Lyapunov function from the limiting point of the ODE that arises from the TD update scheme.

Analysis of Target Networks Existing analysis of the theoretical properties of target networks are limited, usually involving algorithmic changes or restrictive assumptions. Yang et al. (2019) show convergence of a Q-learning approach using a target network that is updated using Polyak averaging with nonlinear function approximation. However their analysis–which makes use of two-timescale analysis–requires a projection step to limit the magnitude of parameters. Carvalho et al. (2020) show convergence of a related method using two-timescale analysis, though their target network update differs significantly from those used in practice. Zhang et al. (2021) analyse the use of target networks with linear function approximation, but require projection steps on both the target network and value parameters. Lee & He (2019) provide finite-iteration bounds, but are limited to on-policy data, linear function approximation, and near-perfect fitting to the target network between updates. Fan et al. (2020) analyse the use of target networks for deep Q-learning (Mnih et al., 2015) with the simplifying assumption that they are performing some form of Fitted Q Iteration.

None of these efforts yield finite time bounds with target networks, nor do any match the policy evaluation methods used in practice as well as the PFPE analysis studied here. Furthermore, our use of a single target network update, rather than independent target and value updates leads to simpler bounds without the need for a two-timescale analysis.

GTD and TDC Methods While not directly related to PFPE or the use of target networks, GTD-style approaches (Sutton et al., 2008; 2009; Maei et al., 2009) also lead to convergent, TD-style algorithms, even with off-policy sampling or nonlinear function approximation. These methods maintain a second set of parameters which must be optimised at a faster timescale than the value parameters. However, these approaches are commonly found to be ineffective and not used in practice due to the difficulty in tuning the rate of second timescale (see, e.g. Fellows et al. (2021)), and potentially additional variance introduced by the second set of parameters (Ghiassian et al., 2020).

Improving Conditioning of TD Methods Previous work concerning conditioning of TD methods has been largely concerned with approximation of preconditioning approaches to iterative-methods (Saad, 2003). The first such approach was focused on preconditioning of on-policy, linear, least-squares forms of TD (Yao & Liu, 2008). Chen et al. (2020); Romoff et al. (2020) adapt this approach for nonlinear function approximation, though their results are still on-policy. Our work, on the other hand, demonstrates that use of the target network, alongside fixed step sizes, changes the form of parameter iterates to ameliorate the poor conditioning that occurs when directly applying TD or fitted methods, even in off-policy settings.

7. Experiments

We proceed to empirical investigation of our bounds. First, we demonstrate that the use of an infrequently updated target network leads to convergence of off-policy evaluation on the Baird's notorious counterexample. Then, we evaluate the effect of a speculative modified update rule in the Cartpole-v0 "gym" environment (Brockman et al., 2016). Additional implementation details for both experiments can be found in Appendix C.

7.1. Baird's Counterexample

In this experiment, we demonstrate the practicality of our core claim–that for sufficiently high k and low enough α, PFPE will not diverge, even under conditions that TD does. To do so, we evaluate the use of target networks with varying update frequencies on the well known off-policy counterexample due to Baird (1995b).

In this environment, depicted in Appendix C, rewards are zero everywhere, transitions are deterministic, and the true solution lies within the linear function approximation class that we make use of. The behaviour policy is set such that all states are sampled with uniform probability. The target policy, however, always transitions to a specific state, and remains there. Due to undersampling of this absorbing state, conventional TD policy evaluation diverges, demonstrating that even in simple environments, TD can be unstable when applied off policy with function approximation.

We report the stepwise (fitted) error in Figure 2 across different values of k, for fixed step size $\alpha = 0.01$, and fixed discount factor $\gamma = 0.99$. We see that with $k = 1$–which is equivalent to using TD with fixed step sizes–our parameters diverge. Likewise, if k is set to 5 or 10, we are unable to overcome the conditioning of the TD Jacobian and diverge, albeit at a slower rate. Once we take $k \geq 50$, however, conditioning has improved enough to lead to convergence.

This supports our theoretical conclusion: that PFPE can be used to improve the convergence conditions of TD.

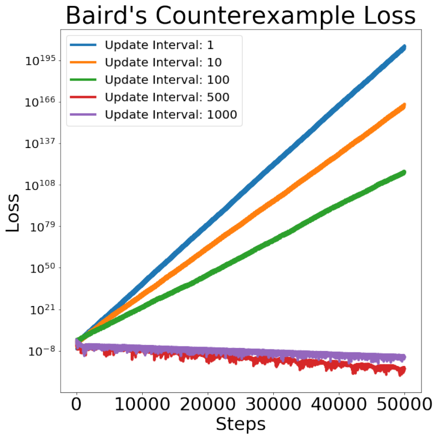

Figure 2: Experiment on Baird's counterexample. Decreasing the frequency of target network updates improves conditioning and leads to convergence of PFPE for suitable choices of hyperparameters.

7.2. Cartpole Experiment

One important insight of our analysis is that we can view the entire optimisation process as a sequence of updates to the target network only. This suggests investigation into alternative forms or acceleration of target network updates. Inspired by the use of optimisation methods with momentum in RL settings (Sarigül & Avci, 2018; Haarnoja et al., 2018), we investigate the effects of a target network that is updated using momentum.

Unlike the standard periodic target network update in Equation (2), we postulate that there may be settings in which a periodic update with momentum may accelerate or stabilise convergence. This update works as follows:

$$\bar{\omega} = \begin{cases} (1-\mu)\omega_i + \mu(\omega_{i-k} - \omega_{i-2k}), & i \bmod k = 0, \\ \bar{\omega}, & \text{otherwise.} \end{cases}$$

We investigate the effects of this momentum update on the Cartpole domain. For this experiment, we use control results in which the policy is continuously learned. This is because control problems are inherently off-policy, and induce additional instability, and thus benefit from faster and more stable convergence of values. We implement the standard DQN (Mnih et al., 2015) algorithm, with our modified target network update in order to examine its effect. The results are shown in Figure 3. Our proposed update indeed leads to improved learning and stability, at least for the hyperparameter ranges tested, suggesting that the momentum update has merit. As a result, we propose investigation of more sophisticated target network update schemes as an avenue for future research.

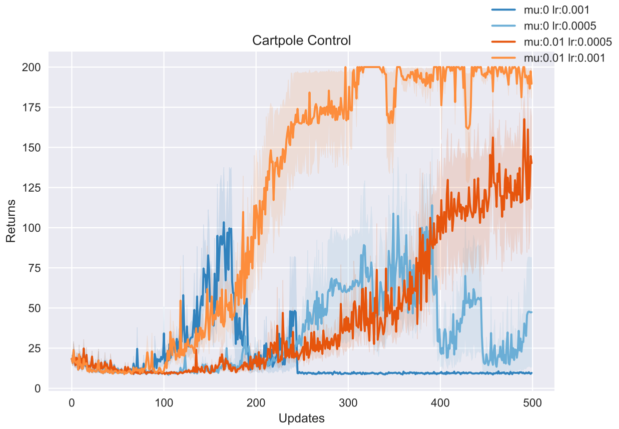

Figure 3: Cartpole Experiment. The agent with the momentum update is significantly more stable and able to consistently learn, while without the modified update, learning collapses.

8. Conclusions

This work analysed the use of target networks through the formulation of a novel class of TD updates, which we refer to as PFPE. These updates generalise traditional TD(0) and fitted policy evaluation methods. Our analysis contributes asymptotic and finite time bounds without additional restrictive assumptions or significant changes to the algorithms used in practice. In our main result, we uncovered novel insight as to when and how target networks are useful: provided step-sizes don't tend to zero and FPE is stable, there always exists a finite number of update steps k and non-zero upper bound over stepsizes such that PFPE can improve conditioning to ensure learning is stable when classic TD methods fail. Our focus on the target network update as the object of concern in terms of optimisation suggests that novel, accelerated methods for updating target networks may help speed up and stabilise learning. Our initial experiments support this notion. Moreover, our analysis reveals that regularisation may be key to determining the stability of PFPE, opening a promising avenue for future research.

Acknowledgements

Mattie Fellows is funded by a generous grant from Waymo. We would like to thank Valentin Thomas for providing a helpful discussion.

References

Allasonniere, S., Kuhn, E., and Trouve, A. Construction of bayesian deformable models via a stochastic approximation algorithm: A convergence study. *Bernoulli*, 16(3):641–678, 2010. ISSN 13507265. URL `http://www.jstor.org/stable/25735007`. D

Andradottir, S. A projected stochastic approximation algorithm. In *1991 Winter Simulation Conference Proceedings.*, pp. 954–957, 1991. doi: 10.1109/WSC.1991.185710. D

Andrieu, C., Moulines, E., and Priouret, P. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44(1):283–312, 2005. doi: 10.1137/S0363012902417267. URL https://doi.org/10.1137/S0363012902417267. D

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In Prieditis, A. and Russell, S. J. (eds.), *Proceedings of the Twelfth International Conference on Machine Learning (ICML 1995)*, pp. 30–37, San Francisco, CA, USA, 1995a. Morgan Kauffman. ISBN 1-55860-377-8. URL http://leemon.com/papers/1995b.pdf. 4.1

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Elsevier, 1995b. 7.1

Bertsekas, D. P. and Ioffe, S. Temporal differences-based policy iteration and applications in neuro-dynamic programming. *Lab. for Info. and Decision Systems Report LIDS-P-2349, MIT, Cambridge, MA*, 14, 1996. 6

Bertsekas, D. P., Borkar, V. S., and Nedic, A. Improved temporal difference methods with linear function approximation. *Learning and Approximate Dynamic Programming*, pp. 231–255, 2004. 6

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation, 2018. 6

Borkar, V. S. and Meyn, S. P. The o.d. e. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469, jan 2000. ISSN 0363-0129. doi: 10.1137/S0363012997331639. URL https://doi.org/10.1137/S0363012997331639. B.2, 2, 2

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018. 5.2

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016. 7

Brooms, A. C. Stochastic approximation and recursive algorithms with applications, 2nd edn by h. j. kushner and g. g. yin. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):654–654, 2006. doi: https://doi.org/10.1111/j.1467-985X.2006.00430_6.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2006.00430_6.x. D

Carvalho, D., Melo, F. S., and Santos, P. A new convergent variant of q-learning with linear function approximation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19412–19421. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e1696007be4eefb81b1a1d39ce48681b-Paper.pdf. 6

Chen, S., Devraj, A. M., Lu, F., Busic, A., and Meyn, S. Zap q-learning with nonlinear function approximation. *Advances in Neural Information Processing Systems*, 33:16879–16890, 2020. 6

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analysis for td (0) with linear function approximation. *arXiv preprint arXiv:1704.01161*, 2017. 6

Debavelaere, V., Durrleman, S., and Allassonnière, S. On the convergence of stochastic approximations under a subgeometric ergodic Markov dynamic. *Electronic Journal of Statistics*, 15(1):1583 – 1609, 2021. doi: 10.1214/21-EJS1827. URL https://doi.org/10.1214/21-EJS1827. D

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005. 6

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning, 2020. 1, 6

Fellows, M., Hartikainen, K., and Whiteson, S. Bayesian bellman operators. *Advances in Neural Information Processing Systems*, 34:13641–13656, 2021. 6

Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html. 1

Ghiassian, S., Patterson, A., Garg, S., Gupta, D., White, A., and White, M. Gradient temporal-difference learning with regularized corrections. In *International Conference on Machine Learning*, pp. 3524–3534. PMLR, 2020. 6

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of* *the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/haarnoja17a.html`. 1

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/haarnoja18b.html`. 1, 7.2

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019. 1, 6

Lee, D. and He, N. Target-based temporal difference learning, 2019. 1, 6

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In Bengio, Y. and LeCun, Y. (eds.), *ICLR*, 2016. 1

Maei, H., Szepesvari, C., Bhatnagar, S., Precup, D., Silver, D., and Sutton, R. S. Convergent temporal-difference learning with arbitrary smooth function approximation. *Advances in neural information processing systems*, 22, 2009. 6

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. 2013. URL `http://arxiv.org/abs/1312.5602`. cite arxiv:1312.5602Comment: NIPS Deep Learning Workshop 2013. 1

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015. 1, 6, 7.2

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008. 6

Nedić, A. and Bertsekas, D. P. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1):79–110, 2003. 6

Pemantle, R. Nonconvergence to Unstable Points in Urn Models and Stochastic Approximations. *The Annals of Probability*, 18(2):698 – 712, 1990. doi: 10.1214/aop/1176990853. URL `https://doi.org/10.1214/aop/1176990853`. 4

Polyak, B. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4:1–17, 12 1964. doi: 10.1016/0041-5553(64)90137-5. 1

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014. 2.1

Romoff, J., Henderson, P., Kanaa, D., Bengio, E., Touati, A., Bacon, P.-L., and Pineau, J. Tdprop: Does jacobi preconditioning help temporal difference learning? *arXiv preprint arXiv:2007.02786*, 2020. 6

Saad, Y. *Iterative methods for sparse linear systems*. SIAM, 2003. 6

Sarigül, M. and Avci, M. Performance comparison of different momentum techniques on deep reinforcement learning. *Journal of Information and Telecommunication*, 2 (2):205–216, 2018. 7.2

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation and td learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019. 6

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, Aug 1988. ISSN 1573-0565. doi: 10.1007/BF00115009. 1, 2.1

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL `http://incompleteideas.net/book/the-book-2nd.html`. 1, 4.1

Sutton, R. S., Szepesvári, C., and Maei, H. R. A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in neural information processing systems*, 21(21):1609–1616, 2008. 6

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., and Wiewiora, E. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, pp. 993–1000, 2009. 6

Tsitsiklis, J. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874. 3.1, 4.1, 6

Vidyasagar, M. Convergence of stochastic approximation via martingale and converse lyapunov methods, 2022. URL `https://arxiv.org/abs/2205.01303`. 4, 2

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation?, 2020. 4.1

Wang, R., Wu, Y., Salakhutdinov, R., and Kakade, S. M. Instabilities of offline rl with pre-trained neural representation, 2021. 4.1

Williams, R. J. and Baird, L. C. Analysis of some incremental variants of policy iteration: First steps toward understanding actor-cr. 1993. 4.1

Yang, Z., Fu, Z., Zhang, K., and Wang, Z. Convergent reinforcement learning with function approximation: A bilevel optimization perspective, 2019. URL https://openreview.net/forum?id=ryfcCo0ctQ. 6

Yao, H. and Liu, Z.-Q. Preconditioned temporal difference learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 1208–1215, 2008. 6

Zhang, S., Yao, H., and Whiteson, S. Breaking the deadly triad with a target network, 2021. 1, 5.1, 6

A. Derivations

A.1. Derivation of Assumption 4 from low distributional shift

Starting from Assumption 4 and the definition of negative definiteness, we need to show:

$$\omega^\top(\gamma\Phi'-\Phi)\omega<0,$$

whenever $\gamma\|Q_\omega\|_{P^\mu,\pi}<\|Q_\omega\|_{d,\mu}$, for all ω. Investigating the first term by expanding the expectations we see:

$$\begin{aligned}\gamma\omega^\top\Phi'\omega&=\gamma\mathbb{E}_{s\sim d,a\sim\pi(s)}\left[\omega^\top\phi(s,a)\mathbb{E}_{s'\sim P(s,a),a'\sim\pi(s')}\left[\phi(s',a')^\top\omega\right]\right],\\&=\gamma\mathbb{E}_{d,\pi,P^\mu}\left[\omega^\top\phi(s,a)\phi(s',a')^\top\omega\right],\\&\le\gamma\sqrt{\mathbb{E}_{d,\pi,P^\mu}\left[(\phi(s,a)^\top\omega)^2\right]\mathbb{E}_{d,\pi,P^\mu}\left[(\phi(s',a')^\top\omega)^2\right]},\\&\le\gamma\sqrt{\mathbb{E}_{d,\pi}\left[(\phi(s,a)^\top\omega)^2\right]\mathbb{E}_{d,\pi,P^\mu}\left[(\phi(s',a')^\top\omega)^2\right]},\\&\le\gamma\|Q_\omega\|_d\|Q_\omega\|_{P^\mu,\pi}.\end{aligned}$$

This allows us to apply our assumption:

$$\omega^\top(\gamma\Phi'-\Phi)\omega\le\gamma\|Q_\omega\|_{P^\mu,\pi}\|Q_\omega\|_d-\|Q_\omega\|_d^2\le\gamma\|Q_\omega\|_d^2-\|Q_\omega\|_{d,\mu}^2<0.$$

A.2. Nonlinear Jacobian Analysis

We start by bounding the maximum eigenvalue:

$$\begin{aligned}\sup\lambda\left(\bar{J}_{\text{TD}}(\omega,\omega^\star)\right)&=\sup_\omega\frac{\omega^\top\bar{J}_{\text{TD}}(\omega,\omega^\star)\omega}{\omega^\top\omega},\\&=\sup_\omega\int_0^1\frac{\omega^\top J_{\text{TD}}(\omega'-t(\omega'-\omega^\star))\omega}{\omega\omega^\top}dt,\\&\le\int_0^1\sup_\omega\frac{\omega^\top J_{\text{TD}}(\omega'-t(\omega'-\omega^\star))\omega}{\omega\omega^\top}dt,\\&\le\int_0^1\sup_{t\in[0,1]}\sup_\omega\frac{\omega^\top J_{\text{TD}}(\omega'-t(\omega'-\omega^\star))\omega}{\omega\omega^\top}dt,\\&=\sup_{t\in[0,1]}\sup_\omega\frac{\omega^\top J_{\text{TD}}(\omega'-t(\omega'-\omega^\star))\omega}{\omega\omega^\top}\underbrace{\int_0^1dt}_{=1},\\&\le\sup_{\omega'}\sup_\omega\frac{\omega^\top J_{\text{TD}}(\omega'-t(\omega'-\omega^\star))\omega}{\omega\omega^\top},\\&=\sup_\omega\sup\lambda\left(J_{\text{TD}}(\omega,\omega^\star)\right).\end{aligned}$$

We now substitute for the definition of the TD Jacobian, yielding:

$$\begin{aligned}J_{\text{TD}}(\omega,\omega^\star)&=\nabla_\omega\delta(\omega,\omega),\\&=\nabla_\omega\mathbb{E}_{\varsigma\sim P_\varsigma}\left[(r+\gamma Q_\omega(s',a')-Q_\omega(s,a))\nabla_\omega Q_\omega(s,a)\right],\\&=\mathbb{E}_{\varsigma\sim P_\varsigma}\left[(\gamma\nabla_\omega Q_\omega(s',a')-\nabla_\omega Q_\omega(s,a))\nabla_\omega Q_\omega(s,a)+(r+\gamma Q_\omega(s',a')-Q_\omega(s,a))\nabla_\omega^2Q_\omega(s,a)\right],\\&=\mathbb{E}_{\varsigma\sim P_\varsigma}\left[(\gamma\nabla_\omega Q_\omega(s',a')-\nabla_\omega Q_\omega(s,a))\nabla_\omega Q_\omega(s,a)+((\mathcal{T}^\pi[Q_\omega](s,a)-Q_\omega(s,a))\nabla_\omega^2Q_\omega(s,a)\right],\end{aligned}$$

as required.

B. Proofs

B.1. FPE Analysis

Lemma 1. *Under Assumption 2, the FPE update $\bar{\omega}_{l+1}\in\arg\inf_\omega\mathcal{L}(\omega,\bar{\omega}_l)$ satisfies:*

$$\bar{\omega}_l^\star-\omega^\star=\bar{H}(\bar{\omega}_l^\star,\omega^\star;\bar{\omega}_l)^{-1}\bar{J}_\delta(\bar{\omega}_l,\omega^\star;\omega^\star),\tag{8}$$

Proof. Given $\bar{\omega}_l$, the FPE fixed point $\bar{\omega}_l^\star$ must be an element of the set:

$$\bar{\omega}_l^\star \in \{\omega | \delta(\omega, \bar{\omega}_l) = 0\},$$

which we use to derive a stability condition for the projection operator:

$$\begin{aligned}
&\delta(\bar{\omega}_l^\star, \bar{\omega}_l) = \delta(\omega^\star, \omega^\star) = 0\\
\implies &\delta(\bar{\omega}_l^\star, \bar{\omega}_l) - \delta(\omega^\star, \bar{\omega}_l) = \delta(\omega^\star, \omega^\star) - \delta(\omega^\star, \bar{\omega}_l).
\end{aligned}$$

Let $\ell_1(t) \coloneqq \bar{\omega}_l^\star - t(\bar{\omega}_l^\star - \omega^\star)$ and $\ell_2(t) \coloneqq \bar{\omega}_l - t(\bar{\omega}_l - \omega^\star)$. We introduce the notation:

$$\delta_1(t, \bar{\omega}_l) \coloneqq \delta(\ell_1(t), \bar{\omega}_l), \quad \delta_2(t, \omega^\star) \coloneqq \delta(\omega^\star, \ell_2(t)).$$

We observe that $\delta_1(0, \bar{\omega}_l) = \delta(\bar{\omega}_l^\star, \bar{\omega}_l)$ and $\delta_1(1, \bar{\omega}_l) = \delta(\omega^\star, \bar{\omega}_l)$, and $\delta_2(0, \omega^\star) = \delta(\omega^\star, \bar{\omega}_l)$ and $\delta_2(1, \omega^\star) = \delta(\omega^\star, \omega^\star)$. From the fundamental theorem of calculus and Assumption 2, it follows:

$$\begin{aligned}
&\delta_1(0, \bar{\omega}_l) - \delta_1(1, \bar{\omega}_l) = \delta_2(1, \omega^\star) - \delta_2(0, \omega^\star),\\
\implies &-\int_0^1 \partial_t \delta(\omega = \ell_1(t), \bar{\omega}_l) dt = \int_0^1 \partial_t \delta(\omega^\star, \omega = \ell_2(t)) dt,\\
\implies &\int_0^1 \nabla_\omega \delta(\omega = \ell_1(t), \bar{\omega}_l)(\bar{\omega}_l^\star - \omega^\star) dt = -\int_0^1 \nabla_\omega \delta(\omega^\star, \omega = \ell_2(t))(\bar{\omega}_l - \omega^\star) dt,\\
\implies &-\int_0^1 \nabla_\omega^2 \mathcal{L}(\omega = \ell_1(t); \bar{\omega}_l)(\bar{\omega}_l^\star - \omega^\star)) dt = -\int_0^1 \nabla_\omega \delta(\omega^\star, \omega = \ell_2(t))(\bar{\omega}_l - \omega^\star) dt,\\
\implies &\int_0^1 \nabla_\omega^2 \mathcal{L}(\omega = \ell_1(t); \bar{\omega}_l) dt (\bar{\omega}_l^\star - \omega^\star) = \int_0^1 \nabla_\omega \delta(\omega^\star, \omega = \ell_2(t)) dt (\bar{\omega}_l - \omega^\star),\\
\implies &\bar{H}(\bar{\omega}_l^\star, \omega^\star; \bar{\omega}_l)(\bar{\omega}_l^\star - \omega^\star) = \bar{J}_\delta(\bar{\omega}_l, \omega^\star; \omega^\star)(\bar{\omega}_l - \omega^\star),\\
\implies &(\bar{\omega}_l^\star - \omega^\star) = \bar{H}(\bar{\omega}_l^\star, \omega^\star; \bar{\omega}_l)^{-1} \bar{J}_\delta(\bar{\omega}_l, \omega^\star; \omega^\star),
\end{aligned}$$

as required. □

Theorem 1. *Under Assumption 2, the sequence of FPE updates $\bar{\omega}_{l+1}^\star \in \arg\inf_\omega \mathcal{L}(\omega, \bar{\omega}_l^\star)$ satisfy:*

$$\bar{\omega}_l^\star - \omega^\star = \prod_{i=0}^{l-1} \left(\bar{H}(\bar{\omega}_{i+1}^\star, \omega^\star; \bar{\omega}_i^\star)^{-1} \bar{J}_\delta(\bar{\omega}_i^\star, \omega^\star; \omega^\star) \right) (\bar{\omega}_0 - \omega^\star).$$

Proof. From Equation (8) of Lemma 1, it follows:

$$\bar{\omega}_{i+1}^\star - \omega^\star = \bar{H}(\bar{\omega}_{i+1}^\star, \omega^\star; \bar{\omega}_i^\star)^{-1} \bar{J}_\delta(\bar{\omega}_i^\star, \omega^\star; \omega^\star)(\bar{\omega}_i^\star - \omega^\star).$$

Recursively applying the result l times, our result follows immediately. □

B.2. Asymptotic Analysis

For this section, we define a Martingale difference sequence that captures the behaviour of our updates. Let $\{\omega_i\}_{i=0}^k$ denote the intermediate function approximation parameters between target parameter updates $\bar{\omega}_{l+1}$ and $\bar{\omega}_l$, with $\omega_0 = \bar{\omega}_l$ and

$\omega_k = \bar{\omega}_{l+1}$. We start by writing our target parameter updates as:

$$\begin{aligned}
\omega_1 &= \bar{\omega}_l + \alpha_l \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_0),\\
\omega_2 &= \omega_1 + \alpha_l \delta(\omega_1, \bar{\omega}_l, \varsigma_1),\\
&= \bar{\omega}_l + \alpha_l \left(\delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_0) + \delta(\bar{\omega}_l + \alpha_l \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_0), \bar{\omega}_l, \varsigma_1)\right),\\
\omega_3 &= \omega_2 + \alpha_l \delta(\omega_2, \bar{\omega}_l, \varsigma_2),\\
&= \bar{\omega}_l + \alpha_l \left(\delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_0) + \delta(\bar{\omega}_l + \alpha_l \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_0), \bar{\omega}_l, \varsigma_1)\right)\\
&\qquad + \alpha_l (\delta(\bar{\omega}_l + \alpha_l \left(\delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_0) + \delta(\bar{\omega}_l + \alpha_l \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_0), \bar{\omega}_l, \varsigma_1)\right), \bar{\omega}_l, \varsigma_2),\\
&\vdots\\
\omega_k &= \bar{\omega}_l + \alpha_l \sum_{i=0}^{k-1} \delta(\bar{\omega}_l + \alpha_l h_i(\bar{\omega}_l, \mathcal{D}, \alpha_l), \bar{\omega}_l, \varsigma_i),\\
&= \bar{\omega}_l + \alpha_l h_k(\bar{\omega}_l, \mathcal{D}, \alpha_l),
\end{aligned}$$

where we define $h_i(\bar{\omega}_l, \mathcal{D}, \alpha_l)$ recursively as:

$$h_i(\bar{\omega}_l, \mathcal{D}, \alpha_l) \coloneqq \sum_{j=0}^{i-1} \delta(\bar{\omega}_l + \alpha_l h_j(\bar{\omega}_l, \mathcal{D}, \alpha_l), \bar{\omega}_l, \varsigma_j).$$

and remark that $h_0(\bar{\omega}_l, \mathcal{D}, \alpha_l) = 0$ trivially. We write our target parameters updates as:

$$\bar{\omega}_{l+1} = \omega_k = \bar{\omega}_l + \alpha_l \left(k\delta(\bar{\omega}_l, \bar{\omega}_l) + \mathcal{M}_{l+1} + \varepsilon_{l+1}\right),$$

where

$$\varepsilon_{l+1} \coloneqq h_k(\bar{\omega}_l, \mathcal{D}_l, \alpha_l) - \sum_{i=0}^{k-1} \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_i),$$

and $\mathcal{M}_{l+1}$ defines the Martingale sequence:

$$\mathcal{M}_{l+1} \coloneqq \sum_{i=0}^{k-1} \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_i) - k\delta(\bar{\omega}_l, \bar{\omega}_l)$$

In this section, we demonstrate that the proof of Borkar & Meyn (2000, Theorem 2.2) can be adapted to account for the additional term ε_{l+1} that arises due to the use of target networks in the updates. Lemma 2 demonstrates that as stepsizes tend to zero, the effect of ϵ_{l+1} becomes negligible, hence the inclusion of ε_{l+1} negligible to our analysis of the underlying ODE defined by the TD updates.

Lemma 2. *Let* $\nu_{n,n+m} \coloneqq \sum_{l=n}^{m+n-1} \alpha_l \epsilon_{l+1}$ *for* $m \geq 1$. *Under Assumptions 1 to 3,* $\lim_{n\to\infty} \sup_m \|\nu_{n,n+m}\| = 0$ *almost surely.*

Proof. We start by bounding each $\|\epsilon_{i+1}\|$ using the the Lipschitzness of δ from Assumption 2:

$$\begin{aligned}
\|\epsilon_{l+1}\| &= \left\|\sum_{i=0}^{k-1} \left(\delta(\bar{\omega}_l + \alpha_l h_i(\bar{\omega}_l, \mathcal{D}, \alpha_l), \bar{\omega}_l, \varsigma_i) - \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_i)\right)\right\|,\\
&\leq \sum_{i=0}^{k-1} \left\|\delta(\bar{\omega}_l + \alpha_l h_i(\bar{\omega}_l, \mathcal{D}, \alpha_l), \bar{\omega}_l, \varsigma_i) - \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_i)\right\|,\\
&\leq \sum_{i=0}^{k-1} L \left\|\bar{\omega}_l + \alpha_l h_i(\bar{\omega}_l, \mathcal{D}, \alpha_l) - \bar{\omega}_l\right\|,\\
&= \alpha_l L \sum_{i=0}^{k-1} \left\|h_i(\bar{\omega}_l, \mathcal{D}, \alpha_l)\right\|,
\end{aligned}$$

To proceed, we recognise that each $\|h_i(\bar{\omega}_l, \mathcal{D}_l, \alpha_l)\| \leq c_h < \infty$ almost surely where c_h is a finite positive constant - otherwise:

$$P(\|h_i(\bar{\omega}_l, \mathcal{D}_l, \alpha_l)\| = \infty) > 0 \implies \mathbb{E}[\|h_i(\bar{\omega}_l, \mathcal{D}_l, \alpha_l)\|] = \infty \implies \mathbb{E}[\|h_i(\bar{\omega}_l, \mathcal{D}_l, \alpha_l)\|^2] = \infty$$
$$\implies \mathbb{E}[\|\delta(\bar{\omega}_l + \alpha_l h_j(\bar{\omega}, \mathcal{D}, \alpha_l), \bar{\omega}_l, \varsigma_j)\|^2] = \infty,$$

for at least one $i > j$, hence $\mathbb{V}_{\varsigma \sim P_\varsigma}[\delta(\omega, \omega', \varsigma)] = \infty$ for some ω, ω' thereby violating Assumption 2. Using c_h, we bound $\|\epsilon_{l+1}\|$:

$$\|\epsilon_{l+1}\| \leq \alpha_l L \sum_{i=0}^{k-1} c_h = \alpha_l c_h k L,$$

almost surely. We use this result to bound $\|\nu_{n,n+m}\|$:

$$\|\nu_{n,n+m}\| \leq \sum_{l=n}^{m+n-1} \alpha_l \|\epsilon_{l+1}\| \leq c_h k L \sum_{l=n}^{m+n-1} {\alpha_l}^2. \tag{9}$$

Now, under Assumption 3,

$$\lim_{n \to \infty} \sup_m \sum_{l=n}^{m+n-1} {\alpha_l}^2 = 0,$$

hence by the bound established in Equation (9):

$$\lim_{n \to \infty} \sup_m \|\nu_{n,n+m}\| = 0,$$

almost surely, as required. □

Theorem 2. *Under Assumptions 1- 4, the sequence of target parameter updates in Equation* (2) *converge almost surely to* $\omega^\star$.

Proof. Our update

$$\bar{\omega}_{l+1} = \bar{\omega}_l + \alpha_l \left(k\delta(\bar{\omega}_l, \bar{\omega}_l) + \mathcal{M}_{l+1} + \varepsilon_{l+1} \right),$$

is identical to the update presented in Borkar & Meyn (2000, Eq. 2.1.1) with an additional term ε_{l+1}. Proof of convergence to the ODE is given by Borkar & Meyn (2000, Lemma 1), which is predicated on the convergence of:

$$\Delta_{n,n+m} \coloneqq \zeta_{n+m} - \zeta_n,$$

from Borkar & Meyn (2000, Eq. 2.1.6) where

$$\zeta_n = \sum_{l=0}^{n-1} \alpha_l \mathcal{M}_{l+1},$$

for $n \geq 1$, that is $\lim_{n \to \infty} \sup_m \|\Delta_{n,n+m}\| = 0$, almost surely. To adapt our updates so that Borkar & Meyn (2000, Lemma 1) still applies, we recognise that the term ζ_n is now replaced in our updates with:

$$\bar{\zeta}_n = \sum_{l=0}^{n-1} \alpha_l (\mathcal{M}_{l+1} + \epsilon_{l+1}),$$

and hence $\Delta_{n,n+m}$ is replaced in our updates with:

$$\begin{aligned}\bar{\Delta}_{n,n+m} \coloneqq & \bar{\zeta}_{n+m} - \bar{\zeta}_n, \\ = & \zeta_{n+m} - \zeta_n + \left(\sum_{l=0}^{n+m-1} \alpha_l \epsilon_{l+1}\right) - \left(\sum_{l=0}^{n-1} \alpha_l \epsilon_{l+1}\right), \\ = & \zeta_{n+m} - \zeta_n + \sum_{l=n}^{n+m-1} \alpha_l \epsilon_{l+1}, \\ = & \zeta_{n+m} - \zeta_n + \nu_{n,n+m}, \\ = & \Delta_{n,n+m} + \nu_{n,n+m},\end{aligned}$$

where $\nu_{n,n+m}$ is defined as Lemma 2. All arguments of Borkar & Meyn (2000, Lemma 1) remain unchanged, except Eq. 2.1.9, where we must now show that $\lim_{n\to\infty}\sup_m \|\bar{\Delta}_{n,n+m}\| = 0$:

$$\begin{aligned}\lim_{n\to\infty}\sup_m \|\bar{\Delta}_{n,n+m}\| &\le \lim_{n\to\infty}\sup_m \left(\|\Delta_{n,n+m}\| + \|\nu_{n,n+m}\|\right), \\ &\le \lim_{n\to\infty}\left(\sup_m \|\Delta_{n,n+m}\| + \sup_m \|\nu_{n,n+m}\|\right), \\ &= \lim_{n\to\infty}\sup_m \|\Delta_{n,n+m}\| + \lim_{n\to\infty}\sup_m \|\nu_{n,n+m}\|.\end{aligned}$$

Applying Lemma 2 yields $\lim_{n\to\infty}\sup_m \|\nu_{n,n+m}\| = 0$ almost surely, hence

$$\lim_{n\to\infty}\sup_m \|\bar{\Delta}_{n,n+m}\| \le \lim_{n\to\infty}\sup_m \|\Delta_{n,n+m}\|,$$

which is proved in Borkar & Meyn (2000, Lemma 1). Convergence of our algorithm is thus only predicated on the convergence of the update:

$$\bar{\omega}_{l+1} = \bar{\omega}_l + \alpha_l \left(k\delta(\bar{\omega}_l, \bar{\omega}_l) + \mathcal{M}_{l+1}\right). \tag{10}$$

Borkar & Meyn (2000, Theorem 2.2) proves convergence of Equation (10) almost surely to $\omega^\star$ given the following four conditions hold:

I $k\delta(\omega, \omega)$ is Lipschitz in ω,

II Stepsizes α_l satisfy Assumption 3,

III The sequence $\{\mathcal{M}_l, \mathcal{F}_l\}_{l\ge 0}$ is a Martingale difference sequence with respect to the increasing family of σ-algebras: $\mathcal{F}_l \coloneqq \sigma(\{\bar{\omega}_i, \mathcal{M}_i\}_{i\in\{0:l\}})$ where $\mathbb{E}\left[\mathcal{M}_{l+1}|\mathcal{F}_l\right] = 0$ and $\mathbb{E}\left[\|\mathcal{M}_{l+1}\|^2|\mathcal{F}_l\right] \le C(1 + \|\bar{\omega}_l\|^2)$ for some positive $C < \infty$.

IV The sequence of iterates remain bounded, that is $\sup_l \|\bar{\omega}_l\| < \infty$ almost surely.

Conditions I and II hold trivially.

For Condition III, we can take expectations of the Martingale difference:

$$\begin{aligned}\mathbb{E}\left[\mathcal{M}_{l+1}|\mathcal{F}_l\right] &= \mathbb{E}\left[\mathcal{M}_{l+1}|\mathcal{F}_l\right], \\ &= \mathbb{E}\left[\sum_{i=0}^{k-1} \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_i) - k\delta(\bar{\omega}_l) \middle| \mathcal{F}_l\right], \\ &= \mathbb{E}\left[k\delta(\bar{\omega}_l, \bar{\omega}_l) - k\delta(\bar{\omega}_l, \bar{\omega}_l) \middle| \mathcal{F}_l\right], \\ &= 0,\end{aligned}$$

as required. We now show that the variance is bounded using Assumption 2:

$$\begin{aligned}
\|\mathcal{M}_{l+1}\|^2 &= \left\| \sum_{i=0}^{k-1} \left(\delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_i) - \delta(\bar{\omega}_l, \bar{\omega}_l)\right) \right\|^2, \\
&\le k \left\| \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_i) - \delta(\bar{\omega}_l, \bar{\omega}_l) \right\|^2, \\
\implies \mathbb{E}\left[\|\mathcal{M}_{l+1}\|^2 | \mathcal{F}_l\right] &\le k^2 \mathbb{E}\left[\left\| \delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma_i) - \delta(\bar{\omega}_l, \bar{\omega}_l) \right\|^2 \middle| \mathcal{F}_l \right], \\
&= k \mathbb{V}_{\varsigma \sim P_\varsigma}[\delta(\bar{\omega}_l, \bar{\omega}_l, \varsigma)], \\
&\le k\sigma_\delta^2,
\end{aligned}$$

thereby satisfying Condition III.

Finally, we prove Condition IV using Vidyasagar (2022, Theorem 5), which states iterates remain bounded almost surely if:

(a) Conditions I and III hold;

(b) there exists some Lyapunov function $V : \Omega \mapsto \mathbb{R}^+$ such that $a\|\omega - \omega^\star\|^2 \le V(\omega) \le b\|\omega - \omega^\star\|^2$ for constants $a, b > 0$ and $\|\nabla_\omega^2 V(\omega)\|$ is bounded, and;

(c) $\nabla_\omega V(\omega)^\top \delta(\omega, \omega) < 0$ for all $\omega \in \mathcal{X}_{\text{TD}}(\omega^\star)$.

We propose $V(\omega) = \frac{1}{2}\|\omega - \omega^\star\|^2$ as a candidate Lyapunov function, which trivially satisfies (b). We now show (c) holds by applying the fundamental theorem of calculus to $\delta(\omega, \omega)$. Let $\ell(t) := \omega - t(\omega - \omega^\star)$. Like in Theorem 1, it follows:

$$\begin{aligned}
\delta(\omega, \omega) &= \delta(\omega, \omega) - \underbrace{\delta(\omega^\star, \omega^\star)}_{=0}, \\
&= \delta \circ l(t = 0) - \delta \circ l(t = 1), \\
&= -\int_0^1 \partial_t \delta \circ l(t) dt, \\
&= \int_0^1 \nabla_\omega \delta \circ l(t) dt (\omega - \omega^\star),
\end{aligned}$$

hence:

$$\begin{aligned}
\nabla_\omega V(\omega)^\top \delta(\omega, \omega) &= (\omega - \omega^\star)^\top \int_0^1 \nabla_\omega \delta \circ l(t) dt (\omega - \omega^\star), \\
&= (\omega - \omega^\star)^\top \bar{J}_{TD}(\omega^\star, \omega)(\omega - \omega^\star), \\
&< 0,
\end{aligned}$$

for all $\omega \in \mathcal{X}_{\text{TD}}(\omega^\star)$ under Assumption 4, as required. □

B.3. Stabilising FPE

Proposition 1. *Using the regularised TD vector:*

$$\delta_{Reg}(\omega, \omega') = \delta(\omega, \omega') - \eta\omega\mathbb{I}(\omega \neq \omega') - \mu(\omega - \omega')$$

the path-mean Jacobians are:

$$\begin{aligned}
\bar{H}_{Reg}(\omega, \omega^\star; \bar{\omega}_l) &= \bar{H}(\omega, \omega^\star; \bar{\omega}_l) + (\mu + \eta)I, \\
\bar{J}_{\delta,Reg}(\omega, \omega^\star; \bar{\omega}_l) &= \bar{J}_\delta(\omega, \omega^\star; \bar{\omega}_l) + \mu I,
\end{aligned}$$

Assumption 5 is satisfied if:

$$\sup_{\omega, \omega' \in \mathcal{X}_{FPE}(\omega^\star)} \left\| \bar{H}_{Reg}(\omega', \omega^\star; \omega)^{-1} \bar{J}_{\delta,Reg}(\omega, \omega^\star; \omega^\star) \right\| < 1. \tag{11}$$

There exists a finite η, μ such that $Equation$ (11) holds.

Proof. Taking derivatives of $\delta_{\text{Reg}}(\omega, \omega')$:

$$\begin{aligned}
&-\nabla_\omega \delta_{\text{Reg}}(\omega, \omega') = -\nabla_\omega \delta(\omega, \omega') + \mathbb{I}(\omega \neq \omega')\eta + I\mu, \\
\implies\ &\bar{H}_{\text{Reg}}(\omega, \omega^\star; \bar{\omega}_l) = -\int_0^1 \nabla_{\omega'} \delta_{\text{Reg}}(\omega' = \omega - t(\omega - \omega^\star), \bar{\omega}_l) dt = \bar{H}(\omega, \omega^\star; \bar{\omega}_l) + (\mu + \eta)I, \\
&\nabla_{\omega'} \delta_{\text{Reg}}(\omega, \omega') = \nabla_{\omega'} \delta(\omega, \omega') + \mu I, \\
\implies\ &\bar{J}_{\delta,\text{Reg}}(\omega, \omega^\star; \bar{\omega}_l) = \int_0^1 \nabla_{\omega'} \delta_{\text{Reg}}(\bar{\omega}_l, \omega' = \omega - t(\omega - \omega^\star)) dt = \bar{J}_\delta(\omega, \omega^\star; \bar{\omega}_l) + \mu I.
\end{aligned}$$

Without loss of generality, assume $\mu = na$ and $\eta = nb$ for some $0 < a, b$. Hence:

$$\left\| \bar{H}_{\text{Reg}}(\omega', \omega^\star; \omega)^{-1} \bar{J}_{\delta,\text{Reg}}(\omega, \omega^\star; \omega^\star) \right\| = \left\| (\bar{H}(\omega', \omega^\star; \omega) + n(a+b)I)^{-1} (\bar{J}_\delta(\omega, \omega^\star; \omega^\star) + naI) \right\|.$$

From the continuity of the norm, it thus follows:

$$\lim_{n \to \infty} \left\| (\bar{H}(\omega', \omega^\star; \omega) + n(a+b)I)^{-1} (\bar{J}_\delta(\omega, \omega^\star; \omega^\star) + naI) \right\| = \left| \frac{a}{a+b} \right| < 1.$$

From the definition of the limit, there exists some finite n' such that

$$\left\| \bar{H}_{\text{Reg}}(\omega', \omega^\star; \omega)^{-1} \bar{J}_{\delta,\text{Reg}}(\omega, \omega^\star; \omega^\star) \right\| < \left| \frac{a}{a+b} \right| + \epsilon,$$

for all $n > n'$. As ϵ is arbitrary, it can be chosen such that

$$\left\| \bar{H}_{\text{Reg}}(\omega', \omega^\star; \omega)^{-1} \bar{J}_{\delta,\text{Reg}}(\omega, \omega^\star; \omega^\star) \right\| < 1,$$

for all $n > n'$, as required. □

B.4. Nonasymptotic Analysis

Lemma 3. *Under Assumption 2, for $i > 0$ the expected updates can be factored as:*

$$\begin{aligned}
\mathbb{E}_{P_\varsigma}[\omega_{i+1} - \bar{\omega}_l^\star] &= \left(I - \alpha_l \bar{H}(\omega_i, \bar{\omega}_l^\star; \bar{\omega}_l)\right)(\omega_i - \bar{\omega}_l^\star), \\
\mathbb{E}_{P_\varsigma}[\omega_{i+1} - \omega^\star] &= \left(I - \alpha_l \bar{H}(\omega_i, \bar{\omega}_l^\star; \bar{\omega}_l)\right)(\omega_i - \bar{\omega}_l^\star) + \bar{\omega}_l^\star - \omega^\star.
\end{aligned}$$

and for $i = 0$:

$$\mathbb{E}_{P_\varsigma}[\omega_1 - \bar{\omega}_l^\star] = (I + \alpha \bar{J}_{TD}(\bar{\omega}_l, \omega^\star))(\bar{\omega}_l - \omega^\star) + \omega^\star - \bar{\omega}_l^\star$$

Proof. By the definition of the expected update ω_{i+1}:

$$\mathbb{E}_{P_\varsigma}[\omega_{i+1} - \bar{\omega}_l^\star] = \omega_i - \bar{\omega}_l^\star + \alpha_l \delta(\omega_i, \bar{\omega}_l) - \alpha_l \underbrace{\delta(\bar{\omega}_l^\star, \bar{\omega}_l)}_{=0}.$$

Like in Theorem 1, let $\ell(t) \coloneqq \omega_i - t(\omega_i - \bar{\omega}_l^\star)$ define the line connecting ω_i to $\bar{\omega}_l^\star$. Using this notation we re-write the expected update as:

$$\mathbb{E}_{P_\varsigma}\left[\omega_{i+1} - \bar{\omega}_l^\star\right] = \omega_i - \bar{\omega}_l^\star + \alpha_l \left(\delta(\omega = \ell(0), \bar{\omega}_l) - \delta(\omega = \ell(1), \bar{\omega}_l)\right).$$

Applying the fundamental theorem of calculus under Assumption 2 and the chain rule yields our desired result:

$$\begin{aligned}
\mathbb{E}_{P_\varsigma}\left[\omega_{i+1} - \bar{\omega}_l^\star\right] &= \omega_i - \bar{\omega}_l^\star - \alpha_l \int_0^1 \partial_t \delta(\omega = \ell(t), \bar{\omega}_l) dt, \\
&= \omega_i - \bar{\omega}_l^\star - \alpha_l \int_0^1 \nabla_\omega \delta(\omega, \bar{\omega}_l)_{\omega = \ell(t)} \partial_t \ell(t) dt, \\
&= \omega_i - \bar{\omega}_l^\star + \alpha_l \left(\int_0^1 \nabla_\omega \delta(\omega, \bar{\omega}_l)_{\omega = \ell(t)} dt \right)(\omega_i - \bar{\omega}_l^\star), \\
&= \left(I - \alpha_l \bar{H}(\omega_i, \bar{\omega}_l^\star; \bar{\omega}_l)\right)(\omega_i - \bar{\omega}_l^\star).
\end{aligned}$$

Our second result follows immediately:

$$\begin{aligned}\mathbb{E}_{P_\varsigma}[\omega_{i+1}-\omega^\star] &= \mathbb{E}_{P_\varsigma}[\omega_{i+1}-\bar{\omega}_l^\star]+\bar{\omega}_l^\star-\omega^\star,\\ &= \left(I-\alpha_l\bar{H}(\omega_i,\bar{\omega}_l^\star;\bar{\omega}_l)\right)(\omega_i-\bar{\omega}_l^\star)+\bar{\omega}_l^\star-\omega^\star.\end{aligned}$$

For our final result:

$$\begin{aligned}\mathbb{E}_{P_\varsigma}[\omega_1-\bar{\omega}_l^\star] &= \mathbb{E}_{P_\varsigma}[\omega_1-\omega^\star+\omega^\star-\bar{\omega}_l^\star],\\ &= \mathbb{E}_{P_\varsigma}[\omega_1-\omega^\star]+\omega^\star-\bar{\omega}_l^\star.\end{aligned}$$

By the definition of the expected update:

$$\mathbb{E}_{P_\varsigma}[\omega_1-\omega^\star]=\bar{\omega}_l-\omega^\star+\alpha_l\delta(\bar{\omega}_l,\bar{\omega}_l)-\alpha_l\underbrace{\delta(\omega^\star,\omega^\star)}_{=0}.$$

Let $\ell(t)\coloneqq\bar{\omega}_l-t(\bar{\omega}_l-\omega^\star)$ define the line connecting $\bar{\omega}_l$ to $\omega^\star$. Using this notation we re-write the expected update as:

$$\mathbb{E}_{P_\varsigma}[\omega_1-\omega^\star]=\bar{\omega}_l-\omega^\star+\alpha_l\left(\delta(\omega=\ell(0),\omega=\ell(t))-\delta(\omega=\ell(1),\omega=\ell(1))\right).$$

Applying the fundamental theorem of calculus under Assumption 2 and the chain rule yields our desired result:

$$\begin{aligned}\mathbb{E}_{P_\varsigma}\left[\omega_{i+1}-\bar{\omega}_l^\star\right] &= \bar{\omega}_l-\omega^\star-\alpha_l\int_0^1\partial_t\delta(\omega=\ell(t),\omega=\ell(t))dt,\\ &= \bar{\omega}_l-\omega^\star-\alpha_l\int_0^1\nabla_\omega\delta(\omega,\omega)|_{\omega=\ell(t)}\partial_t\ell(t)dt,\\ &= \bar{\omega}_l-\omega^\star+\alpha_l\left(\int_0^1\nabla_\omega\delta(\omega,\omega)|_{\omega=\ell(t)}dt\right)(\bar{\omega}_l-\omega^\star),\\ &= \left(I+\alpha_l\bar{J}_{\mathrm{TD}}(\bar{\omega}_l^\star,\omega^\star)\right)(\omega_i-\bar{\omega}_l^\star).\end{aligned}$$

□

Lemma 4. *Under Assumption 2,*

$$\mathbb{E}_{P_{\varsigma_i}}\left[\|\omega_{i+1}-\omega^\star\|\right]\le|1-\alpha_l\lambda_H^\star|\,\|\omega_i-\bar{\omega}_l^\star\|+\|\bar{\omega}_l^\star-\omega^\star\|+\alpha_l\sigma_\delta.$$

Proof. We start by bounding the expected norm term using Jensen's inequality: $\mathbb{E}_X[\sqrt{X^2}]\le\sqrt{\mathbb{E}_X[X^2]}$:

$$\begin{aligned}\mathbb{E}_{P_{\varsigma_i}}\left[\|\omega_{i+1}-\omega^\star\|\right] &\le \sqrt{\mathbb{E}_{P_{\varsigma_i}}\left[\|\omega_{i+1}-\omega^\star\|^2\right]},\\ &= \sqrt{\left\|\mathbb{E}_{P_{\varsigma_i}}\left[\omega_{i+1}-\omega^\star\right]\right\|^2+\mathbb{V}_{P_{\varsigma_i}}\left[\omega_{i+1}-\omega^\star\right]},\\ &= \sqrt{\left\|\mathbb{E}_{P_{\varsigma_i}}\left[\omega_{i+1}-\omega^\star\right]\right\|^2+\mathbb{V}_{P_{\varsigma_i}}\left[\omega_{i+1}\right]},\\ &\le \left\|\mathbb{E}_{P_{\varsigma_i}}\left[\omega_{i+1}-\omega^\star\right]\right\|+\sqrt{\mathbb{V}_{P_{\varsigma_i}}\left[\omega_{i+1}\right]}\end{aligned}$$

where we applied the triangle inequality to derive the final line. We bound the variance term by substituting $\omega_{i+1}=\omega_i+\alpha_l\delta(\omega_i,\bar{\omega}_l,\varsigma_i)$:

$$\begin{aligned}\mathbb{V}_{P_{\varsigma_i}}\left[\omega_{i+1}\right] &= (\alpha_l)^2\mathbb{E}_{P_{\varsigma_i}}\left[\left\|\delta(\omega_i,\bar{\omega}_l,\varsigma_i)-\mathbb{E}_{P_{\varsigma_i}}\left[\delta(\omega_i,\bar{\omega}_l,\varsigma_i)\right]\right\|^2\right],\\ &= (\alpha_l)^2\mathbb{V}_{P_{\varsigma_i}}\left[\delta(\omega_i,\bar{\omega}_l,\varsigma_i)\right],\\ &\le (\alpha_l\sigma_\delta)^2,\\ \implies \mathbb{E}_{P_{\varsigma_i}}\left[\|\omega_{i+1}-\omega^\star\|\right] &\le \left\|\mathbb{E}_{P_{\varsigma_i}}\left[\omega_{i+1}-\omega^\star\right]\right\|+\alpha_l\sigma_\delta \qquad (12)\end{aligned}$$

Applying Lemma 3 to the expectation and using the triangle inequality yields our desired result:

$$\begin{aligned}
\mathbb{E}_{P_{\varsigma_i}}\left[\|\omega_{i+1}-\omega^\star\|\right] &\le \left\|\left(I-\alpha_l\bar{H}(\omega_i,\bar{\omega}_l^\star;\bar{\omega}_l)\right)(\omega_i-\bar{\omega}_l^\star)+(\bar{\omega}_l^\star-\omega^\star)\right\|+\alpha_l\sigma_\delta,\\
&\le \left\|I-\alpha_l\bar{H}(\omega_i,\bar{\omega}_l^\star;\bar{\omega}_l)\right\|\|\omega_i-\bar{\omega}_l^\star\|+\|\bar{\omega}_l^\star-\omega^\star\|+\alpha_l\sigma_\delta,\\
&\le \sup_{\omega_i,\bar{\omega}_l^\star,\bar{\omega}_l}\left\|I-\alpha_l\bar{H}(\omega_i,\bar{\omega}_l^\star;\bar{\omega}_l)\right\|\|\omega_i-\bar{\omega}_l^\star\|+\|\bar{\omega}_l^\star-\omega^\star\|+\alpha_l\sigma_\delta,\\
&= |1-\alpha_l\lambda_H^\star|\,\|\omega_i-\bar{\omega}_l^\star\|+\|\bar{\omega}_l^\star-\omega^\star\|+\alpha_l\sigma_\delta.
\end{aligned}$$

□

Theorem 3. *Define*

$$\sigma_k\coloneqq\left(1-|1-\alpha_l\lambda_H^\star|^k\right)\frac{\sigma_\delta}{\lambda_H^\star},$$

Let Assumptions 1 and 2 hold, then:

$$\mathbb{E}\left[\|\bar{\omega}_{l+1}-\omega^\star\|\right]\le\mathcal{C}(\alpha_l,k)\mathbb{E}\left[\|\bar{\omega}_l-\omega^\star\|\right]+\alpha_l\sigma_k.$$

Proof. Let $\{\omega_i\}_{i=0}^k$ denote the intermediate function approximation parameters between target parameter updates $\bar{\omega}_{l+1}$ and $\bar{\omega}_l$, with $\omega_0=\bar{\omega}_l$ and $\omega_k=\bar{\omega}_{l+1}$. We define the set of samples up to i as: $\mathcal{D}_i\coloneqq\{\varsigma_j\}_{j=0}^i$ with distribution $P_{\mathcal{D}_i}$, with sample ς_j having distribution P_{ς_j}. Under this notation, we must show:

$$\mathbb{E}_{P_{\mathcal{D}_{k-1}}}\left[\|\omega_k-\omega^\star\|\right]\le\mathcal{C}(\alpha_l,k)\|\omega_0-\omega^\star\|+\alpha_l\sigma_k.$$

Applying Lemma 4 to the inner expectation:

$$\begin{aligned}
\mathbb{E}_{P_{\mathcal{D}_{k-1}}}\left[\|\omega_k-\omega^\star\|\right] &= \mathbb{E}_{P_{\mathcal{D}_{k-2}}}\left[\mathbb{E}_{P_{\varsigma_{k-1}}}\left[\|\omega_k-\omega^\star\|\right]\right],\\
&\le \mathbb{E}_{P_{\mathcal{D}_{k-2}}}\left[|1-\alpha_l\lambda_H^\star|\,\|\omega_{k-1}-\bar{\omega}_l^\star\|+\|\bar{\omega}_l^\star-\omega^\star\|+\alpha_l\sigma_\delta\right],\\
&= |1-\alpha_l\lambda_H^\star|\,\mathbb{E}_{P_{\mathcal{D}_{k-2}}}\left[\|\omega_{k-1}-\bar{\omega}_l^\star\|\right]+\|\bar{\omega}_l^\star-\omega^\star\|+\alpha_l\sigma_\delta,\\
&= |1-\alpha_l\lambda_H^\star|\,\mathbb{E}_{P_{\mathcal{D}_{k-3}}}\left[\mathbb{E}_{P_{\varsigma_{k-2}}}\left[\|\omega_{k-1}-\bar{\omega}_l^\star\|\right]\right]+\|\bar{\omega}_l^\star-\omega^\star\|+\alpha_l\sigma_\delta. \qquad (13)
\end{aligned}$$

Applying Equation (12) from Lemma 4 to the inner expectation and applying Lemma 3 yields:

$$\begin{aligned}
\mathbb{E}_{P_{\varsigma_{k-2}}}\left[\|\omega_{k-1}-\bar{\omega}_l^\star\|\right] &\le \left\|\mathbb{E}_{P_{\varsigma_{k-2}}}\left[\omega_{k-1}-\omega^\star\right]\right\|+\alpha_l\sigma_\delta,\\
&\le \left\|\left(I-\alpha_l\bar{H}(\omega_{k-2},\bar{\omega}_l^\star;\bar{\omega}_l)\right)(\omega_{k-2}-\bar{\omega}_l^\star)\right\|+\alpha_l\sigma_\delta,\\
&\le \sup_{\omega_{k-2},\bar{\omega}_l^\star,\bar{\omega}_l}\left\|I-\alpha_l\bar{H}(\omega_{k-2},\bar{\omega}_l^\star;\bar{\omega}_l)\right\|\|\omega_{k-2}-\bar{\omega}_l^\star\|+\alpha_l\sigma_\delta,\\
&= |I-\alpha_l\lambda_H^\star|\,\|\omega_{k-2}-\bar{\omega}_l^\star\|+\alpha_l\sigma_\delta. \qquad (14)
\end{aligned}$$

Recursively applying Equation (15) to Equation (13) $k-1$ times yields:

$$\begin{aligned}
\mathbb{E}_{P_{\mathcal{D}_{k-1}}}\left[\|\omega_k-\omega^\star\|\right] &\le \mathbb{E}_{P_{\varsigma_0}}\left[|1-\alpha_l\lambda_H^\star|^{k-1}\|\omega_1-\bar{\omega}_l^\star\|\right]+\|\bar{\omega}_l^\star-\omega^\star\|+\sum_{i=0}^{k-2}|1-\alpha_l\lambda_H^\star|^i\alpha_l\sigma_\delta,\\
&= |1-\alpha_l\lambda_H^\star|^{k-1}\,\mathbb{E}_{P_{\varsigma_0}}\left[\|\omega_1-\omega_l^\star\|\right]+\|\bar{\omega}_l^\star-\omega^\star\|+\sum_{i=0}^{k-2}|1-\alpha_l\lambda_H^\star|^i\alpha_l\sigma_\delta. \qquad (15)
\end{aligned}$$

Now, applying Equation (12) and Lemma 3 to the expectation:

$$\begin{aligned}
\mathbb{E}_{P_{\varsigma_0}}\left[\|\omega_1-\omega_l^\star\|\right] &\le \left\|\mathbb{E}_{P_{\varsigma_0}}\left[\omega_1-\omega_l^\star\right]\right\|+\alpha_l\sigma_\delta,\\
&= \left\|(I+\alpha\bar{J}_{\text{TD}}(\bar{\omega}_l,\omega^\star))(\bar{\omega}_l-\omega^\star)+\omega^\star-\bar{\omega}_l^\star\right\|+\alpha_l\sigma_\delta,\\
&\le \left\|I+\alpha\bar{J}_{\text{TD}}(\bar{\omega}_l,\omega^\star)\right\|\|\bar{\omega}_l-\omega^\star\|+\|\bar{\omega}_l^\star-\omega^\star\|+\alpha_l\sigma_\delta,\\
&= \left\|\bar{J}_{\text{TD}}^\star\right\|\|\bar{\omega}_l-\omega^\star\|+\|\omega_l^\star-\bar{\omega}_l\|+\alpha_l\sigma_\delta.
\end{aligned}$$

Substituting into Equation (15):

$$\begin{aligned}
\mathbb{E}_{P_{\mathcal{D}_{k-1}}}\left[\|\omega_k-\omega^\star\|\right] &\le |1-\alpha_l\lambda_H^\star|^{k-1}\left\|\bar{J}_{\text{TD}}^\star\right\|\|\bar{\omega}_l-\omega^\star\| + (1+|1-\alpha_l\lambda_H^\star|^{k-1})\|\bar{\omega}_l^\star-\omega^\star\| + \sum_{i=0}^{k-1}|1-\alpha_l\lambda_H^\star|^i\,\alpha_l\sigma_\delta. \\
&= |1-\alpha_l\lambda_H^\star|^{k-1}\left\|\bar{J}_{\text{TD}}^\star\right\|\|\bar{\omega}_l-\omega^\star\| + (1+|1-\alpha_l\lambda_H^\star|^{k-1})\|\bar{\omega}_l^\star-\omega^\star\| + \frac{1-|1-\alpha_l\lambda_H^\star|^k}{1-|1-\alpha_l\lambda_H^\star|}\alpha_l\sigma_\delta, \\
&= |1-\alpha_l\lambda_H^\star|^{k-1}\left\|\bar{J}_{\text{TD}}^\star\right\|\|\bar{\omega}_l-\omega^\star\| + (1+|1-\alpha_l\lambda_H^\star|^{k-1})\|\bar{\omega}_l^\star-\omega^\star\| + \left(1-|1-\alpha_l\lambda_H^\star|^k\right)\frac{\sigma_\delta}{\lambda_H^\star}, \\
&\le |1-\alpha_l\lambda_H^\star|^{k-1}\left\|\bar{J}_{\text{TD}}^\star\right\|\|\bar{\omega}_l-\omega^\star\| + (1+|1-\alpha_l\lambda_H^\star|^{k-1})\|\bar{\omega}_l^\star-\omega^\star\| + \sigma_k.
\end{aligned}$$

Finally, we apply Theorem 1 to yield our desired result:

$$\begin{aligned}
&\mathbb{E}_{P_{\mathcal{D}_{k-1}}}\left[\|\omega_k-\omega^\star\|\right] \\
&\quad\le |1-\alpha_l\lambda_H^\star|^{k-1}\left\|\bar{J}_{\text{TD}}^\star\right\|\|\bar{\omega}_l-\omega^\star\| + \left(1+|1-\alpha_l\lambda_H^\star|^{k-1}\right)\left\|\bar{H}(\bar{\omega}_l^\star,\omega^\star;\bar{\omega}_l)^{-1}\bar{J}_\delta(\bar{\omega}_l,\omega^\star;\omega^\star)(\bar{\omega}_l-\omega^\star)\right\| + \sigma_k, \\
&\quad\le |1-\alpha_l\lambda_H^\star|^{k-1}\left\|\bar{J}_{\text{TD}}^\star\right\|\|\bar{\omega}_l-\omega^\star\| + \left(1+|1-\alpha_l\lambda_H^\star|^{k-1}\right)\left\|\bar{H}(\bar{\omega}_l^\star,\omega^\star;\bar{\omega}_l)^{-1}\bar{J}_\delta(\bar{\omega}_l,\omega^\star;\omega^\star)\right\|\|\bar{\omega}_l-\omega^\star\| + \sigma_k, \\
&\quad\le |1-\alpha_l\lambda_H^\star|^{k-1}\left\|\bar{J}_{\text{TD}}^\star\right\|\|\bar{\omega}_l-\omega^\star\| + \left(1+|1-\alpha_l\lambda_H^\star|^{k-1}\right)\left\|\bar{J}_{\text{FPE}}^\star\right\|\|\bar{\omega}_l-\omega^\star\| + \sigma_k, \\
&\quad= \mathcal{C}(\alpha_l,k)\,\|\bar{\omega}_l-\omega^\star\| + \sigma_k.
\end{aligned}$$

□

Corollary 3.1. *Let Assumptions 1, 2, 5 and 6 hold. For a fixed stepsize $\alpha_l=\alpha>0$. For a fixed stepsize $\alpha_l=\alpha>0$,*

$$\mathbb{E}\left[\|\bar{\omega}_l-\omega^\star\|\right] \le \frac{\alpha\sigma_k}{1-c} + \exp(-l(1-c))\left(\|\bar{\omega}_0-\omega^\star\| - \frac{\sigma_k}{1-c}\right).$$

Proof. We start by applying Theorem 3:

$$\mathbb{E}\left[\|\bar{\omega}_l-\omega^\star\|\right] \le \mathcal{C}(\alpha_l,k)\mathbb{E}\left[\|\bar{\omega}_{l-1}-\omega^\star\|\right] + \alpha_l\sigma_k.$$

As $\mathcal{X}_{\text{FPE}}(\omega^\star)$ is a region of contraction and $\bar{\omega}_l\in\mathcal{X}_{\text{FPE}}(\omega^\star)$ for all $l\ge 0$, there exists a positive $c<1$ under Assumption 6 such that $\mathcal{C}(\alpha_l,k)\le c$, hence:

$$\mathbb{E}\left[\|\bar{\omega}_l-\omega^\star\|\right] \le c\mathbb{E}\left[\|\bar{\omega}_{l-1}-\omega^\star\|\right] + \alpha_l\sigma_k. \tag{16}$$

Now, for a fixed constant stepsize $\alpha_l=\alpha$, we can apply Equation (16) l times, yielding:

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\omega}_l-\omega^\star\|\right] &\le c^l\,\|\bar{\omega}_0-\omega^\star\| + \alpha\sigma_k\sum_{i=0}^{l-1}c^i, \\
&= c^l\,\|\bar{\omega}_0-\omega^\star\| + \alpha\sigma_k\frac{1-c^l}{1-c} \\
&= c^l\left(\|\bar{\omega}_0-\omega^\star\| - \frac{\alpha\sigma_k}{1-c}\right) + \frac{\alpha\sigma_k}{1-c}, \\
&= (1-(1-c))^l\left(\|\bar{\omega}_0-\omega^\star\| - \frac{\alpha\sigma_k}{1-c}\right) + \frac{\alpha\sigma_k}{1-c}.
\end{aligned}$$

Now we apply the bound $1-x\le\exp(-x)$, yielding our desired result:

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\omega}_l-\omega^\star\|\right] &\le \exp(-(1-c))^l\left(\|\bar{\omega}_0-\omega^\star\| - \frac{(\alpha\sigma_k)}{1-c}\right) + \frac{\alpha\sigma_k}{1-c}, \\
&= \exp(-l(1-c))\left(\|\bar{\omega}_0-\omega^\star\| - \frac{\alpha\sigma_k}{1-c}\right) + \frac{\alpha\sigma_k}{1-c}.
\end{aligned}$$

□

B.5. Breaking the Deadly Triad

Theorem 4. *Let Assumption 7 hold over $\mathcal{X}_{FPE}(\omega^\star)$ from Definition 1. For any $\frac{1}{\alpha_l} > \frac{\lambda_1^{\min}+\lambda_1^{\max}}{2}$ such that $\alpha_l > 0$, any*

$$k > 1 + \frac{\log(1 - \|\bar{J}^\star_{FPE}\|) - \log(\|\bar{J}^\star_{TD}\| + \|\bar{J}^\star_{FPE}\|)}{\log(1-\alpha\lambda^{min})},$$

ensures that $\mathcal{X}_{FPE}(\omega^\star)$ is a region of contraction satisfying Assumption 6.

Proof. Now, as $|1-\alpha_l\lambda'|$ is a symmetric function of λ with a minima at $\lambda = \frac{1}{\alpha_l}$ and $\frac{\lambda_1^{\min}+\lambda_1^{\max}}{2}$ is the mid point of $\lambda_1^{\min}$ and $\lambda_1^{\max}$, it follows:

$$\lambda_H^\star \coloneqq \sup_{\omega,\omega' \in \mathcal{X}_{\text{FPE}}(\omega^\star)} \ \operatorname*{arg\,sup}_{\lambda' \in \lambda(\nabla^2_\omega \mathcal{L}(\omega,\omega'))} |1-\alpha_l\lambda'| = \lambda_1^{\min}.$$

Now,

$$\alpha_l < \frac{2}{\lambda_1^{\min}+\lambda_1^{\max}} \implies \lambda_H^\star \le \frac{2}{\alpha_l} \implies |1-\alpha_l\lambda_H^\star| < 1,$$

hence

$$\lim_{k\to\infty} \mathcal{C}(\alpha_l, k) = \lim_{k\to\infty} \left|1-\alpha_l\lambda^{\text{min}}\right|^{k-1} \|J^\star_{\text{TD}}\| + \lim_{k\to\infty}\left(1+\left|1-\alpha_l\lambda^{\text{min}}\right|^{k-1}\right)\left\|\bar{J}^\star_{\text{FPE}}\right\| = \left\|\bar{J}^\star_{\text{FPE}}\right\| < 1.$$

Let $\left\|\bar{J}^\star_{\text{FPE}}\right\| = 1-\epsilon$ where $0<\epsilon<1$. From the definition of a limit, this implies that for ϵ there exists some finite k' such that whenever $k > k'$:

$$\left|\mathcal{C}(\alpha_l,k) - \left\|\bar{J}^\star_{\text{FPE}}\right\|\right| < \epsilon \implies |\mathcal{C}(\alpha_l,k) - (1-\epsilon)| < \epsilon \implies \mathcal{C}(\alpha_l,k) < 1,$$

as required. To find the value of k for which $\mathcal{C}(\alpha_l,k)<1$, we set $\mathcal{C}(\alpha_l,k)=1$ and solve:

$$\begin{aligned}
1 &= \left|1-\alpha_l\lambda^{\text{min}}\right|^{k-1}\left\|\bar{J}^\star_{\text{TD}}\right\| + \left(1+\left|1-\alpha_l\lambda^{\text{min}}\right|^{k-1}\right)\left\|\bar{J}^\star_{\text{FPE}}\right\|,\\
\implies \left|1-\alpha_l\lambda^{\text{min}}\right|^{k-1} &= \frac{1-\left\|\bar{J}^\star_{\text{FPE}}\right\|}{\|\bar{J}^\star_{\text{TD}}\| + \left\|\bar{J}^\star_{\text{FPE}}\right\|},\\
\implies (k-1)\log(\left|1-\alpha_l\lambda^{\text{min}}\right|) &= \log(1-\left\|\bar{J}^\star_{\text{FPE}}\right\|) - \log(\|\bar{J}^\star_{\text{TD}}\| + \left\|\bar{J}^\star_{\text{FPE}}\right\|),\\
\implies k &= 1 + \frac{\log(1-\|\bar{J}^\star_{\text{FPE}}\|) - \log(\|\bar{J}^\star_{\text{TD}}\| + \|\bar{J}^\star_{\text{FPE}}\|)}{\log(1-\alpha\lambda^{\text{min}})}.
\end{aligned}$$

□

C. Additional Experiment Information

For both plots, each configuration was run over 5 random seeds, with the central tendency given by the mean, and the shaded errors representing the standard error of the mean. Hyperparameters that are not varied in the plots were optimised by grid search across either linear or logarithmic hyperparameter ranges, as is suitable. Parameters were chosen that led to the highest performance as averaged across random seeds, then relevant hyperparameters were varied, using the optimal fixed hyperparameters. Hyperparameters that were varied are denoted as lists in the tables below.

C.1. Baird's Counterexample

Figure Figure 4 shows the counterexample. The behaviour policy chooses between the action represented by the wavy line with probability 6/7, and the solid line with probability 1/7. The behaviour policy always chooses the solid line. The linear function approximation scheme is shown in terms of the value function weights. Sampling off policy in this way leads to divergence of TD, but PFPE converges, as seen in Figure 2.

C.2. Cartpole Experiment

For the Cartpole experiment, we use a simple DQN-style setup with a small multilayer perceptron (MLP) representing the value function. A small adjustment is made from PFPE as characterised by the paper. Instead of updating value parameters on single data points, parameter updates are averaged across a small batch. This was found to increase stability of learning in both settings, with no notable effects when comparing across independent variables. This means that, in addition to our target network, we also make use of a replay buffer which stores observed transitions. As such, data used in updates was sampled uniformly from previous transitions. The policy was ϵ-greedy, with the estimated optimal action taken with probability $1 - \epsilon$. The environment is maintained by OpenAI as part of the gym suite, and falls under MIT licensing.

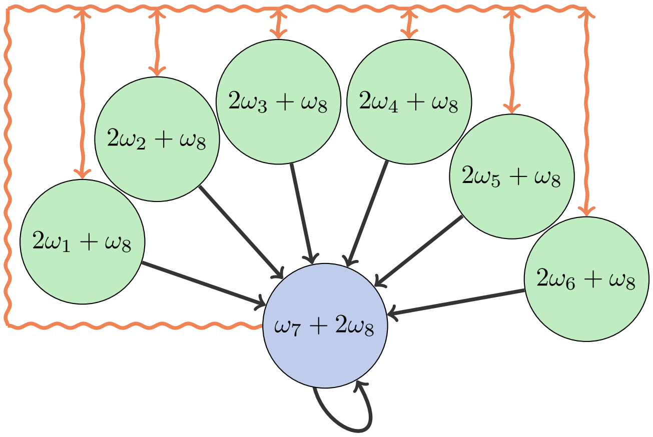

Figure 4: Baird's Counterexample. The solid (grey) action moves the agent to the lower state deterministically. The wavy (orange) action puts the agent into one of the upper states with equal probability

Parameter	Value
Environment Parameters	
γ	0.99
Architecture Parameters	
MLP Hidden Layers	2
Hidden Layer Size	32
Nonlinearity	ReLU
ϵ	0.05
Training Parameters	
Total Target Network Updates	500
Learning Rate	[0.001, 0.0005]
Momentum (μ)	[0, 0.01]
Batch Size	500
Steps per Target Network Update (k)	5
Data Gathering Steps per Update	5
Replay Buffer Size	2500

Table 1: Relevant Parameters for Cartpole Experiment

D. Extensions

As discussed in Section 4, once we can establish Assumption 4 then there are several theoretical tools that become applicable from stochastic approximation to prove convergence under a range of assumptions. Brooms (2006) provide a comprehensive overview of classic methods. In particular, stochastic approximation has been shown to converge when sampling from an ergodic Markov chain under specific regularity assumptions (Allasonniere et al., 2010). Perhaps the easiest to verify in our context is those of Andrieu et al. (2005), who provides a series of assumptions that can be checked in practice. Moreover, this theory was recently extended to Markov chains that converge sub-geometrically to their station distributions by Debavelaere et al. (2021). Adherence of the updates to remain in a contractive region can be ensured by projection into an ever increasing subset of Ω until convergence occurs, which is detailed and analysed in Andradottir (1991).