Decentralized Deterministic Multi-Agent Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

[Zhang, ICML 2018] provided the first decentralized actor-critic algorithm for multi-agent reinforcement learning (MARL) that offers convergence guarantees. In that work, policies are stochastic and are defined on finite action spaces. We extend those results to offer a provably-convergent decentralized actor-critic algorithm for learning deterministic policies on continuous action spaces. Deterministic policies are important in real-world settings. To handle the lack of exploration inherent in deterministic policies, we consider both off-policy and on-policy settings. We provide the expression of a local deterministic policy gradient, decentralized deterministic actor-critic algorithms and convergence guarantees for linearly-approximated value functions. This work will help enable decentralized MARL in high-dimensional action spaces and pave the way for more widespread use of MARL.

1 Introduction

Cooperative multi-agent reinforcement learning (MARL) has seen considerably less use than its single-agent analog, in part because often no central agent exists to coordinate the cooperative agents. As a result, decentralized architectures have been advocated for MARL. Recently, decentralized architectures have been shown to admit convergence guarantees comparable to their centralized counterparts under mild network-specific assumptions (see Zhang et al. [2018], Suttle et al. [2019]). In this work, we develop a decentralized actor-critic algorithm with deterministic policies for multi-agent reinforcement learning. Specifically, we extend results for actor-critic with stochastic policies (Bhatnagar et al. [2009], Degris et al. [2012], Maei [2018], Suttle et al. [2019]) to handle deterministic policies. Indeed, theoretical and empirical work has shown that deterministic algorithms outperform their stochastic counterparts in high-dimensional continuous action settings (Silver et al. [January 2014b], Lillicrap et al. [2015], Fujimoto et al. [2018]). Deterministic policies further avoid estimating the complex integral over the action space. Empirically this allows for lower variance of the critic estimates and faster convergence. On the other hand, deterministic policy gradient methods suffer from reduced exploration. For this reason, we provide both off-policy and on-policy versions of our results, the off-policy version allowing for significant improvements in exploration. The contributions of this paper are three-fold: (1) we derive the expression of the gradient in terms of the long-term average reward, which is needed in the undiscounted multi-agent setting with deterministic policies; (2) we show that the deterministic policy gradient is the limiting case, as policy variance tends to zero, of the stochastic policy gradient; and (3) we provide a decentralized deterministic multi-agent actor critic algorithm and prove its convergence under linear function approximation.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

2 Background

Consider a system of N agents denoted by $\mathcal{N} = [N]$ in a decentralized setting. Agents determine their decisions independently based on observations of their own rewards. Agents may however communicate via a possibly time-varying communication network, characterized by an undirected graph $\mathcal{G}_t = (\mathcal{N}, \mathcal{E}_t)$, where $\mathcal{E}_t$ is the set of communication links connecting the agents at time $t \in \mathbb{N}$. The networked multi-agent MDP is thus characterized by a tuple $(\mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, P, \{R^i\}_{i\in\mathcal{N}}, \{\mathcal{G}_t\}_{t\geq 0})$ where $\mathcal{S}$ is a finite global state space shared by all agents in $\mathcal{N}$, $\mathcal{A}^i$ is the action space of agent i, and $\{\mathcal{G}_t\}_{t\geq 0}$ is a time-varying communication network. In addition, let $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}^i$ denote the joint action space of all agents. Then, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability of the MDP, and $R^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the local reward function of agent i. States and actions are assumed globally observable whereas rewards are only locally observable. At time t, each agent i chooses its action $a_t^i \in \mathcal{A}^i$ given state $s_t \in \mathcal{S}$, according to a local parameterized policy $\pi^i_{\theta^i} : \mathcal{S} \times \mathcal{A}^i \to [0, 1]$, where $\pi^i_{\theta^i}(s, a^i)$ is the probability of agent i choosing action a^i at state s, and $\theta^i \in \Theta^i \subseteq \mathbb{R}^{m_i}$ is the policy parameter. We pack the parameters together as $\theta = [(\theta^1)^\top, \cdots, (\theta^N)^\top]^\top \in \Theta$ where $\Theta = \prod_{i\in\mathcal{N}} \Theta^i$. We denote the joint policy by $\pi_\theta : \mathcal{S} \times \mathcal{A} \to [0, 1]$ where $\pi_\theta(s, a) = \prod_{i\in\mathcal{N}} \pi^i_{\theta^i}(s, a^i)$. Note that decisions are decentralized in that rewards are observed locally, policies are evaluated locally, and actions are executed locally. We assume that for any $i \in \mathcal{N}$, $s \in \mathcal{S}$, $a^i \in \mathcal{A}^i$, the policy function $\pi^i_{\theta^i}(s, a^i) > 0$ for any $\theta^i \in \Theta^i$ and that $\pi^i_{\theta^i}(s, a^i)$ is continuously differentiable with respect to the parameters θ^i over Θ^i. In addition, for any $\theta \in \Theta$, let $P^\theta : \mathcal{S} \times \mathcal{S} \to [0, 1]$ denote the transition matrix of the Markov chain $\{s_t\}_{t\geq 0}$ induced by policy π_θ, that is, for any $s, s' \in \mathcal{S}$, $P^\theta(s'|s) = \sum_{a\in\mathcal{A}} \pi_\theta(s, a) \cdot P(s'|s, a)$. We make the standard assumption that the Markov chain $\{s_t\}_{t\geq 0}$ is irreducible and aperiodic under any π_θ and denote its stationary distribution by d_θ.

Our objective is to find a policy π_θ that maximizes the long-term average reward over the network. Let r^i_{t+1} denote the reward received by agent i as a result of taking action a^i_t. Then, we wish to solve:

$$\max_\theta J(\pi_\theta) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i\in\mathcal{N}} r^i_{t+1}\right] = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} d_\theta(s)\pi_\theta(s, a)\bar{R}(s, a),$$

where $\bar{R}(s, a) = (1/N) \cdot \sum_{i\in\mathcal{N}} R^i(s, a)$ is the globally averaged reward function. Let $\bar{r}_t = (1/N) \cdot \sum_{i\in\mathcal{N}} r^i_t$, then $\bar{R}(s, a) = \mathbb{E}[\bar{r}_{t+1}|s_t = s, a_t = a]$, and therefore, the global relative action-value function is: $Q_\theta(s, a) = \sum_{t\geq 0} \mathbb{E}[\bar{r}_{t+1} - J(\theta)|s_0 = s, a_0 = a, \pi_\theta]$, and the global relative state-value function is: $V_\theta(s) = \sum_{a\in\mathcal{A}} \pi_\theta(s, a)Q_\theta(s, a)$. For simplicity, we refer to V_θ and Q_θ as simply the state-value function and action-value function. We define the advantage function as $A_\theta(s, a) = Q_\theta(s, a) - V_\theta(s)$.

Zhang et al. [2018] provided the first provably convergent MARL algorithm in the context of the above model. The fundamental result underlying their algorithm is a local policy gradient theorem:

$$\nabla_{\theta^i} J(\mu_\theta) = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\nabla_{\theta^i} \log \pi^i_{\theta^i}(s, a^i) \cdot A^i_\theta(s, a)\right],$$

where $A^i_\theta(s, a) = Q_\theta(s, a) - \tilde{V}^i_\theta(s, a^{-i})$ is a local advantage function and $\tilde{V}^i_\theta(s, a^{-i}) = \sum_{a^i\in\mathcal{A}^i} \pi^i_{\theta^i}(s, a^i)Q_\theta(s, a^i, a^{-i})$. This theorem has important practical value as it shows that the policy gradient with respect to each local parameter θ^i can be obtained locally using the corresponding score function $\nabla_{\theta^i} \log \pi^i_{\theta^i}$ provided that agent i has an unbiased estimate of the advantage functions A^i_θ or A_θ. With only local information, the advantage functions A^i_θ or A_θ cannot be well estimated since the estimation requires the rewards $\{r^i_t\}_{i\in\mathcal{N}}$ of all agents. Therefore, they proposed a consensus based actor-critic that leverages the communication network to share information between agents by placing a weight $c_t(i, j)$ on the message transmitted from agent j to agent i at time t. Their action-value function Q_θ was approximated by a parameterized function $\hat{Q}_\omega : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and each agent i maintains its own parameter ω^i, which it uses to form a local estimate $\hat{Q}_{\omega^i}$ of the global Q_θ. At each time step t, each agent i shares its local parameter ω^i_t with its neighbors on the network, and the shared parameters are used to arrive at a consensual estimate of Q_θ over time.

3 Local Gradients of Deterministic Policies

While the use of a stochastic policy facilitates the derivations of convergence proofs, most real-world control tasks require a deterministic policy to be implementable. In addition, the quantities estimated in the deterministic critic do not involve estimation of the complex integral over the action space found in the stochastic version. This offers lower variance of the critic estimates and faster convergence. To address the lack of exploration that comes with deterministic policies, we provide both off-policy and on-policy versions of our results. Our first requirement is a local deterministic policy gradient theorem.

We assume that $\mathcal{A}^i = \mathbb{R}^{n_i}$. We make standard regularity assumptions on our MDP. That is, we assume that for any $s, s' \in \mathcal{S}$, $P(s'|s,a)$ and $R^i(s,a)$ are bounded and have bounded first and second derivatives. We consider local deterministic policies $\mu^i_{\theta^i} : \mathcal{S} \to \mathcal{A}^i$ with parameter vector $\theta^i \in \Theta^i$, and denote the joint policy by $\mu_\theta : \mathcal{S} \to \mathcal{A}$, where $\mu_\theta(s) = (\mu^1_{\theta^1}(s), \ldots, \mu^N_{\theta^N}(s))$ and $\theta = [(\theta^1)^\top, \ldots, (\theta^N)^\top]^\top$. We assume that for any $s \in \mathcal{S}$, the deterministic policy function $\mu^i_{\theta^i}(s)$ is twice continuously differentiable with respect to the parameter θ^i over Θ^i. Let P^θ denote the transition matrix of the Markov chain $\{s_t\}_{t\geq 0}$ induced by policy μ_θ, that is, for any $s, s' \in \mathcal{S}$, $P^\theta(s'|s) = P(s'|s, \mu_\theta(s))$. We assume that the Markov chain $\{s_t\}_{t\geq 0}$ is irreducible and aperiodic under any μ_θ and denote its stationary distribution by d^{μ_θ}.

Our objective is to find a policy μ_θ that maximizes the long-run average reward:

$$\max_\theta J(\mu_\theta) = \mathbb{E}_{s\sim d^{\mu_\theta}}[\bar{R}(s,\mu_\theta(s))] = \sum_{s\in\mathcal{S}} d^{\mu_\theta}(s)\bar{R}(s,\mu_\theta(s)).$$

Analogous to the stochastic policy case, we denote the action-value function by $Q_\theta(s,a) = \sum_{t\geq 0}\mathbb{E}[\bar{r}_{t+1} - J(\mu_\theta)|s_0 = s, a_0 = a, \mu_\theta]$, and the state-value function by $V_\theta(s) = Q_\theta(s,\mu_\theta(s))$. When there is no ambiguity, we will denote $J(\mu_\theta)$ and d^{μ_θ} by simply $J(\theta)$ and d^θ, respectively. We present three results for the long-run average reward: (1) an expression for the local deterministic policy gradient in the on-policy setting $\nabla_{\theta^i} J(\mu_\theta)$, (2) an expression for the gradient in the off-policy setting, and (3) we show that the deterministic policy gradient can be seen as the limit of the stochastic one.

On-Policy Setting

Theorem 1 (Local Deterministic Policy Gradient Theorem - On Policy)**.** *For any $\theta \in \Theta$, $i \in \mathcal{N}$, $\nabla_{\theta^i} J(\mu_\theta)$ exists and is given by*

$$\nabla_{\theta^i} J(\mu_\theta) = \mathbb{E}_{s\sim d^{\mu_\theta}}\left[\nabla_{\theta^i}\mu^i_{\theta^i}(s)\nabla_{a^i}\, Q_\theta(s,\mu^{-i}_{\theta^{-i}}(s),a^i)\big|_{a^i=\mu^i_{\theta^i}(s)}\right].$$

The first step of the proof consists in showing that $\nabla_\theta J(\mu_\theta) = \mathbb{E}_{s\sim d^\theta}\left[\nabla_\theta\mu_\theta(s)\nabla_a\, Q_\theta(s,a)\big|_{a=\mu_\theta(s)}\right]$. This is an extension of the well-known stochastic case, for which we have $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim d_\theta}\left[\nabla_\theta \log(\pi_\theta(a|s))Q_\theta(s,a)\right]$, which holds for a long-term averaged return with stochastic policy (e.g Theorem 1 of Sutton et al. [2000a]). See the Appendix for the details.

Off-Policy Setting In the off-policy setting, we are given a behavior policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, and our goal is to maximize the long-run average reward under state distribution d^π:

$$J_\pi(\mu_\theta) = \mathbb{E}_{s\sim d^\pi}\left[\bar{R}(s,\mu_\theta(s))\right] = \sum_{s\in\mathcal{S}} d^\pi(s)\bar{R}(s,\mu_\theta(s)). \tag{1}$$

Note that we consider here an excursion objective (Sutton et al. [2009], Silver et al. [January 2014a], Sutton et al. [2016]) since we take the average over the state distribution of the behaviour policy π of the state-action reward when selecting action given by the target policy μ_θ. We thus have:

Theorem 2 (Local Deterministic Policy Gradient Theorem - Off Policy)**.** *For any $\theta \in \Theta$, $i \in \mathcal{N}$, $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ a fixed stochastic policy, $\nabla_{\theta^i} J_\pi(\mu_\theta)$ exists and is given by*

$$\nabla_{\theta^i} J_\pi(\mu_\theta) = \mathbb{E}_{s\sim d^\pi}\left[\nabla_{\theta^i}\mu^i_{\theta^i}(s)\nabla_{a^i}\, \bar{R}(s,\mu^{-i}_{\theta^{-i}}(s),a^i)\big|_{a^i=\mu^i_{\theta^i}(s)}\right].$$

Proof. Since d^π is independent of θ we can take the gradient on both sides of (1)

$$\nabla_\theta J_\pi(\mu_\theta) = \mathbb{E}_{s\sim d^\pi}\left[\nabla_\theta \mu_\theta(s)\ \nabla_a \bar{R}(s,\mu_\theta(s))\big|_{a=\mu_\theta(s)}\right].$$

Given that $\nabla_{\theta^i}\mu^j_\theta(s) = 0$ if $i \neq j$, we have $\nabla_\theta\mu_\theta(s) = \mathrm{Diag}(\nabla_{\theta^1}\mu^1_{\theta_1}(s), \ldots, \nabla_{\theta^N}\mu^N_{\theta_N}(s))$ and the result follows. □

This result implies that, off-policy, each agent needs access to $\mu^{-i}_{\theta^{-i}_t}(s_t)$ for every t .

Limit Theorem As noted by Silver et al. [January 2014b], the fact that the deterministic gradient is a limit case of the stochastic gradient enables the standard machinery of policy gradient, such as compatible-function approximation (Sutton et al. [2000b]), natural gradients (Kakade [2001]), on-line feature adaptation (Prabuchandran et al. [2016],) and actor-critic (Konda [2002]) to be used with deterministic policies. We show that it holds in our setting. The proof can be found in the Appendix.

Theorem 3 (Limit of the Stochastic Policy Gradient for MARL). *Let $\pi_{\theta,\sigma}$ be a stochastic policy such that $\pi_{\theta,\sigma}(a|s) = \nu_\sigma(\mu_\theta(s), a)$, where σ is a parameter controlling the variance, and ν_σ satisfy Condition 1 in the Appendix. Then,*

$$\lim_{\sigma\downarrow 0} \nabla_\theta J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma}) = \nabla_\theta J_{\mu_\theta}(\mu_\theta)$$

where on the l.h.s the gradient is the standard stochastic policy gradient and on the r.h.s. the gradient is the deterministic policy gradient.

4 Algorithms

We provide two decentralized deterministic actor-critic algorithms, one on-policy and the other off-policy and demonstrate their convergence in the next section; assumptions and proofs are provided in the Appendix.

On-Policy Deterministic Actor-Critic

Algorithm 1 Networked deterministic on-policy actor-critic

Initialize: step $t = 0$; parameters $\hat{J}^i_0, \omega^i_0, \widetilde{\omega}^i_0, \theta^i_0, \forall i \in \mathcal{N}$; state s_0; stepsizes $\{\beta_{\omega,t}\}_{t\geq 0}, \{\beta_{\theta,t}\}_{t\geq 0}$
Draw $a^i_0 = \mu^i_{\theta^i_0}(s_0)$ and compute $\widetilde{a}^i_0 = \nabla_{\theta^i}\mu^i_{\theta^i_0}(s_0)$
Observe joint action $a_0 = (a^1_0, \ldots, a^N_0)$ and $\widetilde{a}_0 = (\widetilde{a}^1_0, \ldots, \widetilde{a}^N_0)$
repeat
  for $i \in \mathcal{N}$ **do**
    Observe s_{t+1} and reward $r^i_{t+1} = r^i(s_t, a_t)$
    Update $\hat{J}^i_{t+1} \leftarrow (1-\beta_{\omega,t})\cdot \hat{J}^i_t + \beta_{\omega,t}\cdot r^i_{t+1}$
    Draw action $a_{t+1} = \mu^i_{\theta^i_t}(s_{t+1})$ and compute $\widetilde{a}^i_{t+1} = \nabla_{\theta^i}\mu^i_{\theta^i_t}(s_{t+1})$
  end for
  Observe joint action $a_{t+1} = (a^1_{t+1}, \ldots, a^N_{t+1})$ and $\widetilde{a}_{t+1} = (\widetilde{a}^1_{t+1}, \ldots, \widetilde{a}^N_{t+1})$
  for $i \in \mathcal{N}$ **do**
    Update: $\delta^i_t \leftarrow r^i_{t+1} - \hat{J}^i_t + \hat{Q}_{\omega^i_t}(s_{t+1}, a_{t+1}) - \hat{Q}_{\omega^i_t}(s_t, a_t)$
    Critic step: $\widetilde{\omega}^i_t \leftarrow \omega^i_t + \beta_{\omega,t}\cdot\delta^i_t\cdot\nabla_\omega \hat{Q}_{\omega^i}(s_t,a_t)\Big|_{\omega=\omega^i_t}$
    Actor step: $\theta^i_{t+1} = \theta^i_t + \beta_{\theta,t}\cdot\nabla_{\theta^i}\mu^i_{\theta^i_t}(s_t)\ \nabla_{a^i}\hat{Q}_{\omega^i_t}(s_t, a^{-i}_t, a^i)\Big|_{a^i=a^i_t}$
    Send $\widetilde{\omega}^i_t$ to the neighbors $\{j \in \mathcal{N} : (i,j) \in \mathcal{E}_t\}$ over $\mathcal{G}_t$
    Consensus step: $\omega^i_{t+1} \leftarrow \sum_{j\in\mathcal{N}} c^{ij}_t \cdot \widetilde{\omega}^j_t$
  end for
  Update $t \leftarrow t+1$
until end

Consider the following on-policy algorithm. The actor step is based on an expression for $\nabla_{\theta^i} J(\mu_\theta)$ in terms of $\nabla_{a^i} Q_\theta$(see Equation (15) in the Appendix). We approximate the action-value function Q_θ using a family of functions $\hat{Q}_\omega : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ parameterized by ω, a column vector in $\mathbb{R}^K$. Each agent i maintains its own parameter ω^i and uses $\hat{Q}_{\omega^i}$ as its local estimate of Q_θ. The parameters ω^i are updated in the critic step using consensus updates through a weight matrix $C_t = \left(c_t^{ij}\right)_{i,j} \in \mathbb{R}^{N \times N}$ where c_t^{ij} is the weight on the message transmitted from i to j at time t, namely:

$$\hat{J}_{t+1}^i = (1 - \beta_{\omega,t}) \cdot \hat{J}_t^i + \beta_{\omega,t} \cdot r_{t+1}^i \tag{2}$$

$$\widetilde{\omega}_t^i = \omega_t^i + \beta_{\omega,t} \cdot \delta_t^i \cdot \nabla_\omega \hat{Q}_{\omega^i}(s_t, a_t)\Big|_{\omega=\omega_t^i} \tag{3}$$

$$\omega_{t+1}^i = \sum_{j \in \mathcal{N}} c_t^{ij} \cdot \widetilde{\omega}_t^j \tag{4}$$

with

$$\delta_t^i = r_{t+1}^i - \hat{J}_t^i + \hat{Q}_{\omega_t^i}(s_{t+1}, a_{t+1}) - \hat{Q}_{\omega_t^i}(s_t, a_t).$$

For the actor step, each agent i improves its policy via:

$$\theta_{t+1}^i = \theta_t^i + \beta_{\theta,t} \cdot \nabla_{\theta^i} \mu_{\theta_t^i}^i(s_t) \cdot \nabla_{a^i} \hat{Q}_{\omega_t^i}(s_t, a_t^{-i}, a^i)\Big|_{a^i = a_t^i}. \tag{5}$$

Since Algorithm 1 is an on-policy algorithm, each agent updates the critic using only (s_t, a_t, s_{t+1}), at time t knowing that $a_{t+1} = \mu_{\theta_t}(s_{t+1})$. The terms in blue are additional terms that need to be shared when using compatible features (this is explained further in the next section).

Off-Policy Deterministic Actor-Critic We further propose an off-policy actor-critic algorithm, defined in Algorithm 2 to enable better exploration capability. Here, the goal is to maximize $J_\pi(\mu_\theta)$ where π is the behavior policy. To do so, the globally averaged reward function $\bar{R}(s,a)$ is approximated using a family of functions $\hat{\bar{R}}_\lambda : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ that are parameterized by λ, a column vector in $\mathbb{R}^K$. Each agent i maintains its own parameter λ^i and uses $\hat{\bar{R}}_{\lambda^i}$ as its local estimate of $\bar{R}$. Based on (1), the actor update is

$$\theta_{t+1}^i = \theta_t^i + \beta_{\theta,t} \cdot \nabla_{\theta^i} \mu_{\theta_t^i}^i(s_t) \cdot \nabla_{a^i} \hat{\bar{R}}_{\lambda_t^i}(s_t, \mu_{\theta_t^{-i}}^{-i}(s_t), a^i)\Big|_{a^i = \mu_{\theta_t^i}(s_t)}, \tag{6}$$

which requires each agent i to have access to $\mu_{\theta_t^j}^j(s_t)$ for $j \in \mathcal{N}$.

The critic update is

$$\widetilde{\lambda}_t^i = \lambda_t^i + \beta_{\lambda,t} \cdot \delta_t^i \cdot \nabla_\lambda \hat{\bar{R}}_{\lambda^i}(s_t, a_t)\Big|_{\lambda=\lambda_t^i} \tag{7}$$

$$\lambda_{t+1}^i = \sum_{j \in \mathcal{N}} c_t^{ij} \widetilde{\lambda}_t^j, \tag{8}$$

with

$$\delta_t^i = r^i(s_t, a_t) - \hat{\bar{R}}_{\lambda_t^i}(s_t, a_t). \tag{9}$$

In this case, δ_t^i was motivated by distributed optimization results, and is not related to the local TD-error (as there is no "temporal" relationship for R). Rather, it is simply the difference between the sample reward and the bootstrap estimate. The terms in blue are additional terms that need to be shared when using compatible features (this is explained further in the next section).

5 Convergence

To show convergence, we use a two-timescale technique where in the actor, updating deterministic policy parameter θ^i occurs more slowly than that of ω^i and $\hat{J}^i$ in the critic. We study the asymptotic behaviour of the critic by freezing the joint policy μ_θ, then study the behaviour of θ_t under convergence of the critic. To ensure stability, projection is often assumed since it is not clear how boundedness of

Algorithm 2 Networked deterministic off-policy actor-critic

Initialize: step $t=0$; parameters $\lambda_0^i, \widetilde{\lambda}_0^i, \theta_0^i, \forall i \in \mathcal{N}$; state s_0; stepsizes $\{\beta_{\lambda,t}\}_{t\geq 0}, \{\beta_{\theta,t}\}_{t\geq 0}$
Draw $a_0^i \sim \pi^i(s_0)$, compute $\dot{a}_0^i = \mu^i_{\theta_0^i}(s_0)$ and $\widetilde{a}_0^i = \nabla_{\theta^i}\mu^i_{\theta_0^i}(s_0)$
Observe joint action $a_0 = (a_0^1, \ldots, a_0^N)$, $\dot{a}_0 = (\dot{a}_0^1, \ldots, \dot{a}_0^N)$ and $\widetilde{a}_0 = (\widetilde{a}_0^1, \ldots, \widetilde{a}_0^N)$
repeat
  for $i \in \mathcal{N}$ **do**
    Observe s_{t+1} and reward $r_{t+1}^i = r^i(s_t, a_t)$
  end for
  for $i \in \mathcal{N}$ **do**
    Update: $\delta_t^i \leftarrow r_{t+1}^i - \hat{\bar{R}}_{\lambda_t^i}(s_t, a_t)$
    Critic step: $\widetilde{\lambda}_t^i \leftarrow \lambda_t^i + \beta_{\lambda,t}\cdot\delta_t^i \cdot \nabla_\lambda \hat{\bar{R}}_{\lambda^i}(s_t,a_t)\Big|_{\lambda=\lambda_t^i}$
    Actor step: $\theta_{t+1}^i = \theta_t^i + \beta_{\theta,t}\cdot\nabla_{\theta^i}\mu^i_{\theta_t^i}(s_t)\cdot\nabla_{a^i}\hat{\bar{R}}_{\lambda_t^i}(s_t, \mu^{-i}_{\theta_t^{-i}}(s_t), a^i)\Big|_{a^i=\mu_{\theta_t^i}(s_t)}$
    Send $\widetilde{\lambda}_t^i$ to the neighbors $\{j \in \mathcal{N} : (i,j) \in \mathcal{E}_t\}$ over $\mathcal{G}_t$
  end for
  for $i \in \mathcal{N}$ **do**
    Consensus step: $\lambda_{t+1}^i \leftarrow \sum_{j\in\mathcal{N}} c_t^{ij}\cdot\widetilde{\lambda}_t^j$
    Draw action $a_{t+1} \sim \pi(s_{t+1})$, compute $\dot{a}_{t+1}^i = \mu^i_{\theta_{t+1}^i}(s_{t+1})$ and compute $\widetilde{a}_{t+1}^i = \nabla_{\theta^i}\mu^i_{\theta_{t+1}^i}(s_{t+1})$
  end for
  Observe joint action $a_{t+1} = (a_{t+1}^1, \ldots, a_{t+1}^N)$, $\dot{a}_{t+1} = (\dot{a}_{t+1}^1, \ldots, \dot{a}_{t+1}^N)$ and $\widetilde{a}_{t+1} = (\widetilde{a}_{t+1}^1, \ldots, \widetilde{a}_{t+1}^N)$
  Update $t \leftarrow t+1$
until end

$\{\theta_t^i\}$ can otherwise be ensured (see Bhatnagar et al. [2009]). However, in practice, convergence is typically observed even without the projection step (see Bhatnagar et al. [2009], Degris et al. [2012], Prabuchandran et al. [2016], Zhang et al. [2018], Suttle et al. [2019]). We also introduce the following technical assumptions which will be needed in the statement of the convergence results.

Assumption 1 (Linear approximation, average-reward)**.** For each agent i, the average-reward function $\bar{R}$ is parameterized by the class of linear functions, i.e., $\hat{\bar{R}}_{\lambda^i,\theta}(s,a) = w_\theta(s,a)\cdot\lambda^i$ where $w_\theta(s,a) = \big[w_{\theta,1}(s,a), \ldots, w_{\theta,K}(s,a)\big] \in \mathbb{R}^K$ is the feature associated with the state-action pair (s,a). The feature vectors $w_\theta(s,a)$, as well as $\nabla_a w_{\theta,k}(s,a)$ are uniformly bounded for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $k \in [\![1,K]\!]$. Furthermore, we assume that the feature matrix $W_\pi \in \mathbb{R}^{|S|\times K}$ has full column rank, where the k-th column of $W_{\pi,\theta}$ is $\big[\int_{\mathcal{A}}\pi(a|s)w_{\theta,k}(s,a)\mathrm{d}a, s\in\mathcal{S}\big]$ for any $k \in [\![1,K]\!]$.

Assumption 2 (Linear approximation, action-value)**.** For each agent i, the action-value function is parameterized by the class of linear functions, i.e., $\hat{Q}_{\omega^i}(s,a) = \phi(s,a)\cdot\omega^i$ where $\phi(s,a) = \big[\phi_1(s,a), \ldots, \phi_K(s,a)\big] \in \mathbb{R}^K$ is the feature associated with the state-action pair (s,a). The feature vectors $\phi(s,a)$, as well as $\nabla_a\phi_k(s,a)$ are uniformly bounded for any $s\in\mathcal{S}$, $a\in\mathcal{A}$, $k\in\{1,\ldots,K\}$. Furthermore, we assume that for any $\theta\in\Theta$, the feature matrix $\Phi_\theta \in \mathbb{R}^{|\mathcal{S}|\times K}$ has full column rank, where the k-th column of Φ_θ is $\big[\phi_k(s,\mu_\theta(s)), s\in\mathcal{S}\big]$ for any $k\in[\![1,K]\!]$. Also, for any $u\in\mathbb{R}^K$, $\Phi_\theta u \neq \mathbf{1}$.

Assumption 3 (Bounding θ)**.** The update of the policy parameter θ^i includes a local projection by $\Gamma^i : \mathbb{R}^{m_i} \to \Theta^i$ that projects any θ_t^i onto a compact set Θ^i that can be expressed as $\{\theta^i | q_j^i(\theta^i) \leq 0,\ j = 1, \ldots, s^i\} \subset \mathbb{R}^{m_i}$, for some real-valued, continuously differentiable functions $\{q_j^i\}_{1\leq j\leq s^i}$ defined on $\mathbb{R}^{m_i}$. We also assume that $\Theta = \prod_{i=1}^N \Theta^i$ is large enough to include at least one local minimum of $J(\theta)$.

We use $\{\mathcal{F}_t\}$ to denote the filtration with $\mathcal{F}_t = \sigma(s_\tau, C_{\tau-1}, a_{\tau-1}, r_{\tau-1}, \tau\leq t)$.

Assumption 4 (Random matrices)**.** The sequence of non-negative random matrices $\{C_t = (c_t^{ij})_{ij}\}$ satisfies:

1. C_t is row stochastic and $\mathbb{E}(C_t|\mathcal{F}_t)$ is a.s. column stochastic for each t, i.e., $C_t\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top \mathbb{E}(C_t|\mathcal{F}_t) = \mathbf{1}^\top$ a.s. Furthermore, there exists a constant $\eta \in (0,1)$ such that, for any $c_t^{ij} > 0$, we have $c_t^{ij} \geq \eta$.
2. C_t respects the communication graph $\mathcal{G}_t$, i.e., $c_t^{ij} = 0$ if $(i,j) \notin \mathcal{E}_t$.
3. The spectral norm of $\mathbb{E}\big[C_t^\top \cdot (I - \mathbf{1}\mathbf{1}^\top/N) \cdot C_t\big]$ is smaller than one.
4. Given the σ-algebra generated by the random variables before time t, C_t, is conditionally independent of s_t, a_t and r_{t+1}^i for any $i \in \mathcal{N}$.

Assumption 5 (Step size rules, on-policy)**.** The stepsizes $\beta_{\omega,t}, \beta_{\theta,t}$ satisfy:

$$\sum_t \beta_{\omega,t} = \sum_t \beta_{\theta,t} = \infty$$
$$\sum_t (\beta_{\omega,t}^2 + \beta_{\theta,t}^2) < \infty$$
$$\sum_t |\beta_{\theta,t+1} - \beta_{\theta,t}| < \infty.$$

In addition, $\beta_{\theta,t} = o(\beta_{\omega,t})$ and $\lim_{t\to\infty} \beta_{\omega,t+1}/\beta_{\omega,t} = 1$.

Assumption 6 (Step size rules, off-policy)**.** The step-sizes $\beta_{\lambda,t}$, $\beta_{\theta,t}$ satisfy:

$$\sum_t \beta_{\lambda,t} = \sum_t \beta_{\theta,t} = \infty, \qquad \sum_t \beta_{\lambda,t}^2 + \beta_{\theta,t}^2 < \infty$$
$$\beta_{\theta,t} = o(\beta_{\lambda,t}), \qquad \lim_{t\to\infty} \beta_{\lambda,t+1}/\beta_{\lambda,t} = 1.$$

On-Policy Convergence To state convergence of the critic step, we define $D_\theta^s = \mathrm{Diag}\big[d^\theta(s), s \in \mathcal{S}\big]$, $\bar{R}_\theta = \big[\bar{R}(s, \mu_\theta(s)), s \in \mathcal{S}\big]^\top \in \mathbb{R}^{|\mathcal{S}|}$ and the operator $T_\theta^Q : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ for any action-value vector $Q \in \mathbb{R}^{|\mathcal{S}|}$ (and not $\mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ since there is a mapping associating an action to each state) as:

$$T_\theta^Q(Q') = \bar{R}_\theta - J(\mu_\theta) \cdot \mathbf{1} + P^\theta Q'.$$

Theorem 4. *Under Assumptions 3, 4, and 5, for any given deterministic policy μ_θ, with $\{\hat{J}_t\}$ and $\{\omega_t\}$ generated from (2), we have $\lim_{t\to\infty} \frac{1}{N}\sum_{i\in\mathcal{N}} \hat{J}_t^i = J(\mu_\theta)$ and $\lim_{t\to\infty} \omega_t^i = \omega_\theta$ a.s. for any $i \in \mathcal{N}$, where*

$$J(\mu_\theta) = \sum_{s\in\mathcal{S}} d^\theta(s) \bar{R}(s, \mu_\theta(s))$$

is the long-term average return under μ_θ, and ω_θ is the unique solution to

$$\Phi_\theta{}^\top D_\theta^s \big[T_\theta^Q(\Phi_\theta \omega_\theta) - \Phi_\theta \omega_\theta\big] = 0. \tag{10}$$

Moreover, ω_θ is the minimizer of the Mean Square Projected Bellman Error (MSPBE), i.e., the solution to

$$\underset{\omega}{\text{minimize}}\ \|\Phi_\theta \omega - \Pi T_\theta^Q(\Phi_\theta \omega)\|_{D_\theta^s}^2,$$

where Π is the operator that projects a vector to the space spanned by the columns of Φ_θ, and $\|\cdot\|_{D_\theta^s}^2$ denotes the euclidean norm weighted by the matrix D_θ^s.

To state convergence of the actor step, we define quantities $\psi_{t,\theta}^i$, ξ_t^i and $\xi_{t,\theta}^i$ as

$$\psi_{t,\theta}^i = \nabla_{\theta^i} \mu_{\theta^i}^i(s_t) \quad \text{and} \quad \psi_t^i = \psi_{t,\theta_t}^i = \nabla_{\theta^i} \mu_{\theta_t^i}^i(s_t),$$
$$\xi_{t,\theta}^i = \nabla_{a_i} \hat{Q}_{\omega_\theta}(s_t, a_t^{-i}, a_i)\Big|_{a_i = a_i = \mu_{\theta_t^i}^i(s_t)} = \nabla_{a_i} \phi(s_t, a_t^{-i}, a_i)\big|_{a_i = a_i = \mu_{\theta_t^i}^i(s_t)} \omega_\theta,$$
$$\xi_t^i = \nabla_{a_i} \hat{Q}_{\omega_t^i}(s_t, a_t^{-i}, a_i)\Big|_{a_i = \mu_{\theta^i}^i(s_t)} = \nabla_{a_i} \phi(s_t, a_t^{-i}, a_i)\big|_{a_i = \mu_{\theta^i}^i(s_t)} \omega_t^i.$$

Additionally, we introduce the operator $\hat{\Gamma}(\cdot)$ as

$$\hat{\Gamma}^i\left[g(\theta)\right] = \lim_{0<\eta\to 0} \frac{\Gamma^i\left[\theta^i + \eta \cdot g(\theta)\right] - \theta^i}{\eta} \tag{11}$$

for any $\theta \in \Theta$ and $g : \Theta \to \mathbb{R}^{m_i}$ a continuous function. In case the limit above is not unique we take $\hat{\Gamma}^i\left[g(\theta)\right]$ to be the set of all possible limit points of (11).

Theorem 5. *Under Assumptions 2, 3, 4, and 5, the policy parameter θ_t^i obtained from (5) converges a.s. to a point in the set of asymptotically stable equilibria of*

$$\dot{\theta}^i = \hat{\Gamma}^i\left[\mathbb{E}_{s_t\sim d^\theta,\mu_\theta}\left[\psi_{t,\theta}^i \cdot \xi_{t,\theta}^i\right]\right], \quad \text{for any } i \in \mathcal{N}. \tag{12}$$

In the case of multiple limit points, the above is treated as a differential inclusion rather than an ODE.

The convergence of the critic step can be proved by taking similar steps as that in Zhang et al. [2018]. For the convergence of the actor step, difficulties arise from the projection (which is handled using Kushner-Clark Lemma Kushner and Clark [1978]) and the state-dependent noise (that is handled by "natural" timescale averaging Crowder [2009]). Details are provided in the Appendix.

Remark. Note that that with a linear function approximator Q_θ, $\psi_{t,\theta} \cdot \xi_{t,\theta} = \nabla_\theta \mu_\theta(s_t)\, \nabla_a \hat{Q}_{\omega_\theta}(s_t, a)\Big|_{a=\mu_\theta(s_t)}$ may not be an unbiased estimate of $\nabla_\theta J(\theta)$:

$$\mathbb{E}_{s\sim d^\theta}\left[\psi_{t,\theta}\cdot\xi_{t,\theta}\right] = \nabla_\theta J(\theta) + \mathbb{E}_{s\sim d^\theta}\left[\nabla_\theta\mu_\theta(s)\cdot\left(\nabla_a \hat{Q}_{\omega_\theta}(s,a)\Big|_{a=\mu_\theta(s)} - \nabla_a Q_{\omega_\theta}(s,a)\big|_{a=\mu_\theta(s)}\right)\right].$$

A standard approach to overcome this approximation issue is via compatible features (see, for example, Silver et al. [January 2014a] and Zhang and Zavlanos [2019]), i.e. $\phi(s,a) = a \cdot \nabla_\theta \mu_\theta(s)^\top$, giving, for $\omega \in \mathbb{R}^m$,

$$\hat{Q}_\omega(s,a) = a\cdot\nabla_\theta\mu_\theta(s)^\top\omega = (a - \mu_\theta(s))\cdot\nabla_\theta\mu_\theta(s)^\top\omega + \hat{V}_\omega(s),$$
$$\text{with } \hat{V}_\omega(s) = \hat{Q}_\omega(s,\mu_\theta(s)) \;\text{ and }\; \nabla_a \hat{Q}_\omega(s,a)\Big|_{a=\mu_\theta(s)} = \nabla_\theta\mu_\theta(s)^\top\omega.$$

We thus expect that the convergent point of (5) corresponds to a small neighborhood of a local optimum of $J(\mu_\theta)$, i.e., $\nabla_{\theta^i} J(\mu_\theta) = 0$, provided that the error for the gradient of the action-value function $\nabla_a \hat{Q}_\omega(s,a)\Big|_{a=\mu_\theta(s)} - \nabla_a Q_\theta(s,a)\big|_{a=\mu_\theta(s)}$ is small. However, note that using compatible features requires computing, at each step t, $\phi(s_t,a_t) = a_t \cdot \nabla_\theta\mu_\theta(s_t)^\top$. Thus, in Algorithm 1, each agent observes not only the joint action $a_{t+1} = (a_{t+1}^1, \dots, a_{t+1}^N)$ but also $(\nabla_{\theta^1}\mu_{\theta_t^1}^1(s_{t+1}), \dots, \nabla_{\theta^N}\mu_{\theta_t^N}^N(s_{t+1}))$ (see the parts in blue in Algorithm 1).

Off-Policy Convergence

Theorem 6. *Under Assumptions 1, 4, and 6, for any given behavior policy π and any $\theta \in \Theta$, with $\{\lambda_t^i\}$ generated from (7), we have $\lim_{t\to\infty}\lambda_t^i = \lambda_\theta$ a.s. for any $i \in \mathcal{N}$, where λ_θ is the unique solution to*

$$B_{\pi,\theta}\cdot\lambda_\theta = A_{\pi,\theta}\cdot d_\pi^s \tag{13}$$

where $d_\pi^s = \left[d^\pi(s), s \in \mathcal{S}\right]^\top$, $A_{\pi,\theta} = \left[\int_{\mathcal{A}}\pi(a|s)\bar{R}(s,a)w(s,a)^\top \mathrm{d}a, s \in \mathcal{S}\right] \in \mathbb{R}^{K\times|\mathcal{S}|}$ and $B_{\pi,\theta} = \left[\sum_{s\in\mathcal{S}} d^\pi(s)\int_{\mathcal{A}}\pi(a|s)w_i(s,a)\cdot w(s,a)^\top\mathrm{d}a, 1 \le i \le K\right] \in \mathbb{R}^{K\times K}$.

From here on we let

$$\xi_{t,\theta}^i = \nabla_{a_i}\hat{\bar{R}}_{\lambda_\theta}(s_t, \mu_{\theta_t^{-i}}^{-i}(s_t), a_i)\Big|_{a_i=\mu_{\theta_t^i}^i(s_t)} = \nabla_{a_i} w(s_t, \mu_{\theta_t^{-i}}^{-i}(s_t), a_i)\Big|_{a_i=\mu_{\theta_t^i}^i(s_t)}\lambda_\theta$$

$$\xi_t^i = \nabla_{a_i}\hat{\bar{R}}_{\lambda_t^i}(s_t, \mu_{\theta_t^{-i}}^{-i}(s_t), a_i)\Big|_{a_i=\mu_{\theta_t^i}^i(s_t)} = \nabla_{a_i} w(s_t, \mu_{\theta^{-i}}^{-i}(s_t), a_i)\Big|_{a_i=\mu_{\theta^i}^i(s_t)}\lambda_t^i$$

and we keep

$$\psi_{t,\theta}^i = \nabla_{\theta^i}\mu_{\theta^i}^i(s_t), \quad \text{and} \quad \psi_t^i = \psi_{t,\theta_t}^i = \nabla_{\theta^i}\mu_{\theta_t^i}^i(s_t).$$

Theorem 7. *Under Assumptions 1, 3, 4, and 6, the policy parameter θ_t^i obtained from (6) converges a.s. to a point in the asymptotically stable equilibria of*

$$\dot{\theta}^i = \Gamma^i \left[\mathbb{E}_{s\sim d^\pi} \left[\psi_{t,\theta}^i \cdot \xi_{t,\theta}^i \right] \right]. \tag{14}$$

We define compatible features for the action-value and the average-reward function in an analogous manner: $w_\theta(s,a) = (a - \mu_\theta(s)) \cdot \nabla_\theta \mu_\theta(s)^\top$. For $\lambda \in \mathbb{R}^m$,

$$\hat{\hat{R}}_{\lambda,\theta}(s,a) = (a - \mu_\theta(s)) \cdot \nabla_\theta \mu_\theta(s)^\top \cdot \lambda$$
$$\nabla_a \hat{\hat{R}}_{\lambda,\theta}(s,a) = \nabla_\theta \mu_\theta(s)^\top \cdot \lambda$$

and we have that, for $\lambda^* = \underset{\lambda}{\operatorname{argmin}}\, \mathbb{E}_{s\sim d^\pi} \left[\|\nabla_a \hat{\hat{R}}_{\lambda,\theta}(s,\mu_\theta(s)) - \nabla_a \bar{R}(s,\mu_\theta(s))\|^2 \right]$:

$$\nabla_\theta J_\pi(\mu_\theta) = \mathbb{E}_{s\sim d^\pi} \left[\nabla_\theta \mu_\theta(s) \cdot \nabla_a \bar{R}(s,a)\big|_{a=\mu_\theta(s)} \right] = \mathbb{E}_{s\sim d^\pi} \left[\nabla_\theta \mu_\theta(s) \cdot \nabla_a \hat{\hat{R}}_{\lambda^*,\theta}(s,a)\Big|_{a=\mu_\theta(s)} \right].$$

The use of compatible features requires each agent to observe not only the joint action taken $a_{t+1} = (a_{t+1}^1, \ldots, a_{t+1}^N)$ and the "on-policy action" $\dot{a}_{t+1} = (\dot{a}_{t+1}^1, \ldots, \dot{a}_{t+1}^N)$, but also $\widetilde{a}_{t+1} = (\nabla_{\theta^1} \mu_{\theta_t^1}^1(s_{t+1}), \ldots, \nabla_{\theta^N} \mu_{\theta_t^N}^N(s_{t+1}))$ (see the parts in blue in Algorithm 2).

We illustrate algorithm convergence on multi-agent extension of a continuous bandit problem from Sec. 5.1 of Silver et al. [January 2014b]. Details are in the Appendix. Figure 2 shows the convergence of Algorithms 1 and 2 averaged over 5 runs. In all cases, the system converges and the agents are able to coordinate their actions to minimize system cost.

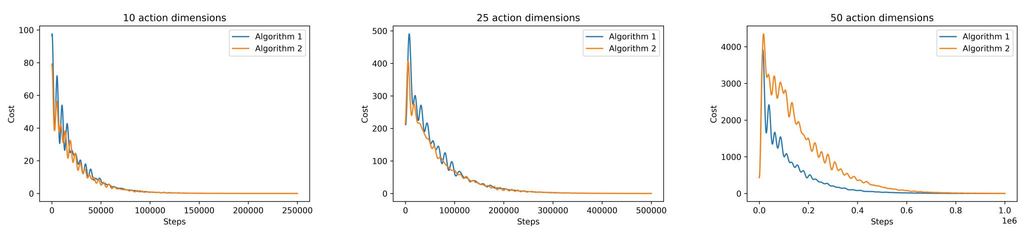

Figure 1: Convergence of Algorithms 1 and 2 on the multi-agent continuous bandit problem.

6 Conclusion

We have provided the tools needed to implement decentralized, deterministic actor-critic algorithms for cooperative multi-agent reinforcement learning. We provide the expressions for the policy gradients, the algorithms themselves, and prove their convergence in on-policy and off-policy settings. We also provide numerical results for a continuous multi-agent bandit problem that demonstrates the convergence of our algorithms. Our work differs from Zhang and Zavlanos [2019] as the latter was based on policy consensus whereas ours is based on critic consensus. Our approach represents agreement between agents on every participants' contributions to the global reward, and as such, provides a consensus scoring function with which to evaluate agents. Our approach may be used in compensation schemes to incentivize participation. An interesting extension of this work would be to prove convergence of our actor-critic algorithm for continuous state spaces, as it may hold with assumptions on the geometric ergodicity of the stationary state distribution induced by the deterministic policies (see Crowder [2009]). The expected policy gradient (EPG) of Ciosek and Whiteson [2018], a hybrid between stochastic and deterministic policy gradient, would also be interesting to leverage. The Multi-Agent Deep Deterministic Policy Gradient algorithm (MADDPG) of Lowe et al. [2017] assumes partial observability for each agent and would be a useful extension, but it is likely difficult to extend our convergence guarantees to the partially observed setting.

References

Albert Benveniste, Pierre Priouret, and Michel Métivier. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, Berlin, Heidelberg, 1990. ISBN 0-387-52894-6.

Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, November 2009. ISSN 0005-1098. doi: 10.1016/j.automatica.2009.07.008. URL `http://dx.doi.org/10.1016/j.automatica.2009.07.008`.

Kamil Ciosek and Shimon Whiteson. Expected Policy Gradients for Reinforcement Learning. *arXiv e-prints*, art. arXiv:1801.03326, Jan 2018.

Martin Crowder. Stochastic approximation: A dynamical systems viewpoint by vivek s. borkar. *International Statistical Review*, 77(2):306–306, 2009.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. *CoRR*, abs/1205.4839, 2012. URL `http://arxiv.org/abs/1205.4839`.

Scott Fujimoto, Herke van Hoof, and Dave Meger. Addressing function approximation error in actor-critic methods. *CoRR*, abs/1802.09477, 2018. URL `http://arxiv.org/abs/1802.09477`.

Sham Kakade. A natural policy gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 1531–1538, Cambridge, MA, USA, 2001. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2980539.2980738`.

Vijaymohan Konda. *Actor-critic Algorithms*. PhD thesis, Cambridge, MA, USA, 2002. AAI0804543.

Harold J. (Harold Joseph) Kushner and (joint author.) Clark, Dean S. *Stochastic approximation methods for constrained and unconstrained systems*. New York : Springer-Verlag, 1978. ISBN 0387903410.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Neural Information Processing Systems (NIPS)*, 2017.

Hamid Reza Maei. Convergent actor-critic algorithms under off-policy training and function approximation. *CoRR*, abs/1802.07842, 2018. URL `http://arxiv.org/abs/1802.07842`.

P. Marbach and J. N. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, Feb 2001. ISSN 0018-9286. doi: 10.1109/9.905687.

K. J. Prabuchandran, Shalabh Bhatnagar, and Vivek S. Borkar. Actor-critic algorithms with online feature adaptation. *ACM Trans. Model. Comput. Simul.*, 26(4):24:1–24:26, February 2016. ISSN 1049-3301. doi: 10.1145/2868723. URL `http://doi.acm.org/10.1145/2868723`.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994. ISBN 0471619779.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic Policy Gradient Algorithms. *International Conference on Machine Learning*, pages 387–395, January 2014a.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic Policy Gradient Algorithms. *International Conference on Machine Learning*, pages 387–395, January 2014b.

Wesley Suttle, Zhuoran Yang, Kaiqing Zhang, Zhaoran Wang, Tamer Basar, and Ji Liu. A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning. *CoRR*, abs/1903.06372, 2019. URL `http://arxiv.org/abs/1903.06372`.

Richard S Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000a.

Richard S Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000b.

Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 993–1000, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

Richard S. Sutton, A. Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *J. Mach. Learn. Res.*, 17(1):2603–2631, January 2016. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2946645.3007026`.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. 80:5872–5881, 10–15 Jul 2018.

Yan Zhang and Michael M. Zavlanos. Distributed off-policy actor-critic reinforcement learning with policy consensus. *CoRR*, abs/1903.09255, 2019.

Numerical experiment details

We demonstrate the convergence of our algorithm in a continuous bandit problem that is a multi-agent extension of the experiment in Section 5.1 of Silver et al. (2014). Each agent chooses an action $a^i \in \mathbb{R}^m$. We assume all agents have the same reward function given by $R^i(a) = -\left(\sum_i a^i - a^*\right)^\mathsf{T} C \left(\sum_i a^i - a^*\right)$. The matrix C is positive definite with eigenvalues chosen from $\{0.1, 1\}$, and $a^* = [4, \ldots, 4]^\mathsf{T}$. We consider 10 agents and action dimensions $m = 10, 20, 50$. Note that there are multiple possible solutions for this problem, requiring the agents to coordinate their actions to sum to a^*. We assume a target policy of the form $\mu_{\theta^i} = \theta^i$ for each agent i and a Gaussian behaviour policy $\beta(\cdot) \sim \mathcal{N}(\theta^i, \sigma_\beta^2)$ where $\sigma_\beta = 0.1$. We use the Gaussian behaviour policy for both Algorithms 1 and 2. Strictly speaking, Algorithm 1 is on-policy, but in this simplified setting where the target policy is constant, the on-policy version would be degenerate such that the Q estimate does not affect the TD-error. Therefore, we add a Gaussian behaviour policy to Algorithm 1. Each agent maintains an estimate $Q^{\omega^i}(a)$ of the critic using a linear function of the compatible features $a - \theta$ and a bias feature. The critic is recomputed from each successive batch of $2m$ steps and the actor is updated once per batch. The critic step size is 0.1 and the actor step size is 0.01. Performance is evaluated by measuring the cost of the target policy (without exploration). Figure 2 shows the convergence of Algorithms 1 and 2 averaged over 5 runs. In all cases, the system converges and the agents are able to coordinate their actions to minimize system cost. The jupyter notebook will be made available for others to use. In fact, in this simple experiment, we also observe convergence under discounted rewards.

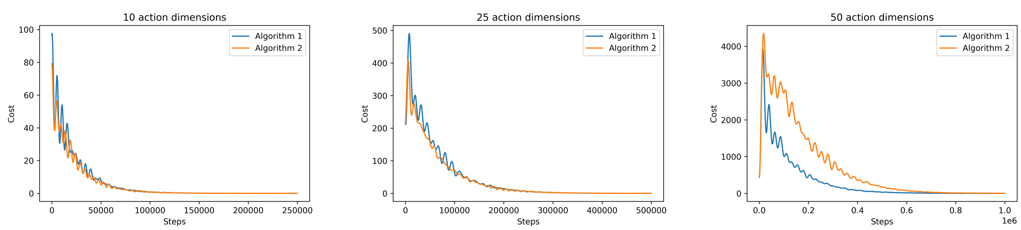

Figure 2: Convergence of Algorithms 1 and 2 on the multi-agent continuous bandit problem.

Proof of Theorem 1

The proof follows the same scheme as Sutton et al. [2000a], naturally extending their results for a deterministic policy μ_θ and a continuous action space $\mathcal{A}$.

Note that our regularity assumptions ensure that, for any $s \in \mathcal{S}$, $V_\theta(s)$, $\nabla_\theta V_\theta(s)$, $J(\theta)$, $\nabla_\theta J(\theta)$, d^θ(s) are Lipschitz-continuous functions of θ (since μ_θ is twice continuously differentiable and Θ is compact), and that $Q_\theta(s, a)$ and $\nabla_a Q_\theta(s, a)$ are Lipschitz-continuous functions of a (Marbach and Tsitsiklis [2001]).

We first show that $\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d^\theta}\left[\nabla_\theta \mu_\theta(s) \nabla_a \left. Q_\theta(s, a)\right|_{a=\mu_\theta(s)}\right]$.

The Poisson equation under policy μ_θ is given by Puterman [1994]

$$Q_\theta(s, a) = \bar{R}(s, a) - J(\theta) + \sum_{s' \in \mathcal{S}} P(s'|s, a) V_\theta(s').$$

So,

$$
\begin{aligned}
\nabla_\theta V_\theta(s) &= \nabla_\theta Q_\theta(s, \mu_\theta(s)) \\
&= \nabla_\theta \big[\bar{R}(s, \mu_\theta(s)) - J(\theta) + \sum_{s' \in \mathcal{S}} P(s'|s, \mu_\theta(s)) V_\theta(s')\big] \\
&= \nabla_\theta \mu_\theta(s) \; \nabla_a \bar{R}(s,a)\big|_{a=\mu_\theta(s)} - \nabla_\theta J(\theta) + \nabla_\theta \sum_{s' \in \mathcal{S}} P(s'|s, \mu_\theta(s)) V_\theta(s') \\
&= \nabla_\theta \mu_\theta(s) \; \nabla_a \bar{R}(s,a)\big|_{a=\mu_\theta(s)} - \nabla_\theta J(\theta) \\
&\quad + \sum_{s' \in \mathcal{S}} \nabla_\theta \mu_\theta(s) \; \nabla_a P(s'|s,a)\big|_{a=\mu_\theta(s)} V_\theta(s') + \sum_{s' \in \mathcal{S}} P(s'|s, \mu_\theta(s)) \nabla_\theta V_\theta(s') \\
&= \nabla_\theta \mu_\theta(s) \nabla_a \left[\bar{R}(s,a) + \sum_{s' \in \mathcal{S}} P(s|s',a) V_\theta(s')\right]\Bigg|_{a=\mu_\theta(s)} \\
&\quad - \nabla_\theta J(\theta) + \sum_{s' \in \mathcal{S}} P(s'|s, \mu_\theta(s)) \nabla_\theta V_\theta(s') \\
&= \nabla_\theta \mu_\theta(s) \nabla_a \; Q_\theta(s,a)\big|_{a=\mu_\theta(s)} + \sum_{s' \in \mathcal{S}} P(s'|s, \mu_\theta(s)) \nabla_\theta V_\theta(s') - \nabla_\theta J(\theta)
\end{aligned}
$$

Hence,

$$
\nabla_\theta J(\theta) = \nabla_\theta \mu_\theta(s) \nabla_a \; Q_\theta(s,a)\big|_{a=\mu_\theta(s)} + \sum_{s' \in \mathcal{S}} P(s'|s, \mu_\theta(s)) \nabla_\theta V_\theta(s') - \nabla_\theta V_\theta(s)
$$

$$
\begin{aligned}
\sum_{s \in \mathcal{S}} d^\theta(s) \nabla_\theta J(\theta) &= \sum_{s \in \mathcal{S}} d^\theta(s) \nabla_\theta \mu_\theta(s) \nabla_a \; Q_\theta(s,a)\big|_{a=\mu_\theta(s)} \\
&\quad + \sum_{s \in \mathcal{S}} d^\theta(s) \sum_{s' \in \mathcal{S}} P(s'|s, \mu_\theta(s)) \nabla_\theta V_\theta(s') - \sum_{s \in \mathcal{S}} d^\theta(s) \nabla_\theta V_\theta(s).
\end{aligned}
$$

Using stationarity property of d^θ, we get

$$
\sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} d^\theta(s) P(s'|s, \mu_\theta(s)) \nabla_\theta V_\theta(s') = \sum_{s' \in \mathcal{S}} d^\theta(s') \nabla_\theta V_\theta(s').
$$

Therefore, we get

$$
\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} d^\theta(s) \nabla_\theta \mu_\theta(s) \; \nabla_a Q_\theta(s,a)\big|_{a=\mu_\theta(s)} = \mathbb{E}_{s \sim d^\theta}\big[\nabla_\theta \mu_\theta(s) \; \nabla_a Q_\theta(s,a)\big|_{a=\mu_\theta(s)}\big].
$$

Given that $\nabla_{\theta^i} \mu_\theta^j(s) = 0$ if $i \neq j$, we have $\nabla_\theta \mu_\theta(s) = \mathrm{Diag}(\nabla_{\theta^1} \mu^1_{\theta_1}(s), \ldots, \nabla_{\theta^N} \mu^N_{\theta_N}(s))$, which implies

$$
\nabla_{\theta^i} J(\theta) = \mathbb{E}_{s \sim d^\theta}\big[\nabla_{\theta^i} \mu^i_{\theta^i}(s) \nabla_{a^i} \; Q_\theta(s, \mu^{-i}_{\theta^{-i}}(s), a^i)\big|_{a^i = \mu^i_{\theta^i}(s)}\big]. \tag{15}
$$

Proof of Theorem 3

We extend the notation for off-policy reward function to stochastic policies as follows. Let β be a behavior policy under which $\{s_t\}_{t \geq 0}$ is irreducible and aperiodic, with stationary distribution d^β. For a stochastic policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, we define

$$
J_\beta(\pi) = \sum_{s \in \mathcal{S}} d^\beta(s) \int_{\mathcal{A}} \pi(a|s) \bar{R}(s,a) \mathrm{d}a.
$$

Recall that for a deterministic policy $\mu : \mathcal{S} \to \mathcal{A}$, we have

$$
J_\beta(\mu) = \sum_{s \in \mathcal{S}} d^\beta(s) \bar{R}(s, \mu(s)).
$$

We introduce the following conditions which are identical to **Conditions B1** from Silver et al. [January 2014a].

Conditions 1. Functions ν_σ parametrized by σ are said to be regular delta-approximation on $\mathcal{R} \subset \mathcal{A}$ if they satisfy the following conditions:

1. The distributions ν_σ converge to a delta distribution: $\lim_{\sigma \downarrow 0} \int_{\mathcal{A}} \nu_\sigma(a', a) f(a) \mathrm{d}a = f(a')$ for $a' \in \mathcal{R}$ and suitably smooth f. Specifically we require that this convergence is uniform in a' and over any class $\mathcal{F}$ of L-Lipschitz and bounded functions, $\|\nabla_a f(a)\| < L < \infty$, $\sup_a f(a) < b < \infty$, i.e.:
$$\lim_{\sigma \downarrow 0} \sup_{f \in \mathcal{F}, a' \in \mathcal{R}} \left| \int_{\mathcal{A}} \nu_\sigma(a', a) f(a) \mathrm{d}a - f(a') \right| = 0.$$
2. For each $a' \in \mathcal{R}$, $\nu_\sigma(a', \cdot)$ is supported on some compact $\mathcal{C}_{a'} \subseteq \mathcal{A}$ with Lipschitz boundary $\mathrm{bd}(\mathcal{C}_{a'})$, vanishes on the boundary and is continuously differentiable on $\mathcal{C}_{a'}$.
3. For each $a' \in \mathcal{R}$, for each $a \in \mathcal{A}$, the gradient $\nabla_{a'} \nu_\sigma(a', a)$ exists.
4. Translation invariance: for all $a \in \mathcal{A}, a' \in \mathcal{R}$, and any $\delta \in \mathbb{R}^n$ such that $a + \delta \in \mathcal{A}$, $a' + \delta \in \mathcal{A}$, $\nu_\sigma(a', a) = \nu_\sigma(a' + \delta, a + \delta)$.

The following lemma is an immediate corollary of **Lemma 1** from Silver et al. [January 2014a].

Lemma 1. *Let ν_σ be a regular delta-approximation on $\mathcal{R} \subseteq \mathcal{A}$. Then, wherever the gradients exist*
$$\nabla_{a'} \nu(a', a) = -\nabla_a \nu(a', a).$$

Theorem 3 is a less technical restatement of the following result.

Theorem 8. *Let $\mu_\theta : \mathcal{S} \to \mathcal{A}$. Denote the range of μ_θ by $\mathcal{R}_\theta \subseteq \mathcal{A}$, and $\mathcal{R} = \cup_\theta \mathcal{R}_\theta$. For each θ, consider $\pi_{\theta,\sigma}$ a stochastic policy such that $\pi_{\theta,\sigma}(a|s) = \nu_\sigma(\mu_\theta(s), a)$, where ν_σ satisfy Conditions 1 on $\mathcal{R}$. Then, there exists $r > 0$ such that, for each $\theta \in \Theta$, $\sigma \mapsto J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma})$, $\sigma \mapsto J_{\pi_{\theta,\sigma}}(\mu_\theta)$, $\sigma \mapsto \nabla_\theta J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma})$, and $\sigma \mapsto \nabla_\theta J_{\pi_{\theta,\sigma}}(\mu_\theta)$ are properly defined on $[0, r]$ (with $J_{\pi_{\theta,0}}(\pi_{\theta,0}) = J_{\pi_{\theta,0}}(\mu_\theta) = J_{\mu_\theta}(\mu_\theta)$ and $\nabla_\theta J_{\pi_{\theta,0}}(\pi_{\theta,0}) = \nabla_\theta J_{\pi_{\theta,0}}(\mu_\theta) = \nabla_\theta J_{\mu_\theta}(\mu_\theta)$), and we have:*
$$\lim_{\sigma \downarrow 0} \nabla_\theta J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma}) = \lim_{\sigma \downarrow 0} \nabla_\theta J_{\pi_{\theta,\sigma}}(\mu_\theta) = \nabla_\theta J_{\mu_\theta}(\mu_\theta).$$

To prove this result, we first state and prove the following Lemma.

Lemma 2. *There exists $r > 0$ such that, for all $\theta \in \Theta$ and $\sigma \in [0, r]$, stationary distribution $d^{\pi_{\theta,\sigma}}$ exists and is unique. Moreover, for each $\theta \in \Theta$, $\sigma \mapsto d^{\pi_{\theta,\sigma}}$ and $\sigma \mapsto \nabla_\theta d^{\pi_{\theta,\sigma}}$ are properly defined on $[0, r]$ and both are continuous at 0.*

Proof of Lemma 2. For any policy β, we let $\left(P^\beta_{s,s'}\right)_{s,s' \in \mathcal{S}}$ be the transition matrix associated to the Markov Chain $\{s_t\}_{t \geq 0}$ induced by β. In particular, for each $\theta \in \Theta$, $\sigma > 0$, $s, s' \in \mathcal{S}$, we have
$$\begin{aligned} P^{\mu_\theta}_{s,s'} &= P(s'|s, \mu_\theta(s)), \\ P^{\pi_{\theta,\sigma}}_{s,s'} &= \int_{\mathcal{A}} \pi_{\theta,\sigma}(a|s) P(s'|s, a) \mathrm{d}a = \int_{\mathcal{A}} \nu_\sigma(\mu_\theta(s), a) P(s'|s, a) \mathrm{d}a. \end{aligned}$$

Let $\theta \in \Theta$, $s, s' \in \mathcal{S}$, $(\theta_n) \in \Theta^{\mathbb{N}}$ such that $\theta_n \to \theta$ and $(\sigma_n)_{n \in \mathbb{N}} \in {\mathbb{R}^+}^{\mathbb{N}}$, $\sigma_n \downarrow 0$:
$$\left| P^{\pi_{\theta_n,\sigma_n}}_{s,s'} - P^{\mu_\theta}_{s,s'} \right| \leq \left| P^{\pi_{\theta_n,\sigma_n}}_{s,s'} - P^{\mu_{\theta_n}}_{s,s'} \right| + \left| P^{\mu_{\theta_n}}_{s,s'} - P^{\mu_\theta}_{s,s'} \right|.$$

Applying the first condition of Conditions 1 with $f : a \mapsto P(s'|s, a)$ belonging to $\mathcal{F}$:
$$\begin{aligned} \left| P^{\pi_{\theta_n,\sigma_n}}_{s,s'} - P^{\mu_{\theta_n}}_{s,s'} \right| &= \left| \int_{\mathcal{A}} \nu_{\sigma_n}(\mu_{\theta_n}(s), a) P(s'|s, a) \mathrm{d}a - P(s'|s, \mu_{\theta_n}(s)) \right| \\ &\leq \sup_{f \in \mathcal{F}, a' \in \mathcal{R}} \left| \int_{\mathcal{A}} \nu_{\sigma_n}(a', a) f(a) \mathrm{d}a - f(a') \right| \underset{n \to \infty}{\longrightarrow} 0. \end{aligned}$$

By regularity assumptions on $\theta \mapsto \mu_\theta(s)$ and $P(s'|s, \cdot)$, we have
$$\left| P^{\mu_{\theta_n}}_{s,s'} - P^{\mu_\theta}_{s,s'} \right| = |P(s'|s, \mu_{\theta_n}(s)) - P(s'|s, \mu_\theta(s))| \underset{n \to \infty}{\longrightarrow} 0.$$

Hence,

$$\left| P^{\pi_{\theta_n},\sigma_n}_{s,s'} - P^{\mu_\theta}_{s,s'} \right| \underset{n\to\infty}{\longrightarrow} 0.$$

Therefore, for each $s, s' \in \mathcal{S}$, $(\theta, \sigma) \mapsto P^{\pi_\theta,\sigma}_{s,s'}$, with $P^{\pi_\theta,0}_{s,s'} = P^{\mu_\theta}_{s,s'}$, is continuous on $\Theta \times \{0\}$. Note that, for each $n \in \mathbb{N}$, $P \mapsto \prod_{s,s'} (P^n)_{s,s'}$ is a polynomial function of the entries of P. Thus, for each $n \in \mathbb{N}$, $f_n : (\theta, \sigma) \mapsto \prod_{s,s'} (P^{\pi_\theta,\sigma\, n})_{s,s'}$, with $f_n(\theta, 0) = \prod_{s,s'} (P^{\mu_\theta\, n})_{s,s'}$ is continuous on $\Theta \times \{0\}$. Moreover, for each $\theta \in \Theta, \sigma \geq 0$, from the structure of $P^{\pi_\theta,\sigma}$, if there is some $n^* \in \mathbb{N}$ such that $f_{n^*}(\theta, \sigma) > 0$ then, for all $n \geq n^*$, $f_n(\theta, \sigma) > 0$.

Now let us suppose that there exists $(\theta_n) \in \Theta^{\mathbb{N}^*}$ such that, for each $n > 0$ there is a $\sigma_n \leq n^{-1}$ such that $f_n(\theta_n, \sigma_n) = 0$. By compacity of Θ, we can take (θ_n) converging to some $\theta \in \Theta$. For each $n^* \in \mathbb{N}$, by continuity we have $f_{n^*}(\theta, 0) = \lim_{n\to\infty} f_{n^*}(\theta_n, \sigma_n) = 0$. Since P^{μ_θ} is irreducible and aperiodic, there is some $n \in \mathbb{N}$ such that for all $s, s' \in \mathcal{S}$ and for all $n^* \geq n$, $\left(P^{\mu_\theta\, n^*}\right)_{s,s'} > 0$, i.e. $f_{n^*}(\theta, 0) > 0$. This leads to a contradiction.

Hence, there exists $n^* > 0$ such that for all $\theta \in \Theta$ and $\sigma \leq n^{*-1}$, $f_n(\theta, \sigma) > 0$. We let $r = n^{*-1}$. It follows that, for all $\theta \in \Theta$ and $\sigma \in [0, r]$, $P^{\pi_\theta,\sigma}$ is a transition matrix associated to an irreducible and aperiodic Markov Chain, thus $d^{\pi_\theta,\sigma}$ is well defined as the unique stationary probability distribution associated to $P^{\pi_\theta,\sigma}$. We fix $\theta \in \Theta$ in the remaining of the proof.

Let β a policy for which the Markov Chain corresponding to P^β is irreducible and aperiodic. Let $s_* \in \mathcal{S}$, as asserted in Marbach and Tsitsiklis [2001], considering stationary distribution d^β as a vector $\left(d^\beta_s\right)_{s\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$, d^β is the unique solution of the balance equations:

$$\sum_{s\in\mathcal{S}} d^\beta_s P^\beta_{s,s'} = d^\beta_{s'} \quad s' \in \mathcal{S}\backslash\{s_*\},$$
$$\sum_{s\in\mathcal{S}} d^\beta_s = 1.$$

Hence, we have A^β an $|\mathcal{S}| \times |\mathcal{S}|$ matrix and $a \neq 0$ a constant vector of $\mathbb{R}^{|\mathcal{S}|}$ such that the balance equations is of the form

$$A^\beta d^\beta = a \tag{16}$$

with $A^\beta_{s,s'}$ depending on $P^\beta_{s',s}$ in an affine way, for each $s, s' \in \mathcal{S}$. Moreover, A^β is invertible, thus d^β is given by

$$d^\beta = \frac{1}{\det(A^\beta)} \mathrm{adj}(A^\beta)^\top a.$$

Entries of adj(A^β) and $\det(A^\beta)$ are polynomial functions of the entries of P^β.

Thus, $\sigma \mapsto d^{\pi_\theta,\sigma} = \frac{1}{\det(A^{\pi_\theta,\sigma})} \mathrm{adj}(A^{\pi_\theta,\sigma})^\top a$ is defined on $[0, r]$ and is continuous at 0.

Lemma 1 and integration by parts imply that, for $s, s' \in \mathcal{S}$, $\sigma \in [0, r]$:

$$\begin{aligned}\int_{\mathcal{A}} \nabla_{a'} \nu_\sigma(a', a)|_{a'=\mu_\theta(s)}\, P(s'|s,a)\mathrm{d}a &= -\int_{\mathcal{A}} \nabla_a \nu_\sigma(\mu_\theta(s), a) P(s'|s,a)\mathrm{d}a \\ &= \int_{\mathcal{C}_{\mu_\theta(s)}} \nu_\sigma(\mu_\theta(s), a) \nabla_a P(s'|s,a)\mathrm{d}a + \text{boundary terms} \\ &= \int_{\mathcal{C}_{\mu_\theta(s)}} \nu_\sigma(\mu_\theta(s), a) \nabla_a P(s'|s,a)\mathrm{d}a\end{aligned}$$

where the boundary terms are zero since ν_σ vanishes on the boundary due to Conditions 1.

Thus, for $s, s' \in \mathcal{S}$, $\sigma \in \left[0, r\right]$:

$$\begin{aligned}
\nabla_\theta P^{\pi_{\theta,\sigma}}_{s,s'} &= \nabla_\theta \int_{\mathcal{A}} \pi_{\theta,\sigma}(a|s) P(s'|s,a) \mathrm{d}a \\
&= \int_{\mathcal{A}} \nabla_\theta \pi_{\theta,\sigma}(a|s) P(s'|s,a) \mathrm{d}a \\
&= \int_{\mathcal{A}} \nabla_\theta \mu_\theta(s)\ \nabla_{a'} \nu_\sigma(a', a)\big|_{a'=\mu_\theta(s)}\ P(s'|s,a) \mathrm{d}a \\
&= \nabla_\theta \mu_\theta(s) \int_{\mathcal{C}_{\mu_\theta(s)}} \nu_\sigma(\mu_\theta(s), a) \nabla_a P(s'|s,a) \mathrm{d}a
\end{aligned} \tag{17}$$

where exchange of derivation and integral in (17) follows by application of Leibniz rule with:

- $\forall a \in \mathcal{A}$, $\theta \mapsto \pi_{\theta,\sigma}(a|s) P(s'|s,a)$ is differentiable, and $\nabla_\theta \pi_{\theta,\sigma}(a|s) P(s'|s,a) = \nabla_\theta \mu_\theta(s)\ \nabla_{a'} \nu_\sigma(a', a)\big|_{a'=\mu_\theta(s)}$.

- Let $a^* \in \mathcal{R}$, $\forall \theta \in \Theta$,

$$\begin{aligned}
\left\|\nabla_\theta \pi_{\theta,\sigma}(a|s) P(s'|s,a)\right\| &= \left\|\nabla_\theta \mu_\theta(s)\ \nabla_{a'} \nu_\sigma(a', a)\big|_{a'=\mu_\theta(s)}\right\| \\
&\leq \left\|\nabla_\theta \mu_\theta(s)\right\|_{\mathrm{op}} \left\|\nabla_{a'} \nu_\sigma(a', a)\big|_{a'=\mu_\theta(s)}\right\| \\
&\leq \sup_{\theta\in\Theta} \left\|\nabla_\theta \mu_\theta(s)\right\|_{\mathrm{op}} \left\|\nabla_a \nu_\sigma(\mu_\theta(s), a)\right\| \\
&= \sup_{\theta\in\Theta} \left\|\nabla_\theta \mu_\theta(s)\right\|_{\mathrm{op}} \left\|\nabla_a \nu_\sigma(a^*, a - \mu_\theta(s) + a^*)\right\| \qquad (18) \\
&\leq \sup_{\theta\in\Theta} \left\|\nabla_\theta \mu_\theta(s)\right\|_{\mathrm{op}} \sup_{a\in\mathcal{C}_{a^*}} \left\|\nabla_a \nu_\sigma(a^*, a)\right\| \mathbf{1}_{a\in\mathcal{C}_{a^*}}
\end{aligned}$$

  where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm, and (18) comes from translation invariance (we take $\nabla_a \nu_\sigma(a^*, a) = 0$ for $a \in \mathbb{R}^n \backslash \mathcal{C}_{a^*}$). $a \mapsto \sup_{\theta\in\Theta} \left\|\nabla_\theta \mu_\theta(s)\right\|_{\mathrm{op}} \sup_{a\in\mathcal{C}_{a^*}} \left\|\nabla_a \nu_\sigma(a^*, a)\right\| \mathbf{1}_{a\in\mathcal{C}_{a^*}}$ is measurable, bounded and supported on $\mathcal{C}_{a^*}$, so it is integrable on $\mathcal{A}$.

- Dominated convergence ensures that, for each $k \in [\![1, m]\!]$, partial derivative $g_k(\theta) = \partial_{\theta_k} \int_{\mathcal{A}} \nabla_\theta \pi_{\theta,\sigma}(a|s) P(s'|s,a) \mathrm{d}a$ is continuous: let $\theta_n \downarrow \theta$, then

$$\begin{aligned}
g_k(\theta_n) &= \partial_{\theta_k} \int_{\mathcal{A}} \nabla_\theta \pi_{\theta_n,\sigma}(a|s) P(s'|s,a) \mathrm{d}a \\
&= \partial_{\theta_k} \mu_{\theta_n}(s) \int_{\mathcal{C}_{a^*}} \nu_\sigma(a^*, a - \mu_{\theta_n}(s) + a^*) \nabla_a P(s'|s,a) \mathrm{d}a \\
&\underset{n\to\infty}{\longrightarrow} \partial_{\theta_k} \mu_\theta(s) \int_{\mathcal{C}_{a^*}} \nu_\sigma(a^*, a - \mu_\theta(s) + a^*) \nabla_a P(s'|s,a) \mathrm{d}a = g_k(\theta)
\end{aligned}$$

  with the dominating function $a \mapsto \sup_{a\in\mathcal{C}_{a^*}} |\nu_\sigma(a^*, a)| \sup_{a\in\mathcal{A}} \left\|\nabla_a P(s'|s,a)\right\| \mathbf{1}_{a\in\mathcal{C}_{a^*}}$.

Thus $\sigma \mapsto \nabla_\theta P^{\pi_{\theta,\sigma}}_{s,s'}$ is defined for $\sigma \in \left[0, r\right]$ and is continuous at 0, with $\nabla_\theta P^{\pi_{\theta,0}}_{s,s'} = \nabla_\theta \mu_\theta(s)\ \nabla_a P(s'|s,a)\big|_{a=\mu_\theta(s)}$. Indeed, let $(\sigma_n)_{n\in\mathbb{N}} \in \left[0, r\right]^{+\mathbb{N}}$, $\sigma_n \downarrow 0$, then, applying the first condition of Conditions 1 with $f : a \mapsto \nabla_a P(s'|s,a)$ belonging to $\mathcal{F}$, we get

$$\begin{aligned}
&\left\|\nabla_\theta P^{\pi_{\theta,\sigma_n}}_{s,s'} - \nabla_\theta P^{\mu_\theta}_{s,s'}\right\| \\
&= \left\|\nabla_\theta \mu_\theta(s)\right\|_{\mathrm{op}} \left\|\int_{\mathcal{C}_{\mu_\theta(s)}} \nu_{\sigma_n}(\mu_\theta(s), a) \nabla_a P(s'|s,a) \mathrm{d}a - \nabla_a P(s'|s,a)\big|_{a=\mu_\theta(s)}\right\| \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Since $d^{\pi_{\theta,\sigma}} = \frac{1}{\det(A^{\pi_{\theta,\sigma}})} \mathrm{adj}\left(A^{\pi_{\theta,\sigma}}\right)^\top a$ with $|\det\left(A^{\pi_{\theta,\sigma}}\right)| > 0$ for all $\sigma \in \left[0, r\right]$ and since entries of $\mathrm{adj}\left(A^{\pi_{\theta,\sigma}}\right)$ and $\det\left(A^{\pi_{\theta,\sigma}}\right)$ are polynomial functions of the entries of $P^{\pi_{\theta,\sigma}}$, it follows that

$\sigma \mapsto \nabla_\theta d^{\pi_{\theta,\sigma}}$ is properly defined on $\left[0, r\right]$ and is continuous at 0, which concludes the proof of Lemma 2. $\square$

We now proceed to prove Theorem 8.

Let $\theta \in \Theta$, π_θ as in Theorem 3, and $r > 0$ such that $\sigma \mapsto d^{\pi_{\theta,\sigma}}$, $\sigma \mapsto \nabla_\theta d^{\pi_{\theta,\sigma}}$ are well defined on $\left[0, r\right]$ and are continuous at 0. Then, the following two functions

$$\begin{aligned}
\sigma &\mapsto J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma}) = \sum_{s\in\mathcal{S}} d^{\pi_{\theta,\sigma}}(s) \int_{\mathcal{A}} \pi_{\theta,\sigma}(a|s)\bar{R}(s,a)\mathrm{d}a, \\
\sigma &\mapsto J_{\pi_{\theta,\sigma}}(\mu_\theta) = \sum_{s\in\mathcal{S}} d^{\pi_{\theta,\sigma}}(s)\bar{R}(s,\mu_\theta(s)),
\end{aligned}$$

are properly defined on $\left[0, r\right]$ (with $J_{\pi_{\theta,0}}(\pi_{\theta,0}) = J_{\pi_{\theta,0}}(\mu_\theta) = J_{\mu_\theta}(\mu_\theta)$). Let $s \in \mathcal{S}$, by taking similar arguments as in the proof of Lemma 2, we have

$$\begin{aligned}
\nabla_\theta \int_{\mathcal{A}} \pi_{\theta,\sigma}(a|s)\bar{R}(s,a)\mathrm{d}a &= \int_{\mathcal{A}} \nabla_\theta \pi_{\theta,\sigma}(a,s)\bar{R}(s,a)\mathrm{d}a, \\
&= \nabla_\theta\mu_\theta(s) \int_{\mathcal{C}_{\mu_\theta(s)}} \nu_\sigma(\mu_\theta(s),a)\nabla_a\bar{R}(s,a)\mathrm{d}a.
\end{aligned}$$

Thus, $\sigma \mapsto \nabla_\theta J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma})$ is properly defined on $\left[0, r\right]$ and

$$\begin{aligned}
\nabla_\theta J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma}) &= \sum_{s\in\mathcal{S}} \nabla_\theta d^{\pi_{\theta,\sigma}}(s) \int_{\mathcal{A}} \pi_{\theta,\sigma}(a|s)\bar{R}(s,a)\mathrm{d}a \\
&\quad + \sum_{s\in\mathcal{S}} d^{\pi_{\theta,\sigma}}(s)\nabla_\theta \int_{\mathcal{A}} \pi_{\theta,\sigma}(a|s)\bar{R}(s,a)\mathrm{d}a \\
&= \sum_{s\in\mathcal{S}} \nabla_\theta d^{\pi_{\theta,\sigma}}(s) \int_{\mathcal{A}} \nu_\sigma(\mu_\theta(s),a)\bar{R}(s,a)\mathrm{d}a \\
&\quad + \sum_{s\in\mathcal{S}} d^{\pi_{\theta,\sigma}}(s)\nabla_\theta\mu_\theta(s) \int_{\mathcal{C}_{\mu_\theta(s)}} \nu_\sigma(\mu_\theta(s),a)\nabla_a\bar{R}(s,a)\mathrm{d}a.
\end{aligned}$$

Similarly, $\sigma \mapsto \nabla_\theta J_{\pi_{\theta,\sigma}}(\mu_\theta)$ is properly defined on $\left[0, r\right]$ and

$$\nabla_\theta J_{\pi_{\theta,\sigma}}(\mu_\theta) = \sum_{s\in\mathcal{S}} \nabla_\theta d^{\pi_{\theta,\sigma}}(s)\bar{R}(s,\mu_\theta(s)) + \sum_{s\in\mathcal{S}} d^{\pi_{\theta,\sigma}}(s)\nabla_\theta\mu_\theta(s)\ \nabla_a\bar{R}(s,a)\big|_{a=\mu_\theta(s)}$$

To prove continuity at 0 of both $\sigma \mapsto \nabla_\theta J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma})$ and $\sigma \mapsto \nabla_\theta J_{\pi_{\theta,\sigma}}(\mu_\theta)$ (with $\nabla_\theta J_{\pi_{\theta,0}}(\pi_{\theta,0}) = \nabla_\theta J_{\pi_{\theta,0}}(\mu_\theta) = \nabla_\theta J_{\mu_\theta}(\mu_\theta)$), let $(\sigma_n)_{n\geq 0} \downarrow 0$:

$$\begin{aligned}
&\left\|\nabla_\theta J_{\pi_{\theta,\sigma_n}}(\pi_{\theta,\sigma_n}) - \nabla_\theta J_{\pi_{\theta,0}}(\pi_{\theta,0})\right\| \\
&\leq \left\|\nabla_\theta J_{\pi_{\theta,\sigma_n}}(\pi_{\theta,\sigma_n}) - \nabla_\theta J_{\pi_{\theta,\sigma_n}}(\mu_\theta)\right\| + \left\|\nabla_\theta J_{\pi_{\theta,\sigma_n}}(\mu_\theta) - \nabla_\theta J_{\mu_\theta}(\mu_\theta)\right\|. \qquad (19)
\end{aligned}$$

For the first term of the r.h.s we have

$$\begin{aligned}
&\left\|\nabla_\theta J_{\pi_{\theta,\sigma_n}}(\pi_{\theta,\sigma_n}) - \nabla_\theta J_{\pi_{\theta,\sigma_n}}(\mu_\theta)\right\| \\
&\leq \sum_{s\in\mathcal{S}} \|\nabla_\theta d^{\pi_{\theta,\sigma_n}}(s)\| \left| \int_{\mathcal{A}} \nu_{\sigma_n}(\mu_\theta(s),a)\bar{R}(s,a)\mathrm{d}a - \bar{R}(s,\mu_\theta(s)) \right| \\
&\quad + \sum_{s\in\mathcal{S}} d^{\pi_{\theta,\sigma_n}}(s)\|\nabla_\theta\mu_\theta(s)\|_{\mathrm{op}} \left\| \int_{\mathcal{A}} \nu_{\sigma_n}(\mu_\theta(s),a)\nabla_a\bar{R}(s,a)\mathrm{d}a - \nabla_a\bar{R}(s,a)\big|_{a=\mu_\theta(s)} \right\|.
\end{aligned}$$

Applying the first assumption in Condition 1 with $f : a \mapsto \bar{R}(s,a)$ and $f : a \mapsto \nabla_a \bar{R}(s,a)$ belonging to $\mathcal{F}$ we have, for each $s \in \mathcal{S}$:

$$\left| \int_{\mathcal{A}} \nu_{\sigma_n}(\mu_\theta(s), a)\bar{R}(s,a)\mathrm{d}a - \bar{R}(s,\mu_\theta(s)) \right| \underset{n\to\infty}{\longrightarrow} 0 \quad \text{and}$$

$$\left\| \int_{\mathcal{A}} \nu_{\sigma_n}(\mu_\theta(s), a)\nabla_a \bar{R}(s,a)\mathrm{d}a - \nabla_a \bar{R}(s,a)\big|_{a=\mu_\theta(s)} \right\| \underset{n\to\infty}{\longrightarrow} 0.$$

Moreover, for each $s \in \mathcal{S}$, $d^{\pi_{\theta,\sigma_n}}(s) \underset{n\to\infty}{\longrightarrow} d^{\mu_\theta}(s)$ and $\nabla_\theta d^{\pi_{\theta,\sigma_n}}(s) \underset{n\to\infty}{\longrightarrow} \nabla_\theta d^{\mu_\theta}(s)$ (by Lemma 2), and $\|\nabla_\theta \mu_\theta(s)\|_{\mathrm{op}} < \infty$, so

$$\left\| \nabla_\theta J_{\pi_{\theta,\sigma_n}}(\pi_{\theta,\sigma_n}) - \nabla_\theta J_{\pi_{\theta,\sigma_n}}(\mu_\theta) \right\| \underset{n\to\infty}{\longrightarrow} 0.$$

For the second term of the r.h.s of (19), we have

$$\begin{aligned} \left\| \nabla_\theta J_{\pi_{\theta,\sigma_n}}(\mu_\theta) - \nabla_\theta J_{\mu_\theta}(\mu_\theta) \right\| &\leq \sum_{s\in\mathcal{S}} \left\| \nabla_\theta d^{\pi_{\theta,\sigma_n}}(s) - \nabla_\theta d^{\mu_\theta}(s) \right\| \left| \bar{R}(s,\mu_\theta(s)) \right| \\ &\quad + \sum_{s\in\mathcal{S}} \left| d^{\pi_{\theta,\sigma_n}}(s) - d^{\mu_\theta}(s) \right| \left\| \nabla_\theta \mu_\theta(s) \right\|_{\mathrm{op}} \left\| \nabla_a \bar{R}(s,a)\big|_{a=\mu_\theta(s)} \right\|. \end{aligned}$$

Continuity at 0 of $\sigma \mapsto d^{\pi_{\theta,\sigma}}(s)$ and $\sigma \mapsto \nabla_\theta d^{\pi_{\theta,\sigma}}(s)$ for each $s \in \mathcal{S}$, boundedness of $\bar{R}(s,\cdot)$, $\nabla_a \bar{R}(s,\cdot)$ and $\nabla_\theta(s)\mu_\theta(s)$ implies that

$$\left\| \nabla_\theta J_{\pi_{\theta,\sigma_n}}(\mu_\theta) - \nabla_\theta J_{\mu_\theta}(\mu_\theta) \right\| \underset{n\to\infty}{\longrightarrow} 0.$$

Hence,

$$\left\| \nabla_\theta J_{\pi_{\theta,\sigma_n}}(\pi_{\theta,\sigma_n}) - \nabla_\theta J_{\pi_{\theta,0}}(\pi_{\theta,0}) \right\| \underset{n\to\infty}{\longrightarrow} 0.$$

So, $\sigma \mapsto \nabla_\theta J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma})$ and $\nabla_\theta J_{\pi_{\theta,\sigma}}(\mu_\theta)$ are continuous at 0:

$$\lim_{\sigma\downarrow 0} \nabla_\theta J_{\pi_{\theta,\sigma}}(\pi_{\theta,\sigma}) = \lim_{\sigma\downarrow 0} \nabla_\theta J_{\pi_{\theta,\sigma}}(\mu_\theta) = \nabla_\theta J_{\mu_\theta}(\mu_\theta).$$

Proof of Theorem 4

We will use the two-time-scale stochastic approximation analysis . We let the policy parameter θ_t fixed as $\theta_t \equiv \theta$ when analysing the convergence of the critic step. Thus we can show the convergence of ω_t towards an ω_θ depending on θ, which will then be used to prove the convergence for the slow time-scale.

Lemma 3. *Under Assumptions 3 – 5, the sequence ω_t^i generated from (2) is bounded a.s., i.e., $\sup_t \|\omega_t^i\| < \infty$ a.s., for any $i \in \mathcal{N}$.*

The proof follows the same steps as that of Lemma B.1 in the PMLR version of Zhang et al. [2018].

Lemma 4. *Under Assumption 5, the sequence $\{\hat{J}_t^i\}$ generated as in 2 is bounded a.s, i.e., $\sup_t |\hat{J}_t^i| < \infty$ a.s., for any $i \in \mathcal{N}$.*

The proof follows the same steps as that of Lemma B.2 in the PMLR version of Zhang et al. [2018].

The desired result holds since **Step 1** and **Step 2** of the proof of Theorem 4.6 in Zhang et al. [2018] can both be repeated in the setting of deterministic policies.

Proof of Theorem 5

Let $\mathcal{F}_{t,2} = \sigma(\theta_\tau, s_\tau, \tau \leq t)$ a filtration. In addition, we define

$$\begin{aligned} H(\theta, s, \omega) &= \nabla_\theta \mu_\theta(s) \cdot \nabla_a Q_\omega(s,a)\big|_{a=\mu_\theta(s)}, \\ H(\theta, s) &= H(\theta, s, \omega_\theta), \\ h(\theta) &= \mathbb{E}_{s\sim d^\theta}\left[H(\theta, s)\right]. \end{aligned}$$

Then, for each $\theta \in \Theta$, we can introduce $\nu_\theta : \mathcal{S} \to \mathbb{R}^n$ the solution to the Poisson equation:

$$\left(I - P^\theta\right) \nu_\theta(\cdot) = H(\theta, \cdot) - h(\theta)$$

that is given by $\nu_\theta(s) = \sum_{k \geq 0} \mathbb{E}_{s_{k+1} \sim P^\theta(\cdot|s_k)} \left[H(\theta, s_k) - h(\theta) | s_0 = s\right]$ which is properly defined (similar to the differential value function V).

With projection, actor update (5) becomes

$$\begin{aligned}
\theta_{t+1} &= \Gamma\left[\theta_t + \beta_{\theta,t} H(\theta_t, s_t, \omega_t)\right] \quad (20)\\
&= \Gamma\left[\theta_t + \beta_{\theta,t} h(\theta_t) - \beta_{\theta,t}\left(h(\theta_t) - H(\theta_t, s_t)\right) - \beta_{\theta,t}\left(H(\theta_t, s_t) - H(\theta_t, s_t, \omega_t)\right)\right]\\
&= \Gamma\left[\theta_t + \beta_{\theta,t} h(\theta_t) + \beta_{\theta,t}\left((I - P^{\theta_t})\nu_{\theta_t}(s_t)\right) + \beta_{\theta,t} A_t^1\right]\\
&= \Gamma\left[\theta_t + \beta_{\theta,t} h(\theta_t) + \beta_{\theta,t}\left(\nu_{\theta_t}(s_t) - \nu_{\theta_t}(s_{t+1})\right) + \beta_{\theta,t}\left(\nu_{\theta_t}(s_{t+1}) - P^{\theta_t}\nu_{\theta_t}(s_t)\right) + \beta_{\theta,t} A_t^1\right]\\
&= \Gamma\left[\theta_t + \beta_{\theta,t}\left(h(\theta_t) + A_t^1 + A_t^2 + A_t^3\right)\right]
\end{aligned}$$

where

$$\begin{aligned}
A_t^1 &= H(\theta_t, s_t, \omega_t) - H(\theta_t, s_t),\\
A_t^2 &= \nu_{\theta_t}(s_t) - \nu_{\theta_t}(s_{t+1}),\\
A_t^3 &= \nu_{\theta_t}(s_{t+1}) - P^{\theta_t}\nu_{\theta_t}(s_t).
\end{aligned}$$

For $r < t$ we have

$$\begin{aligned}
\sum_{k=r}^{t-1} \beta_{\theta,k} A_k^2 &= \sum_{k=r}^{t-1} \beta_{\theta,k}\left(\nu_{\theta_k}(s_k) - \nu_{\theta_k}(s_{k+1})\right)\\
&= \sum_{k=r}^{t-1} \beta_{\theta,k}\left(\nu_{\theta_k}(s_k) - \nu_{\theta_{k+1}}(s_{k+1})\right) + \sum_{k=r}^{t-1} \beta_{\theta,k}\left(\nu_{\theta_{k+1}}(s_{k+1}) - \nu_{\theta_k}(s_{k+1})\right)\\
&= \sum_{k=r}^{t-1} \left(\beta_{\theta,k+1} - \beta_{\theta,k}\right)\nu_{\theta_{k+1}}(s_{k+1}) + \beta_{\theta_r}\nu_{\theta_r}(s_r) - \beta_{\theta_t}\nu_{\theta_t}(s_t) + \sum_{k=r}^{t-1} \epsilon_k^{(2)}\\
&= \sum_{k=r}^{t-1} \epsilon_k^{(1)} + \sum_{k=r}^{t-1} \epsilon_k^{(2)} + \eta_{r,t}
\end{aligned}$$

where

$$\begin{aligned}
\epsilon_k^{(1)} &= \left(\beta_{\theta,k+1} - \beta_{\theta,k}\right)\nu_{\theta_{k+1}}(s_{k+1}),\\
\epsilon_k^{(2)} &= \beta_{\theta,k}\left(\nu_{\theta_{k+1}}(s_{k+1}) - \nu_{\theta_k}(s_{k+1})\right),\\
\eta_{r,t} &= \beta_{\theta_r}\nu_{\theta_r}(s_r) - \beta_{\theta_t}\nu_{\theta_t}(s_t).
\end{aligned}$$

Lemma 5. $\sum_{k=0}^{t-1} \beta_{\theta,k} A_k^2$ *converges a.s. for* $t \to \infty$

Proof of Lemma 5. Since $\nu_\theta(s)$ is uniformly bounded for $\theta \in \Theta$, $s \in \mathcal{S}$, we have for some $K > 0$

$$\sum_{k=0}^{t-1} \left\|\epsilon_k^{(1)}\right\| \leq K \sum_{k=0}^{t-1} \left|\beta_{\theta,k+1} - \beta_{\theta,k}\right|$$

which converges given Assumption 5.

Moreover, since $\mu_\theta(s)$ is twice continuously differentiable, $\theta \mapsto \nu_\theta(s)$ is Lipschitz for each s, and so we have

$$\begin{aligned}
\sum_{k=0}^{t-1} \left\|\epsilon_k^{(2)}\right\| &\leq \sum_{k=0}^{t-1} \beta_{\theta,k}\left\|\nu_{\theta_k}(s_{k+1}) - \nu_{\theta_{k+1}}(s_{k+1})\right\|\\
&\leq K^2 \sum_{k=0}^{t-1} \beta_{\theta,k}\left\|\theta_k - \theta_{k+1}\right\|\\
&\leq K^3 \sum_{k=0}^{t-1} \beta_{\theta,k}^2.
\end{aligned}$$

Finally, $\lim\limits_{t\to\infty} \|\eta_{0,t}\| = \beta_{\theta,0} \|\nu_{\theta_0}(s_0)\| < \infty$ a.s.

Thus, $\sum_{k=0}^{t-1} \left\|\beta_{\theta,k} A_k^2\right\| \leq \sum_{k=0}^{t-1} \left\|\epsilon_k^{(1)}\right\| + \sum_{k=0}^{t-1} \left\|\epsilon_k^{(2)}\right\| + \|\eta_{0,t}\|$ converges a.s. □

Lemma 6. *$\sum_{k=0}^{t-1} \beta_{\theta,k} A_k^3$ converges a.s. for $t \to \infty$.*

Proof of Lemma 6. We set

$$Z_t = \sum_{k=0}^{t-1} \beta_{\theta,k} A_k^3 = \sum_{k=0}^{t-1} \beta_{\theta,k} \left(\nu_{\theta_k}(s_{k+1}) - P^{\theta_k} \nu_{\theta_k}(s_k)\right).$$

Since Z_t is $\mathcal{F}_t$-adapted and $\mathbb{E}\left[\nu_{\theta_t}(s_{t+1})|\mathcal{F}_t\right] = P^{\theta_t}\nu_{\theta_t}(s_t)$, Z_t is a martingale. The remaining of the proof is now similar to the proof of Lemma 2 on page 224 of Benveniste et al. [1990]. □

Let $g^i(\theta_t) = \mathbb{E}_{s_t \sim d^{\theta_t}}\left[\psi_t^i \cdot \xi_{t,\theta_t}^i | \mathcal{F}_{t,2}\right]$ and $g(\theta) = \left[g^1(\theta), \ldots, g^N(\theta)\right]$. We have

$$g^i(\theta_t) = \sum_{s_t \in \mathcal{S}} d^{\theta_t}(s_t) \cdot \psi_t^i \cdot \xi_{t,\theta_t}^i.$$

Given (10), $\theta \mapsto \omega_\theta$ is continuously differentiable and $\theta \mapsto \nabla_\theta \omega_\theta$ is bounded so $\theta \mapsto \omega_\theta$ is Lipschitz-continuous. Thus $\theta \mapsto \xi_{t,\theta}^i$ is Lipschitz-continuous for each $s_t \in \mathcal{S}$. Due to our regularity assumptions, $\theta \mapsto \psi_{t,\theta_t}^i$ is also continuous for each $i \in \mathcal{N}, s_t \in \mathcal{S}$. Moreover, $\theta \mapsto d^\theta(s)$ is also Lipschitz continuous for each $s \in \mathcal{S}$. Hence, $\theta \mapsto g(\theta)$ is Lipschitz-continuous in θ and the ODE (12) is well-posed. This holds even when using compatible features.

By critic faster convergence, we have $\lim_{t\to\infty} \|\xi_t^i - \xi_{t,\theta_t}^i\| = 0$ so $\lim_{t\to\infty} A_t^1 = 0$.

Hence, by Kushner-Clark lemma Kushner and Clark [1978] (pp 191-196) we have that the update in (20) converges a.s. to the set of asymptotically stable equilibria of the ODE (12).

Proof of Theorem 6

We use the two-time scale technique: since critic updates at a faster rate than the actor, we let the policy parameter θ_t to be fixed as θ when analysing the convergence of the critic update.

Lemma 7. *Under Assumptions 4, 1 and 6, for any $i \in \mathcal{N}$, sequence $\{\lambda_t^i\}$ generated from (7) is bounded almost surely.*

To prove this lemma we verify the conditions for **Theorem A.2** of Zhang et al. [2018] to hold. We use $\{\mathcal{F}_{t,1}\}$ to denote the filtration with $\mathcal{F}_{t,1} = \sigma(s_\tau, C_{\tau-1}, a_{\tau-1}, r_\tau, \lambda_\tau, \tau \leq t)$. With $\lambda_t = \left[(\lambda_t^1)^\top, \ldots, (\lambda_t^N)^\top\right]^\top$, critic step (7) has the form:

$$\lambda_{t+1} = (C_t \otimes I)\left(\lambda_t + \beta_{\lambda,t} \cdot y_{t+1}\right) \tag{21}$$

with $y_{t+1} = \left(\delta_t^1 w(s_t, a_t)^\top, \ldots, \delta_t^N w(s_t, a_t)^\top\right)^\top \in \mathbb{R}^{KN}$, $\otimes$ denotes Kronecker product and I is the identity matrix. Using the same notation as in **Assumption A.1** from Zhang et al. [2018], we have:

$$h^i(\lambda_t^i, s_t) = \mathbb{E}_{a\sim\pi}\left[\delta_t^i w(s_t, a)^\top | \mathcal{F}_{t,1}\right] = \int_{\mathcal{A}} \pi(a|s_t)(R^i(s_t, a) - w(s_t, a) \cdot \lambda_t^i) w(s_t, a)^\top \mathrm{d}a,$$

$$M_{t+1}^i = \delta_t^i w(s_t, a_t)^\top - \mathbb{E}_{a\sim\pi}\left[\delta_t^i w(s_t, a)^\top | \mathcal{F}_{t,1}\right],$$

$$\bar{h}^i(\lambda_t) = A_{\pi,\theta}^i \cdot d_\pi^s - B_{\pi,\theta} \cdot \lambda_t, \qquad \text{where } A_{\pi,\theta}^i = \left[\int_{\mathcal{A}} \pi(a|s) R^i(s, a) w(s, a)^\top \mathrm{d}a, s \in \mathcal{S}\right].$$

Since feature vectors are uniformly bounded for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, h^i is Lipschitz continuous in its first argument. Since, for $i \in \mathcal{N}$, the r^i are also uniformly bounded, $\mathbb{E}\left[\|M_{t+1}\|^2 | \mathcal{F}_{t,1}\right] \leq K \cdot (1 + \|\lambda_t\|^2)$ for some $K > 0$. Furthermore, finiteness of $|\mathcal{S}|$ ensures that, a.s., $\|\bar{h}(\lambda_t) - h(\lambda_t, s_t)\|^2 \leq K' \cdot (1 + \|\lambda_t\|^2)$. Finally, $h_\infty(y)$ exists and has the form

$$h_\infty(y) = -B_{\pi,\theta} \cdot y.$$

From Assumption 1, we have that $-B_{\pi,\theta}$ is a Hurwitcz matrix, thus the origin is a globally asymptotically stable attractor of the ODE $\dot{y} = h_\infty(y)$. Hence **Theorem A.2** of Zhang et al. [2018] applies, which concludes the proof of Lemma 7.

We introduce the following operators as in Zhang et al. [2018]:

- $\langle \cdot \rangle : \mathbb{R}^{KN} \to \mathbb{R}^{K}$
$$\langle \lambda \rangle = \frac{1}{N}(\mathbf{1}^\top \otimes I)\lambda = \frac{1}{N}\sum_{i \in \mathcal{N}} \lambda^i.$$
- $\mathcal{J} = \left(\frac{1}{N}\mathbf{1}\mathbf{1}^\top \otimes I\right) : \mathbb{R}^{KN} \to \mathbb{R}^{KN}$ such that $\mathcal{J}\lambda = \mathbf{1} \otimes \langle \lambda \rangle$.
- $\mathcal{J}_\perp = I - \mathcal{J} : \mathbb{R}^{KN} \to \mathbb{R}^{KN}$ and we note $\lambda_\perp = \mathcal{J}_\perp \lambda = \lambda - \mathbf{1} \otimes \langle \lambda \rangle$.

We then proceed in two steps as in Zhang et al. [2018], firstly by showing the convergence a.s. of the disagreement vector sequence $\{\lambda_{\perp,t}\}$ to zero, secondly showing that the consensus vector sequence $\{\langle \lambda_t \rangle\}$ converges to the equilibrium such that $\langle \lambda_t \rangle$ is solution to (13).

Lemma 8. *Under Assumptions 4, 1 and 6, for any $M > 0$, we have*
$$\sup_t \mathbb{E}\Big[\|\beta_{\lambda,t}^{-1}\lambda_{\perp,t}\|^2 \mathbb{1}_{\{\sup_t \|\lambda_t\| \le M\}}\Big] < \infty.$$

Since dynamic of $\{\lambda_t\}$ described by (21) is similar to (5.2) in Zhang et al. [2018] we have
$$\mathbb{E}\Big[\|\beta_{\lambda,t+1}^{-1}\lambda_{\perp,t+1}\|^2 | \mathcal{F}_{t,1}\Big] = \frac{\beta_{\lambda,t}^2}{\beta_{\lambda,t+1}^2}\rho\left(\|\beta_{\lambda,t}^{-1}\lambda_{\perp,t}\|^2 + 2\cdot\|\beta_{\lambda,t}^{-1}\lambda_{\perp,t}\|\cdot\mathbb{E}(\|y_{t+1}\|^2|\mathcal{F}_{t,1})^{\frac{1}{2}} + \mathbb{E}(\|y_{t+1}\|^2|\mathcal{F}_{t,1})\right) \tag{22}$$
where ρ represents the spectral norm of $\mathbb{E}\big[C_t^\top \cdot (I - \mathbf{1}\mathbf{1}^\top/N) \cdot C_t\big]$, with $\rho \in [0, 1)$ by Assumption 4. Since $y_{t+1}^i = \delta_t^i \cdot w(s_t, a_t)^\top$ we have
$$\begin{aligned}\mathbb{E}\Big[\|y_{t+1}\|^2 | \mathcal{F}_{t,1}\Big] &= \mathbb{E}\Big[\sum_{i \in \mathcal{N}} \|(r^i(s_t, a_t) - w(s_t, a_t)\lambda_t^i) \cdot w(s_t, a_t)^\top\|^2 | \mathcal{F}_{t,1}\Big] \\ &\le 2 \cdot \mathbb{E}\Big[\sum_{i \in \mathcal{N}} \|r^i(s_t, a_t) w(s_t, a_t)^\top\|^2 + \|w(s_t, a_t)^\top\|^4 \cdot \|\lambda_t^i\|^2 | \mathcal{F}_{t,1}\Big].\end{aligned}$$

By uniform boundedness of $r(s, \cdot)$ and $w(s, \cdot)$ (Assumptions 1) and finiteness of $\mathcal{S}$, there exists $K_1 > 0$ such that
$$\mathbb{E}\Big[\|y_{t+1}\|^2 | \mathcal{F}_{t,1}\Big] \le K_1(1 + \|\lambda_t\|^2).$$
Thus, for any $M > 0$ there exists $K_2 > 0$ such that, on the set $\{\sup_{\tau \le t}\|\lambda_\tau\| < M\}$,
$$\mathbb{E}\Big[\|y_{t+1}\|^2 \mathbb{1}_{\{\sup_{\tau \le t}\|\lambda_\tau\| < M\}} | \mathcal{F}_{t,1}\Big] \le K_2. \tag{23}$$

We let $v_t = \|\beta_{\lambda,t}^{-1}\lambda_{\perp,t}\|^2 \mathbb{1}_{\{\sup_{\tau \le t}\|\lambda_\tau\| < M\}}$. Taking expectation over (22), noting that $\mathbb{1}_{\{\sup_{\tau \le t+1}\|\lambda_\tau\| < M\}} \le \mathbb{1}_{\{\sup_{\tau \le t}\|\lambda_\tau\| < M\}}$ we get
$$\mathbb{E}(v_{t+1}) \le \frac{\beta_{\lambda,t}^2}{\beta_{\lambda,t+1}^2}\rho\left(\mathbb{E}(v_t) + 2\sqrt{\mathbb{E}(v_t)} \cdot \sqrt{K_2} + K_2\right)$$
which is the same expression as (5.10) in Zhang et al. [2018]. So similar conclusions to the ones of **Step 1** of Zhang et al. [2018] holds:
$$\sup_t \mathbb{E}\Big[\|\beta_{\lambda,t}^{-1}\lambda_{\perp,t}\|^2 \mathbb{1}_{\{\sup_t \|\lambda_t\| \le M\}}\Big] < \infty \tag{24}$$
$$\text{and} \qquad \lim_t \lambda_{\perp,t} = 0 \text{ a.s.} \tag{25}$$

We now show convergence of the consensus vector $\mathbf{1} \otimes \langle \lambda_t \rangle$. Based on (21) we have
$$\begin{aligned}\langle \lambda_{t+1} \rangle &= \langle (C_t \otimes I)(\mathbf{1} \otimes \langle \lambda_t \rangle + \lambda_{\perp,t} + \beta_{\lambda,t} y_{t+1}) \rangle \\ &= \langle \lambda_t \rangle + \langle \lambda_{\perp,t} \rangle + \beta_{\lambda,t}\langle (C_t \otimes I)(y_{t+1} + \beta_{\lambda,t}^{-1}\lambda_{\perp,t}) \rangle \\ &= \langle \lambda_t \rangle + \beta_{\lambda,t}(h(\lambda_t, s_t) + M_{t+1})\end{aligned}$$

where $h(\lambda_t, s_t) = \mathbb{E}_{a_t\sim\pi}\big[\langle y_{t+1}\rangle|\mathcal{F}_t\big]$ and $M_{t+1} = \langle (C_t\otimes I)(y_{t+1}+\beta_{\lambda,t}^{-1}\lambda_{\perp,t})\rangle - \mathbb{E}_{a_t\sim\pi}\big[\langle y_{t+1}\rangle|\mathcal{F}_t\big]$.
Since $\langle\delta_t\rangle = \bar{r}(s_t, a_t) - w(s_t, a_t)\langle\lambda_t\rangle$, we have

$$h(\lambda_t, s_t) = \mathbb{E}_{a_t\sim\pi}(\bar{r}(s_t, a_t)w(s_t, a_t)^\top|\mathcal{F}_t) + \mathbb{E}_{a_t\sim\pi}(w(s_t, a_t)\langle\lambda_t\rangle\cdot w(s_t, a_t)^\top|\mathcal{F}_{t,1})$$

so h is Lipschitz-continuous in its first argument. Moreover, since $\langle\lambda_{\perp,t}\rangle = 0$ and $\mathbf{1}^\top\mathbb{E}(C_t|\mathcal{F}_{t,1}) = \mathbf{1}^\top$ a.s.:

$$\begin{aligned}
\mathbb{E}_{a_t\sim\pi}\big[\langle (C_t\otimes I)(y_{t+1}+\beta_{\lambda,t}^{-1}\lambda_{\perp,t})\rangle|\mathcal{F}_{t,1}\big] &= \mathbb{E}_{a_t\sim\pi}\Big[\frac{1}{N}(\mathbf{1}^\top\otimes I)(C_t\otimes I)(y_{t+1}+\beta_{\lambda,t}^{-1}\lambda_{\perp,t})|\mathcal{F}_{t,1}\Big]\\
&= \frac{1}{N}(\mathbf{1}^\top\otimes I)(\mathbb{E}(C_t|\mathcal{F}_{t,1})\otimes I)\mathbb{E}_{a_t\sim\pi}\big[y_{t+1}+\beta_{\lambda,t}^{-1}\lambda_{\perp,t}|\mathcal{F}_{t,1}\big]\\
&= \frac{1}{N}(\mathbf{1}^\top\mathbb{E}(C_t|\mathcal{F}_{t,1})\otimes I)\mathbb{E}_{a_t\sim\pi}\big[y_{t+1}+\beta_{\lambda,t}^{-1}\lambda_{\perp,t}|\mathcal{F}_{t,1}\big]\\
&= \mathbb{E}_{a_t\sim\pi}\big[\langle y_{t+1}\rangle|\mathcal{F}_{t,1}\big] \text{ a.s.}
\end{aligned}$$

So $\{M_t\}$ is a martingale difference sequence. Additionally we have

$$\mathbb{E}\big[\|M_{t+1}\|^2|\mathcal{F}_{t,1}\big] \le 2\cdot\mathbb{E}\big[\|y_{t+1}+\beta_{\lambda,t}^{-1}\lambda_{\perp,t}\|_{G_t}^2|\mathcal{F}_{t,1}\big] + 2\cdot\|\mathbb{E}\big[\langle y_{t+1}\rangle|\mathcal{F}_{t,1}\big]\|^2$$

with $G_t = N^{-2}\cdot C_t^\top\mathbf{1}\mathbf{1}^\top C_t\otimes I$ whose spectral norm is bounded for C_t is stochastic. From (23) and (24) we have that, for any $M > 0$, over the set $\{\sup_t\|\lambda_t\|\le M\}$, there exists $K_3, K_4 < \infty$ such that

$$\mathbb{E}\big[\|y_{t+1}+\beta_{\lambda,t}^{-1}\lambda_{\perp,t}\|_{G_t}^2|\mathcal{F}_{t,1}\big]\mathbb{1}_{\{\sup_t\|\lambda_t\|\le M\}} \le K_3\cdot\mathbb{E}\big[\|y_{t+1}\|^2+\|\beta_{\lambda,t}^{-1}\lambda_{\perp,t}\|^2|\mathcal{F}_{t,1}\big]\mathbb{1}_{\{\sup_t\|\lambda_t\|\le M\}} \le K_4.$$

Besides, since r_{t+1}^i and w are uniformly bounded, there exists $K_5 < \infty$ such that $\|\mathbb{E}\big[\langle y_{t+1}\rangle|\mathcal{F}_{t,1}\big]\|^2 \le K_5\cdot(1+\|\langle\lambda_t\rangle\|^2)$. Thus, for any $M > 0$, there exists some $K_6 < \infty$ such that over the set $\{\sup_t\|\lambda_t\|\le M\}$

$$\mathbb{E}\big[\|M_{t+1}\|^2|\mathcal{F}_{t,1}\big] \le K_6\cdot(1+\|\langle\lambda_t\rangle\|^2).$$

Hence, for any $M > 0$, assumptions (a.1) - (a.5) of B.1. from Zhang et al. [2018] are verified on the set $\{\sup_t\|\lambda_t\|\le M\}$. Finally, we consider the ODE asymptotically followed by $\langle\lambda_t\rangle$:

$$\dot{\langle\lambda_t\rangle} = -B_{\pi,\theta}\cdot\langle\lambda_t\rangle + A_{\pi,\theta}\cdot d^\pi$$

which has a single globally asymptotically stable equilibrium $\lambda^*\in\mathbb{R}^K$, since $B_{\pi,\theta}$ is positive definite: $\lambda^* = B_{\pi,\theta}^{-1}\cdot A_{\pi,\theta}\cdot d^\pi$. By Lemma 7, $\sup_t\|\langle\lambda_t\rangle\| < \infty$ a.s., all conditions to apply **Theorem B.2.** of Zhang et al. [2018] hold a.s., which means that $\langle\lambda_t\rangle \underset{t\to\infty}{\longrightarrow} \lambda^*$ a.s. As $\lambda_t = \mathbf{1}\otimes\langle\lambda_t\rangle + \lambda_{\perp,t}$ and $\lambda_{\perp,t} \underset{t\to\infty}{\longrightarrow} 0$ a.s., we have for each $i\in\mathcal{N}$, a.s.,

$$\lambda_t^i \underset{t\to\infty}{\longrightarrow} B_{\pi,\theta}^{-1}\cdot A_{\pi,\theta}\cdot d^\pi.$$

Proof of Theorem 7

Let $\mathcal{F}_{t,2} = \sigma(\theta_\tau, \tau\le t)$ be the σ-field generated by $\{\theta_\tau, \tau\le t\}$, and let

$$\zeta_{t,1}^i = \psi_t^i\cdot\xi_t^i - \mathbb{E}_{s_t\sim d^\pi}\left[\psi_t^i\cdot\xi_t^i|\mathcal{F}_{t,2}\right], \qquad \zeta_{t,2}^i = \mathbb{E}_{s_t\sim d^\pi}\left[\psi_t^i\cdot(\xi_t^i - \xi_{t,\theta_t}^i)|\mathcal{F}_{t,2}\right].$$

With local projection, actor update (6) becomes

$$\theta_{t+1}^i = \Gamma^i\left[\theta_t^i + \beta_{\theta,t}\mathbb{E}_{s_t\sim d^\pi}\left[\psi_t^i\cdot\xi_{t,\theta_t}^i|\mathcal{F}_{t,2}\right] + \beta_{\theta,t}\zeta_{t,1}^i + \beta_{\theta,t}\zeta_{t,2}^i\right]. \tag{26}$$

So with $h^i(\theta_t) = \mathbb{E}_{s_t\sim d^\pi}\left[\psi_t^i\cdot\xi_{t,\theta_t}^i|\mathcal{F}_{t,2}\right]$ and $h(\theta) = \left[h^1(\theta),\ldots,h^N(\theta)\right]$, we have

$$h^i(\theta_t) = \sum_{s_t\in\mathcal{S}} d^\pi(s_t)\cdot\psi_t^i\cdot\xi_{t,\theta_t}^i.$$

Given (10), $\theta\mapsto\omega_\theta$ is continuously differentiable and $\theta\mapsto\nabla_\theta\omega_\theta$ is bounded so $\theta\mapsto\omega_\theta$ is Lipschitz-continuous. Thus $\theta\mapsto\xi_{t,\theta}^i$ is Lipschitz-continuous for each $s_t\in\mathcal{S}$. Our regularity assumptions ensure that $\theta \mapsto \psi^i_{t,\theta_t}$ is continuous for each $i \in \mathcal{N}, s_t \in \mathcal{S}$. Moreover, $\theta \mapsto d^\theta(s)$ is also Lipschitz continuous for each $s \in \mathcal{S}$. Hence, $\theta \mapsto g(\theta)$ is Lipschitz-continuous in θ and the ODE (12) is well-posed. This holds even when using compatible features.

By critic faster convergence, we have $\lim_{t\to\infty}\|\xi^i_t - \xi^i_{t,\theta_t}\|= 0$.

Let $M^i_t = \sum_{\tau=0}^{t-1}\beta_{\theta,\tau}\zeta^i_{\tau,1}$. M^i_t is a martingale sequence with respect to $\mathcal{F}_{t,2}$. Since $\{\omega_t\}_t, \{\nabla_a\phi_k(s,a)\}_{s,k}$, and $\{\nabla_\theta\mu_\theta(s)\}_s$ are bounded (Lemma 3, Assumption 2), it follows that the sequence $\left\{\zeta^i_{t,1}\right\}$ is bounded. Thus, by Assumption 5, $\sum_t \mathbb{E}\left[\left\|M^i_{t+1} - M^i_t\right\|^2 |\mathcal{F}_{t,2}\right] = \sum_t \left\|\beta_{\theta,t}\zeta^i_{t,1}\right\|^2 < \infty$ a.s. The martingale convergence theorem ensures that $\left\{M^i_t\right\}$ converges a.s. Thus, for any $\epsilon > 0$,

$$\lim_t \mathbb{P}\left(\sup_{n\geq t}\left\|\sum_{\tau=t}^{n}\beta_{\theta,\tau}\zeta^i_{\tau,1}\right\| \geq \epsilon\right) = 0.$$

Hence, by Kushner-Clark lemma Kushner and Clark [1978] (pp 191-196) we have that the update in (26) converges a.s. to the set of asymptotically stable equilibria of the ODE (12).