

## APPENDIX

## 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  $R^i(a) = -(\sum_i a^i - a^*)^\top C (\sum_i a^i - a^*)$ . The matrix  $C$  is positive definite with eigenvalues chosen from  $\{0.1, 1\}$ , and  $a^* = [4, \dots, 4]^\top$ . 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.

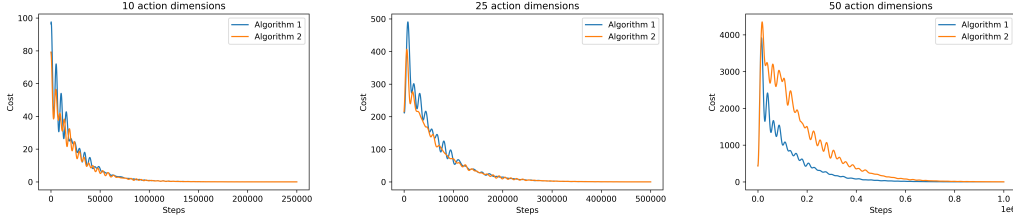


Figure 2: Convergence of Algorithms 1 and 2 on the multi-agent continuous bandit problem.

## ASSUMPTIONS

**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  $w_\theta(s, a) = [w_{\theta,1}(s, a), \dots, w_{\theta,K}(s, a)] \in \mathbb{R}^K$  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  $s \in \mathcal{S}$ ,  $a \in \mathcal{A}$ ,  $k \in \llbracket 1, K \rrbracket$ . Furthermore, we assume that the feature matrix  $W_\pi \in \mathbb{R}^{|\mathcal{S}| \times K}$  has full column rank, where the  $k$ -th column of  $W_\pi$  is  $[\int_{\mathcal{A}} \pi(a|s) w_{\theta,k}(s, a) da, s \in \mathcal{S}]$  for any  $k \in \llbracket 1, K \rrbracket$ .

**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  $\phi(s, a) = [\phi_1(s, a), \dots, \phi_K(s, a)] \in \mathbb{R}^K$  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  $s \in \mathcal{S}$ ,  $a \in \mathcal{A}$ ,  $k \in \{1, \dots, K\}$ . 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  $\Phi_\theta$  is  $[\phi_k(s, \mu_\theta(s)), s \in \mathcal{S}]$  for any  $k \in \llbracket 1, K \rrbracket$ . Also, for any  $u \in \mathbb{R}^K$ ,  $\Phi_\theta u \neq \mathbf{1}$ .

**Assumption 3** (Bounding  $\theta$ ). The update of the policy parameter  $\theta^i$  includes a local projection by  $\Gamma^i : \mathbb{R}^{m_i} \rightarrow \Theta^i$  that projects any  $\theta_t^i$  onto a compact set  $\Theta^i$  that can be expressed as  $\{\theta^i | q_j^i(\theta^i) \leq 0, j = 1, \dots, s^i\} \subset \mathbb{R}^{m_i}$ , 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  $\mathbb{E}[C_t^\top \cdot (I - \mathbf{1}\mathbf{1}^\top/N) \cdot C_t]$  is smaller than one.
4. Given the  $\sigma$ -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:

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

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

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

$$\begin{aligned} \sum_t \beta_{\lambda,t} &= \sum_t \beta_{\theta,t} = \infty, & \sum_t \beta_{\lambda,t}^2 + \beta_{\theta,t}^2 &< \infty \\ \beta_{\theta,t} &= o(\beta_{\lambda,t}), & \lim_{t \rightarrow \infty} \beta_{\lambda,t+1}/\beta_{\lambda,t} &= 1. \end{aligned}$$

## PROOF OF THEOREM 1

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

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

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

The Poisson equation under policy  $\mu_\theta$  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,

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

Hence,

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

Using stationarity property of  $d^\theta$ , 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

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

Given that  $\nabla_{\theta^i} \mu_\theta^j(s) = 0$  if  $i \neq j$ , we have  $\nabla_\theta \mu_\theta(s) = \text{Diag}(\nabla_{\theta^1} \mu_{\theta^1}^1(s), \dots, \nabla_{\theta^N} \mu_{\theta^N}^N(s))$ , which implies

$$\nabla_{\theta^i} J(\theta) = \mathbb{E}_{s \sim d^\theta} [\nabla_{\theta^i} \mu_{\theta^i}^i(s) \nabla_{a^i} Q_\theta(s, \mu_{\theta^{-i}}^{-i}(s), a^i)|_{a^i=\mu_{\theta^i}^i(s)}]. \quad (15)$$

### PROOF OF THEOREM 3

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

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

Recall that for a deterministic policy  $\mu : \mathcal{S} \rightarrow \mathcal{A}$ , 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  $\nu_\sigma$  parametrized by  $\sigma$  are said to be regular delta-approximation on  $\mathcal{R} \subset \mathcal{A}$  if they satisfy the following conditions:

1. The distributions  $\nu_\sigma$  converge to a delta distribution:  $\lim_{\sigma \downarrow 0} \int_{\mathcal{A}} \nu_\sigma(a', a) f(a) da = f(a')$  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.:

$$\lim_{\sigma \downarrow 0} \sup_{f \in \mathcal{F}, a' \in \mathcal{R}} \left| \int_{\mathcal{A}} \nu_\sigma(a', a) f(a) da - f(a') \right| = 0.$$

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  $\text{bd}(\mathcal{C}_{a'})$ , 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  $\nu_\sigma$  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  $\mu_\theta : \mathcal{S} \rightarrow \mathcal{A}$ . Denote the range of  $\mu_\theta$  by  $\mathcal{R}_\theta \subseteq \mathcal{A}$ , and  $\mathcal{R} = \cup_\theta \mathcal{R}_\theta$ . For each  $\theta$ , consider  $\pi_{\theta, \sigma}$  a stochastic policy such that  $\pi_{\theta, \sigma}(a|s) = \nu_\sigma(\mu_\theta(s), a)$ , where  $\nu_\sigma$  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  $\beta$ , we let  $(P_{s, s'}^\beta)_{s, s' \in \mathcal{S}}$  be the transition matrix associated to the Markov Chain  $\{s_t\}_{t \geq 0}$  induced by  $\beta$ . In particular, for each  $\theta \in \Theta$ ,  $\sigma > 0$ ,  $s, s' \in \mathcal{S}$ , we have

$$\begin{aligned} P_{s, s'}^{\mu_\theta} &= P(s'|s, \mu_\theta(s)), \\ P_{s, s'}^{\pi_{\theta, \sigma}} &= \int_{\mathcal{A}} \pi_{\theta, \sigma}(a|s) P(s'|s, a) da = \int_{\mathcal{A}} \nu_\sigma(\mu_\theta(s), a) P(s'|s, a) da. \end{aligned}$$

Let  $\theta \in \Theta$ ,  $s, s' \in \mathcal{S}$ ,  $(\theta_n) \in \Theta^\mathbb{N}$  such that  $\theta_n \rightarrow \theta$  and  $(\sigma_n)_{n \in \mathbb{N}} \in \mathbb{R}^{+\mathbb{N}}$ ,  $\sigma_n \downarrow 0$ :

$$\left| P_{s, s'}^{\pi_{\theta_n, \sigma_n}} - P_{s, s'}^{\mu_\theta} \right| \leq \left| P_{s, s'}^{\pi_{\theta_n, \sigma_n}} - P_{s, s'}^{\mu_{\theta_n}} \right| + \left| P_{s, s'}^{\mu_{\theta_n}} - P_{s, s'}^{\mu_\theta} \right|.$$

Applying the first condition of Conditions 1 with  $f : a \mapsto P(s'|s, a)$  belonging to  $\mathcal{F}$ :

$$\begin{aligned} \left| P_{s, s'}^{\pi_{\theta_n, \sigma_n}} - P_{s, s'}^{\mu_{\theta_n}} \right| &= \left| \int_{\mathcal{A}} \nu_{\sigma_n}(\mu_{\theta_n}(s), a) P(s'|s, a) da - P(s'|s, \mu_{\theta_n}(s)) \right| \\ &\leq \sup_{f \in \mathcal{F}, a' \in \mathcal{R}} \left| \int_{\mathcal{A}} \nu_{\sigma_n}(a', a) f(a) da - f(a') \right| \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

By regularity assumptions on  $\theta \mapsto \mu_\theta(s)$  and  $P(s'|s, \cdot)$ , we have

$$\left| P_{s,s'}^{\mu_{\theta_n}} - P_{s,s'}^{\mu_\theta} \right| = |P(s'|s, \mu_{\theta_n}(s)) - P(s'|s, \mu_\theta(s))| \xrightarrow{n \rightarrow \infty} 0.$$

Hence,

$$\left| P_{s,s'}^{\pi_{\theta_n, \sigma_n}} - P_{s,s'}^{\mu_\theta} \right| \xrightarrow{n \rightarrow \infty} 0.$$

Therefore, for each  $s, s' \in \mathcal{S}$ ,  $(\theta, \sigma) \mapsto P_{s,s'}^{\pi_{\theta, \sigma}}$ , with  $P_{s,s'}^{\pi_{\theta, 0}} = P_{s,s'}^{\mu_\theta}$ , 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 compactness of  $\Theta$ , we can take  $(\theta_n)$  converging to some  $\theta \in \Theta$ . For each  $n^* \in \mathbb{N}$ , by continuity we have  $f_{n^*}(\theta, 0) = \lim_{n \rightarrow \infty} f_{n^*}(\theta_n, \sigma_n) = 0$ . Since  $P^{\mu_\theta}$  is irreducible and aperiodic, there is some  $n \in \mathbb{N}$  such that for all  $s, s' \in \mathcal{S}$  and for all  $n^* \geq n$ ,  $(P^{\mu_\theta})^{n^*}_{s,s'} > 0$ , 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  $\beta$  a policy for which the Markov Chain corresponding to  $P^\beta$  is irreducible and aperiodic. Let  $s_* \in \mathcal{S}$ , as asserted in Marbach & Tsitsiklis (2001), considering stationary distribution  $d^\beta$  as a vector  $(d_s^\beta)_{s \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ ,  $d^\beta$  is the unique solution of the balance equations:

$$\begin{aligned} \sum_{s \in \mathcal{S}} d_s^\beta P_{s,s'}^\beta &= d_{s'}^\beta \quad s' \in \mathcal{S} \setminus \{s_*\}, \\ \sum_{s \in \mathcal{S}} d_s^\beta &= 1. \end{aligned}$$

Hence, we have  $A^\beta$  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  $A_{s,s'}^\beta$  depending on  $P_{s',s}^\beta$  in an affine way, for each  $s, s' \in \mathcal{S}$ . Moreover,  $A^\beta$  is invertible, thus  $d^\beta$  is given by

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

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

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

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

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

where the boundary terms are zero since  $\nu_\sigma$  vanishes on the boundary due to Conditions 1.

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

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

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  $\nabla_\theta \pi_{\theta,\sigma}(a|s) P(s'|s, a) = \nabla_\theta \mu_\theta(s) \nabla_{a'} \nu_\sigma(a', a)|_{a'=\mu_\theta(s)}$ .
- Let  $a^* \in \mathcal{R}$ ,  $\forall \theta \in \Theta$ ,

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

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

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

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

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

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

$$\begin{aligned}
&\|\nabla_\theta P_{s,s'}^{\pi_{\theta,\sigma_n}} - \nabla_\theta P_{s,s'}^{\mu_\theta}\| \\
&= \|\nabla_\theta \mu_\theta(s)\|_{\text{op}} \left\| \int_{\mathcal{C}_{\mu_\theta(s)}} \nu_{\sigma_n}(\mu_\theta(s), a) \nabla_a P(s'|s, a) da - \nabla_a P(s'|s, a)|_{a=\mu_\theta(s)} \right\| \xrightarrow{n \rightarrow \infty} 0.
\end{aligned}$$

Since  $d^{\pi_{\theta}, \sigma} = \frac{1}{\det(A^{\pi_{\theta}, \sigma})} \text{adj}(A^{\pi_{\theta}, \sigma})^\top a$  with  $|\det(A^{\pi_{\theta}, \sigma})| > 0$  for all  $\sigma \in [0, r]$  and since entries of  $\text{adj}(A^{\pi_{\theta}, \sigma})$  and  $\det(A^{\pi_{\theta}, \sigma})$  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  $[0, r]$  and is continuous at 0, which concludes the proof of Lemma 2.  $\square$

We now proceed to prove Theorem 8.

Let  $\theta \in \Theta$ ,  $\pi_{\theta}$  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  $[0, r]$  and are continuous at 0. Then, the following two functions

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

are properly defined on  $[0, r]$  (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

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

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

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

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

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

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  $(\sigma_n)_{n \geq 0} \downarrow 0$ :

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

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

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

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  $s \in \mathcal{S}$ :

$$\left| \int_{\mathcal{A}} \nu_{\sigma_n}(\mu_\theta(s), a) \bar{R}(s, a) da - \bar{R}(s, \mu_\theta(s)) \right| \xrightarrow{n \rightarrow \infty} 0 \quad \text{and}$$

$$\left\| \int_{\mathcal{A}} \nu_{\sigma_n}(\mu_\theta(s), a) \nabla_a \bar{R}(s, a) da - \nabla_a \bar{R}(s, a)|_{a=\mu_\theta(s)} \right\| \xrightarrow{n \rightarrow \infty} 0.$$

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

$$\|\nabla_\theta J_{\pi_\theta, \sigma_n}(\pi_\theta, \sigma_n) - \nabla_\theta J_{\pi_\theta, \sigma_n}(\mu_\theta)\| \xrightarrow{n \rightarrow \infty} 0.$$

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

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

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

$$\|\nabla_\theta J_{\pi_\theta, \sigma_n}(\mu_\theta) - \nabla_\theta J_{\mu_\theta}(\mu_\theta)\| \xrightarrow{n \rightarrow \infty} 0.$$

Hence,

$$\|\nabla_\theta J_{\pi_\theta, \sigma_n}(\pi_\theta, \sigma_n) - \nabla_\theta J_{\pi_\theta, 0}(\pi_\theta, 0)\| \xrightarrow{n \rightarrow \infty} 0.$$

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  $\theta_t$  fixed as  $\theta_t \equiv \theta$  when analysing the convergence of the critic step. Thus we can show the convergence of  $\omega_t$  towards an  $\omega_\theta$  depending on  $\theta$ , which will then be used to prove the convergence for the slow time-scale.

**Lemma 3.** *Under Assumptions 3 – 5, the sequence  $\omega_t^i$  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

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



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

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

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

With projection, actor update (5) becomes

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

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

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

where

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

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

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

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

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

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

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

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

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

*Proof of Lemma 6.* We set

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

Since  $Z_t$  is  $\mathcal{F}_t$ -adapted and  $\mathbb{E}[\nu_{\theta_t}(s_{t+1})|\mathcal{F}_t] = P^{\theta_t} \nu_{\theta_t}(s_t)$ ,  $Z_t$  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).  $\square$

Let  $g^i(\theta_t) = \mathbb{E}_{s_t \sim d^{\theta_t}} [\psi_t^i \cdot \xi_{t,\theta_t}^i | \mathcal{F}_{t,2}]$  and  $g(\theta) = [g^1(\theta), \dots, g^N(\theta)]$ . 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  $\theta$  and the ODE (12) is well-posed. This holds even when using compatible features.

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

Hence, by Kushner-Clark lemma Kushner & 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  $\theta_t$  to be fixed as  $\theta$  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  $\lambda_t = [(\lambda_t^1)^\top, \dots, (\lambda_t^N)^\top]^\top$ , critic step (7) has the form:

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

with  $y_{t+1} = (\delta_t^1 w(s_t, a_t)^\top, \dots, \delta_t^N w(s_t, a_t)^\top)^\top \in \mathbb{R}^{KN}$ ,  $\otimes$  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:

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

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

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

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,  $\mathbb{E}[\|M_{t+1}^i\|^2 | \mathcal{F}_{t,1}] \leq K \cdot (1 + \|\lambda_t\|^2)$  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 Hurwitz 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):

- $\langle \cdot \rangle : \mathbb{R}^{KN} \rightarrow \mathbb{R}^K$ 

$$\langle \lambda \rangle = \frac{1}{N} (\mathbf{1}^\top \otimes I) \lambda = \frac{1}{N} \sum_{i \in \mathcal{N}} \lambda^i.$$
- $\mathcal{J} = \left( \frac{1}{N} \mathbf{1} \mathbf{1}^\top \otimes I \right) : \mathbb{R}^{KN} \rightarrow \mathbb{R}^{KN}$  such that  $\mathcal{J} \lambda = \mathbf{1} \otimes \langle \lambda \rangle$ .
- $\mathcal{J}_\perp = I - \mathcal{J} : \mathbb{R}^{KN} \rightarrow \mathbb{R}^{KN}$  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*

$$\sup_t \mathbb{E} \left[ \|\beta_{\lambda,t}^{-1} \lambda_{\perp,t}\|^2 \mathbb{1}_{\{\sup_{\tau \leq t} \|\lambda_\tau\| \leq M\}} \right] < \infty.$$

Since dynamic of  $\{\lambda_t\}$  described by (21) is similar to (5.2) in Zhang et al. (2018) we have

$$\mathbb{E} \left[ \|\beta_{\lambda,t+1}^{-1} \lambda_{\perp,t+1}\|^2 | \mathcal{F}_{t,1} \right] = \frac{\beta_{\lambda,t}^2}{\beta_{\lambda,t+1}^2} \rho \left( \|\beta_{\lambda,t}^{-1} \lambda_{\perp,t}\|^2 + 2 \cdot \|\beta_{\lambda,t}^{-1} \lambda_{\perp,t}\| \cdot \mathbb{E}(\|y_{t+1}\|^2 | \mathcal{F}_{t,1})^{\frac{1}{2}} + \mathbb{E}(\|y_{t+1}\|^2 | \mathcal{F}_{t,1}) \right) \quad (22)$$

where  $\rho$  represents the spectral norm of  $\mathbb{E}[C_t^\top \cdot (I - \mathbf{1} \mathbf{1}^\top / N) \cdot C_t]$ , 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

$$\begin{aligned} \mathbb{E} \left[ \|y_{t+1}\|^2 | \mathcal{F}_{t,1} \right] &= \mathbb{E} \left[ \sum_{i \in \mathcal{N}} \|(r^i(s_t, a_t) - w(s_t, a_t) \lambda_t^i) \cdot w(s_t, a_t)^\top\|^2 | \mathcal{F}_{t,1} \right] \\ &\leq 2 \cdot \mathbb{E} \left[ \sum_{i \in \mathcal{N}} \|r^i(s_t, a_t) w(s_t, a_t)^\top\|^2 + \|w(s_t, a_t)^\top\|^4 \cdot \|\lambda_t^i\|^2 | \mathcal{F}_{t,1} \right]. \end{aligned}$$

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

$$\mathbb{E} \left[ \|y_{t+1}\|^2 | \mathcal{F}_{t,1} \right] \leq K_1 (1 + \|\lambda_t\|^2).$$

Thus, for any  $M > 0$  there exists  $K_2 > 0$  such that, on the set  $\{\sup_{\tau \leq t} \|\lambda_\tau\| < M\}$ ,

$$\mathbb{E} \left[ \|y_{t+1}\|^2 \mathbb{1}_{\{\sup_{\tau \leq t} \|\lambda_\tau\| < M\}} | \mathcal{F}_{t,1} \right] \leq K_2. \quad (23)$$

We let  $v_t = \|\beta_{\lambda,t}^{-1} \lambda_{\perp,t}\|^2 \mathbb{1}_{\{\sup_{\tau \leq t} \|\lambda_\tau\| < M\}}$ . Taking expectation over (22), noting that  $\mathbb{1}_{\{\sup_{\tau \leq t+1} \|\lambda_\tau\| < M\}} \leq \mathbb{1}_{\{\sup_{\tau \leq t} \|\lambda_\tau\| < M\}}$  we get

$$\mathbb{E}(v_{t+1}) \leq \frac{\beta_{\lambda,t}^2}{\beta_{\lambda,t+1}^2} \rho \left( \mathbb{E}(v_t) + 2\sqrt{\mathbb{E}(v_t)} \cdot \sqrt{K_2} + K_2 \right)$$

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:

$$\sup_t \mathbb{E} \left[ \|\beta_{\lambda,t}^{-1} \lambda_{\perp,t}\|^2 \mathbb{1}_{\{\sup_{\tau \leq t} \|\lambda_\tau\| \leq M\}} \right] < \infty \quad (24)$$

$$\text{and} \quad \lim_t \lambda_{\perp,t} = 0 \text{ a.s.} \quad (25)$$

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

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

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

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

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  $\{\sup_t \|\lambda_t\| \leq M\}$ , there exists  $K_3, K_4 < \infty$  such that

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

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

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

Hence, for any  $M > 0$ , assumptions (a.1) - (a.5) of B.1. from Zhang et al. (2018) are verified on the set  $\{\sup_t \|\lambda_t\| \leq M\}$ . 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  $\langle \lambda_t \rangle \xrightarrow[t \rightarrow \infty]{} \lambda^*$  a.s. As  $\lambda_t = \mathbf{1} \otimes \langle \lambda_t \rangle + \lambda_{\perp,t}$  and  $\lambda_{\perp,t} \xrightarrow[t \rightarrow \infty]{} 0$  a.s., we have for each  $i \in \mathcal{N}$ , a.s.,

$$\lambda_t^i \xrightarrow[t \rightarrow \infty]{} B_{\pi,\theta}^{-1} \cdot A_{\pi,\theta} \cdot d^\pi.$$

#### PROOF OF THEOREM 7

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

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

With local projection, actor update (6) becomes

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

So with  $h^i(\theta_t) = \mathbb{E}_{s_t \sim d^\pi} [\psi_t^i \cdot \xi_{t,\theta_t}^i | \mathcal{F}_{t,2}]$  and  $h(\theta) = [h^1(\theta), \dots, h^N(\theta)]$ , 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  $\theta \mapsto \psi_{t,\theta_t}^i$  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  $\theta$  and the ODE (12) is well-posed. This holds even when using compatible features.

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

Let  $M_t^i = \sum_{\tau=0}^{t-1} \beta_{\theta,\tau} \zeta_{\tau,1}^i$ .  $M_t^i$  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  $\{\zeta_{t,1}^i\}$  is bounded. Thus, by Assumption 5,  $\sum_t \mathbb{E} [\|M_{t+1}^i - M_t^i\|^2 | \mathcal{F}_{t,2}] = \sum_t \|\beta_{\theta,t} \zeta_{t,1}^i\|^2 < \infty$  a.s. The martingale convergence theorem ensures that  $\{M_t^i\}$  converges a.s. Thus, for any  $\epsilon > 0$ ,

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

Hence, by Kushner-Clark lemma Kushner & 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).