Supplementary Materials for "Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents"

A. Pseudocode of Algorithm 1 and Algorithm 2

In this section, we present the pseudocode of Algorithm 1 and Algorithm 2. Algorithm 1 follows the updates (3.3)-(3.5), which is based on the estimate of the local advantage function A_θ^i. This can be achieved by maintaining a consensual approximation of the global action-value function Q_θ at each agent.

Algorithm 1 The networked actor-critic algorithm based on action-value function approximation

Input: Initial values of the parameters μ_0^i, ω_0^i, $\widetilde{\omega}_0^i$, θ_0^i, $\forall i \in \mathcal{N}$; the initial state s_0 of the MDP, and stepsizes $\{\beta_{\omega,t}\}_{t\geq 0}$ and $\{\beta_{\theta,t}\}_{t\geq 0}$.
Each agent $i \in \mathcal{N}$ executes action $a_0^i \sim \pi_{\theta_0^i}^i(s_0, \cdot)$ and observes joint actions $a_0 = (a_0^1, \ldots, a_0^N)$.
Initialize the iteration counter $t \leftarrow 0$.
Repeat:
  for all $i \in \mathcal{N}$ **do**
    Observe state s_{t+1}, and reward r_{t+1}^i.
    Update $\mu_{t+1}^i \leftarrow (1-\beta_{\omega,t}) \cdot \mu_t^i + \beta_{\omega,t} \cdot r_{t+1}^i$.
    Select and execute action $a_{t+1}^i \sim \pi_{\theta_t^i}^i(s_{t+1}, \cdot)$.
  end for
  Observe joint actions $a_{t+1} = (a_{t+1}^1, \ldots, a_{t+1}^N)$.
  for all $i \in \mathcal{N}$ **do**
    Update $\delta_t^i \leftarrow r_{t+1}^i - \mu_t^i + Q_{t+1}(\omega_t^i) - Q_t(\omega_t^i)$.
    Critic step: $\widetilde{\omega}_t^i \leftarrow \omega_t^i + \beta_{\omega,t} \cdot \delta_t^i \cdot \nabla_\omega Q_t(\omega_t^i)$.
    Update $A_t^i \leftarrow Q_t(\omega_t^i) - \sum_{a^i \in \mathcal{A}^i} \pi_{\theta_t^i}^i(s_t, a^i) \cdot Q(s_t, a^i, a^{-i}; \omega_t^i)$, $\quad \psi_t^i \leftarrow \nabla_{\theta^i} \log \pi_{\theta_t^i}^i(s_t, a_t^i)$.
    Actor step: $\theta_{t+1}^i \leftarrow \theta_t^i + \beta_{\theta,t} \cdot A_t^i \cdot \psi_t^i$.
    Send $\widetilde{\omega}_t^i$ to the neighbors $\{j \in \mathcal{N} \colon (i,j) \in \mathcal{E}_t\}$ over the communication network $\mathcal{G}_t$.
  end for
  for all $i \in \mathcal{N}$ **do**
    Consensus step: $\omega_{t+1}^i \leftarrow \sum_{j \in \mathcal{N}} c_t(i,j) \cdot \widetilde{\omega}_t^j$.
  end for
  Update the iteration counter $t \leftarrow t+1$.
Until Convergence

In addition, Algorithm 2 follows the updates (3.8), (3.11), and (3.12), which is based on the estimate of the global advantage function A_θ. This can be achieved by maintaining consensual approximation of both the global state-value function V_θ and the globally averaged reward function $\overline{R}$, at each agent.

Algorithm 2 The networked actor-critic algorithm based on state-value function approximation

Input: Initial values of μ_0^i, $\widetilde{\mu}_0^i$, v_0^i, $\widetilde{v}_0^i$, λ_0^i, $\widetilde{\lambda}_0^i$, θ_0^i, $\forall i \in \mathcal{N}$; the initial state s_0 of the MDP, and stepsizes $\{\beta_{v,t}\}_{t\geq 0}$ and $\{\beta_{\theta,t}\}_{t\geq 0}$.
Each agent i implements $a_0^i \sim \pi_{\theta_0^i}(s_0, \cdot)$.
Initialize the step counter $t \leftarrow 0$.
Repeat:
  for all $i \in \mathcal{N}$ **do**
    Observe state s_{t+1}, and reward r_{t+1}^i.
    Update $\widetilde{\mu}_t^i \leftarrow (1-\beta_{v,t})\cdot \mu_t^i + \beta_{v,t}\cdot r_{t+1}^i$, $\widetilde{\lambda}_t^i \leftarrow \lambda_t^i + \beta_{v,t}\cdot[r_{t+1}^i - \overline{R}_t(\lambda_t^i)]\cdot \nabla_\lambda \overline{R}_t(\lambda_t^i)$.
    Update $\delta_t^i \leftarrow r_{t+1}^i - \mu_t^i + V_{t+1}(v_t^i) - V_t(v_t^i)$
    Critic step: $\widetilde{v}_t^i \leftarrow v_t^i + \beta_{v,t}\cdot \delta_t^i \cdot \nabla_v V_t(v_t^i)$.
    Update $\widetilde{\delta}_t^i \leftarrow \overline{R}_t(\lambda_t^i) - \mu_t^i + V_{t+1}(v_t^i) - V_t(v_t^i)$, $\psi_t^i \leftarrow \nabla_{\theta^i} \log \pi_{\theta_t^i}^i(s_t, a_t^i)$.
    Actor step: $\theta_{t+1}^i = \theta_t^i + \beta_{\theta,t}\cdot \widetilde{\delta}_t^i \cdot \psi_t^i$.
    Send $\widetilde{\mu}_t^i$, $\widetilde{\lambda}_t^i$, $\widetilde{v}_t^i$ to the neighbors over $\mathcal{G}_t$.
  end for
  for all $i \in \mathcal{N}$ **do**
    Consensus step: $\mu_{t+1}^i \leftarrow \sum_{j\in\mathcal{N}} c_t(i,j)\cdot\widetilde{\mu}_t^j$, $\lambda_{t+1}^i \leftarrow \sum_{j\in\mathcal{N}} c_t(i,j)\cdot\widetilde{\lambda}_t^j$, $v_{t+1}^i \leftarrow \sum_{j\in\mathcal{N}} c_t(i,j)\cdot\widetilde{v}_t^j$.
  end for
  Update the iteration counter $t \leftarrow t+1$.
Until Convergence

B. Proofs of the Main Results

In this section, we provide the proofs of the theoretical results presented in the paper. We first prove the policy gradient theorem for MARL, and then provide proofs for the convergence results in §4. Detailed proofs for the convergence of Algorithm 1 are given, followed by succinct proofs for the convergence of Algorithm 2 that draw parallels with the first ones.

Notation. For any vector $x \in \mathbb{R}^n$ and matrix $Y \in \mathbb{R}^{m\times n}$, we use $\|x\|$ and $\|Y\|$ to denote the Euclidean norm of x and the induced 2-norm of Y, respectively. We also use $\|x\|_\infty$ and $\|Y\|_\infty$ to denote the infinite norm of x and the induced infinite-norm of Y, respectively.

B.1. Proof of Theorem 3.1

The proof of this theorem follows the proof of the policy gradient theorem in single-agent reinforcement learning (Sutton et al., 2000), which shows that

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a)\cdot Q_\theta(s,a)\right]\\ &= \sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\theta(s)\pi_\theta(s,a)\left[\nabla_\theta \sum_{i\in\mathcal{N}} \log \pi^i_{\theta^i}(s,a^i)\right]\cdot Q_\theta(s,a),\end{aligned} \tag{B.1}$$

where Q_θ is the action-value function defined in (2.3), and d_θ denotes the stationary distribution of the Markov chain induced by policy π_θ. Here the second equality in (B.1) holds because π_θ is the product of local policy functions. Hence, the gradient with respect to the parameter θ^i becomes

$$\nabla_{\theta^i} J(\theta) = \sum_{s\in\mathcal{S},a\in\mathcal{A}} d_\theta(s)\pi_\theta(s,a)\cdot\nabla_{\theta^i}\log\pi^i_{\theta^i}(s,a^i)\cdot Q_\theta(s,a), \tag{B.2}$$

which proves the first equality in (3.2). Moreover, since $\sum_{a^i\in\mathcal{A}^i}\pi^i_{\theta^i}(s,a^i) = 1$, we have $\nabla_{\theta^i}\left[\sum_{a^i\in\mathcal{A}^i}\pi^i_{\theta^i}(s,a^i)\right] = 0$. To simplify the notation, for each $i\in\mathcal{N}$, we define a^{-i} as the joint actions of all agents except i, and let $\mathcal{A}^{-i} = \prod_{j\neq i}\mathcal{A}^j$. Thus, for any function $F\colon \mathcal{S}\times\mathcal{A}^{-i}\to\mathbb{R}$ which does not rely on a^i, we have for any $s\in\mathcal{S}$ that

$$\begin{aligned}&\sum_{a\in\mathcal{A}}\pi_\theta(s,a)\cdot\left[\nabla_{\theta^i}\log\pi^i_{\theta^i}(s,a^i)\right]\cdot F(s,a^{-i})\\ &\quad= \sum_{a^{-i}\in\mathcal{A}^{-i}} F(s,a^{-i})\cdot\Big[\prod_{j\in\mathcal{N},j\neq i}\pi^j_{\theta^j}(s,a^j)\Big]\cdot\Big[\sum_{a^i\in\mathcal{A}^i}\nabla_{\theta^i}\pi^i_{\theta^i}(s,a^i)\Big] = 0.\end{aligned} \tag{B.3}$$

Thus, replacing F in (B.3) by the value function V_θ and function $\widetilde{V}^i_\theta$ defined in (3.1) and combined with (B.2), we establish (3.2), which concludes the proof. □

B.2. Proof of Theorem 4.6

To proceed with the proof, we first establish the stability of the update $\{\omega_t\}$. This stability condition serves as an assumption in the original two-time-scale SA analysis (Borkar, 2008, Chapter 6.1). It is usually verified separately using some other sufficient conditions (Borkar & Meyn, 2000; Andrieu et al., 2005). We will directly use the lemma in the convergence analysis to follow and defer its proof to Appendix §C.

Lemma B.1. Under Assumptions 2.2, and 4.2-4.5, the sequence $\{\omega^i_t\}$ generated from (3.3) is bounded almost surely, i.e., $\sup_t\|\omega^i_t\| < \infty$ a.s., for any $i\in\mathcal{N}$.

As in the classical two-time-scale SA analysis (Borkar, 2008), we let the policy parameter θ_t be fixed as $\theta_t\equiv\theta$ when analyzing the convergence of the critic step. This allows us to show that ω_t will converge to some ω_θ depending on θ, which can be further utilized to simplify the proof of convergence for the slower time scale. In fact, with linear function approximation, one can rewrite the actor step (3.5) for Algorithm 1 as

$$\theta^i_{t+1} = \Gamma^i\left(\theta^i_t + \beta_{\omega,t}\cdot\frac{\beta_{\theta,t}}{\beta_{\omega,t}}\cdot A^i_t\cdot\psi^i_t\right), \tag{B.4}$$

where the projection Γ^i follows from Assumption 4.1. Note that A_t^i is also bounded a.s., since the parameter ω_t^i is bounded by Lemma B.1, and the feature ϕ_t is bounded by and Assumption 4.5. Moreover, ψ_t^i is bounded a.s. by Assumption 2.2 since it is a continuous function over a bounded set Θ^i. Therefore, $\sup_t \|A_t^i \cdot \psi_t^i\| < \infty$ a.s. Now since $\beta_{\theta,t} \cdot \beta_{\omega,t}^{-1} \to 0$ by Assumption 4.3, it follows that $\beta_{\theta,t} \cdot \beta_{\omega,t}^{-1} \cdot A_t^i \cdot \psi_t^i \to 0$ as $t \to \infty$. Now the update in (B.4) can be viewed to track the ordinary differential equation (ODE) $\dot{\theta}^i(t) = 0$. Hence, one may let θ_t be a constant when analyzing the faster update of ω_t^i. For notational simplicity, we eliminate the notations associated with θ unless otherwise noted.

Let $\{\mathcal{F}_{t,1}\}$ be the filtration with $\mathcal{F}_{t,1} = \sigma(r_\tau, \mu_\tau, \omega_\tau, s_\tau, a_\tau, \mathrm{C}_{\tau-1}, \tau \le t)$, which is an increasing σ-algebra over time t. For notational convenience, let $r_t = (r_t^1, \cdots, r_t^N)^\top$, $\mu_t = (\mu_t^1, \cdots, \mu_t^N)^\top$, $\omega_t = [(\omega_t^1)^\top, \cdots, (\omega_t^N)^\top]^\top$, and $\delta_t = [(\delta_t^1)^\top, \cdots, (\delta_t^N)^\top]^\top$. The update of ω_t in (3.3) can be rewritten in a compact form as

$$\omega_{t+1} = (\mathrm{C}_t \otimes \mathrm{I})\big(\omega_t + \beta_{\omega,t} \cdot y_{t+1}\big), \tag{B.5}$$

where $\otimes$ denotes the Kronecker product, $\mathrm{I} \in \mathbb{R}^{K\times K}$ is the identity matrix, and $y_{t+1} = (\delta_t^1 \phi_t^\top, \cdots, \delta_t^N \phi_t^\top)^\top \in \mathbb{R}^{KN}$. Define the operator $\langle\cdot\rangle : \mathbb{R}^{KN} \to \mathbb{R}^K$ by letting

$$\langle\omega\rangle = \frac{1}{N}(\mathbb{1}^\top \otimes \mathrm{I})\omega = \frac{1}{N}\sum_{i\in\mathcal{N}} \omega^i \tag{B.6}$$

for any $\omega = [(\omega^1)^\top, \cdots, (\omega^N)^\top]^\top \in \mathbb{R}^{KN}$ and $\omega^i \in \mathbb{R}^K$ with $i \in \mathcal{N}$. That is, $\langle\omega\rangle \in \mathbb{R}^K$ represents the average of the vectors in $\{\omega^1, \cdots, \omega^N\}$, which are local to individual agents. Let $\mathcal{J} = (1/N \cdot \mathbb{1}\mathbb{1}^\top) \otimes \mathrm{I}$ be the projection operator that projects the vector into the *consensus* subspace $\{\mathbb{1} \otimes u : u \in \mathbb{R}^K\}$. Thus we have $\mathcal{J}\omega = \mathbb{1} \otimes \langle\omega\rangle$. Moreover, we define $\mathcal{J}_\perp$ as the operator that projects the vector to the *disagreement* subspace, i.e., $\mathcal{J}_\perp = \mathrm{I} - \mathcal{J}$. Thus the disagreement vector $\omega_\perp = \mathcal{J}_\perp\omega$ is written as

$$\omega_\perp = \mathcal{J}_\perp\omega = \omega - \mathbb{1} \otimes \langle\omega\rangle. \tag{B.7}$$

The proof of Theorem 4.6 then consists of two steps. In particular, we separate the iteration ω_t as the sum of a vector in this *consensus* space and a vector in the *disagreement* space, i.e., $\omega_t = \omega_{\perp,t} + \mathbb{1} \otimes \langle\omega_t\rangle$. We first show the a.s. convergence of the disagreement vector sequence $\{\omega_{\perp,t}\}$ to zero. Then, we prove that the consensus vector sequence $\{\mathbb{1} \otimes \langle\omega_t\rangle\}$ converges to the equilibrium such that $\langle\omega_t\rangle$ satisfies (4.2).

Step 1. In this step, we establish that $\lim_t \omega_{\perp,t} = 0$ a.s. To this end, we first have the following lemma on the boundedness of the sequence $\{\mu_t^i\}$ for any $i \in \mathcal{N}$.

Lemma B.2. Under Assumptions 2.2 and 4.2, the sequence $\{\mu_t^i\}$ generated as in (3.3) is bounded almost surely, i.e., $\sup_t |\mu_t^i| < \infty$ a.s., for any $i \in \mathcal{N}$.

Proof. The local update in (3.3) forms a stochastic approximation iteration, whose asymptotic behavior can be captured by the ODE

$$\dot{\mu}^i = -\mu^i + \sum_{s\in\mathcal{S}} d_\theta(s) \sum_{a\in\mathcal{A}} \pi_\theta(s,a) R^i(s,a). \tag{B.8}$$

Let $f(\mu^i)$ denote the right hand side (RHS) of (B.8), which is Lipschitz continuous in μ^i. Moreover, define $f_c(\mu^i) = f(c\mu^i) \cdot c^{-1}$, then $f_\infty(\mu^i) = \lim_c f(c\mu^i) \cdot c^{-1} = -\mu^i$ exists. Therefore, the ODE $\dot{\mu}^i = f_\infty(\mu^i)$ has origin as the unique asymptotically stable equilibrium. In addition, since r_t^i is uniformly bounded, we have

$$\mathbb{E}\big[|r_{t+1}^i - \mathbb{E}(r_{t+1}^i \,|\, \mathcal{F}_{t,1})|^2 \,\big|\, \mathcal{F}_{t,1}\big] \le K_0 \cdot (1 + |\mu_t^i|^2)$$

for some $K_0 < \infty$. Therefore, the conditions (a.1) and (a.4) in Assumption D.1 are satisfied. (See Appendix §D.1 for details.) By Assumption 2.2, (a.2) in Assumption D.1 also holds. We thus conclude that $\sup_t |\mu_t^i| < \infty$ from Theorem D.3 (see also Theorem 9 on page 74-75 in Borkar (2008)). □

Let $z_t^i = [\mu_t^i, (\omega_t^i)^\top]^\top$ and $z_t = [(z_t^1)^\top, \cdots, (z_t^N)^\top]^\top$. By Lemma B.1, we have $\mathbb{P}(\sup_t \|z_t\| < \infty) = 1$, which means that $\mathbb{P}(\bigcup_{M\in\mathbb{Z}^+}\{\sup_t \|z_t\| \le M\}) = 1$, with $\mathbb{Z}^+$ denoting the set of positive integers. Hence, it suffices to show that $\lim_t \omega_{\perp,t}\mathbb{I}_{\{\sup_t \|z_t\| \le M\}} = 0$, for any $M \in \mathbb{Z}^+$, where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. We then establish that $\mathbb{E}(\|\beta_{\omega,t}^{-1}\omega_{\perp,t}\|^2)$ is bounded on $\{\sup_t \|z_t\| \le M\}$, for any $M > 0$.

Lemma B.3. Under Assumptions 4.2-4.5, for any $M > 0$, we have

$$\sup_t \mathbb{E}\big(\|\beta_{\omega,t}^{-1}\omega_{\perp,t}\|^2 \mathbb{I}_{\{\sup_t \|z_t\| \le M\}}\big) < \infty.$$

Proof. First note that by (a.1) in Assumption 4.4 and the fact that $(\mathrm{A}\otimes\mathrm{B})(\mathrm{C}\otimes\mathrm{D}) = (\mathrm{AC})\otimes(\mathrm{BD})$, we have

$$(\mathrm{C}_t \otimes \mathrm{I})(\mathbb{1}\otimes\langle\omega\rangle) = (\mathrm{C}_t\mathbb{1})\otimes\langle\omega\rangle = \mathbb{1}\otimes\langle\omega\rangle.$$

Hence, $\omega_{\perp,t+1}$ has the form $\omega_{\perp,t+1} = \mathcal{J}_\perp[(\mathrm{C}_t\otimes\mathrm{I})(\omega_t + \beta_{\omega,t}y_{t+1})] = \mathcal{J}_\perp[(\mathrm{C}_t\otimes\mathrm{I})(\omega_{\perp,t} + \beta_{\omega,t}y_{t+1})]$, since $\mathcal{J}_\perp(\mathbb{1}\otimes\langle\omega\rangle)$ is zero. Thus, by the definition of $\mathcal{J}_\perp$ in (B.7), the vector $\omega_{\perp,t+1}$ satisfies

$$\omega_{\perp,t+1} = [(\mathrm{I} - \mathbb{1}\mathbb{1}^\top/N)\otimes\mathrm{I}](\mathrm{C}_t\otimes\mathrm{I})(\omega_{\perp,t} + \beta_{\omega,t}y_{t+1}) = [(\mathrm{I}-\mathbb{1}\mathbb{1}^\top/N)\mathrm{C}_t\otimes\mathrm{I}](\omega_{\perp,t} + \beta_{\omega,t}y_{t+1}). \quad \text{(B.9)}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}\big(\big\|\beta_{\omega,t+1}^{-1}\omega_{\perp,t+1}\big\|^2 \,\big|\, \mathcal{F}_{t,1}\big)\\
&= \frac{\beta_{\omega,t}^2}{\beta_{\omega,t+1}^2}\cdot\mathbb{E}\big\{\big(\beta_{\omega,t}^{-1}\omega_{\perp,t} + y_{t+1}\big)^\top\big[\mathrm{C}_t^\top(\mathrm{I}-\mathbb{1}\mathbb{1}^\top/N)\mathrm{C}_t\otimes\mathrm{I}\big]\big(\beta_{\omega,t}^{-1}\omega_{\perp,t} + y_{t+1}\big)\,\big|\,\mathcal{F}_{t,1}\big\}\\
&\le \frac{\beta_{\omega,t}^2}{\beta_{\omega,t+1}^2}\cdot\rho\cdot\mathbb{E}\big[\big(\beta_{\omega,t}^{-1}\omega_{\perp,t} + y_{t+1}\big)^\top\big(\beta_{\omega,t}^{-1}\omega_{\perp,t} + y_{t+1}\big)\,\big|\,\mathcal{F}_{t,1}\big]\\
&\le \frac{\beta_{\omega,t}^2}{\beta_{\omega,t+1}^2}\cdot\rho\cdot\Big\{\big\|\beta_{\omega,t}^{-1}\omega_{\perp,t}\big\|^2 + 2\cdot\big\|\beta_{\omega,t}^{-1}\omega_{\perp,t}\big\|\cdot\big[\mathbb{E}\big(\|y_{t+1}\|^2\,\big|\,\mathcal{F}_{t,1}\big)\big]^{\frac{1}{2}} + \mathbb{E}\big(\|y_{t+1}\|^2\,\big|\,\mathcal{F}_{t,1}\big)\Big\}, \quad \text{(B.10)}
\end{aligned}$$

where ρ represents the spectral norm of $\mathbb{E}[\mathrm{C}_t^\top(\mathrm{I}-\mathbb{1}\mathbb{1}^\top/N)\mathrm{C}_t]$. By (a.2) in Assumption 4.4, we have $\rho\in[0,1)$. The first inequality in (B.10) is due to the conditional independence of C_t and r_{t+1}^i for all $i\in\mathcal{N}$, and thus y_{t+1}, by (a.3) in Assumption 4.4, and the second inequality is due to the Cauchy-Schwarz inequality. Moreover, by the definition of y_{t+1}, we have

$$\begin{aligned}
\mathbb{E}\big(\|y_{t+1}\|^2\,|\,\mathcal{F}_{t,1}\big) &= \mathbb{E}\bigg(\sum_{i\in\mathcal{N}}\big\|\delta_t^i\phi_t\big\|^2\,\bigg|\,\mathcal{F}_{t,1}\bigg) = \mathbb{E}\bigg[\sum_{i\in\mathcal{N}}\big\|\big(r_{t+1}^i - \mu_t^i + \phi_{t+1}^\top\omega_t^i - \phi_t^\top\omega_t^i\big)\phi_t\big\|^2\,\bigg|\,\mathcal{F}_{t,1}\bigg]\\
&\le 3\cdot\mathbb{E}\bigg[\sum_{i\in\mathcal{N}}\big\|r_{t+1}^i\phi_t\big\|^2 + \big\|\mu_t^i\phi_t\big\|^2 + \big\|\phi_t\big(\phi_{t+1}^\top - \phi_t^\top\big)\big\|^2\big\|\omega_t^i\big\|^2\,\bigg|\,\mathcal{F}_{t,1}\bigg], \quad \text{(B.11)}
\end{aligned}$$

where the inequality follows from that by Assumption 4.5, $\mathbb{E}[\|\phi_t(\phi_{t+1}^\top - \phi_t^\top)\|^2\,|\,\mathcal{F}_{t,1}]$ and $\mathbb{E}(\|\phi_t\|^2\,|\,\mathcal{F}_{t,1})$ are both uniformly bounded for any $s_t\in\mathcal{S}$ and $a_t\in\mathcal{A}$. Moreover, by Assumption 4.2, we have $\mathbb{E}(|r_{t+1}^i|^2\,|\,\mathcal{F}_{t,1}) = \mathbb{E}(|r_{t+1}^i|^2\,|\,s_t,a_t)$ also uniformly bounded. Thus the RHS of (B.11) is bounded on the set $\{\sup_{\tau\le t}\|z_\tau\|\le M\}$ for any $M>0$ as follows, i.e., there exists $K_1<\infty$, such that

$$\mathbb{E}\big(\|y_{t+1}\|^2\mathbb{I}_{\{\sup_{\tau\le t}\|z_\tau\|\le M\}}\,\big|\,\mathcal{F}_{t,1}\big) \le K_1\cdot\bigg[1 + \sum_{i\in\mathcal{N}}\mathbb{E}\big(|r_{t+1}^i|^2\,\big|\,\mathcal{F}_{t,1}\big)\bigg]. \quad \text{(B.12)}$$

Let $\eta_t = \|\beta_{\omega,t}^{-1}\omega_{\perp,t}\|^2\mathbb{I}_{\{\sup_{\tau\le t}\|z_\tau\|\le M\}}$ and note that $\mathbb{I}_{\{\sup_{\tau\le t+1}\|z_\tau\|\le M\}} \le \mathbb{I}_{\{\sup_{\tau\le t}\|z_\tau\|\le M\}}$. Then by taking expectation over both sides of (B.10), we obtain that there exists $K_2 = K_1\cdot[1+\mathbb{E}(\sum_{i\in\mathcal{N}}|r_{t+1}^i|^2)] < \infty$ such that

$$\mathbb{E}(\eta_{t+1}) \le \frac{\beta_{\omega,t}^2}{\beta_{\omega,t+1}^2}\cdot\rho\cdot\big[\mathbb{E}(\eta_t) + 2\sqrt{\mathbb{E}(\eta_t)}\cdot\sqrt{K_2} + K_2\big]. \quad \text{(B.13)}$$

Since $\lim_t \beta_{\omega,t}^2\cdot\beta_{\omega,t+1}^{-2} = 1$ and $\rho<1$, for any $\delta>0$, there exists a large enough t_0 such that for any $t>t_0$, $\beta_{\omega,t}^2\cdot\beta_{\omega,t+1}^{-2}\cdot\rho \le 1-\delta$. Hence, there exist positive constants K_3 and b such that for any $t\ge t_0$,

$$\mathbb{E}(\eta_{t+1}) \le (1-\delta)\cdot\big[\mathbb{E}(\eta_t) + 2\sqrt{\mathbb{E}(\eta_t)}\cdot\sqrt{K_2} + K_2\big] \le \big(1-\delta/2\big)\cdot\mathbb{E}(\eta_t) + b\cdot\mathbb{I}_{\{\mathbb{E}(\eta_t)\le K_3\}}.$$

By induction, we obtain that $\mathbb{E}(\eta_t) \leq (1-\delta/2)^{t-t_0}\mathbb{E}(\eta_{t_0}) + 2b/\delta$. Hence, we have $\sup_t \mathbb{E}(\eta_t) < \infty$. In addition, since $\mathbb{I}_{\{\sup_t \|z_t\| \leq M\}} \leq \mathbb{I}_{\{\sup_{\tau \leq t} \|z_\tau\| \leq M\}}$, we further obtain

$$\sup_t \mathbb{E}\big(\|\beta_{\omega,t}^{-1}\omega_{\perp,t}\|^2 \mathbb{I}_{\{\sup_t \|z_t\| \leq M\}}\big) < \infty,$$

which concludes the proof. □

Therefore, by Lemma B.3, we obtain that for any $M > 0$, there exists a constant $K_4 < \infty$, such that for any $t \geq 0$, $\mathbb{E}(\|\omega_{\perp,t}\|^2) \leq K_4 \cdot \beta_{\omega,t}^2$ on the set $\{\sup_t \|z_t\| \leq M\}$. Since $\sum_t \beta_{\omega,t}^2 < \infty$ by Assumption 4.3, we have that $\sum_t \mathbb{E}(\|\omega_{\perp,t}\|^2 \mathbb{I}_{\{\sup_t \|z_t\| \leq M\}})$ is finite by Fubini's theorem. This shows that $\sum_t \|\omega_{\perp,t}\|^2 \mathbb{I}_{\{\sup_t \|z_t\| \leq M\}} < \infty$ a.s., which further yields $\lim_t \omega_{\perp,t} \mathbb{I}_{\{\sup_t \|z_t\| \leq M\}} = 0$ a.s. By Lemmas B.1 and B.2, $\{\sup_t \|z_t\| < \infty\}$ holds with probability 1. This shows that $\lim_t \omega_{\perp,t} = 0$ a.s., and thus concludes **Step 1**.

Step 2. We now proceed to show the convergence of the consensus vector $\mathbb{1} \otimes \langle \omega_t \rangle$. According to the update in (B.5) and definition (B.6), the iteration of $\langle \omega_t \rangle$ has the form

$$\langle \omega_{t+1} \rangle = \frac{1}{N}(\mathbb{1}^\top \otimes \mathrm{I})(\mathrm{C}_t \otimes \mathrm{I})(\mathbb{1} \otimes \langle \omega_t \rangle + \omega_{t,\perp} + \beta_{\omega,t} y_{t+1}) = \langle \omega_t \rangle + \beta_{\omega,t} \langle (\mathrm{C}_t \otimes \mathrm{I})(y_{t+1} + \beta_{\omega,t}^{-1}\omega_{\perp,t}) \rangle.$$

Hence, we write the updates for $\langle \omega_t \rangle$ and $\langle \mu_t \rangle$ as

$$\langle \mu_{t+1} \rangle = \langle \mu_t \rangle + \beta_{\omega,t} \cdot \mathbb{E}\big(\overline{r}_{t+1} - \langle \mu_t \rangle \,\big|\, \mathcal{F}_{t,1}\big) + \beta_{\omega,t} \cdot \xi_{t+1,1}, \tag{B.14}$$

$$\langle \omega_{t+1} \rangle = \langle \omega_t \rangle + \beta_{\omega,t} \cdot \mathbb{E}\big(\langle \delta_t \rangle \phi_t \,\big|\, \mathcal{F}_{t,1}\big) + \beta_{\omega,t} \cdot \xi_{t+1,2}, \tag{B.15}$$

where $\xi_{t+1,1} = \overline{r}_{t+1} - \mathbb{E}(\overline{r}_{t+1} \,|\, \mathcal{F}_{t,1})$ and $\xi_{t+1,2}$ is

$$\xi_{t+1,2} = \langle (\mathrm{C}_t \otimes \mathrm{I})(y_{t+1} + \beta_{\omega,t}^{-1}\omega_{\perp,t}) \rangle - \mathbb{E}\big(\langle \delta_t \rangle \phi_t \,\big|\, \mathcal{F}_{t,1}\big).$$

Note that $\mathbb{E}(\overline{r}_{t+1} - \langle \mu_t \rangle \,|\, \mathcal{F}_{t,1})$ is Lipschitz continuous in $\langle \mu_t \rangle$. Recall that $\langle \delta_t \rangle$ has the form

$$\langle \delta_t \rangle = \frac{1}{N}\sum_{i \in \mathcal{N}} r_{t+1}^i - \mu_t^i + \phi_{t+1}^\top \omega_t^i - \phi_t^\top \omega_t^i = \overline{r}_{t+1} - \langle \mu_t \rangle + \phi_{t+1}^\top \langle \omega_t \rangle - \phi_t^\top \langle \omega_t \rangle.$$

Hence, $\mathbb{E}(\langle \delta_t \rangle \phi_t \,|\, \mathcal{F}_{t,1})$ is Lipschitz continuous in both $\langle \omega_t \rangle$ and $\langle \mu_t \rangle$, and thus the condition (a.1) in Assumption D.1 (See Appendix §D.1) is satisfied.

Note that $\xi_{t,1}$ is a martingale difference sequence and satisfies

$$\mathbb{E}\big(\|\xi_{t+1,1}\|^2 \,\big|\, \mathcal{F}_{t,1}\big) \leq K_5 \cdot \big(1 + \|\langle \omega_t \rangle\|^2 + \|\langle \mu_t \rangle\|^2\big), \tag{B.16}$$

for some $K_5 < \infty$, since $\overline{r}_{t+1}$ is uniformly bounded. In addition, the term $\xi_{t,2}$ is also a martingale difference sequence, since

$$\mathbb{E}\big[\langle (\mathrm{C}_t \otimes \mathrm{I})(y_{t+1} + \beta_{\omega,t}^{-1}\omega_{\perp,t}) \rangle \,\big|\, \mathcal{F}_{t,1}\big] = \mathbb{E}\big[\langle (\mathrm{C}_t \otimes \mathrm{I}) y_{t+1} \rangle \,\big|\, \mathcal{F}_{t,1}\big] = \mathbb{E}\big(\langle y_{t+1} \rangle \,\big|\, \mathcal{F}_{t,1}\big) = \mathbb{E}\big(\langle \delta_t \rangle \phi_t \,\big|\, \mathcal{F}_{t,1}\big),$$

which results from the facts that $\langle \omega_{\perp,t} \rangle = 0$ and that $\mathbb{1}^\top \mathbb{E}(\mathrm{C}_t) = \mathbb{1}^\top$. Moreover, we have

$$\mathbb{E}\big(\|\xi_{t+1,2}\|^2 \,\big|\, \mathcal{F}_{t,1}\big) \leq 2 \cdot \mathbb{E}\big(\big\|y_{t+1} + \beta_{\omega,t}^{-1}\omega_{\perp,t}\big\|_{\mathrm{G}_t}^2 \,\big|\, \mathcal{F}_{t,1}\big) + 2 \cdot \big\|\mathbb{E}\big(\langle \delta_t \rangle \phi_t \,\big|\, \mathcal{F}_{t,1}\big)\big\|^2, \tag{B.17}$$

where $\mathrm{G}_t = \mathrm{C}_t^\top \mathbb{1}\mathbb{1}^\top \mathrm{C}_t \otimes \mathrm{I} \cdot N^{-2}$. Note that G_t has bounded spectral norm since C_t is a stochastic matrix. Thus the first term in (B.17) can be further bounded over the set $\{\sup_t \|z_t\| \leq M\}$, for any $M > 0$. Notably, there exist $K_6, K_7 < \infty$ such that

$$\begin{aligned}
&\mathbb{E}\big(\big\|y_{t+1} + \beta_{\omega,t}^{-1}\omega_{\perp,t}\big\|_{\mathrm{G}_t}^2 \,\big|\, \mathcal{F}_{t,1}\big) \cdot \mathbb{I}_{\{\sup_t \|z_t\| \leq M\}} \\
&\quad \leq K_6 \cdot \mathbb{E}\big(\big\|y_{t+1}\big\|^2 + \big\|\beta_{\omega,t}^{-1}\omega_{\perp,t}\big\|^2 \,\big|\, \mathcal{F}_{t,1}\big) \cdot \mathbb{I}_{\{\sup_t \|z_t\| \leq M\}} < K_7,
\end{aligned}$$

where the second inequality follows from (B.12) and Lemma B.3. Moreover, the second term in (B.17) can be bounded by $\|\mathbb{E}(\langle\delta_t\rangle\phi_t \,|\, \mathcal{F}_{t,1})\|^2 \le K_8 \cdot (1 + \|\langle\omega_t\rangle\|^2 + \|\langle\mu_t\rangle\|^2)$ with some $K_8 < \infty$, due to the boundedness of r^i_{t+1} and ϕ_t (from Assumptions 4.2 and 4.5). Hence, for any $M > 0$, it follows that

$$\mathbb{E}\big(\|\xi_{t+1,1}\|^2 \,\big|\, \mathcal{F}_{t,1}\big) \le K_9 \cdot \big(1 + \|\langle\omega_t\rangle\|^2 + \|\langle\mu_t\rangle\|^2\big), \tag{B.18}$$

over the set $\{\sup_t \|z_t\| \le M\}$ for some $K_9 < \infty$. This verifies that on the set $\{\sup_t \|z_t\| \le M\}$ for any $M > 0$, the condition (a.4) in Assumption D.1 is satisfied.

Now consider the following ODE that captures the asymptotic behavior of (B.14) and (B.15)

$$\dot{\langle z\rangle} = \begin{pmatrix}\dot{\langle\mu\rangle}\\ \dot{\langle\omega\rangle}\end{pmatrix} = \begin{pmatrix}-1 & 0\\ -\Phi^\top \mathrm{D}^{s,a}_\theta \mathbb{1} & \Phi^\top \mathrm{D}^{s,a}_\theta (P^\theta - \mathrm{I})\Phi\end{pmatrix}\begin{pmatrix}\langle\mu\rangle\\ \langle\omega\rangle\end{pmatrix} + \begin{pmatrix}J(\theta)\\ \Phi^\top \mathrm{D}^{s,a}_\theta \overline{R}\end{pmatrix}. \tag{B.19}$$

Recall that $\mathrm{D}^{s,a}_\theta = \mathrm{diag}[d_\theta(s)\cdot\pi_\theta(s,a), s\in\mathcal{S}, a\in\mathcal{A}]$. Let the RHS of the ODE (B.19) be $h(\langle z\rangle)$, then $h(\langle z\rangle)$ is Lipschitz continuous in $\langle z\rangle$, which satisfies the condition (a.1) in Assumption D.1. By the Perron-Frobenius theorem and Assumption 2.2, the stochastic matrix P^θ has a simple eigenvalue of 1, and its remaining eigenvalues have real parts less than 1. Hence, $(P^\theta - \mathrm{I})$ has all eigenvalues with negative real parts but one zero, so does the matrix $\Phi^\top \mathrm{D}^{s,a}_\theta (P^\theta - \mathrm{I})\Phi$, since Φ is full column rank by Assumption 4.5. The simple eigenvalue of zero has eigen-vector ν that satisfies $\Phi\nu = \alpha\mathbb{1}$ for some $\alpha \ne 0$, since $\alpha\mathbb{1}$ lies in the eigen-space of $\mathrm{D}^{s,a}_\theta(P^\theta - \mathrm{I})$ associated with zero. By Assumption 4.5, however, this will not happen with any choice of Φ since $\Phi\nu \ne \alpha\mathbb{1}$ for any $\nu \in \mathbb{R}^K$. Hence, the ODE (B.19) is globally asymptotically stable and has its equilibrium satisfying

$$-\langle\mu\rangle = J(\theta), \qquad \Phi^\top \mathrm{D}^{s,a}_\theta \big[\overline{R} - \langle\mu\rangle\mathbb{1} + P^\theta\Phi\langle\omega\rangle - \Phi\langle\omega\rangle\big] = 0. \tag{B.20}$$

Note that the corresponding solution for $\langle\mu\rangle$ at equilibrium is $J(\theta)$, whereas the solution for $\langle\omega\rangle$ has the form $\omega_\theta + \alpha\nu$ with any $\alpha\in\mathbb{R}$ and $\nu\in\mathbb{R}^K$ such that $\Phi\nu = \mathbb{1}$. While by Assumption 4.5, $\Phi\nu \ne \mathbb{1}$, thus the term ω_θ is unique, and it follows that $\Phi^\top \mathrm{D}^{s,a}_\theta [T^Q_\theta(\Phi\omega_\theta) - \Phi\omega_\theta] = 0$ with T^Q_θ as defined in (4.1).

Recall from Lemmas B.1 and B.2 that $\{z_t\}$ is bounded a.s., so is the sequence $\{\langle z_t\rangle\}$. Hence all conditions for Theorem D.2 to hold are satisfied. For the concatenated vector $\langle z_t\rangle = (\langle\mu_t\rangle, \langle\omega_t\rangle^\top)^\top$, we thus have $\lim_t\langle\mu_t\rangle = J(\theta)$ and $\lim_t\langle\omega_t\rangle = \omega_\theta$ over the set $\{\sup_t\|z_t\| \le M\}$ for any $M > 0$. By Lemmas B.1 and B.2, this holds with probability 1, which concludes **Step 2**. Combined with **Step 1**, we arrive at the conclusion that $\lim_t \omega^i_t = \omega_\theta$ for any $i\in\mathcal{N}$, which completes the proof for Theorem 4.6. □

B.3. Proof of Theorem 4.7

Let $\mathcal{F}_{t,2} = \sigma(\theta_\tau, \tau \le t)$ be the σ-field generated by $\{\theta_\tau, \tau \le t\}$. In addition, we define

$$\zeta^i_{t+1,1} = A^i_t\cdot\psi^i_t - \mathbb{E}_{s_t\sim d_{\theta_t}, a_t\sim\pi_{\theta_t}}\big(A^i_t\cdot\psi^i_t \,\big|\, \mathcal{F}_{t,2}\big), \qquad \zeta^i_{t+1,2} = \mathbb{E}_{s_t\sim d_{\theta_t}, a_t\sim\pi_{\theta_t}}\big[\big(A^i_t - A^i_{t,\theta_t}\big)\cdot\psi^i_t \,\big|\, \mathcal{F}_{t,2}\big],$$

where A^i_{t,θ_t} is as defined in (4.3) with $\theta = \theta_t$. Then the actor update in (3.5) with a local projection becomes

$$\theta^i_{t+1} = \Gamma^i\Big[\theta^i_t + \beta_{\theta,t}\mathbb{E}_{s_t\sim d_{\theta_t}, a_t\sim\pi_{\theta_t}}\big(A^i_{t,\theta_t}\cdot\psi^i_t \,\big|\, \mathcal{F}_{t,2}\big) + \beta_{\theta,t}\zeta^i_{t+1,1} + \beta_{\theta,t}\zeta^i_{t+1,2}\Big]. \tag{B.21}$$

Note that $\zeta^i_{t+1,2} = o(1)$ since the critic converges, i.e., $A^i_t \to A^i_{t,\theta_t}$, at the faster time scale. Moreover, letting $M^i_t = \sum_{\tau=0}^t \beta_{\theta,\tau}\zeta^i_{\tau+1,1}$, $\{M^i_t\}$ is a martingale sequence. Since the sequences $\{\omega^i_t\}$, $\{\psi^i_t\}$, and $\{\phi_t\}$ are all bounded, the sequence $\{\zeta^i_{t,1}\}$ is also bounded. Hence, by Assumption 4.3, we have

$$\sum_t \mathbb{E}\big(\big\|M^i_{t+1} - M^i_t\big\|^2 \,\big|\, \mathcal{F}_{t,2}\big) = \sum_{t\ge1}\big\|\beta_{\theta,t}\zeta^i_{t+1,1}\big\|^2 < \infty \text{ a.s.}$$

By the martingale convergence theorem (Proposition VII-2-3(c) on page 149 of Neveu (1975)), the martingale sequence $\{M^i_t\}$ converges a.s. Thus, for any $\epsilon > 0$, we have

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

In addition, let

$$g^i(\theta_t) = \mathbb{E}_{s_t\sim d_{\theta_t}, a_t\sim\pi_{\theta_t}}\left(A^i_{t,\theta_t}\cdot\psi^i_t \,|\, \mathcal{F}_{t,2}\right) = \sum_{s_t\in\mathcal{S}, a_t\in\mathcal{A}} d_{\theta_t}(s_t)\cdot\pi_{\theta_t}(s_t,a_t)\cdot A^i_{t,\theta_t}\cdot\psi^i_{t,\theta_t},$$

then we show that $g^i(\theta_t)$ is continuous in θ^i_t as follows. First, ψ^i_{t,θ_t} is continuous by Assumption 2.2. Also, the term $d_{\theta_t}(s_t)\cdot\pi_{\theta_t}(s_t,a_t)$ is continuous in θ^i_t since it is the stationary distribution and thus is the solution to $d_{\theta_t}(s)\cdot\pi_{\theta_t}(s,a) = \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} P^{\theta_t}(s',a'\,|\,s,a)\cdot d_{\theta_t}(s')\cdot\pi_{\theta_t}(s',a')$ and $\sum_{s\in\mathcal{S},a\in\mathcal{A}} d_{\theta_t}(s)\cdot\pi_{\theta_t}(s,a) = 1$, where $P^{\theta_t}(s',a'\,|\,s,a) = P(s'\,|\,s,a)\cdot\pi_{\theta_t}(s',a')$. The unique solution to this set of linear equations can be verified to be continuous in θ_t, noting that $\pi_{\theta_t}(s,a) > 0$ by Assumption 2.2. In addition, A^i_{t,θ_t} is continuous in θ^i_t since ω_{θ_t} is the unique solution to the linear equation $\Phi^\top \mathrm{D}^{s,a}_\theta[T^Q_\theta(\Phi\omega_\theta) - \Phi\omega_\theta] = 0$ and can also be verified to be continuous in θ_t. Therefore, by Kushner-Clark lemma (Kushner & Clark, 1978, page 191-196) (see also Theorem D.5 in Appendix §D.2), the update in (B.21) converges a.s. to the set of asymptotically stable equilibria of the ODE (4.5) for each $i\in\mathcal{N}$, which concludes the proof. □

The proof for the convergence of Algorithm 2 is similar to the proofs in B.2 and B.3. To avoid duplication, we leave out some of the details in the proofs to follow.

B.4. Proof of Theorem 4.9

Let $z^i_t = [\mu^i_t, (\lambda^i_t)^\top, (v^i_t)^\top]^\top \in \mathbb{R}^{1+M+L}$. We first have the following lemma on the stability of the updates of $\{z^i_t\}$, as in Lemma B.1, the proof of which is provided in Appendix §C.

Lemma B.4. Under Assumptions 2.2, 4.2-4.4, and 4.8, the sequence $\{z^i_t\}$ generated from (3.8) and (3.11) satisfies $\sup_t \|z^i_t\| < \infty$ a.s., for any $i\in\mathcal{N}$.

Note that $\widetilde{\delta}^i_t\cdot\psi^i_t$ is bounded by Assumptions 2.2, 4.8, and Lemma B.4. Thus the actor step (3.12) can be viewed to track ODE $\dot{\theta}^i = 0$ when analyzing the faster time scale update. Thus by the same argument as in §B.2, we fix the value of θ_t as a constant θ. For notational convenience, let $v_t = [(v^1_t)^\top,\cdots,(v^N_t)^\top]^\top$, $\delta_t = [(\delta^1_t)^\top,\cdots,(\delta^N_t)^\top]^\top$, and $z_t = [(z^1_t)^\top,\cdots,(z^N_t)^\top]^\top$. By little abuse of notation, we here use $\{\mathcal{F}_{t,1}\}$ to denote the filtration with $\mathcal{F}_{t,1} = \sigma(r_\tau, z_\tau, s_\tau, \mathrm{C}_{\tau-1}, \tau\le t)$, an increasing σ-algebra. Then the updates of z_t in (3.8) and (3.11) have the following compact form

$$z_{t+1} = (\mathrm{C}_t\otimes\mathrm{I})\big(z_t + \beta_{v,t}\cdot y_{t+1}\big), \tag{B.22}$$

where $y_t = [(y^1_t)^\top,\cdots,(y^N_t)^\top]^\top \in \mathbb{R}^{(1+M+L)N}$. Here we denote $[r^i_{t+1}-\mu^i_t, (r^i_{t+1}-f_t^\top\lambda^i_t)f_t^\top, \delta^i_t\varphi_t^\top]^\top$ by y^i_{t+1}. Recall that $f_t = f(s_t,a_t)$ and $\varphi_t = \varphi(s_t)$. With the same definitions for $\langle\cdot\rangle$, $\mathcal{J}$, and $\mathcal{J}_\perp$, we can also separate the iteration of z_t as the sum of the consensus vector and the disagreement vector, i.e., $z_t = z_{\perp,t} + \mathbb{1}\otimes\langle z_t\rangle$, with $z_{\perp,t} = \mathcal{J}_\perp z_t$. Then the proof proceeds again in two steps as follows.

Step 1. We first establish that $\lim_t z_{\perp,t} = 0$ a.s. By Lemma B.4, it suffices to show that for any $M\in\mathbb{Z}^+$, $\lim_t z_{\perp,t}\mathbb{I}_{\{\sup_t\|z_t\|\le M\}} = 0$. We first establish the boundedness of $\mathbb{E}(\|\beta^{-1}_{v,t}z_{\perp,t}\|^2)$ over the set $\{\sup_t\|z_t\|\le M\}$, for any $M>0$.

Lemma B.5. Under Assumptions 4.2- 4.4, and 4.8, for any $M>0$, we have

$$\sup_t \mathbb{E}\big(\|\beta^{-1}_{v,t}z_{\perp,t}\|^2\mathbb{I}_{\{\sup_t\|z_t\|\le M\}}\big) < \infty.$$

Proof. Following the derivation of (B.9), we obtain the iteration of $z_{\perp,t}$ as

$$z_{\perp,t+1} = [(\mathrm{I}-\mathbb{1}\mathbb{1}^\top/N)\otimes\mathrm{I}](\mathrm{C}_t\otimes\mathrm{I})(z_{\perp,t}+\beta_{v,t}y_{t+1}) = [(\mathrm{I}-\mathbb{1}\mathbb{1}^\top/N)\mathrm{C}_t\otimes\mathrm{I}](z_{\perp,t}+\beta_{v,t}y_{t+1}). \tag{B.23}$$

Thus, similar to the derivation of (B.10), we obtain

$$\begin{aligned}\mathbb{E}\big(\|\beta^{-1}_{v,t+1}z_{\perp,t+1}\|^2\,|\,\mathcal{F}_{t,1}\big) \le \rho\cdot\frac{\beta^2_{v,t}}{\beta^2_{v,t+1}}\cdot\Big\{&\|\beta^{-1}_{v,t}z_{\perp,t}\|^2 + 2\|\beta^{-1}_{v,t}z_{\perp,t}\| \\ &\cdot\big[\mathbb{E}\big(\|y_{t+1}\|^2\,|\,\mathcal{F}_{t,1}\big)\big]^{\frac{1}{2}} + \mathbb{E}\big(\|y_{t+1}\|^2\,|\,\mathcal{F}_{t,1}\big)\Big\},\end{aligned} \tag{B.24}$$

where ρ represents the spectral norm of $\mathbb{E}[\mathrm{C}_t^\top(\mathrm{I}-\mathbb{1}\mathbb{1}^\top/N)\mathrm{C}_t]$ and $\rho\in[0,1)$. Then we have

$$\mathbb{E}\big(\|y_{t+1}\|^2\,|\,\mathcal{F}_{t,1}\big)=\mathbb{E}\bigg[\sum_{i\in\mathcal{N}}\big|r_{t+1}^i-\mu_t^i\big|^2+\big\|(r_{t+1}^i-f_t^\top\lambda_t^i)f_t\big\|^2+\big\|\delta_t^i\varphi_t\big\|^2\,\bigg|\,\mathcal{F}_{t,1}\bigg]. \tag{B.25}$$

By Assumption 4.8, we have $\mathbb{E}[\|\varphi_t(\varphi_{t+1}^\top-\varphi_t^\top)\|^2\,|\,\mathcal{F}_{t,1}]$ and $\mathbb{E}(\|\varphi_t\|^2\,|\,\mathcal{F}_{t,1})$ uniformly bounded for any $s_t\in\mathcal{S}$. Moreover, by Assumption 4.2, we have $\mathbb{E}(|r_{t+1}^i|^2\,|\,\mathcal{F}_{t,1})=\mathbb{E}(|r_{t+1}^i|^2\,|\,s_t,a_t)$ also uniformly bounded for any $s_t\in\mathcal{S},a_t\in\mathcal{A}$. Thus for any $M>0$, there exist $K_1,K_2<\infty$[1] such that (B.25) is further bounded as

$$\mathbb{E}\big(\|y_{t+1}\|^2\mathbb{I}_{\{\sup_{\tau\le t}\|z_\tau\|\le M\}}\,\big|\,\mathcal{F}_{t,1}\big)\le K_1\cdot\bigg[1+\sum_{i\in\mathcal{N}}\mathbb{E}\big(|r_{t+1}^i|^2\,\big|\,\mathcal{F}_{t,1}\big)\bigg]<K_2. \tag{B.26}$$

Let $\eta_t=\|\beta_{v,t}^{-1}z_{\perp,t}\|^2\mathbb{I}_{\{\sup_{\tau\le t}\|z_\tau\|\le M\}}$. By taking expectation on both sides of (B.24), we obtain

$$\mathbb{E}(\eta_{t+1})\le\frac{\beta_{v,t}^2}{\beta_{v,t+1}^2}\cdot\rho\cdot\big[\mathbb{E}(\eta_t)+2\sqrt{\mathbb{E}(\eta_t)}\cdot\sqrt{K_2}+K_2\big].$$

Following the same argument as in the proof of Lemma B.3, we obtain $\sup_t\mathbb{E}(\eta_t)<\infty$ and thus

$$\sup_t\mathbb{E}\big(\|\beta_{v,t}^{-1}z_{\perp,t}\|^2\mathbb{I}_{\{\sup_t\|z_t\|\le M\}}\big)<\infty,$$

which concludes the proof. □

Therefore, by Lemma B.5 and Assumption 4.3, we arrive at $\lim_t z_{\perp,t}\mathbb{I}_{\{\sup_t\|z_t\|\le M\}}=0$ a.s. for any $M>0$. By Lemma B.4, $\{\sup_t\|z_t\|<\infty\}$ holds with probability 1. This shows that $\lim_t z_{\perp,t}=0$ a.s., and thus concludes **Step 1**.

Step 2. We now proceed to show the convergence of the consensus vector $\mathbb{1}\otimes\langle z_t\rangle$. The iteration of $\langle z_t\rangle$ has the form

$$\langle z_{t+1}\rangle=\frac{1}{N}(\mathbb{1}^\top\otimes\mathrm{I})(\mathrm{C}_t\otimes\mathrm{I})(\mathbb{1}\otimes\langle z_t\rangle+z_{t,\perp}+\beta_{v,t}y_{t+1})=\langle z_t\rangle+\beta_{v,t}\langle(\mathrm{C}_t\otimes\mathrm{I})(y_{t+1}+\beta_{v,t}^{-1}z_{\perp,t})\rangle.$$

Hence, the update for $\langle z_t\rangle$ becomes

$$\langle z_{t+1}\rangle=\langle z_t\rangle+\beta_{v,t}\cdot\mathbb{E}(\langle y_{t+1}\rangle\,|\,\mathcal{F}_{t,1})+\beta_{v,t}\cdot\xi_{t+1}, \tag{B.27}$$

where ξ_{t+1} and $\langle y_{t+1}\rangle$ have the form

$$\begin{aligned}\xi_{t+1}&=\langle(\mathrm{C}_t\otimes\mathrm{I})(y_{t+1}+\beta_{\omega,t}^{-1}\omega_{\perp,t})\rangle-\mathbb{E}(\langle y_{t+1}\rangle\,|\,\mathcal{F}_{t,1}),\\ \langle y_{t+1}\rangle&=[\overline{r}_{t+1}-\langle\mu_t\rangle,(\overline{r}_{t+1}-f_t^\top\langle\lambda_t\rangle)f_t^\top,\langle\delta_t\rangle\varphi_t^\top]^\top,\end{aligned}$$

respectively. Recall that $\langle\delta_t\rangle=\overline{r}_{t+1}-\langle\mu_t\rangle+\varphi_{t+1}^\top\langle v_t\rangle-\varphi_t^\top\langle v_t\rangle$. Note that $\mathbb{E}(\langle y_{t+1}\rangle\,|\,\mathcal{F}_{t,1})$ is Lipschitz continuous in $\langle z_t\rangle=(\langle\mu_t\rangle,\langle\lambda_t\rangle^\top,\langle v_t\rangle^\top)^\top$, and thus the condition (a.1) in Assumption D.1 is satisfied. Moreover, one can verify that the term ξ_t is a martingale difference sequence. The conditional second moment of ξ_t can be bounded as

$$\mathbb{E}\big(\|\xi_{t+1}\|^2\,\big|\,\mathcal{F}_{t,1}\big)\le2\cdot\mathbb{E}\big(\big\|y_{t+1}+\beta_{v,t}^{-1}z_{\perp,t}\big\|_{\mathrm{G}_t}^2\,\big|\,\mathcal{F}_{t,1}\big)+2\cdot\big\|\mathbb{E}\big(\langle y_{t+1}\rangle\,\big|\,\mathcal{F}_{t,1}\big)\big\|^2, \tag{B.28}$$

where $\mathrm{G}_t=\mathrm{C}_t^\top\mathbb{1}\mathbb{1}^\top\mathrm{C}_t\otimes\mathrm{I}\cdot N^{-2}$ has bounded spectral norm. Thus the first term in (B.28) is bounded over the set $\{\sup_t\|z_t\|\le M\}$, for any $M>0$, i.e., there exist $K_3<\infty$ such that

$$\mathbb{E}\big(\big\|y_{t+1}+\beta_{v,t}^{-1}z_{\perp,t}\big\|_{\mathrm{G}_t}^2\,\big|\,\mathcal{F}_{t,1}\big)\cdot\mathbb{I}_{\{\sup_t\|z_t\|\le M\}}\le K_3\cdot\mathbb{E}\big(\big\|y_{t+1}\big\|^2+\big\|\beta_{v,t}^{-1}z_{\perp,t}\big\|^2\,\big|\,\mathcal{F}_{t,1}\big)\cdot\mathbb{I}_{\{\sup_t\|z_t\|\le M\}}.$$

[1] We note that K_1 and K_2 here are absolute constant values, with slight abuse of notation, we use the same notation as in the proof of Theorem 4.6. The same abuse applies to other constants with notation K_j, for any $j\in\mathbb{N}$.

From (B.26) and Lemma B.5, we obtain that the RHS can be further bounded by some $K_4 < \infty$. Moreover, the second term in (B.28) can be bounded by

$$\left\|\mathbb{E}\big(\langle y_{t+1}\rangle \,\big|\, \mathcal{F}_{t,1}\big)\right\|^2 \le \mathbb{E}\big(\left\|\langle y_{t+1}\rangle\right\|^2 \,\big|\, \mathcal{F}_{t,1}\big) \le K_5 \cdot \big(1 + \|\langle \mu_t\rangle\|^2 + \|\langle \lambda_t\rangle\|^2 + \|\langle v_t\rangle\|^2\big) = K_5 \cdot \big(1 + \|\langle z_t\rangle\|^2\big)$$

with some $K_5 < \infty$ due to the boundedness of r^i_{t+1}, f_t, and φ_t. Hence, for any $M > 0$, it follows that

$$\mathbb{E}\big(\|\xi_{t+1}\|^2 \,\big|\, \mathcal{F}_{t,1}\big) \le K_6 \cdot \big(1 + \|\langle z_t\rangle\|^2\big), \tag{B.29}$$

over the set $\{\sup_t \|z_t\| \le M\}$ for some $K_6 < \infty$. This verifies the condition (a.4) in Assumption D.1.

Then the ODE associated with (B.27) has the form

$$\dot{\langle z\rangle} = \begin{pmatrix} \dot{\langle \mu\rangle} \\ \dot{\langle \lambda\rangle} \\ \dot{\langle v\rangle} \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -\mathrm{F}^\top \mathrm{D}^{s,a}_\theta \mathrm{F} & 0 \\ -\Phi^\top \mathrm{D}^s_\theta \mathbb{1} & 0 & \Phi^\top \mathrm{D}^s_\theta (P^\theta - \mathrm{I})\Phi \end{pmatrix} \begin{pmatrix} \langle \mu\rangle \\ \langle \lambda\rangle \\ \langle v\rangle \end{pmatrix} + \begin{pmatrix} J(\theta) \\ \mathrm{F}^\top \mathrm{D}^{s,a}_\theta \overline{R} \\ \Phi^\top \mathrm{D}^s_\theta \overline{R}_\theta \end{pmatrix}. \tag{B.30}$$

Letting the RHS of the ODE (B.30) be $h(\langle z\rangle)$, we have $h(\langle z\rangle)$ Lipschitz continuous in $\langle z\rangle$. Similar to the proof in **Step 2** of §B.2, one can verify that the ODE has a unique globally asymptotically stable equilibrium $[J(\theta), \lambda_\theta^\top, v_\theta^\top]^\top$, by Assumption 4.8 on the feature matrices F and Φ. Here λ_θ and v_θ are the unique solutions to $\mathrm{F}^\top \mathrm{D}^{s,a}_\theta\big(\overline{R} - \mathrm{F}\lambda_\theta\big) = 0$ and $\Phi^\top \mathrm{D}^s_\theta\big[T^V_\theta(\Phi v_\theta) - \Phi v_\theta\big] = 0$, respectively. Recall the operator T^V_θ defined in (4.6). Moreover, the sequence $\{z_t\}$ is bounded almost surely by Assumption B.4. Hence all conditions for Theorem D.2 to hold are satisfied. We thus have $\lim_t \langle \mu_t\rangle = J(\theta)$, $\lim_t \langle \lambda_t\rangle = \lambda_\theta$, and $\lim_t \langle v_t\rangle = v_\theta$ over the set $\{\sup_t \|z_t\| \le M\}$ for any $M > 0$. By Lemma B.4 and the results from **Step 1**, we obtain that $\lim_t \mu^i_t = J(\theta)$, $\lim_t \lambda^i_t = \lambda_\theta$, and $\lim_t v^i_t = v_\theta$ for any $i \in \mathcal{N}$ a.s., which completes the proof. □

B.5. Proof of Theorem 4.10

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

$$\zeta^i_{t+1,1} = \widetilde{\delta}^i_t \cdot \psi^i_t - \mathbb{E}_{s_t\sim d_{\theta_t}, a_t\sim\pi_{\theta_t}}\big(\widetilde{\delta}^i_t \cdot \psi^i_t \,\big|\, \mathcal{F}_{t,2}\big), \qquad \zeta^i_{t+1,2} = \mathbb{E}_{s_t\sim d_{\theta_t}, a_t\sim\pi_{\theta_t}}\big[\big(\widetilde{\delta}^i_t - \widetilde{\delta}^i_{t,\theta_t}\big) \cdot \psi^i_t \,\big|\, \mathcal{F}_{t,2}\big],$$

where $\widetilde{\delta}^i_{t,\theta_t}$ is as defined in (4.9) with $\theta = \theta_t$. Then the actor update in (3.12) with a local projection becomes

$$\theta^i_{t+1} = \Gamma^i\Big[\theta^i_t + \beta_{\theta,t}\mathbb{E}_{s_t\sim d_{\theta_t}, a_t\sim\pi_{\theta_t}}\big(\widetilde{\delta}^i_t \cdot \psi^i_t \,\big|\, \mathcal{F}_{t,2}\big) + \beta_{\theta,t}\zeta^i_{t+1,1} + \beta_{\theta,t}\zeta^i_{t+1,2}\Big]. \tag{B.31}$$

Note that $\zeta^i_{t+1,2} = o(1)$ since the critic converges, i.e., $\widetilde{\delta}^i_t \to \widetilde{\delta}^i_{t,\theta_t}$, at the faster time scale. Moreover, letting $M^i_t = \sum_{\tau=0}^t \beta_{\theta,\tau}\zeta^i_{\tau+1,1}$, we have $\{M^i_t\}$ a martingale sequence. Note that the sequences $\{z^i_t\}$, $\{\psi^i_t\}$, and $\{\phi_t\}$ are all bounded, and so is the sequence $\{\zeta^i_{t,1}\}$. Hence, we have $\sum_t \mathbb{E}(\|M^i_{t+1} - M^i_t\|^2 \,|\, \mathcal{F}_{t,2}) < \infty$ a.s., and further obtain that the martingale sequence $\{M^i_t\}$ converges a.s. (Neveu, 1975, page 149)). Thus the condition (a.4) in Assumption D.4 is satisfied. (See Appendix §D.2 for details.) One can also verify that $\mathbb{E}_{s_t\sim d_{\theta_t}, a_t\sim\pi_{\theta_t}}(\widetilde{\delta}^i_t \cdot \psi^i_t \,|\, \mathcal{F}_{t,2})$ is continuous in θ^i_t, similar as the argument in §B.3. Therefore, we can apply the Kushner-Clark lemma (Theorem D.5 in Appendix §D.2) to show that the update in (B.31) converges a.s. to the set of asymptotically stable equilibria of the ODE (4.10), for each $i \in \mathcal{N}$, which concludes the proof. □

C. Proofs of the Stability of Consensus Updates

As mentioned before, the stability of the updates in stochastic approximation is usually proved separately. Here we provide the proof for the stability of the slower time scale update $\{\omega_t^i\}$ in Algorithm 1 and $\{z_t^i\}$ in Algorithm 2, i.e., Lemmas B.1 and B.4. In particular, we provide a sufficient condition for the stability of the consensus-based SA updates following the spirit of the results in Borkar & Meyn (2000) and Mathkar & Borkar (2016). We first state a main theorem for stability and then verify that Lemmas B.1 and B.4 are two special cases of it.

Let $\mathcal{N}$ be the set of agents with $|\mathcal{N}| = N$ and $x^i \in \mathbb{R}^d$ for any $i \in \mathcal{N}$. Consider the consensus update for $x_n^i \in \mathbb{R}^d$ as[2]

$$x_{n+1}^i = \sum_{j\in\mathcal{N}} c_n(i,j)\big\{x_n^j + \gamma_n\big[h^j(x_n, Y_n) + M_{n+1}^j\big]\big\}, \quad \text{for any } \ i \in \mathcal{N}, \tag{C.1}$$

where $\{Y_n\}_{n\geq 0}$ is an irreducible and aperiodic Markov chain over the finite set A. Let η denote the stationary distribution of $\{Y_n\}$ and $\overline{h}^i(x_n) = \mathbb{E}_{Y_n\sim\eta}[h^i(x_n, Y_n)]$ denote the expectation of $h^i(x_n, Y_n)$ over η. Let $x_n = [(x_n^1)^\top, \cdots, (x_n^N)^\top]^\top \in \mathbb{R}^{dN}$, $\mathrm{C}_n = [c_n(i,j)]_{N\times N}$, $h = [(h^1)^\top, \cdots, (h^N)^\top]^\top \in \mathbb{R}^{dN}$, $\overline{h}(x) = [(\overline{h}^1)^\top, \cdots, (\overline{h}^N)^\top]^\top \in \mathbb{R}^{dN}$, and $M_n = [(M_n^1)^\top, \cdots, (M_n^N)^\top]^\top \in \mathbb{R}^{dN}$. Let $\{\mathcal{F}_n\}$ be the filtration with $\mathcal{F}_n = \sigma(x_m, M_m, Y_m, \mathrm{C}_{m-1}, m \leq n)$.

Assumption C.1. We make the following assumptions:

(a.1) The consensus weight matrices $\{\mathrm{C}_n\}$ satisfy Assumption 4.4;

(a.2) $h^i : \mathbb{R}^n \times A \to \mathbb{R}^n$ is Lipschitz continuous in its first argument for any $i \in \mathcal{N}$;

(a.3) $\{M_n\}$ is a martingale difference sequence satisfying

$$\mathbb{E}\big(\|M_{n+1}\|^2 \,|\, \mathcal{F}_n\big) \leq K \cdot \big(1 + \|x_n\|^2\big),$$

for some $K > 0$;

(a.4) The difference $\zeta_{n+1} = \overline{h}(x_n) - h(x_n, Y_n)$ satisfies

$$\|\zeta_{n+1}\|^2 \leq K' \cdot \big(1 + \|x_n\|^2\big) \ \text{ a.s.},$$

for some $K' > 0$;

(a.5) The stepsize sequence $\{\gamma_n\}$ satisfies $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$;

(a.6) Define $h_c : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ as $h_c(x) = \overline{h}(cx) \cdot c^{-1}$ with some $c > 0$, and $\widetilde{h}_c(y) : \mathbb{R}^d \to \mathbb{R}^d$ be $\widetilde{h}_c(y) = \langle h_c(\mathbb{1} \otimes y)\rangle$. Then $\widetilde{h}_c(y) \to h_\infty(y)$ as $c \to \infty$ uniformly on compact sets for some $h_\infty(y) : \mathbb{R}^d \to \mathbb{R}^d$. Also, for some $\epsilon < N^{-1/2}$, $B^\epsilon = \{y \,|\, \|y\| < \epsilon\}$ contains a globally asymptotically stable attractor of the ODE $\dot{y} = h_\infty(y)$.

Note that the definitions of h_c and h_∞ in (a.6) in Assumption C.1 are different from those in Mathkar & Borkar (2016). Here we consider the averaged ODE in the consensus subspace for each agent, while that reference considers the overall ODE associated with (C.1), i.e., define $\widetilde{h} : \mathbb{R}^{dN} \to \mathbb{R}^{dN}$ as $\widetilde{h}(x) = \mathrm{C}_*\overline{h}(x)$ and let $h_\infty(x) = \lim_c \widetilde{h}(cx) \cdot c^{-1}$, where $\mathrm{C}_* = \lim_n \prod_{m=1}^n \mathrm{C}_m$. In fact, from Nedic & Ozdaglar (2009); Nedich et al. (2016), the limit C_* exists and has identical rows and rank one, provided the sequence $\{\mathrm{C}_t\}$ satisfies Assumption 4.4. Therefore, the globally asymptotical stability of the ODE $\dot{x} = h_\infty(x)$ (see Assumption (A5) in Mathkar & Borkar (2016)) does not hold for the linear ODE we consider in the convergence proof of the critic steps in both algorithms. In contrast, we can verify our condition (a.6) in Assumption C.1 later in the proof of Lemmas B.1 and B.4. We then have the following theorem on the stability of the sequence $\{x_n\}$.

Theorem C.2. Under Assumption C.1, the sequence $\{x_n\}$ generated from (C.1) is bounded almost surely, i.e., $\sup_n \|x_n^i\| < \infty$ a.s. for any $i \in \mathcal{N}$.

[2] To avoid possible confusion with the notation of continuous time t needed in the stability analysis, we use subscript n to denote the iteration index in the proof of Theorem C.2.

C.1. Proof of Theorem C.2

Let $\vartheta_c(y,t)$ denote the solution to the ODE

$$\dot{y} = \widetilde{h}_c(y),\ y(0) = y \tag{C.2}$$

We first have the following lemma, which is similar to Lemma 5 in Mathkar & Borkar (2016), and thus we leave out its proof here.

Lemma C.3. There exist constants $c_0 > 0$ and $T > 0$ such that for all initial conditions y within the sphere $\{y \,|\, \|y\| \leq N^{-1/2}\}$ and all $c \geq c_0$, we have $\|\vartheta_c(y,t)\| \leq (1-\epsilon') \cdot N^{-1/2}$ for $t \in [T, T+1]$, for some $0 < \epsilon' < 1$, where N is the number of agents.

Now, before stating the next set of lemmas, we introduce some notations and terminology. First, by the convention adopted in Borkar & Meyn (2000), we define $t_0 = 0$ and $t_n = \sum_{i=0}^{n} \gamma_i, n \geq 0$. Then we define $\overline{x}(t), t \geq 0$ as $\overline{x}(t_n) = x_n, n \geq 0$ with linear interpolation on each interval $[t_n, t_{n+1}]$. Moreover, we let $T_0 = 0$ and $T_{n+1} = \min\{t_m : t_m \geq T_n + T\}$ for any $n \geq 0$. Then $T_{n+1} \in [T_n + T, T_n + T + \sup_n \gamma_n]$. Let $m(n)$ be such that $T_n = t_{m(n)}$, for $n \geq 0$. Define the piecewise continuous trajectory $\hat{x}(t) = \overline{x}(t) \cdot r_n^{-1}$ for $t \in [T_n, T_{n+1})$, where $r_n = \max_{i \in \mathcal{N}}\{\|\overline{x}(T_n)\|, 1\}$. This implies that $\|\hat{x}(T_n)\| \leq 1$ for any $n \geq 0$. We also define $\hat{x}(T_{n+1}^-) = \overline{x}(T_{n+1}) \cdot r_n^{-1}$, $\hat{M}_{k+1} = M_{k+1} \cdot r_n^{-1}$, and $\hat{\zeta}_{k+1} = \zeta_{k+1} \cdot r_n^{-1}$ for $k \in [m(n), m(n+1))$.

Note that $\{\hat{M}_k\}$ is also a martingale difference sequence as $\{M_k\}$. We first establish boundedness of $\mathbb{E}[\|\hat{x}(t)\|^2]$ as follows.

Lemma C.4. Under Assumption C.1, $\sup_t \mathbb{E}[\|\hat{x}(t)\|^2] < \infty$.

Proof. It suffices to show that $\sup_{m(n) \leq k < m(n+1)} \mathbb{E}[\|\hat{x}(t_k)\|^2] < M$ for some $M > 0$ independent of n. We first write the update of $\hat{x}(t_k)$ for $k \in [m(n), m(n+1))$ in a compact form as

$$\hat{x}(t_{k+1}) = (\mathrm{C}_k \otimes \mathrm{I})\big(\hat{x}(t_k) + \gamma_k\big\{h_{r_n}[\hat{x}(t_k)] + \hat{M}_{k+1} + \hat{\zeta}_{k+1}\big\}\big). \tag{C.3}$$

Note that the additional term $\hat{\zeta}_{k+1}$ also satisfies $\mathbb{E}(\|\hat{\zeta}_{k+1}\|^2 \,|\, \mathcal{F}_k) \leq K' \cdot (1 + \|\hat{x}(t_k)\|^2)$ by condition (a.4) in Assumption C.1 since $r_n \geq 1$. Moreover, since $\mathrm{C}_k \otimes \mathrm{I}$ has bounded norm, it follows similarly as in the proof for Lemma 4 on page 25 in Borkar (2008) that

$$\mathbb{E}[\|\hat{x}(t_{k+1})\|^2]^{1/2} \leq \mathbb{E}[\|\hat{x}(t_k)\|^2]^{1/2}(1 + \gamma_k K_1) + \gamma_k K_2,$$

for some $K_1, K_2 > 0$ and $k \in [m(n), m(n+1))$. Then, by Grönwall inequality, we have the desired boundedness of $\mathbb{E}[\|\hat{x}(t)\|^2]$. □

By Lemma C.4, we immediately have the following result.

Lemma C.5 (Lemma 5 on page 25 in Borkar (2008))**.** The sequence $\{\sum_{k=0}^{n-1} \gamma_k \hat{M}_{k+1}\}$ converges almost surely.

We thus obtain the almost sure boundedness of the trajectory $\{\hat{x}(t)\}$.

Lemma C.6. Under Assumption C.1, $\sup_t \|\hat{x}(t)\| < \infty$ a.s.

Proof. Recall the update in (C.3) and note that $\|(\mathrm{C}_k \otimes \mathrm{I})x\|_\infty \leq \|x\|_\infty$ since C_k is a row stochastic matrix, where $\|\cdot\|_\infty$ denotes the infinite norm of a vector. Thus we have

$$\|\hat{x}(t_{k+1})\|_\infty \leq \|\hat{x}(t_k)\|_\infty + \gamma_k\big\|h_{r_n}[\hat{x}(t_k)] + \hat{M}_{k+1} + \hat{\zeta}_{k+1}\big\|_\infty. \tag{C.4}$$

By iterating (C.4), we obtain

$$\begin{aligned}\|\hat{x}(t_{k+1})\|_\infty &\leq \|\hat{x}(t_{m(n)})\|_\infty + \sum_{l=0}^{k-m(n)} \gamma_{m(n)+l}\big(\big\|h_{r_n}[\hat{x}(t_{m(n)+l})]\big\|_\infty + \big\|\hat{M}_{m(n)+l+1}\big\|_\infty + \big\|\hat{\zeta}_{m(n)+l+1}\big\|_\infty\big) \\ &\leq \|\hat{x}(t_{m(n)})\|_\infty + \sum_{l=0}^{k-m(n)} \gamma_{m(n)+l} \cdot K_3\big[1 + \big\|\hat{x}(t_{m(n)+l})\big\|_\infty\big] + \sum_{l=0}^{k-m(n)} \gamma_{m(n)+l}\big\|\hat{M}_{m(n)+l+1}\big\|_\infty, \end{aligned} \tag{C.5}$$

for some $K_3 > 0$. The second inequality is due to the Lipschitz continuity of h_{r_n} and condition (a.4) on ζ_{n+1}, by the equivalence of vector norms. Moreover, by Lemma C.5, the third term on the RHS of (C.5) is bounded a.s. since $\{\sum_{k=0}^{n-1} \gamma_k \hat{M}_{k+1}\}$ converges a.s. Recall that $\sum_{l=0}^{k-m(n)} \gamma_{m(n)+l} \leq T + \sup_n \gamma_n < \infty$ by definition of $m(n)$ and T_n. Thus, there exist $K_4, K_5 > 0$ such that

$$\|\hat{x}(t_{k+1})\|_\infty \leq K_4 + K_5 \sum_{l=0}^{k-m(n)} \gamma_{m(n)+l} \|\hat{x}(t_{m(n)+l})\|_\infty.$$

By discrete-time Grönwall inequality, we have

$$\sup_{m(n)\leq k<m(n+1)} \|\hat{x}(t_k)\|_\infty \leq K_4 \cdot \exp\big[K_5 \cdot \big(T + \sup_n \gamma_n\big)\big], \tag{C.6}$$

where the RHS of (C.6) is a (random) constant independent of n. Hence by equivalence of vector norms, we further obtain $\sup_t \|\hat{x}(t)\| < \infty$, which concludes the proof. □

The stability of $\|\hat{x}(t)\|$ is essential in showing the convergence of the consensus update in (C.3). For $n \geq 0$, let $y^n(t)$ denote the trajectory of $\dot{y} = \widetilde{h}_c(y)$ with $c = r_n$ and $y^n(T_n) = \langle \hat{x}(T_n)\rangle$, for $t \in [T_n, T_{n+1})$. Then we have the following lemma.

Lemma C.7. Under Assumption C.1, $\lim_n \sup_{t\in[T_n,T_{n+1})} \|\hat{x}(t) - \mathbb{1} \otimes y^n(t)\| = 0$.

Proof. Since $\hat{x}(t)$ is bounded a.s. on $[T_n, T_{n+1})$, we can mimic our proofs for Theorems 4.6 and 4.9 to show the convergence of $\hat{x}(t_k)$ for $k \in [T_n, T_{n+1})$ as $n \to \infty$. We will provide here only a sketch. One can first show that over the set $\{\sup_k \|\hat{x}(t_k)\| \leq M\}$ for any $M > 0$, $\lim_k \|\mathcal{J}_\perp \hat{x}(t_k)\| = 0$. The iteration of $\mathcal{J}_\perp \hat{x}(t_k)$ has the form (similar to (B.9))

$$\mathcal{J}_\perp \hat{x}(t_{k+1}) = [(\mathrm{I} - \mathbb{1}\mathbb{1}^\top/N)\mathrm{C}_k \otimes \mathrm{I}][\mathcal{J}_\perp \hat{x}(t_k) + \gamma_k y_{k+1}], \tag{C.7}$$

where $y_{k+1} = h_{r_n}[\hat{x}(t_k)] + \hat{M}_{k+1} + \hat{\zeta}_{k+1}$ here. One can easily verify that $\mathbb{E}\big(\|y_{k+1}\|^2 \mathbb{I}_{\{\sup_k \|\hat{x}(t_k)\|\leq M\}} \,|\, \mathcal{F}_k\big) < K_6$ for some $K_6 > 0$, due to Lipschitz continuity of h_{r_n} and the conditions (a.3) and (a.4) in Assumption C.1. Hence, by similar arguments as in the proof of Lemma B.3, we obtain $\lim_k \|\mathcal{J}_\perp \hat{x}(t_k)\| = 0$ almost surely, i.e., the vector $\hat{x}(t_k)$ reaches consensus as $k \to \infty$. Then we proceed to show the convergence of the sequence $\{\langle \hat{x}(t_k)\rangle\}$. Define $\hat{h}_c : \mathbb{R}^d \times A \to \mathbb{R}^d$ as $\hat{h}_c(y, Y_k) = \langle h(c \cdot \mathbb{1} \otimes y) \cdot c^{-1}\rangle$; then the iteration can be written as follows

$$\begin{aligned}\langle \hat{x}(t_{k+1})\rangle &= \langle \hat{x}(t_k)\rangle + \gamma_k \cdot \mathbb{E}(\langle y_{k+1}\rangle \,|\, \mathcal{F}_k) + \gamma_k \cdot \xi_{k+1} \\ &= \langle \hat{x}(t_k)\rangle + \gamma_k \cdot \hat{h}_{r_n}[\langle \hat{x}(t_k)\rangle, Y_k] + \gamma_k \cdot \xi_{k+1} + \gamma_k \cdot \beta_{k+1},\end{aligned}$$

where $\xi_{k+1} = \langle (\mathrm{C}_k \otimes \mathrm{I})(y_{k+1} + \gamma_k^{-1} \mathcal{J}_\perp \hat{x}(t_k))\rangle - \mathbb{E}(\langle y_{k+1}\rangle \,|\, \mathcal{F}_k)$, $\beta_{k+1} = \mathbb{E}(\langle y_{k+1}\rangle \,|\, \mathcal{F}_k) - \hat{h}_{r_n}[\langle \hat{x}(t_k)\rangle, Y_k]$, and $\langle y_{k+1}\rangle = \langle h_{r_n}[\hat{x}(t_k)]\rangle + \hat{M}_{k+1} + \langle \hat{\zeta}_{k+1}\rangle$. One can verify that $\{\xi_{k+1}\}$ is a martingale difference sequence satisfying $\mathbb{E}\big(\|\xi_{t+1}\|^2 \,\big|\, \mathcal{F}_{t,1}\big) < K_7 \cdot (1 + \|\langle \hat{x}(t_k)\rangle\|^2)$ for some $K_7 < \infty$ over the set $\{\sup_k \|\hat{x}(t_k)\| \leq M\}$. In addition, note that $\mathbb{E}(\hat{M}_{k+1} \,|\, \mathcal{F}_k) = 0$ and thus $\mathbb{E}(\langle y_{k+1}\rangle \,|\, \mathcal{F}_k) = \langle h[\hat{x}(t_k), Y_k]\rangle \cdot r_n^{-1}$. Thus we have $\|\beta_{k+1}\| \leq L \cdot \|\mathcal{J}_\perp \hat{x}(t_k)\| \cdot r_n^{-1}$ for some $L < \infty$ due to the Lipschitz continuity of h. Hence $\beta_k \to 0$ a.s. since $\|\mathcal{J}_\perp \hat{x}(t_k)\| \to 0$ a.s. and $r_n \geq 1$. Moreover, $\hat{h}_{r_n}[\langle \hat{x}(t_k)\rangle, Y_k]$ is Lipschitz continuous in $\langle \hat{x}(t_k)\rangle$. Therefore, by Theorem D.2, we obtain that $\langle \hat{x}(t_k)\rangle \to y^n(t)$ as $n \to \infty$, namely $k \to \infty$. Further we obtain that $\hat{x}^i(t_k) \to y^n(t)$ for any $i \in \mathcal{N}$, which concludes the proof following Theorem 2 in Chapter 2 of Borkar (2008). □

Now suppose that $\|x_n^i\| \to \infty$ for some $i \in \mathcal{N}$; then there exists a subsequence of $\{n_q\}$ such that $\|\overline{x}^i(T_{n_q})\| \to \infty$ and thus $\|\overline{x}(T_{n_q})\| \to \infty$. Hence $r_{n_q} \to \infty$. If $r_n > c_0 \geq 1$, then $\|\hat{x}(T_n)\| = 1$, and thus $\|y^n(T_n)\| = \|\langle \hat{x}(T_n)\rangle\| \leq N^{-1/2}$. By Lemma C.3, we have $\|\mathbb{1} \otimes y^n(T_n^-)\| = N^{1/2} \cdot \|y^n(T_n^-)\| \leq 1 - \epsilon'$. Thus by Lemma C.7, we have $\|\hat{x}(T_{n+1}^-)\| < 1 - \epsilon''$ for some $0 < \epsilon'' < \epsilon'$. Hence for $r_n > c_0$ and sufficiently large n,

$$\frac{\|\overline{x}(T_{n+1})\|}{\|\overline{x}(T_n)\|} = \frac{\|\hat{x}(T_{n+1}^-)\|}{\|\hat{x}(T_n)\|} < 1 - \epsilon''.$$

It thus follows that if $\|\overline{x}(T_n)\| > 1$, $\|\overline{x}(T_k)\|$ falls back to the unit ball at an exponential rate for $k \geq n$. The rest of the argument follows directly from the proof of Theorem 2 in Mathkar & Borkar (2016), which concludes the proof. □

Now we are ready to apply Theorem C.2 to prove Lemmas B.1 and B.4. We will return to the notations in §4.

C.2. Proof of Lemma B.1

The proof follows by verifying the conditions for Theorem C.2 to hold. Recall that the critic step in (3.3) has the form

$$\omega_{t+1} = (\mathrm{C}_t \otimes \mathrm{I})\big(\omega_t + \beta_{\omega,t} \cdot y_{t+1}\big),$$

with $y_{t+1} = (\delta_t^1 \phi_t^\top, \cdots, \delta_t^N \phi_t^\top)^\top \in \mathbb{R}^{KN}$. Thus the terms corresponding to (C.1) are

$$h^i(\omega_t^i, \mu_t^i, s_t, a_t) = \mathbb{E}(\delta_t^i \phi_t^\top \,|\, \mathcal{F}_{t,1}), \quad M_{t+1}^i = \delta_t^i \phi_t^\top - \mathbb{E}(\delta_t^i \phi_t^\top \,|\, \mathcal{F}_{t,1}). \tag{C.8}$$

Since the Markov chain $\{(s_t, a_t)\}_{t\geq 0}$ is irreducible and aperiodic given policy π_θ, we have $\overline{h}^i(\omega_t^i, \mu_t^i) = \Phi^\top \mathrm{D}_\theta^{s,a}[R^i - \mu_t^i \mathbb{1} + P^\theta \Phi \omega_t^i - \Phi \omega_t^i]$. By Lemma B.2, it is established that $\{\mu_t^i\}$ is bounded a.s. Hence, over the set $\{\sup_t \|\mu_t\| \leq M\}$ for any $M > 0$, there exists some $K' > 0$ such that $\|\overline{h}(\omega_t, \mu_t) - h(\omega_t, \mu_t, s_t, a_t)\|^2 \leq K' \cdot (1 + \|\omega_t\|^2)$, since the Markov chain is finite. This verifies the condition (a.4) in Assumption C.1. Moreover, since r_{t+1}^i and $\|\phi_t\|$ are uniformly bounded, $\mathbb{E}(\|M_{t+1}\|^2 \,|\, \mathcal{F}_{t,1}) \leq K \cdot (1 + \|\omega_t\|^2)$ is also verified for some $K > 0$. More importantly, over the set $\{\sup_t \|\mu_t\| \leq M\}$, $h_\infty(y)$ exists and has the form

$$h_\infty(y) = \lim_c \widetilde{h}_c(y) = \Phi^\top \mathrm{D}_\theta^{s,a}(P^\theta - \mathrm{I})\Phi y. \tag{C.9}$$

Clearly $\dot{y} = h_\infty(y)$ has origin as its globally asymptotically stable attractor (see the proof of Theorem 4.6 in §B.2). Hence we apply Theorem C.2 to conclude the proof. □

C.3. Proof of Lemma B.4

Recall that the critic step from (3.8) and (3.11) has the compact form

$$z_{t+1} = (\mathrm{C}_t \otimes \mathrm{I})\big(z_t + \beta_{v,t} \cdot y_{t+1}\big), \tag{C.10}$$

where $z_t^i = [\mu_t^i, (\lambda_t^i)^\top, (v_t^i)^\top]^\top$ and $y_t = [(y_t^1)^\top, \cdots, (y_t^N)^\top]^\top \in \mathbb{R}^{(1+M+L)N}$. Here y_{t+1}^i denotes $y_{t+1}^i = [r_{t+1}^i - \mu_t^i, (r_{t+1}^i - f_t^\top \lambda_t^i) f_t^\top, \delta_t^i \varphi_t^\top]^\top$. Thus the terms corresponding to (C.1) are

$$h^i(z_t^i, s_t, a_t) = \mathbb{E}(y_{t+1}^i \,|\, \mathcal{F}_{t,1}), \quad M_{t+1}^i = y_{t+1}^i - \mathbb{E}(y_{t+1}^i \,|\, \mathcal{F}_{t,1}). \tag{C.11}$$

Furthermore, we have

$$\overline{h}^i(z_t^i) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -\mathrm{F}^\top \mathrm{D}_\theta^{s,a} \mathrm{F} & 0 \\ -\Phi^\top \mathrm{D}_\theta^s \mathbb{1} & 0 & \Phi^\top \mathrm{D}_\theta^s (P^\theta - \mathrm{I}) \Phi \end{pmatrix} \begin{pmatrix} \mu_t^i \\ \lambda_t^i \\ v_t^i \end{pmatrix} + \begin{pmatrix} J^i(\theta) \\ \mathrm{F}^\top \mathrm{D}_\theta^{s,a} R^i \\ \Phi^\top \mathrm{D}_\theta^s R_\theta^i \end{pmatrix},$$

where $J^i(\theta) = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} d_\theta(s,a) \cdot R^i(s,a)$ and $R_\theta^i(s) = \sum_a \pi_\theta(s,a) R^i(s,a)$. Therefore, one can verify that both conditions (a.3) and (a.4) in Assumption C.1 are satisfied. In addition, $h_\infty(y)$ exists and has the form

$$h_\infty(y) = \lim_c \widetilde{h}_c(y) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -\mathrm{F}^\top \mathrm{D}_\theta^{s,a} \mathrm{F} & 0 \\ -\Phi^\top \mathrm{D}_\theta^s \mathbb{1} & 0 & \Phi^\top \mathrm{D}_\theta^s (P^\theta - \mathrm{I}) \Phi \end{pmatrix} \cdot y.$$

Clearly $\dot{y} = h_\infty(y)$ has origin as its globally asymptotically stable attractor (see the proof of Theorem 4.9 in §B.4), which completes the proof. □

D. Technical Background

D.1. A Basic Result of Stochastic Approximation

For the sake of completeness, we reproduce here a key result from Borkar (2008) that has been used repeatedly in our proofs. The results follow by specializing Corollary 8 and Theorem 9 on page 74-75 in Borkar (2008). We note that this is actually an extension of Theorem 2.1 and Theorem 2.2 in Borkar & Meyn (2000) to the case with irreducible Markovian state and diminishing noise in the update. More general conclusions can also be found in Benaïm (1999); Kushner & Yin (2003).

Consider the n-dimensional stochastic approximation iteration

$$x_{t+1} = x_t + \gamma_t[h(x_t, Y_t) + M_{t+1} + \beta_{t+1}],\ t \geq 0, \tag{D.1}$$

where $\gamma_t > 0$ and $\{Y_t\}_{t\geq 0}$ is a Markov chain on a finite set A.

Assumption D.1. We make the following assumptions:

(a.1) $h : \mathbb{R}^n \times A \to \mathbb{R}^n$ is Lipschitz continuous in its first argument;

(a.2) $\{Y_t\}_{t\geq 0}$ is an irreducible Markov chain with stationary distribution π;

(a.3) The stepsize sequence $\{\gamma_t\}$ satisfies $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$;

(a.4) $\{M_t\}$ is a martingale difference sequence, i.e., $\mathbb{E}[M_{t+1} \,|\, x_\tau, M_\tau, Y_\tau, \tau \leq t] = 0$, satisfying that for some $K > 0$ and $t \geq 0$

$$\mathbb{E}\big(\|M_{t+1}\|^2 \,|\, x_\tau, M_\tau, Y_\tau, \tau \leq t\big) \leq K \cdot \big(1 + \|x_t\|^2\big).$$

(a.5) The sequence $\{\beta_t\}$ is a bounded random sequence with $\beta_t \to 0$ almost surely as $t \to \infty$.

Then the asymptotic behavior of the iteration (D.1) is related to the behavior of the solution to the ODE

$$\dot{x} = \overline{h}(x) = \sum_i \pi(i) h(x, i). \tag{D.2}$$

Suppose (D.2) has a unique globally asymptotically stable equilibrium x^*, we then have the following two theorems.

Theorem D.2. Under Assumption D.1, if $\sup_t \|x_t\| < \infty$ a.s., we have $x_t \to x^*$.

Theorem D.3. Under Assumption D.1, suppose that

$$\lim_{c\to\infty} \frac{\overline{h}(cx)}{c} = h_\infty(x)$$

exists uniformly on compact sets for some $h_\infty \in C(\mathbb{R}^n)$. If the ODE $\dot{y} = h_\infty(y)$ has origin as the unique globally asymptotically stable equilibrium, then

$$\sup_t \|x_t\| < \infty \ \text{ a.s.}$$

D.2. Kushner-Clark Lemma

We state here the well-known Kushner-Clark Lemma (Kushner & Clark, 1978; Metivier & Priouret, 1984; Prasad et al., 2014) in the sequel.

Let Γ be an operator that projects a vector onto a compact set $\mathcal{X} \subseteq \mathbb{R}^N$. Define a vector $\hat{\Gamma}(\cdot)$ as

$$\hat{\Gamma}[h(x)] = \lim_{0<\eta\to 0} \left\{ \frac{\Gamma[x + \eta h(x)] - x}{\eta} \right\},$$

for any $x \in \mathcal{X}$ and with $h : \mathcal{X} \to \mathbb{R}^N$ continuous. Consider the following recursion in N dimensions

$$x_{t+1} = \Gamma\big\{x_t + \gamma_t[h(x_t) + \xi_t + \beta_t]\big\}. \tag{D.3}$$

The ODE associated with (D.3) is given by

$$\dot{x} = \hat{\Gamma}[h(x)]. \tag{D.4}$$

Assumption D.4. We make the following assumptions:

(a.1) $h(\cdot)$ is a continuous $\mathbb{R}^N$-valued function.
(a.2) The sequence $\{\beta_t\}$, $t \geq 0$ is a bounded random sequence with $\beta_t \to 0$ almost surely as $t \to \infty$.
(a.3) The stepsizes γ_t, $t \geq 0$ satisfy $\gamma_t \to 0$ as $t \to \infty$ and $\sum_t \gamma_t = \infty$.
(a.4) The sequence ξ_t, $t \geq 0$ satisfies for any $\epsilon > 0$

$$\lim_t \mathbb{P}\left(\sup_{n \geq t} \left\| \sum_{\tau=t}^{n} \gamma_\tau \xi_\tau \right\| \geq \epsilon \right) = 0.$$

Then the Kushner-Clark Lemma says the following.

Theorem D.5. Under Assumption D.4, suppose that the ODE (D.4) has a compact set $\mathcal{K}^*$ as its set of asymptotically stable equilibria. Then x_t in (D.3) converges almost surely to $\mathcal{K}^*$ as $t \to \infty$.

E. Experiment Details

In this section, we provide the details of the simulation settings in §5.

E.1. Linear Function Approximation

We first consider the setting with linear function approximation to corroborate our theoretical results. Consider in total $N = 20$ agents, each has a binary-valued action space, i.e., $\mathcal{A}^i = \{0, 1\}$, for all $i \in \mathcal{N}$. Thus the cardinality of the set of actions $\mathcal{A}$ is 2^{20}. In addition, there are in total $|\mathcal{S}| = 20$ states. The following ways of selecting the model and algorithm parameters, including transition probabilities, rewards, and features, follow from those in Dann et al. (2014). The elements in the transition probability matrix P are uniformly sampled from the interval $[0, 1]$ and normalized to be stochastic. We also add a small constant 10^{-5} onto each element in the matrix to ensure ergodicity of the MDP such that Assumption 2.2 is satisfied. For each agent i and each state-action pair (s, a), the mean reward $R^i(s, a)$ is sampled uniformly from $[0, 4]$, which varies among agents. The instantaneous rewards r_t^i are sampled from the uniform distribution $[R^i(s, a) - 0.5, R^i(s, a) + 0.5]$. The policy $\pi^i_{\theta^i}(s, a^i)$ is parametrized following the Boltzman policies, i.e.,

$$\pi^i_{\theta^i}(s, a^i) = \frac{\exp\left(q^\top_{s,a^i}\theta^i\right)}{\sum\limits_{b^i \in \mathcal{A}^i} \exp\left(q^\top_{s,b^i}\theta^i\right)}$$

where $q_{s,b^i} \in \mathbb{R}^{m_i}$ is the feature vector with the same dimension as θ^i, for any $s \in \mathcal{S}$ and $i \in \mathcal{N}$. Here we set $m_1 = m_2 = \cdots = m_N = 5$. The elements of q_{s,b^i} are also uniformly sampled from $[0, 1]$. In particular, the gradient of the score function thus has the form

$$\nabla_{\theta^i} \log \pi^i_{\theta^i}(s, a^i) = q_{s,a^i} - \sum_{b^i \in \mathcal{A}^i} \pi^i_{\theta^i}(s, a^i) q_{s,b^i}.$$

The feature vectors $\phi \in \mathbb{R}^K$ for the action-value function $Q(\cdot, \cdot; \omega)$ in Algorithm 1, $\varphi \in \mathbb{R}^L$ for the state-value function $V(\cdot; v)$ and $f \in \mathbb{R}^M$ for the globally averaged reward function $\overline{R}(\cdot, \cdot; \lambda)$ in Algorithm 2, are all uniformly sampled from $[0, 1]$, of dimensions $K = 10 \ll |\mathcal{S}| \cdot |\mathcal{A}|$, $L = 5 < |\mathcal{S}|$, and $M = 10 \ll |\mathcal{S}| \cdot |\mathcal{A}|$. Moreover, the selected feature matrices Φ, $\mathit{\Phi}$, and F are all ensured to have full column rank as required in Assumptions 4.5 and 4.8.

The consensus weight matrix C_t are chosen independent and identically distributed along time t, by normalizing the absolute Laplacian matrix of a connected random graph $\mathcal{G}_t$ over agents $\mathcal{N}$ to be doubly stochastic[3]. The graph $\mathcal{G}_t$ is generated by randomly placing communication links among agents such that the connectivity ratio[4] is $4/N$. The stepsizes are selected as $\beta_{\omega,t} = \beta_{v,t} = 1/t^{0.65}$ and $\beta_{\theta,t} = 1/t^{0.85}$, which satisfy Assumption 4.3.

The performances of the fully decentralized algorithms are compared with those of the centralized algorithms in which the rewards r_t^i of all agents are available at a centralized controller and the global policy π_θ is also updated there. These centralized version algorithms thus reduce to single-agent AC algorithms with linear function approximation. We refer the two centralized AC algorithms for comparison as Central-1 and Central-2, respectively. The algorithm Central-1 has the following critic step, which is based on action-value function approximation as in Algorithm 1

$$\begin{cases} \mu_{t+1} = (1 - \beta_{\omega,t}) \cdot \mu_t + \beta_{\omega,t} \cdot \overline{r}_{t+1}, \\ \delta_t = \overline{r}_{t+1} - \mu_t + Q_{t+1}(\omega_t) - Q_t(\omega_t) \\ \omega_{t+1} = \omega_t + \beta_{\omega,t} \cdot \delta_t \cdot \nabla_\omega Q_t(\omega_t). \end{cases} \tag{E.1}$$

Recall that $Q_t(\omega) = Q(s_t, a_t; \omega)$ with $Q(\cdot, \cdot; \omega)$ the estimate for the global action-value function Q_θ, $\beta_{\omega,t} > 0$ is the stepsize, and $\overline{r}_t = \sum_{i \in \mathcal{N}} r_t^i \cdot N^{-1}$. Accordingly, the central controller improves the policy for each agent i, which results in the actor step

$$\theta^i_{t+1} = \theta^i_t + \beta_{\theta,t} \cdot A_t \cdot \psi^i_t, \quad \forall i \in \mathcal{N}, \tag{E.2}$$

[3] A stochastic matrix P is doubly stochastic if it is both row and column stochastic.

[4] Note that the connectivity ratio is defined as the ratio between the total degree of the graph and the degree of the complete graph, i.e., $2E/[N(N-1)]$, where E is the number of edges.

where $\beta_{\theta,t} > 0$ is the stepsize, A_t and ψ_t^i are defined as

$$A_t = Q_t(\omega_t) - \sum_{a\in\mathcal{A}} \pi_{\theta_t}(s_t, a) \cdot Q(s_t, a; \omega_t), \quad \psi_t^i = \nabla_{\theta^i} \log \pi_{\theta_t^i}^i(s_t, a_t^i). \tag{E.3}$$

The algorithm Central-2 follows the updates of Algorithm 1 in Bhatnagar et al. (2009), based on state-value approximation as in Algorithm 2 here. In particular, it has the following critic step

$$\begin{cases} \mu_{t+1} = (1 - \beta_{v,t}) \cdot \mu_t + \beta_{v,t} \cdot \overline{r}_{t+1}, \\ \delta_t = \overline{r}_{t+1} - \mu_t + V_{t+1}(v_t) - V_t(v_t), \\ v_{t+1} = v_t + \beta_{v,t} \cdot \delta_t \cdot \nabla_v V_t(v_t), \end{cases} \tag{E.4}$$

where we recall that $V_t(v) = V(s_t; v)$ for any $v \in \mathbb{R}^L$ and $\beta_{v,t} > 0$ is the stepsize satisfying Assumption 4.3. Since the rewards of all agents $\{r_t^i\}_{i\in\mathcal{N}}$ are available to the controller, no estimation for the globally averaged

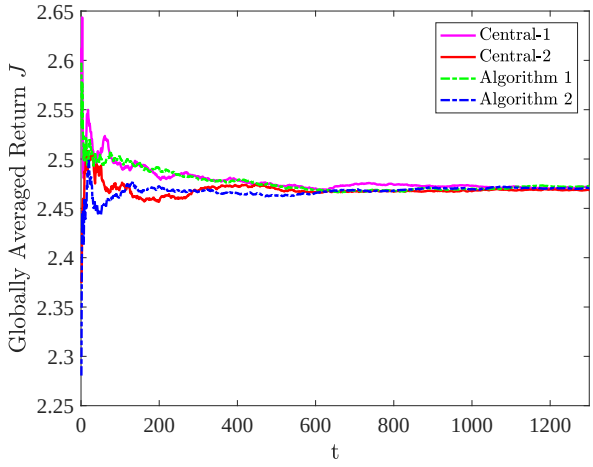
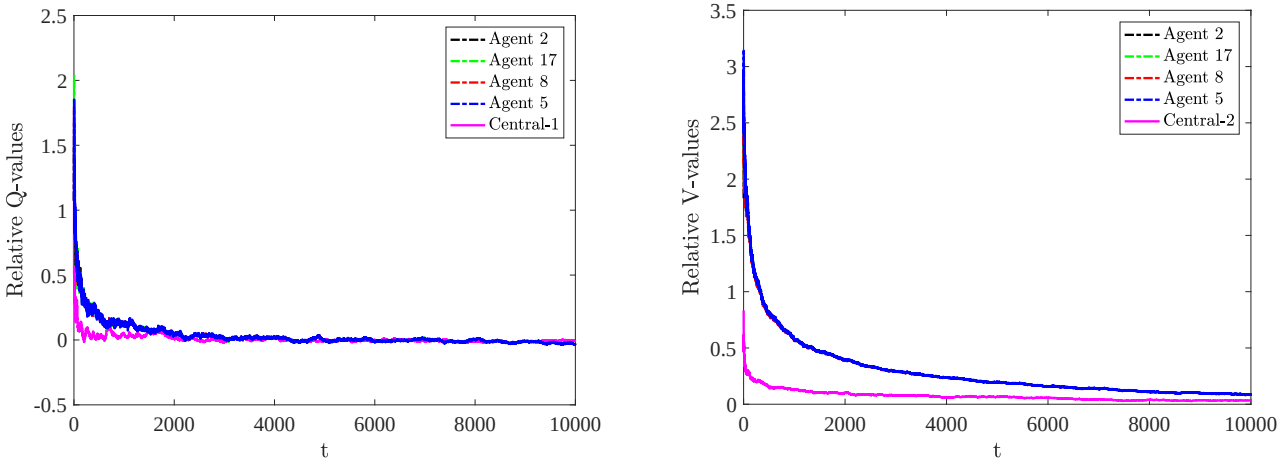


Figure 3. The convergence of globally averaged returns, when linear function approximation is used. We plot the returns achieved by both Algorithm 1 and Algorithm 2, along with their centralized counterparts Central-1 and Central-2.



(a) Algorithm 1 v.s. Central-1 (b) Algorithm 2 v.s. Central-2

Figure 4. The convergence of relative value functions at four randomly selected agents, when linear function approximation is used. We randomly select the agents 2, 5, 8, and 17. In (a), we plot the convergence curve of the relative action-value at a randomly selected state-action pair, obtained from Central-1 and Algorithm 1. In (b), we plot the convergence curve of the relative state-value at a randomly selected state, obtained from Central-2 and Algorithm 2.

reward function $\overline{R}$ is needed in the update. Thus, the global state-value TD-error δ_t can be computed immediately and used in the actor step as

$$\theta_{t+1}^i = \theta_t^i + \beta_{\theta,t} \cdot \delta_t \cdot \psi_t^i, \quad \forall i \in \mathcal{N}, \tag{E.5}$$

where ψ_t^i is as defined in (E.3).

Convergence of the globally averaged returns, the relative action-value function, and the relative state-value function are reported in Figure 3 and Figure 4, respectively. Figure 3 shows that both decentralized algorithms converge to the globally averaged return as achieved by the two centralized counterparts. Moreover, Figure 4 illustrates that for each agent, the approximation of global value functions in Algorithms 1 and 2 reach consensus much faster than the AC algorithm itself converges. In addition, the convergence of the value functions is relatively slower for the decentralized algorithms than for the centralized ones, possibly due to the delay of information diffusion across the network.

E.2. Nonlinear Function Approximation

We also evaluate the performance of Algorithm 1 and Algorithm 2 when nonlinear function approximators, for example, neural networks, are adopted. Although it seems difficult to establish convergence guarantees in this case, we believe that the empirical results are of independent interest, which justify the effectiveness of the proposed fully decentralized algorithms in a more sophisticated environment.

To this end, we consider the simulation environment of the *Cooperative Navigation* task in Lowe et al. (2017). In this environment, agents need to reach a set of L landmarks through physical movement. Agents are able to observe the position of the landmarks and other agents, and are rewarded based on the proximity of any agent to each landmark (Lowe et al., 2017). To fit in our networked MDP model, we modify the environment there in the following aspects. First, we assume the state is globally observable, i.e., the position of the landmarks and other agents are observable to each agent. Moreover, each agent has a certain target landmark to cover, and the individual reward is determined by the proximity to that certain landmark, as well as the penalty from collision with other agents. In this way, the reward function varies between agents. The reward is further scaled by different positive coefficients, representing the different priority/preferences of different agents. In addition, agents are connected via a time-varying communication network with several other agents nearby. The collaborative goal of the agents is then to maximize the network-wide averaged long-term return. The illustration of the modified Cooperative Navigation environment is provided in Figure 5.

Specifically, we consider $N = 10$ agents moving in a rectangular region of size 2×2. Each agent has a single target landmark, i.e., $\mathrm{L} = N = 10$, which is randomly located in the region. The action set for each agent is the

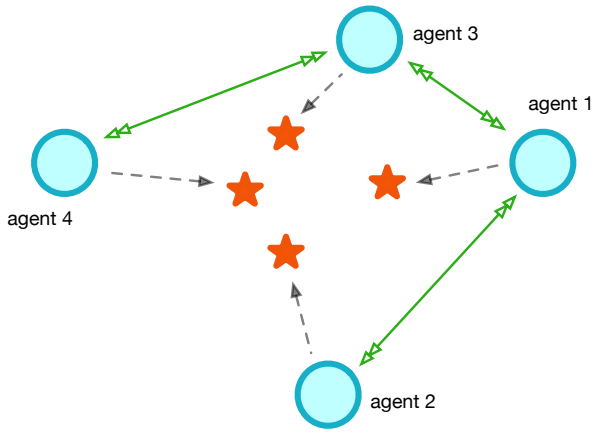


Figure 5. Illustration of the experimental environment for the Cooperative Navigation task we consider, modified from Lowe et al. (2017). In particular, the blue circles represent the agents, the orange stars represent the landmarks, the green arrows represent the communication links between agents, and the gray arrows show the target landmark each agent need to cover.

movement set $\{left, right, up, down, stay\}$, and thus $|\mathcal{A}_i| = 5$ for any $i \in \mathcal{N}$. The state s includes the position of the landmarks and other agents, which thus has a dimension of $2(N + \mathrm{L}) = 40$. The reward of agent i is the negative number of the distance to the target landmark, plus -1 if agent i collides with any other agents. The coefficients that scale the reward of each agent are selected randomly from a uniform distribution over $[0, 2]$. Each agent maintains two neural networks for actor and critic, respectively. Both neural networks have one hidden layer containing 24 neural units, which all use ReLU as the activation function. The output layer for the actor network is softmax, and that for the critic network is linear.

The time-varying network $\mathcal{G}_t$ and consensus matrix C_t are constructed in the same way as in §E.2. The stepsizes for the actor and critic step are set as constants 0.001 and 0.01, respectively. For each episode, the algorithms terminate either all agents reach the target landmarks or after 1000 iterations, and we run in total 200 episodes in each test run. We report the globally averaged return from 10 test runs in Figure 2, where the algorithms Central-1 and Central-2 follow the update rules in §E.1, but with nonlinear function approximation.

To better illustrate the necessity of cooperation via communication among agents, we compare the results of Algorithms 1 and 2 with the *non-cooperative* counterparts, where each agent performs single-agent RL, with no awareness of the existence of other agents in the environment. These non-cooperative algorithms are equivalent to our proposed consensus-based algorithms, when the communication network is disconnected and contains only a self-loop at each node. Thus, we demonstrate the performance of the non-cooperative counterparts of Algorithms 1 and 2 in Figure 6. Note that the results of Algorithms 1 and 2 follow from those in Figure 2. It is shown that the non-cooperative algorithms are unstable, and achieve much worse long-term return with much larger variance than the cooperative counterparts. These observations showcase the necessity of communication in fully decentralized cooperative MARL.

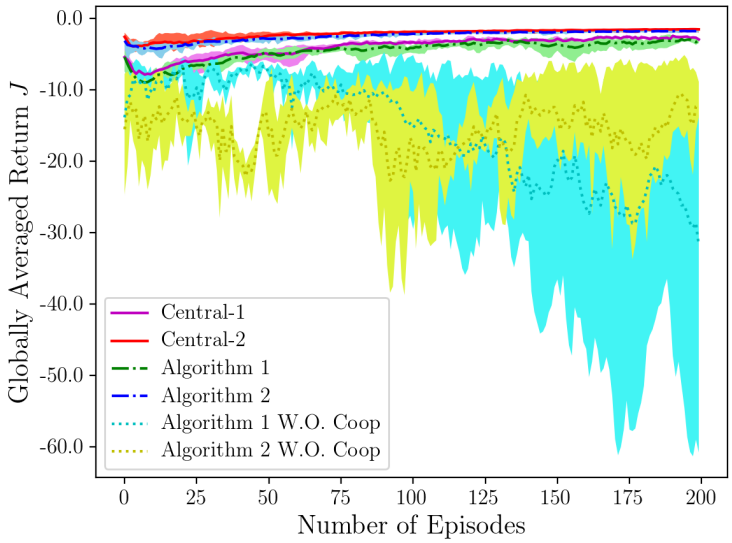


Figure 6. The globally averaged returns for Cooperative Navigation, when neural networks are used for function approximation.

F. Comparison with Existing Work on Multi-Agent Systems and MARL

In this section, we compare both our *model* and *algorithms* with related work on multi-agent systems and collaborative MARL in details.

Our framework of networked multi-agent systems finds a broad range of applications in distributed cooperative control problems, including formation control of unmanned vehicles (Fax & Murray, 2004), cooperative navigation of robots (Corke et al., 2005), load management in energy networks (Dall'Anese et al., 2013), and flocking of mobile sensor networks (Cortes et al., 2004), etc. Previously, the collective goal of the multi-agent system is to either reach a stable and consensus state for all agents (Fax & Murray, 2004; Corke et al., 2005), or solve a static optimization problem in a distributed fashion (Dall'Anese et al., 2013; Nedic & Ozdaglar, 2009). In the first line of work, including formation control and consensus problems, the objective is not formulated explicitly as an optimization problem, and most of the work focuses on continuous-time dynamic systems. Whereas in the second line of work, the problem is approached in a static setting, in the sense that the optimization objective is deterministic and there is no control input affecting the transition of the system, see recent efforts in Nedic & Ozdaglar (2009); Agarwal & Duchi (2011); Chen & Sayed (2012). In contrast, we here model the interaction of multiple agents and evolution of the system as an MDP, a dynamic setting, and explicitly use the network-wide long-term return as the collaborative goal of all agents. In this regard, our framework is pertinent to the cooperative/distributed optimal control problems (Lewis et al., 2013; Movric & Lewis, 2014), but focuses on the discrete-time setting and falls into the realm of reinforcement learning, where the model of the system may be unknown. One recent work (Macua et al., 2017) for multi-task RL, which is almost concurrent to ours, is also based on the model with networked agents. Nonetheless, the MDP problem solved by different agents are totally decoupled, which excludes the work from the realm of MARL with interactive agents as we consider here.

Our framework also departs from the existing framework on collaborative MARL models in the following aspects. In contrast to the canonical multi-agent MDP (MMDP) model proposed in Boutilier (1996); Lauer & Riedmiller (2000), our model allows the agents to exchange information over a communication network with possibly sparse connectivity at each agent. This improves the *scalability* of the multi-agent model with a high population of agents, which is one of the long-standing challenges in general MARL problems (Shoham et al., 2003). Moreover, we allow heterogeneous agents to have various individual reward functions, while the canonical MMDP assumes a common reward function for all agents. The latter setting greatly simplifies the problem since no information exchange among agents is necessary to approximate the value function for each agent. Our model not only fits in the *multi-task* setting which has gained increasing popularity in MARL (Omidshafiei et al., 2017; Teh et al., 2017), but also applies to the general multi-agent RL setting. One of the few models that also consider heterogeneous reward functions in collaborative MARL is Kar et al. (2013), where the global action is assumed to be actuated by a remote controller, but in our case, the agents are fully decentralized and have local control capabilities. Besides, the models in Guestrin et al. (2002); Kok & Vlassis (2006) also consider heterogeneous rewards, but with a strong assumption that the global Q-function can be factorized as several local Q-functions that depend on the actions of only a subset of agents, which simplified the general setting we considered where the Q-function is affected by the joint action of all agents. It is also worth noting that our model generalizes the team Markov game model for collaborative MARL, see Littman (2001); Wang & Sandholm (2003); Arslan & Yüksel (2017), where all agents have individual action sets but share a common payoff function as in the canonical MMDP.

Moreover, our algorithms designed for networked MMDP are distinct from the existing collaborative MARL algorithms in the following aspects. First, our MARL algorithms belong to the type of actor-critic algorithms, whereas several of the existing MARL algorithms are designed based on Q-learning type (critic-based) algorithms only (Boutilier, 1996; Lauer & Riedmiller, 2000; Guestrin et al., 2002; Kok & Vlassis, 2006; Kar et al., 2013). Moreover, these algorithms assume either the rewards are common to all agents (Boutilier, 1996; Lauer & Riedmiller, 2000), or there exists a remote central controller to take actions for the agents (Kar et al., 2013). The rest of them (Guestrin et al., 2002; Kok & Vlassis, 2006) utilize the factorized structure of the Q-function as mentioned above, which fail to handle our general MARL setting. More recently, some actor-critic type MARL algorithms with *distributed/decentralized* structures have gained increasing attention (Gupta et al., 2017; Lowe et al., 2017; Omidshafiei et al., 2017). They are developed for more complicated settings where both cooperation and competition may appear among agents. However, they all rely on a central controller to perform

the critic step, which are closer to the *hierarchical* structure rather than the *fully decentralized* structure we consider here. Second, our algorithms apply function approximation to handle the setting with massively large state and action spaces, while enjoying theoretical guarantees for convergence as we show in §4. However, the existing collaborative MARL algorithms are either guaranteed to converge only for tabular cases (Hu & Wellman, 2003; Wang & Sandholm, 2003; Kar et al., 2013; Prasad et al., 2014), or only have empirical convergence when function approximation is applied (Foerster et al., 2016; Gupta et al., 2017; Lowe et al., 2017; Omidshafiei et al., 2017). The recent work on multi-task RL with networked agents (Macua et al., 2017) also focuses on empirical results only, with no complete convergence analysis.