SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Data-driven containment control of discrete-time multi-agent systems via value iteration

Zhinan PENG[1], Jiangping HU[1,2*] & Bijoy Kumar GHOSH[1,2]

[1]*School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu* 611731*, China;*
[2]*Department of Mathematics and Statistics, Texas Tech University, Lubbock* 79409*, USA*

Appendix A The proof of Theorems 1 and 2

Before giving the proofs of Theorems 1 and 2, we needs the following two important lemmas.

Lemma 1. Let $\mu_i^l(k)$ be an arbitrary stabilizing policy, and its associated iterative performance index function be $\Lambda_i^l(\varepsilon_i(k))$ which is defined by $\Lambda_i^{l+1}(\varepsilon_i(k)) = c_i(\varepsilon_i(k), \mu_i^l(k)) + \alpha\Lambda_i^l(\varepsilon_i(k+1))$. Define the control policy $u_i^l(k)$ generated by Algorithm 1 with an associated performance function $J_i^l(k)$. Starting with $0 \leqslant J_i^0(\varepsilon_i(k)) \leqslant \Lambda_i^0(\varepsilon_i(k))$, one has

$$0 \leqslant J_i^l(k) \leqslant \Lambda_i^l(\varepsilon_i(k)). \tag{A1}$$

Proof. For each agent i, the arbitrary stabilizing control policy $\mu_i^l(k)$ can be written as $\mu_i^l(k) = (u_i^l(k) + (\mu_i^l(k) - u_i^l(k)))$ and its iterative performance index function can be given by

$$\begin{aligned}
\Lambda_i^{l+1}(\varepsilon_i(k)) =& Y_i^l(\Lambda_i^l, \mu_i^l)\\
=& c_i(\varepsilon_i(k), \mu_i^l(k)) + \alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l)\\
=& \varepsilon_i^\top(k)Q_{ii}\varepsilon_i(k) + \mu_i^l(k)^\top R_{ii}\mu_i^l(k) + \sum_{j\in\mathcal{N}(i)} u_j^\top(k)S_{ij}u_j(k) + \alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l)\\
=& \varepsilon_i^\top(k)Q_{ii}\varepsilon_i(k) + \big[u_i^l(k) + (\mu_i^l(k) - u_i^l(k))\big]^\top R_{ii}\big[u_i^l(k)\\
&+ (\mu_i^l(k) - u_i^l(k))\big] + \alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l) + \sum_{j\in\mathcal{N}(i)} u_j^\top(k)S_{ij}u_j(k)\\
=& Y_i^l(\Lambda_i^l, u_i^l) + (\mu_i^l(k) - u_i^l(k))^\top R_{ii}(\mu_i^l(k) - u_i^l(k)) +\\
&\alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l) + 2{\mu_i^l}^\top(k)R_{ii}u_i^l(k) - \big[\alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l) + 2{u_i^l}^\top(k)R_{ii}u_i^l(k)\big],
\end{aligned} \tag{A2}$$

where

$$Y_i^l(\Lambda_i^l, u_i^l) = \varepsilon_i^\top(k)Q_{ii}\varepsilon_i(k) + u_i^l(k)^\top R_{ii}u_i^l(k) + \sum_{j\in\mathcal{N}(i)} u_j^\top(k)S_{ij}u_j(k) + \alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l).$$

Furthermore, it is not difficult to have

$$\begin{aligned}
&(\mu_i^l(k) - u_i^l(k))^\top R_{ii}(\mu_i^l(k) - u_i^l(k)) + \alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l) + 2{\mu_i^l}^\top(k)R_{ii}u_i^l(k)\\
&- \big[\alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l) + 2{u_i^l}^\top(k)R_{ii}u_i^l(k)\big]\\
=& {\mu_i^l}^\top(k)R_{ii}\mu_i^l(k) + \alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l) - \big[{u_i^l}^\top(k)R_{ii}u_i^l(k) + \alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l)\big] \geqslant 0.
\end{aligned} \tag{A3}$$

In fact, since the iterative controller $u_i^l(k)$ for each agent i at each iteration step l in Algorithm 1 is given by

$$u_i^l(k) = \arg\min_{u_i(k)} \{c_i(\varepsilon_i(k), u_i(k)) + \alpha\Lambda_i^l(\varepsilon_i(k+1))\},$$

so this controller makes the associated performance index function $\Lambda_i^{l+1}(\varepsilon_i(k)|u_i^l)$ be minimized at each iteration. However, the iterative controller $\mu_i^l(k)$ is arbitrary for the performance index function $\Lambda_i^{l+1}(\varepsilon_i(k)|\mu_i^l)$, thus we have $\Lambda_i^{l+1}(\varepsilon_i(k)|u_i^l) \leqslant \Lambda_i^{l+1}(\varepsilon_i(k)|\mu_i^l)$ at each iteration, then we have $\Lambda_i^{l+1}(\varepsilon_i(k)|u_i^l) = c_i(\varepsilon_i(k), u_i(k)) + \alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l) \leqslant \Lambda_i^{l+1}(\varepsilon_i(k)|\mu_i^l) =$

* Corresponding author (email: hujp@uestc.edu.cn)

$c_i(\varepsilon_i(k),\mu_i(k))+\alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l)$, i.e., $\varepsilon_i^\top(k)Q_{ii}\varepsilon_i(k)+u_i^l(k)^\top R_{ii}u_i^l(k)+\sum_{j\in\mathcal{N}(i)}u_j^\top(k)S_{ij}u_j(k)+\alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l)\leqslant \varepsilon_i^\top(k)Q_{ii}\varepsilon_i(k)+\mu_i^l(k)^\top R_{ii}\mu_i^l(k)+\sum_{j\in\mathcal{N}(i)}u_j^\top(k)S_{ij}u_j(k)+\alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l)$. Furthermore, we have ${\mu_i^l}^\top(k)R_{ii}\mu_i^l(k)+\alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l)-\left[{u_i^l}^\top(k)R_{ii}u_i^l(k)+\alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l)\right]\geqslant 0$. Therefore, we have $(\mu_i^l(k)-u_i^l(k))^\top R_{ii}(\mu_i^l(k)-u_i^l(k))+\alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l)+2{\mu_i^l}^\top(k)R_{ii}u_i^l(k)-\left[\alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l)+2{u_i^l}^\top(k)R_{ii}u_i^l(k)\right]={\mu_i^l}^\top(k)R_{ii}\mu_i^l(k)+\alpha\Lambda_i^l(\varepsilon_i(k+1)|\mu_i^l)-\left[{u_i^l}^\top(k)R_{ii}u_i^l(k)+\alpha\Lambda_i^l(\varepsilon_i(k+1)|u_i^l)\right]\geqslant 0$ and thus the formula (A3) follows.

Subsisting (A3) into (A2) yields

$$\Lambda_i^{l+1}(\varepsilon_i(k))=Y_i^l(\Lambda_i^l,\mu_i^l)\geqslant Y_i^l(\Lambda_i^l,u_i^l)=\hat{\Lambda}_i^{l+1}(\varepsilon_i(k)). \tag{A4}$$

Similarly, one has

$$J_i^{l+1}(\varepsilon_i(k))=Y_i^l(J_i^l,\mu_i^l)\geqslant Y_i^l(J_i^l,u_i^l)=\hat{J}_i^{l+1}(\varepsilon_i(k)).$$

Since $0\leqslant J_i^0(\varepsilon_i(k))\leqslant\Lambda_i^0(\varepsilon_i(k))$, by the induction method, one has $0\leqslant J_i^l(k)\leqslant\Lambda_i^l(\varepsilon_i(k))$. The inequality (A4) gives the lower bound on $\Lambda_i^{l+1}(\varepsilon_i(k))$ such that

$$0\leqslant J_i^{l+1}(\varepsilon_i(k))\leqslant\hat{\Lambda}_i^{l+1}(\varepsilon_i(k))\leqslant\Lambda_i^{l+1}(\varepsilon_i(k)).$$

Then we have

$$0\leqslant J_i^l(k)\leqslant\hat{\Lambda}_i^l(\varepsilon_i(k))\leqslant\Lambda_i^l(\varepsilon_i(k)).$$

The proof is thus completed.

Lemma 2. For the iterative performance index function $J_i^l(k)$, there exists an upper bound $\overline{U}(k)$ such that

$$0\leqslant J_i^l(k)\leqslant\overline{U}(k). \tag{A5}$$

Proof. Let $\mu_i^l(k)$ be an arbitrary stabilizing policy, and its corresponding iterative performance index function $\Lambda_i^l(\varepsilon_i(k))$ be defined by

$$\Lambda_i^{l+1}(\varepsilon_i(k))=c_i(\varepsilon_i(k),\mu_i^l(k))+\alpha\Lambda_i^l(\varepsilon_i(k+1)).$$

Assume that $\Lambda_i^0(\varepsilon_i(k))=J_i^0(\varepsilon_i(k))=0$, then we have

$$\begin{aligned}\Lambda_i^{l+1}(\varepsilon_i(k))-\Lambda_i^l(\varepsilon_i(k))=&\Lambda_i^l(\varepsilon_i(k+1))-\Lambda_i^{l-1}(\varepsilon_i(k+1))\\=&\Lambda_i^{l-1}(\varepsilon_i(k+2))-\Lambda_i^{l-2}(\varepsilon_i(k+2))\\&\vdots\\=&\Lambda_i^1(\varepsilon_i(k+l))-\Lambda_i^0(\varepsilon_i(k+l)).\end{aligned} \tag{A6}$$

From (A6), one has

$$\begin{aligned}\Lambda_i^{l+1}(\varepsilon_i(k))=&\Lambda_i^1(\varepsilon_i(k+l))+\Lambda_i^l(\varepsilon_i(k))\\=&\Lambda_i^1(\varepsilon_i(k+l))+\Lambda_i^1(\varepsilon_i(k+l-1))+\Lambda_i^{l-1}(\varepsilon_i(k))\\=&\Lambda_i^1(\varepsilon_i(k+l))+\Lambda_i^1(\varepsilon_i(k+l-1))+\Lambda_i^1(\varepsilon_i(k+l-2))+\cdots+\Lambda_i^1(\varepsilon_i(k)),\end{aligned} \tag{A7}$$

which can be further written as

$$\begin{aligned}\Lambda_i^{l+1}(\varepsilon_i(k))=&\sum_{z=0}^{l}\Lambda_i^1(\varepsilon_i(k+z))\\=&\sum_{z=0}^{l}\Big(\varepsilon_i^\top(k)Q_{ii}\varepsilon_i(k+z)+\mu_i^\top(k)R_{ii}\mu_i(k+z)+\sum_{j\in\mathcal{N}(i)}\mu_j^\top(k+z)S_{ij}\mu_j(k+z)\Big).\end{aligned}$$

Since the control polices $\mu_i^l(k)$ are admissible, we have

$$\Lambda_i^{l+1}(\varepsilon_i(k))=\sum_{z=0}^{l}\Lambda_i^1(\varepsilon_i(k+z))\leqslant\sum_{z=0}^{\infty}\Lambda_i^1(\varepsilon_i(k+z))=\overline{U}(k).$$

Then, according to Lemma 1, it follows that $0\leqslant J_i^l(k)\leqslant\Lambda_i^l(\varepsilon_i(k))\leqslant\overline{U}(k)$. Thus the proof is completed.

We now present the proofs of the two main theorems. Theorem 1 states the convergence of the value iteration algorithm.

Proof of Theorem 1

Proof. According to Lemmas 1 and 2, one has the following inequality

$$0\leqslant J_i^l(k)\leqslant\Lambda_i^l(\varepsilon_i(k))\leqslant\overline{U}(k). \tag{A8}$$

It is noted that the performance index functions $\Lambda_i^l(\varepsilon_i(k))$ and $J_i^l(\varepsilon_i(k))$ are respectively given by

$$\begin{aligned}\Lambda_i^{l+1}(\varepsilon_i(k))=&c_i(\varepsilon_i(k),\mu_i^l(k))+\alpha\Lambda_i^l(\varepsilon_i(k+1)),\\J_i^{l+1}(\varepsilon_i(k))=&c_i(\varepsilon_i(k),u_i^l(k))+\alpha J_i^l(\varepsilon(k+1)).\end{aligned} \tag{A9}$$

Starting with $l=0$, let $\Lambda_i^0(\varepsilon_i(k))=J_i^0(\varepsilon_i(k))=0$. Since $J_i^1(\varepsilon_i(k))=c_i(\varepsilon_i(k),u_i^0(k))\geqslant 0$. Then, according to (A9), one has

$$\Lambda_i^{l+1}(\varepsilon_i(k))=c_i(\varepsilon_i(k),u_i^{l+1}(k))+\alpha\Lambda_i^l(\varepsilon(k+1)). \tag{A10}$$

Next, we will prove that if $\Lambda_i^0(\varepsilon_i(k))=J_i^0(\varepsilon_i(k))=0$, then $\Lambda_i^l(\varepsilon_i(k))\leqslant J_i^{l+1}(\varepsilon_i(k))$ by the induction method. According to (A10), one has

$$\begin{aligned}J_i^1(\varepsilon_i(k))-\Lambda_i^0(\varepsilon_i(k))=c_i(\varepsilon_i(k),\mu_i^0(k))\geqslant 0,\\ J_i^1(\varepsilon_i(k))\geqslant\Lambda_i^0(\varepsilon_i(k)).\end{aligned}$$

Now, assume that $\Lambda_i^l(\varepsilon_i(k))\leqslant J_i^{l+1}(\varepsilon_i(k))$ holds for $l-1$ such that $\Lambda_i^{l-1}(\varepsilon_i(k))\leqslant J_i^l(\varepsilon_i(k))$. Then one has

$$J_i^{l+1}(\varepsilon_i(k))-\Lambda_i^l(\varepsilon_i(k))=J_i^l(k)-\Lambda_i^{l-1}(\varepsilon_i(k))\geqslant 0.$$

or, equivalently,

$$\Lambda_i^l(\varepsilon_i(k))\leqslant J_i^{l+1}(\varepsilon_i(k)),\forall l.$$

From the fact that $\Lambda_i^l(\varepsilon_i(k))\leqslant J_i^{l+1}(\varepsilon_i(k))$ and $J_i^l(k)\leqslant\Lambda_i^l(\varepsilon_i(k))$, one has the following monotonic sequence

$$\begin{aligned}J_i^{l+1}(\varepsilon_i(k))&\geqslant\Lambda_i^l(\varepsilon_i(k))\geqslant J_i^l(k)\geqslant\cdots\\ &\geqslant J_i^1(\varepsilon_i(k))\geqslant\Lambda_i^0(\varepsilon_i(k))\geqslant J_i^0(\varepsilon_i(k))\geqslant 0.\end{aligned} \tag{A11}$$

From Lemma 2, $\overline{U}(k)$ is an upper bound of $J_i^{l+1}(\varepsilon_i(k))$, thus one has

$$0\leqslant J_i^l(k)\leqslant\Lambda_i^l(\varepsilon_i(k))\leqslant J_i^{l+1}(\varepsilon_i(k))\leqslant\overline{U}(k). \tag{A12}$$

According to (A12), $J_i^l(k)$ is a monotonically increasing sequence and has an upper bound $\overline{U}(k)$. From Lemma 1, let $\mu_i^l(k)=u_i^*(k)$, then one has $\Lambda_i(\varepsilon_i(k))=J_i^*(\varepsilon_i(k))$, as $l\to\infty$, which follows that $J_i^\infty(\varepsilon_i(k))\leqslant J_i^*(\varepsilon_i(k))$. From Lemma 2, one also has $J_i^*(\varepsilon_i(k))\leqslant\overline{U}(k)=J_i^\infty(\varepsilon_i(k))$. Then one has $J_i^*(\varepsilon_i(k))\leqslant J_i^\infty(\varepsilon_i(k))\leqslant J_i^*(\varepsilon_i(k))$, which implies that $J_i^l(k)$ will converge to the optimal solution $J_i^*(\varepsilon_i(k))$ monotonically. Once the performance index function converges to the optimal value $J_i^*(\varepsilon_i(k))$, the iterative control laws $u_i^l(k)$ also converge to the optimal value $u_i^*(k)$ given by (5). The proof is completed.

Proof of Theorem 2

Proof. Since the optimal performance index function $J_i^*(\varepsilon_i(k))$ satisfies the DT-HJB equation (4), then one has

$$c_i(\varepsilon_i(k),u_i^*(k))=J_i^*(\varepsilon_i(k))-\alpha J_i^*(e_i(k+1)). \tag{A13}$$

Multiplying both sides of (A13) by α^k leads to

$$\alpha^k c_i(\varepsilon_i(k),u_i^*(k))=\alpha^k J_i^*(\varepsilon_i(k))-\alpha^{k+1}J_i^*(\varepsilon_i(k+1)). \tag{A14}$$

For the closed-loop multi-agent system, we define the difference of the Lyapunov function as

$$\triangle(\alpha^k J_i^*(\varepsilon_i(k)))=\alpha^{k+1}J_i^*(\varepsilon_i(k+1))-\alpha^k J_i^*(\varepsilon_i(k)). \tag{A15}$$

Combining (A14), the equation (A15) can be rewritten as

$$\triangle(\alpha^k J^*(\varepsilon_i(k)))=-\alpha^k c_i(\varepsilon_i(k),u_i^*(k))<0. \tag{A16}$$

Form (A16), we conclude that the closed-loop multi-agent system is asymptotically stable, i.e., $\lim_{k\to\infty}\varepsilon_i(k)=0$. Thus the containment control problem is solved.

Appendix B Online actor-critic algorithm

The following algorithm is used for the implementation process of the actor-critic NNs.

Algorithm B1 Online actor-critic algorithm

Initialization:

For $\forall i$, choose a computation precision ϵ;

The initial values of the critic weights $\hat{W}_{ci}$ are chosen as zero and the actor weights $\hat{W}_{ai}$ are initialized in $[0,1]$ randomly;

Let κ_{ai} and κ_{ci} are learning rates.

Iteration:

1: Let the iteration index $l=0$; Start with the initial state $\varepsilon_i(0)$; Calculate the iterative performance index functions and the control laws according to the schemes (8)-(10);

2: **repeat**

3: Update the critic weights by

$$\hat{W}_{ci}^{(l+1)\top} = \hat{W}_{ci}^{l\top} - \kappa_{ci}\left(\mathcal{T}_{J_i}(k) - X_i^{\top}(k)\hat{W}_{ic}^{l}X_i(k)\right)X_i(k)X_i^{\top}(k);$$

where $\mathcal{T}_{J_i}(k)$ is given by (9);

4: Update the actor weights by

$$\hat{W}_{ai}^{(l+1)\top} = \hat{W}_{ai}^{l\top} - \kappa_{ai}\left(\hat{W}_{ai}^{l\top}X_i(k) - \overline{u}_i^{l}(k)\right)X_i^{\top}(k);$$

5: **until** $|\hat{J}_i^{l+1}(k) - \hat{J}_i^{l}(k)| \leqslant \epsilon$; End.

Appendix C Numerical experiments

From the interaction network, we know that the adjacency matrix $\mathcal{A}$ is given by $a_{12}=a_{14}=a_{21}=a_{31}=a_{42}=1$, and the leader adjacency matrices are given by $\mathcal{B}_5 = diag\{1,0,0,0\}$, $\mathcal{B}_6 = diag\{0,1,0,0\}$, $\mathcal{B}_7 = diag\{0,0,0,1\}$, respectively. The system matrices of the leaders and the followers are given by $A=\begin{bmatrix} 0.9950 & 0.0998 \\ -0.09983 & 0.9950 \end{bmatrix}$, $B_1=\begin{bmatrix} 0.20 \\ 0.08 \end{bmatrix}$, $B_2=\begin{bmatrix} 0.21 \\ 0.28 \end{bmatrix}$, $B_3=\begin{bmatrix} 0.20 \\ 0.18 \end{bmatrix}$. $B_4=\begin{bmatrix} 0.20 \\ 0.10 \end{bmatrix}$.

In the performance index functions $J_i(\varepsilon_i(k),u_i(k))$, the weighting matrices are selected as follows: $Q_1=Q_2=Q_3=Q_4=I_{2\times 2}$, $R_{11}=R_{22}=R_{33}=R_{44}=1$, $S_{12}=S_{14}=S_{21}=S_{31}=S_{42}=1$, $S_{13}=S_{23}=S_{24}=S_{32}=S_{34}=S_{41}=S_{43}=0$.

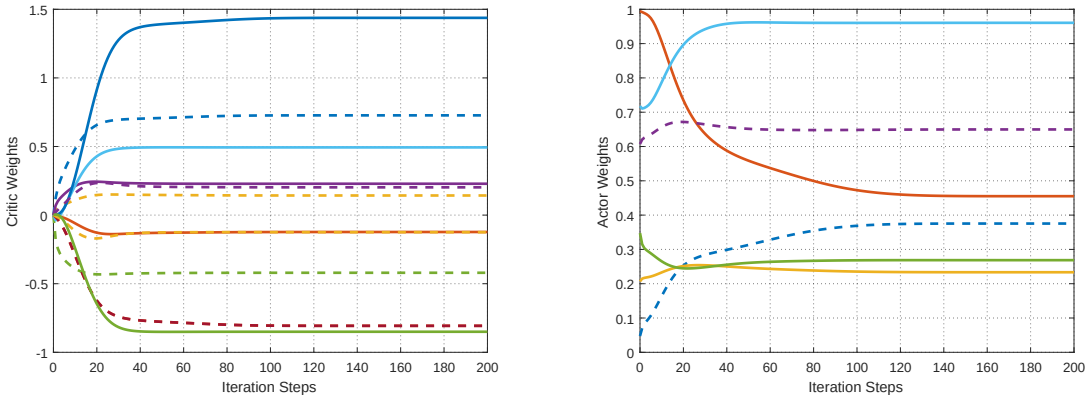

Figure C1 Convergence of the critic NN weights and the actor NN weights.

In Algorithm B1, we choose the discount factor $\alpha=0.92$, and the learning rates $\kappa_{ci}=\kappa_{ai}=0.04$ for $\forall i=1,2,3,4$. Then, during the iterative process, the critic NNs and actor NNs are trained until the network approximation errors reach the precision $\varepsilon=10^{-5}$. The initial values of the critic weights $\hat{W}_{ci}(0)\in\mathcal{R}^{2\times 6}$ are chosen as zero and the actor weights $\hat{W}_{ai}(0)\in\mathcal{R}^{6\times 1}$ are initialized in $[0,1]$ randomly. The initial states of the followers are given by $x_1(0)=[0.0637,0.3229]^{\top}$, $x_2(0)=[0.0984,0.1700]^{\top}$, $x_3(0)=[0.3712,0.0398]^{\top}$, and $x_4(0)=[0.7092,0.6413]^{\top}$.

Figure C1 gives the evolution of the critic and actor weights for agent 1. The evolution of the disagreement vectors ε_{i1} and ε_{i2} are given in Figure C2 (a) and (b), respectively, which shows that the disagreement vectors converge to zero as k is larger than 200. The state trajectories of the followers and the leaders are shown in Figure C3, where the red dashed lines represent the envelope formed by the trajectories of the leaders. It can be seen that the trajectories of the followers will stay in the envelope formed by the leaders.

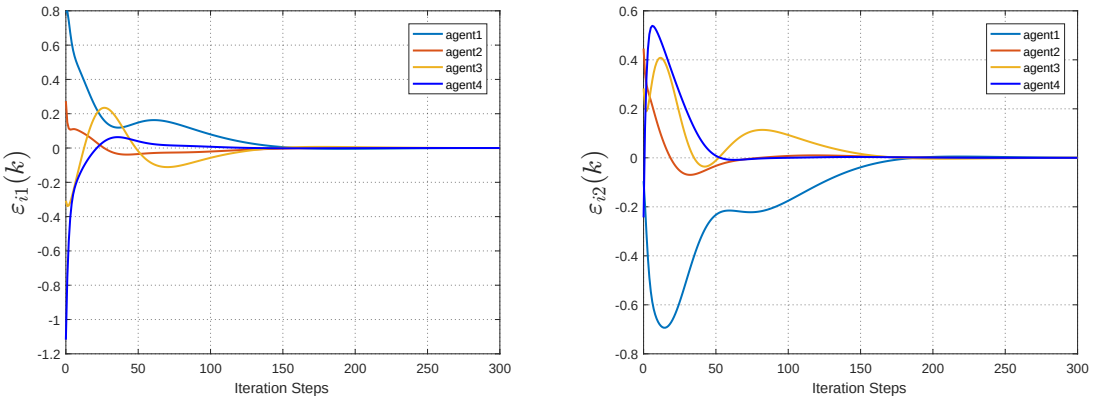

Figure C2 Evolution of the disagreement vectors $\varepsilon_{i1}(k)$ and $\varepsilon_{i2}(k)$ for $i = 1, 2, 3, 4$.

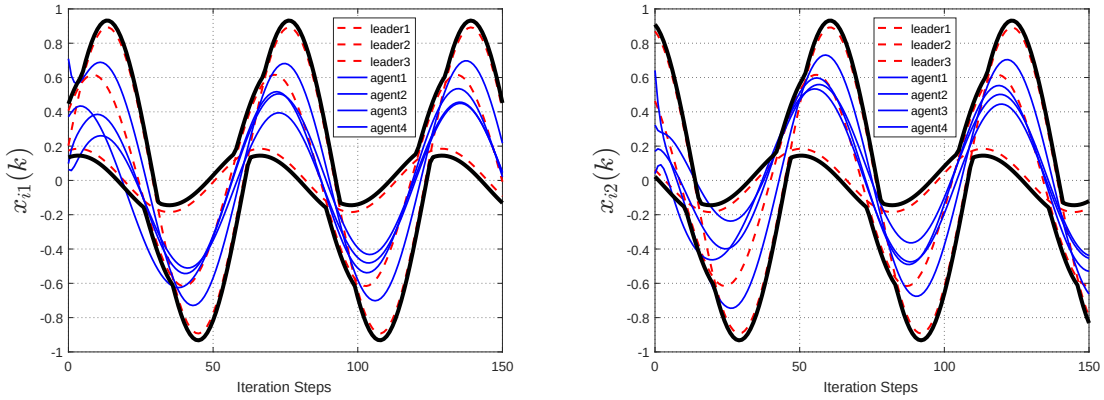

Figure C3 The state evolution of the followers and leaders.