Bridging Learning and Passivity: Integral Reinforcement Learning with SPR Transformation for Multi-Agent Coordination

Jie Ren Bahram Shafai

Electrical & Computer Engineering, Northeastern University, Boston, MA, United States of America (USA)

DOI: https://doi.org/10.51244/IJRSI.2026.1307000205

Received: 22 July 2026; Accepted: 27 July 2026; Published: 07 August 2026

ABSTRACT

This paper presents a framework that enables multi-agent systems to learn controllers independently through integral reinforcement learning while guaranteeing modular stability when connected. The framework addresses a gap: existing modular control methods require analytical models and cannot handle learned controllers, while multi-agent learning methods require network coupling during training. The proposed approach combines Integral Reinforcement Learning for independent training with a Strictly Positive Real transformation via a Proportional-Integral observer for post-learning certification. We establish necessary and sufficient conditions for the PI observer construction and prove that SPR-transformed agents achieve output synchronization, regulation to zero, and global asymptotic stability under strongly connected graphs. Unlike standard passivity-based coordination that synchronizes to arbitrary constants, the positive definiteness condition required by strict positive realness ensures regulation to the origin. We derive a performance bound showing that the network convergence rate and the stability margin depend on the graph topology, the agent dynamics, and the controller quality. The framework trades local optimality for modular guarantees: controllers optimized for isolated operation become conservative when connected, but the performance bound quantifies this cost, showing that better local controllers reduce the penalty. Simulations on 10 heterogeneous agents validate the framework across 800 Monte Carlo trials, demonstrating that IRL achieves statistical equivalence to optimal LQR despite 20% model uncertainty and unstable agents.

Department of Electrical Engineering, Northeastern University, Boston, MA 02115, USA

Keywords: Cooperative control, multi-agent systems, reinforcement learning, passivity, optimal control

INTRODUCTION

Multi-agent coordination problems arise in applications such as unmanned aerial vehicle formations, cooperative robotics, autonomous vehicle platoons, and distributed sensor networks [24]. Designing these systems under uncertain or time-varying dynamics, distributed implementation, and plug-and-play needs creates trade-offs. One is between local performance and global coordination. The other is between learning-based methods and modular stability. Classical centralized optimal control can simultaneously achieve performance, stability, and coordination when complete system models are available. However, practical multi-agent systems often need distributed design in which agents are built independently and system models are unknown or inaccurate. Solving these design problems has led to two distinct research methods with different approaches.

Existing modular control methods guarantee stability and synchronization with agent system structures. Qu and Simaan developed a passivity-short method enabling agents with bounded passivity indices to achieve plug-and-play consensus regardless of the network topology [20]. Their framework separates agent-level controller design (rendering individual agents passivity-short) and network-level distributed control, which needs only agents' worst-case passivity bounds rather than detailed dynamics models. Arcak established passivity as a modular design tool, proving that bidirectional coupling of passive components preserves stability and enables distributed formation stabilizing control with neighbor information [2]. Ren and Beard, and Olfati-Saber and Murray, proved that topology-agnostic consensus protocols guarantee agreement [18, 23]. Papers [6, 13, 25, 32] achieve

output synchronization for heterogeneous systems relying on state-space models of agent dynamics. Recent work extends these ideas through Agarwal et al.'s distributed synthesis using QSR-dissipativity certificates [1], Tanaka and Langbort's retrofit control for modular system modifications [27], Lee and Shim's analysis of heterogeneous teams under strong coupling [14], and extensions of passivity theory for sampled-data systems and output consensus [10, 11]. These methods provide modular stability proofs for plug-and-play operation, but they use only fixed controllers designed from analytical models. When system dynamics are uncertain, time-varying, or too complex for analytical modeling, these model-based methods become intractable or infeasible.

Existing multi-agent learning methods, in contrast, need network connections during training to achieve coordination with convergence guarantees. Independent Learning trains agents without coordination and thus lacks convergence guarantees. Jin et al. show through finite-sample analysis that Independent Learning achieves only approximate convergence with error terms quantifying agent coupling [9]. Because independent training fails to guarantee coordination, the field has adopted networked training methods. In the control community, researchers solve coupled Hamilton–Jacobi equations for coordination using differential graphical games [8, 16, 28–30]. Other RL methods also achieve cooperative output regulation [7] and consensus [34]. In the machine learning community, Centralized Training with Decentralized Execution methods use value decomposition and actor-critic architectures [17, 21, 33]. Because these methods achieve coordination through learned behaviors, network coupling during training is essential. Comprehensive surveys of both communities confirm that network coupling during training is standard practice and independent learning with coordination guarantees is absent from the literature [5, 19].

Currently, no framework enables agents to learn controllers independently through reinforcement learning, then apply a post-learning Strictly Positive Real (SPR) transformation to guarantee modular stability when connected. This gap exists between modular control theory and multi-agent learning: modular methods cannot handle learned controllers without an analyzable structure, while learning methods cannot decompose network-coupled objectives into independent problems that guarantee coordination when composed. The core challenge is enabling independent learning, which is needed for privacy and scalability, while achieving modular stability guarantees when agents connect.

This paper presents a three-stage framework that bridges independent reinforcement learning with modular passivity guarantees. First, each agent trains independently using Integral Reinforcement Learning (IRL) to solve its Linear Quadratic Regulator (LQR) problem from trajectory data without system models. After training, a SPR transformation via a Proportional–Integral (PI) observer is applied to each agent, creating a passive output. Finally, network coupling on the SPR-transformed outputs ensures that agents synchronize, regulate, and remain stable. This enables independent learning with plug-and-play guarantees. The framework trades local optimality for modular stability guarantees. Controllers optimized for isolated operation become conservative when connected, with degradation bounded by the network topology and the controller quality.

This paper is organized as follows. The Preliminaries section reviews IRL and SPR transformation methods. The Framework Design and Analysis section presents the theoretical results, including conditions for SPR transformation via a PI observer and multi-agent synchronization, regulation, stability guarantees, and a performance bound. The Simulations section demonstrates synchronization, regulation, and stability across network topologies. The Conclusion section outlines future research directions.

PRELIMINARIES

Integral reinforcement learning

Consider the linear time-invariant system

$$\dot{x} = Ax + Bu$$

where $x \in \mathbb{R}^n$ is the state and $u \in \mathbb{R}^m$ is the control input. The matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are unknown. We assume (A, B) is stabilizable.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

The LQR problem seeks the control policy that minimizes the infinite-horizon quadratic cost

$$V(x(t_0), t_0) = \int_{t_0}^{\infty} (x^T(\tau)Qx(\tau) + u^T(\tau)Ru(\tau))\, d\tau \quad (1)$$

with $Q \geq 0$, $R > 0$ and $(Q^{1/2}, A)$ detectable. The optimal control is $u(t) = -Kx(t)$ with

$$K = R^{-1}B^T P$$

where P is the unique positive definite solution of the algebraic Riccati equation

$$A^T P + PA - PBR^{-1}B^T P + Q = 0.$$

IRL solves the LQR problem without knowing A through policy iteration. Given an initial stabilizing gain K_1, the algorithm alternates between policy evaluation and policy improvement [31].

Policy Evaluation: For the current policy K_i, solve for P_i satisfying

$$x^T(t)P_i x(t) = \int_t^{t+T} x^T(\tau)(Q + K_i^T R K_i)x(\tau)d\tau + x^T(t+T)P_i x(t+T) \quad (2)$$

for any $T > 0$ along the closed-loop trajectory $\dot{x} = (A - BK_i)x$. We solve this equation with a Kronecker product parameterization [4]. Let $\bar{x}(t)$ denote the quadratic polynomial basis vector with elements $x_j(t)x_k(t)$ for $j \leq k$, and let $\bar{p}_i$ be the vector of unknown parameters in P_i where off-diagonal elements are $2P_{jk}$. Then (2) becomes

$$\bar{p}_i^T(\bar{x}(t) - \bar{x}(t+T)) = \int_t^{t+T} x^T(\tau)(Q + K_i^T R K_i)x(\tau)d\tau. \quad (3)$$

The integral on the right side is measured by augmenting the system with a state $\dot{V}(t) = x^T(t)Qx(t) + u^T(t)Ru(t)$, so the integral equals $V(t+T) - V(t)$. Collecting data at $N \geq n(n+1)/2$ points along a single trajectory yields the least-squares problem

$$\bar{p}_i = (XX^T)^{-1}XY$$

where

$$X = [\bar{x}_1^{\Delta}\ \bar{x}_2^{\Delta}\ \cdots\ \bar{x}_N^{\Delta}], \quad \bar{x}_j^{\Delta} = \bar{x}_j(t) - \bar{x}_j(t+T),$$
$$Y = [d(\bar{x}_1, K_i)\ d(\bar{x}_2, K_i)\ \cdots\ d(\bar{x}_N, K_i)]^T$$

with $d(\bar{x}_j, K_i) = \int_t^{t+T} x^T(\tau)(Q + K_i^T R K_i)x(\tau)d\tau$. This requires persistent excitation to ensure that XX^T is invertible.

Policy Improvement:

$$K_{i+1} = R^{-1}B^T P_i. \quad (4)$$

The policy update only needs the B matrix in (4). The system A matrix is embedded in the observed states $x(t)$ and $x(t+T)$, so A is never required. The iteration terminates when $\| K_{i+1} - K_i \| < \epsilon$ for a tolerance $\epsilon > 0$.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

SPR transformation via PI observer

Lemma 1 (Kalman-Yakubovich-Popov [12].] Let $G(s) = C(sI - A)^{-1}B + D$ be a $p \times p$ transfer function matrix, where (A, B) is controllable and (A, C) is observable. Then, $G(s)$ is strictly positive real if and only if there exist matrices $P = P^T > 0$, L, and W, and a positive constant ϵ such that

$$PA + A^TP = -L^TL - \epsilon P$$
$$PB = C^T - L^TW$$
$$W^TW = D + D^T.$$

For the special case $D = 0$, the conditions simplify to the existence of $P = P^T > 0$ and $Q > 0$ satisfying

$$PA + A^TP = -Q - \epsilon P$$
$$PB = C^T.$$

Lemma 2 ([12).] The linear time-invariant minimal realization $\dot{x} = Ax + Bu$, $y = Cx + Du$ with $G(s) = C(sI - A)^{-1}B + D$ is

- passive if $G(s)$ is positive real;
- strictly passive if $G(s)$ is strictly positive real.

A PI observer for the system is [3, 26]

$$\dot{\hat{x}} = (A - L_PC)\hat{x} + L_Py + Bu + Bw$$
$$\dot{w} = L_I(y - C\hat{x})$$

where $\hat{x} \in \mathbb{R}^n$ is the state estimate, $w \in \mathbb{R}^m$ is the integral term, and $L_P \in \mathbb{R}^{n\times p}$ and $L_I \in \mathbb{R}^{m\times p}$ are the proportional and integral gains.

Introducing $u = -K\hat{x} + v$ where $K \in \mathbb{R}^{m\times n}$ is a stabilizing feedback gain and $v \in \mathbb{R}^m$ is an external input signal, the system becomes

$$\dot{x} = Ax + B(-K\hat{x} + v) = (A - BK)x - BKe + Bv \quad (5)$$

where $e := \hat{x} - x$. The PI observer becomes

$$\dot{\hat{x}} = (A - L_PC)\hat{x} + L_Py + B(-K\hat{x} + v) + Bw$$
$$\dot{w} = -L_ICe.$$

Combining the system and error dynamics, the augmented system is

$$\begin{bmatrix}\dot{x}\\ \dot{e}\\ \dot{w}\end{bmatrix} = \begin{bmatrix}A - BK & -BK & 0\\ 0 & A - L_PC & B\\ 0 & -L_IC & 0\end{bmatrix}\begin{bmatrix}x\\ e\\ w\end{bmatrix} + \begin{bmatrix}B\\ 0\\ 0\end{bmatrix}v. \quad (6)$$

Theorem 1 ([22).] Let the strictly proper transfer function $G(s)$ with m inputs and p outputs be represented by its state-space realization $\dot{x} = Ax + Bu$, $y = Cx$ where (A, B) is controllable and (C, A) is observable. Then, there exist proportional and integral gain matrices L_P and L_I, with stabilizable state-feedback gain K, as well as a matrix $C_{new} = B^TP_K$, such that the transfer function matrix between input v and new output $\gamma = C_{new}\hat{x}$ is SPR.

Augmented system realization

For analysis and multi-agent coordination, we express the augmented system (6) in compact form. Define the augmented state

$$\xi = [x^T \quad e^T \quad w^T]^T \in \mathbb{R}^{2n+m}$$

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

and the augmented system matrices

$$A_0 = \begin{bmatrix} A - BK & -BK & 0 \\ 0 & A - L_P C & B \\ 0 & -L_I C & 0 \end{bmatrix},$$
$$B_0 = [B^T \quad 0 \quad 0]^T.$$

where $A_K = A - BK$ and $A_z = \begin{bmatrix} A - L_P C & B \\ -L_I C & 0 \end{bmatrix}$ as defined previously.

By the construction in [22], there exists $P_0 = P_0^T > 0$ satisfying the KYP conditions for the augmented system. The SPR output matrix is

$$C_0 = B_0^T P_0.$$

The augmented system is

$$\dot{\xi} = A_0 \xi + B_0 v, \quad (7)$$
$$\gamma = C_0 \xi. \quad (8)$$

Since $e = \hat{x} - x$, we have $\hat{x} = x + e$, so the output becomes

$$\gamma = C_0 \xi = B_0^T P_0 \begin{bmatrix} x \\ e \\ w \end{bmatrix} = B^T P_K (x + e) = B^T P_K \hat{x} = C_{new} \hat{x}.$$

The augmented realization (A_0, B_0, C_0) realizes the SPR transfer function from v to γ described by the system (A, B, C_{new}).

FRAMEWORK DESIGN AND ANALYSIS

Framework architecture

Consider a network of N agents, each with linear dynamics

$$\dot{x}_i = A_i x_i + B_i u_i, \quad i = 1, \ldots, N$$

where $x_i \in \mathbb{R}^{n_i}$ is the plant state and $u_i \in \mathbb{R}^{m_i}$ is the control input. The matrix A_i is unknown.

The framework proceeds in three stages. First, each agent applies IRL to learn its optimal LQR controller K_i^* from trajectory data. Agents operate independently during this training phase without network connectivity or communication. The learned K_i^* minimizes the local cost (1).

Second, after training completes, each agent implements a PI observer with augmented dynamics (7) and SPR output (8).

Third, agents connect via the network coupling. The external input v_i from (7) is defined by

$$v_i = \sum_{j \in \mathcal{N}_i} a_{ij} (\gamma_j - \gamma_i) \quad (9)$$

where $\mathcal{N}_i$ denotes the neighbors of agent i in the communication graph and $a_{ij} > 0$ are the coupling weights.

Because agents train independently with $v_i = 0$, the learned controller K_i^* is optimal only for the isolated system. When connected, the network input $v_i \neq 0$ degrades local optimality. The framework trades local optimality for modular guarantees of stability, synchronization, and regulation.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Feasibility conditions for SPR transformation

The PI observer construction requires that each agent's system satisfy specific structural properties.

Lemma 3 (PI Observer Feasibility). Consider the system $\dot{x} = Ax + Bu$ with output $y = Cx$. There exist gains $L_P \in \mathbb{R}^{n\times p}$ and $L_I \in \mathbb{R}^{m\times p}$ such that the PI observer error dynamics matrix A_z is Hurwitz if and only if

1. (A, C) is observable, and
2. $\text{rank}(CA^{-1}B) = m$ when A is invertible.

Proof. See Appendix A. □

Remark 1. When A is singular, condition (2) generalizes to $\text{rank}\begin{bmatrix} A & B \\ C & 0 \end{bmatrix} = n + m$, as shown in Appendix A.

The second condition requires each input channel to influence the measured outputs at steady state. If $\text{rank}(CA^{-1}B) < m$, the augmented system has an unobservable eigenvalue at $s = 0$ that cannot be relocated by any choice of gains, preventing stabilization.

Synchronization and regulation

After SPR transformation, each agent has augmented dynamics $(A_{0,i}, B_{0,i}, C_{0,i})$ that satisfy the KYP conditions for strict positive realness. The following result establishes the network behavior under Laplacian coupling.

Theorem 2 (Regulation to Zero). Consider N agents with dynamics after SPR transformation via PI observer: $\dot{\xi}_i = A_{0,i}\xi_i + B_{0,i}v_i$, where $\xi_i = [x_i^T, e_i^T, w_i^T]^T \in \mathbb{R}^{n_{aug,i}}$ is the augmented state, $v_i \in \mathbb{R}^{m_i}$ is the network input, $\gamma_i = C_{0,i}\xi_i$,
and $\gamma_i \in \mathbb{R}^{m_i}$ is the SPR output. Assume $(A_{0,i}, B_{0,i}, C_{0,i})$ is strictly positive real with $A_{0,i}$ Hurwitz. Let the communication graph be strongly connected with Laplacian L. Under the coupling (9), the network achieves:

1. Output synchronization: $\lim_{t\to\infty} \| \gamma_i(t) - \gamma_j(t) \| = 0$ for all i, j
2. Regulation to zero: $\lim_{t\to\infty} \gamma_i(t) = 0$ for all i
3. Global asymptotic stability: The origin $\xi = 0$ is globally asymptotically stable

Proof. Since each agent is SPR, by the Kalman–Yakubovich–Popov lemma there exist $P_{0,i} = P_{0,i}^T > 0$ and $Q_i = Q_i^T > 0$ such that

$$A_{0,i}^T P_{0,i} + P_{0,i} A_{0,i} = -Q_i,$$
$$P_{0,i} B_{0,i} = C_{0,i}^T.$$

Define the storage function $V_i(\xi_i) = \frac{1}{2}\xi_i^T P_{0,i}\xi_i$. Its derivative along the augmented dynamics is

$$\begin{aligned} \dot{V}_i &= \xi_i^T P_{0,i}(A_{0,i}\xi_i + B_{0,i}v_i) \\ &= \frac{1}{2}\xi_i^T (A_{0,i}^T P_{0,i} + P_{0,i}A_{0,i})\xi_i + \gamma_i^T v_i \\ &= -\frac{1}{2}\xi_i^T Q_i \xi_i + \gamma_i^T v_i. \end{aligned}$$

Let $p = [p_1, \dots, p_N]^T$ be the positive left eigenvector of L satisfying $p^T L = 0$. Consider the weighted Lyapunov function

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

$$V = \sum_{i=1}^{N} p_i V_i(\xi_i) = \frac{1}{2}\sum_{i=1}^{N} p_i \xi_i^T P_{0,i}\xi_i.$$

Its derivative is

$$\dot{V} = \sum_{i=1}^{N} p_i \dot{V}_i = \sum_{i=1}^{N} p_i \left(-\frac{1}{2}\xi_i^T Q_i \xi_i + \gamma_i^T v_i\right)$$
$$= -\frac{1}{2}\sum_{i=1}^{N} p_i \xi_i^T Q_i \xi_i + \sum_{i=1}^{N} p_i \gamma_i^T v_i.$$

Substituting (9) yields

$$\sum_{i=1}^{N} p_i \gamma_i^T v_i = \sum_{i=1}^{N} p_i \gamma_i^T \sum_{j=1}^{N} a_{ij}(\gamma_j - \gamma_i)$$
$$= \sum_{i,j} p_i a_{ij}(\gamma_i^T \gamma_j - \| \gamma_i \|^2)$$
$$= -\frac{1}{2}\sum_{i,j} p_i a_{ij} \| \gamma_i - \gamma_j \|^2.$$

The last equality follows from $p^T L = 0$ and symmetry of the quadratic form [15].

Thus

$$\dot{V} = -\frac{1}{2}\sum_{i=1}^{N} p_i \xi_i^T Q_i \xi_i - \frac{1}{2}\sum_{i,j} p_i a_{ij} \| \gamma_i - \gamma_j \|^2 \leq 0.$$

Since V is positive definite and radially unbounded, all trajectories are bounded. By LaSalle's invariance principle, the largest invariant set where $\dot{V} = 0$ satisfies

$$p_i \xi_i^T Q_i \xi_i = 0 \text{ for all } i,$$
$$\gamma_i = \gamma_j \text{ for all } i, j.$$

Because $p_i > 0$ and $Q_i > 0$, the first condition requires $\xi_i = 0$ for all i. Hence $\gamma_i = C_{0,i}\xi_i = 0$. Therefore, the origin is globally asymptotically stable, achieving both synchronization and regulation to zero. □

Remark 2. Unlike standard passivity-based synchronization, strict positive realness ensures that $A_{0,i}$ is Hurwitz, and the $Q_i > 0$ condition ensures that the invariant set equals the origin. This guarantees regulation to zero rather than synchronization to an arbitrary constant.

Performance degradation bound

While local optimality is lost when agents connect, the following result quantifies the degradation and shows that controller quality affects network performance. The spectral abscissa $\alpha_{net} = \max \mathrm{Re}\,\lambda(A_{net})$ determines the network stability margin and the convergence rate. Trajectories decay with an exponential rate α_{net}, so a more negative α_{net} indicates faster convergence and larger stability margins.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Lemma 4 (SPR Output Factorization). Let an augmented agent after PI observer SPR transformation be realized as $\begin{matrix}\dot{\xi}_i = A_{0,i}\xi_i + B_{0,i}v_i,\\ \gamma_i = C_{0,i}\xi_i,\end{matrix}$ with $\xi_i = [x_i^T, e_i^T, w_i^T]^T \in \mathbb{R}^{2n_i+m_i}$ and observer estimate $\hat{x}_i = x_i + e_i \in \mathbb{R}^{n_i}$. Define the selection matrix $S_{\hat{x},i} = [I_{n_i}, I_{n_i}, 0_{n_i\times m_i}] \in \mathbb{R}^{n_i\times(2n_i+m_i)}$, which satisfies $\hat{x}_i = S_{\hat{x},i}\xi_i$ and $S_{\hat{x},i}S_{\hat{x},i}^T = 2I_{n_i}$. By the SPR construction, there exists a plant-dimension matrix $P_{K,i} \in \mathbb{R}^{n_i\times n_i}$ such that $\gamma_i = C_{0,i}\xi_i = B_i^T P_{K,i}\hat{x}_i = (B_i^T P_{K,i})S_{\hat{x},i}\xi_i \quad \forall\xi_i,$ (10) hence $C_{0,i} = (B_i^T P_{K,i})S_{\hat{x},i} \in \mathbb{R}^{m_i\times(2n_i+m_i)}$. Consequently, $\| C_{0,i} \|_2 \le \| B_i^T P_{K,i} \|_2 \| S_{\hat{x},i} \|_2 = \sqrt{2} \| B_i^T P_{K,i} \|_2$.

Proof. Equality (10) states that the SPR output γ_i computed from the augmented realization equals the SPR output from the plant realization with output matrix $C_{new,i} = B_i^T P_{K,i}$ acting on $\hat{x}_i = S_{\hat{x},i}\xi_i$. Since the expressions agree for all ξ_i, the output matrices are equal: $C_{0,i} = (B_i^T P_{K,i})S_{\hat{x},i}$. The norm bound follows from submultiplicativity and $\| S_{\hat{x},i} \|_2 = \sqrt{\lambda_{max}(S_{\hat{x},i}S_{\hat{x},i}^T)} = \sqrt{2}$. □

Theorem 3 (Performance Degradation Bound). Consider N agents with augmented dynamics $(A_{0,i}, B_{0,i}, C_{0,i})$ after SPR transformation on a strongly connected directed graph with Laplacian L. The network spectral abscissa satisfies

$$\alpha_{net} \le \alpha_{max}^{local} + \kappa_p^2\lambda_{max}(L_s)\beta_B\sqrt{2}(\beta_{KR} + \Delta_{K\text{-}lqr}) \quad (11)$$

where:

- $\alpha_{net} = \max \mathrm{Re}\,\lambda(A_{net})$ is the spectral abscissa of the network matrix $A_{net} = \mathcal{A} - \mathcal{B}(L \otimes I)\mathcal{C}$ with $\mathcal{A} = \mathrm{diag}(A_{0,1}, \dots, A_{0,N})$, $\mathcal{B} = \mathrm{diag}(B_{0,1}, \dots, B_{0,N})$, $\mathcal{C} = \mathrm{diag}(C_{0,1}, \dots, C_{0,N})$
- $\alpha_{max}^{local} = \max_i \lambda_{max}(He(A_{0,i}))$ is the maximum local spectral bound
- $\beta_B = \max_i \| B_{0,i} \|_2$ and $\beta_C = \max_i \| C_{0,i} \|_2$ are worst-case system norms
- $L_s = \frac{1}{2}(D_p L + L^T D_p)$ is the symmetrized Laplacian with $D_p = \mathrm{diag}(p)$ and $p^T L = 0$
- $\kappa_p = \sqrt{\max_i p_i / \min_i p_i}$ is the graph conditioning factor
- $\beta_{KR} = \max_i \| R_i K_i^* \|_2$ and $\Delta_{K\text{-}lqr} = \max_i \delta_i$ where $\delta_i = \| B_i^T (P_{K,i} - P_{lqr,i}) \|_2$

Here $P_{K,i} \in \mathbb{R}^{n_i\times n_i}$ is the plant-dimension matrix from Lemma 4 and $P_{lqr,i}$ solves the algebraic Riccati equation for $K_i^* = R_i^{-1}B_i^T P_{lqr,i}$.

Proof. **(i) Spectral abscissa upper bound.** For any real matrix M with an eigenvalue λ and normalized eigenvector v, we have $v^T He(M) v = \mathrm{Re}(\lambda)$. Since $v^T He(M) v \le \lambda_{max}(He(M))$ for any normalized v, we obtain $\mathrm{Re}(\lambda) \le \lambda_{max}(He(M))$ for all eigenvalues. Taking the maximum gives

$$\alpha_{net} \le \lambda_{max}(He(A_{net})).$$

(ii) Hermitian decomposition. Since $\mathcal{A} = \mathrm{diag}(A_{0,i})$ is block-diagonal,

$$He(A_{net}) = \mathrm{diag}(He(A_{0,1}), \dots, He(A_{0,N})) - He(\mathcal{B}(L \otimes I)\mathcal{C}).$$

(iii) Eigenvalue perturbation. For Hermitian matrices A and B, the eigenvalues satisfy $\lambda_{max}(A - B) \le \lambda_{max}(A) + \| B \|_2$ by the Weyl perturbation theorem. Applying this yields

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

$$\begin{aligned}\lambda_{\max}(\mathrm{He}(A_{net})) &\le \lambda_{\max}(\mathrm{diag}(\mathrm{He}(A_{0,i})))+\| \mathrm{He}(\mathcal{B}(L\otimes I)\mathcal{C}) \|_2 \\ &= \max_i \lambda_{\max}(\mathrm{He}(A_{0,i}))+\| \mathrm{He}(\mathcal{B}(L\otimes I)\mathcal{C}) \|_2 \\ &= \alpha_{\max}^{local}+\| \mathrm{He}(\mathcal{B}(L\otimes I)\mathcal{C}) \|_2.\end{aligned}$$

(iv) Weighted Hermitian bound. By Lemma 8 (Appendix B),

$$\| \mathrm{He}(\mathcal{B}(L\otimes I)\mathcal{C}) \|_2 \le \kappa_p^2 \lambda_{\max}(L_s)\beta_B\beta_C.$$

(v) Bound β_C using controller quality. From Lemma 4, $\| C_{0,i} \|_2 \le \sqrt{2} \| B_i^T P_{K,i} \|_2$. Adding and subtracting $B_i^T P_{lqr,i}$ gives

$$\begin{aligned}\| B_i^T P_{K,i} \|_2 &\le \| B_i^T P_{lqr,i} \|_2 + \| B_i^T (P_{K,i} - P_{lqr,i}) \|_2 \\ &= \| R_i K_i^* \|_2 + \delta_i.\end{aligned}$$

Taking the maximum over all agents yields $\beta_C \le \sqrt{2}(\beta_{KR} + \Delta_{K\text{-}lqr})$.

Combining (i)-(v) establishes (11). □

Remark 3. For identical agents, $\beta_B = \| B_0 \|_2$, $\beta_C = \| C_0 \|_2$, and $\alpha_{\max}^{local} = \lambda_{\max}(\mathrm{He}(A_0))$, giving $\alpha_{net} \le \lambda_{\max}(\mathrm{He}(A_0)) + \kappa_p^2 \lambda_{\max}(L_s) \| B_0 \|_2 \sqrt{2}(\| RK^* \|_2 + \delta_{spr\text{-}lqr})$.

Remark 4. The bound separates the performance penalty into graph effects ($\kappa_p^2 \lambda_{\max}(L_s)$), system structure ($\beta_B$), and controller quality ($\sqrt{2}(\beta_{KR} + \Delta_{K\text{-}lqr})$). The conditioning factor κ_p equals 1 for balanced graphs and grows with directedness. The term $\beta_{KR} = \max_i \| R_i K_i^* \|_2$ depends on the learned controllers, showing that better local controllers across all agents reduce the network degradation penalty.

Remark 5. This result justifies using IRL to learn optimal K_i^* rather than arbitrary stabilizing controllers. While local optimality is lost when agents connect, the bound shows that the controller quality directly affects the network performance through the β_{KR} term. Optimal controllers minimize this penalty, improving the worst-case network stability margin even in heterogeneous teams.

Framework completeness

The preceding results combine to establish the complete framework guarantees.

Theorem 4 (IRL-SPR Framework). Consider the three-stage framework where each agent:

1. Applies IRL from initial stabilizing gain $K_i^{(1)}$ to learn optimal K_i^*
2. Implements SPR transformation via PI observer using learned K_i^*
3. Connects via the Laplacian coupling (9)

Assume each system satisfies $\mathrm{rank}(C_i A_i^{-1} B_i) = m_i$ and the communication graph is strongly connected. Then:

1. Local optimality: Each K_i^* minimizes its local LQR cost for the isolated system
2. Network stability: The coupled network is globally asymptotically stable
3. Synchronization: $\lim_{t\to\infty} \| \gamma_i(t) - \gamma_j(t) \| = 0$ for all i, j
4. Regulation: $\lim_{t\to\infty} \gamma_i(t) = 0$ for all i

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

5. Bounded degradation: $\alpha_{net} \leq \alpha_{max}^{local} + \kappa_p^2 \lambda_{max}(L_s)\beta_B\sqrt{2}(\beta_{KR} + \Delta_{K\text{-}lqr})$

Proof. Property (1) follows from Vrabie's convergence proof for IRL, which establishes that policy iteration from any stabilizing initial controller converges to the optimal LQR solution for the isolated system. Properties (2)-(4) follow from Theorem 2 applied to the post-learning SPR-transformed system. Property (5) follows from Theorem 3. The rank condition from Lemma 3 ensures that the PI observer construction is feasible. □

Remark 6. The framework achieves independent learning with modular stability guarantees. Agents train without network connectivity (enabling privacy and scalability), then gain stability guarantees through post-learning transformation. This separation contrasts with existing multi-agent RL methods that require network coupling during training.

Remark 7. The framework trades local optimality for modular stability. Locally optimal controllers K_i^* become conservative when interconnected because they were designed for $v_i = 0$. Theorem 3 quantifies this trade-off, showing that better local controllers reduce the performance penalty even after degradation.

SIMULATIONS

Simulation setup

We validate the framework on a network of 10 heterogeneous agents with second-order dynamics $\dot{x}_i = A_i x_i + B_i u_i$ where $x_i \in \mathbb{R}^2$ and $u_i \in \mathbb{R}$. Each agent has an output $y_i = [1,0]x_i$ and seeks to minimize the local LQR cost with $Q = \text{diag}(100,1)$ and $R = 1$. The agents have diverse eigenvalues: [0.2, -0.5] (Agent 1, unstable real), [0.2 $\pm$ 0.98j] (Agent 2, unstable oscillatory), [-0.3 $\pm$ 0.5j], [-0.4 $\pm$ 0.63j], [-0.25 $\pm$ 0.89j] (Agents 3-5, stable oscillatory), [-0.25, -0.25], [-0.5, -0.5], [-0.2, -0.8], [-0.3, -0.6], [-0.4, -0.5] (Agents 6-10, stable damped). This diversity demonstrates the framework's robustness: only stabilizability is required, not stability of the open-loop systems.

Nominal models from system identification, linearization, or engineering estimates introduce errors. LQR controllers designed from these nominal models are suboptimal for the true system. For each agent, we create a nominal model $A_{nom,i}$ by multiplying each entry of the true A_i by a random variable uniformly distributed over [0.8, 1.2], representing 20% model uncertainty. Following the IRL policy iteration algorithm in the Preliminaries section, we compute an initial controller $K_{nom,i}$ for each agent by solving the LQR problem for $A_{nom,i}$. We then apply policy iteration starting from this nominal controller to refine it using trajectory data from the true system. The PI observer gains are designed using $A_{nom,i}$ following the PI observer construction in the Preliminaries section. This setup demonstrates the framework's robustness to model uncertainty.

The network is tested on two topologies. The ring topology uses directed coupling $1 \to 2 \to \cdots \to 10 \to 1$ with an asymmetric Laplacian. The path topology uses undirected coupling $1 \leftrightarrow 2 \leftrightarrow \cdots \leftrightarrow 10$ with a symmetric Laplacian. Both topologies are strongly connected as required by Theorem 2. The coupling gain is $\alpha = 0.5$. Figure 1 illustrates both communication structures.

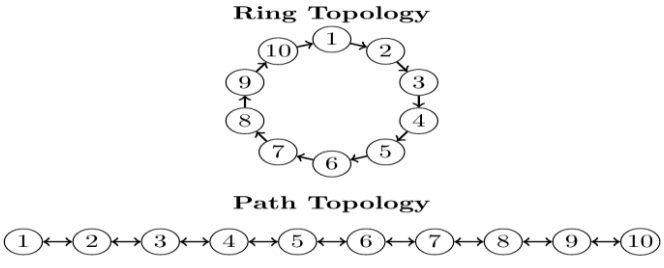

Figure 1. Communication topologies. Ring: directed coupling $1 \to 2 \to \cdots \to 10 \to 1$ (unidirectional, all nodes degree 1). Path: undirected coupling $1 \leftrightarrow 2 \leftrightarrow \cdots \leftrightarrow 10$ (bidirectional, interior nodes degree 2). Both use uniform coupling gain $\alpha = 0.5$.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Four controller types are compared. The IRL controller applies policy iteration starting from $K_{nom,i}$ until convergence. The LQR controller is the optimal controller for the true A_i. The Random controller uses stable random pole placement. The Aggressive controller uses stable pole placement with poles near -10.

We perform 100 Monte Carlo trials for each controller and topology combination. Initial conditions are drawn with a radius uniformly distributed over (0, 1] and an angle uniformly distributed over $[0, 2\pi)$. The simulation horizon is 30 seconds with a time step of 0.01 seconds. Agents are coupled from $t = 0$ with no delay. Three performance metrics are measured: convergence time (time to reach $\max_i \| x_i(t) \| < 0.001$), cumulative control effort $\sum_i \int_0^T \| u_i(t) \|^2 \, dt$, and final synchronization error $\max_{i,j} \| \gamma_i(T) - \gamma_j(T) \|$.

RESULTS

Table 1. Statistical Comparison vs LQR (p-values from t-tests, $n = 100$)

Controller	Sync Error	Effort	Peak Dev
IRL	$p = 1.000$	$p = 1.000$	$p = 1.000$
Random	$p < 0.001$	$p < 0.001$	$p < 0.001$
Aggressive	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table 2. Statistical Comparison vs LQR (p-values from t-tests, $n = 100$)

Controller	Sync Error	Effort	Peak Dev
IRL	$p = 1.000$	$p = 1.000$	$p = 1.000$
Random	$p < 0.001$	$p < 0.001$	$p < 0.001$
Aggressive	$p < 0.001$	$p < 0.001$	$p < 0.001$

Table 1 and Table 2 show a statistical comparison between each controller type and the LQR baseline. IRL achieves statistical equivalence to LQR with $p = 1.000$ for synchronization error, control effort, and peak deviation in both topologies. This validates Theorem 4 property (1): each learned controller K_i^* achieves local optimality for the isolated system.

Additionally, the learned controllers converge rapidly from the nominal initialization despite 2 agents having unstable open-loop dynamics, with an error from the true optimal ranging from 2×10^{-11} to 4×10^{-7}, well within the specified convergence tolerance. The framework demonstrates robustness: IRL achieves optimality despite 20% model uncertainty in nominal models and the presence of unstable open-loop agents. In contrast, Random and Aggressive controllers show statistically significant differences from LQR with $p < 0.001$ for all metrics.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

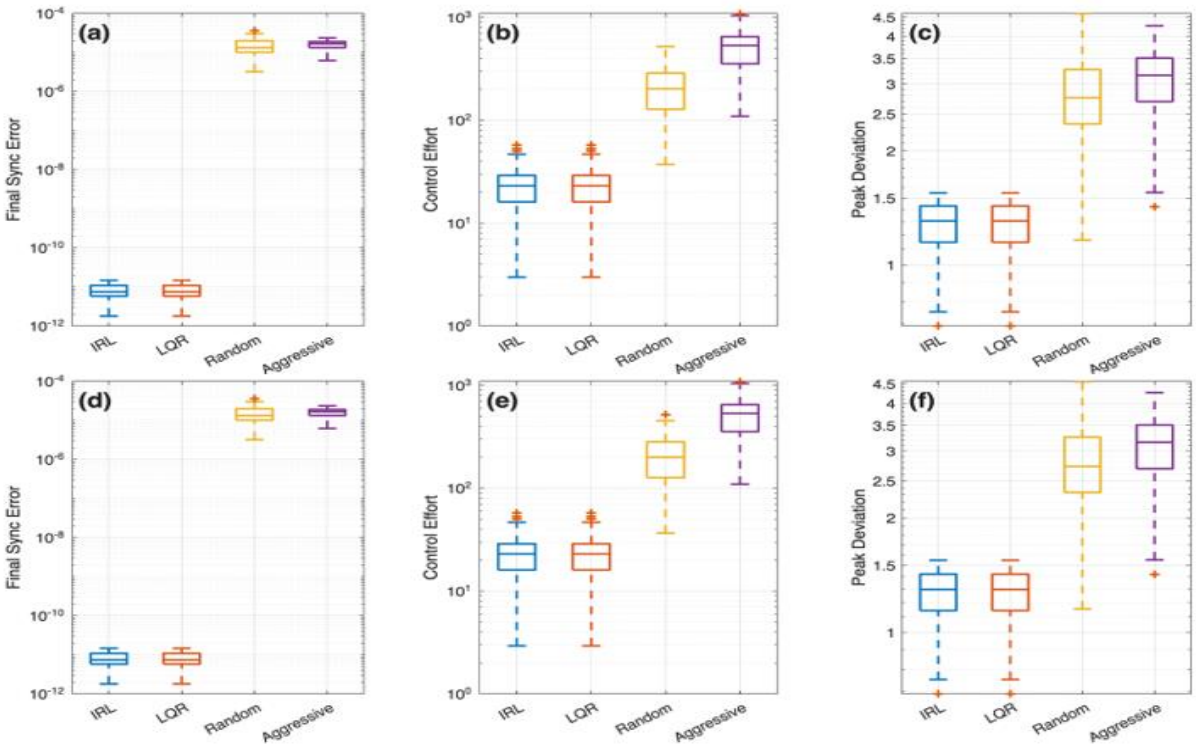

Figure 2. Performance metrics. Box plots over 100 trials showing median, quartiles, and outliers. (a) Ring final synchronization error: IRL/LQR achieve 10^{-12} (machine precision), Random/Aggressive 10^{-5}. (b) Ring control effort: IRL/LQR median 23, Random 203 (9×), Aggressive 535 (23×). (c) Ring peak deviation. (d)-(f) Path metrics show identical distributions to ring topology. Controller quality substantially affects performance despite identical asymptotic behavior.

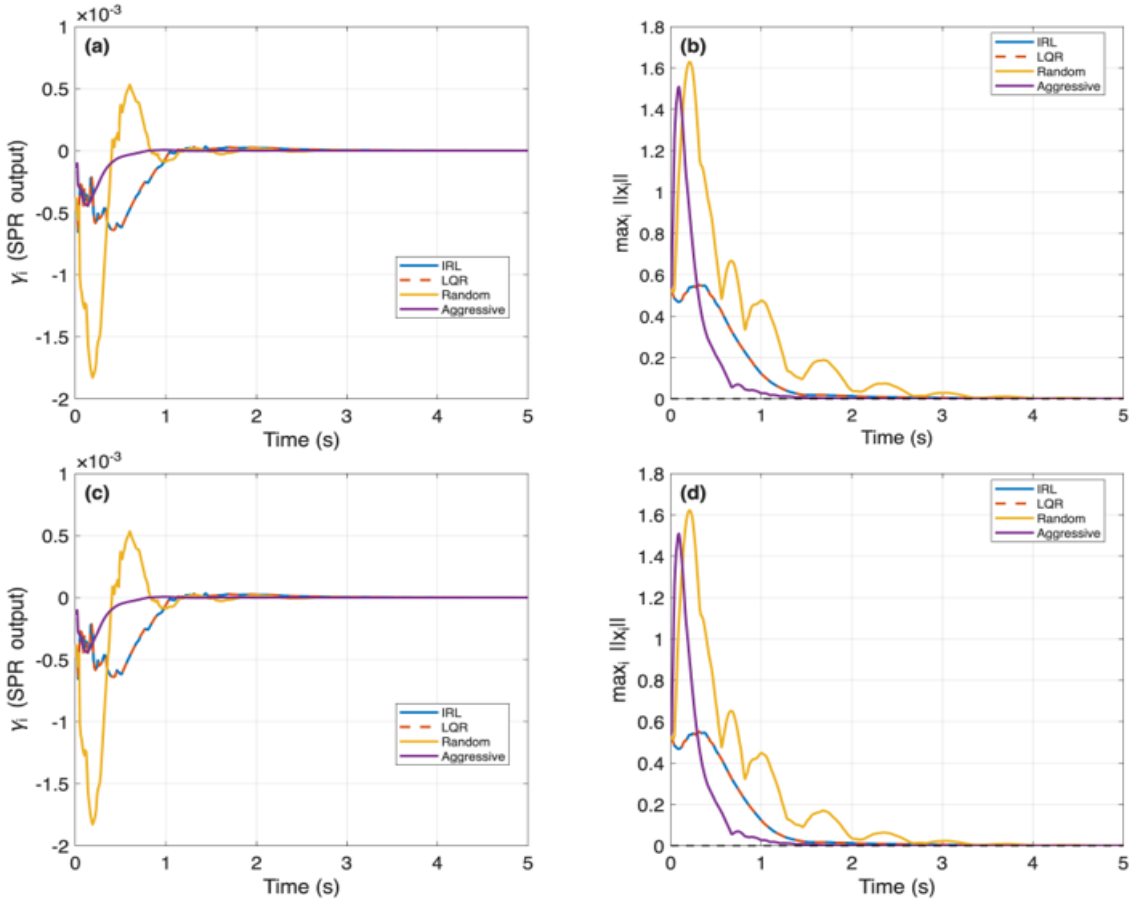

Figure 3. Network dynamics. (a) Ring: Mean SPR output $\gamma_i(t)$ across 10 agents, where each agent's trajectory is the element-wise median across 100 trials, for four controller types. Mean outputs are regulated to zero within 2-3s. (b) Ring: Maximum state norm $max_i \parallel x_i(t) \parallel$ (element-wise median across 100 trials) converges to zero. (c) Path: Mean SPR output shows identical regulation. (d) Path: Maximum state norm shows identical convergence. Four curves per panel: IRL, LQR, Random, Aggressive. Validates Theorem 2 properties (2) and (3): regulation to zero and global asymptotic stability.

Figure 2 panels (a) and (d) show the final synchronization error $\max_{i,j} \parallel \gamma_i(T) - \gamma_j(T) \parallel$ at $T = 30$s. For the ring topology, IRL and LQR achieve median synchronization error of 7.5×10^{-12}, while Random shows a larger error of 1.3×10^{-5} and Aggressive 1.7×10^{-5}. The path topology shows nearly identical results: IRL/LQR error 7.5×10^{-12}, Random error 1.3×10^{-5}, and Aggressive error 1.7×10^{-5}. These results validate Theorem 2 property (1): output synchronization $\lim_{t\to\infty} \parallel \gamma_i(t) - \gamma_j(t) \parallel = 0$. All controllers achieve synchronization asymptotically, but differ in convergence speed, as discussed below.

Figure 3 panels (a) and (c) show SPR output $\gamma_i(t)$ averaged across 10 agents, where each agent's trajectory is the element-wise median across 100 trials. All controllers achieve regulation to zero: the mean SPR outputs converge to zero within 2-3 seconds. This validates Theorem 2 property (2): regulation to zero $\lim_{t\to\infty}\gamma_i(t) = 0$. The regulation behavior is nearly identical between ring and path topologies.

Figure 3 panels (b) and (d) show the maximum state norm $\max_i \parallel x_i(t) \parallel$ across all agents, where each agent's trajectory is the element-wise median across 100 trials. All trajectories converge to zero, demonstrating global asymptotic stability of the origin. This validates Theorem 2 property (3): the origin $\xi = 0$ is globally asymptotically stable. The stability behavior is nearly identical between ring and path topologies. All 800 trials across both topologies and all controller types achieved regulation and stability. This confirms the framework's guarantees depend only on strong connectivity, not specific topological structure.

Figure 2 panels (b) and (e) quantify control effort differences $\sum_i \int_0^T \parallel u_i(t) \parallel^2 dt$. For the ring topology, IRL and LQR achieve median effort of 23.1, while Random shows higher effort of 202.5 (9× worse) and Aggressive 535.0 (23× worse). The path topology shows nearly identical results: IRL/LQR median effort 23.0, Random 199.7, and Aggressive 534.7. Since all controllers converge to equilibrium well before $T = 30$s, the measured finite-horizon effort effectively equals the infinite-horizon cost. LQR and IRL were designed to minimize this cost, which includes the control effort term $u^T Ru$. After network coupling, they lose strict optimality, but remain qualitatively better than Random and Aggressive, which were never designed for cost minimization. These differences are statistically significant as shown at the beginning of the results.

Figure 2 panels (c) and (f) show peak deviation of states from equilibrium during the simulation $\max_{t\in[0,T]}\max_i \parallel x_i(t) \parallel$. IRL and LQR show median peak deviation around 1.3-1.5, while Random and Aggressive show higher peaks of 2.5-3.3 and 3.0-3.5, respectively. Better controllers keep transient responses closer to equilibrium. Statistical significance is confirmed in the results.

Table 3. Performance Degradation Bound Validation (10-Agent Ring)

Controller	α_{net}	α_{max}^{local}	β_{KR}	$\lambda_{max}(L_s)$	Bound	Slack
IRL	-0.66	3.9	10.0	0.20	9.6	10
LQR	-0.66	3.9	10.0	0.20	9.6	10
Random	-0.10	36.5	51.6	0.20	53.9	54
Aggressive	-0.04	86.2	125.1	0.20	124.4	124
Upper bound: $\alpha_{net} \le \alpha_{max}^{local} + \kappa_p^2\lambda_{max}(L_s)\beta_B\sqrt{2}(\beta_{KR} + \Delta_{K-lqr})$						

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Ring: unidirectional coupling, all nodes degree 1. $\kappa_p = 1.0$.

Table 4. Performance Degradation Bound Validation (10-Agent Path)

Controller	α_{net}	α_{max}^{local}	β_{KR}	$\lambda_{max}(L_s)$	Bound	Slack
IRL	-0.66	3.9	10.0	0.39	14.9	16
LQR	-0.66	3.9	10.0	0.39	14.9	16
Random	-0.10	36.5	51.6	0.39	70.4	71
Aggressive	-0.04	86.2	125.1	0.39	160.7	161
Upper bound: $\alpha_{net} \leq \alpha_{max}^{local} + \kappa_p^2 \lambda_{max}(L_s)\beta_B\sqrt{2}(\beta_{KR} + \Delta_{K-lqr})$						
Path: bidirectional coupling, interior nodes degree 2. $\kappa_p = 1.0$.						

Table 3 and Table 4 explain the synchronization error differences observed in Figure 2 panels (a) and (d) through the Theorem 3 bound. For each controller type, we measure the network spectral abscissa α_{net} and verify that the upper bound holds. Controller quality affects network performance through two mechanisms: better controllers create more stable augmented systems (lower α_{max}^{local}) and minimize the coupling penalty (lower β_{KR}), as shown in the table values. IRL and LQR have $\alpha_{max}^{local} = 3.9$ and $\beta_{KR} = 10.0$, while Random degrades to $\alpha_{max}^{local} = 36.5$ and $\beta_{KR} = 51.6$, and Aggressive further degrades to $\alpha_{max}^{local} = 86.2$ and $\beta_{KR} = 125.1$. Better controllers with lower α_{max}^{local} and β_{KR} produce a more negative network α_{net}: IRL/LQR achieve $\alpha_{net} = -0.66$, Random achieves -0.10, and Aggressive achieves -0.04.

The network spectral abscissa determines the convergence speed. At the finite measurement time $T = 30s$, the observed synchronization error reflects convergence progress: IRL/LQR with $\alpha_{net} = -0.66$ have nearly completed convergence with an error $\sim 10^{-12}$, while Random with $\alpha_{net} = -0.10$ and Aggressive with $\alpha_{net} = -0.04$ are still converging with an error $\sim 10^{-5}$, as shown in Figure 2 panels (a) and (d).

The measured α_{net} values are identical across topologies despite different Laplacian structures. However, the bounds differ: path topology bounds are 1.56× larger than ring bounds due to the topology-dependent term $\kappa_p^2 \lambda_{max}(L_s)$ in Theorem 3. Both topologies have $\kappa_p = 1$, but the path's bidirectional coupling creates $\lambda_{max}(L_s) = 0.39$ compared to the ring's unidirectional coupling with $\lambda_{max}(L_s) = 0.20$. The path Laplacian has interior agents with degree 2, while the ring has all agents with degree 1, directly causing the 2× larger maximum eigenvalue. Despite these bound differences, actual performance differs by less than 2% between topologies. This demonstrates the bound's conservatism: slack ranges from 10× to 161×, with topology affecting conservatism while actual behavior remains topology-agnostic.

CONCLUSION

This paper presented a framework that enables agents to learn controllers independently through integral reinforcement learning, then connect with modular stability guarantees. The framework addresses a gap: existing modular control methods require analytical models and cannot handle learned controllers, while multi-agent learning methods require network coupling during training. The proposed three-stage approach—independent IRL training, post-learning SPR transformation via a PI observer, and Laplacian network coupling—achieves both independent learning and plug-and-play coordination.

The theoretical contributions establish conditions under which this integration succeeds. Lemma 3 provides necessary and sufficient conditions for PI observer feasibility, requiring observability and sufficient input-output coupling at steady state. Theorem 2 proves that SPR-transformed agents achieve output synchronization, regulation to zero, and global asymptotic stability under strongly connected communication graphs. Unlike

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

standard passivity-based methods that synchronize to arbitrary constants, the $Q > 0$ condition required by strict positive realness ensures that the invariant set equals the origin, guaranteeing regulation to zero. Theorem 3 bounds the network spectral abscissa, which determines the convergence rate and the stability margin, in terms of the graph topology, the agent dynamics, and the controller quality.

Simulations on 10 heterogeneous agents with diverse dynamics validated the framework. Monte Carlo experiments across 800 trials demonstrated that IRL achieves statistical equivalence to optimal LQR controllers despite 20% model uncertainty and unstable open-loop agents. All trials achieved synchronization, regulation, and stability on both ring and path topologies, confirming topology-agnostic guarantees. Performance metrics showed that controller quality substantially affects transient behavior and control effort, with optimal controllers requiring 9-23× lower effort than poorly-designed controllers while all converge to the same equilibrium.

The framework trades local optimality for modular guarantees. Controllers optimized for isolated operation become conservative when connected because they were designed without network coupling. Independent training necessitates this trade-off: it enables privacy and scalability but prevents anticipating network effects during optimization. The performance bound quantifies this cost, showing that better local controllers reduce the penalty.

Future work includes extending the framework to communication delays. Sequential learning approaches that refine controllers after connection could recover network optimality while maintaining stability during adaptation. Adaptive methods that adjust SPR transformation parameters online could reduce conservatism. Investigating tighter performance bounds or alternative passivity certificates may improve the optimality-stability trade-off.

Conflict of Interest: The authors declare no conflict of interest.

APPENDIX A: PROOF OF LEMMA 3

We prove the PI observer feasibility condition.

Reformulation as output injection

The PI observer error dynamics can be written as

$$\frac{d}{dt}\begin{bmatrix} e \\ w \end{bmatrix} = \begin{bmatrix} A - L_P C & B \\ -L_I C & 0 \end{bmatrix}\begin{bmatrix} e \\ w \end{bmatrix} =: A_z \begin{bmatrix} e \\ w \end{bmatrix}.$$

Define the augmented system matrices

$$A_{aug} := \begin{bmatrix} A & B \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)},$$
$$C_{aug} := [C \quad 0] \in \mathbb{R}^{p\times(n+m)},$$

and the augmented gain

$$L_{aug} := \begin{bmatrix} L_P \\ L_I \end{bmatrix} \in \mathbb{R}^{(n+m)\times p}.$$

Then the error dynamics matrix can be written as

$$A_z = A_{aug} - L_{aug} C_{aug}.$$

This reformulation shows that stabilizing A_z is equivalent to placing the eigenvalues of A_{aug} by output injection L_{aug}. By pole placement theory, this is possible if and only if (A_{aug}, C_{aug}) is observable.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

PBH observability test

We apply the Popov-Belevitch-Hautus (PBH) test for observability.

Lemma 5 (PBH Test). A pair $(\mathcal{A}, \mathcal{C})$ with $\mathcal{A} \in \mathbb{C}^{N\times N}$ and $\mathcal{C} \in \mathbb{C}^{q\times N}$ is observable if and only if $\operatorname{rank}\begin{bmatrix} sI - \mathcal{A} \\ \mathcal{C} \end{bmatrix} = N$ for all $s \in \mathbb{C}$.

Applying this to (A_{aug}, C_{aug}), observability requires

$$\operatorname{rank}\begin{bmatrix} sI - A_{aug} \\ C_{aug} \end{bmatrix} = n + m \quad \text{for all } s \in \mathbb{C}. \quad (12)$$

Case 1: $s \neq 0$

For $s \neq 0$, the PBH matrix is

$$\begin{bmatrix} sI - A_{aug} \\ C_{aug} \end{bmatrix} = \begin{bmatrix} sI_n - A & -B \\ 0 & sI_m \\ C & 0 \end{bmatrix}.$$

This matrix has $n + m$ columns and $n + m + p$ rows. We need to show it has full column rank $n + m$.

Consider the last m columns. These form the submatrix

$$\begin{bmatrix} -B \\ sI_m \\ 0 \end{bmatrix}.$$

Since $s \neq 0$, the block sI_m has full rank m, so the last m columns are linearly independent.

For the matrix to have full column rank $n + m$, we need the first n columns to have rank n when restricted to a complementary subspace. The first n columns form

$$\begin{bmatrix} sI_n - A \\ 0 \\ C \end{bmatrix}.$$

Removing the zero rows (middle m rows) gives

$$\begin{bmatrix} sI_n - A \\ C \end{bmatrix}.$$

By the PBH test, this has rank n for all $s \in \mathbb{C}$ if and only if (A, C) is observable. Therefore, condition (12) holds for all $s \neq 0$ if and only if (A, C) is observable.

Case 2: $s = 0$

For $s = 0$, the PBH matrix becomes

$$\begin{bmatrix} 0 \cdot I - A_{aug} \\ C_{aug} \end{bmatrix} = \begin{bmatrix} -A & -B \\ 0 & 0 \\ C & 0 \end{bmatrix}.$$

The middle m rows are all zeros and do not contribute to the rank. Removing them and multiplying the first n rows by -1 gives the equivalent rank condition

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

$$\operatorname{rank}\begin{bmatrix} A & B \\ C & 0 \end{bmatrix} = n + m. \quad (13)$$

Equivalence to $\operatorname{rank}(CA^{-1}B) = m$

When A is invertible, we show that (13) is equivalent to $\operatorname{rank}(CA^{-1}B) = m$ using Gaussian elimination.

Starting with the block matrix in (13), multiply the first n rows by A^{-1}:

$$\begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \sim \begin{bmatrix} I_n & A^{-1}B \\ C & 0 \end{bmatrix}.$$

Next, subtract C times the first n rows from the last p rows:

$$\begin{bmatrix} I_n & A^{-1}B \\ C & 0 \end{bmatrix} \sim \begin{bmatrix} I_n & A^{-1}B \\ 0 & -CA^{-1}B \end{bmatrix}.$$

The rank of this matrix is $n + \operatorname{rank}(-CA^{-1}B) = n + \operatorname{rank}(CA^{-1}B)$.

For full column rank $n + m$, we require

$$n + \operatorname{rank}(CA^{-1}B) = n + m,$$

which simplifies to $\operatorname{rank}(CA^{-1}B) = m$.

CONCLUSION

Combining both cases, (A_{aug}, C_{aug}) is observable if and only if:

1. (A, C) is observable (from the $s \neq 0$ case), and
2. $\operatorname{rank}(CA^{-1}B) = m$ when A is invertible (from the $s = 0$ case).

Since the spectrum of A_z can be assigned arbitrarily by choice of L_P and L_I if and only if (A_{aug}, C_{aug}) is observable, these conditions are necessary and sufficient for the existence of gains making A_z Hurwitz. This completes the proof of Lemma 3.

Appendix B: Weighted Hermitian Identity For Directed Graphs

This appendix establishes the weighted Hermitian identity used in the proof of Theorem 3.

Commutation identities for block-diagonal scaling

For block-diagonal matrices $\mathcal{B} = \operatorname{diag}(B_{0,1}, \dots, B_{0,N})$ and $\mathcal{C} = \operatorname{diag}(C_{0,1}, \dots, C_{0,N})$, and diagonal weighting $S = D_p^{1/2} = \operatorname{diag}(\sqrt{p_1}, \dots, \sqrt{p_N})$, the following commutation properties hold:

Lemma 6 (Block-Diagonal Commutation). For appropriately sized identity matrices on input, state, and output spaces:
$(S \otimes I_{state})\mathcal{B} = \mathcal{B}(S \otimes I_{input})$,
$(S \otimes I_{output})\mathcal{C} = \mathcal{C}(S \otimes I_{state})$.

Proof. On block i, the scalar $s_i = \sqrt{p_i}$ commutes with any matrix: $(s_i I)B_i = B_i(s_i I)$ and $(s_i I)C_i = C_i(s_i I)$. Lifting via Kronecker product $\otimes I$ preserves these equalities across all blocks. □

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Change of variables in weighted operator norms

Lemma 7 (Invariance under Change of Variables). Let $S = D_p^{1/2}$ and define the weighted norm on $\mathbb{R}^n$ by $\| z \|_{D_p} := \| (S \otimes I)z \|_2$. For any linear operator $\mathcal{C}$, its induced norm in the weighted metric is $\| \mathcal{C} \|_{D_p} := \sup_{\|z\|_{D_p}=1} \| \mathcal{C}z \|_{D_p}$. Under the change of variables $\tilde{z} := (S \otimes I)z$, $\| \mathcal{C} \|_{D_p} = \sup_{\|\tilde{z}\|_2=1} \| (S \otimes I)\mathcal{C}(S^{-1} \otimes I)\tilde{z} \|_2 = \| (S \otimes I)\mathcal{C}(S^{-1} \otimes I) \|_2$.

Proof. By definition, $\| \mathcal{C} \|_{D_p} = \sup_{\|(S\otimes I)z\|_2=1} \| (S \otimes I)\mathcal{C}z \|_2$. Set $\tilde{z} := (S \otimes I)z$. Since $S \otimes I$ is invertible, this mapping is bijective with $\| (S \otimes I)z \|_2 = \| \tilde{z} \|_2$ and $(S \otimes I)\mathcal{C}z = (S \otimes I)\mathcal{C}(S^{-1} \otimes I)\tilde{z}$. The constraint $\| z \|_{D_p} = 1$ becomes $\| \tilde{z} \|_2 = 1$, and the supremum over the weighted unit sphere equals the supremum over the Euclidean unit sphere of the similarity-transformed operator. □

Weighted Hermitian bound for heterogeneous agents

Lemma 8 (Weighted Hermitian Identity for Heterogeneous Agents). Let $\mathcal{B} = \mathrm{diag}(B_{0,1}, \dots, B_{0,N})$ and $\mathcal{C} = \mathrm{diag}(C_{0,1}, \dots, C_{0,N})$ be block-diagonal matrices with potentially different block sizes. Define the weighted Hermitian operator $\mathrm{He}_{D_p \otimes I}(M) := \frac{1}{2}((D_p \otimes I)M + M^T(D_p \otimes I))$ and the weighted operator norm $\| X \|_{D_p} := \| (S \otimes I)X(S^{-1} \otimes I) \|_2$ where $S = D_p^{1/2}$.

For $M = \mathcal{B}(L \otimes I)\mathcal{C}$, $\| \mathrm{He}_{D_p \otimes I}(M) \|_{D_p} \le \lambda_{\max}(L_s) \| D_p^{1/2}\mathcal{B} \|_2 \| \mathcal{C}D_p^{-1/2} \|_2$. (14)

Consequently, converting to Euclidean norm via $\| X \|_2 \le \kappa_p \| X \|_{D_p}$, $\| \mathrm{He}(M) \|_2 \le \kappa_p^2 \lambda_{\max}(L_s)\beta_B\beta_C$ (15) where $\beta_B = \max_i \| B_{0,i} \|_2$, $\beta_C = \max_i \| C_{0,i} \|_2$, and $\kappa_p = \sqrt{\max_i p_i / \min_i p_i}$.

Proof. For test vector z with $\| z \|_{D_p} = 1$, define $u = (S \otimes I)\mathcal{B}^T z$ and $v = (S^{-1} \otimes I)\mathcal{C}z$. By the commutation identities from Lemma 6, the quadratic form satisfies

$$z^T \mathrm{He}_{D_p \otimes I}(M) z = \mathrm{Re}(u^T (SLS^{-1} \otimes I)v).$$

Define $M_s = \mathrm{He}(SLS^{-1}) = S^{-T}L_sS^{-1} \succcurlyeq 0$. Applying Cauchy-Schwarz with respect to the semi-inner product induced by $M_s \otimes I$ yields

$$\begin{aligned} \mathrm{Re}(u^T (SLS^{-1} \otimes I)v) &\le \sqrt{(u^T(M_s \otimes I)u)(v^T(M_s \otimes I)v)} \\ &\le \lambda_{\max}(M_s) \| u \|_2 \| v \|_2. \end{aligned}$$

Since $\| u \|_2 \le \| S\mathcal{B} \|_2 \| z \|_2 = \| D_p^{1/2}\mathcal{B} \|_2 \| z \|_2$ and $\| v \|_2 \le \| \mathcal{C}S^{-1} \|_2 \| z \|_2 = \| \mathcal{C}D_p^{-1/2} \|_2 \| z \|_2$, and $\lambda_{\max}(M_s) = \lambda_{\max}(S^{-T}L_sS^{-1}) \le \| S^{-1} \|_2^2 \lambda_{\max}(L_s)$, we obtain

$$|z^T \mathrm{He}_{D_p \otimes I}(M) z| \le \| S^{-1} \|_2^2 \lambda_{\max}(L_s) \| D_p^{1/2}\mathcal{B} \|_2 \| \mathcal{C}D_p^{-1/2} \|_2 \| z \|_2^2.$$

Working in the weighted norm where $\| z \|_{D_p} = 1$ implies $\| z \|_2 \le \| S^{-1} \|_2$, and passing to the induced operator norm yields

$$\| \mathrm{He}_{D_p \otimes I}(M) \|_{D_p} \le \lambda_{\max}(L_s) \| D_p^{1/2}\mathcal{B} \|_2 \| \mathcal{C}D_p^{-1/2} \|_2,$$

establishing (14).

For block-diagonal matrices, $\| D_p^{1/2}\mathcal{B} \|_2 \le \sqrt{\max_i p_i}\beta_B$ and $\| \mathcal{C}D_p^{-1/2} \|_2 \le \beta_C / \sqrt{\min_i p_i}$. Converting to Euclidean norm via $\| X \|_2 \le \kappa_p \| X \|_{D_p}$ gives

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

$$\| \mathrm{He}(M) \|_2 \leq \kappa_p \lambda_{\max}(L_s) \sqrt{\max_i p_i} \frac{\beta_B \beta_C}{\sqrt{\min_i p_i}}$$
$$= \kappa_p^2 \lambda_{\max}(L_s) \beta_B \beta_C,$$

establishing (15). □

Remark 8. The weighted Hermitian operator $\mathrm{He}_{D_p \otimes I}$ arises in Lyapunov analysis for directed graphs. For heterogeneous agents, the factorization identity from the homogeneous case does not hold, but the operator norm bound remains valid through the variational Cauchy-Schwarz approach.

REFERENCES

1. Etika Agarwal, Shuvomoy Das Gupta, Necmiye Ozay, and Petros Voulgaris. Distributed synthesis of local controllers for networked systems with arbitrary interconnection topologies. IEEE Transactions on Automatic Control, 66 (2): 683–698, 2021. doi: 10.1109/TAC.2020.3028593.
2. Murat Arcak. Passivity as a design tool for group coordination. IEEE Transactions on Automatic Control, 52 (8): 1380–1390, 2007. doi: 10.1109/TAC.2007.902733.
3. S. Beale and B. Shafai. Robust control system design with a proportional integral observer. International Journal of Control, 50 (1): 97–111, 1989.
4. W. Brewer. Kronecker products and matrix calculus in system theory. IEEE Transactions on Circuits and Systems, 25 (9): 772–781, 1978.
5. Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 55 (2): 895–943, 2022. doi: 10.1007/s10462-021-09996-w.
6. Changchun Hua, Kuo Liu, and Xinping Guan. Semi-global/global output consensus for nonlinear multiagent systems with time delays. Automatica, 103: 480–489, 2019. doi: 10.1016/j.automatica.2019.02.022.
7. Yutao Jiang, Weinan Gao, Jing Wu, Tianyou Chai, and Frank L. Lewis. Reinforcement learning and cooperative H_∞ output regulation of linear continuous-time multi-agent systems. Automatica, 150: 110581, 2023. doi: 10.1016/j.automatica.2022.10634.
8. Qing Jiao, Hamidreza Modares, Shengyuan Xu, Frank L. Lewis, and Kyriakos G. Vamvoudakis. Multi-agent zero-sum differential graphical games for disturbance rejection in distributed control. Automatica, 69: 24–34, 2016. doi: 10.1016/j.automatica.2016.02.002.
9. Ruiyang Jin, Zaiwei Chen, Yiheng Lin, Jie Song, and Adam Wierman. Approximate global convergence of independent learning in multi-agent systems. In Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS), volume 258 of Proceedings of Machine Learning Research, pages 2818–2826. PMLR, 2025.
10. Yu Kawano, Krishna Chaitanya Kosaraju, and Jacquelien M. A. Scherpen. Krasovskii and shifted passivity-based control. IEEE Transactions on Automatic Control, 66 (10): 4926–4932, 2021. doi: 10.1109/TAC.2020.3040252.
11. Yu Kawano, Alessio Moreschini, and Michele Cucuzzella. Krasovskii passivity for sampled-data stabilization and output consensus. IEEE Transactions on Automatic Control, 70 (2): 1038–1053, 2025. doi: 10.1109/TAC.2024.3438749.
12. Hassan K. Khalil. Nonlinear Control. Pearson, 3rd edition, 2015.
13. Hyungbo Kim, Hyungbo Shim, Juhoon Back, and Jin Heon Seo. Consensus of output-coupled linear multi-agent systems under fast switching network: Averaging approach. Automatica, 49 (1): 267–272, 2013. doi: 10.1016/j.automatica.2012.09.025.
14. Jin Gyu Lee and Hyungbo Shim. A tool for analysis and synthesis of heterogeneous multi-agent systems under rank-deficient coupling. Automatica, 120: 109118, 2020. doi: 10.1016/j.automatica.2020.109118.
15. Frank L. Lewis, Hongwei Zhang, Kristian Hengster-Movric, and Abhijit Das. Cooperative Control of Multi-Agent Systems: Optimal and Adaptive Design Approaches. Communications and Control Engineering. Springer-Verlag London, 2014. doi: 10.1007/978-1-4471-5574-4.

16. Jianguo Li, Hamidreza Modares, Tianyou Chai, Frank L. Lewis, and Lihua Xie. Off-policy reinforcement learning for synchronization in multiagent graphical games. IEEE Transactions on Neural Networks and Learning Systems, 28 (10): 2434–2445, 2017. doi: 10.1109/TNNLS.2016.2609500.
17. Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017), pages 6379–6390, 2017.
18. Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. Consensus and cooperation in networked multi-agent systems. Proceedings of the IEEE, 95 (1): 215–233, 2007. doi: 10.1109/JPROC.2006.887293.
19. Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. Applied Intelligence, 53 (11): 13677–13722, 2023. doi: 10.1007/s10489-022-04105-y.
20. Zhihua Qu and Marwan A. Simaan. Modularized design for cooperative control and plug-and-play operation of networked heterogeneous systems. Automatica, 50 (9): 2405–2414, 2014. doi: 10.1016/j.automatica.2014.07.003.
21. Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. Journal of Machine Learning Research, 21 (178): 1–51, 2020.
22. Jie Ren and Bahram Shafai. Transformation of linear systems into strictly positive real systems with a proportional-integral observer. In The 2025 Modeling, Estimation and Control Conference, 2025.
23. Wei Ren and Randal W. Beard. Consensus seeking in multiagent systems under dynamically changing interaction topologies. IEEE Transactions on Automatic Control, 50 (5): 655–661, 2005. doi: 10.1109/TAC.2005.846556.
24. Stefano Riverso, Marcello Farina, and Giancarlo Ferrari-Trecate. Plug-and-Play Model Predictive Control based on robust control invariant sets. Automatica, 50 (8): 2179–2186, 2014. doi: 10.1016/j.automatica.2014.06.005.
25. Georg S. Seyboth, Wei Ren, and Frank Allgöwer. Cooperative control of linear multi-agent systems via distributed output regulation and transient synchronization. Automatica, 68: 132–139, 2016. doi: 10.1016/j.automatica.2016.01.068.
26. B. Shafai and M. Saif. Proportional-integral observer in robust control, fault detection, and decentralized control of dynamic systems. In Robust Control and Fault Detection, volume 27, pages 13–43. Springer International Publishing, Cham, 2015.
27. Takashi Tanaka and Cedric Langbort. Retrofit control with approximate environment modeling. Automatica, 108: 108488, 2019. doi: 10.1016/j.automatica.2019.06.031.
28. Farzaneh Tatari, Mohammad Bagher Naghibi-Sistani, and Kyriakos G. Vamvoudakis. Distributed learning algorithm for non-linear differential graphical games. Transactions of the Institute of Measurement and Control, 39 (2): 173–182, 2017. doi: 10.1177/0142331215603791.
29. Kyriakos G. Vamvoudakis and Frank L. Lewis. Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton–Jacobi equations. Automatica, 47 (8): 1556–1569, 2011. doi: 10.1016/j.automatica.2011.03.005.
30. Kyriakos G. Vamvoudakis, Frank L. Lewis, and Greg R. Hudas. Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality. Automatica, 48 (8): 1598–1611, 2012. doi: 10.1016/j.automatica.2012.03.005.
31. Draguna Vrabie, Octavian Pastravanu, Murad Abu-Khalaf, and Frank L. Lewis. Adaptive optimal control for continuous-time linear systems based on policy iteration. Automatica, 45 (2): 477–484, 2009. doi: 10.1016/j.automatica.2008.08.017.
32. Peter Wieland, Rodolphe Sepulchre, and Frank Allgöwer. An internal model principle is necessary and sufficient for linear output synchronization. Automatica, 47 (5): 1068–1074, 2011. doi: 10.1016/j.automatica.2011.01.081.
33. Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre M. Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022), 2022.
34. Huaguang Zhang, He Jiang, Yanhong Luo, and Geyang Xiao. Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method. IEEE Transactions on Industrial Electronics, 64 (5): 4091–4100, 2017. doi: 10.1109/TIE.2016.2542134.