Hybrid Iteration ADP Algorithm to Solve Cooperative, Optimal Output Regulation Problem for Continuous-Time, Linear, Multiagent Systems: Theory and Application in Islanded Modern Microgrids With IBRs

Omar Qasem, *Graduate Student Member, IEEE*, Masoud Davari, *Senior Member, IEEE*, Weinan Gao, *Senior Member, IEEE*, Daniel R. Kirk, and Tianyou Chai, *Life Fellow, IEEE*

***Abstract*—In this article, we propose a novel adaptive dynamic programming (ADP) algorithm, named hybrid iteration (HI), to solve the cooperative, optimal output regulation problem (CO^2RP) for continuous-time, linear, multiagent systems. Unlike the traditional ADP algorithms, i.e., policy iteration (PI) and value iteration (VI), HI does not need an initial stabilizing control policy required by PI. At the same time, it maintains a faster convergence rate compared with VI. First, a model-based HI algorithm is proposed to solve the CO^2RP. Based on the proposed HI algorithm, a data-driven, adaptive, optimal controller is developed to solve the cooperative, adaptive, and optimal output regulation problem without using any information about the physics of the system. Instead, the states/input information collected along the trajectories of the dynamic system is employed. The proposed data-driven HI is applied to the adaptive, optimal secondary voltage control (also known as voltage restoration control) of an islanded modern microgrid based on the inverter-based resources. Compared with the VI and PI algorithms, comparative simulation results demonstrate that the proposed HI approach is significantly able to save the convergence time of the central processing unit (also known as CPU) deployed, reduce the number of learning iterations, and remove the requirement of the initial stabilizing control policy. Comparative experiments reveal the practicality and superiority of the proposed methodology.**

***Index Terms*—Adaptive dynamic programming (ADP), continuous-time, cooperative, linear, multiagent systems (MASs), optimal output regulation, reinforcement learning.**

Manuscript received 4 August 2022; revised 6 December 2022 and 20 January 2023; accepted 4 February 2023. Date of publication 28 February 2023; date of current version 10 July 2023. This work was supported in part by the Science and Technology Major Project 2020 of Liaoning Province under Grant 2020JH1/10100008, in part by the National Natural Science Foundation of China under Grant 61991404, and in part by 111 Project 2.0 (No. B08015). The work of Masoud Davari was supported in part by the U.S. National Science Foundation under ECCS-EPCN Awards #1808279 and #1902787 and in part by the dSPACE company, Verivolt company, the professional development part of Masoud Davari's Discovery and Innovation Award from the 2020–2021 University Awards of Excellence at Georgia Southern University, and his 2022 Impact Area Accelerator Grant partially funded by Georgia Southern University—at which all experiments were conducted. *(Corresponding author: Weinan Gao.)*

Omar Qasem and Daniel R. Kirk are with the College of Engineering and Science, Florida Institute of Technology, Melbourne, FL 32901 USA (e-mail: oqasem2021@my.fit.edu; dkirk@fit.edu).

Masoud Davari is with the Department of Electrical and Computer Engineering, Georgia Southern University (Statesboro Campus), Statesboro, GA 30460 USA (e-mail: davari@ualberta.ca).

Weinan Gao and Tianyou Chai are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China (e-mail: weinan.gao@nyu.edu; tychai@mail.neu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIE.2023.3247734.

Digital Object Identifier 10.1109/TIE.2023.3247734

I. Introduction

OVER the past decade, the cooperative output regulation problem (CORP) has been widely investigated due to its massive impact and importance in engineering applications, including distributed energy resources and inverter-based resources (IBRs) in modern microgrids (M^2Gs), connected and autonomous vehicles, cooperative robot reconnaissance, and satellite clustering [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. The CORP is employed in the design of distributed controllers to achieve asymptotic tracking of a class of reference inputs and disturbance rejection in leader–follower multiagent systems (MASs) while maintaining the stability of the closed-loop system [11]. Two major strategies are usually used in addressing the CORPs: feedback-feedforward control [12], [13] and the internal model principle [14], [15]. Existing studies of the leader–follower consensus problem usually assume the availability of the states of the exosystem (leader) to all other agents (followers); this assumption is restricted. In practice, such as in M^2Gs, communication channels are employed to transmit the reference's state information. However, failures or delays in transferring the exact information will destabilize the overall MAS [7]. Therefore, distributed observers were developed in [16] and [17] to estimate the leader's states, enabling the distributed controller to maintain the asymptotic input tracking and disturbance rejection. It is noteworthy that the output regulation problem is different from the output-feedback controller

design problems—see [18], [19], and [20]—in the sense that an exosystem generates the desired trajectories in the output regulation problems. In addition, the exosystem is considered in order to create disturbances to the system.

Besides the accessibility issue of the leader, the CORP by itself does not optimize the performance of the closed-loop MAS. Therefore, CORP and optimal control problem have to be addressed together. Dynamic programming (DP) [21] and Bellman's principle of optimality are the backbones of solving optimal control problems [22]. Since DP suffers from the curse of dimensionality in practice, learning-based methods, including reinforcement learning [23] and adaptive dynamic programming (ADP) [24], were developed in order to provide solutions to optimal control and decision-making problems without using the modeling information. ADP approaches are mainly built upon policy iteration (PI) and value iteration (VI). These strategies have been developed to control continuous-time and discrete-time systems [25], [26], [27], [28], [29], [30], [31], [32], [33], [34].

Based on Kleinman's PI [35] and its data-driven implementation [36], the gap between optimality and cooperative output regulation is filled in [37], wherein the cooperative, adaptive, optimal output regulation problem is solved by developing distributed feedback-feedforward controllers through originally combining the theories of ADP; adaptive, optimal control; and cooperative output regulation. The presented data-driven method converges to the optimal policy with a quadratic rate. However, its major drawback is that a stabilizing control policy is required to initiate the learning process—thus making it expensive to implement without the MASs dynamics. In order to overcome this barrier, and based on the results in [38], the cooperative, optimal output regulation problem (CO^{2}RP) is solved by VI in [19] and [39], wherein the initial stabilizing control policy is no more required. Nevertheless, VI requires more learning time and iterations for convergence, which restricts its use in practice due to the delay incurred in its learning process. The issue becomes more severe in applications requiring quick decisions and actions.

The main objective of this work is to develop an innovative ADP method, named hybrid iteration (HI), to solve the CO^{2}RP. The HI will enable us to bring the advantages of PI and VI together and remove their drawbacks simultaneously. Notably, the optimal solution for the CORP will be developed by ADP without an initial stabilizing control policy. In addition, the method will quadratically converge to the optimal solution. Therefore, this article's contributions are as follows.

1) As the first contribution, an innovative successive approximation algorithm is proposed in order to achieve cooperative, optimal output regulation of continuous-time, linear MASs (hereinafter referred to as MASs for ease of reference) by obtaining an optimal distributed control policy based on the knowledge of the system dynamics of each agent.
2) As the second contribution, an efficient, nonmodel-based learning method is developed to implement HI using the states/input data collected online along the MAS trajectories, with completely unknown model information.
3) Last but not least, this article is the first attempt to apply data-driven HI strategies to control islanded M^{2}Gs based on IBRs.

The rest of this article is organized as follows. In Section II, the problem statement is provided by recalling the formulation of the CO^{2}RP and the existing DP solutions. The new HI solution to the CO^{2}RP is proposed in Section III, wherein the model-based HI is first presented with a rigorous proof of convergence. Following that, this article's new data-driven HI algorithm for MASs is proposed in order to solve the CO^{2}RP discussed earlier. Additionally, the convergence analysis of this algorithm is provided. Simulation results are given in Section IV by applying data-driven HI to an application of M^{2}Gs and comparing its performance with PI and VI. Comparative experiments are also conducted in order to reveal the proposed methodology's practicality and effectiveness. Finally, Section V concludes this article.

Notations. Throughout this article, the following notations are denoted. $\mathbb{Z}_+$ denotes the set of nonnegative integers. I_n denotes the identity matrix of dimension n and $0_{n\times m}$ denotes an $n \times m$ zero matrix. $\mathbf{1}_n$ represents a vector of dimension n, with all its elements being 1. $||\cdot||$ denotes the induced norm operator for matrices and the Euclidean norm operator for vectors. $\otimes$ denotes the Kronecker product operator. Given a matrix $A \in \mathbb{R}^{n\times m}$, with $a_i \in \mathbb{R}^n$ are the columns of A, $\text{vec}(A) = [a_1^{\mathrm{T}}, a_2^{\mathrm{T}}, \ldots, a_m^{\mathrm{T}}]^{\mathrm{T}}$. Given $A^{\mathrm{T}}A$ is invertible, $A^{\dagger} = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ indicates the pseudoinverse of A. Given a vector $z \in \mathbb{R}^n$ and a symmetric matrix $P = P^{\mathrm{T}} \in \mathbb{R}^{m\times m}$, $\text{vecs}(P) = [p_{11}, 2p_{12}, \ldots, 2p_{1\,m}, p_{22}, 2p_{23}, \ldots, 2p_{m-1,m}, p_{mm}]^{\mathrm{T}} \in \mathbb{R}^{\frac{1}{2}m(m+1)}$, and $\text{vecv}(z) = \text{vecs}(zz^{\mathrm{T}})$. $P \succ (\succeq) 0$ and $P \prec (\preceq) 0$ indicate that P is the positive definite (semidefinite) and negative definite (semidefinite), respectively. $\text{diag}(c_1, c_2, c_3)$ denotes a diagonal matrix with c_1, c_2, and c_3 as its diagonal elements. For a matrix $A \in \mathbb{R}^{n\times n}$, $\sigma(A)$ denotes the spectrum of A. $\text{Re}(\lambda)$ represents the real part of the eigenvalue $\lambda \in \sigma(A)$. $\mathcal{J}^m$ denotes all $m \times m$ real symmetric matrices normed space, equipped with the induced matrix norm. $\mathcal{J}_+^m = \{P \in \mathcal{J}^m : P \succeq 0\}$.

II. Problem Statement and Formulation

This section presents the problem to be studied. Also, the preliminaries are shown as a preface for the proposed solution to the CO^{2}RP. Consider the following MAS:

$$\dot{v} = Ev \tag{1}$$

$$\dot{x}_i = A_i x_i + B_i u_i + D_i v \tag{2}$$

$$e_i = C_i x_i + F_i v, \quad i \in \mathcal{T} \tag{3}$$

where for each ith subsystem, $x_i \in \mathbb{R}^{n_i}$ is the state, $u_i \in \mathbb{R}^{m_i}$ is the control input, $e_i \in \mathbb{R}^{p_i}$ is the tracking error, and $v \in \mathbb{R}^q$ is the exostate in the exosystem (1). $D_i v$ and $-F_i v$ are generated by the exostate as the ith subsystem disturbance and the reference signal, respectively. Given the exosystem (1) and the plant (2) and (3), a diagraph $\mathcal{G}$ is defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the sets of nodes and edges are represented by $\mathcal{V} = \{0, 1, \ldots, N\}$ and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, respectively. The leader modeled via the exosystem (1) is represented by node 0. The followers described by (2)

and (3) are identified by a set of nodes $\mathcal{T} = \{1, 2, \ldots, N\}$. The adjacency matrix $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{(N+1)\times(N+1)}$ is defined by the weight a_{ij} such that $a_{ij} > 0$ if $(j, i) \in \mathcal{E}$; otherwise, $a_{ij} = 0$. $\mathcal{N}_i$ denotes the set of all the nodes j such that $(j, i) \in \mathcal{E}$. The Laplacian $\mathcal{L}$ of the diagraph $\mathcal{G}$ is defined as follows:

$$\mathcal{L} = \begin{bmatrix} \sum_{j=1}^{N} a_{0j} & -[a_{01}, \ldots, a_{0N}] \\ -\Delta \mathbf{1}_N & \mathcal{H} \end{bmatrix} \tag{4}$$

where $\Delta = \text{diag}(a_{10}, a_{20}, \ldots, a_{N0})$, and $\mathcal{H} = [h_{ij}] \in \mathbb{R}^{N\times N}$ is defined by $h_{ii} = (\sum_{j=0}^{N} a_{ij}) - a_{ii}$ and $h_{ij} = -a_{ij}\ \forall\, i \neq j$.

This article develops a data-driven learning algorithm for MASs to solve the CO^{2}RP, i.e., the tracking error of all the followers asymptotically converges to zero in an optimal sense with guaranteed stability. Second, the developed algorithm should not rely on the knowledge of an initial stabilizing policy or the system dynamics in the state equation. Third, this algorithm should converge in fewer iterations than what is required by VI.

In order to solve the CO^{2}RP and traditional CORP, some standard assumptions are considered for the system expressed in (1)–(3).

Assumption 1: The diagraph $\mathcal{G}$ contains a directed spanning tree with the node 0 as the root.

Assumption 2: The pair (A_i, B_i) is stabilizable $\forall i \in \mathcal{T}$.

Assumption 3: $\text{Rank}\left(\begin{bmatrix} A_i - \lambda I & B_i \\ C_i & 0 \end{bmatrix}\right) = n_i + p_i \forall \lambda \in \sigma(E)$, $i \in \mathcal{T}$.

First, recalling the solution to the CORP is presented in the following lemma.

Lemma 1 ([40], [41]): Under Assumptions 1–3, choose a large enough constant $\gamma > 0$ such that $\text{Re}(\lambda_i(E) - \gamma\lambda_j(\mathcal{H})) < 0\ \forall\ i = 1, 2, \ldots, q$ and any $j \in \mathcal{T}$. Let K_i be a stabilizing control gain matrix $\forall i \in \mathcal{T}$, and let $L_i = K_i X_i + U_i$, where the following regulator equations are solved by the pair (X_i, U_i):

$$A_i X_i + B_i U_i + D_i = X_i E \tag{5}$$

$$C_i X_i + F_i = 0. \tag{6}$$

Then, the following distributed control policy solves the CORP:

$$\dot{\zeta} = E\zeta_i + \gamma \left[\sum_{j\in\mathcal{N}_i} a_{ij}(\zeta_j - \zeta_i) + a_{i0}(v - \zeta_i) \right] \tag{7}$$

$$u_i = -K_i x_i + L_i \zeta_i\ \forall i \in \mathcal{T}. \tag{8}$$

In order to guarantee both the transient- and steady-state responses of the agents, one can design a control policy to achieve the (cooperative) output regulation in an optimal sense—i.e., the (cooperative) optimal output regulation [41], [42], [43]. In order to solve the CO^{2}RP, the following two optimization problems are to be addressed.

Problem 1:

$$\min_{(X_i, U_i)} \text{Tr}\left(X_i^{\mathrm{T}} \bar{Q}_i X_i + U_i^{\mathrm{T}} \bar{R}_i U_i\right) \tag{9}$$

$$\text{subject to (5) and (6)} \tag{10}$$

where $\bar{Q}_i = \bar{Q}_i^{\mathrm{T}} \succ 0,\ \bar{R}_i = \bar{R}_i^{\mathrm{T}} \succ 0$.

From [44], given any matrices D_i and F_i, Assumption 3 ensures that the regulators (5) and (6) are solvable $\forall i \in \mathcal{T}$. In addition, it has been shown in [43] that the Problem 1 has an optimal solution (X_i^*, U_i^*). Denote $\bar{x}_i := x_i - X_i^* v$ and $\bar{u}_i := u_i - U_i^* v$. Then, the following error system is obtained:

$$\dot{\bar{x}}_i = A_i \bar{x}_i + B_i \bar{u}_i \tag{11}$$

$$e_i = C_i \bar{x}_i. \tag{12}$$

Afterward, a constrained minimization problem described in Problem 2 is solved in order to obtain the optimal feedback controller in the form of $\bar{u}_i^* = -K_i^* \bar{x}_i$.

Problem 2:

$$\min_{\bar{u}_i} \int_0^\infty \left(\bar{x}_i^{\mathrm{T}} Q_i \bar{x}_i + \bar{u}_i^{\mathrm{T}} R_i \bar{u}_i\right) \mathrm{d}t \tag{13}$$

$$\text{subject to (11)} \tag{14}$$

where $Q_i = Q_i^{\mathrm{T}} \succeq 0$ and $R_i = R_i^{\mathrm{T}} \succ 0$, with the pair $(A_i, \sqrt{Q_i})$ being observable $\forall i \in \mathcal{T}$.

Assuming the system matrices are known, the CO^{2}RP is solvable in a suboptimal sense with the development of a distributed controller [41] described by

$$u_i^* = -K_i^* x_i + L_i^* \zeta_i\ \forall i \in \mathcal{T} \tag{15}$$

where K_i^* is computed by solving the Problem 2 such that $K_i^* = R_i^{-1} B_i^{\mathrm{T}} P_i^*$, where $P_i^* = (P_i^*)^{\mathrm{T}} \succ 0$ is the solution to the following algebraic Riccati equation:

$$P_i^* A_i + A_i^{\mathrm{T}} P_i^* + Q_i - P_i^* B_i R_i^{-1} B_i^{\mathrm{T}} P_i^* = 0. \tag{16}$$

It is noticeable that (16) is nonlinear in P_i^*, which makes it a difficult task to find P_i^* directly from (16). Two successive approximation methods are recalled in order to solve (16).

1) Policy Iteration: The PI algorithm is a successive approximation method for solving optimal control problems by alternating policy evaluation and policy improvement. Kleinman [35] has proposed a PI method to approximate P_i^* from (16), which is recalled in the following text.

a) *Policy Evaluation:* Solve $P_{i,k}$ from

$$\mathcal{Z}(P_{i,k}, K_{i,k}) = 0,\ k \in \mathbb{Z}_+ \forall i \in \mathcal{T} \tag{17}$$

where $\mathcal{Z}(\cdot, \cdot)$ is the Lyapunov operator defined by

$$\mathcal{Z}(P_i, K_i) = P_i(A_i - B_i K_i) + (A_i - B_i K_i)^{\mathrm{T}} P_i + Q_i + K_i^{\mathrm{T}} R_i K_i.$$

b) *Policy Improvement:* Update the control gain matrix by

$$K_{i,k+1} = R_i^{-1} B_i^{\mathrm{T}} P_{i,k}. \tag{18}$$

The following lemma summarizes the convergence of Kleinman's PI-based algorithm.

Lemma 2 ([35]): Let $K_{i,0} \in \mathbb{R}^{m_i \times n_i}$ be a stabilizing feedback gain matrix $\forall\ i \in \mathcal{T}$, the matrix $P_{i,k} \succ 0$ be the solution for the Lyapunov equation (17), and the control gain matrix $K_{i,k}$, for $k = 1, 2, \ldots$ is recursively defined by (18). Then, the following properties hold $\forall\ k \in \mathbb{Z}_+, i \in \mathcal{T}$.

i) The matrix $A_i - B_i K_{i,k}$ is Hurwitz.

ii) $P_i^* \preceq P_{i,k} \preceq P_{i,k-1}$.
iii) $\lim_{k\to\infty} K_{i,k} = K_i^*$, $\lim_{k\to\infty} P_{i,k} = P_i^*$.

2) Value Iteration: Different from PI, the VI [38], [39] relaxes the learning process by overcoming the barrier of knowing a stabilizing control gain matrix $K_{i,0}$ for each subsystem. VI starts from an arbitrary initial value matrix $P_{i,0} = (P_{i,0})^{\mathrm{T}} \succ 0$ for all $i \in \mathcal{T}$. The iterative scheme of VI is done by updating the value matrix until a predefined criterion is satisfied. The iterative process of VI is done as follows.

a) *Value Update:* Given $P_{i,0}$, update the value matrix using

$$P_{i,k+1} \leftarrow \epsilon_k \left(P_{i,k}A_i + A_i^{\mathrm{T}}P_{i,k} + Q_i - P_{i,k}B_iR_i^{-1}B_i^{\mathrm{T}}P_{i,k}\right) + P_{i,k}, \qquad k \in \mathbb{Z}_+, i \in \mathcal{T} \tag{19}$$

where $\{\epsilon_k\}_{k=0}^{\infty}$ is a deterministic sequence defined in Section III.

The proof of convergence of the VI algorithm can be found in Theorem 3.3 detailed in [38].

Remark 1: Unlike the research conducted in the state-of-the-art literature that has studied the output-feedback control (see [18], [19], and [20]), the formulation of the system described by (1)–(3) considers the presence of the exosystem (1)—which generates the reference signal $-F_i v$ to the output of the ith subsystem and simultaneously generates the disturbance $D_i v$ to the system (2).

III. Main Results

Although VI is less conservative than PI in the sense that no prior knowledge of an initial stabilizing control policy is required, it usually needs tremendously more learning iterations than PI to converge to an optimal solution. This matter makes the use of VI not applicable to the applications where a quick decision needs to be taken, such as M²Gs. This section proposes an entirely novel approximation strategy, namely HI, to bridge the performance gap between PI and VI for MASs. To be more specific, HI does not require any initial stabilizing control policy and requires much fewer iterations than VI to converge to the optimal solution. Initially, a model-based HI algorithm is introduced. Based upon that, a data-driven HI is developed such that the optimal control policy is learned from the information collected along the trajectories of the dynamic systems.

A. Model-Based HI for CO² RP

The HI algorithm differs from the traditional existing DP algorithms in the sense that the optimal solution convergence process gets completed in two phases. In the first phase, a stabilizing control policy is found for each $i \in \mathcal{T}$ using VI to avoid prior knowledge of a stabilizing control policy for each agent. Moreover, the phase is terminated once a stabilizing control gain matrix is obtained in order to prevent a large number of iterations incurred in VI. Using the obtained stabilizing control gain matrix from Phase 1, PI is leveraged and initiated until the convergence to the optimal solution is achieved. The details of both phases are discussed as follows.

1) Phase 1 to Find a Stabilizing Control Policy: Throughout Phase 1, the value matrix is iteratively updated until a stabilizing control policy is found. It is worth mentioning that stochastic approximation is used in this phase to solve the value update step. To begin with, $\{B_r\}_{r=0}^{\infty}$ is defined as a collection of nonempty interiors bounded sets, which satisfies

$$B_r \subset B_{r+1} \in \mathcal{J}_+^n,\ r \in \mathbb{Z}_+,\ \lim_{r\to\infty} B_r = \mathcal{J}_+^n$$

and $\hat{\varepsilon}_i > 0 \forall i \in \mathcal{T}$ is a small threshold. In addition, select a deterministic sequence $\{\epsilon_k\}_{k=0}^{\infty}$ such that

$$\epsilon_k > 0,\ \sum_{k=0}^{\infty} \epsilon_k = \infty,\ \lim_{k\to 0} \epsilon_k = 0.$$

Since the primary goal of Phase 1 is to seek a stabilizing policy for the ith subsystem and save iterations compared with VI, Q_i is replaced with $\hat{Q}_i \succ Q_i$ and the value update step is repeated until the first stabilizing control policy is obtained. With the obtained stabilizing control policy, Phase 1 is stopped, and that policy is then employed in the following phase.

2) Phase 2 to Explore the Optimal Control Policy for CO² RP: In Phase 2, the PI is initiated using the stabilizing control policy obtained from Phase 1. The policy evaluation in (17) and the policy improvement in (18) are repeated until the value matrix $P_{i,k}$ is close enough to $P_i^*, \forall i \in \mathcal{T}$. Algorithm 1 presents the detailed steps of the model-based HI algorithm, including its proof of convergence shown in Theorem 1.

Algorithm 1: Model-Based HI.

1: $i \leftarrow 1$
2: **repeat**
3: Select $\hat{\varepsilon}_i > 0$, $P_{i,0} = (P_{i,0})^{\mathrm{T}} \succ 0$, and $\hat{Q}_i = (\hat{Q}_i)^{\mathrm{T}} \succ Q_i$. $k, r \leftarrow 0$.
4: **repeat**
5: $\tilde{P}_{i,k+1} \leftarrow P_{i,k} + \epsilon_k(P_{i,k}A_i + A_i^{\mathrm{T}}P_{i,k} + \hat{Q}_i - P_{i,k}B_iR_i^{-1}B_i^{\mathrm{T}}P_{i,k})$
6: **if** $\tilde{P}_{i,k+1} \notin B_r$ **then** $P_{i,k+1} \leftarrow P_{i,0}$, $r \leftarrow r+1$.
7: **else** $P_{i,k+1} \leftarrow \tilde{P}_{i,k+1}$ **endif**
8: $k \leftarrow k+1$
9: **until** $(\tilde{P}_{i,k} - P_{i,k-1})/\epsilon_{k-1} \prec \hat{Q}_i$
10: **repeat**
11: $K_{i,k} \leftarrow R_i^{-1}B_i^{\mathrm{T}}P_{i,k-1}$
12: Solve $P_{i,k}$ from $\mathcal{Z}(P_{i,k}, K_{i,k}) = 0$. $k \leftarrow k+1$.
13: **until** $\|P_{i,k} - P_{i,k-1}\| < \hat{\varepsilon}_i$
14: $i \leftarrow i+1$
15: **until** $i = N+1$

Theorem 1: Consider the sequences $\{P_{i,k}\}_{k=0}^{\infty}$ and $\{K_{i,k}\}_{k=1}^{\infty}$ computed by Algorithm 1 for all $i \in \mathcal{T}$. There exists a $k^* \in \mathbb{Z}_+$ such that the inequalities of $\|P_{i,k^*} - P_i^*\| \le \hat{\varepsilon}_i$ and $\|K_{i,k^*} - K_i^*\| \le \hat{\varepsilon}_i$ hold in which $\hat{\varepsilon}_i > 0$ is a small threshold for any $i \in \mathcal{T}$.

Proof: Based on the article presented in [38], by repeating Steps 5–8, $\|P_{i,k} - P_i^*\| \to 0$ is achieved as $k \to \infty\ \forall\, i \in \mathcal{T}$. By defining $K_{i,k}$ expressed in (18), the convergence of the sequence $\{P_{i,k}\}_{k=0}^{\infty}$ implies the convergence of the sequence $\{K_{i,k}\}_{k=1}^{\infty}$. Therefore, with $k \to \infty$, $\|K_{i,k} - K_i^*\| \to 0$ for all $i \in \mathcal{T}$ can also achieved. The condition in Step 9, i.e.,

$(\tilde{P}_{i,k} - P_{i,k-1})/\epsilon_{k-1} \prec \hat{Q}_i$, is equivalent to

$$P_{i,k-1}A_i + A_i^{\mathrm{T}}P_{i,k-1} - P_{i,k-1}B_iR_i^{-1}B_i^{\mathrm{T}}P_{i,k-1} \prec 0$$

which implies the following inequality:

$$\begin{aligned} &P_{i,k-1}(A_i - B_iK_{i,k}) + (A_i - B_iK_{i,k})^{\mathrm{T}}P_{i,k-1} \\ &\quad \prec -K_{i,k}^{\mathrm{T}}R_iK_{i,k} \preceq 0 \end{aligned} \tag{20}$$

where $K_{i,k} = R_i^{-1}B_i^{\mathrm{T}}P_{i,k-1}$. Therefore, given any $\hat{Q}_i \succ 0$, a stabilizing control gain matrix $K_{i,k_a} = R_i^{-1}B_i^{\mathrm{T}}P_{i,k_a-1}$ can always be obtained such that the matrix $A_i - B_iK_{i,k_a}$ is Hurwitz when Steps 4–9 are finished at iteration $k_a \in \mathbb{Z}_+$. By starting with the stabilizing control gain K_{i,k_a} and considering [35], repeating Steps 10–13 will lead to the convergence to the optimal solution P_i^*. In other words, when Step 13 is satisfied, one can always find an iteration index $k^* \in \mathbb{Z}_+$ such that $\|P_{i,k^*} - P_i^*\| \leq \max\{\hat{\varepsilon}_i/\|R_i^{-1}B_i^{\mathrm{T}}\|, \hat{\varepsilon}_i\}$. This fact implies that $\|K_{i,k^*} - K_i^*\| \leq \|R_i^{-1}B_i^{\mathrm{T}}\|\|P_{i,k^*} - P_i^*\| \leq \hat{\varepsilon}_i$ for any $i \in \mathcal{T}$. The proof is, thus, completed. ■

B. Data-Driven, Cooperative, HI for CO^2 RP

This section extends the HI Algorithm 1 to a data-driven version, where the algorithm relies on the states/input information collected along the trajectories of each subsystem. First, the details of finding a stabilizing control policy using the online data are given. Afterward, the PI-based data-driven suboptimal controller designed in [41] is used with the stabilizing policy obtained in order to converge to the optimal solution.

Considering the ith subsystem, define $\bar{x}_{ij} = x_i - X_{ij}v$ for $0 \leq j \leq h_i + 1$, where $X_{i0} = 0_{n_i \times q}$, $X_{ij} \in \mathbb{R}^{n_i \times q}$ so that $C_iX_{i1} + F_i = 0$. The matrices X_{ij} for $2 \leq j \leq h_i + 1$, where $h_i = (n_i - p_i)q$ is the null space dimension of $I_q \otimes C_i$, are selected such that the basis for $\ker(I_q \otimes C_i)$ is formed by all the vectors $\mathrm{vec}(X_{ij})$. With the above definitions along with (1) and (2), the following differential equation is then obtained:

$$\dot{\bar{x}}_{ij} = A_{i,k}\bar{x}_{ij} + B_i(K_{i,k}\bar{x}_{ij} + u_i) + (D_i - S_i(X_{ij}))v \tag{21}$$

where the Sylvester map $S_i : \mathbb{R}^{n_i \times q} \to \mathbb{R}^{n_i \times q}$ satisfies $S_i(X) = XE - A_iX \; \forall \, X \in \mathbb{R}^{n_i \times q}$, and $A_{i,k} = A_i - B_iK_{i,k}$. For any two vectors $a(t) \in \mathbb{R}^{\boldsymbol{n}}, b(t) \in \mathbb{R}^{\boldsymbol{m}}$, and a sufficiently large $\rho \in \mathbb{Z}_+$, the following matrices are defined:

$$\begin{aligned} \delta_b &= \left[\mathrm{vecv}(b)|_{t_0}^{t_1}, \mathrm{vecv}(b)|_{t_1}^{t_2} \ldots, \mathrm{vecv}(b)|_{t_{\rho-1}}^{t_\rho}\right]^{\mathrm{T}} \\ &\in \mathbb{R}^{\rho \times \boldsymbol{m}(\boldsymbol{m}+1)/2} \end{aligned}$$

$$\begin{aligned} \Gamma_{a,b} &= \left[\int_{t_0}^{t_1} a \otimes b \, \mathrm{d}\tau, \int_{t_1}^{t_2} a \otimes b \, \mathrm{d}\tau, \ldots, \int_{t_{\rho-1}}^{t_\rho} a \otimes b \, \mathrm{d}\tau\right]^{\mathrm{T}} \\ &\in \mathbb{R}^{\rho \times \boldsymbol{nm}}. \end{aligned}$$

1) Phase 1 to Find a Stabilizing Control Policy for CO^{2}RP: First, a data-driven VI approach is proposed in order to solve the CO^{2}RP. This approach is used for the sake of getting a stabilizing control policy. Consider the Lyapunov candidate $V_k(\bar{x}_{ij}) = \bar{x}_{ij}^{\mathrm{T}}P_{i,k}\bar{x}_{ij}$, where $k \in \mathbb{Z}_+$ and $i \in \mathcal{T}$. By taking the time derivative of $V_k(\bar{x}_{ij})$ along with (21), with some mathematical manipulations and rearrangements, one obtains the following:

$$\begin{aligned} \dot{V}_k(\bar{x}_{ij}) &= \dot{\bar{x}}_{ij}^{\mathrm{T}}P_{i,k}\bar{x}_{ij} + \bar{x}_{ij}^{\mathrm{T}}P_{i,k}\dot{\bar{x}}_{ij} \\ &= \bar{x}_{ij}^{\mathrm{T}}(H_{i,k})\bar{x}_{ij} + 2u_i^{\mathrm{T}}R_iK_{i,k+1}\bar{x}_{ij} \\ &\quad + 2v^{\mathrm{T}}(D_i - S_i(X_{ij}))^{\mathrm{T}}P_{i,k}\bar{x}_{ij} \end{aligned} \tag{22}$$

where $H_{i,k} = A_i^{\mathrm{T}}P_{i,k} + P_{i,k}A_i$.

By taking the integral of (22) over $[t_0, t_s]$, where $\{t_l\}_{l=0}^s$ (with $t_l = t_{l-1} + \Delta t$ and $\Delta t > 0$) is an increasing sequence, the result can be written in the following Kronecker product representation:

$$\Theta_{ij}\begin{bmatrix} \mathrm{vec}(H_{i,k}) \\ \mathrm{vec}(K_{i,k+1}) \\ \mathrm{vec}((D_i - S(X_{ij}))^{\mathrm{T}}P_{i,k}) \end{bmatrix} = \delta_{\bar{x}_{ij},\bar{x}_{ij}}\mathrm{vecs}(P_{i,k}) \tag{23}$$

where $\Theta_{ij} = \left[\Gamma_{\bar{x}_{ij},\bar{x}_{ij}}, \quad 2\Gamma_{\bar{x}_{ij},u}(I_{n_i} \otimes R_i), \quad 2\Gamma_{\bar{x}_{ij},v}\right]$.

Lemma 3: For all $j \in \mathbb{Z}_+$, if there exists an $s' \in \mathbb{Z}_+$ such that the following rank condition is satisfied for all $s > s'$:

$$\mathrm{rank}\left(\left[\Gamma_{\bar{x}_{ij},\bar{x}_{ij}}, \Gamma_{\bar{x}_{ij},u_i}, \Gamma_{\bar{x}_{ij},v}\right]\right) = \frac{n_i(n_i+1)}{2} + (m_i + q)n_i \tag{24}$$

the matrix Θ_{ij} has full column rank $\forall \, k \in \mathbb{Z}_+, i \in \mathcal{T}$, for any increasing sequence $\{t_l\}_{l=0}^s$ (with $t_l = t_{l-1} + \Delta t$ and $\Delta t > 0$).

Proof: One can prove this lemma by contradiction. Assume $\Xi_v = [\mathrm{vec}(\Omega_1), \mathrm{vec}(\Omega_2), \mathrm{vec}(\Omega_3)]^{\mathrm{T}}$ is a nonzero solution to

$$\Theta_{ij}\Xi_v = 0. \tag{25}$$

Considering (22) and (21), the following equation is concluded:

$$\Gamma_{\bar{x}_{ij},\bar{x}_{ij}}\mathrm{vec}(\Omega_1) + 2\Gamma_{\bar{x}_{ij},u_i}\mathrm{vec}(\Omega_2) + 2\Gamma_{\bar{x}_{ij},v}\mathrm{vec}(\Omega_3) = 0 \tag{26}$$

where $\Omega_1 = H_{i,k}$, $\Omega_2 = R_iK_{i,k+1}$, and $\Omega_3 = (D_i - S(X_{ij}))^{\mathrm{T}}P_{i,k}$. Then, (26) implies the following equation:

$$\left[\Gamma_{\bar{x}_{ij},\bar{x}_{ij}}, \quad 2\Gamma_{\bar{x}_{ij},u}, \quad 2\Gamma_{\bar{x}_{ij},v}\right]\begin{bmatrix} \mathrm{vec}(\Omega_1) \\ \mathrm{vec}(\Omega_2) \\ \mathrm{vec}(\Omega_3) \end{bmatrix} = 0. \tag{27}$$

Under full rank condition of (24), one concludes that $\mathrm{vec}(\Omega_1) = 0$, $\mathrm{vec}(\Omega_2) = 0$, and $\mathrm{vec}(\Omega_3) = 0$. As a result, the following equations are obtained:

$$H_{i,k} := A_i^{\mathrm{T}}P_{i,k} + P_{i,k}A_i = 0 \tag{28}$$

$$R_iK_{i,k+1} := B_i^{\mathrm{T}}P_{i,k} = 0. \tag{29}$$

By Assumption 2, the pair $(A_i, B_i) \neq (0_{n\times n}, 0_{n\times m})$ is obviously concluded, so one can quickly notice that having $P_{i,k} = 0$ is required. Consequently, $\Xi_v = 0$ is the unique solution to (25), which contradicts our assumption that $\Xi_v \neq 0$. The proof is, thus, completed. ■

Lemma 3 shows that if (24) is satisfied, the existence and uniqueness of the solution to (23) is guaranteed, where the solution can be obtained using the pseudoinverse of Θ_{ij}.

Remark 2: The matrix Θ_{ij} is fixed for all $k \in \mathbb{Z}_+$ and does not require to be updated at each iteration k.

To this end—similar to the model-based HI method—$\hat{Q}_i$ is regarded as the state weights in the value update equation. The value matrix is updated by the stochastic approximation, i.e.,

$$P_{i,k+1} \leftarrow P_{i,k} + \epsilon_k \left(H_{i,k} + \hat{Q}_i - (K_{i,k+1})^{\mathrm{T}} R_i K_{i,k+1} \right)$$

until the condition $(P_{i,k} - P_{i,k-1})/\epsilon_k \prec \hat{Q}_i$ is satisfied. By the latter, it is guaranteed that the obtained control policy is stabilizing, which is then employed to start Phase 2.

2) Phase 2 to Explore the Optimal Control Policy for CO² RP: Now, it is time to start with PI since it is guaranteed that the control policy obtained from the previous phase is stabilizing. Consider a Lyapunov function defined by $V_k(\bar{x}_{ij}) = \bar{x}_{ij}^{\mathrm{T}} P_{i,k} \bar{x}_{ij}$ $\forall i \in \mathcal{T}$ and $k = 1, 2, \ldots$. By its time derivative along with (21), the following is obtained:

$$\begin{aligned} \dot{V}_k(\bar{x}_{ij}) = {} & \bar{x}_{ij}^{\mathrm{T}} (P_{i,k} A_{i,k} + A_{i,k}^{\mathrm{T}} P_{i,k}) \bar{x}_{ij} \\ & + 2(u_i + K_{ik} \bar{x}_{ij})^{\mathrm{T}} B_i^{\mathrm{T}} P_{i,k} \bar{x}_{ij} \\ & + 2v^{\mathrm{T}} (D_i - S_i(X_{ij}))^{\mathrm{T}} P_{i,k} \bar{x}_{ij}. \end{aligned} \tag{30}$$

The following is reached by taking the integral of (30) over $[t_0, t_s]$ and using the fact that $P_{i,k} A_{i,k} + {A_{i,k}}^{\mathrm{T}} P_{i,k} = -Q_i - K_{i,k}^{\mathrm{T}} R_i K_{i,k}$

$$\begin{aligned} V_k(\bar{x}_{ij})|_{t_0}^{t_s} = \int_{t_0}^{t_s} & \left[\bar{x}_{ij}^{\mathrm{T}} (-Q_i - K_{i,k}^{\mathrm{T}} R_i K_{i,k}) \bar{x}_{ij} \right. \\ & + 2(u_i + K_{i,k} \bar{x}_{ij})^{\mathrm{T}} R_i K_{i,k+1} \bar{x}_{ij} \\ & \left. + 2v^{\mathrm{T}} (D_i - S_i(X_{ij}))^{\mathrm{T}} P_{i,k} \bar{x}_{ij} \right] \mathrm{d}\tau. \end{aligned} \tag{31}$$

Equation (31) can be written in the Kronecker representation as follows:

$$\Psi_{ij,k} \begin{bmatrix} \mathrm{vecs}(P_{i,k}) \\ \mathrm{vec}(K_{i,k+1}) \\ \mathrm{vec}((D_i - S_i(X_{ij}))^{\mathrm{T}} P_{i,k}) \end{bmatrix} = \Phi_{ij,k} \tag{32}$$

where

$$\begin{aligned} \Psi_{ij,k} = {} & \left[\delta_{\bar{x}_{ij}, \bar{x}_{ij}}, -2\Gamma_{\bar{x}_{ij}, \bar{x}_{ij}} \left(I_{n_i} \otimes K_{i,k}^{\mathrm{T}} R_i \right) \right. \\ & \left. -2\Gamma_{\bar{x}_{ij}, u_i} \left(I_{n_i} \otimes R_i \right), -2\Gamma_{\bar{x}_{ij}, v} \right] \\ \Phi_{ij,k} = {} & -\Gamma_{\bar{x}_{ij}, \bar{x}_{ij}} \mathrm{vec} \left(Q_i + K_{i,k}^{\mathrm{T}} R_i K_{i,k} \right). \end{aligned}$$

It is noticeable that the satisfaction of (24) also implies the existence and the uniqueness of the solution to (32). The novel data-driven HI algorithm can now be introduced, as all its preliminaries are ready. It is presented in Algorithm 2 with its proof of convergence in Theorem 2.

Theorem 2: If (24) is satisfied, the sequences $\{P_{i,k}\}_{k=0}^{\infty}$ and $\{K_{i,k}\}_{k=1}^{\infty}$ learned by Algorithm 2 converge to P_i^* and K_i^* $\forall$ $i \in \mathcal{T}$, respectively.

Proof: Given that (24) is satisfied, one can guarantee that a unique solution is obtained from (23). In addition, $H_{i,k}$, $K_{i,k+1}$, and $\mathrm{vec}((D_i - S(X_{ij}))^{\mathrm{T}} P_{i,k})$ must satisfy (22) $\forall i \in \mathcal{T}$ such that $H_{i,k} = A_i^{\mathrm{T}} P_{i,k} + P_{i,k} A_i$ and $K_{i,k+1} = R_i^{-1} B_i^{\mathrm{T}} P_{i,k}$. Therefore, both $\tilde{P}_{i,k+1}$ and $P_{i,k+1}$ solved from 10–16 in Algorithm 2 are equivalent to those solved in Algorithm 1 through Steps 4–9. Moreover, by the article presented in [36], given that (24) is satisfied, the pair $(P_{i,k}, K_{i,k+1})$ obtained from solving Steps 18–20 in Algorithm 2 is equivalent to those solved in Algorithm 1 through Steps 10–13. The convergence of the sequences $\{P_{i,k}\}_{k=0}^{\infty}$ and $\{K_{i,k+1}\}_{k=0}^{\infty}$ obtained using Algorithm 1 is proved in Theorem 1. Therefore, the same sequences obtained under Algorithm 2 are also ensured. Moreover, the feedback gain matrix K_{i,k^*} is stabilizing for small $\hat{\varepsilon}_i > 0$. It is guaranteed from Lemma 1 that the closed-loop system is asymptotically stable and the tracking error converges asymptotically to zero using the learned suboptimal controller (7) and (33). This section completes the proof. ■

Algorithm 2: Data-Driven HI for CO²RP.

1: $i \leftarrow 1$
2: Choose $\hat{\varepsilon}_i > 0$, $P_{i,0} = (P_{i,0})^{\mathrm{T}} \succ 0$, and $\hat{Q}_i = (\hat{Q}_i)^{\mathrm{T}} \succ Q_i$.
3: **repeat**
4: Compute the matrices $X_{i0}, X_{i1}, \ldots, X_{i,h_i+1}$.
5: Employ $u_i^0 = -K_{i,0} x + \eta_i$, with arbitrary $K_{i,0}$, and exploration noise η_i over $[t_0, t_s]$. $j \leftarrow 0$.
6: **repeat**
7: Compute $\Gamma_{\bar{x}_{ij}\bar{x}_{ij}}$, $\Gamma_{\bar{x}_{ij}u_i}$, and $\Gamma_{\bar{x}_{ij}v}$ while satisfying (24). $j \leftarrow j + 1$.
8: **until** $j = h_i + 2$
9: $k \leftarrow 0$, $j \leftarrow 0$, $r \leftarrow 0$.
10: **repeat**
11: Solve $H_{i,k}$ and $K_{i,k+1}$ from (23).
12: $\tilde{P}_{i,k+1} \leftarrow P_{i,k} + \epsilon_k (H_{i,k} + \hat{Q}_i - (K_{i,k+1})^{\mathrm{T}} R_i K_{i,k+1})$
13: **if** $\tilde{P}_{i,k+1} \notin B_r$ **then** $P_{k+1} \leftarrow P_{i,0}$, $r \leftarrow r + 1$.
14: **else** $P_{i,k+1} \leftarrow \tilde{P}_{i,k+1}$ **end if**
15: $k \leftarrow k + 1$
16: **until** $(P_{i,k} - P_{i,k-1})/\epsilon_k \prec \hat{Q}_i$
17: $j \leftarrow 0$
18: **repeat**
19: Solve $P_{i,k}$ and $K_{i,k+1}$ from (32). $k \leftarrow k + 1$.
20: **until** $\|P_{i,k} - P_{i,k-1}\| < \hat{\varepsilon}_i$
21: $k \leftarrow k^*$, $j \leftarrow 1$.
22: **repeat**
23: From (32), solve $S_i(X_{ij})$. $j \leftarrow j + 1$.
24: **until** $j = h_i + 2$
From Problem 1, find (X_i^*, U_i^*) using online data.
25: $L_{i,k^*} \leftarrow U_i^* + K_{i,k^*} X_i^*$
26: Obtain the suboptimal controller using (7) and

$$u_i^* = -K_{i,k^*} x_i + L_{i,k^*} \zeta_i. \tag{33}$$

27: $i \leftarrow i + 1$
28: **until** $i = N + 1$

Remark 3: The proposed HI Algorithm 2 is an off-policy learning algorithm. Each agent has its own optimal control policy and is learned independently of the other agents.

Remark 4: An exploration noise is added to the input of the system (2) and (3) during the learning process of Algorithm 2. Such an input is chosen to satisfy the rank condition (24)—which is similar to the persistent excitation condition. The noise

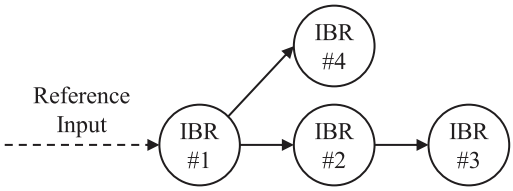

Fig. 1. Sample of an M^{2}G s swarm communication architecture.

selected can be a random noise or a summation of sinusoidal signals with distinct frequencies.

C. Computational Complexity of the HI Algorithm

The computational complexity analysis of the proposed HI algorithm is now analyzed and elaborated. If singular value decomposition (SVD) is employed to compute the pseudoinverse in Steps 11 and 19 of Algorithm 2, it is found that the most intensive steps computationally are Steps 11 and 19 with the computational complexity of $\mathcal{O}(n_i^2\mathcal{Y}_1)$, where $\mathcal{Y}_1$ is the number of rows of Θ_{ij} and $\Psi_{ij,k}$. The SVD method is considered in the analysis since it has less complexity per iteration than the one incurred in the inversion of matrices [45]. From the above analysis, since the number of data to be collected in order to satisfy the rank condition in (24) for HI, PI, and VI are the same, one can conclude that the complexity per iteration of Algorithm 2 cannot be higher than that of the PI algorithm. In case they are to be equal due to the system's dynamics, HI will still have an advantage over PI since the stabilizing policy is not required to begin the learning process, thus preserving the quadratic rate of convergence. Moreover, Phase 2 of HI has higher complexity per iteration compared with VI. Still, due to the quadratic convergence of HI, the time required for convergence is much less than what is necessary for VI since VI needs many iterations to converge to its optimal policy, considering its sublinear convergence rate.

IV. Illustrative Example With Simulations and Experiments

This section presents the efficacy of the proposed HI algorithm by implementing the data-driven HI algorithm on an M^{2}G with communication topology, as depicted in Fig. 1. The results are first validated using MATLAB 2022b. Afterward, the effectiveness of the learned control policy is experimentally tested.

A. Simulation Results

Based on the HI method proposed for MASs in this article, this section derives a secondary voltage control (also known as voltage restoration control) for an islanded M^{2}G containing N IBRs. For such a configuration, one can obtain the following voltage dynamics, which are linear and in the form of second-order dynamical systems according to the article presented in [6] and considering the actuators with partial loss of effectiveness (PLOE) and bias faults occurring in M^{2}Gs [see (2) in [5] for the

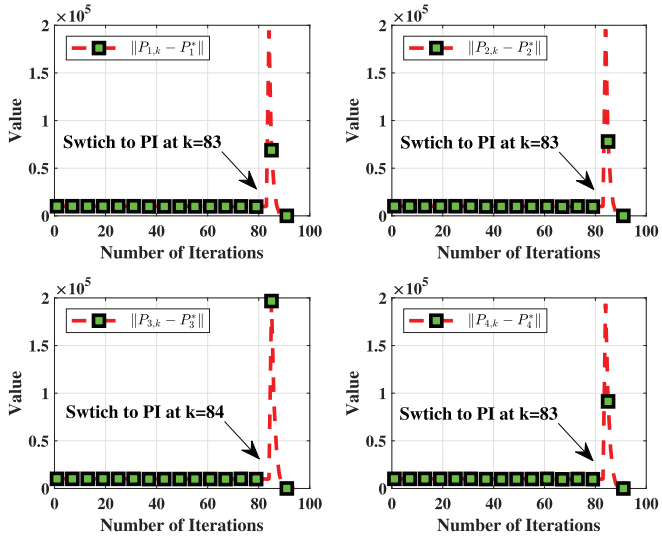

Fig. 2. $\|P_{i,k} - P_i^*\|$ of IBR #i ($i = 1$, 2, 3, and 4) using HI Algorithm 2.

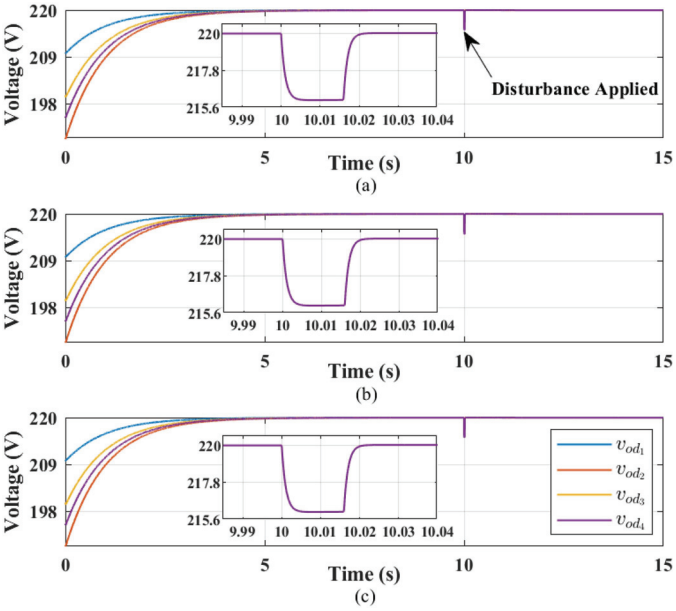

Fig. 3. Trajectories of the voltages v_{od_i} in rms with the learned control policy using (a) HI Algorithm 2, (b) data-driven PI algorithm, and (c) data-driven VI algorithm.

latter]

$$\ddot{v}_{od_i} = \mathcal{Q}_i u_i + \mathcal{P}_i \ \forall i \in \mathcal{T} \tag{34}$$

where $0 < \mathcal{Q}_i \leq 1$ is the PLOE fault factor of the ith IBRs actuator, and $\mathcal{P}_i$ is the time-varying PLOE fault severity of the ith IBR satisfying $\|\mathcal{P}_i\| \leq \mathcal{P}_i^S$ with $\mathcal{P}_i^S > 0$. Using (34) and assuming $x_i = [v_{od_i} \quad \dot{v}_{od_i}]^{\mathrm{T}}$, the dynamical model of the ith IBR can be represented as follows:

$$\dot{x}_i = A_i x_i + B_i(\mathcal{Q}_i u_i + \mathcal{P}_i), \ i \in \mathcal{T} \tag{35}$$

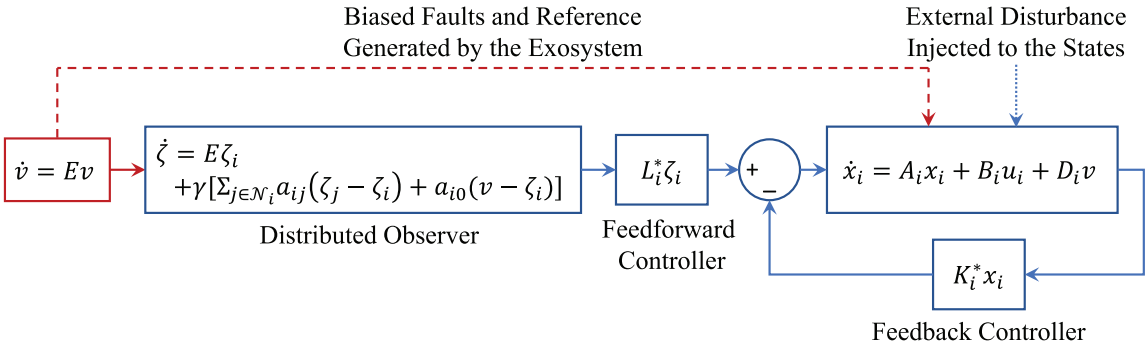

Fig. 4. Closed-loop system with the feedback-feedforward diagram at the moment of applying the disturbance to the system.

where the system matrices obtained are as follows:

$$A_i = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, B_i = \begin{bmatrix} 0 \\ \mathcal{Q}_i \end{bmatrix}, \text{and } C_i = \begin{bmatrix} 1 & 0 \end{bmatrix} \tag{36}$$

in which $\mathcal{Q}_1 = 0.95$, $\mathcal{Q}_2 = 0.90$, $\mathcal{Q}_3 = 0.85$, and $\mathcal{Q}_4 = 0.80$. The disturbances are modeled as sinusoidal signals biased with a constant generated by the exosystem. As a result, (35) can be written in the form of (1)–(3) considering the matrices defined in (36) and the following ones $\forall i \in \mathcal{T}$:

$$E = \begin{bmatrix} 0 & -5 & 0 \\ 5 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, D_2 = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$D_3 = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{5} & \frac{1}{3} & \frac{2}{5} \end{bmatrix}, \quad D_4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad F_i = \begin{bmatrix} 0 & 0 & -1.1 \end{bmatrix}.$$

The goal is to regulate the voltage v_{od_i} of each IBR in an optimal sense—despite the unknown dynamics of all IBRs—this matter requires the designer to solve a cooperative, adaptive, optimal output regulation problem.

The data-driven HI Algorithm 2 is validated by deploying it in the system described by (35) with the communication topology, as depicted in Fig. 1. Also, the data-driven PI and VI algorithms are employed to compare their results with those obtained from the HI method. The rest of the parameters are chosen by $Q_i = 10^4 I_2$, $R_i = 1$, $P_{i,0} = 0.01 I_2$, $\epsilon_k = \frac{4}{k}$, $\hat{\varepsilon}_i = 10^{-4}$ $\forall 1 \leq i \leq 4$, and $B_r = 10(r+1)$. The learning period is from $t = 0$ s to $t = 6$ s such that the bounded input employed to the system is a summation of randomly generated sinusoidal signals with distinct frequencies. This section uses the parameters indicated in Table I, as reported in [5].

Given that the initial stabilizing control policy is known for the PI method, for simulation purposes, the total central processing unit (CPU) time required for convergence is found to be 0.203, 0.125, and 1.922 (all in s) for the HI, PI, and VI methods, respectively. From the results, as shown in Table I and Fig. 2, one can realize the efficiency of HI over VI and PI. PI primarily relies on prior knowledge of an initial stabilizing control policy for each subsystem, which is assumed to be known in this case. One can notice that PI requires fewer learning iterations and less convergence time for each subsystem to learn its own optimal control policy. However, achieving this requirement in practice is not straightforward due to the lack of modeling information. One can see that HI requires significantly less time and iterations than VI to converge.

TABLE I
NUMBER (NO.) OF ITERATIONS AND CPU TIME [IN SECOND (S)] FOR EACH ALGORITHM TO CONVERGE TO THE OPTIMAL SOLUTION FOR EACH IBR WITH PARAMETERS IN [5]

IBR #	HI		PI		VI	
	Iteration No.	CPU Time	Iteration No.	CPU Time	Iteration No.	CPU Time
1	93	0.0781	15	0.0468	3095	0.9220
2	93	0.0937	15	0.0156	2915	0.4531
3	94	0.0156	15	0.0313	2759	0.2656
4	93	0.0156	15	0.0313	2837	0.2813

Additionally, the HI method is less conservative than the PI method, as it removes the condition of prior knowledge of a stabilizing policy for each follower by achieving the optimal control policy with a fast convergence rate. It is worth mentioning that P_i^* is not required to perform the simulations, as shown in Fig 2. This article assumes that it is known in advance for the sake of presenting the results of convergence. In practice, it is enough to have $\|P_{i,k} - P_{i,k-1}\| \leq \hat{\varepsilon}_i$ to guarantee that the learned value matrix P_{i,k^*} is close enough to the actual optimal one, i.e., P_i^*, as shown in the proof of Theorem 2.

Moreover, Phase 1 of Algorithms 1 and 2 generates a monotonically increasing sequence $\{P_{i,k}\}_{k=0}^{\infty}$ since $P_{i,0} \prec P_i^*$. The weighting matrix Q_i was chosen with large weights to ensure a faster trajectories convergence at the cost of requiring a large $P_{i,k_a} \succeq P_i^* \succ 0$, which can ensure that $A_i - B_i K_{i,k_a}$ is Hurwitz. Thus, the control policy K_{i,k_a} is stabilizing, which is obtained by satisfying the condition in Step 16 in Algorithm 2. This fact is evident at the switching points in Fig. 2. The significant jump on $P_{i,k}$ does not affect the state trajectory of the closed-loop system since the off-policy reinforcement learning strategy has been applied. The quadratic convergence of Phase 2 can also be observed. The tracking trajectories, as depicted in Fig. 3, reveal that the HI, PI, and VI methods learn optimal control policies since all three approaches converge to the same optimal control policy—which is unique. Therefore, it is notable that all tracking trajectories have the same behavior.

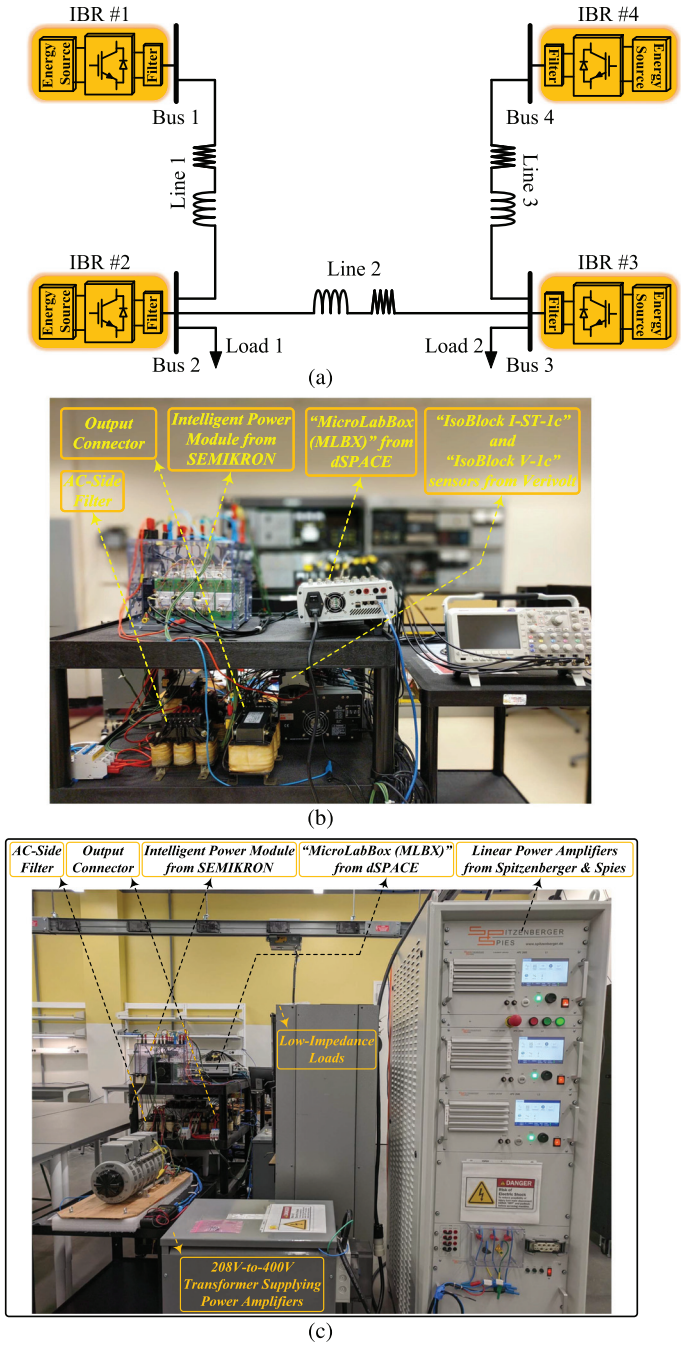

Fig. 5. Test rig deployed in the experiments. (a) M^{2}Gs power system (b) and details of one IBR (parameters are reported in [5]), and (c) amplifiers and low-impedance loads—all housed in the Laboratory for Advanced Power and Energy Systems at Georgia Southern University—where experiments have been conducted.

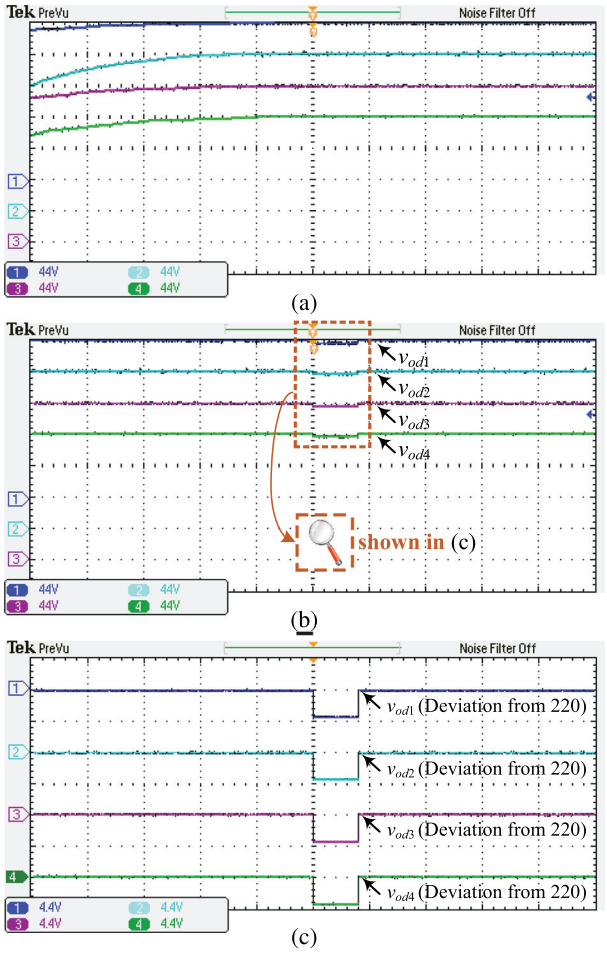

Fig. 6. Experimental results of the proposed method, HI Algorithm 2 with 93, 93, 94, and 93 iterations for IBRs #1–#4 (simulations in Fig. 3). (a) Signals in the transient condition with 44 V/div for the vertical axis and 1 s/div for the horizontal axis. (b) Signals in the steady-state condition with 44 V/div for the vertical axis and 20 ms/div for the horizontal axis. (c) Signal deviations from 220 V with 4.4 V/div for the vertical axis and 20 ms/div for the horizontal axis.

From Fig. 3, one can notice that the learned control policy results in maintaining the optimal voltage of each M^{2}Gs IBR with the error converging to zero. Additionally, in order to examine the functionality of the proposed HI method, an external disturbance is applied at $t = 10$ s. The external disturbances are directly applied to the system states through the signals injected into the systems states; disturbance signals replicate sudden voltage sags that may occur in M^{2}Gs (see Fig. 4 for details). Finally, one can also observe that all the IBRs reject external disturbances and reasonably recover from them.

B. Experimental Results

In order to validate the simulation results and to show the practicability of the proposed method, the optimal control policy learned from the HI Algorithm 2 has been employed to control an M^2 G experimentally. The test rig, as depicted in Fig. 5, is utilized to conduct experimental examinations related to the M^2 G simulated in this article. Its IBRs are implemented by the SEMIKRON intelligent power insulated-gate bipolar transistors (IGBTs), the SKM 50 GB 123 D modules. Besides, the SEMIKRON gate drives, the SKHI 21 A (R) product, and protection circuitry are employed to make the converter functional. The Verivolt current/voltage sensors (the IsoBlock I-ST-1c/IsoBlock

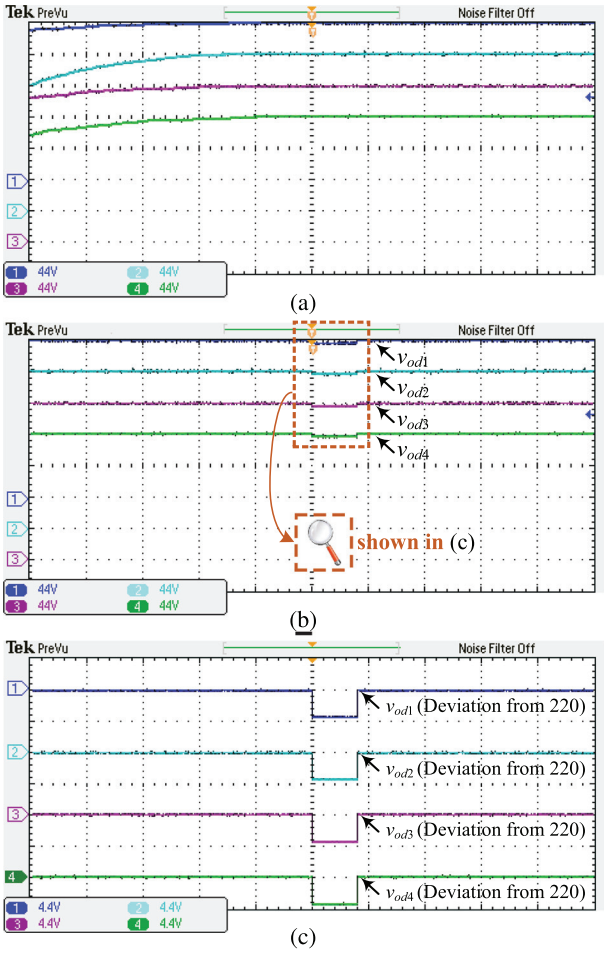

Fig. 7. Comparative experimental results of the data-driven PI algorithm with 15 iterations for IBRs #1–#4 (simulations in Fig. 3). (a) Signals in the transient condition with 44 V/div for the vertical axis and 1 s/div for the horizontal axis. (b) Signals in the steady-state condition with 44 V/div for the vertical axis and 20 ms/div for the horizontal axis. (c) Signal deviations from 220 V with 4.4 V/div for the vertical axis and 20 ms/div for the horizontal axis.

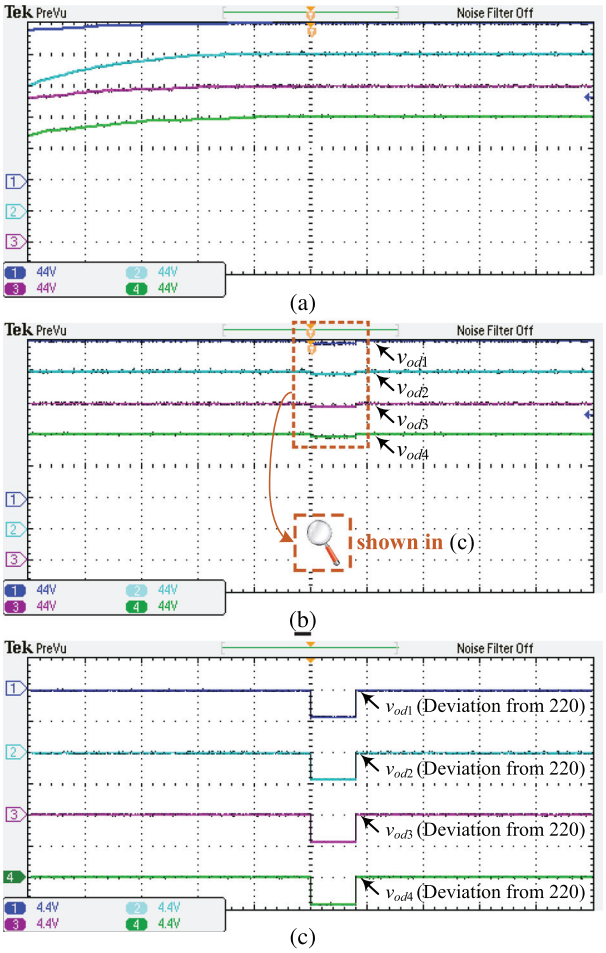

Fig. 8. Comparative experimental results of the data-driven VI algorithm with 3095, 2915, 2759, and 2837 iterations for IBRs #1–#4 (simulations in Fig. 3). (a) Signals in the transient condition with 44 V/div for the vertical axis and 1 s/div for the horizontal axis. (b) Signals in the steady-state condition with 44 V/div for the vertical axis and 20 ms/div for the horizontal axis. (c) Signal deviations from 220 V with 4.4 V/div for the vertical axis and 20 ms/div for the horizontal axis.

V-1c products) have been hooked to digital inputs to measure the currents and the voltages, respectively. In order to practically replicate external disturbance signals applied to the IBRs' voltages (according to Section IV-A and Fig. 4), the voltage sags are generated by programmable power amplifiers and by connecting local low-impedance loads to IBRs. A dSPACE MicroLabBox (MLBX) device using a real-time processor, field-programmable gate arrays, and different inputs/outputs' channels links the IBR under test to the measurement and drive circuitry.

Furthermore, all the parameters of the setup deployed are similar to those of simulations and are found in [5]. Therefore, a reasonably fair comparison between comparative simulations and comparative experiments is feasible. In this regard, comparing Figs. 6–8 with Fig. 3 reveals that simulations and comparative experiments match well. Thus, they demonstrate the effectiveness of the proposed control methodology.

V. Conclusion

In this article, we had solved the CO²RP using a novel computational ADP algorithm called HI. Unlike the existing ADP algorithms, HI has advantages in the sense that prior knowledge of a stabilizing control policy is not required compared with PI. At the same time, the fast convergence speed of PI has still been preserved. Additionally, compared with VI, HI converges much faster to the optimal control policy regarding the number of learning iterations and convergence time. The cooperative, adaptive, optimal controller had been designed via the data-driven HI, and its convergence had been proved. Comparative simulation results had shown the superiority of the proposed HI methodology over the existing ADP methods. Also, comparative experiments had been displayed in order to demonstrate the practicability and excellence of the proposed method, which had been applied to the secondary voltage control (also known as voltage restoration control) of an islanded M²G based on IBRs.

REFERENCES

[1] J. Ploeg, E. Semsar-Kazerooni, G. Lijster, N. van de Wouw, and H. Nijmeijer, "Graceful degradation of cooperative adaptive cruise control," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 488–497, Feb. 2015.

[2] T. Liu and J. Huang, "Cooperative output regulation for a class of nonlinear multi-agent systems with unknown control directions subject to switching networks," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 783–790, Mar. 2018.

[3] M. Lu and L. Liu, "Distributed feedforward approach to cooperative output regulation subject to communication delays and switching networks," *IEEE Trans. Autom. Control*, vol. 62, no. 4, pp. 1999–2005, Apr. 2017.

[4] C. Deng, W.-W. Che, and P. Shi, "Cooperative fault-tolerant output regulation for multiagent systems by distributed learning control approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4831–4841, Nov. 2020.

[5] M. Zhai, Q. Sun, R. Wang, B. Wang, S. Liu, and H. Zhang, "Fully distributed fault-tolerant event-triggered control of microgrids under directed graphs," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 5, pp. 3570–3579, Sep./Oct. 2022.

[6] A. Bidram, A. Davoudi, F. L. Lewis, and J. M. Guerrero, "Distributed cooperative secondary control of microgrids using feedback linearization," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3462–3470, Aug. 2013.

[7] M. Raeispour, H. Atrianfar, M. Davari, and G. B. Gharehpetian, "Fault-tolerant, distributed control for emerging, VSC-based, islanded microgrids–An approach based on simultaneous passive fault detection," *IEEE Access*, vol. 10, pp. 10995–11010, 2022.

[8] C. Deng, C. Wen, W. Wang, X. Li, and D. Yue, "Distributed adaptive tracking control for high-order nonlinear multiagent systems over event-triggered communication," *IEEE Trans. Autom. Control*, vol. 68, no. 2, pp. 1176–1183, Feb. 2023, doi: 10.1109/TAC.2022.3148384.

[9] L. N. Tan, "Omnidirectional-vision-based distributed optimal tracking control for mobile multirobot systems with kinematic and dynamic disturbance rejection," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5693–5703, Jul. 2018.

[10] C. Deng, D. Zhang, and G. Feng, "Resilient practical cooperative output regulation for MASs with unknown switching exosystem dynamics under DoS attacks," *Automatica*, vol. 139, May 2022, Art. no. 110172.

[11] J. Huang and Z. Chen, "A general framework for tackling the output regulation problem," *IEEE Trans. Autom. Control*, vol. 49, no. 12, pp. 2203–2218, Dec. 2004.

[12] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4164–4169, Dec. 2016.

[13] A. Isidori and C. I. Byrnes, "Output regulation of nonlinear systems," *IEEE Trans. Autom. Control*, vol. 35, no. 2, pp. 131–140, Feb. 1990.

[14] B. A. Francis and W. M. Wonham, "The internal model principle of control theory," *Automatica*, vol. 12, no. 5, pp. 457–465, Sep. 1976.

[15] P. Wieland, R. Sepulchre, and F. Allgower, "An internal model principle is necessary and sufficient for linear output synchronization," *Automatica*, vol. 47, no. 5, pp. 1068–1074, May 2011.

[16] H. Cai, F. L. Lewis, G. Hu, and J. Huang, "The adaptive distributed observer approach to the cooperative output regulation of linear multi-agent systems," *Automatica*, vol. 75, pp. 299–305, Jan. 2017.

[17] S. Baldi, I. A. Azzollini, and P. A. Ioannou, "A distributed indirect adaptive approach to cooperative tracking in networks of uncertain single-input single-output systems," *IEEE Trans. Autom. Control*, vol. 66, no. 10, pp. 4844–4851, Oct. 2021.

[18] Y. Yang, C. Xu, D. Yue, X. Zhong, X. Si, and J. Tan, "Event-triggered ADP control of a class of non-affine continuous-time nonlinear systems using output information," *Neurocomputing*, vol. 378, pp. 304–314, Feb. 2020.

[19] X. Wang and H. Su, "Completely model-free RL-based consensus of continuous-timemulti-agent systems," *Appl. Math. Comput.*, vol. 382, Oct. 2020, Art. no. 125312.

[20] F. Zhao, W. Gao, Z.-P. Jiang, and T. Liu, "Event-triggered adaptive optimal control with output feedback: An adaptive dynamic programming approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 5208–5221, Nov. 2021.

[21] R. E. Bellman, *"The theory of dynamic programming*," RAND Corp., Santa Monica, CA, USAJul. 1954.

[22] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. Hoboken, NJ, USA: Wiley, Feb. 2012.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, Nov. 2018.

[24] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Jul./Sep. 2009.

[25] K. van Berkel, B. de Jager, T. Hofman, and M. Steinbuch, "Implementation of dynamic programming for optimal control problems with continuous states," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 3, pp. 1172–1179, May 2015.

[26] Q. Wei, R. Song, Z. Liao, B. Li, and F. L. Lewis, "Discrete-time impulsive adaptive dynamic programming," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4293–4306, Oct. 2020.

[27] Z. P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Foundations Trends Syst. Control*, vol. 8, no. 3, pp. 176–284, Dec. 2020.

[28] D. P. Bertsekas, "Value and policy iterations in optimal control and adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 500–509, Mar. 2017.

[29] S. A. A. Rizvi and Z. Lin, "Reinforcement learning-based linear quadratic regulation of continuous-time systems using dynamic output feedback," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4670–4679, Nov. 2020.

[30] W. Bai, T. Li, and S. Tong, "NN reinforcement learning adaptive control for a class of nonstrict-feedback discrete-time systems," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4573–4584, Nov. 2020.

[31] H. Zhang, H. Jiang, Y. Luo, and G. Xiao, "Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4091–4100, May 2017.

[32] Y. Yang, Z. Liu, Q. Li, and D. C. Wunsch, "Output constrained adaptive controller design for nonlinear saturation systems," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 2, pp. 441–454, Feb. 2021.

[33] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, Feb. 2009.

[34] H. Zargarzadeh, T. Dierks, and S. Jagannathan, "Optimal control of nonlinear continuous-time systems in strict-feedback form," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2535–2549, Oct. 2015.

[35] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Trans. Autom. Control*, vol. 13, no. 1, pp. 114–115, Feb. 1968.

[36] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, Oct. 2012.

[37] W. Gao, Z.-P. Jiang, F. L. Lewis, and Y. Wang, "Leader-to-formation stability of multiagent systems: An adaptive optimal control approach," *IEEE Trans. Autom. Control*, vol. 63, no. 10, pp. 3581–3587, Oct. 2018.

[38] T. Bian and Z.-P. Jiang, "Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design," *Automatica*, vol. 71, pp. 348–360, Sep. 2016.

[39] W. Gao, M. Mynuddin, D. C. Wunsch, and Z.-P. Jiang, "Reinforcement learning-based cooperative optimal output regulation via distributed adaptive internal model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5229–5240, Oct. 2022.

[40] Y. Su and J. Huang, "Cooperative output regulation of linear multi-agent systems," *IEEE Trans. Autom. Control*, vol. 57, no. 4, pp. 1062–1066, Apr. 2012.

[41] W. Gao, Z.-P. Jiang, F. L. Lewis, and Y. Wang, "Cooperative optimal output regulation of multi-agent systems using adaptive dynamic programming," in *Proc. Amer. Control Conf.*, 2017, pp. 2674–2679.

[42] W. Gao and Z.-P. Jiang, "Nonlinear and adaptive suboptimal control of connected vehicles: A global adaptive dynamic programming approach," *J. Intell. Robot. Syst.*, vol. 85, no. 3, pp. 597–611, Mar. 2017.

[43] A. J. Krener, "The construction of optimal linear and nonlinear regulators," in *Systems, Models and Feedback: Theory and Applications*, vol. 12. Boston, MA, USA: Birkhauser, Jun. 1992,, pp. 301–322.

[44] J. Huang, *Nonlinear Output Regulation: Theory and Applications*. Philadelphia, PA, USA: SIAM, Nov. 2004.

[45] G. H. Golub and C. F. van Loan*, Matrix Computations*. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1983.

Omar Qasem (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Kuwait University, Khaldiya, Kuwait, in 2013 and 2018, respectively. He is currently working toward the Ph.D. degree in mechanical engineering with the Florida Institute of Technology, Melbourne, FL, USA.

His research interests include optimal control, reinforcement learning, adaptive dynamic programming, output regulation, cooperative control, and nonlinear control theory.

Masoud Davari (Senior Member, IEEE) was born in Isfahan, Iran, on September 14, 1985. He received the B.Sc. degree (*summa cum laude*) in electrical engineering (power) from the Isfahan University of Technology, Isfahan, Iran, in 2007, the M.Sc. degree (*summa cum laude*) in electrical engineering (power) from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2010, and the Ph.D. degree in electrical engineering (power electronics in energy systems with distinction/honors) from the University of Alberta, Edmonton, AB, Canada, in 2016.

He was with Iran's Grid Secure Operation Research Center and Iran's Electric Power Research Institute, Tehran, Iran, from January 2010 to December 2011. From April 2015 to June 2017, he was a Senior R&D Specialist and Senior Consultant with Quanta-Technology Company, Markham, ON, Canada, in the field of the dynamic interaction of renewables with smart ac/dc grids and control, protection, and automation of microgrids. In July 2017, he joined as a tenure-track Assistant Professor with the Allen E. Paulson College of Engineering and Computing, Department of Electrical and Computer Engineering, Georgia Southern University (GSU), Statesboro, GA, USA—where he was recommended for being granted "*early*" promotion to Associate Professor and award of "*early*" tenure on December 3, 2021, and officially approved for both on February 16, 2022. He is the founder and the Director of the Laboratory for Advanced Power and Energy Systems [LAPES (watch it on https://www.youtube.com/watch?v=mhVHp7uMNKo)] in the state-of-the-art Center for Engineering and Research established in 2021 with GSU. He has developed and implemented several experimental test rigs for research universities and the power and energy industry. He has also authored several papers published in IEEE Transactions and journals. His research interests include the dynamics, controls, and protections of different power electronic converters utilized in the hybrid ac/dc smart grids, and hardware-in-the-loop simulation-based testing of modernized power systems.

Dr. Davari has been an active member and a chapter lead in the IEEE Power and Energy Society Task Force on "*Innovative Teaching Methods for Modern Power and Energy Systems*" since July 2020. He has been an active member and a chapter lead (for Chapter 3) in the IEEE Working Group P2004—a newly established IEEE working group entitled "*Hardware-in-the-Loop Simulation Based Testing of Electric Power Apparatus and Controls*" for IEEE Standards Association since June 2017. He is an invited member of the Golden Key International Honour Society. He was the Chair of the Literature Review Subgroup of DC@Home Standards for the IEEE Standards Association from April 2014 to October 2015. He is an invited reviewer of several IEEE TRANSACTIONS/JOURNALS, IET journals, *Energies* journal, and various IEEE conferences, the invited speaker at different universities and in diverse societies, and the Best Reviewer of the IEEE TRANSACTIONS ON POWER SYSTEMS in 2018 and 2020. He was the recipient of the 2019–2020 Allen E. Paulson College of Engineering and Computing (CEC) Faculty Award for Outstanding Scholarly Activity in the Allen E. Paulson CEC at GSU, the Discovery and Innovation Award from the 2020–2021 University Awards of Excellence at GSU, and one of the awardees of the 2021–2022 Impact Area Accelerator Grants (partially funded) at GSU.

Weinan Gao (Senior Member, IEEE) received the B.Sc. degree in automation from Northeastern University, Shenyang, China, in 2011, the M.Sc. degree in control theory and control engineering from Northeastern University, in 2013, and the Ph.D. degree in electrical engineering from New York University, Brooklyn, NY, USA, in 2017.

He is a Professor with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, China. Previously, he was an Assistant Professor of mechanical and civil engineering with the Florida Institute of Technology, Melbourne, FL, USA, an Assistant Professor of electrical and computer engineering with Georgia Southern University, Statesboro, GA, USA, and a Visiting Professor of Mitsubishi Electric Research Laboratory, Cambridge, MA, USA. His research interests include reinforcement learning, adaptive dynamic programming, optimal control, cooperative adaptive cruise control, intelligent transportation systems, sampled-data control systems, and output regulation theory.

Dr. Gao is the recipient of the Best Paper Award in IEEE International Conference on Real-time Computing and Robotics in 2018, and the David Goodman Research Award at New York University in 2019. He is an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE/CAA JOURNAL OF AUTOMATICA SINICA, *Control Engineering Practice*, *Neurocomputing,* and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS, a member of Editorial Board of Neural Computing and Applications, and a Technical Committee member of IEEE Control Systems Society on Nonlinear Systems and Control and in IFAC TC 1.2 Adaptive and Learning Systems.

Daniel R. Kirk received the B.Sc. degree in mechanical engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1997, and the M.Sc. and Ph.D. degrees in aeronautical and astronautical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1999 and 2002, respectively.

He is a Professor of aerospace engineering with the Florida Institute of Technology, Melbourne, FL, USA. His research interests focus on air-breathing and rocket propulsion, experimental and computational fluid dynamics, and advanced and additive manufacturing for existing and new aerospace applications. He has served as a Visiting Scholar with NASA Marshall Space Flight Center and NASA Kennedy Space Center and has managed research projects with NASA, the United States Air Force, and the Office of Naval Research.

Tianyou Chai (Life Fellow, IEEE) received the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 1985.

Since 1985, he has been with the Research Center of Automation, Northeastern University, where he became a Professor in 1988 and a Chair Professor in 2004. He is the founder and Director of the Center of Automation, which became a National Engineering and Technology Research Center in 1997. His current research interests include adaptive control, intelligent decoupling control, integrated plant control and systems, and the development of control technologies with applications to various industrial processes.

Dr. Chai is a member of the Chinese Academy of Engineering, an academician of International Eurasian Academy of Sciences, and an IFAC Fellow. He is a Distinguished Visiting Fellow of The Royal Academy of Engineering (U.K.) and an Invitational Fellow of Japan Society for the Promotion of Science. For his contributions, he was a recipient of three prestigious awards of National Science and Technology Progress, the 2002 Technological Science Progress Award from the Ho Leung Ho Lee Foundation, the 2007 Industry Award for Excellence in Transitional Control Research from the IEEE Control Systems Society, and the 2010 Yang Jia-Chi Science and Technology Award from the Chinese Association of Automation.