Automatica 142 (2022) 110366

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

# Resilient reinforcement learning and robust output regulation under denial-of-service attacks[☆]

Weinan Gao [a,b], Chao Deng [c,*], Yi Jiang [a,d], Zhong-Ping Jiang [e]

[a] State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, 110004, China
[b] Department of Mechanical and Civil Engineering, Florida Institute of Technology, Melbourne, FL 32901, USA
[c] Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[d] Department of Biomedical Engineering, City University of Hong Kong, Hong Kong Special Administrative Region of China
[e] Department of Electrical and Computer Engineering, New York University, Six MetroTech Center, Brooklyn, NY, 11201, USA

ARTICLE INFO

*Article history:*
Received 8 February 2021
Received in revised form 19 January 2022
Accepted 30 March 2022
Available online xxxx

*Keywords:*
Reinforcement learning
Robust output regulation
Hybrid iteration
Denial-of-service attacks

ABSTRACT

In this paper, we have proposed a novel resilient reinforcement learning approach for solving robust optimal output regulation problems of a class of partially linear systems under both dynamic uncertainties and denial-of-service attacks. Fundamentally different from existing works on reinforcement learning, the proposed approach rigorously analyzes both the resilience of closed-loop systems against attacks and the robustness against dynamic uncertainties. Moreover, we have proposed an original successive approximation approach, named hybrid iteration, to learn the robust optimal control policy, that converges faster than value iteration, and is independent of an initial admissible controller. Simulation results demonstrate the efficacy of the proposed approach.

## 1. Introduction

Reinforcement learning (RL) concerns how an agent makes decisions to minimize a predefined cost criterion through active interactions with unknown environment (Sutton & Barto, 2018). In the control community, RL and adaptive/approximate dynamic programming (ADP) (Powell, 2007) have been employed as direct adaptive optimal control methods to stabilize dynamical systems in both discrete-time and continuous-time; see the recent surveys and tutorial papers (Jiang et al., 2020a; Kiumarsi et al., 2018; Liu et al., 2021; Vamvoudakis & Kokolakis, 2020). As a generalization of traditional RL-based stabilization approaches, Gao and Jiang (2016, 2018, 2022), Jiang et al. (2020b, 2020c), Odekunle et al. (2020) and Zhao et al. (2022) have solved adaptive optimal output regulation problems so that the closed-loop system can asymptotically track the reference, while rejecting the disturbance in a model-free optimal sense. Different from existing RL considering only static uncertainties in the system, robust adaptive dynamic programming (RADP) has been proposed in Gao and Jiang (2015) and Jiang and Jiang (2017) to handle the dynamic uncertainties in the system through developing robust optimal controllers.

Although RL-based control has been extensively studied, there are two important conundrums in practice. Most of existing RL research works consider that the communication in the control system is normal without any attacks. However, the cyberattack is usually unavoidable which significantly threatens the security of closed-loop control systems. Therefore, the first conundrum is, besides stability and robustness analysis, how to analyze the resilience of systems against cyberattacks without the knowledge of system dynamics. It is well known that policy iteration (PI) and value iteration (VI) are two typical successive approximation approaches in RL for learning the optimal control policy (Jiang et al., 2020a). Comparing with VI, the convergence of PI is much faster at the price of a strong assumption that an initial admissible control policy must be available. The second conundrum is to propose a novel successive approximation approach that converges faster than VI, and is independent of an initial admissible controller.

The aim of this paper is to solve these two conundrums in RL-based control under denial-of-service (DoS) attacks. DoS attack is a typical attack pattern in the domain of cyber attacks. Essentially, the DoS attackers block the information transmission among networks (Amin et al., 2009; Teixeira et al., 2015). In De Persis and Tesi (2015), a control framework is introduced to provide an explicit characterization of frequency and duration of the DoS attacks under which closed-loop stability can be achieved by using the feedback control. Under this framework, some valuable

[☆] This work has been supported in part by the U.S. National Science Foundation under Grant EPCN-1903781 and CMMI-2138206, and the Scientific Starting Fund from Nanjing University of Posts and Telecommunications (NUPTSF) under Grant NY221007. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Kyriakos G. Vamvoudakis under the direction of Editor Miroslav Krstic.

[*] Corresponding author.

*E-mail addresses:* wgao@fit.edu (W. Gao), cdeng@ntu.edu.sg (C. Deng), yjian22@cityu.edu.hk (Y. Jiang), zjiang@nyu.edu (Z.-P. Jiang).

https://doi.org/10.1016/j.automatica.2022.110366
0005-1098/© 2022 Elsevier Ltd. All rights reserved.

results are emerged; see, e.g., An and Yang (2018) and Deng and Wen (2020). However, there is neither model-free optimal control design nor resilient analysis for dynamical systems invaded by DoS attacks. Most recently, Galarza-Jimenez et al. (2022) and Zhai and Vamvoudakis (2021) have leveraged learning-based control techniques, such as ADP and extremum seeking, to defend the closed-loop dynamic systems from adversarial attacks in the absence of dynamic uncertainties. In this paper, we will propose a novel resilient reinforcement learning ($R^2L$) based control approach to solve a class of robust optimal output regulation problems with unknown system dynamics, dynamic uncertainties, and denial-of-service (DoS) attacks. The major contributions are listed as follows.

1. This paper has, for the first time, proposed an $R^2L$ approach to solve a robust optimal output regulation problem which bridges the gap between RL, RADP, output regulation, and small-gain theories. With the proposed approach, one can directly analyze the resilience of the closed-loop system with the learned robust optimal controller, and find the lower bound of DoS attack duration criterion that the closed-loop system can handle without the knowledge of system dynamics.
2. We have proposed a novel successive approximation algorithm, named hybrid iteration (HI), to learn the optimal control policy and the corresponding value. Comparing with PI, the HI completely removes the assumption of the admissibility of the initial control policy. Also, the convergence rate of HI is usually faster than that of VI.
3. Most existing works on DoS attacks only consider static uncertainties in systems (De Persis & Tesi, 2015; Feng et al., 2020; Feng & Tesi, 2017; Galarza-Jimenez et al., 2022; Zhai & Vamvoudakis, 2021). Different from them, this paper has considered the existence of both DoS attacks and dynamic uncertainties in systems.
4. Comparing with our previous work (Gao & Jiang, 2015) solving the robust optimal output regulation problem via PI, the proposed approach in this paper can solve that problem without relying the knowledge of an admissible control policy.

**Notations.** Throughout this paper, $\mathbb{R}_+$ denotes the set of nonnegative real numbers, $\mathbb{Z}_+$ the set of nonnegative integers, and $\mathbb{N}_+$ the set of positive integers. A set $\mathbb{C}^-$ indicates the open left-half complex plane. The operator $|\cdot|$ represents the Euclidean norm for vectors and the induced norm for matrices. A set $\mathcal{P}^n$ includes all $n \times n$ real, symmetric and positive semidefinite matrices. A continuous function $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is of class $\mathcal{K}$ if it is strictly increasing and $\alpha(0) = 0$. A function $\beta : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is of class $\mathcal{KL}$ if for each fixed $t$, the function $\beta(\cdot, t)$ is of class $\mathcal{K}$ and, for each fixed $s$, the function $\beta(s, \cdot)$ is non-increasing and tends to 0 at infinity. The symbol $\otimes$ indicates the Kronecker product operator and $\mathrm{vec}(A) = [a_1^T, a_2^T, \dots, a_m^T]^T$, where $a_i \in \mathbb{R}^n$ are the columns of $A \in \mathbb{R}^{n\times m}$. For an arbitrary column vector $v \in \mathbb{R}^n$, $\mathrm{vecv}(v) = [v_1^2, v_1v_2, \dots, v_1v_n, v_2^2, v_2v_3, \dots, v_{n-1}v_n, v_n^2]^T \in \mathbb{R}^{\frac{1}{2}n(n+1)}$. $\mathrm{vecs}(P) = [p_{11}, 2p_{12}, \dots, 2p_{1m}, p_{22}, 2p_{23}, \dots., 2p_{m-1,m}, p_{mm}]^T \in \mathbb{R}^{\frac{1}{2}m(m+1)}$ for a symmetric matrix $P \in \mathbb{R}^{m\times m}$, and $\lambda_M(P)$ (resp. $\lambda_m(P)$) denotes the maximum (resp. minimum) eigenvalue of a real symmetric matrix $P$. $P \succ 0$ (resp. $P \prec 0$) represents that the matrix $P$ is positive (resp. negative) definite. For any piecewise continuous function $u : \mathbb{R}_+ \to \mathbb{R}^m$, $\|u\|$ stands for $\sup_{t\geq 0} |u(t)|$.

## 2. Problem formulation and preliminaries

In this paper, we consider a class of partially linear systems, which includes a linear subsystem interconnected with a nonlinear subsystem known by dynamic uncertainty. The dynamics of the interconnected system is described by

$$\dot{\zeta}(t) = \Delta(\zeta(t), e(t), v(t)), \tag{1}$$

$$\dot{x}(t) = Ax(t) + B[u(t) + \Phi(\zeta(t), e(t), v(t))] + Dv(t), \tag{2}$$

$$e(t) = Cx(t) + Fv(t), \tag{3}$$

where $e(t) \in \mathbb{R}$, $u(t) \in \mathbb{R}$ and $x(t) \in \mathbb{R}^n$ are respectively the measurement output, the control input and the state. $\zeta(t) \in \mathbb{R}^p$ is the state of system (1). $\Delta : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^q \to \mathbb{R}^p$ and $\Phi : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^q \to \mathbb{R}$ are locally Lipschitz functions satisfying $\Delta(0, 0, v) = 0$ and $\Phi(0, 0, v) = 0$ for any fixed $v \in \mathbb{R}^q$. $\Phi$ is called a dynamic uncertainty to the system (2)–(3). The signal $v(t) \in \mathbb{R}^q$ is the exostate of an autonomous system usually referred as exosystem:

$$\dot{v}(t) = Sv(t). \tag{4}$$

The constant matrices $A$, $B$, $C$, $D$, $F$ and $S$ are with proper dimensions. Throughout this paper, Assumptions 1–4 are made on the overall system (1)–(4).

**Assumption 1.** The pair $(A, B)$ is stabilizable and all eigenvalues of S are simple with zero real part.

**Assumption 2.** $\mathrm{rank}\begin{bmatrix} A-\lambda I & B \\ C & 0 \end{bmatrix} = n + 1, \forall \lambda \in \sigma(S)$.

**Assumption 3** (*Strong Unboundedness Observability*)**.** There exist a function $\sigma_s$ of class $\mathcal{KL}$ and a function $\gamma_s$ of class $\mathcal{K}$, both of which are independent of any $v$ such that for any measurable locally essentially bounded $e$ on $[0, T)$ with $0 < T \leq +\infty$ and any $v \in \Sigma_v$, $\zeta(t)$ right maximally defined on $[0, T')(0 < T' \leq T)$ satisfies $|\zeta(t)| \leq \sigma_s(|\zeta(0)|, t) + \gamma_s(\|[e_{[0,t]}, \Phi_{[0,t]}]^T\|), \forall t \in [0, T')$, where $e_{[0,t]}$ and $\Phi_{[0,t]}$ are the truncated functions of $e$ and $\Delta$ over $[0, t]$, respectively.

**Assumption 4** (*Input-to-Output Stability*)**.** There exist a function $\sigma_\Phi$ of class $\mathcal{KL}$ and a function $\gamma_\Phi$ of class $\mathcal{K}$, both of which are independent of any $v$ such that, for any initial state $\zeta(0)$, any measurable locally essentially bounded $e$ on $[0, T)$ with $0 < T \leq +\infty$ and any $v \in \Sigma_v$, $\Phi(t)$ right maximally defined on $[0, T')(0 < T' \leq T)$ satisfies

$$|\Phi(t)| \leq \sigma_\Phi(|\zeta(0)|, t) + \gamma_\Phi\left(\|e_{[0,t]}\|\right), \forall t \in [0, T'). \tag{5}$$

**Remark 1.** Assumptions 1–2 are general conditions for solving output regulation problems; see Huang (2004). Assumptions 3–4 make the system (1) be strong unboundedness observable (Jiang et al., 1994) with zero-offset and input-to-output stable (IOS) (Sontag, 2007). Similar assumptions appear in Huang and Chen (2004) and Jiang and Jiang (2017).

## 3. Model-based robust optimal controller design

In this section, we will design a controller to achieve output regulation in a robust optimal sense; see Jiang and Jiang (2017). Note that the controller design in this section does not take DoS attacks into consideration. To begin with, we choose a $G_2 \in \mathbb{R}^p$ such that the pair $(S, G_2)$ is controllable. Through the classical internal model principle (Huang, 2004; Isidori, 2017; Marino &

Tomei, 2003), the following equation

$$\dot z(t) = Sz(t) + G_2 e(t) \tag{6}$$

serves as an internal model for system (2) with (4) as an exosystem. Next, we show in the following lemma that, in the absence of the dynamic uncertainty, one can develop a state-feedback controller for the system (2)–(4) with an internal model (6) to solve the output regulation problem.

**Lemma 1.** *Under Assumptions 1–2, the augmented system*

$$\begin{aligned} \dot v(t) =& Sv(t) \\ \dot x(t) =& Ax(t) + Bu(t) + Dv(t), \\ e(t) =& Cx(t) + Fv(t), \\ \dot z(t) =& Sz(t) + G_2 e(t) \end{aligned} \tag{7}$$

*in closed-loop with the state-feedback controller*

$$u(t) = -K_x x(t) - K_z z(t), \tag{8}$$

*achieves output regulation (asymptotic tracking with disturbance rejection) if the matrix*

$$A_c = \begin{bmatrix} A - BK_x & -BK_z \\ G_2 C & S \end{bmatrix}$$

*is Hurwitz.*

**Proof.** Under the conditions in Assumptions 1–2, there exists uniquely a pair $(X, U)$ solving the following regulator equations

$$XS = AX + BU + D, \tag{9}$$

$$0 = CX + F. \tag{10}$$

By Huang (2004, Lemma 1.27), the matrix equations (10) combined with

$$XS = (A - BK_x)X - BK_z Z + D, \tag{11}$$

$$ZS = SZ + G_2(CX + F) \tag{12}$$

have a unique solution $\hat X$ and $Z$. This implies that $X = \hat X$, and $U = -K_x X - K_z Z$. Define the following vectors and matrices

$$\begin{aligned} \tilde x =& x - Xv, \tilde z = z - Zv, \tilde u = u - Uv, K = \begin{bmatrix} K_x & K_z \end{bmatrix}, \\ \tilde\xi =& \begin{bmatrix} \tilde x^T & \tilde z^T \end{bmatrix}^T \in \mathbb{R}^{n+q}, \bar C = \begin{bmatrix} C & 0 \end{bmatrix}, \\ \bar A =& \begin{bmatrix} A & 0 \\ G_2 C & S \end{bmatrix}, \bar B = \begin{bmatrix} B \\ 0 \end{bmatrix}, \bar D = \begin{bmatrix} D \\ G_2 F \end{bmatrix}. \end{aligned}$$

Based on (11)–(12), we have the following error system of (7):

$$\begin{aligned} \dot{\tilde x} =& Ax + Bu + Dv - XSv = (A - BK_x)\tilde x - BK_z \tilde z, \\ \dot{\tilde z} =& Sz + G_2 e - ZSv = S\tilde z + G_2 C\tilde x. \end{aligned} \tag{13}$$

We can convert (13) into a more compact form,

$$\begin{aligned} \dot{\tilde\xi} =& (\bar A - \bar B K)\tilde\xi = A_c \tilde\xi, \\ e =& \bar C \tilde\xi. \end{aligned}$$

Due to the fact that $A_c$ is Hurwitz, we conclude that $\lim_{t\to\infty} \tilde\xi(t) = 0$, and $\lim_{t\to\infty} e(t) = 0$. The proof is thus completed. □

In order to take into account the transient performance of the closed-loop system, the optimal output regulation problem is defined as follows.

**Problem 1.** The optimal output regulation problem is solved if a feedback controller is designed to achieve output regulation. Moreover, the following dynamic programming

$$\min_{\tilde u} \int_0^\infty (\tilde\xi^T Q \tilde\xi + \tilde u^2) dt \tag{14}$$

$$\text{s.t.} \quad \dot{\tilde\xi} = \bar A\tilde\xi + \bar B\tilde u \tag{15}$$

is solved, where $Q = Q^T \succ 0$.

It should be mentioned that the dynamic programming problem in Problem 1 is a standard linear quadratic regulator problem. The solution to this problem is an optimal feedback controller of the form

$$\tilde u^* = -K^* \tilde\xi, \tag{16}$$

where the optimal control gain matrix $\bar K^*$ is

$$K^* = \bar B^T P^*. \tag{17}$$

From linear optimal control theory, the matrix $P^* = (P^*)^T > 0$ is the solution to the following algebraic Riccati equation (ARE)

$$\bar A^T P^* + P^* \bar A + Q - P^* \bar B \bar B^T P^* = 0, \tag{18}$$

and (16) is equivalent to

$$u^* = \tilde u^* + Uv := -K_x^* x - K_z^* z. \tag{19}$$

Therefore, (19) is an optimal controller for the original system (2)–(3) augmented by an internal model (6) when $\Phi \equiv 0$. In order to develop a robust optimal controller, the closed-loop system has to be robust against nontrivial $\Phi$. In the following theorem, we show that (19) becomes a robust optimal controller for the overall system (1)–(3) if a small-gain condition is satisfied.

**Theorem 1.** *Under Assumptions 3–4, the system (1)–(4) with the optimal control policy (19) achieves output regulation for any $v(t)$, if the following small-gain condition is satisfied*

$$\gamma_\Phi \gamma_e < 1, \tag{20}$$

*where*

$$\gamma_e = |\bar C| \sqrt{\frac{\lambda_M(P^*)}{\lambda_m(P^*)\lambda_m(Q)}}. \tag{21}$$

**Proof.** Defining $\xi(t) = \begin{bmatrix} x^T(t) & z^T(t) \end{bmatrix}^T$, the original system (2) augmented with (6) can be written by

$$\dot\xi(t) = \bar A\xi(t) + \bar B(u(t) + \Phi(t)) + \bar D v(t). \tag{22}$$

For simplicity, we use $\Phi(t)$ to represent $\Phi(\zeta(t), e(t), v(t))$. Based on Eqs. (9)–(12), one obtains the following error system

$$\begin{aligned} \dot{\tilde\xi}(t) =& (\bar A - \bar B K^*)\tilde\xi(t) + \bar B\Phi(t), \\ e(t) =& \bar C\tilde\xi(t). \end{aligned} \tag{23}$$

The ARE (18) is equivalent to

$$(\bar A - \bar B K^*)^T P^* + P^*(\bar A - \bar B K^*) = -Q - P^* \bar B \bar B^T P^*.$$

Then, differentiating the Lyapunov function $V = \tilde\xi^T P^* \tilde\xi$ along the trajectories of system (23), we have

$$\begin{aligned} \dot V &= \tilde\xi^T [(\bar A - \bar B K^*)^T P^* + P^*(\bar A - \bar B K^*)]\tilde\xi + 2\tilde\xi^T P^* \bar B \Phi \\ &= -\tilde\xi^T [Q + P^* \bar B \bar B^T P^*]\tilde\xi + 2\tilde\xi^T P^* \bar B\Phi \\ &\le -\tilde\xi^T Q\tilde\xi - |\Phi - \bar B^T P^* \tilde\xi|^2 + |\Phi|^2 \\ &\le -\tilde\xi^T Q\tilde\xi + |\Phi|^2 \\ &\le -\lambda_m(Q)|\tilde\xi|^2 + |\Phi|^2. \end{aligned} \tag{24}$$

By the fact that $\lambda_m(P^*)|\tilde\xi|^2 \le V \le \lambda_M(P^*)|\tilde\xi|^2$ for any $t \ge 0$, we have

$$V(t) \le \exp\left(-\frac{\lambda_m(Q)}{\lambda_M(P^*)} t\right) V(0) + \frac{\lambda_M(P^*)}{\lambda_m(Q)} \|\Phi\|^2.$$

An immediate consequence of the previous inequality is

$$|\tilde{\xi}(t)| \leq \exp\left(-\frac{\lambda_m(Q)}{2\lambda_M(P^*)}t\right)\sqrt{\frac{\lambda_M(P^*)}{\lambda_m(P^*)}}|\tilde{\xi}(0)| + \sqrt{\frac{\lambda_M(P^*)}{\lambda_m(P^*)\lambda_m(Q)}}\|\Phi\|, \quad \forall t \geq 0, \tag{25}$$

which implies that the system (23) with $\Phi$ as the input is input-to-state stable (see, e.g., Sontag, 1989). One can write

$$|e(t)| \leq \sigma_e(|\tilde{\xi}(0)|, t) + \gamma_e\|\Phi\|, \tag{26}$$

where

$$\sigma_e(|\tilde{\xi}(0)|, t) = |\bar{C}| \exp\left(-\frac{\lambda_m(Q)}{2\lambda_m(P^*)}t\right)\sqrt{\frac{\lambda_M(P^*)}{\lambda_m(P^*)}}|\tilde{\xi}(0)|$$

is a function of $\mathcal{KL}$ and $\gamma_e$ is defined in (21), which guarantees that (23) with $e$ as output has SUO property with zero-offset and IOS properties (Jiang et al., 1994). Assumptions 3 and 4 indicate that the $\zeta$-system has SUO property with zero-offset and IOS properties with input-to-output gain function $\gamma_\Phi(s)$. Under the small-gain condition (20), the interconnected system (1) with (23) is globally asymptotically stable at the origin for any $v(t)$. We have $\lim_{t\to\infty}\left(x(t) - Xv(t)\right) = 0$, and $\lim_{t\to\infty} e(t) = 0$, which immediately implies that both asymptotic tracking and disturbance rejection are achieved. The proof is thus completed. □

## 4. Online learning of the robust optimal controller under DoS

In this section, we will propose three online learning strategies–PI, VI, and HI–to learn the robust optimal control policy (19) in terms of online data. Moreover, we consider the existence of DoS attack during the learning process, which will be described in Section 4.1.

**Remark 2.** The learning algorithms to be proposed in this section are model-free since they rely neither on the knowledge of system matrices $A, B, C, D, F$, nor the structures of system functions $\Phi(\zeta, e, v)$ and $\Delta(\zeta, e, v)$.

### 4.1. DoS attacks

In this section, both the actuator and sensor attacks on the interconnected system will be considered; see Fig. 1. We use $\mathcal{I}_s = [h_s, h_s + \tau_s)$ to represent the $s$th ($s \in \mathbb{N}_+$) DoS attacks interval. $h_s$, $h_s + \tau_s$ and $\tau_s$ represent the start time, end time and the length of the $s$th DoS attacks. Based on Assumption 5, we define $\Pi_D(t_a, t_b) := (t_a, t_b) \bigcap \bigcup_{s=1}^{\infty} \mathcal{I}_s$ as the set in which the communication is denied under the influenced of DoS attacks during the interval $[t_a, t_b]$. Accordingly, we use $\Pi_N(t_a, t_b) := [t_a, t_b] \setminus \Pi_D(t_a, t_b)$ to denote the normal communication set.

The following assumptions are made regarding the DoS frequency and DoS duration.

**Assumption 5** (*DoS Frequency*)**.** There exist constants $\eta > 1$ and $\tau_D > 0$ such that

$$n(t_a, t_b) \leq \eta + \frac{t_b - t_a}{\tau_D}, \ \forall\, t_b > t_a \geq 0, \tag{27}$$

where $n(t_a, t_b)$ denotes the number of DoS off/on transitions occurring on the interval $[t_a, t_b]$.

**Assumption 6** (*DoS Duration*)**.** There exist constants $T > 1$ and $\kappa > 0$ such that

$$|\Pi_D(t_a, t_b)| \leq \kappa + \frac{t_b - t_a}{T}, \ \forall\, t_b > t_a \geq 0,$$

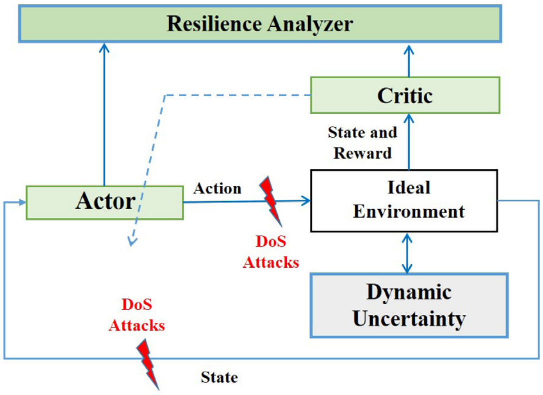

**Fig. 1.** The learning framework for the closed-loop system under DoS attacks.

where $|\Pi_D(t_a, t_b)|$ denotes the Lebesgue measure of the set $\Pi_D(t_a, t_b)$.

In the rest of this section, we will propose several model-free RL algorithms and analyze the robustness and resilience of the closed-loop systems.

### 4.2. Policy iteration

To begin with, we rewrite the augmented system (7) by:

$$\dot{\xi} = \bar{A}_k\xi + \bar{B}\,(K_k\xi + w) + \bar{D}v, \tag{28}$$

where $\bar{A}_k = \bar{A} - \bar{B}K_k$, and $w = u + \Phi$. The idea of policy iteration is to implement both policy evaluation

$$0 = \bar{A}_k^T P_k + P_k\bar{A}_k + Q + K_k^T K_k \tag{29}$$

and policy improvement

$$K_{k+1} = \bar{B}^T P_k \tag{30}$$

using online data by iterations. By Eqs. (28)–(30), we have

$$\xi^T(t_1)P_k\xi(t_1) - \xi^T(t_0)P_k\xi(t_0) = \int_{t_0}^{t_1}[-\xi^T(Q + K_k^T K_k)\xi + 2\xi^T K_{k+1}^T(K_k\xi + w) + 2\xi^T P_k\bar{D}v]d\tau. \tag{31}$$

Based on Assumption 6, there always exists a sequence $\{t_i\}_{i=0}^{\infty}$ such that the communication are allowed in all the following intervals $[t_0, t_1], [t_2, t_3], [t_4, t_5], \ldots$. For any two vectors $a, b$ and a sufficiently large number $s > 0$, define $\delta_a = [\mathrm{vecv}(a(t_1)) - \mathrm{vecv}(a(t_0)), \mathrm{vecv}(a(t_3)) - \mathrm{vecv}(a(t_2)), \ldots, \mathrm{vecv}(a(t_{2s+1})) - \mathrm{vecv}(a(t_{2s}))]^T$, $\Gamma_{a,b} = [\int_{t_0}^{t_1} a\otimes b d\tau, \int_{t_2}^{t_3} a\otimes b d\tau, \ldots, \int_{t_{2s}}^{t_{2s+1}} a\otimes b d\tau]^T$. Eq. (31) implies the following linear equation

$$\Psi_{PI}^{(k)}\begin{bmatrix} \mathrm{vecs}(P_k) \\ \mathrm{vec}(K_{k+1}) \\ \mathrm{vec}\left(\bar{D}^T P_k\right) \end{bmatrix} = \Phi_{PI}^{(k)}, \tag{32}$$

where $\Psi_{PI}^{(k)} = [\delta_\xi, -2\Gamma_{\xi,\xi}(I \otimes K_k^T) - 2\Gamma_{\xi,w}, -2\Gamma_{\xi_i,v}]$, $\Phi_{PI}^{(k)} = -\Gamma_{\xi,\xi}\mathrm{vec}(Q + K_k^T K_k)$. The uniqueness of solution to (32) is guaranteed under some rank conditions as shown below. For want of space, we omit the proof of lemma which follows the same line of proofs as in Gao and Jiang (2016) and Jiang and Jiang (2012).

**Lemma 2.** *For all $k \in \mathbb{Z}_+$, if there exists a $s^* \in \mathbb{Z}_+$ such that for all $s > s^*$,*

$$\mathrm{rank}([\Gamma_{\xi,\xi}, \Gamma_{\xi,w}, \Gamma_{\xi,v}]) = \frac{(n+q)(n+3q+3)}{2}, \tag{33}$$

*then the matrix $\Psi_{PI}^{(k)}$ has full column rank for all $k \in \mathbb{Z}_+$.*

Now, we are ready to present a data-driven PI Algorithm 1 to approximate the optimal values $K^*$ and $P^*$.

**Algorithm 1** Online Policy Iteration Algorithm

```
Select a small $c_1 > 0$ and an admissible control gain $K_0$. Apply any locally essentially bounded input $u$ on $[t_0, t_{2s+1}]$ such that (33) holds.
$k \leftarrow 0$
repeat
    Solve $P_k$ and $K_{k+1}$ from (32)
    $k \leftarrow k+1$
until $|P_k - P_{k-1}| < c_1$
```

### 4.3. Value iteration

Based on Bian and Jiang (2016), the nature of VI algorithm is to update the value matrix and control gain by

$$\begin{aligned} P_{k+1} =& \epsilon_k \left(\bar{A}^T P_k + P_k \bar{A} - P_k \bar{B}\bar{B}^T P_k + Q\right) + P_k, \\ K_k =& \bar{B}^T P_k, \end{aligned} \tag{34}$$

where $\epsilon_k$ is the step size satisfying $\epsilon_k > 0$, $\sum_{k=0}^{\infty} \epsilon_k = \infty$, $\sum_{k=0}^{\infty} \epsilon_k^2 < \infty$. Based on (28) and (34), we have

$$\begin{aligned} & \xi^T(t_1) P_k \xi(t_1) - \xi^T(t_0) P_k \xi(t_0) \\ =& \int_{t_0}^{t_1} \xi^T \mathcal{H}_k \xi + 2w^T K_k \xi + 2v^T \bar{D}^T P_k \xi d\tau, \end{aligned} \tag{35}$$

where $\mathcal{H}_k = \bar{A}^T P_k + P_k \bar{A}$.

(35) implies the following equation

$$\Psi_{VI}^{(k)} \begin{bmatrix} \mathrm{vecs}(\mathcal{H}_k) \\ \mathrm{vec}(K_k) \\ \mathrm{vec}\left(\bar{D}^T P_k\right) \end{bmatrix} = \Phi_{VI}^{(k)}, \tag{36}$$

where $\Psi_{VI}^{(k)} = [\delta_\xi, 2\Gamma_{\xi,w}, 2\Gamma_{\xi,v}]$, $\Phi_{VI}^{(k)} = -\Gamma_{\xi,\xi}\mathrm{vec}(P_k)$. By defining a collection of bounded set $\{\mathcal{B}_q\}_{q=0}^{\infty}$ as $\mathcal{B}_q \subset \mathcal{B}_{q+1}, q \in \mathbb{Z}_+$, $\lim\limits_{q\to\infty} \mathcal{B}_q = \mathcal{P}_n$. The online VI algorithm is presented in Algorithm 2.

**Algorithm 2** Online Value Iteration Algorithm

```
Select a small $c_2 > 0$. Apply any locally essentially bounded input $u$ on $[t_0, t_{2s+1}]$ such that (33) holds.
$k \leftarrow 0$, $q \leftarrow 0$. Choose a positive definite $P_0 \succ 0$.
repeat
    Solve $\mathcal{H}_k$ and $K_k$ from (36)
    $\tilde{P}_{k+1} \leftarrow P_k + \epsilon_k(\mathcal{H}_k + Q - K_k^T K_k)$
    if $\tilde{P}_{k+1} \notin \mathcal{B}_q$ then
        $P_{k+1} \leftarrow P_0, q \leftarrow q+1.$
    else $P_{k+1} \leftarrow \tilde{P}_{k+1}$
    end if
    $k \leftarrow k+1$
until $|P_k - P_{k-1}| < c_2 \epsilon_k$
```

### 4.4. Hybrid iteration

The VI can be initialized from any initial control policy to learn, but the convergence rate is usually not satisfactory. The PI is a variant of Newton–Raphson method, which can ensure the quadratic convergence (Jiang et al., 2020a), but its implementation must rely on an admissible control policy. We will propose a novel HI algorithm, Algorithm 3, to combine PI and VI efficiently.

We will show in Theorem 2 that there always exists an upper bound on the number of iterations to achieve an admissible controller using HI. Before that, let us present a Lemma 3 to facilitate the proof of Theorem 2.

**Algorithm 3** Online Hybrid Iteration Algorithm

```
Select a small $c_3 > 0$. Apply any locally essentially bounded input $u$ on $[t_0, t_{2s+1}]$ such that (33) holds.
$k \leftarrow 0$, $q \leftarrow 0$. $\underline{k} \leftarrow 0$, $\bar{k}_q \leftarrow \left(\frac{\sup_{P\in\mathcal{B}_q} |P|}{\epsilon\lambda_m(Q)}\right)^2 + 2$. Choose a $P_0 \succ 0$.
repeat
    Solve $\mathcal{H}_k$ and $K_k$ from (36)
    $\tilde{P}_{k+1} \leftarrow P_k + \epsilon(\mathcal{H}_k + Q - K_k^T K_k)$
    if $\tilde{P}_{k+1} \notin \mathcal{B}_q$ then
        $P_{k+1} \leftarrow P_0, q \leftarrow q+1.$ $\epsilon \leftarrow \epsilon/2$, $\underline{k} \leftarrow k$, $\bar{k}_q \leftarrow \left(\frac{\sup_{P\in\mathcal{B}_q} |P|}{\epsilon\lambda_m(Q)}\right)^2 + 2$
    else $P_{k+1} \leftarrow \tilde{P}_{k+1}$
    end if
    $k \leftarrow k+1$
until $(k > \underline{k} + \bar{k}_q + 1)$ or $\left(\mathcal{H}_{k-1} \prec 2K_{k-1}^T K_{k-1}\right.$ and $\left.P_{k-1} \succ 0\right)$
$k \leftarrow k-1$
repeat
    Solve $P_k$ and $K_{k+1}$ from (32)
    $k \leftarrow k+1$
until $|P_k - P_{k-1}| < c_3$
```

**Lemma 3.** *Let sequences $\{P_k^\epsilon\}_{k=0}^{\infty}$ and $\{K_k^\epsilon\}_{k=1}^{\infty}$ determined by the following equations*

$$\begin{aligned} P_{k+1}^\epsilon =& A_\epsilon^T P_k^\epsilon A_\epsilon + \epsilon Q - \frac{A_\epsilon^T P_k^\epsilon B_\epsilon B_\epsilon^T P_k^\epsilon A_\epsilon}{\epsilon + B_\epsilon^T P_k^\epsilon B_\epsilon}, \\ K_k^\epsilon =& \frac{B_\epsilon^T P_k^\epsilon A_\epsilon}{\epsilon + B_\epsilon^T P_k^\epsilon B_\epsilon} \end{aligned} \tag{37}$$

*where $A_\epsilon = I + \bar{A}\epsilon$ and $B_\epsilon = \bar{B}\epsilon$, and $P_0^\epsilon = P_0 \succ 0$. For any finite $\bar{k} \in \mathbb{Z}_+$ and any small $\delta \in \mathbb{R}_+$, there exists a $\epsilon^* \in \mathbb{R}_+$ such that $|K_{\bar{k}}^\epsilon - K_{\bar{k}}| < \delta$ and $|P_{\bar{k}}^\epsilon - P_{\bar{k}}| < \delta$ for any $\epsilon \in (0, \epsilon^*]$, where sequences $\{P_k\}_{k=0}^{\infty}$ and $\{K_k\}_{k=1}^{\infty}$ are determined by (34) starting from $P_0$.*

**Proof.** We will prove by induction.

1. Letting $k = 1$, Eqs. (37) can be rewritten by
$$\begin{aligned} P_1^\epsilon =& \epsilon\left(\bar{A}^T P_0 + P_0 \bar{A} - P_0 \bar{B}\bar{B}^T P_0 + Q\right) + P_0 + O_1(\epsilon) \\ =& P_1 + O_1(\epsilon), \\ K_1^\epsilon =& \bar{B}^T P_0 + O_2(\epsilon) = K_1 + O_2(\epsilon) \end{aligned} \tag{38}$$
where, for $i = 1, 2$, $\limsup_{\epsilon\to 0} |O_i(\epsilon)/\epsilon| < \infty$.
2. Suppose, for $k = l > 1$, we have $P_l^\epsilon - P_l = O_3(\epsilon)$ and $K_l^\epsilon - K_l = O_4(\epsilon)$, where $\limsup_{\epsilon\to 0} |O_i(\epsilon)/\epsilon| < \infty$, for $i = 3, 4$. By (37), we have
$$\begin{aligned} P_{l+1}^\epsilon =& \epsilon\left(\bar{A}^T P_l^\epsilon + P_l^\epsilon \bar{A} - P_l^\epsilon \bar{B}\bar{B}^T P_l^\epsilon + Q\right) + P_l^\epsilon + O_5(\epsilon) \\ =& \epsilon\left(\bar{A}^T P_l + P_l \bar{A} - P_l \bar{B}\bar{B}^T P_l + Q\right) + P_l + O_6(\epsilon) \\ =& P_{l+1} + O_6(\epsilon), \\ K_{l+1}^\epsilon =& \bar{B}^T P_l^\epsilon + O_7(\epsilon) = \bar{B}^T P_l + \bar{B}^T O_3(\epsilon) + O_7(\epsilon) \\ =& K_{l+1} + \bar{B}^T O_3(\epsilon) + O_7(\epsilon) \end{aligned} \tag{39}$$
where $\limsup_{\epsilon\to 0} |O_i(\epsilon)/\epsilon| < \infty$, for $i = 5, 6, 7$.

The proof is thus completed. □

**Theorem 2.** *There always exists a $\epsilon \in (0, \epsilon^*]$ such that, for any iteration $k > k_{a1}$ the control gain $K_k$ determined by the HI Algorithm 3 is admissible, where $k_{a1} = \underline{k} + \bar{k}_q$.*

**Proof.** Inspired from the fact that VI provides a finite horizon optimal solution (Granzotto et al., 2021), we construct the following discrete-time optimal control problem

$$\min_{\tilde{u}_k}\left(\tilde{\xi}_N^T P_0 \tilde{\xi}_N + \sum_{k=0}^{N-1} \epsilon\left(\tilde{\xi}_k^T Q \tilde{\xi}_k + \tilde{u}_k^2\right)\right)$$
$$\text{s.t.}\quad \tilde{\xi}_{k+1} = A_\epsilon \tilde{\xi}_k + B_\epsilon \tilde{u}_k, \forall k = 0, 1, \ldots, N-1.$$

Based on Assumption 1, one can always choose a small enough $\epsilon$ such that the pair $(A_\epsilon, B_\epsilon)$ is stabilizable. The corresponding optimal feedback control gain is $K_N^\epsilon$, which can be solved through updating Eqs. (37) by iterations. By Grimm et al. (2005, Corollary 3), we have the matrix $A_\epsilon - B_\epsilon K_N^\epsilon$ is Schur for any

$$N > \frac{\bar{a}(\bar{a} + \bar{a}_W)}{a_W^2} + 1 \tag{40}$$

where $\bar{a}_W = 0$, $a_W = \lambda_m(\epsilon Q)$ and $\bar{a} = \max_{0 \le k \le N} |P_k^\epsilon| = \max_{0 \le k \le N} |P_k| + \delta$ in this paper. From Bian and Jiang (2016), if the step size $\epsilon$ is small enough, we observe that $\bar{a} \le \sup_{P \in \mathcal{B}_q} |P| + \delta$, where the set $\mathcal{B}_q$ is bounded. Then, one can always find a small enough $\delta$ such that the right side of Eq. (40) is no greater than $\bar{k}_q$. Based on Lemma 3, there always exists a small pair $(\epsilon, \delta)$ such that the matrix $A_\epsilon - B_\epsilon K_N$ is Schur. It is immediate to have that the closed-loop matrix $\bar{A} - \bar{B}K_N$ of the continuous-time system (15) is Hurwitz, and the cost (14) is finite with respect to a controller $\tilde{u} = -K_N \tilde{\xi}$. In other words, $K_{\underline{k}+N}$ is an admissible control gain for the original continuous-time optimal control problem (14)–(15) for any $N > \bar{k}_q$, which completes the proof. □

By introducing the following Lemma, one can find another sufficient condition to ensure the admissibility of the learned control gain via the HI Algorithm 3.

**Lemma 4.** *Considering a pair $(\mathcal{H}_k, K_k)$ solved from Eq. (36) at an iteration $k \in \mathbb{Z}_+$. If both inequalities $\mathcal{H}_k - 2K_k^T K_k \prec 0$ and $P_k \succ 0$ hold, the control gain $K_k$ must be admissible. Moreover, there always exists a $k_{a2} \in \mathbb{Z}_+$ such that these inequalities hold when $k = k_{a2}$.*

**Proof.** Based on Bian and Jiang (2016), we have $\mathcal{H}_k = \bar{A}^T P_k + P_k \bar{A}$ and $K_k = \bar{B}^T P_k$, where $P_k \in \mathcal{B}_q$. One can observe that

$$\begin{aligned} 0 \succ & \mathcal{H}_k - 2K_k^T K_k \\ = & \bar{A}^T P_k + P_k \bar{A} - 2\bar{B}^T P_k \\ = & (\bar{A} - \bar{B}K_k)^T P_k + P_k(\bar{A} - \bar{B}K_k). \end{aligned}$$

Combining with the condition that $P_k \succ 0$, it is sufficient to show that $\bar{A} - \bar{B}K_k$ is Hurwitz through Lyapunov stability analysis, and the cost is finite with respect to the closed-loop system. Therefore, $K_k$ is admissible if $\mathcal{H}_k - 2K_k^T K_k \prec 0$ and $P_k \succ 0$. Moreover, based on the fact that $\lim_{k\to\infty} P_k \succ 0$ and $\lim_{k\to\infty} \mathcal{H}_k - 2K_k^T K_k = -Q - (K^*)^T K^* \prec 0$, there must exist a finite integer $k_{a2}$ such that $\mathcal{H}_{k_{a2}} - 2K_{k_{a2}}^T K_{k_{a2}} \prec 0$ and $P_{k_{a2}} \succ 0$. The proof is thus completed. □

Now, we are ready to prove the convergence of the HI Algorithm 3 in the following theorem.

**Theorem 3.** *Sequences $\{P_k\}_{k=0}^\infty$ and $\{K_k\}_{k=1}^\infty$ learned by the HI Algorithm 3 converge to $P^*$ and $K^*$, quadratically.*

**Proof.** Based on Theorem 2 and Lemma 4, the condition in the step 11 of the Algorithm 3 triggers at the iteration $k_a \le \min\{k_{a1} + 1, k_{a2} + 1\}$. This is sufficient to ensure that the learned control gain $K_{k_a - 1}$ is admissible. After finding out this admissible control gain, we switch to PI methods in steps 12–15. Given an admissible control gain $K_k$, one can solve $P_k$ and $K_{k+1}$ from (32) which are equivalent to the solution to (29)–(30). From (29)–(30), we have

$$\begin{aligned} &(A - BK_{k+1})^T(P_{k+1} - P^*) + (P_{k+1} - P^*)(A - BK_{k+1}) \\ =& -Q - K_{k+1}^T R K_{k+1} - (A - BK^*)^T P^* - P^*(A - BK^*) \\ & - (K^* - K_{k+1})^T B^T P^* - P^* B(K^* - K_{k+1}) \\ =& -Q - K_{k+1}^T R K_{k+1} + Q + K^* R K^* \\ & - (K^* - K_{k+1})^T R K^* - K^* R(K^* - K_{k+1}) \\ =& -(K_{k+1} - K^*)^T R(K_{k+1} - K^*) \\ =& -(P_k - P^*)BR^{-1}B^T(P_k - P^*). \end{aligned} \tag{41}$$

Based on (41), we have

$$P_{k+1} - P^* = \int_0^\infty e^{(A - BK_{k+1})\tau}(P_k - P^*)BR^{-1}B^T (P_k - P^*)e^{(A - BK_{k+1})\tau} d\tau,$$

which implies that there exist constants $c_p$ and $c_k$ such that

$$\lim_{k\to\infty} \frac{|P_{k+1} - P^*|}{|P_k - P^*|^2} = c_p, \lim_{k\to\infty} \frac{|K_{k+1} - K^*|}{|K_k - K^*|^2} = c_k. \tag{42}$$

Therefore, both sequences $\{P_k\}_{k=0}^\infty$ and $\{K_k\}_{k=1}^\infty$ converge quadratically to their optimal values. The proof is thus completed. □

**Remark 3.** Similar to existing RADP algorithms (Jiang & Jiang, 2017), we use $\Phi(t)$ as a measurable signal during the learning phase in the Algorithm 3. After the learning completes, the implementation of controller no longer depends on $\Phi(t)$. Note that it is unnecessary to measure $\Phi(t)$ exactly. It can be relaxed, for instance, by the measurement of a biased signal $\Phi_b(t) = \Phi(t) + \tilde{\Phi}(t)$, where $|\tilde{\Phi}(t)|$ is upper bounded for $t \in [t_0, t_{2s+1}]$. In terms of robust dynamic programming and robust PI techniques (Bian & Jiang, 2019; Pang et al., 2022), it can be shown that the approximated solution learned by HI will eventually enter a small neighborhood of the optimal solution under the relaxed condition.

### 4.5. Resilience and robustness analysis of closed-loop systems under DoS

Considering the effect of DoS, the control input and internal model applied to the process can be expressed as

$$u(t) = -K^* \xi(t_{k(t)}), \tag{43}$$
$$\dot{z}(t) = Sz(t) + G_2 e(t_k(t)), \tag{44}$$

where $t_{k(t)}$ represents the last time instant that receives the updated information. Define the error between the last successfully received values and actual values as $\epsilon_\xi(t) = \xi(t_{k(t)}) - \xi(t)$, $\epsilon_e(t) = e(t_{k(t)}) - e(t)$. We propose the following Theorem to seek a lower bound of DoS duration parameter $T$ to ensure the robust optimal output regulation under DoS.

**Theorem 4.** *Under Assumptions 1–4, the system (1)–(4) in closed-loop with the optimal controller (43) and internal model (44) under DoS attacks achieves global output regulation if*

1. *The DoS duration criterion $T$ satisfies*

$$T > 1 + \frac{4\lambda_M(P^*)(|K^*|^2 + |P^*||G_2\bar{C}|)}{\lambda_m(Q)\lambda_m(P^*)} := T^*; \tag{45}$$

2. *The small-gain condition* $\gamma_\Phi \gamma_e < 1$ *is satisfied where*

$$\gamma_e = |\bar{C}| \sqrt{\frac{1 + 2\exp\left(\frac{\kappa T^* \lambda_m(Q)}{\lambda_M(P^*)} + \eta\right) \frac{e^{\beta\tau_D}}{1-e^{-\beta\tau_D}}}{\min\left\{\frac{\lambda_m(Q)\lambda_m(P^*)}{\lambda_M(P^*)}, 4(|K^*|^2 + |P^*||G_2\bar{C}|)\right\}}} \tag{46}$$

*with* $\beta = \frac{\lambda_m(Q)}{\lambda_M(P^*)}\left(1 - \frac{T^*}{T}\right)$.

**Proof.** The system (2) in closed-loop with the controller (43) and internal model (44) is

$$\begin{aligned} \dot{\xi}(t) =& (\bar{A} - \bar{B}\bar{K}^*)\xi(t) + \bar{B}(-K^*\epsilon_\xi(t) + \Phi(t)) \\ &+ \bar{D}v(t) + \begin{bmatrix} 0 \\ G_2\epsilon_e(t) \end{bmatrix}. \end{aligned} \tag{47}$$

Letting $\Xi = \begin{bmatrix} X^T & Z^T \end{bmatrix}^T$, we have

$$\epsilon_\xi(t) = \xi(t_{k(t)}) - \xi(t) = \tilde{\xi}(t_{k(t)}) - \tilde{\xi}(t), \tag{48}$$

$$\begin{aligned} \epsilon_e(t) =& e(t_{k(t)}) - e(t) \\ =& \bar{C}(\xi(t_{k(t)}) - \Xi v(t_{k(t)})) - \bar{C}(\xi(t) - \Xi v(t)) \\ =& \bar{C}\epsilon_\xi(t). \end{aligned} \tag{49}$$

One can obtain the following error system

$$\begin{aligned} \dot{\tilde{\xi}}(t) =& (\bar{A} - \bar{B}\bar{K}^*)\tilde{\xi}(t) + \bar{B}(-K^*\epsilon_\xi(t) + \Phi(t)) + \begin{bmatrix} 0 \\ G_2\bar{C}\epsilon_\xi(t) \end{bmatrix}, \\ e(t) =& \bar{C}\tilde{\xi}(t). \end{aligned} \tag{50}$$

Before analyzing the stability of the interconnected system, we give and prove the following Lemma to show that the system (50) is IOS.

**Lemma 5.** *For any DoS duration criterion* $T > T^*$*, the system* (50) *regarding* $\Phi(t)$ *as the input and* $e(t)$ *as the output is IOS and SUO with zero offset.*

**Proof.** By taking $V = \tilde{\xi}^T P^* \tilde{\xi}$ as a Lyapunov function, along the closed-loop system (50), we have

$$\begin{aligned} \frac{d}{dt}V \leq& -\tilde{\xi}^T[Q + (K^*)^T K^*]\tilde{\xi} - 2(K^*\tilde{\xi})^T K^*\epsilon_\xi \\ &+ 2(K^*\tilde{\xi})^T\Phi + 2|\tilde{\xi}||P^*||G_2\bar{C}\epsilon_\xi| \\ \leq& -\lambda_m(Q)|\tilde{\xi}|^2 + 2(|K^*|^2 + |P^*||G_2\bar{C}|)|\tilde{\xi}||\epsilon_\xi| + |\Phi|^2. \end{aligned} \tag{51}$$

Considering the internal $[h_s + \tau_s, h_{s+1})$ where communications are normal, i.e., $\epsilon_\xi = 0$, and by the fact that $\lambda_m(P^*)|\tilde{\xi}|^2 \leq V \leq \lambda_M(P^*)|\tilde{\xi}|^2$, we have $\frac{d}{dt}V \leq -\frac{\lambda_m(Q)}{\lambda_M(P^*)}V + |\Phi|^2$, which implies that

$$V(\tilde{\xi}(t)) \leq e^{-w_1(t-h_s-\tau_s)}V(\tilde{\xi}(h_s + \tau_s)) + \gamma_1\|\Phi\|^2, \tag{52}$$

where $w_1 = \frac{\lambda_m(Q)}{\lambda_M(P^*)}$, and $\gamma_1 = \frac{1}{w_1} = \frac{\lambda_M(P^*)}{\lambda_m(Q)}$.

During the interval $[h_s, h_s + \tau_s)$ that communications are denied, the error is bounded by $|\epsilon_\xi(t)| \leq |\tilde{\xi}(h_s)| + |\tilde{\xi}(t)|$. (51) is equivalent to

$$\begin{aligned} \frac{d}{dt}V \leq& [2(|K^*|^2 + |P^*||G_2\bar{C}|) - \lambda_m(Q)]|\tilde{\xi}|^2 + |\Phi|^2 \\ &+ 2(|K^*|^2 + |P^*||G_2\bar{C}|)|\tilde{\xi}||\tilde{\xi}(h_s)| \\ \leq& 2(|K^*|^2 + |P^*||G_2\bar{C}|)(|\tilde{\xi}|^2 + |\tilde{\xi}||\tilde{\xi}(h_s)|) + |\Phi|^2 \\ \leq& w_2 \max\{V(\tilde{\xi}(t)), V(\tilde{\xi}(h_s))\} + |\Phi|^2, \end{aligned} \tag{53}$$

where $w_2 = \frac{4(|K^*|^2 + |P^*||G_2\bar{C}|)}{\lambda_m(P^*)}$. Then, for any $t \in [h_s, h_s + \tau_s)$, we have

$$V(\tilde{\xi}(t)) \leq e^{w_2(t-h_s)}V(\tilde{\xi}(h_s)) + \gamma_2 e^{w_2(t-h_s)}\|\Phi\|^2, \tag{54}$$

where $\gamma_2 = \frac{1}{w_2} = \frac{\lambda_m(P^*)}{4(|K^*|^2 + |P^*||G_2\bar{C}|)}$.

By De Persis and Tesi (2015, Lemma 3), for all $t \geq 0$, the Lyapunov function satisfies

$$\begin{aligned} V(\tilde{\xi}(t)) \leq& e^{-w_1|\Pi_N(0,t)|}e^{w_2|\Pi_D(0,t)|}V(\tilde{\xi}_0) \\ &+ \gamma_3\left[1 + 2\sum_{s\in\mathbb{Z}_+\backslash 0; h_s\leq t} e^{-w_1|\Pi_N(h_s+\tau_s,t)|}e^{w_2|\Pi_D(h_s,t)|}\right]\|\Phi\|^2, \end{aligned} \tag{55}$$

where $\gamma_3 = \max\{\gamma_1, \gamma_2\}$, and $\tilde{\xi}_0 = \tilde{\xi}(0)$. Based on Assumption 6, we have $|\Pi_D(h_s, t)| \leq \kappa + \frac{t-h_s}{T}, \forall t \geq h_s$. For the normal communication duration, we have $|\Pi_N(h_s + \tau_s, t)| = t - h_s - |\Pi_D(h_s, t)|, \forall t \geq h_s$. Therefore, one can observe that

$$\begin{aligned} &\sum_{s\in\mathbb{Z}_+\backslash 0; h_s\leq t} e^{-w_1|\Pi_N(h_s+\tau_s,t)|}e^{w_2|\Pi_D(h_s,t)|} \\ \leq& e^{w_3\kappa}\sum_{s\in\mathbb{Z}_+\backslash 0; h_s\leq t} e^{-\beta(t-h_s)}, \end{aligned} \tag{56}$$

where

$$\begin{aligned} w_3 =& w_1 + w_2 \\ =& \frac{\lambda_m(Q)\lambda_m(P^*) + 4\lambda_M(P^*)(|K^*| + |P^*||G_2\bar{C}|)}{\lambda_M(P^*)\lambda_m(P^*)}, \\ \beta =& w_1 - \frac{w_3}{T} \\ =& \frac{\lambda_m(Q)\lambda_m(P^*)(T-1) - 4\lambda_M(P^*)(|K^*|^2 + |P^*||G_2\bar{C}|)}{\lambda_M(P^*)\lambda_m(P^*)T}. \end{aligned}$$

It is easy to check that $T^*$ defined in (45) is equivalent to $T^* = w_3/w_1$. By De Persis and Tesi (2015, Lemma 4) and Assumption 5, we have

$$\sum_{s\in\mathbb{Z}_+\backslash 0; h_s\leq t} e^{-\beta(t-h_s)} \leq \frac{e^{-\beta\tau_D\eta}}{1 - e^{-\beta\tau_D}}.$$

Finally, we have the Lyapunov function along the trajectory of the closed-loop system satisfies

$$V(\tilde{\xi}(t)) \leq e^{\kappa w_3 - \beta t}V(\tilde{\xi}_0) + \gamma_3\left(1 + 2e^{\kappa w_3}\frac{e^{\beta\tau_D\eta}}{1 - e^{-\beta\tau_D}}\right)\|\Phi\|^2.$$

This implies that

$$\begin{aligned} |\tilde{\xi}(t)| \leq& \sqrt{e^{\kappa w_3 - \beta t}\frac{\lambda_M(P^*)}{\lambda_m(P^*)}}|\tilde{\xi}_0| + \gamma_4\|\Phi\|^2, \\ |e(t)| \leq& |\bar{C}|\sqrt{e^{\kappa w_3 - \beta t}\frac{\lambda_M(P^*)}{\lambda_m(P^*)}}|\tilde{\xi}_0| + \gamma_e\|\Phi\|^2, \end{aligned} \tag{57}$$

where

$$\begin{aligned} \gamma_4 =& \left[\frac{\gamma_3}{\lambda_m(P^*)}\left(1 + 2e^{\kappa w_3}\frac{e^{\beta\tau_D\eta}}{1 - e^{-\beta\tau_D}}\right)\right]^{1/2}, \\ \gamma_e =& |\bar{C}|\sqrt{\frac{1 + 2\exp(\kappa w_1 T^* + \eta)\frac{e^{\beta\tau_D}}{1-e^{-\beta\tau_D}}}{\min\left\{\frac{\lambda_m(Q)\lambda_m(P^*)}{\lambda_M(P^*)}, 4(|K^*|^2 + |P^*||G_2\bar{C}|)\right\}}}, \end{aligned}$$

which is equivalent to (46).

From (57), we see that a sufficient condition to make the system input-to-state stable is letting $\beta > 0$. In this setting, one can obtain that the DoS duration criterion $T$ has to satisfy the inequality (45). By Jiang et al. (1994), we conclude that the input-to-state stability guarantees that (50) with $e$ as output has SUO property with zero-offset and IOS properties with IOS-gain as $\gamma_e$. □

Assumptions 3 and 4 indicate that the $\zeta$-system (1) is SUO with zero-offset and IOS with IOS gain $\gamma_\Phi$. By the small-gain theory (Jiang & Liu, 2018; Jiang et al., 1994), under the small-gain condition, the interconnected system (1) and (50) is globally asymptotically stable at the origin for any $v \in \Sigma_v$. We have $\lim_{t\to\infty}\left(x(t) - Xv(t)\right) = 0$, and $\lim_{t\to\infty} e(t) = 0$, which immediately implies that both disturbance rejection and asymptotic tracking are achieved in global sense. The proof is thus completed. □

**Remark 4.** One can observe from (46) that the IOS-gain $\gamma_e$ is monotonically increasing with $\kappa$ and $\eta$, but not with $\tau_D$. Actually, when $\tau_D = \log(2)/\beta := \tau_D^*$, $\gamma_e$ achieves its minimum, i.e.,

$$\gamma_e^* = |\bar{C}| \sqrt{\frac{1+8\exp\left(\frac{\kappa T^*\lambda_m(Q)}{\lambda_M(P^*)}+\eta\right)}{\min\left\{\frac{\lambda_m(Q)\lambda_m(P^*)}{\lambda_M(P^*)},\, 4(|K^*|^2+|P^*||G_2\bar{C}|)\right\}}} \qquad (58)$$

where $1/\tau_D^*$ is named by DoS critical frequency.

## 5. Simulation

In order to validate the propose control approach, we consider an interconnection of two synchronous generators wherein the generator 2 is regarded as the dynamic uncertainty of generator 1 (Gao et al., 2016). The interconnected system can be modeled using (1)–(4), where $x, \zeta \in \mathbb{R}^2$ are the angle, rotor speed of the generators 1 and 2, respectively. The input $u$ is the mechanical power of the generator 1. $\Delta$ is a locally Lipschitz function such that the system (1) is IOS and SUO with $\gamma_\Delta = \alpha s$. $v$ is the desired rotor angle of the generator 1. System matrices and functions are

$$A = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{D_1}{2H_1} \end{bmatrix}, B = \begin{bmatrix} 0 \\ \frac{\omega_0}{2H_i} \end{bmatrix}, D = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$
$$C = \begin{bmatrix} 1 & 0 \end{bmatrix}, F = -1, S = 0, G_2 = 1,$$
$$\Phi(e, \zeta, v) = -\frac{E_1E_2}{X}[\sin(e + v - \zeta_1) - \sin(v)],$$

where the definition and the value of all parameters can be found in Jiang and Jiang (2017, Chapter 5).

For the purpose of simulation, we select $\kappa = 0.35$, $\tau_D = 15$, $T = 21$, $\alpha = 0.1$, and $\eta = 1$. We implement HI Algorithm 3 to learn the optimal control gain. We choose the initial control gain as $K_0 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$. We find that the updated control gain is admissible after 91 iterations. This enables the transition from VI to PI to significantly accelerate the convergence speed such that the convergence is achieved after 2 more iterations. The approximated optimal control gain learned by HI is

$$K_{93}^{HI} = \begin{bmatrix} 1.75669, 1.02985, 1.002 \end{bmatrix}.$$

For reference, we calculate the optimal control gain through solving algebraic Riccati equation

$$K^* = \begin{bmatrix} 1.75095 & 1.02974 & 1.0 \end{bmatrix}.$$

Compared with the initial control policy, we find from Fig. 2 that the transient response and the tracking performance are much better after updating the controller by the learned approximated optimal one under DoS attacks. Based on the learned control gain and the value matrix, one can find the bound of DoS duration parameter $T^* = 4.21 \times 10^3$, and the IOS gain $\gamma_e = 1.04 \times 10^5$. Similar to De Persis and Tesi (2015), these are sufficient conditions to guarantee the resilience and stability of the closed-loop system. As shown in Fig. 2, the DoS duration bound $T^*$ can in practice be much smaller than the theoretic ones.

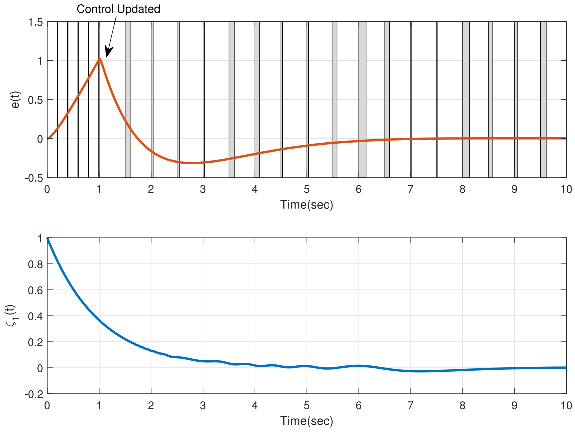

**Fig. 2.** The trajectories of angle differences of generator 1 ($e$) and generator 2 ($\zeta_1$) under DoS attack (shaded areas).

**Table 1**
Performance comparison of algorithms 1–3.

| Algorithm | PI | VI | HI |
|---|---|---|---|
| Need An Admissible $K_0$ | Yes | No | No |
| No. of iterations | 12 | 8469 | 116 |
| CPU time (sec) | 0.2064 | 1.7526 | 0.2549 |

In order to compare the computational complexity of different iteration algorithms, we randomly generate 200 dynamical systems and implement Algorithms 1–3 to learn the optimal control respectively. The average CPU time and the number of iterations of each algorithm needed for convergence are illustrated in Table 1 wherein one can see that the PI and HI algorithms significantly outperform the VI algorithm. The PI relies on a strong assumption of an initial stabilizing control gain which is hard to satisfy in practice.

## 6. Conclusion

In this paper, we studied the $LO^2RP$ with assured convergence rate requirement under the challenges from the unknown system and exosystem dynamics. Without relying on the knowledge of system dynamics and initial stabilizing feedback control gain, a novel VI algorithm is proposed, which is capable of learning the optimal regulator using online data with a guaranteed convergence rate. An application to a LCL coupled inverter-based distributed generation system shows the efficiency of the learning-based output regulation approach.

## References

Amin, S., Cárdenas, A. A., & Sastry, S. S. (2009). Safe and secure networked control systems under denial-of-service attacks. In *International workshop on hybrid systems: Computation and control* (pp. 31–45). Springer.

An, L., & Yang, G.-H. (2018). Decentralized adaptive fuzzy secure control for nonlinear uncertain interconnected systems against intermittent DoS attacks. *IEEE Transactions on Cybernetics*, *49*(3), 827–838.

Bian, T., & Jiang, Z. P. (2016). Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. *Automatica*, *71*, 348–360.

Bian, T., & Jiang, Z. P. (2019). Continuous-time robust dynamic programming. *SIAM Journal on Control and Optimization*, *57*(6), 4150–4174.

De Persis, C., & Tesi, P. (2015). Input-to-state stabilizing control under denial-of-service. *IEEE Transactions on Automatic Control*, *60*(11), 2930–2944.

Deng, C., & Wen, C. (2020). Distributed resilient observer-based fault-tolerant control for heterogeneous multiagent systems under actuator faults and DoS attacks. *IEEE Transactions on Control of Network Systems*, *7*(3), 1308–1318.

Feng, S., Cetinkaya, A., Ishii, H., Tesi, P., & De Persis, C. (2020). Networked control under DoS attacks: Tradeoffs between resilience and data rate. *IEEE Transactions on Automatic Control*, *66*(1), 460–467.

Feng, S., & Tesi, P. (2017). Resilient control under denial-of-service: Robust design. *Automatica*, *79*, 42–51.

Galarza-Jimenez, F., Poveda, J. I., Bianchin, G., & Dall'Anese, E. (2022). Extremum seeking under persistent gradient deception: A switching systems approach. *IEEE Control Systems Letters*, *6*, 133–138.

Gao, W., & Jiang, Z. P. (2015). Global optimal output regulation of partially linear systems via robust adaptive dynamic programming. In *Proceedings of the 1st Conference on Modelling, Identification and Control of Nonlinear Systems, Vol. 48* (11), (pp. 742–747). Saint-Petersburg, Russia.

Gao, W., & Jiang, Z. P. (2016). Adaptive dynamic programming and adaptive optimal output regulation of linear systems. *IEEE Transactions on Automatic Control*, *61*(12), 4164–4169.

Gao, W., & Jiang, Z. P. (2018). Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2614–2624.

Gao, W., & Jiang, Z. P. (2022). Learning-based adaptive optimal output regulation of linear and nonlinear systems: An overview. *Control Theory and Technology*, *20*, 1–19.

Gao, W., Jiang, Y., Jiang, Z. P., & Chai, T. (2016). Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming. *Automatica*, *72*, 37–45.

Granzotto, M., Postoyan, R., Buşoniu, L., Nešić, D., & Daafouz, J. (2021). Finite-horizon discounted optimal control: Stability and performance. *IEEE Transactions on Automatic Control*, *66*(2), 550–565.

Grimm, G., Messina, M., Tuna, S., & Teel, A. (2005). Model predictive control: for want of a local control Lyapunov function, all is not lost. *IEEE Transactions on Automatic Control*, *50*(5), 546–558.

Huang, J. (2004). *Nonlinear output regulation: Theory and applications*. Philadelphia, PA: SIAM.

Huang, J., & Chen, Z. (2004). A general framework for tackling the output regulation problem. *IEEE Transactions on Automatic Control*, *49*(12), 2203–2218.

Isidori, A. (2017). *Lectures in feedback design for multivariable systems*. Switzerland: Springer.

Jiang, Z. P., Bian, T., & Gao, W. (2020). Learning-based control: A tutorial and some recent results. *Foundations and Trends in Systems and Control*, *8*(3), 176–284.

Jiang, Y., Fan, J., Gao, W., Chai, T., & Lewis, F. L. (2020). Cooperative adaptive optimal output regulation of discrete-time nonlinear multi-agent systems. *Automatica*, *121*, Article 109149.

Jiang, Y., & Jiang, Z. P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, *48*(10), 2699–2704.

Jiang, Y., & Jiang, Z. P. (2017). *Robust adaptive dynamic programming*. Hoboken. NJ: Wiley.

Jiang, Y., Kiumarsi, B., Fan, J. L., Chai, T. Y., Li, J. N., & Lewis, F. L. (2020). Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning. *IEEE Transactions on Cybernetics*, *50*(7), 3147–3156.

Jiang, Z. P., & Liu, T. (2018). Small-gain theory for stability and control of dynamical networks: A survey. *Annual Reviews in Control*, *46*, 58–79.

Jiang, Z. P., Teel, A. R., & Praly, L. (1994). Small-gain theorem for ISS systems and applications. *Mathematics of Control, Signals, and Systems*, *7*(2), 95–120.

Kiumarsi, B., Vamvoudakis, K. G., Modares, H., & Lewis, F. L. (2018). Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2042–2062.

Liu, D., Xue, S., Zhao, B., Luo, B., & Wei, Q. (2021). Adaptive dynamic programming for control: A survey and recent advances. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(1), 142–160.

Marino, R., & Tomei, P. (2003). Output regulation for linear systems via adaptive internal model. *IEEE Transactions on Automatic Control*, *48*(12), 2199–2202.

Odekunle, A., Gao, W., Davari, M., & Jiang, Z. P. (2020). Reinforcement learning and non-zero-sum game output regulation for multi-player linear uncertain systems. *Automatica*, *112*, Article 108672.

Pang, B., Bian, T., & Jiang, Z. P. (2022). Robust policy iteration for continuous-time linear quadratic regulation. *IEEE Transactions on Automatic Control*, *67*(1), 504–511.

Powell, W. B. (2007). *Approximate dynamic programming: Solving the curse of dimensionality*. New York, NY: John Wiley & Sons.

Sontag, E. D. (1989). Smooth stabilization implies coprime factorization. *IEEE Transactions on Automatic Control*, *34*(4), 435–443.

Sontag, E. D. (2007). Input to state stability: Basic concepts and results. In *Nonlinear and optimal control theory* (pp. 163–220). Berlin: Springer-Verlag.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA: The MIT Press.

Teixeira, A., Shames, I., Sandberg, H., & Johansson, K. H. (2015). A secure control framework for resource-limited adversaries. *Automatica*, *51*, 135–148.

Vamvoudakis, K. G., & Kokolakis, N.-M. T. (2020). Synchronous reinforcement learning-based control for cognitive autonomy. *Foundations and Trends in Systems and Control*, *8*(1–2), 1–175.

Zhai, L., & Vamvoudakis, K. G. (2021). Data-based and secure switched cyber-physical systems. *Systems & Control Letters*, *148*, Article 104826.

Zhao, F., Gao, W., Liu, T., & Jiang, Z. P. (2022). Adaptive optimal output regulation of linear discrete-time systems based on event-triggered output-feedback. *Automatica*, *131*, Article 110103.

**Weinan Gao** 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, Shenyang, China, in 2013, and the Ph.D. degree in Electrical Engineering from New York University, Brooklyn, NY, in 2017. 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. He is the recipient of the US NSF Engineering Research Initiation Award, the Best Paper Award in IEEE International Conference on Real-time Computing and Robotics, and the David Goodman Research Award at New York University. Dr. Gao is an Associate Editor or a Guest Editor of IEEE/CAA Journal of Automatica Sinica, IEEE Transactions on Circuits and Systems II: Express Briefs, Neurocomputing, IEEE Transactions on Neural Network and Learning Systems, and Complex & Intelligent Systems, a member of Editorial Board of Neural Computing and Applications and Control Engineering Practice, and a Technical Committee Member in IEEE Control Systems Society on Nonlinear Systems and Control and in IFAC TC 1.2 Adaptive and Learning Systems.

**Chao Deng** received the Ph.D. degree in Control Engineering from Northeastern University, China, in 2018. From May 2018 to May 2021, he was a Research Fellow at the School of Electrical and Electronic Engineering at Nanyang Technological University, Singapore. Currently, he is a faculty at the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, China. His research interests include distributed secondary control of microgrid, and cyber–physical systems.

**Yi Jiang** was born in Ezhou, Hubei, China. He received the B.Eng. degree in automation, M.S. and Ph.D. degrees in control theory and control engineering from Information Science and Engineering College and State Key Laboratory of Synthetical Automation for Process Industries in Northeastern University, Shenyang, Liaoning, China in 2014, 2016 and 2020, respectively. From January to July 2017, he was a Visiting Scholar with the University of Texas at Arlington, TX, USA, and from March 2018 to March 2019, he was a Research Assistant with the University of Alberta, Edmonton, Canada. Currently, he is a Postdoc with the City University of Hong Kong, Hong Kong, China. His research interests include networked control systems, industrial process operational control, reinforcement learning and event-triggered control. Dr. Jiang is an Associate Editor for Advanced Control for Applications: Engineering and Industrial Systems and the recipient of Excellent Doctoral Dissertation Award from Chinese Association of Automation (CAA) in 2021.

**Zhong-Ping Jiang** received the M.Sc. degree in statistics from the University of Paris XI, France, in 1989, and the Ph.D. degree in automatic control and mathematics from the Ecole des Mines de Paris (now, called ParisTech-Mines), France, in 1993, under the direction of Prof. Laurent Praly.

Currently, he is a Professor of Electrical and Computer Engineering at the Tandon School of Engineering, New York University. His main research interests include stability theory, robust/adaptive/distributed nonlinear control, robust adaptive dynamic programming, reinforcement learning and their applications to information, mechanical and biological systems. In these fields, he has written six books and is author/co-author of over 500 peer-reviewed journal and conference papers.

Prof. Jiang is a recipient of the prestigious Queen Elizabeth II Fellowship Award from the Australian Research Council, CAREER Award from the U.S. National Science Foundation, JSPS Invitation Fellowship from the Japan Society for the Promotion of Science, Distinguished Overseas Chinese Scholar Award from the NSF of China, and several best paper awards. He has served as Deputy Editor-in-Chief, Senior Editor and Associate Editor for numerous journals. Prof. Jiang is a Fellow of the IEEE, a Fellow of the IFAC, a Fellow of the CAA and is among the Clarivate Analytics Highly Cited Researchers. In 2021, he is elected as a foreign member of the Academia Europaea.