2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Resilient Learning-Based Control Under Denial-of-Service Attacks

Sayan Chakraborty[1], Weinan Gao[2], Kyriakos G. Vamvoudakis[3], Zhong-Ping Jiang[1]

Abstract— **In this paper, we have proposed a resilient reinforcement learning method for discrete-time linear systems with unknown parameters, under denial-of-service (DoS) attacks. The proposed method is based on policy iteration that learns the optimal controller from input-state data amidst DoS attacks. We achieve an upper bound for the DoS duration to ensure closed-loop stability. The resilience of the closed-loop system, when subjected to DoS attacks with the learned controller and an internal model, has been thoroughly examined. The effectiveness of the proposed methodology is demonstrated on an inverted pendulum on a cart.**

I. INTRODUCTION

Reinforcement learning (RL) outlines strategies for an agent to adjust its actions when interacting with an unfamiliar environment, aiming to fulfill a long-term objective [1]. Researchers from the control community have used ideas from RL and adaptive/approximate dynamic programming (ADP) [2], [3], [4], [5] to develop data-driven adaptive optimal control methods to address the stabilization problem of dynamical systems (see [6], [7], [8], [9], [10], [11], [12]). As a generalization, the authors in [13] combined ADP with output regulation theory for asymptotic tracking and disturbance rejection, later subsequently extending this framework to a data-driven approach for non-linear systems in [14]. This approach has been extended to multi-agent systems in [15], [16] and references therein.

However, the existing ADP studies usually rely on the assumption that communication channels for control and measurement are ideal, which makes the designed controller vulnerable to cyberattacks. Consequently, it becomes crucial to extend the analysis of control systems beyond stability and robustness to include resilience, ensuring that systems can effectively withstand and recover from cyber threats. When a system is under DoS attack, the transmission of information is blocked among networks [17], [18]. An explicit characterization of DoS frequency and duration has been given in [19] such that the closed-loop system remains robustly stable under DoS attacks. Other research in this direction can be found in [20], [21], [22], [23]. However, the authors in the aforementioned references consider neither model-free controller design nor resilient analysis for the closed-loop system under DoS attacks. Most recently, learning-based approaches have been adopted in [24], [25], [26], [27] by using RL, ADP, and extremum seeking to defend the closed-loop system under adversarial attacks.

Most of the above-mentioned works have considered continuous-time systems. Recently, researchers have begun to explore learning-based approaches to study the resilience of closed-loop discrete-time systems under DoS attacks. In order to guarantee resilience under DoS attacks, many authors have adopted data-driven predictive control [28], [29] and adaptive control [30], [31] techniques to address the problem of learning-based control to guarantee the resilience of closed-loop discrete-time systems under DoS attacks. In this paper, we address the data-driven optimal output regulation problem for discrete-time systems with unknown parameters under DoS attacks using RL and ADP. The problem of output regulation focuses on the development of a feedback control law that ensures asymptotic tracking and disturbance rejection. To our knowledge, learning-based optimal output regulation for discrete-time systems under DoS attacks has been proposed for the first time. Our approach enables direct analysis of the closed-loop system's resilience with the optimal controller and achieves an upper bound for the DoS attack duration that the system can withstand while maintaining stability

The rest of the paper is organized as follows. Section II formulates the control objective and controller design in the absence of DoS attacks. Section III provides the resilience analysis of the closed-loop system under DoS attacks. Section IV provides the online learning method using policy iteration when the system is under DoS attacks. Section V presents the simulation results. Finally, the conclusion and future work are mentioned in Section VI.

Facts and notations: $\mathbb{R}_+$ denotes the set of non-negative real numbers. $\mathbb{Z}_+$ the set of non-negative integers. $|x|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^n$. $|A|$ denotes the induced matrix norm for a matrix $A \in \mathbb{R}^{m\times n}$. For a square matrix A, $\sigma(A)$ denotes the spectrum of A. For a real symmetric matrix A, $\lambda_m(A)$ and $\lambda_M(A)$ denote the minimum and maximum eigenvalues of A, respectively. For a symmetric positive definite matrix $P \in \mathbb{R}^{m\times m}$ and $x \in \mathbb{R}^m$, we have $\lambda_m(P)|x|^2 \leq x^{\mathrm{T}}Px \leq \lambda_M(P)|x|^2$. For any function $\zeta\colon \mathbb{Z}_+ \to \mathbb{R}^n$, $\|\zeta\| = \sup\{|\zeta_k|\colon k \in \mathbb{Z}_+\} \leq \infty$. A function $\alpha\colon \mathbb{R}_+ \to \mathbb{R}_+$ belongs to class $\mathcal{K}$ if it is continuous, strictly increasing and $\alpha(0) = 0$. A function α belongs to class $\mathcal{K}_\infty$ if it is of class $\mathcal{K}$ and also $\alpha(r) \to \infty$ as $r \to \infty$. A function $\beta\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ belongs to class $\mathcal{KL}$

[1]S. Chakraborty and Z.-P. Jiang are with the CAN Lab, New York University, Brooklyn, NY 11201 USA, sc8804@nyu.edu, zjiang@nyu.edu

[2]W. Gao is with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China, gaown@mail.neu.edu.cn

[3]K. G. Vamvoudakis is with The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, GA 30332-0150 USA, kyriakos@gatech.edu

This work was supported in part by the NSF under grant nos. CNS-2148309, EPCN-2210320, CPS-2227185, S&AS-1849198, CPS-1851588, and CPS-2227153, and by Minerva under grant No. N00014 − 18 − 1 − 2874.

if, for each fixed $k \geq 0$, the function $\beta(.,k)$ belongs to class $\mathcal{K}$, and for each fixed $r \geq 0$, the function $\beta(r,.)$ is decreasing and $\beta(r,k) \to 0$ as $k \to \infty$. For any $x,\ y \in \mathbb{R}^n$, and for any $\epsilon > 0$, we have $x^{\mathrm{T}}y \leq \frac{1}{4\epsilon}x^{\mathrm{T}}x + \epsilon y^{\mathrm{T}}y$. $\otimes$ indicates the Kronecker product, $\mathrm{vec}(T) = \left[t_1^{\mathrm{T}}, t_2^{\mathrm{T}}, \cdots, t_m^{\mathrm{T}}\right]^{\mathrm{T}}$ with $t_i \in \mathbb{R}^r$ being the columns of $T \in \mathbb{R}^{r\times m}$. For a symmetric matrix $P \in \mathbb{R}^{m\times m}$, $\mathrm{vecs}(P) = [p_{11}, 2p_{12}, \cdots, 2p_{1m}, p_{22}, 2p_{23}, \cdots, 2p_{(m-1)m}, p_{mm}]^{\mathrm{T}} \in \mathbb{R}^{(1/2)m(m+1)}$, for a column vector $v \in \mathbb{R}^n$, $\mathrm{vecv}(v) = [v_1^2, v_1v_2, \cdots, v_1v_n, v_2^2, v_2v_3, \cdots, v_{n-1}v_n, v_n^2]^{\mathrm{T}} \in \mathbb{R}^{(1/2)n(n+1)}$. For any two sequence of vectors $\{a_i\}_{i=k_0}^{k_s}$, $\{b_i\}_{i=k_0}^{k_s}$, define $\Xi_a = \left[\mathrm{vecv}(a_{k_0+1}) - \mathrm{vecv}(a_{k_0}), \cdots, \mathrm{vecv}(a_{k_s}) - \mathrm{vecv}(a_{k_s-1})\right]^{\mathrm{T}}$, $J_{a,b} = \left[a_{k_0} \otimes b_{k_0}, \cdots, a_{k_s} \otimes b_{k_s}\right]^{\mathrm{T}}$, $J_a = \left[\mathrm{vecv}(a_{k_0}), \cdots, \mathrm{vecv}(a_{k_s})\right]^{\mathrm{T}}$. I_n and 0_n are the identity matrix and the zero matrix of dimension $n \times n$, respectively.

II. Preliminaries and problem formulation

Consider the following discrete-time cascade system:

$$x_{k+1} = Ax_k + Bu_k + Dw_k, \tag{1}$$
$$w_{k+1} = Ew_k, \tag{2}$$
$$e_k = Cx_k + Fw_k, \tag{3}$$

where $k \in \mathbb{Z}_+$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^n$, $C \in \mathbb{R}^{1\times n}$, $D \in \mathbb{R}^{n\times q}$, $E \in \mathbb{R}^{q\times q}$, $F \in \mathbb{R}^{1\times q}$ are constant matrices, $e_k \in \mathbb{R}$ is the measurement output, $u_k \in \mathbb{R}$ the control input, $x_k \in \mathbb{R}^n$ is the state, $w_k \in \mathbb{R}^q$ is the state of the exosystem and $y_{dk} = -Fw_k$ is the reference signal.

Assumption 2.1: The pair (A, B) is stabilizable and the eigenvalues of E are simple on the unit circle. □

Assumption 2.2: $\mathrm{rank}\begin{pmatrix}\begin{bmatrix} A-\lambda I & B \\ C & 0\end{bmatrix}\end{pmatrix} = n+1,\ \forall \lambda \in \sigma(E)$. □

Definition 2.1: A dynamic compensator of the form

$$z_{k+1} = \mathcal{G}_1 z_k + \mathcal{G}_2 e_k, \quad \forall k \in \mathbb{Z}_+ \tag{4}$$

is called an internal model of the system (1)-(3) if the pair $(\mathcal{G}_1, \mathcal{G}_2)$ incorporates an internal model of the exosystem matrix E [32]. □

Remark 2.1: In this work, we let $\mathcal{G}_1 = E$, and choose $\mathcal{G}_2 \in \mathbb{R}^q$ such that the pair $(E, \mathcal{G}_2)$ is controllable. □

In the absence of DoS attacks, we show in Lemma 2.1 that it is possible to develop a state-feedback controller for the discrete-time system (1)-(3) with an internal model (4) that solves the output regulation problem. Consider the following augmented system $\forall k \in \mathbb{Z}_+$

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Dw_k, \\ w_{k+1} &= Ew_k, \\ e_k &= Cx_k + Fw_k, \\ z_{k+1} &= Ez_k + \mathcal{G}_2 e_k. \end{aligned} \tag{5}$$

Lemma 2.1: Under Assumptions 2.1-2.2, if there exists a state-feedback controller

$$u_k = -K_x x_k - K_z z_k, \quad \forall k \in \mathbb{Z}_+ \tag{6}$$

such that the closed-loop system matrix

$$A_c = \begin{bmatrix} A - BK_x & -BK_z \\ \mathcal{G}_2 C & E \end{bmatrix} \tag{7}$$

of the augmented system (5) is Schur. Then, the controller (6) solves the output regulation problem.

Proof: Assumptions 2.1-2.2 guarantee the existence of a unique pair (X, U) solving the following regulator equations

$$XE = AX + BU + D, \tag{8}$$
$$0 = CX + F. \tag{9}$$

Also, (9) combined with

$$XE = (A - BK_x)X - BK_zZ + D, \tag{10}$$
$$ZE = EZ + \mathcal{G}_2(CX + F), \tag{11}$$

have unique solutions $\hat{X}$ and Z (see Lemma 1.38 in [32]). This implies that $X = \hat{X}$ and $U = -K_xX - K_zZ$. By defining the error states as $\tilde{x}_k = x_k - Xw_k$ and $\tilde{z}_k = z_k - Zw_k$, the following error system of (5) can be derived using (10)-(11)

$$\tilde{x}_{k+1} = (A - BK_x)\tilde{x}_k - BK_z\tilde{z}_k, \tag{12}$$
$$\tilde{z}_{k+1} = \mathcal{G}_2 C\tilde{x}_k + E\tilde{z}_k. \tag{13}$$

Next, by defining $\tilde{\zeta}_k = [\tilde{x}_k^{\mathrm{T}} \quad \tilde{z}_k^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{n+q}$, and using (12)-(13) one can obtain

$$\tilde{\zeta}_{k+1} = A_c\tilde{\zeta}_k, \tag{14}$$
$$e_k = \bar{C}\tilde{\zeta}_k, \tag{15}$$

where $A_c = \bar{A} - \bar{B}K$, $\bar{C} = [C \quad 0]$, $\bar{B} = [B^{\mathrm{T}} \quad 0]^{\mathrm{T}}$, and $\bar{A} = \begin{bmatrix} A & 0 \\ \mathcal{G}_2 C & E \end{bmatrix}$, $K = [K_x \quad K_z]$.

Since A_c is Schur, we have $\lim_{k\to\infty}\tilde{\zeta}_k = 0$ and $\lim_{k\to\infty} e_k = 0$. Thus, Lemma 2.1 is proved. ■

As evident from Lemma 2.1, the output regulation properties of (5) are guaranteed under a state-feedback controller of the form (6). Furthermore, the optimal output regulation problem can be posed as follows.

Problem 2.1: In order to obtain the optimal state-feedback controller that solves the output regulation problem for (5), the following dynamic programming problem is solved

$$\min_{\tilde{u}} \sum_{k=0}^{\infty} (\tilde{\zeta}_k^{\mathrm{T}} Q \tilde{\zeta}_k + \tilde{u}_k^2) \tag{16}$$
$$\text{s.t} \quad \tilde{\zeta}_{k+1} = \bar{A}\tilde{\zeta}_k + \bar{B}\tilde{u}_k, \tag{17}$$

where $Q = Q^{\mathrm{T}} \succ 0$, $\tilde{u}_k = u_k - Uw_k$. □

Problem 2.1 is a standard discrete-time linear quadratic regulator problem. The solution to this problem is an optimal feedback controller of the form

$$\tilde{u}_k^\star = -K^\star\tilde{\zeta}_k, \tag{18}$$

where $K^\star = (1 + \bar{B}^{\mathrm{T}}P^\star\bar{B})^{-1}\bar{B}^{\mathrm{T}}P^\star\bar{A}$ and $P^\star = P^{\star\mathrm{T}} \succ 0$ solves the following discrete-time algebraic Riccati equation

$$\bar{A}^{\mathrm{T}}P^\star\bar{A} - P^\star + Q - \bar{A}^{\mathrm{T}}P^\star\bar{B}(1 + \bar{B}^{\mathrm{T}}P^\star\bar{B})^{-1}\bar{B}^{\mathrm{T}}P^\star\bar{A} = 0. \tag{19}$$

The optimal controller for (5) can be obtained $\forall k \in \mathbb{Z}_+$ as

$$u_k^\star = \tilde{u}_k^\star + Uw_k := -K_x^\star x_k - K_z^\star z_k. \tag{20}$$

III. Resilience analysis under DoS Attacks

In this paper, we examine scenarios where DoS attacks simultaneously impact both the measurement and control channels of the augmented system described by (5). It is assumed that, during DoS attacks, the transmission and reception of data are both disrupted. Let $\{h_m\}_{m\in\mathbb{Z}_+}$ denote the sequence of off/on transitions of DoS, where $h_0 \geq 0$. The m^{th} DoS attack interval of length τ_m is represented as $\mathcal{J}_m := [h_m, h_m + \tau_m)$. For each interval $[k_1, k_2]$, let $\Lambda_N(k_1, k_2)$ and $\Lambda_D(k_1, k_2)$ denote the set of time instants where communication is allowed and denied, respectively. Thus, $\Lambda_N(k_1, k_2)$ and $\Lambda_D(k_1, k_2)$ can be defined as follows

$$\Lambda_D(k_1, k_2) := \bigcup_{m\in\mathbb{Z}_+} \mathcal{J}_m \bigcap [k_1, k_2], \tag{21}$$

$$\Lambda_N(k_1, k_2) := [k_1, k_2] \setminus \Lambda_D(k_1, k_2). \tag{22}$$

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

Assumption 3.1: (*DoS Frequency*) There exist constants $\eta > 1$ and $\tau_D > 0$ such that $\forall k_2 > k_1 \geq 0$,

$$n(k_1, k_2) \leq \eta + \frac{k_2 - k_1}{\tau_D}, \tag{23}$$

where $n(k_1, k_2)$ denotes the number of DoS off/on transitions occurring on the interval $[k_1, k_2]$. □

Assumption 3.2: (*DoS Duration*) There exist constants $T > 1$ and $\kappa > 0$ such that $\forall k_2 > k_1 \geq 0$,

$$|\Lambda_D(k_1, k_2)| \leq \kappa + \frac{k_2 - k_1}{T}, \tag{24}$$

where $|\Lambda_D(k_1, k_2)|$ denotes the Lebesgue measure of the set $\Lambda_D(k_1, k_2)$. □

When the system is under DoS attack, the control input and internal model can be expressed $\forall k \in \mathbb{Z}_+$ as

$$u_k = -K^\star \zeta_{k_{m(k)}}, \tag{25}$$

$$z_{k+1} = Ez_k + \mathcal{G}_2 e_{k_{m(k)}}, \tag{26}$$

where $k_{m(k)}$ represents the most recent time instant at which the updated information is received. Let $\bar{\epsilon}_k = \zeta_{k_{m(k)}} - \zeta_k$ and $\underline{\epsilon}_k = e_{k_{m(k)}} - e_k$ be the error values between last successfully received values and actual values. Using the optimal controller (25) and the internal model (26), the following closed-loop system is obtained

$$\zeta_{k+1} = (\bar{A} - \bar{B}K^\star)\zeta_k - \bar{B}K^\star\bar{\epsilon}_k + \bar{D}w_k + \begin{bmatrix} 0 \\ \mathcal{G}_2\underline{\epsilon}_k \end{bmatrix}, \tag{27}$$

where $\zeta_k = [x_k^{\mathrm{T}} \quad z_k^{\mathrm{T}}]^{\mathrm{T}}$, $\bar{D} = [D^{\mathrm{T}} \quad (\mathcal{G}_2 F)^T]^{\mathrm{T}}$. Defining $\tilde{\zeta}_k = \zeta_k - \Xi w_k$, where $\Xi = [X^{\mathrm{T}} \quad Z^{\mathrm{T}}]^{\mathrm{T}}$, we have

$$\bar{\epsilon}_k = \zeta_{k_{m(k)}} - \zeta_k = \tilde{\zeta}_{k_{m(k)}} - \tilde{\zeta}_k \tag{28}$$

$$\underline{\epsilon}_k = e_{k_{m(k)}} - e_k = \bar{C}\bar{\epsilon}_k. \tag{29}$$

From (27)-(29) the following error system can be obtained

$$\begin{aligned} \tilde{\zeta}_{k+1} &= (\bar{A} - \bar{B}K^\star)\tilde{\zeta}_k - \bar{B}K^\star\bar{\epsilon}_k + \begin{bmatrix} 0 \\ \mathcal{G}_2\bar{C}\bar{\epsilon}_k \end{bmatrix}, \\ e_k &= \bar{C}\tilde{\zeta}_k. \end{aligned} \tag{30}$$

In this work, we seek to give a lower bound on the DoS duration parameter T, such that output regulation is achieved under DoS attacks. This is obtained in the following Theorem.

Theorem 3.1: The error system described in (30) is globally asymptotically stable at the origin if the following condition on the DoS duration parameter T holds

$$T > 1 + \frac{\log(1+\omega_2)}{-\log(1-\omega_1)} := T^\star, \tag{31}$$

where

$$\begin{aligned} \omega_1 &= \frac{\lambda_m(Q)}{\lambda_M(P^\star)}, \omega_2 = \frac{\alpha_1 + 4\alpha_2}{\lambda_m(P^\star)}, \\ \alpha_1 &= 1 + 2|K^{\star\mathrm{T}}\bar{B}^{\mathrm{T}}P^\star\bar{B}K^\star|^2 + 2|\bar{A}^{\mathrm{T}}P^\star\bar{A}|^2, \\ \alpha_2 &= 2 + 4|K^{\star\mathrm{T}}\bar{B}^{\mathrm{T}}P^\star\bar{B}K^\star|^2 + 4|\tilde{D}^{\mathrm{T}}P^\star\tilde{D}|^2. \end{aligned}$$

Proof: Defining the Lyapunov function $V = \tilde{\zeta}_k^{\mathrm{T}} P^\star \tilde{\zeta}_k$, the following can be obtained

$$\begin{aligned} V(\tilde{\zeta}_{k+1}) - V(\tilde{\zeta}_k) &= \tilde{\zeta}_{k+1}^{\mathrm{T}}P^\star\tilde{\zeta}_{k+1} - \tilde{\zeta}_k^{\mathrm{T}}P^\star\tilde{\zeta}_k \\ &\leq -\lambda_m(Q)|\tilde{\zeta}_k|^2 - |\sqrt{P^\star}\bar{B}K^\star\tilde{\zeta}_k \\ &\quad -\sqrt{P^\star}\bar{B}K^\star\bar{\epsilon}_k|^2 - |\sqrt{P^\star}\bar{A}\tilde{\zeta}_k - \sqrt{P^\star}\tilde{D}\bar{\epsilon}_k|^2 \\ &\quad -|\sqrt{P^\star}\bar{B}K^\star\bar{\epsilon}_k + \sqrt{P^\star}\bar{A}\tilde{\zeta}_k|^2 \\ &\quad -|\sqrt{P^\star}\bar{B}K^\star\tilde{\zeta}_k + \sqrt{P^\star}\tilde{D}\bar{\epsilon}_k|^2 \\ &\quad -|\sqrt{P^\star}\bar{B}K^\star\bar{\epsilon}_k + \sqrt{P^\star}\tilde{D}\bar{\epsilon}_k|^2 \\ &\quad +2\tilde{\zeta}_k^{\mathrm{T}}\bar{A}^{\mathrm{T}}P^\star\bar{A}\tilde{\zeta}_k + 2\tilde{\zeta}_k^{\mathrm{T}}K^{\star\mathrm{T}}\bar{B}^{\mathrm{T}}P^\star\bar{B}K^\star\tilde{\zeta}_k \\ &\quad +4\bar{\epsilon}_k^{\mathrm{T}}K^{\star\mathrm{T}}\bar{B}^{\mathrm{T}}P^\star\bar{B}K^\star\bar{\epsilon}_k + 4\bar{\epsilon}_k^{\mathrm{T}}\tilde{D}^{\mathrm{T}}P^\star\tilde{D}\bar{\epsilon}_k, \end{aligned} \tag{32}$$

where the inequality is obtained using completion of squares, $\tilde{D} = [0, \ (\mathcal{G}_2\bar{C})^{\mathrm{T}}]^{\mathrm{T}}$. From (32) and considering the interval $W_m := [h_m + \tau_m, h_{m+1})$ where communications are normal, i.e., $\bar{\epsilon}_k = 0$, we have $V(\tilde{\zeta}_{k+1}) - V(\tilde{\zeta}_k) \leq -\lambda_m(Q)|\tilde{\zeta}_k|^2$, which implies

$$V(\tilde{\zeta}_k) \leq (1-\omega_1)^{k-h_m-\tau_m} V(\tilde{\zeta}_{h_m+\tau_m}), \forall k \in W_m, \tag{33}$$

During the interval $Z_m := [h_m, h_m + \tau_m)$ where communications are denied, the error is bounded by $|\bar{\epsilon}_k| \leq |\tilde{\zeta}_{h_m}| + |\tilde{\zeta}_k|$ and (32) is equivalent to $V(\tilde{\zeta}_{k+1}) - V(\tilde{\zeta}_k) \leq (\alpha_1+\alpha_2)|\tilde{\zeta}_k|^2 + \alpha_2|\tilde{\zeta}_{h_m}|^2 + 2\alpha_2|\tilde{\zeta}_{h_m}||\tilde{\zeta}_k|$, which implies

$$V(\tilde{\zeta}_{k+1}) - V(\tilde{\zeta}_k) \leq \omega_2 \max\{V(\tilde{\zeta}_k), V(\tilde{\zeta}_{h_m})\}, \tag{34}$$

Then, $\forall k \in Z_m$ we have

$$V(\tilde{\zeta}_k) \leq (1+\omega_2)^{k-h_m} V(\tilde{\zeta}_{h_m}). \tag{35}$$

Lemma 3.1: For all $k \in \mathbb{Z}_+$, $V(\tilde{\zeta}_k)$ satisfies

$$V(\tilde{\zeta}_k) \leq (1-\omega_1)^{|\Lambda_N(0,k)|}(1+\omega_2)^{|\Lambda_D(0,k)|} V(\tilde{\zeta}_0). \tag{36}$$

Proof: We use induction to prove the Lemma. Consider the interval $W_{-1} = [0, h_0]$. (36) holds trivially if $h_0 = 0$. If

$h_0 > 0$, over W_{-1}, $V(\tilde{\zeta}_k)$ obeys (33). Thus, (36) follows by noting that, $|\Lambda_N(0,k)| = k$ and $|\Lambda_D(0,k)| = 0$, $\forall k \in W_{-1}$. Next, assume that (36) holds for the interval $[0, h_p], p \in \mathbb{Z}_+$. Then we have,

$$V(\tilde{\zeta}_{h_p}) \leq (1-\omega_1)^{|\Lambda_N(0,h_p)|}(1+\omega_2)^{|\Lambda_D(0,h_p)|}V(\tilde{\zeta}_0). \quad (37)$$

Next, consider the interval $Z_p \coloneqq [h_p, h_p + \tau_p)$. Then, over Z_p, $V(\tilde{\zeta}_k)$ obeys (35) as follows

$$V(\tilde{\zeta}_k) \leq (1+\omega_2)^{k-h_p}V(\tilde{\zeta}_{h_p}). \quad (38)$$

By substituting (37) in (38), (36) follows by noting that $|\Lambda_N(0,k)| = |\Lambda_N(0,h_p)|$, $|\Lambda_D(0,k)| = k - h_p + |\Lambda_D(0,h_p)|$, $\forall k \in Z_p$. Therefore, (36) holds for all $k \in [0, h_p + \tau_p]$.

Next, consider the interval $W_p \coloneqq [h_p + \tau_p, h_{p+1})$. Then, over W_p, $V(\tilde{\zeta}_k)$ obeys (33) as follows $\forall k \in W_m$.

$$V(\tilde{\zeta}_k) \leq (1-\omega_1)^{k-h_p-\tau_p}V(\tilde{\zeta}_{h_p+\tau_p}). \quad (39)$$

In particular, we have,

$$V(\tilde{\zeta}_k) \leq (1-\omega_1)^{k-h_p-\tau_p}(1+\omega_2)^{\tau_p}V(\tilde{\zeta}_{h_p}). \quad (40)$$

Then, by substituting (37) in (40), (36) follows by noting that $|\Lambda_N(0,k)| = k - h_p - \tau_p + |\Lambda_N(0,h_p)|$, $|\Lambda_D(0,k)| = \tau_p + |\Lambda_D(0,h_p)|$, $\forall k \in W_p$. Therefore, (36) holds for all $k \in [0, h_{p+1}]$, where $p \in \mathbb{Z}_+$. ∎

Note that $|\Lambda_N(0,k)| = k - |\Lambda_D(0,k)|$. Then, from (36) and Assumption 3.2, we have

$$(1-\omega_1)^{|\Lambda_N(0,k)|}(1+\omega_2)^{|\Lambda_D(0,k)|} \leq \left[\frac{1+\omega_2}{1-\omega_1}\right]^{\kappa}\Delta^k, \quad (41)$$

where $\Delta = (1-\omega_1)^{\frac{T-1}{T}}(1+\omega_2)^{\frac{1}{T}}$. Under the condition (31), $\Delta < 1$ (for example, choose $T = \frac{T^\star}{\delta}$, where $\delta \in (0,1)$).

Using (41), the following can be obtained from (36)

$$V(\tilde{\zeta}_k) \leq \left[\frac{1+\omega_2}{1-\omega_1}\right]^{\kappa}\Delta^k V(\tilde{\zeta}_0). \quad (42)$$

Thus, the following can be obtained from (42) and (30)

$$|\tilde{\zeta}_k| \leq \beta_{\tilde{\zeta}}(|\tilde{\zeta}_0|, k), \quad (43)$$

$$|e_k| \leq \beta_e(|\tilde{\zeta}_0|, k), \quad (44)$$

where

$\beta_{\tilde{\zeta}}(|\tilde{\zeta}_0|,k) = \sqrt{\left[\frac{1+\omega_2}{1-\omega_1}\right]^{\kappa}\frac{\lambda_M(P^\star)}{\lambda_m(P^\star)}\Delta^k}|\tilde{\zeta}_0|$ and

$\beta_e(|\tilde{\zeta}_0|,k) = |\bar{C}|\sqrt{\left[\frac{1+\omega_2}{1-\omega_1}\right]^{\kappa}\frac{\lambda_M(P^\star)}{\lambda_m(P^\star)}\Delta^k}|\tilde{\zeta}_0|$,

are class $\mathcal{KL}$ functions. From (43) it is clear that (30) has global asymptotic stability property. Thus, we have $\lim_{k\to\infty}(x_k - Xw_k) = 0$, and $\lim_{k\to\infty} e_k = 0$. This implies asymptotic tracking and disturbance rejection. The proof is thus complete. ∎

IV. Learning-based design under DoS Attacks

In this section, we propose an online strategy to learn the optimal controller (20) while the system is under DoS attacks. We assume that the system matrices A, B and D are unknown. We use policy iteration to learn the optimal controller. The idea of policy iteration [33] is to implement both policy evaluation

$$\bar{A}_j^{\mathrm{T}}P_j\bar{A}_j - P_j + Q + K_j^{\mathrm{T}}K_j = 0 \quad (45)$$

and policy improvement

$$K_{j+1} = (1 + \bar{B}^{\mathrm{T}}P_j\bar{B})^{-1}\bar{B}^{\mathrm{T}}P_j\bar{A}, \quad (46)$$

where $\bar{A}_j = \bar{A} - \bar{B}K_j$

Firstly, we rewrite the augmented system (5) as

$$\zeta_{k+1} = \bar{A}_j\zeta_k + \bar{B}(u_k + K_j\zeta_k) + \bar{D}w_k, \quad (47)$$

Along the trajectories of (47), one can obtain that:

$$\begin{aligned}\zeta_{k+1}^{\mathrm{T}}P_j\zeta_{k+1} - \zeta_k^{\mathrm{T}}P_j\zeta_k = &\big[\bar{A}_j\zeta_k + \bar{B}(u_k + K_j\zeta_k) \\ &+ \bar{D}w_k\big]^{\mathrm{T}}P_j\big[\bar{A}_j\zeta_k + \bar{B}(u_k + K_j\zeta_k) + \bar{D}w_k\big] - \zeta_k^{\mathrm{T}}P_j\zeta_k.\end{aligned} \quad (48)$$

Then, using (45), we have:

$$\begin{aligned}\zeta_{k+1}^{\mathrm{T}}P_j\zeta_{k+1} - \zeta_k^{\mathrm{T}}P_j\zeta_k + \zeta_k^{\mathrm{T}}Q_j\zeta_k &= 2\zeta_k^{\mathrm{T}}\Gamma_{1j}^{\mathrm{T}}u_k \\ &+ 2\zeta_k^{\mathrm{T}}\Gamma_{1j}^{\mathrm{T}}K_j\zeta_k - \zeta_k^{\mathrm{T}}K_j^{\mathrm{T}}\Gamma_{2j}K_j\zeta_k + u_k^{\mathrm{T}}\Gamma_{2j}u_k \\ &+ 2\zeta_k^{\mathrm{T}}\Theta_{1j}w_k + 2u_k^{\mathrm{T}}\Theta_{2j}w_k + w_k^{\mathrm{T}}\Theta_{3j}w_k,\end{aligned} \quad (49)$$

where $Q_j = Q + K_j^{\mathrm{T}}K_j$, $\Theta_{1j} = \bar{A}^{\mathrm{T}}P_j\bar{D}$, $\Theta_{2j} = \bar{B}^{\mathrm{T}}P_j\bar{D}$, $\Theta_{3j} = \bar{D}^{\mathrm{T}}P_j\bar{D}$, $\Gamma_{1j} = \bar{B}^{\mathrm{T}}P_j\bar{A}$, $\Gamma_{2j} = \bar{B}^{\mathrm{T}}P_j\bar{B}$. By Assumption 3.2, there always exists a sequence $\{k_s\}_{s=0}^{\infty}$ such that communications are allowed. Then, by collecting online data, the following linear equation can be obtained from (49)

$$\Psi_j\theta_j = -J_{\zeta,\zeta}\mathrm{vec}(Q_j), \quad (50)$$

where

$\Psi_j = \big[\Xi_\zeta, -2J_{\zeta,u} - 2J_{\zeta,\zeta}(I_n \otimes K_j^{\mathrm{T}}), J_{K_j\zeta} - J_u, -2J_{w,\zeta}, -2J_{w,u}, -J_w\big]$, $\theta_j = \big[\mathrm{vecs}(P_j)^{\mathrm{T}}, \mathrm{vec}(\Gamma_{1j})^{\mathrm{T}}, \mathrm{vecs}(\Gamma_{2j})^{\mathrm{T}}, \mathrm{vec}(\Theta_{1j})^{\mathrm{T}}, \mathrm{vec}(\Theta_{2j})^{\mathrm{T}}, \mathrm{vecs}(\Theta_{3j})^{\mathrm{T}}\big]^{\mathrm{T}}$.

One can solve (50) in the least square sense. Under certain choices of E, the matrix J_w may not be full column rank. In such cases, it becomes necessary to reduce the columns of J_w such that Ψ_j is full column rank (see [34], [35]). Denote $\bar{\Psi}_j$ as the matrix which contains the reduced columns of J_w such that $\bar{\Psi}_j$ is full column rank. Since $\bar{\Psi}_j$ has less number of columns, the size of θ_j is also reduced, which is denoted as $\bar{\theta}_j$. Then, the least squares problem (50) can be written as

$$\bar{\Psi}_j\bar{\theta}_j = -J_{\zeta,\zeta}\mathrm{vec}(Q_j). \quad (51)$$

Assumption 4.1: There exists a $s^\star \in \mathbb{Z}_+$ such that for all $s > s^\star$, and for any sequence $k_0 < k_1 < \cdots < k_s$:

$$\begin{aligned}\mathrm{rank}([J_\zeta, J_{\zeta,u}, J_u, J_{w,\zeta}, J_{w,u}, J_w]) &= \frac{n(n+1)}{2} + n \\ &+ 1 + nq + q + \frac{q(q+1)}{2} - N,\end{aligned} \quad (52)$$

where N is the number of linearly dependent columns of J_w. □

***Remark** 4.1:* A typical choice of $s^\star$ can be $s^\star \geq \frac{n(n+1)}{2} + n + 1 + nq + q + \frac{q(q+1)}{2}$. □

***Remark** 4.2:* Under Assumption 4.1, (51) has a unique solution and the sequences $\{P_j\}_{j=0}^{\infty}$ and $\{K_j\}_{j=0}^{\infty}$ obtained using Algorithm 1 converge to a neighborhood of the optimal values $P^\star$ and $K^\star$, respectively [34], [35]. □

Algorithm 1: Online Model-free Policy Iteration

1: Employ $u_k = -K_0\zeta_k + \eta_k$ as the input on the time horizon $[k_0,\ k_s]$, where K_0 is initial stabilizing gain and η_k is the exploration signal.
2: Compute $\bar{\Psi}_j$ until the rank condition in (52) is satisfied. Let $j = 0$.
3: Solve for $\bar{\theta}_j$ from (51).
4: Compute $K_{j+1} = (1 + \Gamma_{2j})^{-1}\Gamma_{1j}$.
5: Let $j \leftarrow j + 1$, repeat Step 3 until $|P_j - P_{j-1}| \leq \epsilon_0$, $j \geq 1$, where $\epsilon_0 > 0$ is a predefined small threshold.

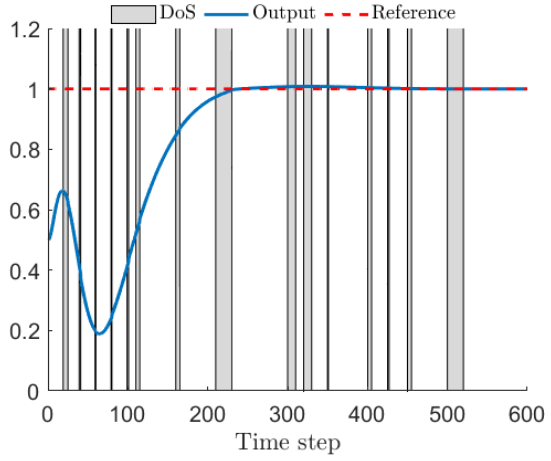

Fig. 1: Tracking and disturbance rejection under DoS attacks.

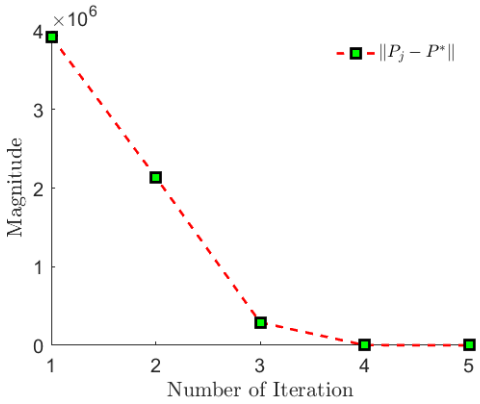

Fig. 2: Convergence of P_j to $P^\star$.

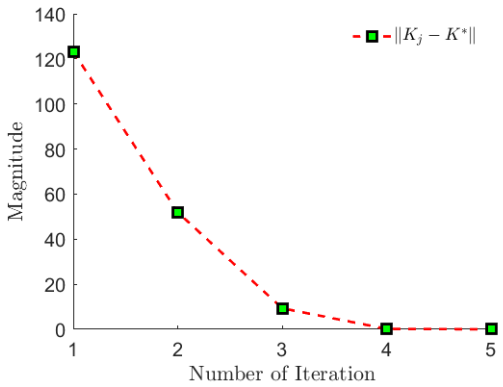

Fig. 3: Convergence of K_j to $K^\star$.

V. Simulation Results and Discussion

In this section, we show the efficacy of the proposed algorithm by applying it to an inverted pendulum on a cart with the following system matrices,

$$A = \begin{bmatrix} 1 & \Delta T & 0 & 0 \\ 0 & 1 - \frac{b\Delta T}{M} & -\frac{mg\Delta T}{M} & 0 \\ 0 & 0 & 1 & \Delta T \\ 0 & \frac{b\Delta T}{lM} & \frac{(M+m)g\Delta T}{lM} & 1 \end{bmatrix},\ B = \begin{bmatrix} 0 \\ \frac{\Delta T}{M} \\ 0 \\ -\frac{\Delta T}{lM} \end{bmatrix},$$

$D = [0 \quad 0.01 \quad 0 \quad 0.01]^{\mathrm{T}}$, $E = 1$, $C = [1\ 0\ 0\ 0]$, $F = -1$, $\mathcal{G}_2 = 0.5$, $\Delta T = 0.01$. For the meaning and value of the parameters please refer to [36]. The initial conditions are given as $x_0 = [0.5,\ 0,\ 0,\ 0]$, and $w_0 = 1$. The weight matrices are chosen as $Q = \mathrm{diag}(1000, 1000, 1000, 1000, 15)$. The DoS parameters are selected as $\kappa = 40$, $\tau_D = 15$, $T = 10$, and $\eta = 1$. The exploration signal in Algorithm 1 is chosen as the summation of sinusoidal waves with different frequencies. Using input-state data for $k \in [0, 100]$, Algorithm 1 converges with a tolerance of $\epsilon_0 = 0.5$ to a neighborhood of the optimal values $P^\star$ and $K^\star$ in five iterations as shown in Figs. 2 and 3, respectively. The optimal controller gain $K^\star$ and the controller gain learned using Algorithm 1 are given in Table I.

TABLE I: Comparison of controller gain values.

	Index				
Controller	1	2	3	4	5
$K^\star$	-153.9801	-99.7489	-283.9957	-56.1038	-2.6548
K_5	-153.9802	-99.7490	-283.9958	-56.1038	-2.6548

We immediately apply the learned controller after $k = 100$. Fig. 1 shows the output and reference trajectories, with the DoS attacks represented as shaded areas. The learned controller can track the reference signal even in the presence of DoS attacks. The DoS duration parameter can be obtained as $T^\star = 6.5295 \times 10^6$. Similar to [19], [24], these are sufficient conditions to guarantee the resilience and stability of the closed-loop system. In practice $T^\star$ can be much smaller. This is demonstrated by applying stronger DoS attacks after 100 time steps.

VI. Conclusion and Future Works

This paper investigates the challenge of achieving optimal output regulation of discrete-time linear systems with unknown parameters while facing denial-of-service (DoS) attacks. We have proposed a resilient online policy iteration algorithm capable of learning the optimal controller using only the input-state data in the presence of DoS attacks. An upper bound on the DoS duration is achieved to guarantee the stability of the closed-loop system. Finally, the proposed technique is applied to an inverted pendulum on a cart.

Future work will focus on extending this technique to discrete-time non-linear systems.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] D. Bertsekas, *Dynamic programming and optimal control*. Athena scientific, 2012, vol. 1.

[3] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[4] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[5] P. Werbos, "Beyond regression:" new tools for prediction and analysis in the behavioral sciences," *Ph. D. dissertation, Harvard University*, 1974.

[6] 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, 2012.

[7] 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, 2009.

[8] K. G. Vamvoudakis, N.-M. T. Kokolakis *et al.*, "Synchronous reinforcement learning-based control for cognitive autonomy," *Foundations and Trends® in Systems and Control*, vol. 8, no. 1–2, pp. 1–175, 2020.

[9] S. Chakraborty, L. Cui, K. Ozbay, and Z.-P. Jiang, "Automated lane changing control in mixed traffic: An adaptive dynamic programming approach," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 1823–1828.

[10] Y. Jiang and Z.-P. Jiang, "Adaptive dynamic programming as a theory of sensorimotor control," *Biological Cybernetics*, vol. 108, no. 4, pp. 459–473, 2014.

[11] ——, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 882–893, 2014.

[12] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

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

[14] ——, "Learning-based adaptive optimal tracking control of strict-feedback nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2614–2624, 2017.

[15] W. Gao, M. Mynuddin, D. C. Wunsch, and Z.-P. Jiang, "Reinforcement learning-based cooperative optimal output regulation via distributed adaptive internal model," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5229–5240, 2021.

[16] A. Odekunle, W. Gao, M. Davari, and Z.-P. Jiang, "Reinforcement learning and non-zero-sum game output regulation for multi-player linear uncertain systems," *Automatica*, vol. 112, p. 108672, 2020.

[17] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Hybrid Systems: Computation and Control: 12th International Conference, HSCC 2009, San Francisco, CA, USA, April 13-15, 2009. Proceedings 12*. Springer, 2009, pp. 31–45.

[18] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

[19] C. De Persis and P. Tesi, "Input-to-state stabilizing control under denial-of-service," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2930–2944, 2015.

[20] C. Deng and C. Wen, "Distributed resilient observer-based fault-tolerant control for heterogeneous multiagent systems under actuator faults and dos attacks," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 3, pp. 1308–1318, 2020.

[21] L. An and G.-H. Yang, "Decentralized adaptive fuzzy secure control for nonlinear uncertain interconnected systems against intermittent dos attacks," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 827–838, 2018.

[22] X.-M. Zhang, Q.-L. Han, X. Ge, and L. Ding, "Resilient control design based on a sampled-data model for a class of networked control systems under denial-of-service attacks," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3616–3626, 2019.

[23] R. Zhang, G. Li, and R. Yang, "Secure control for the discrete-time cpss under dos attacks via a switching strategy," in *2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, 2023, pp. 1678–1683.

[24] W. Gao, C. Deng, Y. Jiang, and Z.-P. Jiang, "Resilient reinforcement learning and robust output regulation under denial-of-service attacks," *Automatica*, vol. 142, p. 110366, 2022.

[25] F. Galarza-Jimenez, J. I. Poveda, G. Bianchin, and E. Dall'Anese, "Extremum seeking under persistent gradient deception: A switching systems approach," *IEEE Control Systems Letters*, vol. 6, pp. 133–138, 2021.

[26] L. Zhai and K. G. Vamvoudakis, "Data-based and secure switched cyber–physical systems," *Systems & Control Letters*, vol. 148, p. 104826, 2021.

[27] Y. Shi, X. Dong, Y. Hua, J. Yu, and Z. Ren, "Distributed output formation tracking control of heterogeneous multi-agent systems using reinforcement learning," *ISA Transactions*, vol. 138, pp. 318–328, 2023.

[28] W. Liu, J. Sun, G. Wang, F. Bullo, and J. Chen, "Data-driven resilient predictive control under denial-of-service," *IEEE Transactions on Automatic Control*, 2022.

[29] X. Zhang, J. Liu, X. Xu, S. Yu, and H. Chen, "Robust learning-based predictive control for discrete-time nonlinear systems with unknown dynamics and state constraints," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 12, pp. 7314–7327, 2022.

[30] Y.-S. Ma, W.-W. Che, C. Deng, and Z.-G. Wu, "Distributed model-free adaptive control for learning nonlinear mass under dos attacks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1146–1155, 2021.

[31] F. Li and Z. Hou, "Learning-based model-free adaptive control for nonlinear discrete-time networked control systems under hybrid cyber attacks," *IEEE Transactions on Cybernetics*, 2022.

[32] J. Huang, *Nonlinear output regulation: theory and applications*. SIAM, 2004.

[33] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Transactions on Automatic Control*, vol. 16, no. 4, pp. 382–384, 1971.

[34] S. Chakraborty, W. Gao, K. G. Vamvoudakis, and Z.-P. Jiang, "Adaptive optimal output regulation of discrete-time linear systems: A reinforcement learning approach," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 7950–7955.

[35] S. Chakraborty, W. Gao, L. Cui, F. L. Lewis, and Z.-P. Jiang, "Learning-based adaptive optimal output regulation of discrete-time linear systems," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 10 283–10 288, 2023.

[36] R. Gurumoorthy and S. Sanders, "Controlling non-minimum phase nonlinear systems-the inverted pendulum on a cart example," in *1993 American Control Conference*. IEEE, 1993, pp. 680–685.