Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 58-14 (2024) 580–585

Cyberattack Detection and Handling for Neural Network-Approximated Economic Model Predictive Control [★]

Jihan Abou Halloun [*] **Helen Durand** [**]

[*] *Wayne State University, Detroit, MI 48202, USA (e-mail: gh5411@wayne.edu)*

[**] *Wayne State University, Detroit, MI 48202, USA (e-mail: helen.durand@wayne.edu)*

Abstract: Cyberattacks on control systems can create unprofitable and unsafe operating conditions. To enhance safety and attack resiliency of control systems, cyberattack detection strategies can be developed. Prior work in our group has sought to develop cyberattack detection strategies that are integrated with an advanced control formulation known as Lyapunov-based economic model predictive control (LEMPC), in the sense that the controller properties can be used to analyze closed-loop stability in the presence or absence of undetected attacks. In this work, we consider neural network-approximated control laws, concepts for mitigating cyberattacks on such control laws, and how these ideas elucidate concepts in how to fight back against cyberattacks. We begin by providing sufficient conditions under which a neural network (NN) that approximates an LEMPC maintains safety for a sampling period after a cyberattack by inheriting safety properties from the LEMPC formulation. Then, we discuss a second concept inspired by neural network repair in the presence of adversarial attacks for attempting to ensure safety of controllers for a time period after undetected attacks, even those not based on a rigorous control law formulation like LEMPC. We examine the potential conservatism differences between the LEMPC-based safety strategy and one based on repairing problematic control actions, and discuss how this concept can inspire ideas for fighting back against attacks.

Keywords: Neural networks in process control, Lyapunov-based economic model predictive control, reachability analysis, cybersecurity.

1. INTRODUCTION

Cybersecurity considerations are becoming more critical due to the high exposure of control systems to a range of threats and security vulnerabilities. One avenue toward enhancing safety and attack resiliency of control systems is developing cyberattack detection and handling strategies. A variety of strategies have been developed for attack detection, which may monitor the state (passive detection) or disrupt it (active detection) to locate attacks. Our group's prior work (e.g., (Oyama and Durand (2020))) has focused on developing both active and passive detection strategies in the context of Lyapunov-based economic model predictive control (LEMPC). Some of these strategies have had the ability to maintain safety for a characterizable period of time after an undetected attack. Despite this potential benefit, there are several important questions which remain unsolved for these designs, including whether they are more conservative than alternative methodologies for attack detection, and also to what extent the restriction of considering LEMPC alone is useful. A variety of alternative means for attempting to promote resilience of a process against cyberattacks have been developed. These include machine learning-based methods (Chen et al. (2020)) and reachability-based techniques (Liu et al. (2021)) for attack detection. In Wu et al. (2018), for example, a neural network-based detection methodology was paired with a Lyapunov-based model predictive controller (LMPC) to mitigate the impact of detected cyberattacks and stabilize the nonlinear system at its steady-state. In Narasimhan et al. (2023), reachable sets are used for linear systems to detect cyberattacks even when transient operation occurs.

A potential benefit of reachable sets toward addressing the two challenges raised above with respect to the LEMPC-based strategies is that reachable sets of states can be computed for any control action (independent of the control law), suggesting that they may be used in safety assessment under attacks in a variety of scenarios, including for other forms of economic model predictive control (EMPC) besides the Lyapunov-based version, or in cases where neural networks (NN's) are used to approximate controllers. NN's have been used to approximate an advanced controller to reduce its computational intensity (Åkesson and Toivonen (2006); Lucia and Karg (2018)). Furthermore, NN outputs have been analyzed from a reachability analysis perspective (e.g., (Yang et al. (2022, 2021); Xiang

[★] Financial support from the the National Aeronautics and Space Administration (NASA), under award number 80NSSC20M0124, Michigan Space Grant Consortium (MSGC), the Air Force Office of Scientific Research (award number FA9550-19-1-0059), the National Science Foundation CNS-1932026 and CBET-2143469, and Wayne State University is gratefully acknowledged.

10.1016/j.ifacol.2024.08.399

and Johnson (2018))). In Yang et al. (2022), for example, reachability analysis was used to compute unsafe NN output regions which can be backtracked to identify the corresponding unsafe NN input space and then repair the NN in the sense that it is re-designed to no longer compute problematic outputs.

Motivated by these considerations, this work seeks to provide deeper insight into the LEMPC-based strategies, and in particular the two questions noted above, with the goal to better understand various cyberattack detection techniques and relationships among them, as well as to motivate directions in steps beyond detection such as fighting back against attackers. To do this, we begin by addressing the question of the extent to which focusing on LEMPC in cyberattack detection design is useful. We do this by first demonstrating that one of the passive detection LEMPC-based methods can be extended to a case where a NN version of that LEMPC is learned, demonstrating that there are some cases where the LEMPC-based methods can be extended beyond the specific implementations described in Oyama and Durand (2020). Next, we discuss how a method based on reachability analysis, with inspiration from the NN repair strategy in Yang et al. (2022), could enable a wide range of control laws (and their NN versions) to obtain the same safety guarantees in the presence of undetected attacks as the passive detection LEMPC-focused method, but with the need to determine and analyze additional sets that are developed *a priori* in the LEMPC strategy. This facilitates a comparison between the techniques. In addition, we discuss how the repair strategy can inspire initial ideas in fighting back against cyberattackers once they are detected in a process. We close with an example that illustrates some of the key features of the reachability and repair procedure.

2. PRELIMINARIES

2.1 Notation

A continuous function $\alpha : [0, a) \longrightarrow [0, \infty)$ is a class $\mathcal{K}$ function if $\alpha(0) = 0$ and it is strictly increasing. Ω_ρ denotes a level set of a scalar-valued function V (i.e., $\Omega_\rho := \{x \in R^n \ : \ V(x) \leq \rho\}$). '/' signifies set subtraction.

2.2 Class of Systems

The class of nonlinear systems considered is:

$$\dot{x}(t) = f(x(t), u(t), w(t)) \tag{1}$$

where $x(t) \in X \subset R^n$, $u(t) \in U \subset R^m$ ($U := \{u \in R^m \ : \ |u| \leq u^{max}\}$), and $w(t) \in W \subset R^z$ ($W := \{w \in R^z \ : \ |w| \leq \theta, \theta > 0\}$) are the bounded state, input, and disturbance vectors. The function $f(\cdot)$ in Eq. (1) is assumed to be locally Lipschitz on $X \times U \times W$. $w \equiv 0$ in the corresponding "nominal" system of Eq. (1). It is assumed that the origin of Eq. (1) is an equilibrium point ($f(0, 0, 0) = 0$).

It is assumed that a control law $u = h(x) \in U$ exists that renders the origin of the nominal system of Eq. (1) asymptotically stable, in the sense that there exists a sufficiently smooth Lyapunov function V, and class $\mathcal{K}$ functions $\alpha_i(\cdot), i = 1, 2, 3, 4$, where:

$$\alpha_1(|x|) \leq V(x) \leq \alpha_2(|x|), \tag{2a}$$

$$\frac{\partial V(x)}{\partial x} f(x, h(x), 0) \leq -\alpha_3(|x|), \tag{2b}$$

$$|\frac{\partial V(x)}{\partial x}| \leq \alpha_4(|x|), \tag{2c}$$

$$h(x) \in U \tag{2d}$$

$\forall x \in D \subset R^n$ where D is an open neighborhood of the origin. The stability region of the closed-loop system under the controller $h(x)$ is defined as $\Omega_\rho \subset D$ and is chosen such that $x \in X, \forall x \in \Omega_\rho$. The Lipschitz property of f combined with the bounds on u and w implies that there exist positive constants $M_f, L_x, L_u, L_w, L'_x, L'_w, L'_u$ such that $\forall x, x_1, x_2 \in \Omega_\rho, u, u_1, u_2 \in U$ and $w \in W$, the following inequalities hold:

$$\begin{aligned} |f(x_1, u_1, w) - f(x_2, u_2, 0)| \leq L_x|x_1 - x_2| \\ + L_u|u_1 - u_2| + L_w|w| \end{aligned} \tag{3a}$$

$$\begin{aligned} |\frac{\partial V(x_1)}{\partial x} f(x_1, u_1, w) - \frac{\partial V(x_2)}{\partial x} f(x_2, u_2, 0)| \\ \leq L'_x|x_1 - x_2| + L'_w|w| + L'_u|u_1 - u_2| \end{aligned} \tag{3b}$$

$$|f(x, u, w)| \leq M_f \tag{3c}$$

2.3 Lyapunov-Based Economic Model Predictive Control

LEMPC (Heidarinejad et al. (2012)) computes control actions via:

$$\min_{u(t) \in S(\Delta)} \int_{t_k}^{t_{k+N}} L_e(\tilde{x}(\tau), u(\tau)) d\tau \tag{4a}$$

$$\text{s.t.} \quad \dot{\tilde{x}}(t) = f(\tilde{x}(t), u(t), 0) \tag{4b}$$

$$\tilde{x}(t_k) = \bar{x}(t_k) \tag{4c}$$

$$\tilde{x}(t) \in X, \forall t \in [t_k, t_{k+N}) \tag{4d}$$

$$u(t) \in U, \forall t \in [t_k, t_{k+N}) \tag{4e}$$

$$\begin{aligned} V(\tilde{x}(t)) \leq \rho_e, \forall t \in [t_k, t_{k+N}) \\ \text{if } \bar{x}(t_k) \in \Omega_{\rho_e} \end{aligned} \tag{4f}$$

$$\begin{aligned} \frac{\partial V(\bar{x}(t_k))}{\partial x} f(\bar{x}(t_k), u(t_k), 0) \\ \leq \frac{\partial V(\bar{x}(t_k))}{\partial x} f(\bar{x}(t_k), h(\bar{x}(t_k)), 0) \\ \text{if } \bar{x}(t_k) \in \Omega_\rho / \Omega_{\rho_e} \end{aligned} \tag{4g}$$

where N is the prediction horizon and $u(t)$ is a piecewise-constant input trajectory with N pieces computed using sample-and-hold where each piece is held constant for a sampling period Δ (denoted $u(t) \in S(\Delta)$). At a sampling time t_k, the LEMPC receives a state measurement $\bar{x}(t_k)$ (Eq. (4c)) and evaluates the economics-based stage cost L_e of Eq. (4a) throughout $N\Delta$ using Eq. (4b). The state predictions and inputs must satisfy the constraints of Eqs. (4d) and (4e). $\Omega_{\rho_e} \subset \Omega_\rho$ makes Ω_ρ forward invariant under the LEMPC of Eq. (4), with the Lyapunov-based stability constraints in Eqs. 4f-4g.

3. A NN-APPROXIMATED VERSION OF AN LEMPC INTEGRATED WITH PASSIVE CYBERATTACK DETECTION

As noted in the Introduction, one goal of this work is to better understand the extent to which the integrated cyberattack detection and control strategies from Oyama and Durand (2020) are limiting due to their focus on LEMPC. Therefore, in this section, we focus on one of the detection strategies and demonstrate that despite its focus on LEMPC, it can have extensions, such as being implemented as a NN, that could achieve the same types of safety guarantees as the original LEMPC (in this case, the guarantee of safety in the sense of boundedness of the closed-loop state in Ω_ρ for at least one sampling period after an undetected attack under sufficient conditions). This provides a first indication that the methods derived in Oyama and Durand (2020) can have applicability in developing further detection strategies beyond those which require an LEMPC during process operation. The particular situation which inspires considering the NN version of the control law is that advanced control laws such as LEMPC might be computationally expensive. One approach that has been explored for attempting to decrease the online computational complexity of advanced controllers is to approximate them using a NN (Lucia and Karg (2018)). To develop such a NN, an LEMPC can be solved a number of times, from various initial conditions, to generate a dataset with a large number of possible sensor measurements for various initial conditions x_0, for the training and validation of the NN. The NN is then trained off-line to fit the NN parameters (i.e., NN weights) where the input layers are the state measurements and the output layers are the control actions generated by the controller. Once the NN is well-trained, it can be used to predict control actions similar to those of the LEMPC when a new state measurement is received. We refer to the resulting NN as a NN-approximated LEMPC (which we will abbreviate as NN-LEMPC in the following).

In Oyama and Durand (2020), a passive cyberattack detection strategy based on comparing state predictions (under the control actions computed by an LEMPC) with state measurements (with bounded sensor noise) was developed. In this section, we will develop a similar detection strategy for the process of Eq. (1) under the NN-LEMPC in the case where the state predictions are made under the NN-LEMPC control input. For these developments, we make the following assumption.

Assumption 1. The errors in the sensor measurements and also in the control actions computed by the LEMPC of Eq. (4) and the NN-LEMPC are bounded. Specifically, the sensor measurement error is bounded such that $|\tilde{x}(t_k) - x(t_k)| \leq \theta_v$, $\theta_v > 0$, at every sampling time, where $\tilde{x}(t_k)$ represents the state measurement at t_k and $x(t_k)$ represents the value of the actual state at t_k. In addition, the difference between the input computed from the LEMPC of Eq. (4) (i.e., u) and that computed by the LEMPC-NN (i.e., $\hat{u}$) is bounded such that $|u(t_k) - \hat{u}(t_k)| \leq \epsilon$, for $\epsilon > 0$.

To derive the detection strategy, we first note that Oyama and Durand (2020) derived sufficient conditions to enable an LEMPC to maintain the closed-loop state within Ω_ρ in the absence of attacks, and for at least one sampling period after an undetected attack (i.e., an attack which kept a false state measurement within a bound of the state prediction from the last non-attacked sensor measurement). A NN-LEMPC, however, may not be able to perfectly represent the LEMPC for which these theoretical results were derived. Specifically, Assumption 1 allows for deviations between the values of $u(t_k)$ and $\hat{u}(t_k)$ (where the magnitude of ϵ may be adjusted by, for example, selecting different neural network architectures or using different amounts of training data). Thus, there may be cases where $u(t_k)$ might cause the closed-loop state to move away from the boundary of Ω_ρ, whereas the slightly different input $\hat{u}(t_k)$ may not. This indicates that, to obtain safety for a sampling period after an undetected attack using the NN-LEMPC, it is not sufficient to design an LEMPC for the state prediction-based detection strategy according to Oyama and Durand (2020) and then to train the NN-LEMPC from the data corresponding to that LEMPC. Instead, the requirements on the LEMPC from which the data for training the NN-LEMPC will be derived need to be made more stringent to enable the NN-LEMPC developed from that data to then have the desired safety properties in the presence and absence of attacks. Thus, we will require that the theoretical requirements in Oyama and Durand (2020) continue to hold, but that additional requirements, discussed in this section, are introduced.

3.1 Designing an LEMPC for NN-LEMPC Attack-Handling

In this section, we describe how the theoretical requirements imposed on the LEMPC of Eq. (4) for the state prediction-based detection strategy in Oyama and Durand (2020) can be augmented to enable the NN-LEMPC trained on the LEMPC data to then inherit similar theoretical qualities. We first introduce two propositions that will be used in the proof of the main result.

Proposition 1. Consider the following systems:

$$\dot{\tilde{x}}(t) = f(\tilde{x}(t), u(t_0), 0) \tag{5}$$

$$\dot{\hat{x}}(t) = f(\hat{x}(t), \hat{u}(t_0), w(t)) \tag{6}$$

where $|\hat{x}(t_0) - \tilde{x}(t_0)| \leq \delta$, $\delta > 0$, $t_0 = 0$, and $|u(t_0) - \hat{u}(t_0)| \leq \epsilon$. Then, if $\hat{x}(t), \tilde{x}(t) \in \Omega_\rho$ for $t \geq 0$:

$$|\tilde{x}(t) - \hat{x}(t)| \leq f_{W,NN}(\delta, \epsilon, t) \tag{7}$$

for $t \geq t_0$, where

$$f_{W,NN}(s,p,\tau) := \left[\left(s + \frac{L_u p + L_w \theta}{L_x} \right) e^{L_x \tau} - \left(\frac{L_u p + L_w \theta}{L_x} \right) \right] \tag{8}$$

Proof 1. Integrating Eqs. (5)-(6) from $t_0 = 0$ to t, we obtain:

$$\tilde{x}(t) = \tilde{x}(t_0) + \int_0^t f(\tilde{x}(s), u(t_0), 0)ds \tag{9}$$

$$\hat{x}(t) = \hat{x}(t_0) + \int_0^t f(\hat{x}(s), \hat{u}(t_0), w(s))ds \tag{10}$$

Subtracting Eq. (10) from Eq. (9), taking the Euclidean norm of both sides, and applying Eq. (3a) as well as the bounds on w and $|u(t_0) - \hat{u}(t_0)|$ gives:

$$\begin{aligned} |\tilde{x}(t) - \hat{x}(t)| &\leq |\tilde{x}(t_0) - \hat{x}(t_0)| \\ &+ \int_0^t [|f(\tilde{x}(s), u(t_0), 0) - f(\hat{x}(s), \hat{u}(t_0), w(s))|]ds \\ &\leq \delta + \int_0^t [L_x|\tilde{x}(s) - \hat{x}(s)| + L_w|w(s)| + \\ &\qquad L_u|u(t_0) - \hat{u}(t_0)|]ds \\ &\leq \delta + (L_u\epsilon + L_w\theta)t + L_x \int_0^t [|\tilde{x}(s) - \hat{x}(s)|]ds \end{aligned} \tag{11}$$

Applying the Gronwall-Bellman inequality gives Eqs. (7)-(8).

Proposition 2. (c.f. Heidarinejad et al. (2012)) Consider the Lyapunov function $V(\cdot)$ of the system of Eq. (1). There exists a quadratic function $f_v(\cdot)$ such that:

$$V(\hat{x}) \le V(x) + f_v(|x-\hat{x}|) \tag{12}$$

for all $x, \hat{x} \in \Omega_\rho$, with $f_v(s) = \alpha_4(\alpha_1^{-1}(\rho))s + M_v s^2$, where M_v is a positive constant.

Theorem 1 below provides sufficient conditions under which the NN-LEMPC guarantees that the state of the closed-loop system of Eq. (1) is always maintained in Ω_ρ before an attack occurs, if the LEMPC of Eq. (4) is also designed according to both these requirements and the additional requirements for the state prediction-based attack detection strategy in Oyama and Durand (2020).

Theorem 1. Consider the system of Eq. (1) under the NN-LEMPC design based on an LEMPC satisfying the state prediction-based detection LEMPC conditions of Oyama and Durand (2020), and a controller $h(\cdot)$ that satisfies Eq. (2). Let Assumption 1 hold, and $\epsilon_w > 0,\ \Delta > 0,\ N \ge 1$, and $\rho > \rho_{samp,NN} > \rho_{NN} > \rho_e > 0$ satisfy:

$$\rho_e + f_v(f_{W,NN}(\delta,\epsilon,\Delta)) \le \rho_{NN} \tag{13}$$

$$L'_x(M\Delta+\delta) + L'_u\epsilon + L'_w\theta - \alpha_3(\alpha_2^{-1}(\rho_e)) \le -\epsilon_w/\Delta \tag{14}$$

$$\begin{aligned}\rho > \max\{V(\tilde{x}(t)) :\ & V(\hat{x}(t_k)) \in \Omega_{\rho_{samp,NN}},\ t \in [t_k, t_{k+1}),\\ & u \in U,\ w \in W\}\end{aligned} \tag{15}$$

$$\begin{aligned}\rho_{samp,NN} > \max\{V(\hat{x}(t)) :\ & V(\tilde{x}(t_k)) \in \Omega_{\rho_e},\ t \in [t_k, t_{k+1}),\\ & u \in U,\ w \in W\}\end{aligned} \tag{16}$$

If $\hat{x}(t_0), \tilde{x}(t_0) \in \Omega_{\rho_e}$, then $\hat{x}(t) \in \Omega_\rho$, and the state measurement $\tilde{x}(t_k) \in \Omega_\rho$, $\forall t, t_k \ge 0$, before a sensor attack occurs.

Proof 2. When $\hat{x}(t_k) \in \Omega_{\rho_e}$, then if the LEMPC of Eq. 4 had been used over the subsequent sampling period (instead of the NN-LEMPC), it would have computed a control action such that $V(\tilde{x}(t)) \le \rho_e$. From Proposition 1, Proposition 2, and Eq. (7):

$$\begin{aligned}V(\hat{x}(t)) &\le V(\tilde{x}(t)) + f_v(|\hat{x}-\tilde{x}|)\\ &\le \rho_e + f_v(f_{W,NN}(\delta,\epsilon,\Delta))\end{aligned} \tag{17}$$

for $t \in [t_k, t_{k+1})$. If the conditions of the theorem hold (with $\delta = \theta_v$ when designed for a case without attacks), then $V(\hat{x}(t)) \in \Omega_{\rho_{NN}} \subset \Omega_\rho$. If instead $\hat{x}(t_k) \in \Omega_\rho/\Omega_{\rho_e}$, then:

$$\begin{aligned}&\frac{\partial V(\hat{x}(\tau))}{\partial x} f(\hat{x}(\tau),\hat{u}(t_0),w(\tau)) = \frac{\partial V(\hat{x}(\tau))}{\partial x} f(\hat{x}(\tau),\hat{u}(t_0),w(\tau))\\ &- \frac{\partial V(\tilde{x}(t_k))}{\partial x} f(\tilde{x}(t_k),u(t_0),0) + \frac{\partial V(\tilde{x}(t_k))}{\partial x} f(\tilde{x}(t_k),u(t_0),0)\\ &\le L'_x|\hat{x}(t)-\tilde{x}(t_k)| + L'_u|\hat{u}(t_k)-u(t_k)|\\ &\quad + L'_w|w| - \alpha_3(|\tilde{x}(t_k)|)\\ &\le L'_x(M\Delta+\delta) + L'_u\epsilon + L'_w\theta - \alpha_3(\alpha_2^{-1}(\rho_e)) \le -\epsilon_w/\Delta\end{aligned} \tag{18}$$

from Eq. (3a) as well as the bounds on w, $|f|$, and $|u(t_k) - \hat{u}(t_k)|$. When Eq. (14) holds, $\dot{V}$ is negative along the closed-loop state trajectory under the NN-LEMPC. Eqs. (15)-(16) ensure that either $\hat{x}(t_k) \in \Omega_{\rho_e}$ or that $\hat{x}(t_k) \in \Omega_\rho/\Omega_{\rho_e}$. Specifically, Eq. (16) ensures that if $\tilde{x}(t_k) \in \Omega_{\rho_e}$, then $\hat{x}(t) \in \Omega_{\rho_{samp,NN}} \subset \Omega_\rho$ for all $t \in [t_k, t_{k+1})$. This ensures that the furthest that the closed-loop state can move away from the origin is to a level set of V where $V(\hat{x}) \le \rho_{samp,NN}$, because subsequently Eq. (18) will apply and $\dot{V}$ will decrease when the conditions of the theorem are met. Because the closed-loop state cannot move farther than $\Omega_{\rho_{samp,NN}}$ from the origin, the condition of Eq. (15), which ensures that the measurement even when $x(t) \in \Omega_{\rho_{samp,NN}}$ is still within Ω_ρ, ensures that the state measurement is always within Ω_ρ.

To demonstrate that the NN-LEMPC obtains safety properties after an undetected attack, we first devise the detection strategy. In this case, we will compare the state prediction under $\hat{u}(t_k)$ (where a prediction of the state at t_{k+1} obtained by numerically integrating the equation $\dot{\tilde{x}}_{NN} = f(\tilde{x}_{NN}, \hat{u}(t_k), 0)$ for the state prediction $\tilde{x}_{NN}$ from a state measurement from t_k is denoted $\tilde{x}_{NN}(t_{k+1}|t_k)$) with the state measurement. Thus, the detection strategy is as follows:

(1) At t_k, operate the process under the NN-LEMPC. Go to Step 2.
(2) Obtain a state measurement $\tilde{x}(t_{k+1})$. Compute $q = |\tilde{x}_{NN}(t_{k+1}|t_k) - \tilde{x}(t_{k+1})|$. If $q > \nu_{th}$, flag an attack and implement mitigating actions. Otherwise, go to Step 3.
(3) $t_k \leftarrow t_{k+1}$. Return to Step 1.

Under this implementation strategy, the following theorem provides conditions under which the NN-LEMPC will achieve safety for a sampling period after an undetected attack.

Theorem 2. Consider that the conditions of Theorem 1 hold. Then, the closed-loop state of the system of Eq. (4) under the NN-LEMPC and the implementation strategy of this section will be maintained within Ω_ρ for at least one sampling period after an undetected attack when $\delta \ge f_{W,NN}(\theta_v, 0, \Delta) + \nu_{th}$ in Eqs. 13 and 14.

Proof 3. To prove this, we note that due to Assumption 1,

$$|\tilde{x}_{NN}(t_{k-1}|t_{k-1}) - \hat{x}(t_{k-1})| \le \theta_v \tag{19}$$

where $\tilde{x}_{NN}(t_{k-1}|t_{k-1})$ denotes a state measurement at t_{k-1} and $\hat{x}(t_{k-1})$ signifies the actual state at t_{k-1}. Considering Proposition 1 with $\epsilon = 0$ (since both the actual closed-loop state and the predictions of the state will use the same input $\hat{u}(t_0)$), we obtain:

$$|\hat{x}(t_k) - \tilde{x}_{NN}(t_k|t_{k-1})| \le f_{W,NN}(\theta_v, 0, \Delta) \tag{20}$$

Then if an attack is not flagged at t_k, this, along with the detection bound threshold ν_{th}, provide the following:

$$\begin{aligned}&|\hat{x}(t_k) - \tilde{x}_{NN}(t_k|t_k)|\\ &\le |\hat{x}(t_k) - \tilde{x}_{NN}(t_k|t_{k-1}) + \tilde{x}_{NN}(t_k|t_{k-1}) - \tilde{x}_{NN}(t_k|t_k)|\\ &\le f_{W,NN}(\theta_v, 0, \Delta) + \nu_{th}\end{aligned} \tag{21}$$

Thus, if $\delta \ge f_{W,NN}(\theta_v, 0, \Delta) + \nu_{th}$ in Theorem 1, then from the result of that theorem, the closed-loop state will be kept within Ω_ρ over the following sampling period.

4. REACHABILITY ANALYSIS FOR EMPC AND NN-EMPC WITH SENSOR ATTACKS

In the previous section, we sought to demonstrate that the restriction of the cyberattack detection considerations in (Oyama and Durand (2020)) to LEMPC may not necessarily preclude their utility for other scenarios, including extension to NN-LEMPC. In this section, we seek to further understand the restrictiveness and conservatism of the LEMPC-focused method by comparing it with a reachability-based strategy. As noted in the Introduction, reachability analysis has played a role in both safety analysis and detection for control system cyberattacks, and has also played a role in NN safety verification. Thus, we consider that a reachability-based method offers a point of comparison between the state prediction-based detection framework in (Oyama and Durand (2020)) (which was demonstrated to extend to a NN case in the prior section), and other economic model predictive control (EMPC) formulations and their NN approximations.

4.1 Cyberattack Detection and Handling with Reachability Analysis for EMPC and NN-EMPC

In this section, we propose a method based on reachability analysis and input repair (inspired by the concept of repairing a NN in Yang et al. (2022)) for ensuring safety (in the sense of boundedness of the closed-loop state in a safe operating region) for a sampling period after an undetected cyberattack when the detection scheme is the same as is used for the LEMPC-based strategy (i.e., at every sampling time t_k, we check $q = |\tilde{x}_{EMPC}(t_k|t_{k-1}) - \tilde{x}(t_k)|$, where $\tilde{x}_{EMPC}(t_k|t_{k-1})$ is a prediction of the state at t_k using the input from an EMPC or NN-EMPC from a state measurement at t_{k-1}). The safe operating region is considered to be a superset of Ω_ρ so that checking whether the state is in Ω_ρ suffices to verify safety, as for the LEMPC-based strategies.

In this strategy, we consider that a measurement of the state is received at t_k. If $q > \nu_{th}$, an attack is detected and mitigating actions (e.g., emergency shut-down) are performed, as would also be done for the LEMPC-based method if an attack was detected. However, when $q \leq \nu_{th}$, the reachability-based method proceeds as follows. First, the set of all states consistent with the measurement is obtained and denoted by $\mathcal{S}$ (i.e., the set of all states such that $|\tilde{x}_{EMPC}(t_k|t_{k-1}) - \tilde{x}(t_k)|$, where $\tilde{x}_{EMPC}(t_k|t_{k-1})$ refers to a prediction of the state at t_k based on a measurement at t_{k-1} under an EMPC or NN-EMPC). Then, $u(t_k)$ is computed by the EMPC or NN-EMPC based on the state measurement (which could be falsified with an undetected attack). Subsequently, the set of all state trajectories which would be generated under $u(t_k)$, $w \in W$, starting from $\mathcal{S}$ can be computed. The safety of these trajectories (i.e., whether they remain within Ω_ρ) can then be assessed. If any is found to leave Ω_ρ, the input determined by the EMPC or NN-EMPC can be "repaired." The idea of the repair is that a search is performed for a different input which can keep the state trajectories over a sampling period, starting from all states in $\mathcal{S}$, under the updated input and with $w \in W$, within Ω_ρ. If such a repaired input can be found, then just as the LEMPC-based strategies were able to ensure safety for a sampling period after an undetected attack, EMPC's and NN-EMPC's with the input repair and reachability analysis approach would then also ensure safety within Ω_ρ for a sampling period after an undetected attack.

We can now utilize this reachability strategy to understand some of the benefits and limitations of the LEMPC-based detection strategy. First, we note that both the LEMPC-based strategy and reachability strategy can be extended to controllers learned with a neural network, and both are able to keep the closed-loop state within Ω_ρ for a sampling period after an undetected sensor attack under sufficient conditions. Both can do this utilizing the same detection metric. However, the sufficient conditions for both are what distinguish the methods. Specifically, for the LEMPC-based method, the conditions are a variety of control-theoretic principles from (Oyama and Durand (2020)), similar to (and augmented by, if a NN-LEMPC is used) those in Theorems 1-2. The sufficient conditions for the reachability technique include that inputs must be determined to exist that can achieve the repair for every possible undetected attack based on the operating region within Ω_ρ in which the EMPC or NN-EMPC should keep the process in the absence of attacks. Thus, while a disadvantage of LEMPC is the need to find a control strategy for which all of the theoretical conditions can be met, it also has the benefit of essentially ensuring the existence of a "repaired" input, at least for the state for which the repair is most relevant (i.e., the actual state). It ensures that even with an undetected attack, so that the state may be in $\mathcal{S}$ and different from the value from the sensor, the input which the LEMPC or NN-LEMPC would compute would prevent the closed-loop state from leaving Ω_ρ in a sampling period.

There is potential that the LEMPC-based strategy may be more conservative that the reachability-based method. Specifically, the LEMPC requires that an input which would keep the closed-loop state in Ω_ρ must be the input which would have been computed by the LEMPC for a false state measurement. However, the repair method adjusts the input away from what an EMPC or NN-EMPC would have computed if the state cannot be maintained within Ω_ρ for all states in $\mathcal{S}$ under that input. This may give flexibility to finding stabilizing inputs throughout state-space.

A final interesting idea that comes from this analysis is considering an extension of the repaired input concept to fighting back against attackers, in the sense of taking actions that reduce their power after they are detected. Our prior LEMPC-focused works have not handled safety after attacks. However, the repaired input option has a flavor of fighting back against attacks, since the idea is to override malicious control actions. One might build on that idea to fight back against attacks by, after detection, determining in what set the actual closed-loop state is most likely to lie, and then seeking to apply control actions that keep all such possible states safe. The goal would be to buy some time after an attack is detected before the attacker can compromise system safety.

5. CHEMICAL PROCESS EXAMPLE

In this section, we demonstrate aspects of the reachability and input repair discussion through a continuous stirred tank reactor (CSTR) with a second-order $A \longrightarrow B$ reaction, creating a process that has the dynamics:

$$\dot{C_A} = \frac{F}{V}(C_{A0} - C_A) - k_0 e^{-\frac{E}{R_g T}} C_A^2 \tag{22a}$$

$$\dot{T} = \frac{F}{V}(T_0 - T) - \frac{\Delta H k_0}{\rho_L C_p} e^{-\frac{E}{R_g T}} C_A^2 + \frac{Q}{\rho_L C_p V} \tag{22b}$$

where the states C_A and T are the concentration of the reactant A and the temperature in the reactor, respectively. The manipulated variables in this case are the feed concentration of the reactant A (i.e., C_{A0}) and the heat rate Q. The values of the parameters in Eq. (22) are taken from (Alanqar et al. (2015)). The operating steady state values are $[C_{As} \quad T_s]^T = [1.22 \text{ kmol/m}^3 \ 438.2 \text{ K}]^T$ and $[C_{A0s} \quad Q_s]^T = [4.0 \text{ kmol/m}^3 \quad 0 \text{ kJ/hr}]$. The controller considered for this process is a model predictive controller (MPC), with the form in Eq. 4, except without the constraints of Eqs. 4f-4g. This MPC used an objective function of $(T - T_s)^2$ and was simulated in Matlab using fmincon with $N = 10$. The Explicit Euler method was used to numerically integrate the differential equations of Eq. (22), with an integration step of 10^{-4} hr in the EMPC (and 10^{-5} hr for the process). The input and state constraints are: 350 K $\leq T \leq$ 500 K; 0.5 kmol/m^3 $\leq C_{A0} \leq$ 7.5 kmol/m^3; -5×10^5 kJ/hr $\leq Q \leq 5 \times 10^5$ kJ/hr. The CSTR was simulated under disturbances that were added to the right side of Eq. (22a) and Eq.(22b) with zero mean and standard deviations of 5 kmol m^{-3}hr^{-1} and 20 K/hr, and bounds of 20 kmol m^{-3}hr^{-1} and 50 K/hr, respectively. An approximation of the set of points that are reachable from $(0.82 \text{ kmol/m}^3, 446.25 \text{ K})$ was determined by simulating the system 1000 times with different seeds for the random number generator, from the same state and different disturbance realizations. A square was then placed around this region, and discretized to have 50 rows and 5 columns. A state within the resulting set of $(1.06 \text{ kmol/m}^3, 4436.21 \text{ K})$ was then selected to be a state measurement, and the corresponding control action from the MPC was computed. Under this input, an approximation of the set of states that could be reached under this input from the points in the discretized rectangle was computed by taking each vertex of the discretized rectangle as an initial state and, using the input computed by the MPC and 1000 simulations from each grid point under different realizations of the disturbances. The resulting set of states is plotted in blue circles in Fig. 1. A box containing these states is shown in red, and a green box represents a potential safe set. As shown, the possible states are far from the boundary of the safe set, indicating that the repair strategy can be simplified to occur only for states near the boundary where there would be a possibility of exiting the safe region.

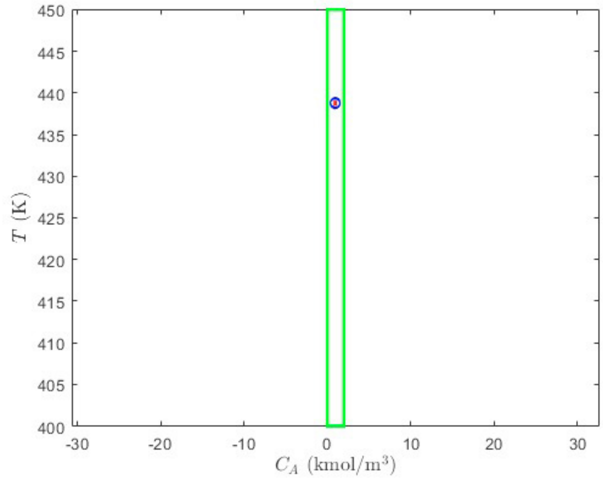

Fig. 1. Reachable set approximation in a safe region.

6. CONCLUSION

In this work, we analyze conditions under which an LEMPC-based cyberattack detection strategy can be extended to a NN-approximated version, and compare LEMPC-based cyberattack detection/handling with a reachable set and input repair approach.

REFERENCES

Åkesson, B.M. and Toivonen, H.T. (2006). A neural network model predictive controller. *Journal of Process Control*, 16(9), 937–946.

Alanqar, A., Ellis, M., and Christofides, P.D. (2015). Economic model predictive control of nonlinear process systems using empirical models. *AIChE Journal*, 61(3), 816–830.

Chen, S., Wu, Z., and Christofides, P.D. (2020). Machine learning-based cyber-attack detection and resilient operation via economic model predictive control for nonlinear processes. In *2020 28th Mediterranean Conference on Control and Automation (MED)*, 794–801. IEEE.

Heidarinejad, M., Liu, J., and Christofides, P.D. (2012). Economic model predictive control of nonlinear process systems using Lyapunov techniques. *AIChE Journal*, 58(3), 855–870.

Liu, H., Niu, B., and Qin, J. (2021). Reachability analysis for linear discrete-time systems under stealthy cyber attacks. *IEEE Transactions on Automatic Control*, 66(9), 4444–4451.

Lucia, S. and Karg, B. (2018). A deep learning-based approach to robust nonlinear model predictive control. *IFAC-PapersOnLine*, 51(20), 511–516.

Narasimhan, S., El-Farra, N.H., and Ellis, M.J. (2023). A reachable set-based scheme for the detection of false data injection cyberattacks on dynamic processes. *Digital Chemical Engineering*, 7, 100100.

Oyama, H. and Durand, H. (2020). Integrated cyber-attack detection and resilient control strategies using Lyapunov-based economic model predictive control. *AIChE Journal*, 66(12), e17084.

Wu, Z., Albalawi, F., Zhang, J., Zhang, Z., Durand, H., and Christofides, P.D. (2018). Detecting and handling cyber-attacks in model predictive control of chemical processes. *Mathematics*, 6(10), 173.

Xiang, W. and Johnson, T.T. (2018). Reachability analysis and safety verification for neural network control systems. *arXiv preprint arXiv:1805.09944*.

Yang, X., Johnson, T.T., Tran, H.D., Yamaguchi, T., Hoxha, B., and Prokhorov, D.V. (2021). Reachability analysis of deep ReLU neural networks using facet-vertex incidence. In *HSCC*, volume 21, 19–21.

Yang, X., Yamaguchi, T., Tran, H.D., Hoxha, B., Johnson, T.T., and Prokhorov, D. (2022). Neural network repair with reachability analysis. In *International Conference on Formal Modeling and Analysis of Timed Systems*, 221–236. Springer.