Lyapunov-Based Cyberattack Detection for Distinguishing Between Sensor and Actuator Attacks

Dominic Messina [*] **Helen Durand** [**]

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

Abstract: Control-theoretic cyberattack detection strategies are control strategies where control theory can be used in the design of the detection policies and analysis of stability properties with and without cyberattacks. This work provides a step toward understanding how to diagnose cyberattacks using control-theroetic cyberattack detection mechanisms. Specifically, we analyze the conditions under which a control-theoretic cyberattack detection strategy developed in our prior work to handle detection of simultaneous actuator and sensor attacks can be extended to distinguish between whether attacks are occurring on sensors or actuators. We present and evaluate heuristic concepts for attempting to diagnose sensor attacks; these again demonstrate the utility of control-theoretic diagnosis policies and lead to further suggestions for such control-theoretic policies.

Keywords: Cyber physical system, Advanced process control, Model predictive and optimization-based control, Nonlinear predictive control, Sensors and actuators

1. INTRODUCTION

Cyber-physical systems (CPSs), characterized by the integration of physical process components with computer and communication networks, can enhance the ability to monitor and control industrial processes, while adding vulnerabilities to cyberattacks on the cyber components. Detection mechanisms Narasimhan et al. (2023) Wu et al. (2018) and resilient control Sundberg and Pourkargar (2023) have been investigated for attempting to improve safety when the possibility of cyberattacks on control systems exists. The goals for securing control systems are similar to fault-tolerant control concepts of detection, diagnosis, isolation, and recovery (Isermann (1997); Patton and Chen (1997); Zhang et al. (2004). However, though abnormal dynamics due to improper functioning of a CPS may be observed in the presence of both cyberattacks and faults, cyberattacks are distinct from faults due to their intentionally malicious and coordinated nature. Due to this, there is a need for cyberattack-specific strategies for detection and diagnosis.

In our prior work Oyama et al. (2022), we investigated detection of attacks when they could occur on both sensors and actuators at the same time. We developed process operation and detection strategies that, when at least one set of sensors used by one of several redundant state estimators is not attacked, could guarantee that the closed-loop state is always maintained within a safe operating region before detection (regardless of whether sensor and actuator attacks are occurring individually or at once). However, despite the benefits of this strategy in maintaining safety before detection and facilitating detection of complex attacks, we did not provide guidance on how it could be used to diagnose which type of attack was occurring. As a result, we assumed that some type of emergency procedure (e.g., plant shut-down) would be required after the attack detection, since the location of the attack in the system was not known. To reduce the impact of an attack on a cyberphysical system, it may be desirable to consider whether the part of the plant under attack could be diagnosed, so that each piece could be dealt with individually. In Oyama et al. (2023), we discussed ideas for attempting to use a version of the detection strategies implemented with distributed control toward diagnosing attacks; however, we did not locate a clear path in moving toward diagnosis without significant redundancy.

Motivated by these considerations, in this work, we provide a new type of analysis of the detection strategies from our prior work, seeking to understand whether the detection policies themselves may carry certain diagnosis capabilities. We demonstrate that under certain conditions, the strategy from Oyama et al. (2022) can also be used in diagnosing cyberattacks. This suggests that cyberattack detection policies may be designed to facilitate not only the detection of attacks, but also their diagnosis. However, because the technique from Oyama et al. (2022) allows diagnosis only in several restrictive scenarios, this suggests that a further understanding of what it takes to develop control/detection strategies that facilitate attack diagnosis is needed. In moving toward this goal, we test a heuristic strategy for diagnosis of attacks in simulation to seek to

[⋆] Financial support from the Air Force Office of Scientific Research (award number FA9550-19-1-0059), the National Science Foundation CBET-1839675, CNS-1932026, and CBET-2143469, and Wayne State University is gratefully acknowledged.

better understand principles behind diagnosis to inspire future work on the development of detection strategies that have both diagnosis and detection capabilities.

2. PRELIMINARIES

2.1 Notation

$|x|$ and x^T denote the Euclidean norm and transpose of a vector x. A class $\mathcal{K}$ function $\alpha : [0, a) \to [0, \infty)$ is strictly increasing with $\alpha(0) = 0$. $x \in A/B$ signifies the set $\{x \in R^n : x \in A, x \notin B\}$. A level set of a positive definite function V is denoted by $\Omega_\rho := \{x \in R^n : V(x) \leq \rho\}$.

2.2 Class of Systems

The following class of nonlinear systems is considered:

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

where $x \in X \subset R^n$, $u \in U \subset R^m$ ($U := \{u \in R^m : |u| \leq u^{\max},\ u^{\max} > 0\}$, and $w \in W \subset R^z$ ($W := \{w \in R^z : |w| \leq \theta_w, \theta_w > 0\}$) are the state, input, and disturbance vectors. f is locally Lipschitz on $X \times U \times W$, and the origin is assumed to be an equilibrium point of the unforced nominal system of Eq. 1 ($f(0, 0, 0) = 0$). We assume that a sufficiently smooth Lyapunov function V exists, as well as class $\mathcal{K}$ functions $\alpha_j(\cdot)$, $j = 1, \ldots, 4$, and an asymptotically stabilizing feedback controller $h(x)$ for the system of Eq. 1 such that:

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

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

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

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

$\forall x \in D \subset R^n$ where D is an open neighborhood of the origin. $\Omega_\rho \subset D$ is called the "stability region" and is chosen such that $x \in X$, $\forall x \in \Omega_\rho$.

The smoothness of V, boundedness of u and w, and local Lipschitz property of f give:

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

$$\begin{aligned}&\left|\frac{\partial V(x_1)}{\partial x} f(x_1, u, w) - \frac{\partial V(x_2)}{\partial x} f(x_2, u, 0)\right|\\ &\quad \leq L'_x|x_1 - x_2| + L'_w|w|\end{aligned} \tag{3b}$$

$$|f(x, u, w)| \leq M_f \tag{4}$$

$\forall x_1, x_2 \in \Omega_\rho$, $u, u_1, u_2 \in U$ and $w \in W$, where L_x, L'_x, L_u, L_w, L'_w, and M_f are positive constants.

In addition, we assume that there are M sets of measurements $y_i \in R^{q_i}, i = 1, ..., M$, available at t_k:

$$y_i(t) = k_i(x(t)) + v_i(t) \tag{5}$$

where k_i is a vector-valued function and v_i represents measurement noise associated with the i-th measurement vector y_i. We assume that the measurement noise is bounded such that $v_i \in V_i := \{v_i \in R^{q_i} : |v_i| \leq \theta_{v,i}, \theta_{v,i} \geq 0\}$, and measurements of y_i are continuously available. For each of the M sets of measurements, we assume that there exists a deterministic observer described by the following dynamic equation:

$$\dot{z}_i = F_i(\epsilon_i, z_i, y_i) \tag{6}$$

where z_i is the estimate of the process state from the $i-th$ observer, $i = 1, .., M$, F_i is a vector-valued function, and $\epsilon_i > 0$. When a controller $h(z_i)$ is used with Eq. 6 to control the closed-loop system of Eq. 1, we assume that Assumption 1 and Assumption 2 below hold.

Assumption 1. There exist positive constants θ^*_w,$\theta^*_{v,i}$, such that for each pair $\{\theta_w, \theta_{v,i}\}$ with $\theta_w \leq \theta^*_w$, $\theta_{v,i} \leq \theta^*_{v,i}$, there exist $0 < \rho_{1,i} < \rho$, $e_{m0i} > 0$, and $\epsilon^*_{L,i} > 0$, $\epsilon^*_{U,i} > 0$ such that if $x(0) \in \Omega_{\rho_{1,i}}$, $|z_i(0) - x(0)| \leq e_{m0i}$, and $\epsilon_i \in (\epsilon^*_{L,i}, \epsilon^*_{U,i})$, the trajectories of the closed-loop system are bounded in Ω_ρ, $\forall t \geq 0$.

Assumption 2. There exists $e^*_{mi} > 0$ such that for each $e_{mi} \geq e^*_{mi}$, there exists $t_{bi}(\epsilon_i)$ such that $|z_i(t) - x(t)| \leq e_{mi}$, $\forall t \geq t_{bi}(\epsilon_i)$.

2.3 Lyapunov-Based Economic Model Predictive Control (LEMPC)

This work utilizes a control design known as LEMPC Heidarinejad et al. (2012), which is formulated as follows:

$$\min_{u(t) \in S(\Delta)} \int_{t_k}^{t_{k+N}} L_e(x_b(\tau), u(\tau))\, d\tau \tag{7a}$$

$$\text{s.t.}\quad \dot{x_b}(t) = f(x_b(t), u(t), 0) \tag{7b}$$

$$x_b(t_k) = x_a(t_k) \tag{7c}$$

$$x_b(t) \in X,\ \forall\, t \in [t_k, t_{k+N}) \tag{7d}$$

$$u(t) \in U,\ \forall\, t \in [t_k, t_{k+N}) \tag{7e}$$

$$\begin{aligned}&V(x_b(t)) \leq \rho_e,\ \ \forall\, t \in [t_k, t_{k+N}),\\ &\quad \text{if } x_b(t_k)\ \in\ \Omega_{\rho_e}\end{aligned} \tag{7f}$$

$$\begin{aligned}&\frac{\partial V(x_b(t_k))}{\partial x} f(x_b(t_k), u(t_k), 0)\\ &\leq \frac{\partial V(x_b(t_k))}{\partial x}\, f(x_b(t_k), h(x(t_k)), 0),\\ &\quad \text{if}\ \ x_b(t_k) \in \Omega_\rho/\Omega_{\rho_e}\end{aligned} \tag{7g}$$

where $u(t) \in S(\Delta)$ signifies that the optimal solution is a piecewise-constant input vector with N pieces. The prediction horizon consists of N sampling periods, where each sampling period has a duration of Δ. The objective function is the time-integral of the economic stage cost L_e of Eq. 7a, evaluated throughout the prediction horizon. The state predictions $x_b(t)$ of Eq. 7b are obtained using the nominal model of Eq. 1 where $w \equiv 0$ (x_a is considered to be x for the actual process from Eq. 1). The constraints of Eqs. 7d-7e are state and input constraints, respectively. The two Lyapunov-based stability constraints are given by Eqs. 7f and 7g where $\Omega_{\rho_e} \subset \Omega_\rho$. LEMPC is implemented in a receding horizon fashion, where the optimal input for $[t_k, t_{k+1})$ (denoted $u^*(t_k|t_k)$) is applied to the process in sample-and-hold at the beginning of each sampling period.

3. EXTENDING LEMPC-BASED CYBERATTACK DETECTION POLICIES TO DIAGNOSIS TASKS

In Oyama et al. (2022), we developed strategies for detecting attacks on sensors, actuators or both, under the condition that not all sensors were under attack. These strategies were investigated by building from three detection strategies originally proposed in Oyama and Durand

(2020) for detecting sensor attacks only (the three policies in Oyama and Durand (2020) were referred to in Oyama et al. (2022) as Detection Strategies 1-S, 2-S, and 3-S, where the number was used to signify one of the three strategies from Oyama and Durand (2020) and the "S" was used to signify that the strategies in Oyama and Durand (2020) were developed for the detection of sensor attacks only). In Oyama et al. (2022), we demonstrated that a similar set of three strategies could be developed for detecting actuator attacks, with different levels of success, when only the actuators were attacked (this set of three actuator attack-focused detection strategies was referred to as Detection Strategies 1-A, 2-A, and 3-A). Some of the sensor attack detection strategies were good at detecting sensor attacks but their corollaries for the actuator attack detection policies worked poorly for detecting the actuator attacks, or vice versa. As a result, our strategy for detecting attacks on sensors and actuators at once involved combining policies that worked well for sensors with those that worked well for actuators. The strategy of combining Detection Strategies 3-S and 2-A will receive focus in this work.

The Detection Strategy 3-S/2-A checks: 1) whether a set of redundant state estimators produced estimates that were consistent with one another (where at least one of these estimators was assumed to not be impacted by an attack); and 2) whether a state prediction at a time t_k (initialized from an estimate at time t_{k-1}) was within a bound of the state estimate received at t_k. If the norm of the difference between two redundant state estimates was greater than a bound, or if the norm of the difference between the prediction and the estimate was greater than a bound, an attack was flagged. When no attacks were flagged, under sufficient conditions (including that not all estimators could be impacted by an attack), we guaranteed that safety was maintained (in the sense of keeping the closed-loop state within Ω_ρ at all times) until an attack was detected.

However, this methodology was limited after the detection of an attack. Specifically, though we assumed that complex attacks could be occurring and provided a means for detecting them before they would create safety issues at a plant, we provided no framework for figuring out which type was occurring (i.e., if only sensors were impacted, or only actuators, or both, and which ones). This limits the responses that a plant can take when an attack is flagged, because it would not be known how to isolate the attack under this strategy. It would be desirable for a strategy that facilitates complex attack detection to also enable understanding of what types of attacks have occurred, and on what components. However, some strategies that are capable of detecting attacks may not be able to reveal the source of the attack with certainty. Thus, there is a need to understand what types of characteristics of an attack detection strategy lend themselves to also diagnosing the attack, to inspire further research on how to design detection strategies for complex attacks that have the properties of guaranteeing safety before detection, detection when safety could be compromised, and diagnosis of the attack to facilitate isolation of attacked equipment after the attack is detected.

An important first step toward understanding how to develop control-theoretic attack detection methods that also permit diagnosis is to probe the extent to which Detection Strategy 3-S/2-A permits diagnosis. This can then be used to help indicate properties of a control-theoretic attack detection method that make it strong or weak with respect to diagnosis, aiding with the development of strategies which overcome weaknesses in future detection strategy developments. The studies indicate that under specific circumstances, Detection Strategy 3-S/2-A can diagnose whether an attack is occurring on the sensors or the actuators, but cannot diagnose all attacks.

To further build understanding toward how to create cyberattack diagnosis characteristics in cyberattack detection policies, we use simulations to evaluate a heuristic concept for diagnosing attacks (one that has no theoretical backing). Despite its heuristic nature (which causes it to not perform well at the diagnosis task), this simulation inspires additional concepts regarding how the control-theoretic detection bounds in Oyama et al. (2022) might be utilized toward diagnosis.

3.1 Evaluating Detection Strategy 3-S/2-A for Diagnosis Properties

In this section, we analyze the ability of Detection Strategy 3-S/2-A to aid with diagnosing cyberattacks. We do not consider any means for diagnosis of undetected attacks (since abnormal behavior is not flagged when an attack is undetected, there is nothing known to diagnose).

Detection Strategy 3-S/2-A requires an output-feedback LEMPC of the form of Eq. 7, with $x(t_k)$ replaced with $z_1(t_k)$. Concurrently, it considers that there exist converged redundant state estimators which are able to estimate the full state of the system within a bound of the actual process state (denoted x_a) through the use of a subset of the state measurements, in the absence of cyberattacks. In addition to the redundant estimators, a state estimate $z_1(t_k)$ is used as an initial condition for numerically integrating the nominal model of Eq. 7b under an input similar to that calculated by the LEMPC to compute a state prediction x_b at t_{k+1}. At each sampling time t_k, state estimates from the redundant estimators, $z_i(t_k)$ and $z_j(t_k), i = 1, ..., M, j = 1, ..., M$, are compared with each other to determine whether a bound indicating a sensor attack is violated. Also, the estimate $z_1(t_k)$ is compared with the state prediction $x_b(t_k|t_{k-1})$ (indicating a prediction of the value of the state at t_k based on an estimate from t_{k-1}) to indicate whether an attack is present.

The diagnosis properties of Detection Strategy 3-S/2-A are strongly tied to a number of theoretical results regarding how bounds are set on the two detection metrics (the metrics focused on a comparison between state estimates and on a comparison between a state prediction and state estimate). Therefore, we review a number of these theoretical results that will be used in presenting the diagnosis properties.

Proposition 3. (c.f. Oyama and Durand (2020)) The following derivation of a bound on the difference between two redundant estimates in the absence of an attack holds:

$$\begin{aligned} |z_i(t) - z_j(t)| &= |z_i(t) - x_a(t) + x_a(t) - z_j(t)| \\ &\le |z_i(t) - x_a(t)| + |x_a(t) - z_j(t)| \\ \le \epsilon_{ij} := (e^*_{mi} + e^*_{mj}) &\le \epsilon_{\max} := \max\{\epsilon_{ij}\} \end{aligned} \tag{8}$$

Proposition 4. (c.f. Oyama and Durand (2020)) Consider the system of Eq. 1 under the output feedback LEMPC of Eq. 7 (with $x_b(t_k)$ replaced with $z_1(t_k)$). $M > 1$ state estimators independently estimate the state of the system, and at least one estimator is not impacted by false state measurements. Assuming attacks do not occur until after the state estimators are converged (denoted by time t_q, when $|z_i(t) - x(t)| \le e^*_{mi} \forall i = 1, ..., M$), if a cyberattack is not flagged at t_k, then the worst-case difference between z_i, $i \ge 1$ and the actual state of the system $x_a(t_k)$ is given by

$$|z_i(t_k) - x_a(t_k)| \le \epsilon^*_M := \epsilon_{max} + \max\{e^*_{mj}\}, j = 1, ..., M \tag{9}$$

Definition 5. (c.f. Oyama et al. (2022)) Consider the state trajectories for the actual process x_a and for the predicted state x_b from $t \in [t_k, t_{k+1})$, which are the solutions of the systems:

$$\begin{aligned} \dot{x}_a &= f(x_a(t), \bar{u}(t), w(t)) \\ \dot{x}_b &= f(x_b(t), \hat{u}(t), 0) \end{aligned} \tag{10}$$

where $|x_a(t_k) - z_1(t_k)| \le \gamma$. $\bar{u}$ is the optimal input for $t \in [t_k, t_{k+1})$ computed from the output-feedback LEMPC of Eq. 7 based on the estimate $z_1(t_k)$ of the actual state at t_k, where $x_b(t_k) = z_1(t_k)$. $\hat{u}$ is an input used to calculate a state prediction that results in the trajectory x_b corresponding to the predicted value of the closed-loop state. The following bound between the optimal input and the input used for state prediction is assumed to hold:

$$|\bar{u}(t) - \hat{u}(t)| \le \epsilon_u \tag{11}$$

Proposition 6. (c.f. Oyama et al. (2022)) Consider the systems defined in Definition 5 operated under the output-feedback LEMPC of Eq. 7 (where $x_b(t_k)$ is replaced with $z_1(t_k)$) and designed based on the controller $h(\cdot)$, (assumed to satisfy Eq. 2 and Lipschitz continuity of each component). The following bound holds:

$$|x_a(t) - x_b(t)| \le f_u(\gamma, t) \tag{12}$$

and initial states $|x_a(t_0) - x_b(t_0)| \le \gamma$, where $x_b(t_0) = z_1(t_0)$ and $t_0 = 0$:

$$f_u(s, \tau) := s e^{L_x t} + (e^{L_x t} - 1)(\frac{L_u \epsilon_u + L_w \theta}{L_x}) \tag{13}$$

Proposition 7. (c.f. Oyama et al. (2022)) Consider the systems x_a and x_b defined in Definition 5 operated under the output-feedback LEMPC of Eq. 7 (where $x_b(t_k)$ is replaced with $z_1(t_k)$). If $|z_i(t_k) - z_j(t_k)| < \epsilon_{max}$ and $|z_i(t_{k+1}) - z_j(t_{k+1})| < \epsilon_{max}$, $i = 1, .., M$, $j = 1, ..., M$, and Eq. 11 holds in the absence of an attack, then the following bound on the error between the state estimate $z_1(t_{k+1})$ and the state prediction $x_b(t_{k+1}|t_k)$ based on an estimate obtained at time t_k in the absence of cyberattacks holds:

$$|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)| \le \epsilon^*_M + f_u(\epsilon^*_M, \Delta) \tag{14}$$

In Oyama et al. (2022), it was suggested that a threshold could be placed on $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)|$ to aid with attack detection. Specifically, in the absence of attacks, $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)|$ should never exceed $\epsilon^*_M + f_u(\epsilon^*_M, \Delta)$. Thus, if ν_u represents a threshold on $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)|$ above which attacks are flagged, setting the threshold to $\epsilon^*_M + f_u(\epsilon^*_M, \Delta)$ would prevent false detections. The two major components of the Detection Strategy 3-S/2-A are therefore: 1) checking whether $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)| > \nu_u$ and 2) checking whether $|z_i(t_k) - z_j(t_k)| > \epsilon_{max}$, $i = 1, .., M$, $j = 1, ..., M$.

We can now begin to analyze the conditions under which those two detection bounds are violated, and assess whether these have implications for diagnosing attacks. For example, consider the detection metric $|z_i(t) - z_j(t)| \le \epsilon_{\max}$, $i, j = 1, \ldots, M$. If we consider that an attack can only occur after the state estimate is converged (and that the state estimates remain converged regardless of which control action is applied, including if rogue control actions are applied in an actuator attack), then violation of this bound can happen only if sensors are attacked. The conclusion is that violation of this metric signifies that at least a sensor attack is occurring (however, violation of this bound does not provide sufficient data to indicate that an actuator attack is occurring or is not occurring).

In addition, we can analyze the conditions under which $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)| > \nu_u \ge \epsilon^*_M + f_u(\epsilon^*_M, \Delta))$. When $|z_i(t) - z_j(t)| \le \epsilon_{\max}$, $i, j = 1, \ldots, M$, and not all estimates are impacted by a sensor attack, ϵ^*_M is an upper bound on terms related to estimate and state measurement deviations used in deriving this expression, indicating that the cause of the violation of the bound is that $|x_a(t_{k+1}) - x_b(t_{k+1}|t_k)| > f_u(\epsilon^*_M, \Delta)$, which would be occurring due to the value of ϵ_u not being respected in f_u. This would suggest that at least an actuator attack is present on the system if this detection metric is violated (and if the other detection bound is not violated). This does not guarantee that there is not an undetected sensor attack. Overall, this indicates that extreme attacks on sensors (enough to cause the first detection bound to be violated) can reveal that at least sensors are attacked (but do not show if actuators are attacked), that extreme attacks on actuators (enough to cause the second detection bound to be violated) can reveal that at least actuators are attacked (but that would not show if there is an undetected sensor attack), but that if both detection bounds are violated, it would be known that at least a sensor attack is occurring but that it could not be clarified whether the sensors are causing the second detection bound to be violated or if there is also an actuator attack. Based on this analysis, an implementation strategy and a theorem that highlights this result are presented below.

An implementation strategy for this sensor and actuator attack distinguishing concept follows, where we consider the set of redundant estimators to be z_1 and z_2 and assume that the process has been run in the absence of attacks for some time t_q, so that $|z_i(t) - x(t)| \le e^*_{mi} \forall i = 1, 2$:

1. At a sampling time t_k, the output-feedback controller receives the state estimate $z_1(t_k)$ and computes inputs for each sampling period from t_k to t_{k+N}. The input $u^*(t_k)$ is applied to the process.
2. Evaluate $|z_1(t_k) - z_2(t_k)|$. If $|z_1(t_k) - z_2(t_k)| > \epsilon_{max}$, go to Step 4a. Else, solve the LEMPC optimization problem and send the first input to the process. Calculate $x_b(t_{k+1}|t_k)$ based on $z_1(t_k)$. Proceed to Step 3.

(3) Evaluate $|z_1(t_{k+1}) - z_2(t_{k+1})|$. If it is greater than ϵ_{max}, go to Step 4a. Else, go to Step 4.
(4) Evaluate $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)|$. If it is greater than ν_u, go to Step 4b. Else, no attacks are detected on the system, proceed to Step 5.
  (a) Detect that at least a sensor is being attacked and apply mitigating actions.
  (b) Detect that an actuator is being attacked and apply mitigating actions.
(5) Set $t_k \leftarrow t_{k+1}$. Go to Step 1.

The following theorem characterizes the conditions under which the strategy above guarantees detection of a sensor attack or the detection of an actuator attack in the case that $M = 2$ and one of the estimates is not impacted by an attack.

Theorem 8. Consider x_a and x_b defined in Definition 5, where $x_b(t_k) = z_1(t_k)$, and one of the two estimates z_1 or z_2 is not impacted by an attack. Assuming the bound of Eq. 11 and bounded measurement noise and disturbances, if $|z_1(t_k) - z_2(t_k)| > \epsilon_{\max}$, or if $|z_1(t_{k+1}) - z_2(t_{k+1})| > \epsilon_{\max}$ but $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)| \leq \nu_u$, a false sensor measurement attack is present in the system. If instead $|z_1(t_{k+1}) - z_2(t_{k+1})| \leq \epsilon_{max}$ but $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)| > \nu_u$, then an actuator attack is present in the control system.

Proof. To prove Theorem 8, we first note that $|z_1(t_k) - z_2(t_k)| > \epsilon_{max}$ may only occur if either $z_1(t_k)$ or $z_2(t_k)$ violates $|z_i(t) - x(t)| \leq e^*_{mi}$ ($i = 1$ or 2) since:

$$\begin{aligned} |z_1(t) - z_2(t)| &\leq |z_1(t) - x_a(t)| + |x_a(t) - z_2(t)| \\ &\leq e^*_{m1} + e^*_{m2} \leq \epsilon_{max} \end{aligned} \tag{15}$$

in the absence of attacks. Thus, either $|z_1(t) - x_a(t)| > e^*_{m1}$ or $|x_a(t) - z_2(t)| > e^*_{m2}$, indicating an attack is occurring on some sensors. Next, assuming no sensor attack is detected on the system, since $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)| \leq |z_1(t_{k+1}) - x_a(t_{k+1})| + |x_a(t_{k+1} - x_b(t_{k+1}|t_k)| \leq \epsilon^*_M + f_u(\epsilon^*_M, \Delta) \leq \nu_u$ when there are no attacks Oyama et al. (2022), and ν_u is selected such that $\nu_u \geq \epsilon^*_M + f_u(\epsilon^*_M, \Delta)$, then if $|z_1(t_{k+1}) - x_b(t_{k+1}|t_k)| > \nu_u$, this implies that somehow the reverse of the situation in the no-attack case occurred. The reason for that reversal could not be that sensors were attacked, since ϵ^*_M is defined as an upper bound on $|z_i(t_k) - x(t_k)|$ when sensor attacks are undetected by checking whether $|z_1(t_k) - z_2(t_k)| \leq \epsilon_{max}$. Thus, the reason for the reversal must lie in an issue with the term $f_u(\epsilon^*_M, \Delta)$ not upper bounding $|x_a(t_{k+1}) - x_b(t_{k+1}|t_k)|$, which could occur if the bound on $|\bar{u}(t) - \hat{u}(t)|$ is not respected (indicating an actuator attack).

3.2 Investigations in Cyberattack Diagnosis Through a Chemical Process Example

While the strategy described in Section 3.1 under certain conditions may provide more information about which set components of a system are affected by a cyberattack, to achieve diagnosis, the specific components which are affected by the attack must be located. Determining strategies which locate attacked components requires analyzing potential methods which meet detection and safety requirements, including under complex attacks, and also facilitate diagnosis. The design of such a strategy improving on detection strategies to enable diagnosis capabilities is not immediately obvious. Therefore, it is necessary to investigate potential directions to elucidate what is required for a cyberattack diagnosis strategy. To inspire concepts for diagnosis, we analyze a simulation of a naive idea for cyberattack diagnosis. The specific concept is inspired by an LEMPC-based sensor attack detection strategy from Oyama and Durand (2020). Specifically, in Oyama and Durand (2020), a method for attack detection is presented that operates a process in a manner that should decrease the Lyapunov function between certain sampling periods. Specifically, the controller for the system is randomly switched to an LEMPC designed around a new steady-state and with the constraint of Eq. 7g always utilized, so that the value of the Lyapunov function for this i-th steady-state (V_i) should decrease over the subsequent sampling period (a lack of decrease would flag an attack). As a naive concept for diagnosis inspired by this detection strategy, we will analyze whether analyzing components of the time derivative of the Lyapunov function (and whether they are increasing or decreasing) could provide any guidance toward attack diagnosis.

A Lyapunov function is scalar-valued (and therefore has a scalar-valued time derivative); however it is constructed from the dot product of two vectors, $\frac{\partial V_i(x_i(t_k))}{\partial x_{j,i}}$ and $f_{j,i}(x_i(t_k), u_i(t_k), 0)$, where, $f_{j,i}$ represents the j-th component of a vector function f_i that represents f in deviation variable form with respect to the i-th steady-state (x_i and u_i represent the state and input vectors in deviation form from this steady-state, where $x_{j,i}$ is the j-th component of this deviation form of the state vector). Here, we explore patterns in the variation of each term in $\dot{V}_i$ under normal operation (to evaluate whether there would be benefits in exploring the variations when a cyberattack occurs on the sensors). Specifically, assuming that we wish to see a decrease in $\dot{V}_i$ (inspired by the cyberattack detection concept mentioned above), we will calculate each term in $\dot{V}_i$ separately to determine if any of these terms are individually contributing positive values to $\dot{V}_i$ that might make it less likely to decrease, whether this is indicative of abnormal behavior, and if this may be used as part of a strategy for diagnosing sensor cyberattacks (e.g., if no positive terms are observed under normal behavior, then observing positive terms may be indicative of a cyberattack on the sensor measuring the state contributing the positive term). In general, $\dot{V}_i$ can be a complex function of interactions of multiple states, so that isolating effects in individual terms may not be possible; however, exploring how this diagnosis idea performs can still give insights to guide further diagnosis ideas.

We analyze this concept with a Lyapunov-based control law using a continuous stirred tank reactor example from Alanqar et al. (2015). The parameters and dynamic model are taken from Alanqar et al. (2015) and the states are the reactant concentration C_A and temperature T. C_{A0} (the reactant feed concentration) is set to 4 kmol/m^3 and Q (the heat rate) is bounded between $-5 \times 10^5 \leq Q \leq 5 \times 10^5$ kJ/h. The Lyapunov-based controller is designed using the Lyapunov function $V_1 = x^T P x$, where $P = [1200\ 5; 5\ 0.1]$. The stabilizing controller for Q was designed via Sontag's control law Lin and Sontag (1991). The process was run for 0.15 h, with an integration step

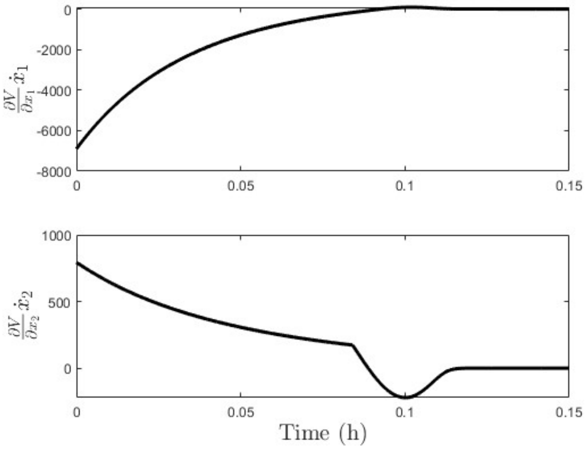

Fig. 1. Time derivative of V along the state trajectories of C_A (above) and T (below).

in the Explicit Euler numerical integration method of 10^{-4} h. The initial condition is $C_A = 0.823$ kmol/m^3 and $T = 446.25$ K. The values of $\frac{\partial V}{\partial x_1}\dot{x}_1$ and $\frac{\partial V}{\partial x_2}\dot{x}_2$) are plotted (with x_1 and x_2 representing the states C_A and T in deviation form from the steady-state corresponding to $C_A = 1.22$ kmol/m^3, $T = 438.2$ K, and $Q = 0$ kJ/h. Though V decreases, the individual terms comprising $\dot{V}$ are not necessarily always negative during normal operation. This is due to a combination of the controller and the process dynamics. For a strategy like this to be able to be used in diagnosis, it would need to be possible to find a process with a compatible control law such that the closed-loop dynamics cause $\dot{V}$ to be negative with each term being negative. This provides insights into how one might begin to conceive of ways of creating diagnosis methods, and the specifications of good diagnosis strategies.

Other ideas for incorporating diagnosis in LEMPC-based methods for cyberattack detection could utilize other detection metrics. For example, if the sensor noise associated with each sensor in the system can be characterized, it may be possible to utilize a strategy similar to the detection strategy from Oyama and Durand (2020) based on detecting attacks when a state prediction and measurement were not sufficiently close to one another. Instead of a bound placed on the Euclidean norm of the difference between the state prediction and state measurement as is done in Oyama and Durand (2020) for detection, one could consider placing separate bounds on the Euclidean norm of the difference of each element in the state prediction and state measurement vectors. Specifically, instead of the bound $|x_b(t_k|t_{k-1} - x_b(t_k|t_k)| > \nu$, developing bounds on individual states of the system (e.g. $|x_{bi}(t_k|t_{k-1} - x_{bi}(t_k|t_k)| > \nu_i, i = 1, ..., n$) may be an idea for diagnosing which sensor in the control system is being attacked. The fact that the heuristic strategy explored in this section does not have a clear pathway to being used without potential false detection indicates that strategies which integrate detection, control, and diagnosis policies with control-theoretic guarantees pose significant benefits in streamlining the attack detection process.

4. CONCLUSIONS

In this work, in moving towards the development of a control-theoretic cyberattack diagnosis strategy, we presented an LEMPC-based cyberattack detection method which provides more information about the location of a cyberattack on a cyber-physical system relative to our prior cyberattack detection strategies in terms of distinguishing whether an attack is present on the set of actuators or on the set of sensors.

REFERENCES

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.

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

Isermann, R. (1997). Supervision, fault-detection and fault-diagnosis methods—an introduction. *Control engineering practice*, 5(5), 639–652.

Lin, Y. and Sontag, E.D. (1991). A universal formula for stabilization with bounded controls. *Systems & Control Letters*, 16, 393–397.

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 cyberattack detection and resilient control strategies using Lyapunov-based economic model predictive control. *AIChE Journal*, 66, e17084.

Oyama, H., Messina, D., Rangan, K.K., and Durand, H. (2022). Lyapunov-based economic model predictive control for detecting and handling actuator and simultaneous sensor/actuator cyberattacks on process control systems. *Frontiers in Chemical Engineering*, 4, 810129.

Oyama, H., Messina, D., Rangan, K.K., Leonard, A.F., Nieman, K., Durand, H., Tyrrell, K., Hinzman, K., and Williamson, M. (2023). Development of directed randomization for discussing a minimal security architecture. *Digital Chemical Engineering*, 6, 100065.

Patton, R.J. and Chen, J. (1997). Observer-based fault detection and isolation: Robustness and applications. *Control Engineering Practice*, 5(5), 671–682.

Sundberg, B. and Pourkargar, D.B. (2023). Cyberattack awareness and resiliency of integrated moving horizon estimation and model predictive control of complex process networks. In *2023 American Control Conference (ACC)*, 3815–3820. IEEE.

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.

Zhang, X., Parisini, T., and Polycarpou, M.M. (2004). Adaptive fault-tolerant control of nonlinear uncertain systems: an information-based diagnostic approach. *IEEE Transactions on automatic Control*, 49(8), 1259–1274.