Detection of Man-in-the-Middle Attacks in Model-Free Reinforcement Learning

Rishi Rani SMR@UCSD.EDU and **Massimo Franceschetti** MFRANCESCHETTI@ENG.UCSD.EDU
Dept. of Electrical and Computer Engineering,
University of California, San Diego
La Jolla, CA-92093

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

This paper proposes a Bellman Deviation algorithm for the detection of man-in-the-middle (MITM) attacks occurring when an agent controls a Markov Decision Process (MDP) system using model-free reinforcement learning. This algorithm is derived by constructing a "Bellman Deviation sequence" and findind stochastic bounds on its running sequence average. We show that an intuitive, necessary and sufficient "informational advantage" condition must be met for the proposed algorithm to guarantee the detection of attacks with high probability, while also avoiding false alarms.

Keywords: Cyber-Physical Systems, Learning Based Attacks, Man-in-the-Middle Attacks, Model-Free Reinforcement Learning.

1. Introduction

Recent advancements in wireless technology and computation have enabled the possibility of performing networked control in cyber-physical systems (CPS), leading to a multitude of applications such as cloud robotics, autonomous navigation and industrial processes (Kehoe et al., 2015). These modern learning and decision making systems are inherently online as they make decisions on the fly, in a closed-loop fashion and based on past observations. However, the distributed nature of CPS leads to security vulnerabilities that drives a need for developing secure optimal control strategies. The consequences of security breaches can be catastrophic as the attackers' target can range from systems for financial gain, to hijacking autonomous vehicles or unmanned aerial vehicles, to breaching life-critical systems as an act of terror (Urbina et al., 2016; Dibaji et al., 2019a; Jamei et al., 2016). Some instances of attacks that were discovered and made public include the Ukraine power grid cyber-attack, the German steel mill cyber-attack, the revenge sewage attack in Australia, the David Besse nuclear power plant attack in Ohio and the Iranian uranium enrichment facility attack by the Stuxnet malware (Sandberg et al., 2015). These recent events motivated several studies on prevention of security breaches at a control-theoretic level (Bai et al., 2017; Dolk et al., 2017; Shoukry et al., 2016; Chen et al., 2016; Shi et al., 2018; Dibaji et al., 2018; R. et al., 2018; Niu et al., 2021; Chong et al., 2019; Tomić et al., 2018; Ding et al., 2019; Teixeira et al., 2015; M. Xue and Das, 2012; Cetinkaya et al., 2017; Brown et al., 2019; Law et al., 2015; Pirani et al., 2021; Hashemi et al., 2018). In this general framework, the "man-in-the-middle" (MITM) class of attacks in CPS is an important paradigm that has been widely studied (Smith, 2011). An adversary overrides the sensor feedback signals transmitted from the physical plant to the legitimate agent with spoofed signals that mimic safe and stable operation. Simultaneously, the plant is pushed towards a catastrophic trajectory by overriding the control signal with malicious inputs. The legitimate agent must therefore constantly monitor the plant outputs and look for statistical anomalies in the spoofed feedback signals to detect such attacks. The adversary, on the other hand, aims to generate spoofed sensor readings in a way that would be indistinguishable, in a statistical sense, from the legitimate ones while at the same time attempting to drive the system to a catastrophic state.

Two special cases of the MITM attack have been studied extensively. The first case is the *replay attack*, in which the adversary observes and records the true system behavior for a given time period and then replays this recording periodically at the agent's input (Mo et al., 2015; Zhu and Martínez, 2014; Miao et al., 2013). The second case is the *statistical-duplicate attack*, here the adversary is assumed to have perfect knowledge of the system dynamics therefore allowing the adversary to construct arbitrarily long trajectories that are statistically identical to the true system (Smith, 2011; Satchidanandan and Kumar, 2017; Hespanhol et al., 2018). The replay attack, by nature, is relatively easy to detect as it assumes no knowledge of system parameters. One strategy to counter replay attacks is to superimpose a watermark signal on the control signal, unbeknownst to the adversary (Hespanhol et al., 2018; Fang et al., 2017; Hosseini et al., 2016; Ferdowsi and Saad, 2019; Liu et al., 2018). The statistical-duplicate attack assumes full knowledge of the system dynamics and parameters. As a consequence, it is barred from observing the control actions, as otherwise it would be omniscient and undetectable. Due to the adversary having complete information, it requires a more sophisticated detection procedure to ensure it can be detected. To combat the adversary's full knowledge, the agent may adopt *moving target* (Weerakkody and Sinopoli, 2015; Kanellopoulos and Vamvoudakis, 2020; Zhang et al., 2020; Griffioen et al., 2019) or *baiting* (Dibaji et al., 2019b; Hoehn and Zhang, 2016) techniques. Alternatively, introducing private randomness through *watermarking* also proves to be a viable strategy (Satchidanandan and Kumar, 2017).

Another class of MITM attacks are *learning-based attacks*, which are related to the broader study of learning based control (Fisac et al., 2019a; Berkenkamp et al., 2017; Fisac et al., 2019b; Yuan and Mo, 2015; Tu and Recht, 2018). In learning based attacks, the adversary initially has no knowledge of the system dynamics, but spends some time learning the system from observation before it hijacks the control signal to achieve catastrophic effects while attempting to remain undetected. This paradigm is more practical, as it is unreasonable to assume perfect knowledge of system models as is done in a statistical duplicate attacks. Yet, it remains powerful, as the adversary learned model may allow sophisticated deception schemes instead of relying on simple techniques like the replay attack. Using an information theoretic approach, upper and lower bounds were drawn on the asymptotic probability of deception for scalar and vector linear time invariant systems (Khojasteh et al., 2021). Similar approaches were used to draw bounds on the time required by an agent to declare a deception attack or no breach with a certain confidence, along with lower bounds on the adversaries training time and energy spent by the agent to guarantee a certain confidence in detection (Rangi et al., 2021).

Our contributions are as follows: we extend the model of learning-based attacks to include the learning of the agent itself. Specifically, we consider a legitimate agent performing model-free control through reinforcement learning (RL). In this context, since the agent has no explicit model of the system, attack detection (AD), which typically occurs through the observation of anomalous behavior, becomes particularly challenging. Detection, in our case, is performed by careful monitoring of the Q-function, which provides an implicit model of the system. We propose an AD algorithm, named the "Bellman Deviation Detection" algorithm. The proposed algorithm asymptotically guarantees AD with high probability while also avoiding false alarms, when an "informational advantage" condition is met. The informational advantage condition relates the error in the agent's Q-function to the adversary's error in the model parameters. The analysis also provides useful insights into the nature of the problem in terms of the information pattern required for successful detection. Finally, we point out that our analysis accounts for errors in the learning techniques of both the agent and the adversary, models the system as an MDP rather than a deterministic system, and assumes that the reward function is unknown and rewards are subject to added white noise. These assumptions are made in an effort to make the analysis closer to real-world scenarios.

2. Mathematical Preliminaries and Notation

A Markov Decision Process is defined by the quadruple $(\mathcal{X}, \mathcal{U}, \mathbf{P}, r)$, where $\mathcal{X}$ is the set of states with cardinality $|\mathcal{X}| = N$ and $\mathcal{U}$ is the set of actions with cardinality $|\mathcal{U}| = M$, while $\mathbf{P}$ is the transition probability matrix and $r()$ is the reward function. The probabilistic transitions from state to state are Markov and are given by

$$Pr(x_{t+1}|x_t, u_t) \sim \mathbf{p}_{x_t,u_t} \equiv [p_{x_t,u_t}(x_1), \dots p_{x_t,u_t}(x_N)] \tag{1}$$
$$\text{and } \mathbf{P} = \begin{bmatrix} \mathbf{p}_{x_1,u_1} \\ \vdots \\ \mathbf{p}_{x_N,u_M} \end{bmatrix}.$$

Similarly, the reward for each transition from state x_t by action u_t is given by

$$r(x_t, u_t) \triangleq r_{x_t,u_t} = \mathbb{E}_{x_{t+1}\sim\mathbf{P}}\ r(x_t, u_t, x_{t+1}). \tag{2}$$

The model-free control objective is to learn a policy function $\pi(x) : \mathcal{X} \to \mathcal{U}$ such that the following discounted reward is maximized

$$\pi^*(x) = \arg\max_{\pi} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t), x_{t+1})\right], x_0 \in \mathcal{X}, \tag{3}$$

where γ is the discount factor and represents how much the future reward is discounted. This problem is termed the *infinite time horizon discounted reward problem.* This objective is achieved by learning the optimal Q-function of the problem, which is

$$Q^*(x) = \max_{\pi} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t), x_{t+1})\right], x_0 \in \mathcal{X}. \tag{4}$$

The optimal Q-function relates to the optimal policy as $\pi^*(x) = \arg\max_u Q^*(x, u)$ and the optimal value function, which describes the total accrued reward of an optimal trajectory, is defined as

$$V^*(x) = \max_{x} Q^*(x, u), \tag{5}$$
$$\mathbf{v} = [V^*(x_1) \dots, V^*(x_N)],$$

where $\mathbf{v}$ denotes the optimal value function as a vector. Finally, we note that the optimal Q-function can be recursively written using the *Bellman equation* as

$$\begin{aligned} Q^*(x,u) =& r(x,u) + \gamma \sum_{x' \in \mathcal{X}} p(x,u,x') \cdot \left(\max_{u'} Q^*(x',u')\right) \\ =& r(x,u) + \gamma \sum_{x' \in \mathcal{X}} p(x,u,x') \cdot V^*(x') \\ =& r(x,u) + \gamma \ \mathbf{p}_{x,u} \mathbf{v}^T. \end{aligned} \tag{6}$$

Throughout out the paper we describe vectors using bold face and vectors are row vector by default (to align with MDP convention). Matrices are bold face and capitalized, $\|\cdot\|_2$ refers to the vector euclidean norm. Finally, we say that an event occurs with high probability (w.h.p.) if its probability p_n tends to one as the parameter n tends to infinity.

3. Problem Setup

The system is modeled as an MDP that is controlled by an agent receiving a reward that is corrupted by additive white noise. The reward noise w_t, is i.i.d., with zero mean and variance maybe infinite. We assume that the agent has learned an estimate of the optimal Q-function of the system using a trajectory $\boldsymbol{\tau}_A$ described as

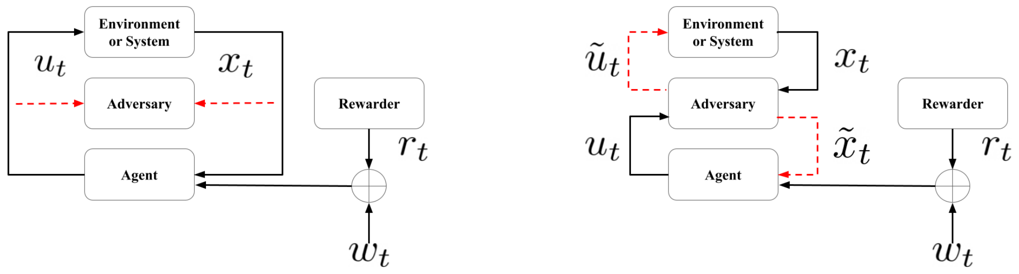
$$\boldsymbol{\tau}_A = (x_1^A, u_1^A, \dots x_{t_A}^A, u_{t_A}^A), \tag{7}$$

where t_A is the agent training time. No additional assumption is made on τ_A itself and the trajectory can be controlled by the agent. The agent has no information about the system model or reward function and uses a generalised learning algorithm with the following stochastic guarantee

$$\begin{gathered} |\hat{Q}_{t_A}(x,u) - Q(x,u)| \leq \varepsilon(t_A), \text{w.h.p} \\ \text{and } \forall x \in \mathcal{X}, u \in \mathcal{U} \\ \text{s.t } \epsilon(t) \to 0 \text{ as } t \to \infty. \end{gathered} \tag{8}$$

As described in Figure 1(*a*)subfigure, the adversary initially is in its learning phase where it observes a trajectory $\boldsymbol{\tau}_B$ and it learns the system giving it an estimate of the transition model $\hat{\mathbf{P}}$. During its learning phase the adversary has no control over its learning trajectory $\boldsymbol{\tau}_B$, as it merely learns by observing and does not control the system. Therefore, no asymptotic convergence guarantees are placed on its estimate $\hat{\mathbf{P}}$. In the attack phase (as described in Figure 1(*b*)subfigure) the agent takes control of the system and feeds the agent a spoofed state feedback signal. This feedback signal is statistically consistent with its transition model estimate $\hat{\mathbf{P}}$. Note that $\hat{\mathbf{P}}$ need not be an be an explicit estimate made by the adversary (for example the adversary may also use model-free learning), however there exists an implicit statistical model it follows. The trajectory $\boldsymbol{\tau}_C$formed during the attack phase is used by the agent to detect for perform AD. The adversary in this phase steers the true system towards catastrophe and the agent is tasked with detecting the attack and declaring a breach. The adversary's strategy to lead the system to catastrophe does not affect AD, namely the adversary's closed feedback with system is not of strict concern to the detection problem.

Problem Statement: Given the agent has a learned estimate of the optimal Q-function $\hat{Q}()$ and the adversary spoofs the system with a transition model estimate $\hat{\mathbf{P}}$, devise a detection algorithm that uses the trajectory during attack $\boldsymbol{\tau}_C$ and provides guarantees on AD as the trajectory length $t_C \to \infty$.



(a) **Adversary Learning Phase:** During this phase, the attacker eavesdrops and learns the system, without altering the feedback signal to the agent.

(b) **Adversary Attack Phase:** During this phase, the adversary hijacks the system and intervenes as a MITM in two places: acting as a fake system to the agent and acting as a fake agent to the system.

Figure 1: Adversary Attack Model

4. A Detection Algorithm Based on "Bellman Deviation"

In this section we describe our proposed algorithm and prove its stochastic guarantees.

4.1. Algorithmic Description

Before we describe the detection algorithm, we start by defining all the required quantities. The trajectory during attack is a tuple of the form

$$\tau_C = (x_1^C, u_1^C, \dots x_{t_C}^C, u_{t_C}^C). \tag{9}$$

Let $t_C(i,j)$ be the number of times the state action pair (i,j) is observed and the sequence $x_{i,j}(k)$ and the $u_{i,j}(k)$ are the respective states and actions that followed them each subsequent time. Similarly let $r_{i,j}$ be the immediate reward doled out at that instant and $w_{i,j}(k)$ be its associated white noise.

Definition 1 (Bellman Deviation Sequence) *Let*

$$\begin{aligned} d_{i,j}(k) =& \hat{Q}(i,j) - r_{i,j} - w_{i,j}(k) \\ & - \gamma \hat{V}(x_{i,j}(k)) \quad , \forall k \in [1, t_C(i,j)], \end{aligned} \tag{10}$$

be the Bellman deviation sequence . This sequence represents the deviations from Bellman like behavior in the observed trajectory during the attack phase.

The Bellman deviation sequence (BDS) is simply the temporal difference (TD) errors separated by state-action pair to form $M \times N$ different sequence. Each representing the sequence of TD errors measured in the trajectory when the system transitioned through the respective state-action pair.

Definition 2 (Bellman Deviation Average) *Let*

$$\bar{d}_{i,j} = \frac{\sum_{k=1}^{t_C(i,j)} d_{i,j}(k)}{t_C(i,j)} \tag{11}$$

be the Bellman deviation averages (BDAs). This average helps us eliminate the disturbances we find due to noise in rewards and the stochastic transitions.

The Bellman deviation average (BDA) is simply an average of the BDS. We use bounds on the BDA to determine if the system is under a MITM attack. A high BDA would suggest that the system is under attack. To draw the exact bounds on the deviation averages however, we need to define useful measures on the system and adversary model estimates as well.

Definition 3 (Maximum System Discernibility) *Given an MDP system $(\mathcal{X}, \mathcal{U}, \mathbf{P}, r)$, we can define its system discernibility as*

$$\Phi(\mathbf{v}) = \frac{\gamma \cdot \|\mathbf{v} - \boldsymbol{\mu}(\mathbf{v})\|_2}{\sqrt{N}}, \tag{12}$$

where $\mathbf{v}$ is the associated optimal value function represented as a vector and the function $\boldsymbol{\mu}(\cdot)$ is a function that returns a vector (of same dimension $1 \times N$) where all the elements are the simple average of the input vector.

The above definition can be understood intuitively as a measure that tells us how easy it is to observe deviation in that system's trajectory during the attack phase. For example, if system with $\Phi(\mathbf{v}) = 0$. This implies that the value function gives us no information about the different trajectories as they have the same accrued reward. This makes the a deviation from optimal trajectory indiscernible and hence AD infeasible. So the system discernibility measure is a key feature of the system and should be kept in mind while designing secure systems.

Finally, we define a quantity to measure the minimum error in an adversary's system model.

Definition 4 (Minimum Adversary Model Error) *Given the system state transition model is $\mathbf{P}$ and the adversary estimate is $\hat{\mathbf{P}}$ we define the minimum adversary model error as*

$$\Delta(\mathbf{P}, \hat{\mathbf{P}}) = \sigma_2(\mathbf{P} - \hat{\mathbf{P}}) = \sigma_2(\tilde{\mathbf{P}}), \tag{13}$$

where the function $\sigma_2(\cdot)$ returns the second smallest singular value of the matrix.

The minimum adversary model error gives us a measure of the minimum error of the adversary's estimate of the conditional distribution $\hat{\mathbf{p}}$ across all state-action pairs. Note that the rows of probability error matrix $\tilde{\mathbf{P}}$ sum to 0, since its the difference of two stochastic matrices. Therefore its smallest singular value is trivially 0 making the second smallest singular value a good measure of minimum error. With the above quantities defined we are now ready to present the Bellman deviation detection algorithm (see Algorithm 1) and prove its correctness.

In Algorithm 1 the division $\frac{\mathbf{D}}{\mathbf{T}}$ is an element-wise division of the two matrices. The algorithm essentially calculates the BDAs $\bar{d}_{i,j}$, takes the maximum value among them and compares it to the bound $\xi = \delta \cdot \phi - (1 + \gamma)\epsilon$. If it crosses this bound a breach is declared. Note that the condition $\delta \cdot \phi >= 2 \cdot (1 + \gamma)\epsilon$ is the informational advantage condition that essentially puts an upper-bound on the adversary errors with respect to the adversary's model error. The algorithm guarantees AD and no false alarms, with high probability, if and only if this condition is met.

Remark 5 *We point out how the algorithm does not need exact estimates of the error bound on the Q-function $\varepsilon(t_A)$, the system discernibility $\Phi(\mathbf{v})$ or minimum adversary model error $\Delta(\mathbf{P}, \hat{\mathbf{P}})$, but only an over estimate (ϵ) or under estimate (ϕ, δ) respectively. This allows for a more practical scenarios where exact values of these quantities would be unavailable and could be obtained by bootstrap methods.*

Algorithm 1: Bellman Deviation Detection

```
require:
$t_C \geq 0$, length($\boldsymbol{\tau}_C$)= $t_C$          // run when trajectory is non-empty
$\epsilon \geq \varepsilon(t_A)$                        // have an over estimate of agent error
$\delta \leq \Delta(\mathbf{P}, \hat{\mathbf{P}})$   // have an under estimate of adversary minimum error
$\phi \leq \Phi(\mathbf{v})$          // have an under estimate of system discernibility
$\delta \cdot \phi \geq 2 \cdot (1+\gamma)\epsilon$          // meet informational advantage condition
initialize:
$\xi \leftarrow \delta \cdot \phi - (1+\gamma)\epsilon$                       // Set Bellman deviation bound
$\mathbf{D} \leftarrow [\mathbf{0}]_{M \times N}$                   // Initialize Bellman deviation averages
$\mathbf{T} \leftarrow [\mathbf{0}]_{M \times N}$               // Initialize counter for state action pairs
for $i \leftarrow 1$ to $t_C$ do
    $i \leftarrow \boldsymbol{\tau}_C[n][0]$                                            // current state
    $j \leftarrow \boldsymbol{\tau}_C[n][1]$                                           // current action
    $k \leftarrow \boldsymbol{\tau}_C[n+1][0]$                                            // next state
    $\mathbf{D}[i,j] \leftarrow \hat{Q}(i,j) - r(i,j,k) - \gamma\hat{V}(k) + \mathbf{D}[i,j]$  // sum TD errors in Bellman
     deviation sequence
    $\mathbf{T}[i,j] \leftarrow \mathbf{T}[i,j] + 1$                                  // increment counter
end
$\mathbf{D} \leftarrow \frac{\mathbf{D}}{\mathbf{T}}$                 // normalize to get Bellman deviation averages
if $\max(\mathbf{D}) > \xi$          // compare largest deviation average with bound
 then
    declare breach
else
    declare no breach
end
```

4.2. Correctness of the Algorithm

In this section we prove the correctness of the proposed algorithm. We first start by proving an asymptotic upper bound on the BDAs if no attack is underway. Complete proofs of the Theorems 6 and 7 can be found in the supplementary material (Rani and Franceschetti, 2022).

Theorem 6

In the case when no attack takes place, we have that the following inequality holds for all BDAs,

$$|\bar{d}_{i,j}| \leq (1+\gamma)\varepsilon(t_A), \textit{w.h.p as} \\ t_C(i,j) \to \infty \;\; \forall (i,j) \in \mathcal{X} \times \mathcal{U}, \tag{14}$$

where $\varepsilon(t_A)$ is the error in the agent's estimate of the optimal Q-function.

Proof Sketch

We rearrange the terms of the Bellman equation (6) and subtract it from (11) to get,

$$
\begin{aligned}
\bar{d}_{i,j} =& \frac{\sum_{k=1}^{t_C(i,j)} \hat{Q}(i,j) - r_{i,j} - w_{i,j}(k) - \gamma \hat{V}(x_{i,j}(k))}{t_C(i,j)} \\
& - Q^*(i,j) + r_{i,j} + \gamma \mathbf{p}_{i,j} \mathbf{v}^T \\
=& \frac{\sum_{k=1}^{t_C(i,j)} \left(\hat{Q}(i,j) - Q^*(i,j) \right)}{t_C(i,j)} - \left(\frac{\sum_{k=1}^{t_C(i,j)} r_{i,j}}{t_C(i,j)} - r_{i,j} \right) \\
& - \gamma \left(\frac{\sum_{k=1}^{t_C(i,j)} \hat{V}(x_{i,j}(k))}{t_C(i,j)} - \mathbf{p}_{i,j} \mathbf{v}^T \right) - \frac{\sum_{k=1}^{t_C(i,j)} w_{i,j}(k)}{t_C(i,j)}.
\end{aligned} \tag{15}
$$

We then use the convergence bound on the Q-function from (8) along with the law of large numbers (LLN) to show that the first term involving the $\hat{Q}(i,j) - Q^*(i,j)$ is bound by ϵ_{t_A} and the third term involving $\hat{V}(x_{i,j}(k)$ and $\mathbf{p}_{i,j}\mathbf{v}^T$ is also bounded by ϵ_{t_A}.

$$
\left| \frac{\sum_{k=1}^{t_C(i,j)} \left(\hat{Q}(i,j) - Q^*(i,j) \right)}{t_C(i,j)} \right| \leq \varepsilon(t_A) \tag{16}
$$

$$
\gamma \left| \frac{\sum_{k=1}^{t_C(i,j)} \hat{V}(x_{i,j}(k))}{t_C(i,j)} - \mathbf{p}_{i,j} \mathbf{v}^T \right| \leq \gamma \varepsilon(t_A) \tag{17}
$$

Clearly the term with rewards is trivially 0 and using the LLN we show that the term involving the reward noise asymptotically tends to 0.

Therefore, by finally using triangular inequalities we can prove that,

$$
\begin{gathered}
|\bar{d}_{i,j}| \leq (1+\gamma)\epsilon(t_A), \text{w.h.p as} \\
t_C(i,j) \to \infty \ \ \forall (i,j) \in \mathcal{X} \times \mathcal{U}.
\end{gathered}
$$

■

Similarly we now prove a theorem that lower-bounds the largest BDA when the system is under attack.

Theorem 7 *Given the system is under attack, the largest BDA can be lower bounded as follows,*

$$
\begin{gathered}
\max_{i,j} |\bar{d}_{i,j}| \geq \Phi(\mathbf{v}) \cdot \Delta(\mathbf{P}, \hat{\mathbf{P}}) - (1+\gamma)\varepsilon(t_A), \\
\textit{w.h.p as } t_C(i,j) \to \infty.
\end{gathered} \tag{18}
$$

Proof Sketch In a manner similar to the proof of Theorem 6 we subtract equation (6) from (11) but also introduce additional terms as,

$$\begin{aligned}\bar{d}_{i,j} =& \frac{\sum_{k=1}^{t_C(i,j)} \hat{Q}(i,j) - r_{i,j} - w_{i,j}(k) - \gamma \hat{V}(x_{i,j}(k))}{t_C(i,j)} \\ & - Q^*(i,j) + r_{i,j} + \gamma \hat{\mathbf{p}}_{i,j}\mathbf{v}^T + \gamma \tilde{\mathbf{p}}_{i,j}\mathbf{v}^T \\ =& \frac{\sum_{k=1}^{t_C(i,j)} \left(\hat{Q}(i,j) - Q^*(i,j)\right)}{t_C(i,j)} - \left(\frac{\sum_{k=1}^{t_C(i,j)} r_{i,j}}{t_C(i,j)} - r_{i,j}\right) \\ & - \gamma \left(\frac{\sum_{k=1}^{t_C(i,j)} \hat{V}(x_{i,j}(k))}{t_C(i,j)} - \hat{\mathbf{p}}_{i,j}\mathbf{v}^T\right) - \frac{\sum_{k=1}^{t_C(i,j)} w_{i,j}(k)}{t_C(i,j)} \\ & + \gamma \tilde{\mathbf{p}}_{i,j}\mathbf{v}^T.\end{aligned} \tag{19}$$

And similar to the proof of the Theorem 6 we show using arguments involving the LLN and the convergence bound on $\hat{Q}(\cdot)$ in (8) that the first term is bounded as in (16) , while

$$\gamma \left| \frac{\sum_{k=1}^{t_C(i,j)} \hat{V}(x_{i,j}(k))}{t_C(i,j)} - \hat{\mathbf{p}}_{i,j}\mathbf{v}^T \right| \leq \gamma \varepsilon(t_A) \tag{20}$$

since the trajectory of the spoofed system being controlled has parameters $\hat{\mathbf{P}}$. And unlike the previous the case the new term $\gamma \tilde{\mathbf{p}}_{i,j}\mathbf{v}^T$ can be lower bounded using the Cauchy- Schwartz inequality and other further analysis as,

$$\max_{i,j} |\gamma \tilde{\mathbf{p}}_{i,j}\mathbf{v}^T| \geq \Phi(\mathbf{v}) \cdot \Delta(\mathbf{P}, \hat{\mathbf{P}}). \tag{21}$$

The term involving the reward is trivially 0 and the reward noise term tends to 0 due too the LLN. Therefore by finally using triangular inequalities we can prove that,

$$\begin{gathered}\max_{i,j} |\bar{d}_{i,j}| \geq \Phi(\mathbf{v}) \cdot \Delta(\mathbf{P}, \hat{\mathbf{P}}) - (1+\gamma)\varepsilon(t_A), \\ \text{w.h.p as } t_C(i,j) \to \infty.\end{gathered}$$

■

With an upperbound on the deviation proven, we finally prove the correctness of Algorithm 1 when the informational advantage condition is met.

Theorem 8 *The informational advantage condition,*

$$\delta \cdot \phi > 2 \cdot (1+\gamma)\epsilon, \tag{22}$$

is necessary and sufficient for Algorithm 1 to guarantee AD while avoiding false alarms with high probability as $t_C \to \infty$. Here δ and ϕ are under-estimates of the adversary minimum model error and maximum system system discernibility as, $\delta \leq \Delta(\mathbf{P}, \hat{\mathbf{P}})$ and $\phi \leq \Phi(\mathbf{v})$, and ϵ is an over-estimate of agent error in Q-function as $\epsilon \geq \varepsilon(t_A)$

Proof

Due to Theorem 6,

$$|\bar{d}_{i,j}| \leq (1+\gamma)\varepsilon(t_A) \leq (1+\gamma)\epsilon \tag{23}$$

with high probability as $t_C \to \infty$, since $\epsilon \geq \varepsilon(t_A)$. Similarly by Theorem 7,

$$\max_{i,j} |\bar{d}_{i,j}| \geq \Phi(\mathbf{v}) \cdot \Delta(\mathbf{P}, \hat{\mathbf{P}}) - (1+\gamma)\varepsilon(t_A) \geq \phi \cdot \delta - (1+\gamma)\epsilon \tag{24}$$

with high probability as $t_C \to \infty$, since $\phi \leq \Phi(\mathbf{v})$ and $\delta \leq \Delta(\mathbf{P}, \hat{\mathbf{P}})$. Therefore, we can guarantee AD with no false alarms as $t_C \to \infty$ for Algorithm 1, if and only if

$$\phi \cdot \delta - (1+\gamma)\epsilon > (1+\gamma)\epsilon.$$

That is, when the lower bound on the largest BDA during attack exceeds the upper bound on all BDAs during no attack. This allows us to detect if an attack takes place when the lower bound is exceeded. We can now rewrite the above equation as

$$\phi \cdot \delta > 2 \cdot (1+\gamma)\epsilon.$$

Since asymptotic AD with no false alarms with high probability can be achieved by Algorithm 1 if and only if Equation (22) is true. This proves that Equation (22) is a necessary and sufficient condition.

■

Remark 9 (On Asynchronous Detection) *We note here that Theorem 8 proves the detection guarantees for when the start of the adversary's attack and the agent's detection algorithm are synchronized. However, it is easy to extend this proof to the case when the start of the attack and detection are offset by finite time (by using the Cesaro Mean theorem).*

5. Conclusion

In this paper we proposed a Bellman Deviation Detection algorithm that is a simple statistical test that can be used by an agent that performs a model-free reinforcement learning to guarantee attack detection in an asymptotic sense. We proved stochastic guarantees of the proposed algorithm which reveal how an informational advantage condition can be exploited by the agent to guarantee detection. Our Bellman Deviation Detection algorithm provides security guarantees against MITM attacks in the context of model-free RL, while also account for the imperfect knowledge of the system at both the agent and the adversary ends.

Acknowledgments

This work was partially supported by National Science Foundation award ENG-2127946.

References

Cheng-Zong Bai, Fabio Pasqualetti, and Vijay Gupta. Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs. *Automatica*, 82:251–260, 2017. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2017.04.047. URL https://www.sciencedirect.com/science/article/pii/S0005109817302418.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/766ebcd59621e305170616ba3d3dac32-Paper.pdf.

Philip N. Brown, Holly P. Borowski, and Jason R. Marden. Security against impersonation attacks in distributed systems. *IEEE Transactions on Control of Network Systems*, 6(1):440–450, 2019. doi: 10.1109/TCNS.2018.2838519.

Ahmet Cetinkaya, Hideaki Ishii, and Tomohisa Hayakawa. Networked control under random and malicious packet losses. *IEEE Transactions on Automatic Control*, 62(5):2434–2449, 2017. doi: 10.1109/TAC.2016.2612818.

Yuan Chen, Soummya Kar, and José M.F. Moura. Cyber physical attacks with control objectives and detection constraints. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1125–1130, 2016. doi: 10.1109/CDC.2016.7798418.

Michelle S. Chong, Henrik Sandberg, and André M.H. Teixeira. A tutorial introduction to security and privacy for cyber-physical systems. In *2019 18th European Control Conference (ECC)*, pages 968–978, 2019. doi: 10.23919/ECC.2019.8795652.

S. M. Dibaji, M. Pirani, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty. Secure control of wide-area power systems: Confidentiality and integrity threats. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 7269–7274, 2018. doi: 10.1109/CDC.2018.8618862.

Seyed Mehran Dibaji, Mohammad Pirani, David Bezalel Flamholz, Anuradha M. Annaswamy, Karl Henrik Johansson, and Aranya Chakrabortty. A systems and control perspective of cps security. *Annual Reviews in Control*, 47:394–411, 2019a. ISSN 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2019.04.011. URL https://www.sciencedirect.com/science/article/pii/S1367578819300185.

Seyed Mehran Dibaji, Mohammad Pirani, David Bezalel Flamholz, Anuradha M. Annaswamy, Karl Henrik Johansson, and Aranya Chakrabortty. A systems and control perspective of cps security. *Annual Reviews in Control*, 47:394–411, 2019b. ISSN 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2019.04.011. URL https://www.sciencedirect.com/science/article/pii/S1367578819300185.

Kemi Ding, Xiaoqiang Ren, Daniel E. Quevedo, Subhrakanti Dey, and Ling Shi. Dos attacks on remote state estimation with asymmetric information. *IEEE Transactions on Control of Network Systems*, 6(2):653–666, 2019. doi: 10.1109/TCNS.2018.2867157.

V. S. Dolk, P. Tesi, C. De Persis, and W. P. M. H. Heemels. Event-triggered control systems under denial-of-service attacks. *IEEE Transactions on Control of Network Systems*, 4(1):93–105, 2017. doi: 10.1109/TCNS.2016.2613445.

Chongrong Fang, Yifei Qi, Peng Cheng, and Wei Xing Zheng. Cost-effective watermark based detector for replay attacks on cyber-physical systems. In *2017 11th Asian Control Conference (ASCC)*, pages 940–945, 2017. doi: 10.1109/ASCC.2017.8287297.

Aidin Ferdowsi and Walid Saad. Deep learning for signal authentication and security in massive internet-of-things systems. *IEEE Transactions on Communications*, 67(2):1371–1387, 2019. doi: 10.1109/TCOMM.2018.2878025.

Jaime F. Fisac, Anayo K. Akametalu, Melanie N. Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J. Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2019a. doi: 10.1109/TAC.2018.2876389.

Jaime F. Fisac, Anayo K. Akametalu, Melanie N. Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J. Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2019b. doi: 10.1109/TAC.2018.2876389.

Paul Griffioen, Sean Weerakkody, and Bruno Sinopoli. An optimal design of a moving target defense for attack detection in control systems. In *2019 American Control Conference (ACC)*, pages 4527–4534, 2019. doi: 10.23919/ACC.2019.8814689.

Navid Hashemi, Carlos Murguia, and Justin Ruths. A comparison of stealthy sensor attacks on control systems. In *2018 Annual American Control Conference (ACC)*, pages 973–979, 2018. doi: 10.23919/ACC.2018.8431300.

Pedro Hespanhol, Matthew Porter, Ram Vasudevan, and Anil Aswani. Statistical watermarking for networked control systems. In *2018 Annual American Control Conference (ACC)*, pages 5467–5472, 2018. doi: 10.23919/ACC.2018.8431569.

Andreas Hoehn and Ping Zhang. Detection of covert attacks and zero dynamics attacks in cyber-physical systems. In *2016 American Control Conference (ACC)*, pages 302–307, 2016. doi: 10.1109/ACC.2016.7524932.

Maryam Hosseini, Takashi Tanaka, and Vijay Gupta. Designing optimal watermark signal for a stealthy attacker. In *2016 European Control Conference (ECC)*, pages 2258–2262, 2016. doi: 10.1109/ECC.2016.7810627.

Mahdi Jamei, Emma Stewart, Sean Peisert, Anna Scaglione, Chuck McParland, Ciaran Roberts, and Alex McEachern. Micro synchrophasor-based intrusion detection in automated distribution systems: Toward critical infrastructure security. *IEEE Internet Computing*, 20(5):18–27, 2016. doi: 10.1109/MIC.2016.102.

Aris Kanellopoulos and Kyriakos G. Vamvoudakis. A moving target defense control framework for cyber-physical systems. *IEEE Transactions on Automatic Control*, 65(3):1029–1043, 2020. doi: 10.1109/TAC.2019.2915746.

Ben Kehoe, Sachin Patil, Pieter Abbeel, and Ken Goldberg. A survey of research on cloud robotics and automation. *IEEE Transactions on Automation Science and Engineering*, 12(2):398–409, 2015. doi: 10.1109/TASE.2014.2376492.

Mohammad Javad Khojasteh, Anatoly Khina, Massimo Franceschetti, and Tara Javidi. Learning-based attacks in cyber-physical systems. *IEEE Transactions on Control of Network Systems*, 8 (1):437–449, 2021. doi: 10.1109/TCNS.2020.3028035.

Yee Wei Law, Tansu Alpcan, and Marimuthu Palaniswami. Security games for risk minimization in automatic generation control. *IEEE Transactions on Power Systems*, 30(1):223–232, 2015. doi: 10.1109/TPWRS.2014.2326403.

Hanxiao Liu, Jiaqi Yan, Yilin Mo, and Karl Henrik Johansson. An on-line design of physical watermarks. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 440–445, 2018. doi: 10.1109/CDC.2018.8619632.

Y. Wan M. Xue, S. Roy and S. K. Das. Security and vulnerability of cyber-physical. In *Handbook on securing cyber-physical critical infrastructure*, page 5, 2012.

Fei Miao, Miroslav Pajic, and George J. Pappas. Stochastic game approach for replay attack detection. In *52nd IEEE Conference on Decision and Control*, pages 1854–1859, 2013. doi: 10.1109/CDC.2013.6760152.

Yilin Mo, Sean Weerakkody, and Bruno Sinopoli. Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems Magazine*, 35(1):93–109, 2015. doi: 10.1109/MCS.2014.2364724.

Luyao Niu, Jie Fu, and Andrew Clark. Optimal minimum violation control synthesis of cyber-physical systems under attacks. *IEEE Transactions on Automatic Control*, 66(3):995–1008, 2021. doi: 10.1109/TAC.2020.2989268.

Mohammad Pirani, Ehsan Nekouei, Henrik Sandberg, and Karl Henrik Johansson. A graph-theoretic equilibrium analysis of attacker-defender game on consensus dynamics under ¡inline-formula¿¡tex-math notation=”latex”¿$\mathcal{H}_2$¡/tex-math¿¡/inline-formula¿ performance metric. *IEEE Transactions on Network Science and Engineering*, 8(3):1991–2000, 2021. doi: 10.1109/TNSE.2020.3035964.

Tunga R., Carlos Murguia, and Justin Ruths. Tuning windowed chi-squared detectors for sensor attacks. In *2018 Annual American Control Conference (ACC)*, pages 1752–1757, 2018. doi: 10.23919/ACC.2018.8431073.

Anshuka Rangi, Mohammad Javad Khojasteh, and Massimo Franceschetti. Learning based attacks in cyber physical systems: Exploration, detection, and control cost trade-offs. In Ali Jadbabaie, John Lygeros, George J. Pappas, Pablo A.nbsp;Parrilo, Benjamin Recht, Claire J. Tomlin, and Melanie N. Zeilinger, editors, *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pages 879–892. PMLR, 07 – 08 June 2021. URL https://proceedings.mlr.press/v144/rangi21a.html.

Rishi Rani and Massimo Franceschetti. Supplementary: Detection of man-in-the-middle attacks in model-free reinforcement learning. 2022. URL https://drive.google.com/file/d/1tGPEATbG1pFG2sy3q6lF4s2FdndagH_6/view?usp=sharing.

Henrik Sandberg, Saurabh Amin, and Karl Henrik Johansson. Cyberphysical security in networked control systems: An introduction to the issue. *IEEE Control Systems Magazine*, 35(1):20–23, 2015. doi: 10.1109/MCS.2014.2364708.

Bharadwaj Satchidanandan and P. R. Kumar. Dynamic watermarking: Active defense of networked cyber–physical systems. *Proceedings of the IEEE*, 105(2):219–240, 2017. doi: 10.1109/JPROC.2016.2575064.

Dawei Shi, Ziyang Guo, Karl Henrik Johansson, and Ling Shi. Causality countermeasures for anomaly detection in cyber-physical systems. *IEEE Transactions on Automatic Control*, 63(2):386–401, 2018. doi: 10.1109/TAC.2017.2714646.

Yasser Shoukry, Michelle Chong, Masashi Wakaiki, Pierluigi Nuzzo, Alberto L. Sangiovanni-Vincentelli, Sanjit A. Seshia, Joao P. Hespanha, and Paulo Tabuada. Smt-based observer design for cyber-physical systems under sensor attacks. In *2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS)*, pages 1–10, 2016. doi: 10.1109/ICCPS.2016.7479119.

Roy S. Smith. A decoupled feedback structure for covertly appropriating networked control systems. *IFAC Proceedings Volumes*, 44(1):90–95, 2011. ISSN 1474-6670. doi: https://doi.org/10.3182/20110828-6-IT-1002.01721. URL https://www.sciencedirect.com/science/article/pii/S1474667016435925. 18th IFAC World Congress.

André Teixeira, Iman Shames, Henrik Sandberg, and Karl Henrik Johansson. A secure control framework for resource-limited adversaries. *Automatica*, 51:135–148, 2015. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2014.10.067. URL https://www.sciencedirect.com/science/article/pii/S0005109814004488.

Ivana Tomić, Michael J. Breza, Greg Jackson, Laksh Bhatia, and Julie A. McCann. Design and evaluation of jamming resilient cyber-physical systems. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 687–694, 2018. doi: 10.1109/Cybermatics_2018.2018.00138.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5005–5014. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/tu18a.html.

David I. Urbina, Jairo A. Giraldo, Alvaro A. Cardenas, Nils Ole Tippenhauer, Junia Valente, Mustafa Faisal, Justin Ruths, Richard Candell, and Henrik Sandberg. Limiting the impact of stealthy attacks on industrial control systems. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 1092–1105, New

York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978388. URL https://doi.org/10.1145/2976749.2978388.

Sean Weerakkody and Bruno Sinopoli. Detecting integrity attacks on control systems using a moving target approach. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5820–5826, 2015. doi: 10.1109/CDC.2015.7403134.

Ye Yuan and Yilin Mo. Security in cyber-physical systems: Controller design against known-plaintext attack. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5814–5819, 2015. doi: 10.1109/CDC.2015.7403133.

Zhenyong Zhang, Ruilong Deng, David K. Y. Yau, Peng Cheng, and Jiming Chen. Analysis of moving target defense against false data injection attacks on power grid. *IEEE Transactions on Information Forensics and Security*, 15:2320–2335, 2020. doi: 10.1109/TIFS.2019.2928624.

Minghui Zhu and Sonia Martínez. On the performance analysis of resilient networked control systems under replay attacks. *IEEE Transactions on Automatic Control*, 59(3):804–808, 2014. doi: 10.1109/TAC.2013.2279896.