UNIVERSITY OF CALIFORNIA SAN DIEGO

Secure Reinforcement Learning And The Detection Of Man-In-The-Middle Attacks

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Electrical Engineering (Communication Theory and Systems)

by

Rishi Rani

Committee in charge:

Professor Massimo Franceschetti, Chair
Professor Tara Javidi
Professor Piya Pal
Professor Jorge Poveda

2023

The thesis of Rishi Rani is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents, family and friends who supported me.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Massimo Franceschetti for his support as the chair of my committee. Through multiple drafts and many long nights, his guidance has proved to be invaluable.

Chapter 1, in full, is a reprint of the material as it has been submitted for publication as it will appear in Proceedings of The 4th Annual Learning for Dynamics and Control Conference, 2023, Rani, Rishi; Franceschetti, Massimo, PMLR, 2009. The thesis author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint as it has been submitted for review of the material as it may appear in 62nd IEEE Conference on Decision and Control, 2023. Rani, Rishi; Franceschetti, Massimo.

ABSTRACT OF THE THESIS

Secure Reinforcement Learning And The Detection Of Man-In-The-Middle Attacks

by

Rishi Rani

Master of Science in Electrical Engineering (Communication Theory and Systems)

University of California San Diego, 2023

Professor Massimo Franceschetti, Chair

In this thesis, we study the the detection of man-in-the-middle (MITM) attacks in model-free reinforcement learning. We consider the problem of a learning-based, where the system may be subject to an adversarial attack that hijacks the feedback signal and the controller actions. The adversary first learns the dynamics of the system in a learning phase before hijacking the system in a attack phase. We then propose simple attack detection algorithms to detect such MITM attacks without for two different system models. Firstly, when the system can be modelled as a Markov decision process. Secondly, when it can modelled as a discrete linear time invariant (LTI) system with stochastic distrubances. We also show that a necessary and sufficient "informational advantage" condition must be met for both systems to guarantee the detection of attacks with

high probability, while also avoiding false alarms.

Chapter 1

Detection of Man-in-The-Middle Attacks for Markov Decision Process Systems

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 finding 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 re-

quired 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

$$\begin{aligned} V^*(x) &= \max_x Q^*(x, u), \\ \mathbf{v} &= [V^*(x_1) \ldots, V^*(x_N)], \end{aligned} \tag{5}$$

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.

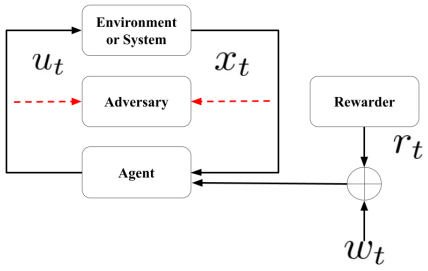
3. Problem Setup

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

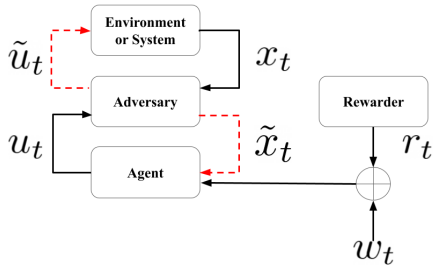
$$\boldsymbol{\tau}_A = (x_1^A, u_1^A, \ldots x_{t_A}^A, u_{t_A}^A), \tag{7}$$

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

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



(*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

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 $t_C \to \infty$.

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 $d_{i,j}$, 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,

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

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,

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

■

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,*

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

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,

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

$$\text{w.h.p as } t_C(i,j) \to \infty.$$

■

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.

Chapter 1, in full, is a reprint of the material as it has been submitted for publication as it will appear in Proceedings of The 4th Annual Learning for Dynamics and Control Conference, 2023, Rani, Rishi; Franceschetti, Massimo, PMLR, 2009. The thesis author was the primary investigator and author of this paper.

Chapter 2

Detection of Man-in-The-Middle Attacks for Discrete Linear Time Invariant Systems

Detection of Man in the Middle Attacks in Model-Free Reinforcement Learning for the Linear Quadratic Regulator

Rishi Rani and Massimo Franceschetti

Abstract

We consider the problem of a learning-based, man-in-the-middle (MITM) attack in a cyber-physical system. We use a simple abstraction where an agent performs linear quadratic regulation (LQR) of a discrete-time, linear, time-invariant (LTI) system with stochastic disturbances, using model-free reinforcement learning. The system may be subject to an adversarial attack that overrides the feedback signal and the controller actions. We propose a "Bellman Deviation" algorithm that can be used by the agent to detect the attack. This algorithm only requires an estimate of the Q-function, and optimal average stage cost, and no explicit information of the system parameters. We show that the proposed algorithm asymptotically guarantees attack detection (AD) with high probability while avoiding false alarms, when an "informational advantage" condition is met. This condition compares the amount of information the agent has aquired about the system with the one aquired by the adversary.

This work was supported by the National Science Foundation

R. Rani is a graduate student with the Department of Electrical and Computer Engineering, University of California, San Diego smr@ucsd.edu

M. Franceschetti is with Faculty of the Department of Electrical and Computer Engineering, University of California, San Diego massimo@ece.ucsd.edu

Keywords: Cyber-physical systems, linear dynamical systems, secure control, system identification, manin-the-middle attack, physical-layer authentication, linear quadratic regulator.

I. INTRODUCTION

The use of networked control in cyber-physical systems (CPS) coupled with advancements in computation and wireless technology has led to several innovative applications in cloud robotics and automation [1]. In all of these contexts, security considerations have become critical. Attacks on CPS can have severe and varied consequences such as hijacking autonomous vehicles and drones, hijacking life-critical infrastructure as an act of terror and attacks on financial systems [2]–[4]. Examples of security breaches that have been made public are the revenge sewage attack in Maroochy Shire, Australia; the German steel mill cyber-attack; the Ukraine power grid cyber-attack; the Davis-Besse nuclear power plant attack in Ohio, USA; and the Iranian uranium-enrichment facility attack via the Stuxnet malware [5]. These attacks have inspired several studies of security from a control-theoretic perspective [6]–[22].

One well studied paradigm is the man-in-the-middle (MITM) attack [23]–[26], where the adversary overrides the feedback channel with spoofed signals that indicate stable and safe operation of the system while the adversary simultaneously steers the system to a catastrophic trajectory. To detect the attack and ensure safety the legitimate agent must monitor the feedback observations with the intent to find statistically anomalous behaviour in an online fashion. Conversely, the attacker's objective is to produce spoofed feedback signals that are statistically indistinguishable from the true system behavior to avoid being detected by the agent. Some of the techniques developed to protect from MITM attacks include *watermarking, moving target and baiting* [27]–[38].

In recent years, there has been much activity in the field of learning based control [39]–[43] and in the context of security this has lead to the study of *learning based attacks*.

The learning based attack is broken into two phases- in the learning phase, the adversary initially has no knowledge of the system dynamics, but learns its dynamics by observing the system's trajectory. In the attack phase, it overrides the control signal to achieve catastrophic effects while attempting to remain undetected. The imperfect knowledge the adversary has of the system is one of the elements of the learning based attack that makes it prone to detection. Asymptotic upper and lower bounds on the probability of deception in scalar and vector LTI systems were obtained in [44]. A similar approach was used to study a control-cost trade-off that analysed the training time required by the adversary to deceive the legitimate agent and the energy required by the agent to detect the attack in vector LTI systems [45].

In our previous work, [46], we considered the problem of detecting MITM attacks for a model-free reinforcement learning based controller in a system whose dynamics are described by a finite Markov decision process (MDP). Our work in this paper also considers attack detection, but in the different context of a Linear Quadratic Regulator (LQR). In a finite MDP, the set of states and actions are finite and state dynamics are defined as Markov transitions. The expected rewards doled are defined for each state-action transition. It follows that finding the optimal control strategy for an MDP is equivalent to solving a linear program, as the objective of maximizing the expected accrued reward can be modeled as a linear objective function. On the other hand, the LQR has state and actions defined over real Euclidean spaces, with dynamics modeled as a discrete linear time-invariant (LTI) system and the cost is modeled as a non-negative quadratic function on the state and action vectors. In this case, the problem of finding the optimal control strategy is equivalent to solving a convex quadratic program. In both scenarios, we tackle the infinite time-horizon problem, however in the MDP case the accrued reward over the infinite time horizon is made analytically tractable and finite by introducing a discount factor $\gamma < 1$. The LQR problem sidesteps this tractability issue by subtracting the optimal average stage cost from each stage cost, making the objective

function finite. These key differences make the LQR problem require a novel analysis to derive an attack detection algorithm. The LQR problem can be viewed as a continuous MDP with quadratic costs and an infinite time horizon problem with a discount factor of $\gamma = 1$.

Our contributions are as follows: we extend the model of learning-based attacks to account for errors the agent has due to limited learning. Specifically, we assume the agent performs model-free reinforcement learning based control on the linear quadratic regulator (LQR) problem. We assume the agent constructs an estimate of the optimal Q-function and optimal average stage cost, and has no other information about the system. In this context, since the agent has no explicit model of the dynamics, attack detection (AD), which typically occurs through the observation of statistically anomalous behavior, becomes particularly challenging. In our case, detection is performed by estimating the temporal difference (TD) error through an approximate Bellman equation and averaging it over time. This method is motivated by the fact that the Q-function estimate has some implicit information on the system dynamics. We propose an AD algorithm, named the "Bellman Deviation" algorithm that asymptotically guarantees AD with high probability while also avoiding false alarms, when an "informational advantage" condition is met. The informational advantage condition determines whether the adversary can avoid detection by relating the error in the agent's Q-function to the adversary's error in the model parameters. The analysis also provides a functional understanding of the nature of the problem in terms of the information pattern required for successful detection.

Notation: in what follows we use the variable x to denote states, u to denote actions and y to denote the stacked state-action vectors. Vectors are column vectors by default. Matrices are capitalized, $\|\cdot\|_2$ refers to the vector euclidean norm or the induced matrix spectral norm and $\|\cdot\|_F$ is the Frobenius matrix norm. $A \succeq 0$ means a square matrix A is positive semi-definite (PSD). Finally, we say that an event occurs with high probability (w.h.p.) if its probability p_n limits to 1 as the index n tends to infinity.

II. System Model & Mathematical Preliminaries

We consider a discrete LTI system with the state dynamics modeled as

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the state vector at time t, $u_t \in \mathbb{R}^m$ is the action taken at time t and w_t is a 0 mean i.i.d noise of finite covariance Σ. We rewrite the above equation as follows,

$$\begin{aligned} x_{t+1} &= \underbrace{\begin{bmatrix} A & B \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} x_t \\ u_t \end{bmatrix}}_{y_t} + w_t \\ &= Cy_t + w_t. \end{aligned} \tag{2}$$

The system is assumed to be fully observable and stabilizabl and for the sake of simplicity we take the feedback signal to be x_{t+1} itself.

The regulation stage cost is defined as

$$c(x_t, y_t) + n_t = x_t^T S x_t + u_t^T R u_t + n_t \tag{3}$$

where $S \succeq 0$ and $R \succeq 0$ (positive semi-definite) and n_t is a 0-mean cost noise that may have infinite variance. The optimization problem is to minimize the average expected stage cost over an infinite time horizon. This is done by finding the optimal stationary policy $\pi^*(x)$ as

$$\lambda = \min_{\pi} \lim_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} x_t^T S x_t + u_t^T R u_t + n_t \right]. \tag{4}$$

We define the optimal Q-function (associated with optimal policy) as

$$Q^*(x, u) = \mathbb{E}\left[\sum_{t=1}^{\infty} (x_t^T S x_t + u_t^T R u_t + n_t - \lambda) \right] \tag{5}$$

$$\text{given } x_1 = x, u_1 = u, \ s.t \ \ u_t = \pi^*(x_t),$$

where λ is the average accrued stage cost over the infinite trajectory. The optimal value function $V^*(x)$, defined as $V^*(x) = \arg\max_u Q^*(x,u)$, can be described as

$$V^*(x) = \mathbb{E}\left[\sum_{t=1}^{\infty}(x_t^T S x_t + u_t^T R u_t + n_t - \lambda)\Big| x_1 = x\right] \tag{6}$$
$$s.t \;\; u_t = \pi^*(x_t).$$

Due to the system being LTI and the cost noise/state disturbance being i.i.d, we find that the optimal policy function is a linear function of the state vector as $\pi^*(x_t) = Kx_t$. Similarly, since the regulation costs are quadratic functions of the state and actions vectors we find that the value function is a quadratic function of the state vector as $V^*(x_t) = x_t^T V x_t$ and the Q-function is quadratic as

$$\begin{aligned} Q^*(x_t,u_t) &= \begin{bmatrix} x_t^T u_t^T \end{bmatrix}\left(\begin{bmatrix} S & 0 \\ 0 & R \end{bmatrix} + \begin{bmatrix} A \; B \end{bmatrix}^T V \begin{bmatrix} A \; B \end{bmatrix}\right)\begin{bmatrix} x_t \\ u_t \end{bmatrix} \\ &= y^T Q y_t. \end{aligned} \tag{7}$$

The optimal average stage cost λ can also be described through the value function as

$$\lambda = \mathbb{E}\left[V^*(w_t)\right] = \mathbb{E}\left[w_t^T V w_t\right] = \mathrm{Tr}(V\Sigma), \tag{8}$$

and we finally rewrite (7) as a recursive relation as follows

$$Q(x_t,u_t) = c(x_t,u_t) - \lambda + \mathbb{E}\left[V^*(x_{t+1})\right]. \tag{9}$$

This is called the *Bellman equation*.

The solution for V can be computed by solving the discrete algebraic Riccati equation

$$V = A^T V A - A^T V B(R + B^T V B)^{-1} B^T V A + Q, \tag{10}$$

and K can computed from V as

$$K = -(R + B^T V B)^{-1} B^T V A. \tag{11}$$

Note, however, that in a model-free RL setting this is not possible as the agent does not know A and B and uses algorithms to estimate Q and V directly from the trajectory.

III. Problem Formulation

We assume an agent uses model-free RL to control the system with the goal of performing LQR. The agent learns an estimate of the optimal Q-function from an observed trajectory τ_A

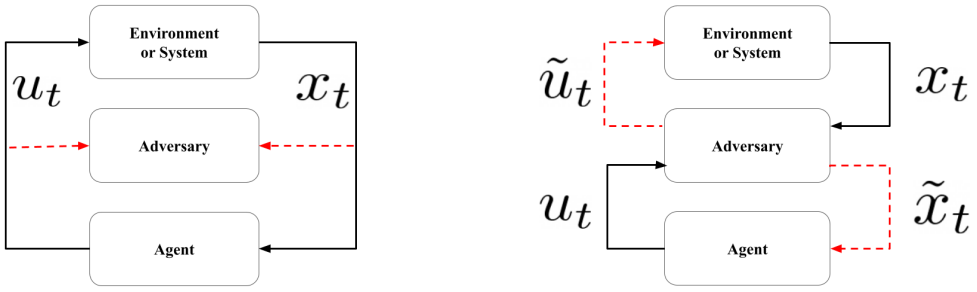
$$\tau_A = (x_1^A, u_1^A, c_1^A \dots x_{t_A}^A, u_{t_A}^A, c_{t_A}^A), \tag{12}$$

where x_t^A is the state vector at time t, y_t^A is the control vector, c_t^A is the stage cost and t_A is the agent training time. No assumption is made on whether the agent has control over τ_A or it merely observes the trajectory. The estimate of the optimal value function can be computed from the Q-function as $\hat{V}(x) = \max_u Q(x,u)$. The agent also has no information about the system model or cost function and uses an arbitrary learning algorithm to estimate the optimal Q-function and average stage cost with the following stochastic guarantee

$$\begin{gathered} \|\hat{Q}_{t_A} - Q\|_2 \leq \varepsilon_1(t_A), \|\hat{V}_{t_A} - V\|_2 \leq \varepsilon_1(t_A) \\ \text{and } |\hat{\lambda} - \lambda| \leq \varepsilon_2(t_A), \text{ w.h.p} \\ \text{s.t } \varepsilon_1(t_A) \to 0, \ \varepsilon_2(t_A) \to 0 \text{ as } t_A \to \infty. \end{gathered} \tag{13}$$

The only other conditions we place on $\hat{V}$ and $\hat{Q}$ are that $\hat{V} \succeq 0$ and $\hat{Q} \succeq 0$ as it is trivial to show that V and Q are PSD.

As described in Fig.1a, during the learning phase, the adversary observes the system trajectory $\boldsymbol{\tau}_B$ and it learns the system, obtaining an estimate of the dynamics model as $(\hat{A}, \hat{B})$. During this phase, the adversary has no control over its learning trajectory $\boldsymbol{\tau}_B$, as



(a) Learn Phase

(b) Attack Phase

Figure 1: a) In the learning phase, the attacker eavesdrops and learns the system, without highjacking the system. b) In the attack phase, the adversary intervenes as a MITM in two places: acting as a fake system to the agent and controlling the system.

it simply observes and does not control the system. Therefore, no asymptotic convergence guarantees are placed on its estimates $(\hat{A}, \hat{B})$. In the attack phase (as described in Fig.1b), the adversary takes control of the system and feeds the agent fake state feedback vectors. These fake signals are generated to be statistically consistent with its dynamics model estimate $(\hat{A}, \hat{B})$. We assume the adversary has full information of Σ so as to introduce the appropriate state disturbances in its fictitious feedback signals. The adversary does not need to have an explicit model estimate $(\hat{A}, \hat{B})$, but as long as the adversary's fictitious state feedback signals $\tilde{x}_t$ and the agent's actions u_t form linear dynamics, there will exist an implicit linear model that describes these dynamics.

The trajectory formed during the attack phase, $\boldsymbol{\tau}_C$, is now used by the agent to detect attacks. The adversary during attack 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 the detection problem, i.e, the adversary's closed feedback with the 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 an estimate of the optimal average stage cost $\hat{\lambda}$, and the adversary spoofs the system with a model estimate $(\hat{A}, \hat{B})$, devise a detection algorithm that uses the trajectory during the attack $\boldsymbol{\tau}_C$ to provide guarantees on AD as the trajectory length $t_C \to \infty$.

IV. Algoritm Description and Proof

In this section we describe the detection algorithm and then prove its stochastic guarantees.

A. Algorithm Description

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

$$\tau_C = (x_1^C, u_1^C, c_1^C \dots x_{t_C}^C, u_{t_C}^C, c_{t_C}^C). \tag{14}$$

where t_C is the length of the trajectory observed and the sequence x_i^C, u_i^C and c_i^C are the respective state, action and cost sequence that form the trajectory. For the rest of the section the superscript C notation is dropped to improve readability and referring to any of the above sequences implies it is from τ_C unless stated otherwise.

Definition IV.1 (Bellman Deviation Sequence)**.** Let d_i be the Bellman deviation sequence

$$d_i = \hat{Q}(x_i, u_i) + \hat{\lambda} - c_i - \hat{V}(x_{i+1}). \tag{15}$$

This sequence represents the deviations from Bellman-like behaviour in the observed trajectory during the attack phase.

The Bellman deviation sequence (BDS) is simply the temporal difference (TD) errors defined over an *approximate* Bellman equation, $\hat{Q}(x_i, u_i) \approx c_i + n_i - \hat{\lambda} - \hat{V}(x_{i+1})$.

Definition IV.2 (Bellman Deviation Average)**.** Let $\bar{d}$ be the Bellman deviation average

$$\bar{d} = \frac{\sum_{i=1}^{t_C} d_i}{\sum_{i=1}^{t_C} \|y_i\|_2^2}. \tag{16}$$

This average helps us eliminate the variability we find due to the disturbance in the state transitions.

The Bellman deviation average (BDA) is a simple average of the BDS normalized by the energy of the combined state-action vectors y_i. Since a system not under attack should display approximate Bellman behaviour a large BDA would suggest that the system is under attack. We use bounds on the BDA to determine if the system is under a MITM 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 IV.3 (System Discernibility Semi-Metric)**.** Given an LQR problem with state transition matrices (A, B) and cost matrices (S, R), we define the system discernibility semi-metric $\Phi_{\mathbf{V}} : \mathbb{R}^{n\times(n+m)} \times \mathbb{R}^{n\times(n+m)} \to \mathbb{R}_+$as

$$\Phi_{\mathbf{V}}(M_1, M_2) = \sigma_{min}(M_1^T V M_1 - M_2^T V M_2), \tag{17}$$

where V is the optimal value function of the system expressed as a PSD matrix and the function $\sigma_{min}(\cdot)$ returns the smallest singular value of the matrix.

The above function is a semi-metric as it has most properties of a metric over the space $\mathbb{R}^{n\times(n+m)}$, though $\Phi_{\mathbf{V}}(M_1, M_2) = 0 \implies M_1 = M_2$ and the triangular inequality does not hold for this semi-metric. Finally, we define a quantity to measure the minimum adversary discernibility.

Definition IV.4 (Minimum Adversary Discernibility)**.** Given the system state transition model is $C = [A\ B]$ and the adversary estimate is $\hat{C} = [\hat{A}\ \hat{B}]$ we define the minimum adversary discernibility as

$$\Phi_{\mathbf{V}}(C, \hat{C}) = \sigma_{min}(C^T V C - \hat{C}^T V \hat{C}), \tag{18}$$

where the function $\sigma_{min}(\cdot)$ returns the smallest singular value of the matrix.

The minimum adversary discernibility gives us a measure of the minimum deviation in the value function for the subsequent stage due to the error in the adversary's estimate of the state transition model C. The definition can be understood intuitively as a measure that tells us how easy it is to observe deviation in the value function during the attack phase. For example, in a system with $\Phi_{\mathbf{V}} = \phi$, we can lower bound the difference in the subsequent state's value for the current state-action vector y as

$$\begin{aligned} (Cy)^T V(Cy) - (\hat{C}y)^T V(\hat{C}y) &= y^T(C^T V C - \hat{C}^T V \hat{C})y \\ |y^T(C^T V C - \hat{C}^T V \hat{C})y| &\geq \phi \|y\|_2^2, \end{aligned} \tag{19}$$

where Cy, $\hat{C}y$ are the next states when there an attack does not underway and is underway respectively. We point out that this example ignores stochastic disturbances in state transitions but these are accounted for in the proof of theorem IV.3. A minimum adversary discernibility of $\Phi_{\mathbf{V}}(C, \hat{C}) = 0$ implies that the value function give us no information about a non-trivial lower bound on the change in values. A non-zero adversary discernibility is required to guarantee a *discernible change* in value of the trajectory that can guarantee AD. It follows that the system discernibility semi-metric is a key property of the system and should be kept in mind while designing secure systems.

With the above quantities defined, we can now present the Bellman deviation detection algorithm (see Algorithm 1) and prove its correctness. Algorithm 1 constructs the BDS and averages this sequence to calculate the BDA. The BDA is then compared with a certain bound ($\bar{d} \geq \phi - (\epsilon_1 + \frac{t_c}{e}\epsilon_2)$), and if the BDA crosses this bound an attack is declared. Note that this lower bound for AD is a function of the error in the agent's estimate of the optimal Q-function and optimal average stage cost but the algorithm only needs an over-estimate of these errors. The lower bound is similarly a function of the minimum adversary discernibility but the agent only requires an under-estimate of this

quantity. This allows for more practical scenarios where exact values of these quantities would be unavailable. The algorithm guarantees AD and no false alarms, w.h.p, if and only if the informational advantage condition is met ($\phi > 2 \cdot (\epsilon_1 + \frac{t_c}{e}\epsilon_2)$). This condition relates an error metric on the adversary's model to an error metric on the agent's estimate of the optimal Q-function and λ. In essence, we condition on the agent having more information on the system than the adversary in order to detect a MITM attack.

Algorithm 1 Bellman Deviaion Detection

Require: $t_C \geq 0$, length(τ_C)= t_C

$\epsilon_1 \geq \varepsilon_1(t_A)$

$\epsilon_2 \geq \varepsilon_2(t_A)$

$\phi \leq \Phi_{\mathbf{V}}(C, \hat{C})$

initialize $d \leftarrow 0$

initialize $e \leftarrow 0$

initialize $n \leftarrow 1$

while $n < t_C$ **do**

$x_n \leftarrow \tau_C[n][0]$

$u_n \leftarrow \tau_C[n][1]$

$c_n \leftarrow \tau_C[n][2]$

$x_{n+1} \leftarrow \tau_C[n+1][0]$

$y_n \leftarrow [x_n^T, y_n^T]^T$

$d \leftarrow y_n^T \hat{Q} y_n + \hat{\lambda} - c_n - x_{n+1}^T \hat{V} x_{n+1} + d$

$e \leftarrow \|y_n\|_2^2 + e$

$n \leftarrow n + 1$

end while

$e \leftarrow \|y_{t_C}\|_2^2 + e$

$d \leftarrow \frac{d}{e}$

$\xi = \phi - (\epsilon_1 + \frac{t_c}{e}\epsilon_2)$

if $\phi > 2 \cdot (\epsilon_1 + \frac{t_c}{e}\epsilon_2)$ **then** ▷ Information Advantage Condition

if $d \geq \xi$ **then**

declare breach

else

declare no breach

end if

end if

B. Correctness of the Algorithm

In this section we prove the stochastic guarantees of the proposed algorithm. The general approach we use for the proof is to derive asymptotic upper and lower bounds on the BDA based on attack conditions. We derive these bounds using stochastic analysis and rely on stochastic linear systems theory, the strong law of large numbers (SLLN) and Kolmogorov's strong law of large numbers (KSLLN), which for completeness we now state as follows.

Theorem IV.1 (Kolmogorov's Strong Law of Large Numbers)**.** *Suppose $X_1, X_2 \dots$ are independent random variables such that $\mathbb{E}[X_n] = 0$ and*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} Var(X_i)}{n^2} < \infty$$

then the sequence mean asymptotically converges as

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{a.s} 0.$$

Proof. Refer to Theorem 2.3.10 in [47]. ■

Complete proofs of the Theorem IV.3 and IV.2 can be found in the supplementary material [48]. We now start by proving an asymptotic upper bound on the BDA when no attack is underway.

Theorem IV.2. *When no attack occurs, the Bellman deviation average can be lower bounded as*

$$|\bar{d}| \leq \varepsilon_1(t_A) + \frac{t_c\ \varepsilon_2(t_A)}{\sum_{i=1}^{t_C} \|y\|_2^2}, \textit{w.h.p as} \tag{20}$$

$$t_C \to \infty,$$

where $\varepsilon_1(t_A)$ is the spectral error norm bound on $\hat{Q}$ and $\hat{V}$, and $\varepsilon_2(t_A)$ is the error bound on the estimate $\hat{\lambda}$.

Proof Sketch: We rearrange the terms of the Bellman equation (9) and subtract it from (16) to get,

$$\begin{aligned}\bar{d} =& \frac{\sum_{i=1}^{t_C} y_i^T \hat{Q} y_i + \hat{\lambda} - c_i - n_t - x_{i+1}^T \hat{V} x_{i+1}}{\sum_{i=1}^{t_C} \|y\|_2^2} \\ & \frac{\sum_{i=1}^{t_C} -y_i^T Q y_i - \lambda + c_i + (Cy_i)^T V (Cy_i) + \lambda}{\sum_{i=1}^{t_C} \|y\|_2^2} \\ =& \frac{\sum_{i=1}^{t_C} y_i^T \tilde{Q} y_i + \tilde{\lambda} - x_{i+1}^T \hat{V} x_{i+1} - n_t}{\sum_{i=1}^{t_C} \|y\|_2^2} \\ & \frac{+(x_{i+1} - w_{i+1})^T V (x_{i+1} - w_{i+1}) + \lambda}{\sum_{i=1}^{t_C} \|y\|_2^2} \\ =& \frac{\sum_{i=1}^{t_C} y_i^T \tilde{Q} y_i + \tilde{\lambda} - x_{i+1}^T \tilde{V} x_{i+1} - n_t}{\sum_{i=1}^{t_C} \|y\|_2^2} \\ & \frac{+\lambda - w_{i+1}^T V w_{i+1} + w_{i+1}^T V C y_i + (Cy_i)^T V w_{i+1}}{\sum_{i=1}^{t_C} \|y\|_2^2},\end{aligned} \tag{21}$$

where the super script tilde denotes the error of the estimates. We then use the convergence bound on the $\hat{Q}$, $\hat{V}$ and $\hat{\lambda}$ from (13) to show that the first terms involving $\tilde{Q}$, $\tilde{V}$ and $\tilde{\lambda}$ are upper bound as

$$\left| \frac{\sum_{i=1}^{t_C} y_i^T \tilde{Q} y_i - x_{i+1}^T \tilde{V} x_{i+1}}{\sum_{i=1}^{t_C} \|y\|_2^2} \right| \leq \varepsilon_1(t_A), \tag{22}$$

$$\left| \frac{\sum_{i=1}^{t_C} \tilde{\lambda}}{\sum_{i=1}^{t_C} \|y\|_2^2} \right| \leq \frac{t_C \, \varepsilon_2(t_A)}{\sum_{i=1}^{t_C} \|y\|_2^2}. \tag{23}$$

We then use the strong law of large numbers to show that the following terms converge as follows,

$$\frac{\sum_{i=1}^{t_C} w_{i+1}^T V w_{i+1}}{\sum_{i=1}^{t_C} \|y\|_2^2} \xrightarrow{\text{a.s}} \frac{t_C \lambda}{\sum_{i=1}^{t_C} \|y\|_2^2}, \tag{24}$$

$$\frac{\sum_{i=1}^{t_C} n_t}{\sum_{i=1}^{t_C} \|y\|_2^2} \xrightarrow{\text{a.s}} 0, \tag{25}$$

as $t_C \to \infty$. We then apply the Kolmogorov's strong law of large numbers (Theorem IV.1) as

$$\frac{\sum_{i=1}^{t_C} w_{i+1}^T V C y_i}{\sum_{i=1}^{t_C} \|y\|_2^2} = \frac{\sum_{i=1}^{t_C} C y_i^T V w_{i+1}}{\sum_{i=1}^{t_C} \|y\|_2^2} \xrightarrow{\text{a.s}} 0, \tag{26}$$

as $t_c \to \infty$.

Finally, using triangular inequalities we can prove (20).

■

With a upper bound on the BDA derived when no attack is under way, we now derive an asymptotic lower-bound on the BDA when the system is under attack. This allows us to detect attacks if the lower bound during attacks strictly exceeds the upper bound when no attack occurs. We claim the agent has an *informational advantage* over the adversary when this condition occurs.

Theorem IV.3. *Given the system is under attack, the Bellman deviation average can be lower bounded as follows,*

$$|\bar{d}| \geq \Phi_{\mathbf{V}}(C, \hat{C}) - \left(\varepsilon_1(t_A) + \frac{t_c \, \varepsilon_2(t_A)}{\sum_{i=1}^{t_C} \|y\|_2^2} \right), \tag{27}$$

$$\textit{w.h.p as } t_C \to \infty,$$

where $\varepsilon_1(t_A)$ is the spectral error norm bound on $\hat{Q}$ and $\hat{V}$, $\varepsilon_2(t_A)$ is the error bound on the estimate $\hat{\lambda}$ and $\Phi_{\mathbf{V}}(C, \hat{C})$ is the minimum adversary discernibility.

Proof Sketch: In a manner similar to the proof of Theorem IV.2 we subtract equation

(9) from (16) but also introduce additional terms as

$$\begin{aligned}
\bar{d} =& \frac{\sum_{i=1}^{t_C} y_i^T \hat{Q} y_i + \hat{\lambda} - c_i - x_{i+1}^T \hat{V} x_{i+1}}{\sum_{i=1}^{t_C} \|y\|_2^2} \\
& \frac{\sum_{i=1}^{t_C} -y_i^T Q y_i - \lambda + c_i + (Cy_i)^T V (Cy_i) + \lambda}{\sum_{i=1}^{t_C} \|y\|_2^2} \\
=& \frac{\sum_{i=1}^{t_C} y_i^T \tilde{Q} y_i + \tilde{\lambda} - x_{i+1}^T \hat{V} x_{i+1} + \lambda}{\sum_{i=1}^{t_C} \|y\|_2^2} \\
& \frac{+(x_{i+1} - w_{i+1})^T V (x_{i+1} - w_{i+1})}{\sum_{i=1}^{t_C} \|y\|_2^2} \\
& \frac{+y_i^T (C^T V C - \hat{C}^T V \hat{C}) y_i}{\sum_{i=1}^{t_C} \|y\|_2^2} \\
=& \frac{\sum_{i=1}^{t_C} y_i^T \tilde{Q} y_i + \tilde{\lambda} - x_{i+1}^T \tilde{V} x_{i+1}}{\sum_{i=1}^{t_C} \|y\|_2^2} \\
& \frac{+\lambda - w_{i+1}^T V w_{i+1} + w_{i+1}^T V C y_i + (Cy_i)^T V w_{i+1}}{\sum_{i=1}^{t_C} \|y\|_2^2} \\
& \frac{+y_i^T (C^T V C - \hat{C}^T V \hat{C}) y_i}{\sum_{i=1}^{t_C} \|y\|_2^2}.
\end{aligned} \tag{28}$$

Similar to the proof of the Theorem IV.2 we can show using arguments involving the convergence bounds in (13) that (22, 23) hold true, using the SLLN we show (24, 25) hold true and finally using the KSLLN that (26) holds true. Now, from Definition IV.4 we know that

$$y_i^T (C^T V C - \hat{C}^T V \hat{C}) y_i \geq \Phi_{\mathbf{V}}(C, \hat{C}) \|y_i\|_2^2.$$

Hence we can show that,

$$\frac{\sum_{i=1}^{t_C} y_i^T (\tilde{C}^T V \tilde{C}) y_i}{\sum_{i=1}^{t_C} \|y\|_2^2} \geq \Phi_{\mathbf{V}}(C, \hat{C}) \tag{29}$$

with the minimum adversary discernibility providing a lower bound for the term. With asymptotic bounds drawn on all terms, all but one term being upper bounded and one term being lower bounded, we can now use the triangular inequality to prove (27).

■

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

Theorem IV.4. *The informational advantage condition,*

$$\phi >= 2 \cdot \left(\epsilon_1 + \frac{t_c}{\sum_{i=1}^{t_C} \|y\|_2^2} \epsilon_2 \right), \tag{30}$$

is necessary and sufficient for Algorithm 1 to guarantee AD while avoiding false alarms with high probability as $t_C \to \infty$. Here ϕ is an under-estimate of the minimum adversary discernibility as $\phi \leq \Phi_{\mathbf{V}}(C, \hat{C})$, and ϵ_1 and ϵ_2 are over-estimates of agent error in Q-function and optimal average stage cost λ as $\epsilon_1 \geq \varepsilon_1(t_A)$ and $\epsilon_2 \geq \varepsilon_2(t_A)$.

Proof. Due to Theorem IV.2,

$$|\bar{d}| \leq \varepsilon_1 + \frac{t_c}{\sum_{i=1}^{t_C} \|y\|_2^2} \varepsilon_2 \tag{31}$$

with high probability as $t_C \to \infty$, since $\epsilon_1 \geq \varepsilon_1(t_A)$ and $\epsilon_2 \geq \varepsilon_2(t_A)$. Similarly by Theorem IV.3,

$$|\bar{d}| \geq \Phi_{\mathbf{V}}(C, \hat{C}) - \left(\varepsilon_1(t_A) + \frac{t_c \, \varepsilon_2(t_A)}{\sum_{i=1}^{t_C} \|y\|_2^2} \right) \tag{32}$$

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

$$\phi - \left(\epsilon_1 + \frac{t_c}{e} \epsilon_2 \right) > \epsilon_1 + \frac{t_c}{\sum_{i=1}^{t_C} \|y\|_2^2} \epsilon_2.$$

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

$$\phi > 2 \cdot \left(\epsilon_1 + \frac{t_c}{\sum_{i=1}^{t_C} \|y\|_2^2} \epsilon_2 \right).$$

Since asymptotic AD with no false alarms with high probability can be achieved by Algorithm 1 if and only if (30) is true. This proves that (30) is a necessary and sufficient condition for the algorithm's correctness. ■

V. CONCLUSIONS AND FUTURE WORKS

A. *Conclusions*

We discussed the problem of a learning based man-in-the-middle (MITM) attack on a CPS in a where the system is a discrete-time, LTI system with stochastic disturbances. This system is subject to an adversarial attack that overrides the system feedback and takes control of the system. The agent, on the other hand, performs model-free reinforcement learning and is constantly on a look-out for an attack; once the agent detects an attack, it declares a breach of the system. We propose a "Bellman Deviation" detection algorithm that can be used by an agent that performs linear quadratic regulation using model-free RL to detect a MITM attack on the system. This algorithm requires only the estimate of the Q-function and the optimal average stage cost, and needs no explicit information on the parameters of the system dynamics to detect an MITM attack. We proved the correctness of the algorithm and showed that the proposed algorithm asymptotically guarantees attack detection (AD) with high probability while avoiding any false alarms, provided that an intuitive informational advantage condition that relates the agent and adversary's learning costs is satisfied. The Bellman Deviation detection algorithm provides security guarantees against MITM attacks in the context of model-free RL for the LQR, while also accounting for the imperfect knowledge of the system at both the agent and the adversary ends.

B. *Future Work*

An open question is whether the proposed algorithm is the most sample efficient or even order efficient. Obtaning an information-theoretic lower bound on the sample

efficiency of such AD algorithm that takes imperfect information of the system into account on both agent and adversary ends would be of great interest.

Additionally, the proposed informational advantage condition relates the errors in the agent's model and the adversary's model using abstract quantities that can be intuitively interpreted. Extending this condition using more explicit quantities, like the training times of the agent and the adversary, would be practically useful as it would describe the cost related to securing the system using tangible quantities.

REFERENCES

[1] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, "A survey of research on cloud robotics and automation," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 2, pp. 398–409, 2015.

[2] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1092–1105.

[3] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of cps security," *Annual Reviews in Control*, vol. 47, pp. 394–411, 2019.

[4] M. Jamei, E. Stewart, S. Peisert, A. Scaglione, C. McParland, C. Roberts, and A. McEachern, "Micro synchrophasor-based intrusion detection in automated distribution systems: Toward critical infrastructure security," *IEEE Internet Computing*, vol. 20, no. 5, pp. 18–27, 2016.

[5] H. Sandberg, S. Amin, and K. H. Johansson, "Cyberphysical security in networked control systems: An introduction to the issue," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 20–23, 2015.

[6] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs," *Automatica*, vol. 82, pp. 251–260, 2017.

[7] 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*, vol. 4, no. 1, pp. 93–105, 2017.

[8] Y. Shoukry, M. Chong, M. Wakaiki, P. Nuzzo, A. L. Sangiovanni-Vincentelli, S. A. Seshia, J. P. Hespanha, and P. Tabuada, "Smt-based observer design for cyber-physical systems under sensor attacks," in *2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS)*, 2016, pp. 1–10.

[9] Y. Chen, S. Kar, and J. M. Moura, "Cyber physical attacks with control objectives and detection constraints," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, 2016, pp. 1125–1130.

[10] D. Shi, Z. Guo, K. H. Johansson, and L. Shi, "Causality countermeasures for anomaly detection in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 386–401, 2018.

[11] 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)*, 2018, pp. 7269–7274.

[12] T. R., C. Murguia, and J. Ruths, "Tuning windowed chi-squared detectors for sensor attacks," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 1752–1757.

[13] L. Niu, J. Fu, and A. Clark, "Minimum violation control synthesis on cyber-physical systems under attacks," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 262–269.

[14] M. S. Chong, H. Sandberg, and A. M. Teixeira, "A tutorial introduction to security and privacy for cyber-physical systems," in *2019 18th European Control Conference (ECC)*, 2019, pp. 968–978.

[15] I. Tomić, M. J. Breza, G. Jackson, L. Bhatia, and J. 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)*, 2018, pp. 687–694.

[16] K. Ding, X. Ren, D. E. Quevedo, S. Dey, and L. Shi, "Dos attacks on remote state estimation with asymmetric information," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 2, pp. 653–666, 2019.

[17] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005109814004488

[18] A. Cetinkaya, H. Ishii, and T. Hayakawa, "Networked control under random and malicious packet losses," *IEEE Transactions on Automatic Control*, vol. 62, no. 5, pp. 2434–2449, 2017.

[19] P. N. Brown, H. P. Borowski, and J. R. Marden, "Security against impersonation attacks in distributed systems," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 1, pp. 440–450, 2019.

[20] Y. W. Law, T. Alpcan, and M. Palaniswami, "Security games for risk minimization in automatic generation control," *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 223–232, 2015.

[21] I. Shames, F. Farokhi, and T. Summers, "Security analysis of cyber-physical systems using h2 norm," *IET Control Theory Applications*, vol. 11, 01 2017.

[22] N. Hashemi, C. Murguia, and J. Ruths, "A comparison of stealthy sensor attacks on control systems," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 973–979.

[23] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 90–95, 2011, 18th IFAC World Congress. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474667016435925

[24] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.

[25] M. Zhu and S. Martínez, "On the performance analysis of resilient networked control systems under replay attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 804–808, 2014.

[26] F. Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," in *52nd IEEE Conference on Decision and Control*, 2013, pp. 1854–1859.

[27] B. Satchidanandan and P. R. Kumar, "Dynamic watermarking: Active defense of networked cyber–physical systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, 2017.

[28] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *2015 54th IEEE Conference on Decision and Control (CDC)*, 2015, pp. 5820–5826.

[29] D. Flamholz, A. Annaswamy, and E. Lavretsky, "Baiting for defense against stealthy attacks on cyber-physical systems," 01 2019.

[30] A. Kanellopoulos and K. G. Vamvoudakis, "A moving target defense control framework for cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 1029–1043, 2020.

[31] Z. Zhang, R. Deng, D. K. Y. Yau, P. Cheng, and J. Chen, "Analysis of moving target defense against false data injection attacks on power grid," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2320–2335, 2020.

[32] P. Griffioen, S. Weerakkody, and B. Sinopoli, "An optimal design of a moving target defense for attack detection in control systems," in *2019 American Control Conference (ACC)*, 2019, pp. 4527–4534.

[33] A. Hoehn and P. Zhang, "Detection of covert attacks and zero dynamics attacks in cyber-physical systems," in *2016 American Control Conference (ACC)*, 2016, pp. 302–307.

[34] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, "Statistical watermarking for networked control systems," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 5467–5472.

[35] C. Fang, Y. Qi, P. Cheng, and W. X. Zheng, "Cost-effective watermark based detector for replay attacks on cyber-physical systems," in *2017 11th Asian Control Conference (ASCC)*, 2017, pp. 940–945.

[36] M. Hosseini, T. Tanaka, and V. Gupta, "Designing optimal watermark signal for a stealthy attacker," in *2016 European Control Conference (ECC)*, 2016, pp. 2258–2262.

[37] A. Ferdowsi and W. Saad, "Deep learning for signal authentication and security in massive internet-of-things systems," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1371–1387, 2019.

[38] H. Liu, J. Yan, Y. Mo, and K. H. Johansson, "An on-line design of physical watermarks," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 440–445.

[39] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737–2752, 2019.

[40] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[41] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737–2752, 2019.

[42] Y. Yuan and Y. Mo, "Security in cyber-physical systems: Controller design against known-plaintext attack," in *2015 54th IEEE Conference on Decision and Control (CDC)*, 2015, pp. 5814–5819.

[43] S. Tu and B. Recht, "Least-squares temporal difference learning for the linear quadratic regulator," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 5005–5014.

[44] M. J. Khojasteh, A. Khina, M. Franceschetti, and T. Javidi, "Learning-based attacks in cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 1, pp. 437–449, 2021.

[45] A. Rangi, M. J. Khojasteh, and M. Franceschetti, "Learning based attacks in cyber physical systems: Exploration, detection, and control cost trade-offs," in *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, ser. Proceedings of Machine Learning Research, vol. 144. PMLR, 07 – 08 June 2021, pp. 879–892.

[46] R. Rani and M. Franceschetti, "Detection of man-in-the-middle attacks in model-free reinforcement learning," 2023, to appear in 5th Annual Learning for Dynamics Control Conference. [Online]. Available: https://drive.google.com/file/d/1AvaVbfT5kJNHEAgJa5dqSko79PPLhCHH/view?usp=sharing

[47] P. K. Sen and J. M. Singer, *Large sample methods in statistics: an introduction with applications*. CRC press, 1994, vol. 25.

[48] R. Rani and M. Franceschetti, "Supplementary: Detection of man in the middle attacks in model-free reinforcement learning for the linear quadratic regulator," 2023. [Online]. Available: https://drive.google.com/file/d/1TdwzVIr5zgjoQgX3LUt-G_nmwbHBl8eF/view?usp=sharing

Chapter 2, in full, is a reprint as it has been submitted for review of the material as it may appear in 62nd IEEE Conference on Decision and Control, 2023. Rani, Rishi; Franceschetti, Massimo.