Optimal Attack against Cyber-Physical Control Systems with Reactive Attack Mitigation[*]

Subhash Lakshminarayana
Advanced Digital Sciences Center, Illinois at Singapore
Singapore 138682
subhash.l@adsc.com.sg

Teo Zhan Teng[†]
GovTech Singapore
Singapore 117438
teozt@hotmail.com

David K.Y. Yau
Singapore University of Technology and Design
Singapore 487372
david_yau@sutd.edu.sg

Rui Tan
Nanyang Technological University
Singapore 639798
tanrui@ntu.edu.sg

ABSTRACT

This paper studies the performance and resilience of a cyber-physical control system (CPCS) with attack detection and reactive attack mitigation. It addresses the problem of deriving an optimal sequence of false data injection attacks that maximizes the state estimation error of the system. The results provide basic understanding about the limit of the attack impact. The design of the optimal attack is based on a Markov decision process (MDP) formulation, which is solved efficiently using the value iteration method. Using the proposed framework, we quantify the effect of false positives and mis-detections on the system performance, which can help the joint design of the attack detection and mitigation. To demonstrate the use of the proposed framework in a real-world CPCS, we consider the voltage control system of power grids, and run extensive simulations using PowerWorld, a high-fidelity power system simulator, to validate our analysis. The results show that by carefully designing the attack sequence using our proposed approach, the attacker can cause a large deviation of the bus voltages from the desired setpoint. Further, the results verify the optimality of the derived attack sequence and show that, to cause maximum impact, the attacker must carefully craft his attack to strike a balance between the attack magnitude and stealthiness, due to the simultaneous presence of attack detection and mitigation.

KEYWORDS

Cyber-physical control system, Reactive attack mitigation, Resilience, Voltage control.

ACM Reference format:
Subhash Lakshminarayana, Teo Zhan Teng, David K.Y. Yau, and Rui Tan. 2017. Optimal Attack against Cyber-Physical Control Systems with Reactive Attack Mitigation. In *Proceedings of e-Energy '17, Shatin, Hong Kong, May 16-19, 2017,* 12 pages.
DOI: http://dx.doi.org/10.1145/3077839.3077852

1 INTRODUCTION

Critical infrastructures such as power grids and transportation systems are witnessing growing adoption of modern information and communication technologies (ICTs) for autonomous operation. While these advancements have improved their operational efficiency, ICTs may also make them vulnerable to cyber attacks. Vulnerabilities in ICT systems were exploited in recent high-profile cybersecurity incidents such as the BlackEnergy [3] and Dragonfly [1] attacks against power grids and the Stuxnet worm [17] against nuclear plants. These attacks injected false sensor data and/or control commands to the industrial control systems and resulted in widespread damage to the physical infrastructures and service outages. These incidents alert us to a general class of attacks called *false data injection* (FDI) against cyber-physical systems (CPS).

Attack detection and mitigation are two basic CPS security research problems, where the *attack detection* makes decisions in real time regarding the presence of an attack and *attack mitigation* isolates a detected attack and/or reduces its adverse impact on the system performance. CPSs often have various built-in anomaly detection methods that are effective in detecting simple fault-like FDI attacks, such as injecting surges, ramps, and random noises. However, critical CPSs (e.g., power grids) are the target of sophisticated attackers (such as hostile national organizations), whose attacks are often well-crafted using detailed knowledge of the system and its anomaly detection methods. To avoid detection, the attacker can inject a sequence of attacks of small magnitude and gradually mislead the system to a sub-optimal and even unsafe state. However, due to the stochastic nature of the physical and measurement processes of CPSs, as well as the adoption of stringent, advanced attack detectors, the well-crafted attacks can be detected probabilistically [26, 32]. Upon detecting an attack, mitigation should be activated to isolate the attack or maintain acceptable system performance in coexisting with the attack.

[*] This work was supported in part by the National Research Foundation (NRF), Prime Minister's Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-31) and administered by the National Cybersecurity R&D Directorate and in part by a Start-up Grant at NTU.

[†] The work was conducted when Teo Zhan Teng was with the Advanced Digital Sciences Center, Illinois at Singapore.

e-Energy '17, Shatin, Hong Kong
© 2017 ACM. 978-1-4503-5036-5/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3077839.3077852

Therefore, attack detection and mitigation are deeply coupled and they jointly define the system resilience against FDI attacks. On the one hand, a conservative detector may miss attacks, causing system performance degradation due to the mis-activation of attack mitigation. On the other hand, an aggressive detector may frequently raise false positives, triggering unnecessary mitigation actions in the absence of attacks, while attack mitigation generally needs to sacrifice the system performance to increase its robustness against attacks. Thus, it is important to understand the joint effect of attack detection and mitigation on the system performance, which serves as a basis for designing satisfactory detection-mitigation mechanisms. However, prior research on FDI attacks mostly study attack detection and mitigation separately [6, 20, 21, 26], and falls short of capturing their joint effect on the system. The studies on attack detection [20, 21, 26] generally ignore the attack mitigation triggered by probabilistic detection of attacks, and its impact on the future system states. On the other hand, the studies on attack mitigation [7, 23, 33] assume that the attack has been detected, and ignore the probabilistic nature of the attack detection and any adverse impact of mis-activation or false activation of mitigation due to misdetections and false alarms.

As an early (but important) effort in closing the gap, we jointly consider attack detection and mitigation in the system defense. In particular, we study their joint effect from an attacker's perspective and investigate the largest system performance degradation that a sophisticated attacker can cause in the presence of such a detection-mitigation defense mechanism. Studying this largest performance degradation helps us quantify the limit of attack impact, and serves as an important basis for designing/comparing detection and mitigation strategies to protect critical infrastructures. However, the attacker faces a fundamental dilemma in designing his attack – a large attack magnitude will result in high detection probability, thus nullifying the attack impact on the system (due to mitigation) whereas a small attack magnitude increases stealthiness but may do little damage. To achieve a significant impact, the attacker's injections must strike a balance between magnitude and stealthiness.

In this paper, we consider a general discrete-time linear time invariant (LTI) system with a feedback controller that computes its control decision based on the system state estimated by a Kalman filter (KF). For each time step, the controller uses a χ^2 attack detector [25], and activates mitigation actions upon detecting an attack. Following the Kerckhoffs's principle, we consider an attacker who accurately knows the system and its attack detection and mitigation methods. The attacker launches FDI attacks on the sensor measurements over an attack time horizon, aiming at misleading the controller into making erroneous control decisions. As the attack detection at each time step is probabilistic, we formulate the attacker's problem as a constrained stochastic optimization problem with an objective of maximizing the state estimation error over the attack time horizon, subject to a general constraint that the energy of the attack signal is upper-bounded. The solution to this problem naturally leads to an attack sequence that strikes a balance between attack magnitude and stealthiness to achieve the largest system performance degradation.

The main challenge in solving the aforementioned attacker's problem lies in the fact that the system state at any time depends on all the past attack detection results, due to reactive attack mitigation. Thus, the optimal attack at any time must exhaustively account for all possible sequences of past detection results, which is computationally complex. Moreover, the probabilistic attack detection introduces additional randomness into the system dynamics. Our key observation to overcome these issues is that the system dynamics is Markovian and the attacker's injections at any time can be computed based on knowledge about it, which captures the impact of all the past detection results. To summarize, the main contributions of our work are as follows:

- We solve the aforementioned attacker's problem using a Markov decision process (MDP) framework. In our formulation, the sequential operations of probabilistic attack detection and mitigation are mapped to the MDP's state transition probabilities. The MDP is solved by state space discretization and using the *value iteration* algorithm [30].
- To illustrate our analysis, we use a real-world CPCS – power grid voltage control – as our case study. The voltage controller adjusts the pilot bus voltages to predefined setpoints based on voltage measurements by applying feedback control on the generators' reactive power outputs. In the presence of attack mitigation, the attacker injects false measurements into the system, aiming at deviating the pilot bus voltages. Extensive simulations using PowerWorld, a high-fidelity power simulator, show that the optimal attack sequence computed using our proposed approach causes the maximum deviation of the pilot bus voltages from the desired setpoint.
- Based on the above framework, we also consider the problem of designing the detection threshold from the defender's perspective. To this end, we quantify the impact of false positives (FP) and misdetections (MD) via an extensive simulation-based study. Based on these costs, the attack detection threshold can be tuned to balance the performance downgrades due to FPs and MDs depending on the accuracy of the mitigation signal.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the system model. Section 4 gives the problem formulation. Section 5 describes the MDP-based solution methodology. Section 6 analyzes the impact of FPs and MDs on the system performance. Section 7 presents the simulation results. Section 8 concludes.

2 RELATED WORK

As mentioned earlier, most of the existing studies treat attack detection and mitigation problems separately. In the category of attack detection, the performance degradation caused by stealthy attacks in a noiseless LTI system has been analyzed [13, 28]. Any deviation from the expected state trajectory in the deterministic system can be considered a fault or an attack. However, non-determinism and measurement noises experienced by real-world systems provide an opportunity for the attacker to masquerade his attack as natural noises, thereby rendering attack detection probabilistic. Research [21], [20], and [26] has studied the impact of stealthy false data

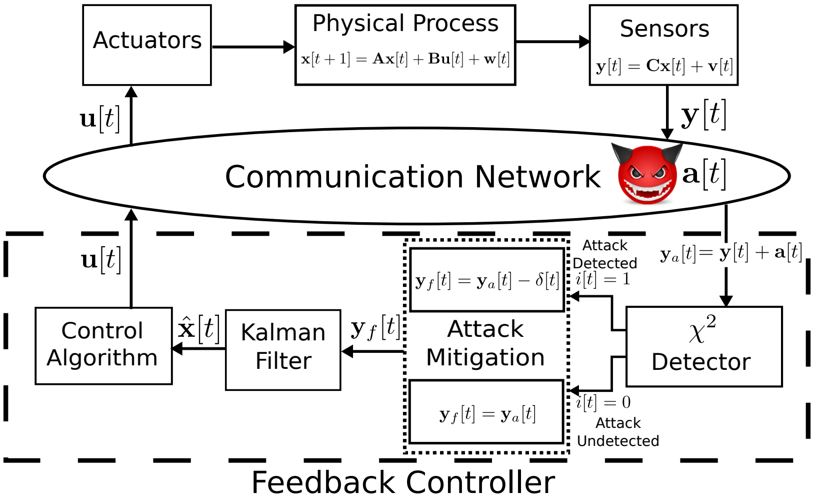


Figure 1: Block diagram of the system model.

injection (FDI) attacks against stochastic LTI systems, and derived optimal attack sequences that can cause the worst system performance degradation. Bai and Gupta [6] characterize a fundamental trade-off between the stealthiness level of an attack and the system performance degradation. However, these studies [6, 20, 21, 26] generally ignore the attack mitigation triggered by probabilistic detection of attacks and its impact on the future system states and attack detection.

In the category of attack mitigation, preventive and reactive mitigation strategies have been proposed [11]. Preventive mitigation identifies vulnerabilities in the system design and removes them to prevent exploitation by attackers. For instance, in a power system, a set of sensors and their data links can be strategically selected and protected such that a bad data detection mechanism cannot be bypassed by FDI attacks against other sensors and their links that are not protected [8, 12]. However, preventive mitigation provides static solutions only, which do not address the adaptability of strategic and knowledgeable attackers against critical infrastructures. Thus, in addition to preventative mitigation, it is important to develop reactive attack mitigation, i.e., countermeasures that are initiated after detecting an attack and tune the system based on the estimated attack activities. Reactive attack mitigation is mainly studied under game-theoretic settings [7, 23]. Specifically, the attacker manipulates a set of sensor/control signals and aims at disrupting the system operation, while the defender responds by tuning the remaining system parameters to negate the attack or minimize its impact. However, most studies on reactive mitigation (e.g., [7, 23, 33]) assume that the attack has been detected, and ignore the impact of uncertain attack detection on the overall attack mitigation. In contrast, our framework captures the interdependence between the attack detection and mitigation, and their joint impact on the system's dynamics and performance.

3 PRELIMINARIES

3.1 System Model

A block diagram of the system model is illustrated in Fig. 1. We consider a general discrete-time LTI system that evolves as

$$\mathbf{x}[t+1] = \mathbf{A}\mathbf{x}[t] + \mathbf{B}\mathbf{u}[t] + \mathbf{w}[t], \tag{1}$$

where $\mathbf{x}[t] \in \mathbb{R}^n$ is the system state vector, $\mathbf{u}[t] \in \mathbb{R}^p$ is the control input, and $\mathbf{w}[t] \in \mathbb{R}^n$ is the process noise at the t-th time slot. Matrices $\mathbf{A}$ and $\mathbf{B}$ denote the propagation and control matrices, respectively. The initial system state $\mathbf{x}[0]$ and process noise $\mathbf{w}[t]$ are independent Gaussian random variables. Specifically, $\mathbf{x}[0] \sim \mathcal{N}(\mathbf{0}, \mathbf{X})$ and $\mathbf{w}[t] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, where $\mathbf{0} = [0, \ldots, 0]^T$ and $\mathbf{X}$ and $\mathbf{Q}$ are the covariance matrices. The process described in (1) is observed through sensors deployed in the system, whose observation at time t, denoted by $\mathbf{y}[t] \in \mathbb{R}^m$, is given by

$$\mathbf{y}[t] = \mathbf{C}\mathbf{x}[t] + \mathbf{v}[t], \tag{2}$$

where $\mathbf{C} \in \mathbb{R}^{m \times n}$ is the measurement matrix and $\mathbf{v}[t] \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ is the measurement noise at time t and $\mathbf{R}$ is the covariance. We assume that $\mathbf{v}[t]$ is independent of $\mathbf{x}[0]$ and $\mathbf{w}[t]$. Moreover, we assume that the system in (1) is controllable and the measurement process in (2) is observable.

The controller uses a Kalman filter (KF) to estimate the system state based on the observations. The KF works as follows [16]:

$$\hat{\mathbf{x}}[t+1] = \mathbf{A}\hat{\mathbf{x}}[t] + \mathbf{B}\mathbf{u}[t] + \mathbf{K}(\mathbf{y}[t+1] - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}[t] + \mathbf{B}\mathbf{u}[t])), \tag{3}$$

where $\hat{\mathbf{x}}[t]$ is the estimate of the system state, $\mathbf{K}$ denotes the steady-state Kalman gain given by $\mathbf{K} = \mathbf{P}_\infty \mathbf{C}^T(\mathbf{C}\mathbf{P}_\infty \mathbf{C}^T + \mathbf{R})^{-1}$, and the matrix $\mathbf{P}_\infty$ is the solution to the algebraic Riccati equation $\mathbf{P}_\infty = \mathbf{A}\mathbf{P}_\infty \mathbf{A}^T + \mathbf{Q} - \mathbf{A}\mathbf{P}_\infty \mathbf{C}^T(\mathbf{C}\mathbf{P}_\infty \mathbf{C} + \mathbf{R})^{-1}\mathbf{C}\mathbf{P}_\infty \mathbf{A}^T$. We denote the KF estimation error at time t by $\mathbf{e}[t] = \mathbf{x}[t] - \hat{\mathbf{x}}[t]$.

LTI Model in Power Systems. The analysis in this paper is based on the general discrete-time LTI model described above. As a number of control loops found in a power system can be modeled using

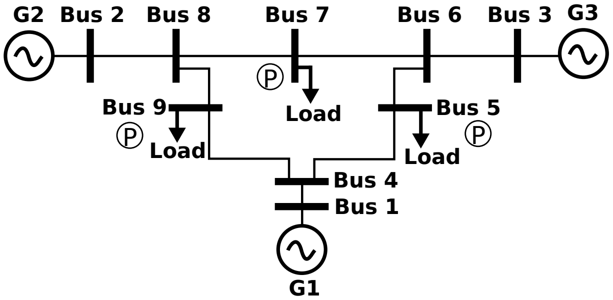


Figure 2: IEEE 9-bus power system.

the LTI model, our analysis applies to these control loops. In the following, we provide examples of a discrete-time LTI system, namely a power system's voltage control and generator swing equations.

A power system consists of a set of buses (nodes) to which generators and loads are connected, and transmission lines that connect these buses. As an example, the IEEE 9-bus test system is illustrated in Fig. 2.

Voltage control: Power system voltage control refers to maintaining the voltages of selected critical buses (called *pilot buses* marked with "P" in Fig. 2) within safe operational limits by adjusting the output voltage of the generator buses [15]. It can be modeled as an LTI system described in Eqs. (1) and (2). Specifically, the state vector $\mathbf{x}[t]$ refers to the voltages of the pilot buses at time t, which should be maintained at a nominal voltage denoted by $\mathbf{x}_0$. The control signal, which is applied at the generator buses, corresponds to the change in the generator bus voltages, i.e., $\mathbf{u}[t] = \mathbf{v}_G[t] - \mathbf{v}_G[t-1]$, where $\mathbf{v}_G[t]$ is a vector of the generator bus voltages. Under this model, the voltage control system can be approximated by an LTI system with $\mathbf{A} = \mathbf{I}$ [29], [15]. The control matrix $\mathbf{B}$ is an unknown parameter that can be estimated from real data traces (more details on estimating the matrix $\mathbf{B}$ will be presented in Section 7). Since the estimation cannot be perfect, the LTI model may be inaccurate, though the inaccuracies are small and can be captured as process noise. Since the system state can be directly measured by voltage sensors deployed at the pilot buses, the measurement matrix is an identity matrix, i.e., $\mathbf{C} = \mathbf{I}$. The system is bounded-input bounded-output stable if the control algorithm satisfies $\mathbf{B}\mathbf{u}[t] = \alpha(\mathbf{x}_0 - \mathbf{x}[t])$ for $\alpha \in (0, 1)$, and this control is adopted in practical systems [29]. However, as the sensor measurements are noisy, the controller cannot have perfect knowledge of the system state $\mathbf{x}[t]$. Rather, the state is estimated using the KF-based technique described in (3). Based on the estimated state $\hat{\mathbf{x}}[t]$, the control can be computed as

$$\mathbf{u}[t] = \alpha\mathbf{B}^{-1}(\mathbf{x}_0 - \hat{\mathbf{x}}[t]). \tag{4}$$

Generator swing equations: The swing equations establish a mathematical relationship between the angles of the mechanical motor and the generated alternating current electricity [18]. The swing equations can be linearized and modeled as an LTI system described by Eqs. (1) and (2) under the assumption of direct current (DC) power flow[27]. For a power network consisting of n generators, the state vector consists of $2n$ entries. The first n entries are the generator's rotor phase angles and the last n entries are the generator's rotor frequency. The control inputs correspond to changes in mechanical input power to the generators, and is responsible for maintaining the generator's rotor angle and frequency within a safe operational range. The entries of the matrix $\mathbf{A}$ depend on the power system's topology (including the transmission lines' susceptances) as well as the generators' mechanical parameters (such as inertia and damping constants). The structure of the matrix $\mathbf{B}$ depends on the type of feedback control used to restrict the rotor angle frequency to within the safety range [18]. The measurement vector $\mathbf{y}[t]$ under the DC power flow model includes nodal real power injections at all the buses, all the branch power flows, and the rotor angles. The observation matrix $\mathbf{C}$ can be constructed based on the power system topology [22].

3.2 Threat Model, Attack Detection & Mitigation

Modern-day critical infrastructure systems extensively use ICT for their operation. For instance, in a power grid, the remote terminal units (RTUs) and many other field devices are connected by the internet protocol (IP). The sensor and control data is transmitted over the Internet using virtual private networks (VPNs) for logical isolation [14]. However it has been demonstrated in the past that software-based protection schemes such as VPNs can be breached by attackers (e.g., see [2]). Additionally, in a power grid, the sensors (such as the voltage and current measurement units) are spread over a large geographical area, making their measurements vulnerable to physical attacks [19, 24]. Such vulnerabilities can be exploited to launch attacks and disrupt the normal power grid operations.

In this paper, we follow Kerckhoffs's principle and consider an attacker who has accurate knowledge of the targeted CPCS and read access to the system state. Such knowledge can be obtained in practice by malicious insiders, long-term data exfiltration [1], or social engineering against employees, contractors, or vendors of a critical infrastructure operator [17]. Specifically, we assume that the attacker knows the matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, as well as the operational details of the KF and the system's method of anomaly detection (including the detection threshold). In addition, the attacker also has read and write accesses to the system sensors.

We consider FDI attacks on the system sensors. Under this attack model, the compromised observations, denoted by $\mathbf{y}_a[t]$, are given by

$$\mathbf{y}_a[t] = \mathbf{y}[t] + \mathbf{a}[t], \tag{5}$$

where $\mathbf{a}[t] \in \mathbb{R}^m$ is the attacker's injection. To model the attacker's energy constraint, we assume that the norm of the injection, $||\mathbf{a}[t]||$, is upper-bounded by a constant $a_{\max}$, i.e., $||\mathbf{a}[t]|| \leq a_{\max}$. Denote by $\mathcal{A}$ the set of all feasible attack vectors that satisfy the above energy constraint.

We assume that the controller uses the χ^2 detector [25] to detect the attack, which has been widely adopted in security analysis of LTI systems [20], [26]. We note that our analysis framework can also be extended to address other attack detectors. The χ^2 detector computes a quantity $g[t] = \mathbf{r}[t]^T\mathbf{P}_r^{-1}\mathbf{r}[t]$, where $\mathbf{r}[t]$ is the residual given by

$$\mathbf{r}[t] = \mathbf{y}_a[t+1] - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}[t] + \mathbf{B}\mathbf{u}[t]), \tag{6}$$

and $\mathbf{P}_r = \mathbf{C}\mathbf{P}_\infty\mathbf{C}+\mathbf{R}$ is a constant matrix that denotes the covariance of the residual in the steady state. Denoted by $i[t] \in \{0, 1\}$ the detection result of the detector. The detector declares an attack if $g[t]$ is greater than a predefined threshold η. Specifically,

$$i[t] = \begin{cases} 0, & \text{if } 0 \leq g[t] \leq \eta; \\ 1, & \text{else.} \end{cases} \tag{7}$$

Based on the detection result, the controller applies a reactive mitigation action. If the χ^2 detector's alarm is triggered, the controller forwards a modified version of the observation $\mathbf{y}_a[t] - \boldsymbol{\delta}[t]$ to the KF, where $\boldsymbol{\delta}[t] \in \mathbb{R}^m$ is an attack mitigation signal; otherwise, the controller directly forwards $\mathbf{y}_a[t]$ to the KF (ref. Fig. 1). Thus, the controller's operation can be expressed as

$$\mathbf{y}_f[t] = \mathbf{y}_a[t] - i[t]\boldsymbol{\delta}[t]. \tag{8}$$

With the controller's mitigation action, the KF estimate is computed as

$$\hat{\mathbf{x}}[t+1] = \mathbf{A}\hat{\mathbf{x}}[t] + \mathbf{B}\mathbf{u}[t] + \mathbf{K}(\mathbf{y}_f[t+1] - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}[t] + \mathbf{B}\mathbf{u}[t])). \tag{9}$$

The mitigation signal $\boldsymbol{\delta}[t]$ can be generated using existing mitigation approaches (e.g., [9], [31]). The main focus of this paper is not the design of the mitigation strategy, but to understand the impact of the detection-mitigation loop on the optimal attack strategy. Thus, in this paper, we do not focus on a specific mitigation approach. Instead, we design a generic framework that admits any mitigation signal. In Section 7, our simulations are based on a perfect mitigation strategy in which the controller can precisely remove the attack signal, as well as a practical mitigation strategy in which the mitigation signal is a noisy version of the attack signal.

Combining (1), (8) and (9), we obtain the dynamics of the KF estimation error with attack mitigation as

$$\begin{aligned} \mathbf{e}[t+1] &= \mathbf{A}_K \mathbf{e}[t] + \mathbf{W}_K \mathbf{w}[t] \\ &\quad - \mathbf{K}(\mathbf{a}[t+1] - i[t+1]\boldsymbol{\delta}[t+1]) - \mathbf{K}\mathbf{v}[t+1], \; t \geq 0, \end{aligned} \tag{10}$$

where $\mathbf{A}_K = \mathbf{A} - \mathbf{KCA}$ and $\mathbf{W}_K = (\mathbf{I} - \mathbf{KC})$. Since the KF is assumed to be in the steady state at time 0, we have $\mathbb{E}[\mathbf{e}[0]] = \mathbf{0}$ and $\mathbb{E}[\mathbf{e}[0]\mathbf{e}[0]^T] = \mathbf{P}_e = (\mathbf{I} - \mathbf{KC})\mathbf{P}_\infty$.

4 PROBLEM FORMULATION

Under the Kerckhoffs's assumption about the attacker's knowledge, we analyze attack strategies that can mislead the controller into making erroneous control decisions. This is accomplished indirectly by increasing the estimation errors. For a given attack detection threshold η and mitigation strategy $\{\boldsymbol{\delta}[t]\}_{t=1}^T$ over a horizon of T time slots, the optimal attack sequence that maximizes the cumulative sum of KF's expected norm of the estimation error over the horizon is given by the following optimization problem:

$$\begin{aligned} \max_{\mathbf{a}[1],\ldots,\mathbf{a}[T]} \quad & \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{e}[t]\|^2] \\ s.t. \quad & \text{KF error dynamics (10)}, \\ & \|\mathbf{a}[t]\| \leq a_{\max}, \forall t. \end{aligned} \tag{11}$$

Maximizing the KF estimation error implies that the controller no longer has an accurate estimate of the system state. In systems that use KF for state estimation (such as positioning systems, power systems, etc.), control input computed based on inaccurate/wrong system state estimates can adversely affect their performance and even result in catastrophic safety incidents. Moreover, the cumulative sum in the objective function implies that the attack has a sustained adverse impact on the system over the entire attack time horizon. We note that similar cumulative metrics have also been widely adopted in control system design to assess the performance of controllers [5]. Thus, with an objective of maximizing the cumulative metric, the optimal attack sequence will bring the largest performance degradation to the control systems that are designed in terms of cumulative metrics.

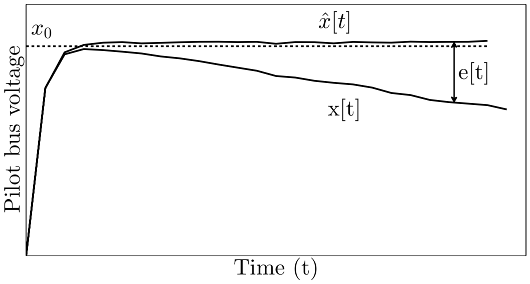


Figure 3: Attack impact for the voltage control problem.

Relevance to Power System. We illustrate the relevance of the optimization problem stated in (11) to power grid's voltage control. Recall that the voltage controller's objective is to adjust the pilot bus voltage to its setpoint $\mathbf{x}_0$ by applying control. Fig. 3 shows the impact of an attack that is able to bypass the χ^2 detector (and consequently the controller's mitigation steps) on the pilot bus voltage. In this figure, the dotted line indicates the voltage setpoint, and the solid lines show the evolution of the system state $\mathbf{x}[t]$ and estimate $\hat{\mathbf{x}}[t]$. The gap between the two curves measures the KF estimation error $\mathbf{e}[t]$. As evident from the figure, if the attacker manages to increase the KF's estimation error using a carefully constructed attack sequence, then he can cause a significant deviation of the system state from the desired setpoint. Interestingly, the estimate $\hat{\mathbf{x}}[t]$ is close to the setpoint $\mathbf{x}_0$ that misleads the controller into believing that the desired setpoint has already been achieved, while the actual pilot bus voltage continues to deviate.

Intuitively, to cause a significant impact, the attack magnitude must be large. But at the same time, it is important that the attack bypasses the controller's detection – otherwise the attack will be mitigated. Thus the solution of the optimization problem (11) must strike a balance between the attack magnitude and stealthiness. In the following section, we solve the optimization problem (11) using an MDP-based approach.

5 MDP SOLUTION

In this section, we cast the optimization problem (11) to an MDP problem [30] and solve it using the value iteration method. Before doing so, we first state the main challenge involved in solving (11).

5.1 Challenge

The main challenge in solving (11) lies in the fact that the KF error dynamics, and consequently the attack detection results are coupled across different time slots. To illustrate this point, we present a

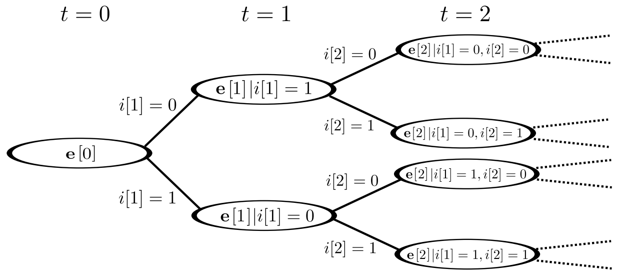


Figure 4: Evolution of KF estimation error conditioned on the attack detection results.

pictorial depiction of the KF error dynamics (10) in Fig. 4. As evident from this figure, the error dynamics of $\mathbf{e}[t]$ depend on the sequential decisions of the χ^2 detector $\mathbf{i}_{[1:t]} = \{i[t]\}_{t=1}^{T}$ due to reactive attack mitigation, which is triggered based on the the attack detection. Thus, to compute the expected error at any time t, the attacker must consider all possible combinations of the past attack detection results $\mathbf{i}_{[1:t]} = \{i[t]\}_{t=1}^{T}$. The complexity of such an approach grows exponentially in terms of the optimization time horizon T (since at any time t, there can be 2^t different combinations of the past attack detection results, see Fig. 4). In the following subsections, we present an efficient solution methodology to solve the attacker's problem (11) by modeling it as an MDP, and propose a value iteration based method to compute the optimal attack sequence.

5.2 Markov Decision Process Model

In this subsection, we develop the MDP modelling of the optimization problem (11). Our key observation is that the dynamics of the KF estimation error in (10) is Markovian. Hence, the knowledge of $\mathbf{e}[t]$ at time t will capture all the past events, and exhaustive search across all the possible past attack detection results is not necessary.

A state in the MDP corresponds to the KF filter estimation error $\mathbf{e}[t]$ and the actions correspond to the attacker's injection $\mathbf{a}[t]$. Our approach is to map the KF error dynamics (10) to the state transition probabilities of the MDP, and the objective function of (11) to the MDP's long-term expected reward. The solution to the MDP is a policy which maps each MDP state to an attacker's action. In particular, the optimal policy maximizes the long-term expected reward of the MDP, and hence solves the optimization problem (11). The mathematical details of the MDP is presented next. The structure of the MDP's solution is illustrated with the help of a numerical example in Section 5.4.

MDP Modeling Details. Formally, the MDP is defined by a tuple $(\mathcal{E}, \mathcal{A}, \mathcal{T}, R)$, where $\mathcal{E} \subseteq \mathbb{R}^n$ is the state space of the problem corresponding to the set of all possible $\mathbf{e}[t]$. $\mathcal{A}$ is the action space of the attacker. $\mathcal{T}(\mathbf{e}, \mathbf{a}, \mathbf{e}')$ is the probability of transition from state $\mathbf{e}$ to $\mathbf{e}'$ (where $\mathbf{e}, \mathbf{e}' \in \mathcal{E}$) under an action $\mathbf{a} \in \mathcal{A}$ of the attacker. Mathematically, $\mathcal{T}(\mathbf{e}, \mathbf{a}, \mathbf{e}') \triangleq \mathbb{P}(\mathbf{e}[t+1] = \mathbf{e}' \big| \mathbf{e}[t] = \mathbf{e}, \mathbf{a}[t+1] = \mathbf{a})$. $R(\mathbf{e}', \mathbf{a}, \mathbf{e})$ is the immediate expected reward for the attacker when it takes an action $\mathbf{a} \in \mathcal{A}$ in state $\mathbf{e} \in \mathcal{E}$.

MDP state transition probabilities: We now compute the state transition probability corresponding to the error dynamics (10).

We adopt the following approach. First, we compute the quantity $\mathbb{P}(\mathbf{e}_{\text{lb}} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{\text{ub}} \big| \mathbf{e}[t] = \mathbf{e}, \mathbf{a}[t+1] = \mathbf{a})$. Then we use the fact that for a random variable X,

$$\mathbb{P}(X = x) \approx \frac{F(-\infty, x+\epsilon) - F(-\infty, x-\epsilon)}{2\epsilon},$$

where $F(x_1, x_2) = \mathbb{P}(x_1 \leq X \leq x_2)$ and $\epsilon > 0$ is a small positive quantity.

The result is stated in the following lemma:

Lemma 5.1. *For a given $\mathbf{e}[t] = \mathbf{e}$ and $\mathbf{a}[t+1] = \mathbf{a}$ the attack detection probability at any time t can be computed as $\mathbb{P}(\mathrm{Y} \geq \eta)$, where $\mathrm{Y} = \mathbf{r}_c[t+1]^T \mathbf{P}_r^{-1} \mathbf{r}_c[t+1]$ is a generalized chi-square distributed random variable. Further, the quantity $\mathbb{P}(\mathbf{e}_{lb} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{ub} \big| \mathbf{e}[t] = \mathbf{e}, \mathbf{a}[t+1] = \mathbf{a})$ can be computed as the sum of the following terms:*

$$\mathbb{P}\left(\begin{bmatrix} 0 \\ \mathbf{e}_{lb} - \mathbf{y}_2 \end{bmatrix} \leq \mathbf{X} \leq \begin{bmatrix} \eta \\ \mathbf{e}_{ub} - \mathbf{y}_2 \end{bmatrix}\right) + \mathbb{P}\left(\begin{bmatrix} \eta \\ \mathbf{e}_{lb} - \mathbf{y}_2 - \mathbf{K}\boldsymbol{\delta} \end{bmatrix} \leq \mathbf{X} \leq \begin{bmatrix} \infty \\ \mathbf{e}_{ub} - \mathbf{y}_2 - \mathbf{K}\boldsymbol{\delta} \end{bmatrix}\right). \tag{12}$$

In (12), $\mathbf{X} \in \mathbb{R}^{n+1}$ is a concatenated variable given by $\mathbf{X} = \begin{bmatrix} \mathrm{Y} & (\mathbf{W}_K \mathbf{w}[t] - \mathbf{K}\mathbf{v}[t+1])^T \end{bmatrix}^T$, $\mathbf{y}_2 = \mathbf{A}_K \mathbf{e} - \mathbf{K}\mathbf{a}$, and $\boldsymbol{\delta}$ is the mitigation signal.

Lemma 5.1 is proved in Appendix A. For a generic system of dimensions $n, m \geq 2$, it is hard to obtain analytical expressions for the probability terms involved in Lemma 5.1 (since they involve a generalized chi-square distribution, as well as the correlations between the random variables Y and $\mathbf{W}_K \mathbf{w}[t] - \mathbf{K}\mathbf{v}[t+1]$, which is hard to quantify analytically). However, for the scalar case i.e. $n = m = 1$, the attack detection and transition probabilities can be computed using the Gaussian distribution, as stated in the following corollary:

Corollary 5.2. *For $n = m = 1$, the attack detection probability at any time t can be computed as*

$$\mathbb{P}\Big(\mathrm{Y} \in (-\infty, -\sqrt{\eta \mathbf{P}_r} - \mathbf{C}\mathbf{A}\mathbf{e} - \mathbf{a}] \cup [\sqrt{\eta \mathbf{P}_r} - \mathbf{C}\mathbf{A}\mathbf{e} - \mathbf{a}, \infty)\Big), \tag{13}$$

where $\mathrm{Y} \sim \mathcal{N}(0, \mathbf{C}\mathbf{Q}\mathbf{W}_K^T + \mathbf{R})$. Further, the quantity $\mathbb{P}(\mathbf{e}_{lb} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{ub} \big| \mathbf{e}[t] = \mathbf{e}, \mathbf{a}[t+1] = \mathbf{a})$ is equal to the sum of the following terms:

$$\mathbb{P}\left(\begin{bmatrix} -\sqrt{\eta \mathbf{P}_r} - \mathbf{y}_1 \\ \mathbf{e}_{lb} - \mathbf{y}_2 \end{bmatrix} \leq \mathbf{X} \leq \begin{bmatrix} \sqrt{\eta \mathbf{P}_r} - \mathbf{y}_1 \\ \mathbf{e}_{ub} - \mathbf{y}_2 \end{bmatrix}\right) + \mathbb{P}\left(\begin{bmatrix} -\infty \\ \mathbf{e}_{lb} - \mathbf{y}_2 - \mathbf{K}\boldsymbol{\delta} \end{bmatrix} \leq \mathbf{X} \leq \begin{bmatrix} -\sqrt{\eta \mathbf{P}_r} - \mathbf{y}_1 \\ \mathbf{e}_{ub} - \mathbf{y}_2 - \mathbf{K}\boldsymbol{\delta} \end{bmatrix}\right) + \mathbb{P}\left(\begin{bmatrix} \sqrt{\eta \mathbf{P}_r} - \mathbf{y}_1 \\ \mathbf{e}_{lb} - \mathbf{y}_2 - \mathbf{K}\boldsymbol{\delta} \end{bmatrix} \leq \mathbf{X} \leq \begin{bmatrix} \infty \\ \mathbf{e}_{ub} - \mathbf{y}_2 - \mathbf{K}\boldsymbol{\delta} \end{bmatrix}\right) \tag{14}$$

where $\mathbf{y}_1 = \mathbf{C}\mathbf{A}\mathbf{e} + \mathbf{a}$ and $\mathbf{y}_2 = \mathbf{A}_K \mathbf{e} - \mathbf{K}\mathbf{a}$ and $\mathbf{X} \in \mathbb{R}^2$ is a zero-mean Gaussian distributed random vector whose covariance matrix is given by

$$Cov(\mathbf{X}) = \begin{bmatrix} \mathbf{C}\mathbf{Q}\mathbf{W}_K^T + \mathbf{R} & \mathbf{C}\mathbf{Q}\mathbf{W}_K^T - \mathbf{R}\mathbf{K}^T \\ \mathbf{W}_K^T \mathbf{Q}\mathbf{C} - \mathbf{K}\mathbf{R}^T & \mathbf{W}_K \mathbf{Q}\mathbf{W}_K^T + \mathbf{K}\mathbf{R}\mathbf{K}^T \end{bmatrix}. \tag{15}$$

The probabilities in (13) and (14) can be computed using the cumulative distribution function (c.d.f.) of Gaussian distribution. Corollary 5.2 is also proved in Appendix A.

MDP reward: We now map the objective function of (11) to the MDP reward function. Accordingly, the immediate expected reward of the MDP is given by

$$R(\mathbf{e}', \mathbf{a}, \mathbf{e}) = \int_{\mathbf{e}' \in \mathcal{E}} \mathcal{T}(\mathbf{e}, \mathbf{a}, \mathbf{e}') ||\mathbf{e}'||^2. \qquad (16)$$

MDP policy and state value function: The solution to the MDP corresponds to a policy π, which is a mapping from a state to an action. The state value function of the MDP for a given policy π is defined as

$$V^{\pi}(\mathbf{e}) = \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} ||\mathbf{e}[t]||^2 \middle| \mathbf{e}[0] = \mathbf{e} \right]. \qquad (17)$$

Optimal policy: The optimal policy π^* maximizes the total expected reward, $\pi^* = \arg\max_{\pi} V^{\pi}(\mathbf{e}), \forall \mathbf{e} \in \mathcal{E}$, and the optimal value function is defined as $V^*(\mathbf{e}) = V^{\pi^*}(\mathbf{e})$.

In the next subsection, we present an algorithm to compute the optimal policy of the MDP described above.

5.3 Solving the MDP

MDPs can be solved efficiently by value/policy iteration methods [30]. However, in this work we are dealing with real-world quantities (for e.g. voltages in a power grid) which are continuous variables. Hence, the MDP described in Section 5.2 has continuous state and action spaces[1], which makes it impractical to apply a value iteration method directly. In order to address this issue, in what follows, we define a *discretized MDP* obtained by discretizing the state space of the original continuous MDP. The optimal policy of the discretized MDP can be used as a near-optimal solution to the continuous MDP. Existing studies (e.g., [10]) adopt similar discretization approaches. In the following, we provide only a sketch of the discretization procedure. More details of the discretized procedure can be found in Appendix B. This is followed by a value iteration algorithm to compute its optimal policy.

The MDP discretization procedure is based on the following three steps:

1. Construct a discretized MDP that mimics the continuous MDP closely.
2. Solve the discretized MDP using value iteration, which gives an optimal policy for the discretized MDP.
3. Map the discterized MDP's optimal policy to a near-optimal policy for the continuous MDP.

Let Ξ denote the discretized version of the original state space $\mathcal{E}$, where $\Xi = \{\xi_1, \ldots, \xi_N\}$, where N is the number descritization levels, and $\overline{\mathcal{T}}(\xi_i, \mathbf{a}, \xi_j)$, $\overline{R}(\xi_i, \mathbf{a}, \xi_j)$ and $\overline{V}(\xi_i)$ denote the state transition probabilities, the reward and value function of the discretized MDP. The mathematical details of their computation are provided in Appendix B. The discretized MDP can be solved using the value iteration method whose steps are given by the following algorithm:

Algorithm 1 (**Value Iteration**).

```
1: Set $\overline{V}_0^*(\xi_i) = 0$ for all $\xi_i \in \Xi$.
2: for $t = 0$ to T-1 do
3:     for all discretized states $\xi_i \in \Xi$ do
           $\overline{V}_{t+1}^*(\xi_i) \leftarrow \max_{\mathbf{a}} \sum_{\xi_j \in \Xi} \overline{\mathcal{T}}(\xi_i, \mathbf{a}, \xi_j) \left[\overline{R}(\xi_i, \mathbf{a}, \xi_j) + \overline{V}_t^*(\xi_j)\right]$,
           $\overline{\pi}_{t+1}^*(\xi_i) \leftarrow \arg\max_{\mathbf{a}} \sum_{\xi_j \in \Xi} \overline{\mathcal{T}}(\xi_i, \mathbf{a}, \xi_j) \left[\overline{R}(\xi_i, \mathbf{a}, \xi_j) + \overline{V}_t^*(\xi_j)\right]$.
4:     end for
5: end for
```

Algorithm 1 gives the optimal policy of the discretized MDP [30].

Note that the optimal policy of the discretized MDP computed in Algorithm 1 cannot be directly applied to the continuous MDP, since we do not know the optimal policy for a state $\mathbf{e} \in \mathcal{E}$ that is not in the discretized state space Ξ. To address this issue, we use the nearest neighbour approximation, i.e., for a state $\mathbf{e} \notin \Xi$, we choose an action based on the policy of its nearest neighbour, $\pi(\mathbf{e}) = \bar{\pi}^*(\xi_i)$, where $\xi_i = \arg\min_{1 \le i \le N} ||\mathbf{e} - \xi_i||$. We lastly make some remarks on the MDP formulation in this section.

• Although in this section we cast the optimization problem (11) as a finite time horizon MDP problem, our framework can be extended to the infinite time horizon MDP problem readily by introducing a discount factor $0 \le \gamma < 1$ in the reward function. The discount factor ensures that the cumulative sum of rewards is finite as well as the convergence of the value iteration algorithm.

• The optimal cost of the discretized MDP is guaranteed to lie within a bounded distance from the optimal cost of the original MDP [10]. As the discretization is finer, the discretized MDP approaches to the original MDP more closely.

5.4 Attack Magnitude and Stealthiness

We now illustrate the structure of the MDP solution using a numerical example. In Fig. 5, we plot the attack detection probability (computed as in (13)) and the attack impact computed in terms of the MDP's immediate expected reward (using the result of (14) and (16)) for different values of attack magnitude $\mathbf{a}$. The system parameters are $n = m = 1$, $\mathbf{A} = 1$, $\mathbf{C} = 1$, $\mathbf{Q} = 1$, $\mathbf{R} = 10$, $\eta = 10$ and $\boldsymbol{\delta} = \mathbf{a}$. It can be observed that while the probability of detection is low for an attack of small magnitude, it also has little impact. On the other hand, the probability of detection is high for an attack of large magnitude, and consequently the expected attack impact is also low. The optimal attack lies in between these two quantities. In this example, the optimal attack that maximizes the expected immediate reward has a magnitude of 10, and a detection probability of 0.3. Thus, the MDP solution strikes a balance between attack magnitude and stealthiness, resulting in maximum impact[2].

6 COST OF FALSE POSITIVES AND MISDETECTIONS

In this section, we use the framework developed thus far to quantify the cost of FPs and MDs in a simulation-based approach. We use

[1] We note that the MDP problem has continuous state and action spaces, but is not a continuous-time MDP (since we only consider discrete-time LTI systems).

[2] Strictly speaking, MDP solution maximizes the long term expected reward. For the ease of illustration, in this example we only considered the immediate expected reward.

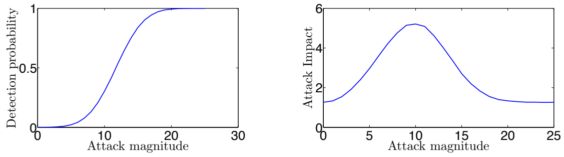


Figure 5: Attack detection probability and the expected attack impact (immediate expected reward of MDP) for different attack magnitudes.

the cumulative state estimation error (objective function of (11)) as the cost metric.

To quantify these costs, we consider an LTI system with an *oracle* attack detector as the reference system. An oracle detector is one that has a perfect detection capability, and hence no FPs or MDs. The cost of FPs is the additional cost incurred due to wrongly triggered mitigations in the original LTI system, compared with the reference system. The cost of MDs is the additional cost incurred due to unmitigated attacks in the original LTI system, compared with the reference system. In particular, we consider optimal attacks as derived in Section 5.2 to characterize the worst-case performance degradation due to MDs. We compute these costs as follows:

Cost of FPs: To quantify the cost of FPs, we compute the state estimation error (objective function of (11)) in the following two systems: (i) the LTI system of (1) and (2) with the χ^2 detector and mitigation modules and no attacks, i.e. $\mathbf{a}[t] = 0,\ \forall t$ (ii) the reference LTI system with $\mathbf{a}[t] = 0,\ \forall t$.

Under setting (i), all the alarms of the χ^2 detector correspond to FPs, which will wrongly trigger a mitigation action. Since the mitigation signal is imperfect, it leads to an increase in the estimation error. Note that for the reference system, there are no FPs, and hence no wrongly triggered mitigations. The difference between the state estimation errors of the two systems quantifies the performance degradation due to FP.

Cost of MDs: To quantify the cost of MDs, we compute the state estimation errors in the following two systems: (i) the LTI system of (1) and (2) with the χ^2 detector and mitigation modules and optimal attacks (computed as in Section 5.2) (ii) the reference LTI system with optimal attacks. The difference between the state estimation errors of the two systems quantifies the performance degradation due to MD, which we define as the cost of MD.

In Section 7.2, we present simulation results to quantify the cost of FP and MD under different attack detection thresholds and mitigation strategies. We also provide guidelines to tune the attack detection threshold based on this quantification.

7 SIMULATION RESULTS

In this section, we present simulation results to examine the system performance under different attack sequences. Throughout this section, we use different notations to denote the attacker's knowledge of the detection and mitigation parameters (η_a and $\{\boldsymbol{\delta}_a[t]\}_{t=1}^T$, respectively), and the actual parameters used by the controller (η_d and $\{\boldsymbol{\delta}_d[t]\}_{t=1}^T$, respectively). While solving the attacker's problem (11), we assume *perfect attack mitigation* in which the attack can be removed precisely, i.e. $\boldsymbol{\delta}_a[t] = \mathbf{a}[t],\ \forall t$. From an attacker's perspective, this assumption gives an underestimate of the performance degradation he can cause (since the value of the objective function of (11) will increase if the controller uses a mitigation strategy different from perfect mitigation).

While evaluating the attack impact, we consider two mitigation strategies used by the controller. First, the perfect attack mitigation $\boldsymbol{\delta}_d[t] = \mathbf{a}[t],\ \forall t$. However perfect mitigation requires the controller to estimate the injected attack vector accurately, which may not be practical. Thus we introduce a *practical attack mitigation approach* under which the attack mitigation is imperfect, i.e., $\boldsymbol{\delta}_d[t] = \mathbf{a}[t] + \mathbf{b}[t]$, where $\mathbf{b}[t] \in \mathbb{R}^n$ models the mismatch between the controller's mitigation action and the actual attack vector (possibly due to inaccuracy in estimating the injected attack). In our simulations, we generate a random vector to model $\mathbf{b}[t]$.

7.1 Optimality of the Attack Sequence

First, we verify the optimality of the attack sequence derived using the MDP-based methodology described in Section 5. We consider a general LTI model described by (1) and (2) with $n = 1$, $\mathbf{A} = 1$, $\mathbf{C} = 1$, $\mathbf{Q} = 1$, $\mathbf{R} = 10$.

We compare the cost function of (11) under three different attack sequences: (i) optimal attack computed using the MDP-based approach of Section 5, (ii) a constant attack sequence of magnitude 10 units as shown in Fig. 6a and (iii) a ramp attack as shown in Fig. 6b. The time horizon of attack T is fixed to 10 units. To implement the discretized MDP, we truncate the state space in the range $[-30, 30]$ and discretize it in equal intervals of 0.25 units. Thus, the state space of the discretized MDP consists of a total of 241 states, i.e., $\{-30, -29.75, \ldots, 0, \ldots, 29.75, 30\}$. All the optimization problems involved in the implementation of the value iteration algorithm are solved using the *fmincon* function in MATLAB.

For attack impact, we compute an empirical value of the objective function of (11) by conducting W simulation runs (where W is a large number). Let $\mathbf{e}_\omega[t]$ denote the state estimation error at a time instant $t \in \{1, 2, \ldots, T\}$ during the simulation run $\omega = \{1, 2, \ldots, W\}$, where $\mathbf{e}_\omega[t]$ follows the dynamics given by

$$\begin{aligned}\mathbf{e}_\omega[t+1] &= \mathbf{A}_K \mathbf{e}_\omega[t] + \mathbf{W}_K \mathbf{w}[t] \\ &- \mathbf{K}(\mathbf{a}^*[t+1] - i[t+1]\boldsymbol{\delta}_d[t+1]) - \mathbf{K}\mathbf{v}[t+1],\ t \geq 0,\end{aligned}$$

where $\mathbf{a}^*[t]$ is the attack derived from the MDP policy, i.e., $\mathbf{a}^*[t] = \pi^*(\mathbf{e}_\omega[t])$. The empirical cost at time t is then computed by averaging over the W simulations, i.e.,

$$\text{Cost}[t] = \frac{1}{W} \sum_{\omega=1}^{W} \sum_{\tau=1}^{t} \|\mathbf{e}_\omega[t]\|^2. \tag{18}$$

In our simulations, we set $W = 10000$. To evaluate the empirical cost under other attack strategies, we use a similar approach and replace the optimal attack with the corresponding attacks (i.e., constant and ramp attacks).

Fig. 7 provides a comparison of the cost at different time slots under the different attack sequences assuming the controller implements perfect mitigation $\boldsymbol{\delta}_d[t] = \mathbf{a}[t],\ \forall t$. It can be seen that the cost is greatest for MDP-based attacks, which validates its optimality. To investigate the attack impact under a practical mitigation strategy, we use a similar approach as described above and set $\boldsymbol{\delta}_d[t] = \mathbf{a}[t] + \mathbf{b}[t]$, where we generate $\mathbf{b}[t]$ as a Gaussian distributed random variable with a standard deviation of 15 units.

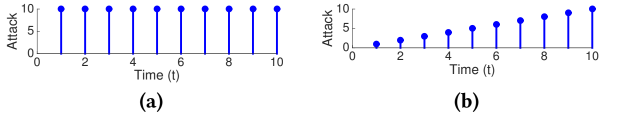


Figure 6: (a) Constant magnitude attack; (b) Ramp attack.

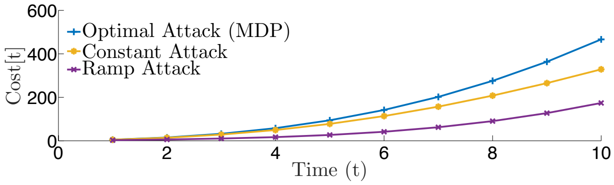


Figure 7: Comparison of cost with perfect attack mitigation and different attack strategies.

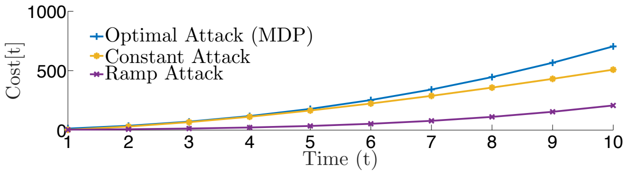


Figure 8: Comparison of cost with practical attack mitigation and different attack strategies.

From Fig. 8, it can be observed that even under the practical mitigation, the cost is greatest for the MDP-based attack sequence. Comparing Fig. 7 and Fig. 8, we also observe that the attack impact is greater for the practical mitigation compared with that of perfect mitigation (since perfect mitigation completely nullifies the attack's impact when it is detected).

7.2 Quantifying the Cost of False Positives and Misdetections

Next, we present simulation results to quantify the cost of FPs and MDs following the approach in Section 6. In our simulations, we consider the aforementioned practical attack mitigation. Fig. 9 shows the cost of FPs and MDs for different detection thresholds η and standard deviations of the attack mitigation signal σ_{mit}. We note that a low value of η represents an aggressive detector, where as a high value of η represents a conservative detector. For the mitigation signal, a low value of σ_{mit} represents accurate mitigation, where as a high value represents inaccurate mitigation. In particular, $\sigma_{\text{mit}} = 0$ corresponds to perfect mitigation.

From these plots, we observe that as the attack detection threshold η is increased, the cost of FP decreases, while the cost of MD increases. This result is intuitive – a low detection threshold detects most attacks but also leads to a high number of FPs. Thus, the wrongly triggered mitigations will result in a high FP cost. On the other hand, a high detection threshold yields a low number of FPs, but also increases the number of MDs. The figures show a basic tradeoff between FPs and MDs, quantified in terms of the cost function.

We also observe that these costs depend on the accuracy of the attack mitigation signal. E.g., when the accuracy is high (e.g., $\sigma_{\text{mit}} = 0, 5$), the cost of FP is very low, even for a low detection threshold. Thus, in this scenario, the system operator can choose a low detection threshold and obtain good system performance overall. However, when the accuracy of the mitigation signal is low (e.g., $\sigma_{\text{mit}} = 15$), the cost of FP is very high for a low detection threshold. E.g., in Fig. 9(d), the cost of FP for $\eta = 0$ is greater than the cost of MD for $\eta = 5$. In this scenario, the system operator must choose a high detection threshold to obtain an acceptable level of system performance. Thus, our result helps the system operator select an appropriate threshold that balances between the costs of FP and MD, depending on the accuracy of the mitigation signal.

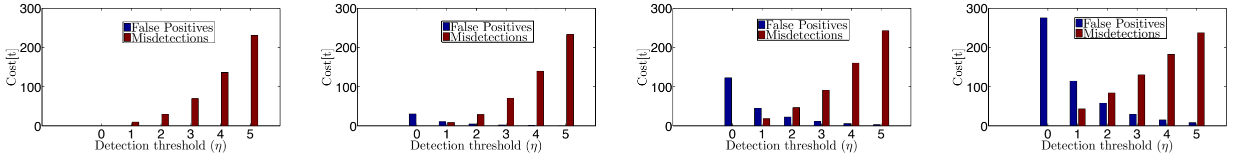
Lastly, we note that for $\sigma_{\text{mit}} = 0$ (perfect mitigation), the cost of FP is zero for all detection thresholds. Under perfect mitigation, even if an FP event occurs, the controller can accurately estimate that the attack magnitude is zero (i.e., no attack). Thus, in this specific case, wrongly triggered mitigations do not increase the cost of FP. We also note that for $\eta = 0$, there are no MDs. Hence, the cost of MD in this case is nearly zero.

7.3 Simulations for Voltage Control System

Next, we perform simulations on the voltage control system using PowerWorld, which is a high fidelity power system simulator widely used in the industry [4]. All the simulations are performed on the IEEE 9-bus system shown in Fig. 2, in which buses 1, 2, and 3 are the generator buses, whereas buses 5, 7, and 9 are the pilot buses. The control matrix $\mathbf{B}$ is estimated using linear regression on the data traces of $\mathbf{x}[t+1] - \mathbf{x}[t]$ and $\mathbf{u}[t]$ obtained in a PowerWorld simulation. We present the simulation results next.

First, we verify the accuracy of the LTI model in approximating the real-world voltage control system by examining the voltage at pilot bus 5. In our simulations, the voltage controller aims to adjust the voltage of this bus from an initial voltage of 1 pu to a setpoint ($\mathbf{x}_0$) of 0.835 pu (base voltage of 230 kV) by applying the control described in (4). Fig. 10 plots the bus voltage from $t = 1$ to $t = 30$ obtained from the PowerWorld simulations, as well as the voltage values obtained from the LTI model. To average the effect of random measurement noise, we repeat the experiment 100 times, and take the mean value. The two curves match well in this figure, thus verifying the accuracy of the proposed LTI model.

Next, we simulate the impact of the proposed attacks on the voltage control system. We assume that the attacker has access to the voltage sensor of bus 5, and injects false measurements to mislead the controller. We compute the optimal attack sequence based on the LTI model using the MDP method implemented in MATLAB. To evaluate the attack impact, we run Monte Carlo simulations using the PowerWorld simulator by injecting the derived optimal attack into the voltage measurements, and implementing the control in (4) based on the corresponding state estimate. Fig. 11 shows the pilot bus voltage (bus 5) for different attack sequences with $\eta = 5$'s and perfect attack mitigation. It can be observed that the pilot bus voltage deviates from the setpoint of 0.835 pu, and the largest voltage deviation is seen under the optimal attack. In particular,



(a) $\sigma_{\text{mit}} = 0$.

(b) $\sigma_{\text{mit}} = 5$.

(c) $\sigma_{\text{mit}} = 10$.

(d) $\sigma_{\text{mit}} = 15$.

Figure 9: Cost of FPs and MDs for different attack detection thresholds and standard deviation of the attack mitigation signal.

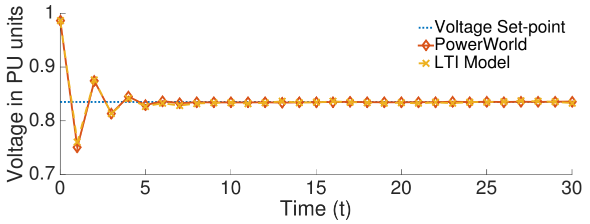


Figure 10: Comparison between PowerWorld and LTI model.

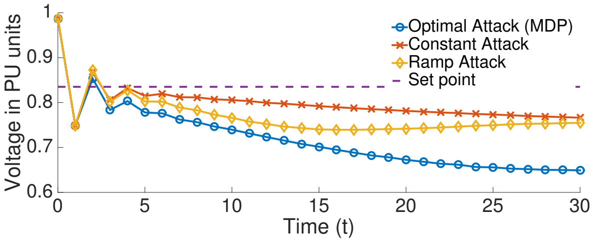


Figure 11: Pilot bus voltage (Bus 5) under different attack sequences.

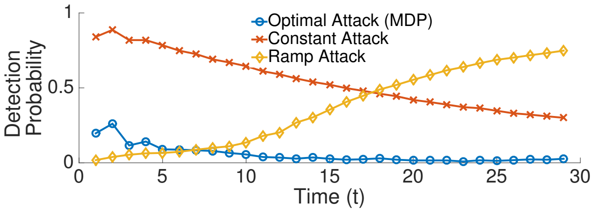
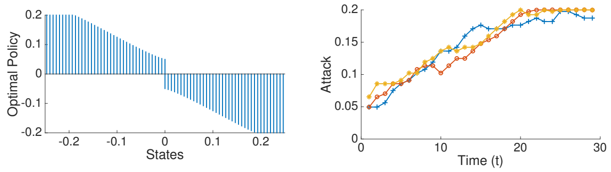


Figure 12: Attack Detection probability for different attacks.

over an attack duration of 30 time slots, we observe that bus 5 voltage deviates to 0.65 pu under the optimal attack, a difference of about 0.2 pu from its setpoint.

Fig. 12 shows the attack detection probability under these attacks at different time instants. We also plot the optimal policy computed by the value iteration algorithm (Algorithm 1) in Fig. 13a, and the optimal attack sequence for three Mote-Carlo instantiations in Fig. 13b. We observe that the attack detection probability for a naive attack sequence (such as the ramp attack) increases with time, which results in nullifying its impact (due to attack mitigation). However, the optimal attack is crafted in a way such that the detection probability decreases over time. Consequently, the optimal attack causes a significant deviation of the pilot bus voltage from its setpoint.



(a)

(b)

Figure 13: (a) Optimal policy for different system states as computed by the value iteration algorithm. (b) Optimal attack sequence for 3 Monte Carlo simulation instantiations.

8 CONCLUSION

In this paper, we studied the performance of a CPCS with attack detection and reactive attack mitigation. We derived the optimal attack sequence that maximizes the state estimation error over the attack time horizon using an MDP framework. Our results show that an arbitrarily constructed attack sequence will have little impact on the system since it will be detected, hence mitigated. The optimal attack sequence must be crafted to strike a balance between the stealthiness and the attack magnitude. Our results are useful for the system operator to assess the limit of attack impact and compare different attack detection and mitigation strategies. We also quantified the impact of FP and MD on the state estimation error, which helps select the right attack detection threshold depending on the accuracy of the attack mitigation signal. We demonstrated the application of our results to the voltage control in a power system.

REFERENCES

[1] 2014. The Dragonfly Attack. (2014). https://bit.ly/1RzGx2P.
[2] 2014. The Heartbleed Bug. (2014). http://heartbleed.com/.
[3] 2016. Confirmation of a Coordinated Attack on the Ukrainian Power Grid. (2016). http://bit.ly/1OmxfnG.
[4] 2017. PowerWorld. (2017). http://www.powerworld.com.
[5] T. Abdelzaher, Y. Diao, J. L. Hellerstein, C. Lu, and X. Zhu. 2008. Introduction to control theory and its application to computing systems. In *Performance Modeling and Engineering*. Springer, 185–215.
[6] C. Z. Bai and V. Gupta. 2014. On Kalman filtering in the presence of a compromised sensor: Fundamental performance bounds. In *Proc. American Control Conference*. 3029–3034.
[7] Carlos Barreto, Alvaro A. Cárdenas, and Nicanor Quijano. 2013. Controllability of Dynamical Systems: Threat Models and Reactive Security. In *Proc. International Conference on Decision and Game Theory for Security*. 45–64.
[8] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye. 2010. Detecting false data injection attacks on DC state estimation. In *Proc. Workshop on Secure Control Systems (SCS)*. https://tinyurl.com/mhqg99c
[9] A. A. Cárdenas, S. Amin, Z. Lin, Y. Huang, C. Huang, and S. Sastry. 2011. Attacks Against Process Control Systems: Risk Assessment, Detection, and Response. In *Proc.ACM Asia Conference on Computer and Communications Security*. 355–366.

[10] C. S. Chow and J. N. Tsitsiklis. 1991. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE Trans. Autom. Control* 36, 8 (Aug 1991), 898–914.

[11] L. F. Combita, J. Giraldo, A. A. Cardenas, and N. Quijano. 2015. Response and reconfiguration of cyber-physical control systems: A survey. In *Proc. IEEE Colombian Conference on Automatic Control (CCAC)*. 1–6.

[12] G. Dan and H. Sandberg. 2010. Stealth Attacks and Protection Schemes for State Estimators in Power Systems. In *Proc. IEEE International Conference on Smart Grid Communications*. 214–219.

[13] H. Fawzi, P. Tabuada, and S. Diggavi. 2014. Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks. *IEEE Trans. Autom. Control* 59, 6 (June 2014), 1454–1467.

[14] A. Hahn, A. Ashok, S. Sridhar, and M. Govindarasu. 2013. Cyber-Physical Security Testbeds: Architecture, Application, and Evaluation for Smart Grid. *IEEE Trans. Smart Grid* 4, 2 (June 2013), 847–855.

[15] M. D. Ilic, Xiaojun Liu, G. Leung, M. Athans, C. Vialas, and P. Pruvot. 1995. Improved secondary and new tertiary voltage control. *IEEE Trans. Power Syst.* 10, 4 (Nov 1995), 1851 – 1862.

[16] T. Kailath, A.H. Sayed, and B. Hassibi. 2000. *Linear Estimation*. Prentice Hall.

[17] S. Karnouskos. 2011. Stuxnet worm impact on industrial cyber-physical system security. In *Conf. IEEE Industrial Electronics Society*.

[18] P. Kundur. 1994. *Power System Stability and Control*. McGraw-Hil, New York, US.

[19] D. F. Kune, J. Backes, S. S. Clark, D. Kramer, M. Reynolds, K. Fu, Y. Kim, and W. Xu. 2013. Ghost talk: Mitigating EMI signal injection attacks against analog sensors. In *IEEE Symp. Security and Privacy*.

[20] C. Kwon, W. Liu, and I. Hwang. 2013. Security analysis for Cyber-Physical Systems against stealthy deception attacks. In *Proc. American Control Conference*. 3344–3349.

[21] W. Liu, C. Kwon, I. Aljanabi, and I. Hwang. 2012. Cyber security analysis for state estimators in air traffic control systems. In *Proc. AIAA Conference on Guidance, Navigation, and Control*.

[22] Y. Liu, P. Ning, and M. K. Reiter. 2009. False Data Injection Attacks Against State Estimation in Electric Power Grids. In *Proc. ACM Conference on Computer and Communications Security*. 21–32.

[23] C. Y. T. Ma, D. K. Y. Yau, X. Lou, and N. S. V. Rao. 2013. Markov Game Analysis for Attack-Defense of Power Networks Under Possible Misinformation. *IEEE Trans. Power Syst.* (May 2013), 1676–1686.

[24] P. McDaniel and S. McLaughlin. 2009. Security and Privacy Challenges in the Smart Grid. *IEEE Security Privacy* 7, 3 (2009), 75–77.

[25] R.K. Mehra and J. Peschon. 1971. An innovations approach to fault detection and diagnosis in dynamic systems. *Automatica* 7, 5 (1971), 637 – 640.

[26] Y. Mo and B. Sinopoli. 2016. On the Performance Degradation of Cyber-Physical Systems Under Stealthy Integrity Attacks. *IEEE Trans. Autom. Control* 61, 9 (Sep. 2016), 2618–2624.

[27] F. Pasqualetti, F. Dorfler, and F. Bullo. 2011. Cyber-Physical Attacks in Power Networks: Models, Fundamental Limitations and Monitor Design. (2011). https://arxiv.org/abs/1103.2795.

[28] F. Pasqualetti, F. DÃűrfler, and F. Bullo. 2013. Attack Detection and Identification in Cyber-Physical Systems. *IEEE Trans. Autom. Control* 58, 11 (Nov 2013), 2715–2729.

[29] J.P. Paul and J.Y. Leost. 1987. Improvements of the secondary voltage control in France. In *Power Systems and Power Plant Control*. 83 – 88.

[30] M. L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA.

[31] S. Sridhar and M. Govindarasu. 2014. Model-Based Attack Detection and Mitigation for Automatic Generation Control. *IEEE Trans. Smart Grid* 5, 2 (March 2014), 580–591.

[32] Q. Dinh Vu, R. Tan, and D. K. Y. Yau. 2016. On applying fault detectors against false data injection attacks in cyber-physical control systems. In *Proc. IEEE INFOCOM*. 1–9.

[33] Q. Zhu and T. Basar. 2015. Game-Theoretic Methods for Robustness, Security, and Resilience of Cyberphysical Control Systems: Games-in-Games Principle for Optimal Cross-Layer Resilient Control Systems. *IEEE Control Syst. Mag* 35, 1 (Feb 2015), 46–65.

APPENDIX A: PROOFS OF LEMMA 5.1 AND COROLLARY 5.2

Attack Detection Probability. We start with the attack detection probability. To this end, we derive the relationship between the residual $\mathbf{r}[t+1]$ and the KF estimation error $\mathbf{e}[t]$. Using (1), (2) and (5) in (6), followed by some algebraic manipulations, we obtain

$$\mathbf{r}[t+1] = \mathbf{CAe}[t] + \mathbf{Cw}[t] + \mathbf{a}[t+1] + \mathbf{v}[t+1]. \tag{19}$$

Let $\mathbf{r}_c[t+1]$ denote the conditional random variable given by

$$\begin{aligned}\mathbf{r}_c[t+1] &\triangleq \mathbf{r}[t+1]\Big|\mathbf{e}[t]=\mathbf{e}, \mathbf{a}[t+1]=\mathbf{a}\\ &= \mathbf{Cw}[t] + \mathbf{v}[t+1] + \mathbf{CAe} + \mathbf{a},\end{aligned} \tag{20}$$

where (20) is obtained from (19). According to the chi-square detection rule, the attack detection probability for a given value of $\mathbf{e}[t]=\mathbf{e}$ and $\mathbf{a}[t+1]=\mathbf{a}$ can be computed as

$$\mathbb{P}(\mathbf{r}_c[t+1]^T\mathbf{P}_r^{-1}\mathbf{r}_c[t+1] \geq \eta). \tag{21}$$

From (20), it follows that the random variable $\mathbf{r}_c[t+1]^T\mathbf{P}_r^{-1}\mathbf{r}_c[t+1]$ follows a generalized chi-square distribution, which can be used to compute the probability in (21).

In particular for $n=m=1$, it follows that the attack is detected if $\mathbf{r}_c[t+1] \geq \sqrt{\eta\mathbf{P}_r}$, which is satisfied if (from (20))

$$\begin{aligned}\mathbf{Cw}[t] + \mathbf{v}[t+1] \in (-\infty, -\sqrt{\eta\mathbf{P}_r} - \mathbf{CAe} - \mathbf{a}]\\ \cup\, [\sqrt{\eta\mathbf{P}_r} - \mathbf{CAe} - \mathbf{a}, -\infty).\end{aligned} \tag{22}$$

Alternately, the chi-square detector misses the attack if the noise terms satisfy

$$\begin{aligned}&\mathbf{Cw}[t] + \mathbf{v}[t+1] \in\\ &\qquad\left[-\sqrt{\eta\mathbf{P}_r} - \mathbf{CAe} - \mathbf{a}, \sqrt{\eta\mathbf{P}_r} - \mathbf{CAe} - \mathbf{a}\right].\end{aligned} \tag{23}$$

The probabilities of the events in (22) and (23) correspond to attack detection and misdetection probabilities, which can be computed using the c.d.f. of the Gaussian distribution.

MDP State Transition Probabilities. Next, we compute the quantity $\mathbb{P}(\mathbf{e}_{\text{lb}} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{\text{ub}}\big|\mathbf{e}[t]=\mathbf{e}, \mathbf{a}[t+1]=\mathbf{a})$. Recall the KF error evolution in (10). Depending on the attack detection result $i[t+1]$, there are two cases:

- Case 1: When $i[t+1]=0$, and $\mathbf{e}_{\text{lb}} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{\text{ub}}$.
- Case 2: When $i[t+1]=1$, and $\mathbf{e}_{\text{lb}} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{\text{ub}}$.

The quantity $\mathbb{P}(\mathbf{e}_{\text{lb}} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{\text{ub}}\big|\mathbf{e}[t]=\mathbf{e}, \mathbf{a}[t+1]=\mathbf{a})$ can be computed as the sum of probabilities of the two cases. We investigate each case separately and derive its probability.

- Case 1: Substituting $i[t+1]=0$, in (10), we obtain

$$\mathbf{e}[t+1] = \mathbf{A}_K\mathbf{e}[t] + \mathbf{W}_K\mathbf{w}[t] - \mathbf{Ka}[t] - \mathbf{Kv}[t+1]. \tag{24}$$

Given $\mathbf{e}[t]=\mathbf{e}$, and $\mathbf{a}[t+1]=\mathbf{a}$, to have $\mathbf{e}_{\text{lb}} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{\text{ub}}$, the noise terms must satisfy (from (24))

$$\begin{aligned}&\mathbf{W}_K\mathbf{w}[t] - \mathbf{Kv}[t+1] \in\\ &\qquad[\mathbf{e}_{\text{lb}} - \mathbf{A}_K\mathbf{e} + \mathbf{Ka}, \mathbf{e}_{\text{ub}} - \mathbf{A}_K\mathbf{e} + \mathbf{Ka}].\end{aligned} \tag{25}$$

In Case 1, the conditions (21) and (25) must be satisfied simultaneously, the probability of which can be computed as the joint probability of the two events given by the result of Lemma 5.1 (first expression of (12)).

In particular, for $n=m=1$, the conditions (23) and (25) must be satisfied simultaneously, the probability of which is given by

$$\mathbb{P}\left(\begin{bmatrix}-\sqrt{\eta\mathbf{P}_r} - \mathrm{y}_1\\ \mathbf{e}_{\text{lb}} - \mathbf{y}_2\end{bmatrix} \leq \mathbf{X} \leq \begin{bmatrix}\sqrt{\eta\mathbf{P}_r} - \mathrm{y}_1\\ \mathbf{e}_{\text{ub}} - \mathbf{y}_2\end{bmatrix}\right) \tag{26}$$

where $\mathbf{X} \in \mathbb{R}^{2n}$ is the concatenated vector given by

$$\mathbf{X} = \begin{bmatrix}\mathbf{Cw}[t] + \mathbf{v}[t]\\ \mathbf{W}_K\mathbf{w}[t] - \mathbf{Kv}[t]\end{bmatrix},$$

and $\mathbf{y}_1 = \mathbf{C}\mathbf{A}\mathbf{e} + \mathbf{a}$ and $\mathbf{y}_2 = \mathbf{A}_K\mathbf{e} - \mathbf{K}\mathbf{a}$. The probability in (26) can be computed using the Gaussian distribution as follows. Since $\mathbf{w}[t]$ and $\mathbf{v}[t+1]$ are Gaussian, the terms $\mathbf{C}\mathbf{w}[t] + \mathbf{v}[t]$ and $\mathbf{W}_K\mathbf{w}[t] - \mathbf{K}\mathbf{v}[t]$ are jointly Gaussian distributed. It is straightforward to note the mean value of the concatenated vector, i.e.,

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[\mathbf{C}\mathbf{w}[t] + \mathbf{v}[t]] \\ \mathbb{E}[\mathbf{W}_K\mathbf{w}[t] - \mathbf{K}\mathbf{v}[t]] \end{bmatrix} = \mathbf{0},$$

and its covariance matrix is given by

$$\mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \mathbf{C}\mathbf{Q}\mathbf{W}_K^T + \mathbf{R} & \mathbf{C}\mathbf{Q}\mathbf{W}_K^T - \mathbf{R}\mathbf{K}^T \\ \mathbf{W}_K^T\mathbf{Q}\mathbf{C} - \mathbf{K}\mathbf{R}^T & \mathbf{W}_K\mathbf{Q}\mathbf{W}_K^T + \mathbf{K}\mathbf{R}\mathbf{K}^T \end{bmatrix}. \quad (27)$$

Following the above arguments, we can compute (26) using the c.d.f. of the Gaussian distribution.

- Case 2: Substituting $i[t+1] = 1$, in (10), we obtain

$$\mathbf{e}[t+1] = \mathbf{A}_K\mathbf{e} + \mathbf{W}_K\mathbf{w}[t] - \mathbf{K}(\mathbf{a} - \boldsymbol{\delta}) - \mathbf{K}\mathbf{v}[t+1]. \quad (28)$$

Given $\mathbf{e}[t] = \mathbf{e}$, $\mathbf{a}[t+1] = \mathbf{a}$, and $\boldsymbol{\delta}[t+1] = \boldsymbol{\delta}$ to have $\mathbf{e}_{\mathrm{lb}} \leq \mathbf{e}[t+1] \leq \mathbf{e}_{\mathrm{ub}}$, the noise terms must satisfy (from (28))

$$\mathbf{W}_K\mathbf{w}[t] - \mathbf{K}\mathbf{v}[t+1] \in \Big[\mathbf{e}_{\mathrm{lb}} - \mathbf{A}_K\mathbf{e} + \mathbf{K}(\mathbf{a} - \boldsymbol{\delta}), \mathbf{e}_{\mathrm{ub}} - \mathbf{A}_K\mathbf{e} + \mathbf{K}(\mathbf{a} - \boldsymbol{\delta})\Big]. \quad (29)$$

In Case 2, the conditions (22) and (29) must be satisfied simultaneously, the probability of which can be computed as the joint probability of the two events, given by the result of Lemma 5.1 (second expression of (12)).

For $n = m = 1$, the conditions (22) and (25) must be satisfied simultaneously, the probability of which is provided in the result of Corollary 5.2 (second and third expressions of (14)). The probabilities can be computed using the c.d.f. of the Gaussian distribution similar to Case 1.

APPNEDIX B: MDP DISCRETIZATION

In this appendix, we provide details of the MDP discretization procedure of Section 5.3. Formally, we define the discretized MDP by a tuple $(\Xi, \mathcal{A}, \overline{\mathcal{T}}, \overline{R})$. Here in, Ξ denotes a discretized version of the original state space $\mathcal{E}$, given by $\Xi = \{\xi_1, \ldots, \xi_N\}$, where N is the number descritization levels, and $\overline{\mathcal{T}}(\xi_i, \mathbf{a}, \xi_j)$, $\overline{R}(\xi_i, \mathbf{a}, \xi_j)$ and $\overline{V}(\xi_i)$ denote respectively the state transition probabilities, the reward, and the value function of the discretized MDP. Next, we elaborate the three steps involved in the MDP discretization as stated in Section 5.3.

We start with Step 1, i.e., construction of the discretized MDP from the original continuous MDP. A pictorial illustration of the discretization procedure is shown in Fig. 14. In this figure, points $\xi_1, \xi_2, \ldots$ represent a discretized version of the original MDP's continuous state space. Note that the points $\xi_1, \xi_2, \ldots$ are a subset of the original state space. The arrows in Fig. 14 show a mapping between state transitions of the continuous MDP to those of the discretized MDP. The mapping is based on the following logic. Consider all the state transitions in the continuous MDP from $\mathbf{e}[t] = \xi_i$, $i = 1, 2, \ldots$, to a state $\mathbf{e}[t+1] = \mathbf{e}' \notin \{\xi_1, \xi_2, \ldots\}$ under an action $\mathbf{a}$. In the discretized MDP, all such transitions are mapped to a state ξ_i, $i = 1, 2, \ldots$, that is nearest to $\mathbf{e}'$. E.g., in Fig. 14, all the state transitions in the continuous MDP from $\mathbf{e}[t] = \xi_1$ to the states $\mathbf{e}'_1$ and $\mathbf{e}'_2$ are mapped to ξ_2 (since ξ_2 is closest to $\mathbf{e}'_1$ and $\mathbf{e}'_2$ in the discretized state space). The state transition probabilities from ξ_1 to ξ_2 in the discretized MDP is computed as the sum (and in the limiting case, the integration) of all such state transition probabilities of the continuous MDP.

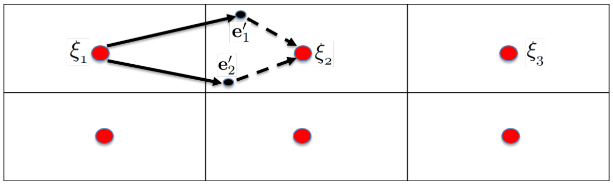


Figure 14: A pictorial representation of the discretization procedure.

Based on this logic, a mathematically rigorous way to compute $\overline{\mathcal{T}}(\xi_i, \mathbf{a}, \xi_j)$ from $\mathcal{T}(\xi_i, \mathbf{a}, \xi_j)$ is given by

$$\overline{\mathcal{T}}(\xi_i, \mathbf{a}, \xi_j) = \mathbb{P}(\xi_j \big| \xi_i, \mathbf{a}) = \int_{\mathbf{e}' \in B(\xi_j)} \mathcal{T}(\xi_i, \mathbf{a}, \mathbf{e}'),$$

where $B(\xi_j)$ denotes the set of points $\mathbf{e}' \in \mathcal{E}$ which are closer to ξ_j than any other point $\xi_k \in \Xi, k \neq j$. Mathematically, $B(\xi_j)$ is given by

$$B(\xi_j) = \{\mathbf{e}' \in \mathcal{E} \big| d(\mathbf{e}', \xi_j) \leq d(\mathbf{e}', \xi_k), 1 \leq k \leq N, k \neq j\},$$

where $d(\mathbf{x}, \mathbf{y})$ denotes the Euclidean distance between the points $\mathbf{x}$ and $\mathbf{y}$, i.e. $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$.

The immediate expected reward in the discretized MDP $\overline{R}(\xi_i, \mathbf{a}, \xi_j)$ can be computed as $\overline{R}(\xi_i, \mathbf{a}, \xi_j) = \sum_{j=1}^{N} \overline{\mathcal{T}}(\xi_i, \mathbf{a}, \xi_j)||\xi_j||^2$.

Next, we proceed to Step 2 of the discrterization procedure, i.e., computing the optimal policy of the discretized MDP. We use the notations $\overline{\pi}^*$ to denote the optimal policy of the discretized MDP and $\overline{V}^*(\xi_i)$ to denote the optimal state value function of state $\xi_i \in \Xi$. They can be computed using the value iteration algorithm specified as Algorithm 1.

Finally, we proceed to Step 3 of the discrterization procedure, i.e., mapping the optimal policy of the discretized MDP to a near-optimal policy of the continuous MDP. First note that the optimal policy of the discretized MDP computed in Algorithm 1 cannot be directly applied to the continuous MDP, since we do not know the optimal policy for a state $\mathbf{e} \in \mathcal{E}$ that is not in the discretized state space Ξ. To address this issue, we use the nearest neighbour approximation, i.e., for a state $\mathbf{e} \notin \Xi$, we choose an action based on the policy of its nearest neighbour, $\pi(\mathbf{e}) = \bar{\pi}^*(\xi_i)$, where $\xi_i = \arg\min_{1 \leq i \leq N} ||\mathbf{e} - \xi_i||$.