Automatica 113 (2020) 108680

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Defensive deception against reactive jamming attacks in remote state estimation[☆]

Kemi Ding[a], Xiaoqiang Ren[b,*], Daniel E. Quevedo[c], Subhrakanti Dey[d], Ling Shi[a]

[a] Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong
[b] School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
[c] Department of Electrical Engineering, Paderborn University, Germany
[d] Hamilton Institute, National University of Ireland, Maynooth, Ireland

ARTICLE INFO

Article history:
Received 25 October 2017
Received in revised form 15 June 2019
Accepted 22 October 2019
Available online 8 November 2019

Keywords:
Kalman filters
System security
Defensive deception
State estimation
Game theory

ABSTRACT

This paper considers a synthetic counter-measure, combining transmission scheduling and defensive deception, to defend against jamming attacks in remote state estimation. In the setup studied, an attacker sabotages packet transmissions from a sensor to a remote estimator by congesting the communication channel between them. In order to efficiently degrade the estimation accuracy, the intelligent attacker tailors its jamming strategy by reacting to the real-time information it collects. In response to the jamming attacks, the sensor with a long-term goal will select the transmission power level at each stage. In addition, by modifying the real-time information intentionally, the sensor creates asymmetric uncertainty to mislead the attacker and thus mitigate attacks. Considering the dynamic nature of the process, we model the strategic interaction between the sensor and the attacker by a general stochastic game with asymmetric information structure. To obtain stationary optimal strategies for each player, we convert this game into a belief-based dynamic game and analyze the existence of its optimal solution. For a tractable implementation, we present an algorithm that finds equilibrium strategies based on multi-agent reinforcement learning for symmetric-information stochastic games. Numerical examples illustrate properties of the proposed algorithm.

1. Introduction

Cyber–physical systems (CPSs), merging communication networks and computational elements into feedback systems, provide great robustness and stability, real-time monitoring and efficient controls to physical processes (Kim & Kumar, 2012). These capabilities impel the utilization of CPSs in various realms, including smart grids, transportation systems, critical infrastructures (e.g., gas supply and water pollution monitoring systems) and ubiquitous wearable medical devices. Despite the enormous advantages brought by communication and information technologies, some cyber components have unfortunately exhibited vulnerabilities for malicious adversaries to exploit, leaving CPS security a great concern (Mo, Kim, Brancik, Dickinson, Lee, Perrig, & Sinopoli, 2012). Deliberate attackers can gain access to wireless connections among sensors, estimators and actuators to launch cyber attacks. For example, the Ukraine blackout has been regarded as the first power outage accident in the world caused by cyber attackers (SANS, 2016). Two typical classes of cyber attacks on CPSs, as summarized in Cardenas, Amin, and Sastry (2008), are: integrity attack and denial-of-service (DoS) attack. DoS attacks compromise the availability of resources, and compared with other cyber attacks, they are most accomplishable and common in practical CPSs (Feng & Tesi, 2017). In this paper, we investigate a remote state estimation problem under reactive jamming attacks, which is a type of DoS attack in Grover, Lim, and Yang (2014). To be more specific, the attacker targets on jeopardizing the transmission of measurements from a sensor to a remote estimator, see Fig. 1.

Literature review of defense mechanisms against jamming attacks and motivation. Security against jamming threats has been amply investigated in traditional communication and information systems (Grover et al., 2014). In summary, existing anti-jamming solutions range from physical-layer defenses (e.g., using directional antennas to maintain communication connectivity in multi-hop wireless networks (Noubir, 2004) or spread spectrum to tolerate

[☆] The work by K. Ding and L. Shi is supported by a Hong Kong RGC theme-based project T23-701/14N. The work by S. Dey is supported by the VR (Swedish Research Council) Project Grant Dnr 621-2013-5395. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Juan C. Aguero under the direction of Editor Torsten Söderström.

[*] Corresponding author.

E-mail addresses: kdingaa@ust.hk (K. Ding), xqren@shu.edu.cn (X. Ren), dquevedo@ieee.org (D.E. Quevedo), subhrakanti.dey@angstrom.uu.se (S. Dey), eesling@ust.hk (L. Shi).

https://doi.org/10.1016/j.automatica.2019.108680
0005-1098/© 2019 Elsevier Ltd. All rights reserved.

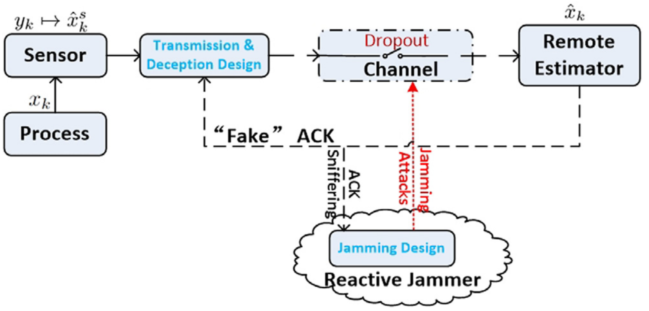


Fig. 1. System model.

interference in Pickholtz, Schilling, and Milstein (1982), to link-layer defenses (e.g., coordinated spatial–temporal randomization in Pajic and Mangharam (2009) and channel hopping in Khattab, Mosse, and Melhem (2008), Lazos, Liu, and Krunz (2009) and Navda, Bohra, Ganguly, and Rubenstein (2007) to improve jamming resiliency in networks), and even network-layer defenses (such as spatial retreats in Xu, Wood, Trappe, and Zhang (2004)). However, these defensive approaches could be insufficient to fully address the security challenges in CPSs, since they leave out a particularly crucial characteristic of CPSs, namely, the tight coupling between the cyber domain and physical processes. A key point to note is that, jamming security in cyber/communication literature is confined to study stationary data sources, rather than dynamic physical systems in CPSs, and hence might struggle to take possible consequences of attacks on physical dynamical systems.

Many existing works in the networked control systems community have made great efforts to analyze and evaluate the vulnerabilities of CPSs to jamming attacks (Cardenas et al., 2008; Zhang, Cheng, Shi, & Chen, 2015). Generally, the designed defense mechanisms are ad hoc in order to fulfill various control objectives for the specific physical systems. For example, a family of impulsive controllers was proposed in Feng and Tesi (2017) to guarantee the closed-loop stability of a linear time-invariant system under a jamming attack with limited frequency and duration. The authors in Befekadu, Gupta, and Antsaklis (2015) derived an optimal stochastic control policy for the risk-sensitive control problem in the presence of a Markov-modulated jamming attack and De Persis and Tesi (2015) provided transmission scheduling under restricted jamming attacks to preserve input-to-output stability of a closed-loop system. Taking into account transmission power limitations, Qin, Li, Shi, and Yu (2018) designed an optimal transmission scheme to improve the remote estimation accuracy under jamming attacks. Note that, in practice, the jammer is capable to obtain real-time information of systems and tailor its jamming schemes accordingly. To avoid such intelligent jammers from degrading estimation performance, Ding, Li, Quevedo, Dey, and Shi (2017) adopted a game-theoretic approach to design an optimal channel-hopping defensive policy. Common to the approaches mentioned above is that it is assumed that the attacker has genuine information about the system. Motivated by this, an emerging concept—defensive deception (Wang & Lu, 2018) was proposed to prevent cyber attacks. The key idea is to manipulate the attacker's behavior via a carefully-crafted deception scheme.

Defensive deception mechanism and contributions. In our current paper, besides a transmission scheme resilient to jamming attacks, we propose an ad-hoc deceptive defense mechanism to disrupt potential jamming attacks in remote state estimation. The key of deceptive defense is to provide plausible-looking, yet misleading, real-time information of the system to deceive the attacker, and thereby cause it to waste jamming resources. The details of our deceptive defense are as follows: the sensor decides its transmission energy based on the acknowledgment signals (ACKs) sent back from the remote estimator, which may also be intercepted by the attacker to adjust its jamming policy deliberately (Ding et al., 2017; Li, Shi, Cheng, Chen, & Quevedo, 2015); in order to confuse the attacker, the sensor may modify the ACKs by sending an ACK-reverse instruction to the estimator before packet transmission/jamming. Assuming the attacker follows a pre-determined jamming tactic, our previous work (Ding, Ren, & Shi, 2016) investigated an optimal deception scheme for the sensor. However, the obtained deception scheme is insufficient. In particular, if the transmission strategies and the coupled deception scheme are jointly considered by the attacker, then it may modify its jamming tactic accordingly. Thus, a sophisticated interaction between the sensor and the attacker is investigated in our current work. Considering an infinite-time horizon, we formulate the strategic interaction as an asymmetric-information stochastic game, in which the attacker needs to handle the unknown environment dynamics induced by the deception "tricks".

The concept of defensive deception has been developed in cyber security recently. For example, the authors in Carroll and Grosu (2011) employed camouflage in a network by disguising honeypots as real systems or revealing real systems as honeypots. They used a static signal game to quantitatively analyze the interaction between the defender and the attacker. Different from this space-domain deception in Carroll and Grosu (2011), our defensive deception in the current work is to schedule the ACK-reverse instruction in a temporal manner. The survey (Pawlick, Colbert, & Zhu, 2017) provided a taxonomy that defines six types of defensive deception in cyber security and classified preliminary works on defensive deception based on their game-theoretic models. Most of them focus on simple game models, such as static Nash games in Zhu and Başar (2013), Stackelberg games in Feng, Zheng, Mohapatra, and Cansever (2017) or signaling games in Pawlick and Zhu (2015), while excluding advanced dynamic games (in which the strategic interaction between the attacker and the defender occurs over multiple stages). The recent work (Horák, Zhu, & Bošanskỳ, 2017) considers the sequential nature of attackers, and utilizes a zero-sum one-sided partially observed stochastic game, played in an infinite-time horizon under deterministic transition, to design optimal dynamic deception approaches for computer networks.

Compared with existing works about defensive deception, our game model is more sophisticated. First, our work concerns a dynamic game with twofold actions for the sensor (i.e., the coupled ACK-deception and transmission scheduling). In each stage, the sensor/attacker with a long-term goal will select the transmission/jamming power level strategically. In addition, the sensor may disclose ACKs truthfully or deceptively to manipulate the attacker's belief to prevent it from jamming efficiently and thus minimize the damage. Second, we employ a more general framework of an asymmetric-information stochastic game, especially with nonzero-sum payoffs and randomized transitions. The generality poses challenges to prove the existence of equilibria (Madani, Hanks, & Condon, 1999). In contrast to Horák et al. (2017) where equilibrium analysis is absent, we provide a mathematical framework to solve this general asymmetric-information stochastic game via resorting to the Markovian belief-state technique. The latter is conventionally used in reformulating a partially observable Markov decision process (POMDP) to a Markov decision process (MDP). Also, we propose an algorithm based on multi-agent reinforcement learning to find optimal defensive deception strategies.

In summary, the main contributions of our work are twofold: (1) Compared with existing works on anti-jamming techniques, we propose a synthetic defense mechanism that combines defensive deception and transmission scheduling to alleviate jamming effects on remote estimation accuracy.

(2) Compared with game models in previous works on defensive deception, we use a general asymmetric-information dynamic game to model the infinite-time strategic interactions between the sensor and the attacker. Moreover, we develop an equivalent belief-based stochastic game to obtain the optimal stationary strategies for each agent. This equivalent game has continuous state space and infinite action space. Therefore, analyzing the game solution becomes very involved. In Theorem Theorem 3.2, we prove the existence of optimal strategies. Moreover, a grid-based multi-agent reinforcement learning algorithm is proposed in Section 3.4, in which each agent gradually learns its optimal strategies through interactions with both the opponent and the (unknown) dynamic environment.

Outline. The remainder of the paper is organized as follows. Section 2 contains mathematical models of the remote state estimation setup of interest. It also presents our anti-jamming scheme, which combines defensive deception and transmission strategies. The framework of the stochastic game between the sensor and the attacker is presented in Section 3, where the existence of an optimal solution is proved and a multi-agent reinforcement learning algorithm is provided to obtain the optimal strategies. Some examples and concluding remarks are presented in Sections 4 and 5, respectively.

Notations: $\mathbb{R}^n$ is the n dimensional Euclidean space. $\mathbb{S}_+^n$ (or $\mathbb{S}_{++}^n$) is the set of n by n positive semi-definite matrices (or positive definite matrices). Let $\mathbb{N}$ denote the set of natural numbers. When $X \in \mathbb{S}_+^n$ (or $X \in \mathbb{S}_{++}^n$), we write $X \geq 0$ (or $X > 0$). For functions h, g, $h \circ g$ is defined as the function composition $h(g(\cdot))$. $\mathbb{E}[\cdot]$ is the expectation of a random variable, $\Delta(\cdot)$ refers to the probability measure space over a set, and $\mathtt{Pr}(\cdot)$ refers to probability. $\mathtt{Tr}(\cdot)$ denotes the trace of a matrix. The superscripts $\top$ and $\star$ stand for transposition and optimal solution, respectively, while the superscripts/subscripts *1* and *2* denote the sensor and the attacker, respectively. The superscripts "e" and "c" in a_k^e and a_k^c stand for energy and cheating actions, respectively. The use of bold, bold capital and calligraphic letters follows the convention in the strategic form games (Fudenberg & Tirole, 1991). We represent the sets by calligraphic letters, random variables by bold capital letters and particular realizations by bold lowercase letters. y_0^k stands for the sequence $\{y_0, \ldots, y_k\}$. $\mathbf{1}(\cdot)$ is the indicator function and the Dirac delta function is defined as:

$$\delta_{kj} = \begin{cases} 1, & \text{if } k = j; \\ 0, & \text{others.} \end{cases}$$

2. Problem formulation

As depicted in Fig. 1, the sensor transmits state information of the process to the remote estimator under jamming attacks. In this section, we introduce the essential components of the overall system structure.

2.1. Kalman filter preliminaries

Consider the following discrete-time linear process:

$$x_{k+1} = Ax_k + w_k, \quad y_k = Cx_k + v_k, \tag{1}$$

where the state vector of the system at time k is $x_k \in \mathbb{R}^n$, and the noisy measurement obtained by the sensor is $y_k \in \mathbb{R}^m$. The process noise $w_k \in \mathbb{R}^n$ and the measurement noise $v_k \in \mathbb{R}^m$ are mutually independent zero-mean i.i.d Gaussian random processes with $\mathbb{E}[w_k w_j^\top] = \delta_{kj} Q$ ($Q \geq 0$), $\mathbb{E}[v_k v_j^\top] = \delta_{kj} R$ ($R > 0$), and $\mathbb{E}[w_k v_j^\top] = 0$, $\forall j, k$. We assume that the initial state x_0 is a zero-mean Gaussian random vector with covariance $\Sigma_0 \geq 0$, and it is uncorrelated with w_k and v_k. It is further assumed that the time-invariant pair (A, C) is detectable and $(A, \sqrt{Q})$ is stabilizable.

A smart sensor (Hovareshti, Gupta, & Baras, 2007) is adopted in Fig. 1: instead of sending the raw measurements y_0^k directly, the sensor in Fig. 1 computes the optimal estimate of state x_k by running a Kalman filter locally. The obtained minimum mean-squared error (MMSE) estimate of the process state is given by $\hat{x}_k^1 = \mathbb{E}[x_k | y_0^k]$, with its corresponding estimation error covariance $P_k^1 \triangleq \mathbb{E}[(x_k - \hat{x}_k^1)(x_k - \hat{x}_k^1)^\top | y_0^k]$. Intuitively, a local estimate contains all the "useful" information of the historical measurements, which can lead to better performance. This is indeed verified in Gupta, Hassibi, and Murray (2007), where sending the estimate results in better estimation performance at the receiver compared to sending the measurements (all else being equal).

These terms are computed recursively by means of a Kalman filter (Anderson & Moore, 2012). For notational simplicity, we define the Lyapunov and Riccati operators h and $\tilde{g}$: $\mathbb{S}_+^n \to \mathbb{S}_+^n$ as

$$h(X) \triangleq AXA^\top + Q, \tilde{g}(X) \triangleq X - XC^\top[CXC^\top + R]^{-1}CX.$$

From the detectability and stabilizability assumption, the estimation error covariance P_k^1 converges exponentially to a unique fixed point $\overline{P}$ of $\tilde{g} \circ h$ (Anderson & Moore, 2012). Without loss of generality, we ignore the transient periods and assume that the Kalman filter at the sensor has entered steady state; i.e.,

$$P_k^1 = \overline{P}, \;\; k \geq 1. \tag{2}$$

The steady-state error covariance $\overline{P}$ has the following property (Li et al., 2015): for $0 \leq t_1 < t_2$,

$$\mathtt{Tr}[\overline{P}] \leq \mathtt{Tr}[h^{t_1}(\overline{P})] < \mathtt{Tr}[h^{t_2}(\overline{P})]. \tag{3}$$

2.2. Communication model and anti-jamming scheme

As demonstrated in Fig. 1, the sensor transmits the local estimate $\hat{x}_k^1$ as a data packet to the remote estimator through a scalar dropout channel, which is vulnerable to jamming attacks. By emitting high-power signals to occupy the communication channel, the attacker is able to sabotage the state information delivery, and hence degrade the estimation quality.

With limited power supplies, each time the sensor intelligently selects the transmission power, denoted by $a_{1,k}$, from the value set $\mathbb{E}_1$. Analogously, the attacker chooses the jamming power taking account of the energy-consumption constraints. Let $a_{2,k} \in \mathbb{E}_2$ denote the power choice made by the attacker at time k. We consider finite discrete value sets; namely, $\mathbb{E}_1 = \{o_1^{(1)}, \ldots, o_1^{(m)}\}$ and $\mathbb{E}_2 = \{o_2^{(1)}, \ldots, o_2^{(n)}\}$ with $o_1^{(1)} \leq \cdots \leq o_1^{(m)}$ and $o_2^{(1)} \leq \cdots \leq o_2^{(n)}$. For example, if $o_1^{(1)} = 0$ and $o_2^{(1)} = 0$, then it is possible that the sensor is at inactive status and no jamming attack is launched.

Suppose that the point-to-point communication network is a memoryless lossy channel. We characterize the packet arrival by a binary random process (Bernoulli process) η_k with $\eta_k = 0$ representing the occurrence of packet loss. Considering the influence imposed by the attacker, we adopt a general function $q(\cdot)$ to characterize the packet arrival rate:

$$\mathtt{Pr}(\eta_k = 1 | a_k^e = a^e) \triangleq q(a^e), \tag{4}$$

where we denote $a_k^e \triangleq (a_{1,k}, a_{2,k})$ as the energy pair selected by the sensor and the attacker. Generally, $q(\cdot)$ is non-decreasing in $a_{1,k}$ and non-increasing in $a_{2,k}$, and its specific form depends on the channel model, modulation and coding techniques used (Tse & Viswanath, 2005).[1] Note that the packet-loss information (i.e.,

[1] For example, the function $q(\cdot)$ adopts the form $q(x_1, x_2) = 1 - \frac{1}{2}(1 - \sqrt{\frac{x}{1+x}})$ for Rayleigh fading scalar channel, in which $x \triangleq L\frac{h_1 x_1}{h_2 x_2 + n_0}$ with channel parameters L, h_1, h_2 and n_0.

η_k), as depicted in Fig. 1, will be causally sent to the sensor via a short ACK frame through a reliable feedback channel. This scenario is typical. For example, the transmission control protocol (TCP) adopts ACK mechanism to achieve transmission reliability and provide flow control. Moreover, equipped with jamming antennas, a powerful attacker is capable of capturing the ACK information by channel eavesdropping technologies and meanwhile use this information to adaptively launch pertinent random noises to override transmitted packets (Zhang et al., 2015).

Next, we present our anti-jamming scheme, which is two-fold:

- *Transmission scheduling.* With the collected ACK information, the sensor can develop a comprehensive understanding of the receipt of packets at the estimator, and then elaborate a real-time transmission schedule resilient to jamming attacks, i.e., $a_{1,k}$ depends on the previous ACK sequence η_0^{k-1} (Li et al., 2015).
- *Defensive deception.* Since the behavior of the attacker depends on the real-time information (ACKs), the sensor will take actions to deceive the attacker into developing a false belief of ACKs and further mitigate the damage of jamming attacks. As for general communication protocols, the event $\eta_k = 0$ represents packet loss, which is common knowledge shared by the three agents (i.e., the sensor, the estimator and the attacker). A "trick" is played by the sensor and the estimator to confuse the attacker: the sensor inserts an additional bit a_k^c containing an ACK-reverse instruction into the preamble and transmits it reliably[2] to the estimator simultaneously when sending the packet $\hat{x}_k^1$; then, the estimator sends back the modified ACK, denoted by $\tilde{\eta}_k$:

  $$\tilde{\eta}_k = \eta_k \oplus a_k^c, \;\; a_k^c \in \mathbb{C} \triangleq \{0, 1\}, \tag{5}$$

  where $\oplus$ represents the XOR operation. For instance, if the packet is lost and $a_k^c = 1$, the attacker overhears the fake ACK $\tilde{\eta}_k = 1$ and then believes that the packet has been received successfully. On the other hand, the real information η_k can be obtained by the sensor.

Note that the transmission scheduling is a well-studied defense reacting to jamming attacks in remote state estimation (Li et al., 2015), and motivated by the deceptive defense concept in cyber security (Pawlick et al., 2017), we introduce ACK-deception as an emerging defense technology. Different from traditional encryption techniques focusing on information hiding, defensive deception may disclose true information and fake information selectively to protect the crucial information (i.e., η_k) and mislead the attacker at the same time. Note that the disclosure of true information may make the deception scheme more convincing to the attacker. Moreover, it might be difficult to implement expensive protection of ACKs through encryption techniques, since off-the-shelf sensors in CPSs have limited resources. In general, encryption algorithms require additional overheads other than computational resources. For example, the use of public key cryptography requires setting up, sharing and maintenance of public key infrastructure, which consumes extra communication resources and program memory and may lead to a prohibitive cost in Rifa-Pous and Herrera-Joancomartí (2011).

2.3. Remote estimation

Let $\hat{x}_k$ denote the MMSE estimate of the process x_k generated by the remote estimator, with error covariance matrix P_k. Similar to Ding et al. (2017), a simple recursion of $\hat{x}_k$ is obtained given by $\hat{x}_k = \eta_k \hat{x}_k^1 + (1-\eta_k)A\hat{x}_{k-1}$.

Moreover, the error covariance P_k at time k is

$$\begin{aligned} P_k &\triangleq \mathbb{E}[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^\top] \\ &= \begin{cases} \overline{P}, & \eta_k = 1, \\ h(P_{k-1}), & \text{otherwise}, \end{cases} \end{aligned} \tag{6}$$

where $\overline{P}$ stands for the steady-state error covariance defined in (2). For notational brevity, we define a random variable $s_k \in \mathbb{Z}$ as the holding time[3]:

$$s_k \triangleq k - \max_{0 \le l \le k}\{l : \eta_l = 1\}, \tag{7}$$

which represents the intervals between the present moment k and the most recent time that the data packet has been successfully received by the estimator. Based on (6), it is easy to obtain that $P_k = h^{s_k}(\overline{P})$, and the iteration of the holding time, $s_k = (1-\eta_k)(s_{k-1}+1)$. Without loss of generality, we suppose that the initial packet $\hat{x}_0^1$ is obtained by the estimator, i.e., $P_0 = \overline{P}$ and $s_0 = 0$. Hence, for any given time k, s_k takes values from the countable set $\mathbb{S}_k = \{s_k : 0, 1, 2, \ldots, k\}$.

At time k, provided the pair of jamming power and transmission power $\{a_{1,k}, a_{2,k}\}$, the evolution of P_k (or equivalently s_k) can be described using a Markov chain. Here, we define the state of the Markov chain as the holding time s_k, and the transition law among the states is characterized by a transition probability matrix:

$$\mathbb{T}(a_{1,k}, a_{2,k}) = \begin{pmatrix} q_k & 1-q_k & & \\ q_k & & 1-q_k & \\ \vdots & & & \ddots \end{pmatrix}, \tag{8}$$

where the entry $\mathbb{T}_{i,j}$ represents the transition probability from state $s_k = i$ to $s_{k+1} = j$, and the other default entries are 0. Notice that the probability $q_k \triangleq q(a_{1,k}, a_{2,k})$ according to (4).

Notice that in practice by employing low-energy microcontrollers and limited random access memory (RAM), the computational and memory of wireless nodes are typically restricted. These limitations preclude the sensor (or the attacker) from adopting innumerable states (i.e., $\mathbb{S}_k$ with $k = \infty$) to generate its transmission and cheating schemes (or jamming schemes). To circumvent this, we truncate the state space $\mathbb{S}_{k=\infty}$ as $\mathbb{S} \triangleq \{s_k : 0, \ldots, N\}$, in which the final state N represents all the state $s_k \ge N$. For the simplified problem, the effect of the truncation on the system performance, measured by the estimation gap $D(N) \triangleq \sum_{k=0}^{+\infty} |\mathtt{Tr}[\mathrm{E}(P_k|\mathbb{S}_k)] - \mathtt{Tr}[\mathrm{E}(P_k|\mathbb{S})]|$, is ignorable. The reasons are as follows. Here, we consider the scenario in which the non-truncated Markov chain is bounded[4] under the sensor's transmission and attacker's jamming strategies. That is, $\sum_{k=0}^{+\infty} |\mathtt{Tr}[\mathrm{E}(P_k|\mathbb{S}_k)] < \infty$. Moreover, we have $\mathtt{Tr}[\mathrm{E}(P_k|\mathbb{S}_k)] - \mathtt{Tr}[\mathrm{E}(P_k|\mathbb{S})] = 0$ for any $k < N$ and otherwise greater than zero based on (3). Hence, $D(N) \le \sum_{k=0}^{+\infty} \mathtt{Tr}[\mathrm{E}(P_k|\mathbb{S}_k)]$ goes to zero as $N \to \infty$.

As for the attacker, by processing its collected information $\tilde{\eta}_k$ following (7), it can obtain the manipulated holding time denoted

[2] Due to its simple one-bit structure, we shall assume that the ACK information can be transmitted reliably by an error correction coding (ECC) with a sufficient coding rate (Kurose & Ross, 2012). Under the ACK reliability assumption, this work provides a benchmark to analyze the asymmetric information structure between the sensor and the attacker, specifically when the sensor has full knowledge of the ACK information. Whereas, the analysis under ACK dropout is more complicated and lies beyond the scope of this paper.

[3] In the rest of this paper, we will omit the subscript of s_k when the underlying time index k is obvious from the context; when it is ambiguous, the subscript will be included.

[4] If the accumulated estimation performance is unbounded under a pair of policies adopted by the sensor and the attacker, the attacker can dominate the game trivially by using the corresponding jamming scheme to obtain an unbounded benefit.

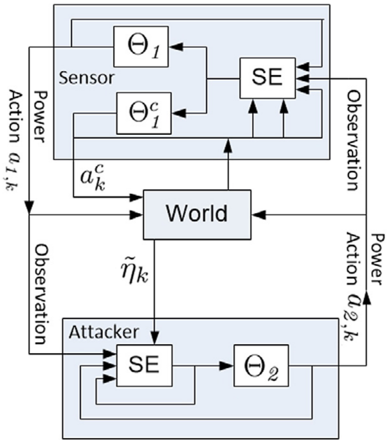


Fig. 2. Dynamic game with asymmetric information.

by $\tilde{s}_k$. Therefore, the sensor has knowledge about the ongoing state information. However, due to the deception "tricks", the attacker is uncertain about the current state, which induces the aforementioned information asymmetry.

2.4. Problem of interest

As mentioned previously, the attacker, uses the feedback information (i.e., ACKs) to deteriorate the estimation accuracy by disrupting the forward communication channel between the sensor and the remote estimator; while, the sensor aimed at alleviating reactive jamming attacks, adopts an ACK deception-based transmission strategy. Notice that, in the previous discussion the attacker has no precise information about the underlying states, which, however, can be captured by the sensor. With these considerations, a dynamic game formulation with an asymmetric information structure can be formulated. In this work, we focus on the game solution (i.e., the equilibrium point). As the asymmetric information structure raises difficulties to the existence of the optimal strategies and their calculation, we develop a belief-based stochastic game with symmetric information and translate the design of optimal strategies for the original asymmetric-information game to this equivalent one.

3. Dynamic sensor–attacker game with cheating information

A dynamic game is played between the sensor and the attacker with the property that, at each time k, two players simultaneously select actions that will be revealed at the end of time k. Notice that the sensor can access more information about the state than the attacker does, in particular, the accurate ACKs. In each iteration of the game, if the jammer's power allocation is pre-determined, then the game degenerates to a classical transmission scheduling problem, which is solved unilaterally from the perspective of the sensor based on Markov decision process (MDP) theory (see Ding et al. (2016)). On the other hand, since the attacker cannot directly obtain the true state with the involvement of cheating actions, we can formulate the decision making process at the attacker's side as a partially observable Markov decision process (POMDP). Synthesizing each side, however, the problem remains difficult as the strategy interaction depends on the choice of two agents with different information sets.

To overcome the aforementioned difficulty, we can capitalize on the conventional technology in the POMDP via generating its equivalent belief-state MDP problem. Specifically, the attacker, based on its observations, develops a probability distribution over the underlying state, called as belief $\mathbf{B}$. As illustrated in Fig. 2, at each stage, the belief is updated by the state estimation (SE) device from $\mathbf{B}$ to the next one denoted by $\mathbf{B}'$, where the transition probability is a function of the current belief and of the observed signals (including the actions selected by the two agents and the modified ACKs). Notice that the belief state $\mathbf{B}$ is a sufficient statistic. By viewing the belief $\mathbf{B}$ as the new state, we can formulate an equivalent MDP with continuous state space and study its solution. Regarding this dynamic sensor–attacker game, we can also transform the original stochastic game with an asymmetric information structure into an symmetric-information Markov game in a similar way. Notice that in the transformed game both the state space and the action space are probability measure spaces, which causes difficulties in analyzing its solution, i.e., the stationary Nash equilibrium (SNE). The main existence result of SNE is given in Theorem 3.2, and a tractable implementation is provided to find the SNE for each player.

3.1. Belief-based stochastic game definition

We introduce a belief-based sensor–attacker cheating game, which is characterized by a quintuplet: $\mathcal{G}^S \triangleq (\mathcal{I}, \mathcal{B}, \mathcal{A}, \mathbf{Q}, r)$; the specific components are defined in the following:

3.1.1. Player

$\mathcal{I} = \{0, 1, \ldots, N+1\}$ is the set of generalized players. The private information collected by the sensor is the holding time s_k, which is called *type* of the sensor (as it is closely related to the objective function of the sensor). Here, we treat each *type* of the sensor as a temporary player/agent.[5] By labeling these agents, $i = N+1$ represents the attacker, and the others are the *type agents* for the sensor, who share the same preferences of the sensor. Furthermore, assume each agent $i \in \mathcal{I}$ is rational.

At the beginning of the game, all the *type agents* plan their *ex ante* strategies,[6] and a temporary agent i is responsible for choosing the action for its original player (i.e., the sensor) when the process of game reaches an interim status (i.e., $s_k = i$). As a whole, at each stage, the sensor will be informed of the *type* (i.e., s_k) and adopted *type-contingent* strategies accordingly.

3.1.2. Belief state space

$\mathcal{B} = \Delta(\mathbb{S})$ represents the continuous belief state space, which is a collection of probability distributions over $\mathbb{S} = \{0, \ldots, N\}$. To be more specific, we denote by $\mathbf{B}_k(s_k = m)$ the probability that s_k equals m. As $\mathbf{B}_k$ is common knowledge shared by all players, we can develop a behavioral strategy (b.s.) for each player based on this shared information structure, which overcomes the asymmetry within the Markov-chain state s_k in the original game. Let $\mathcal{B}$ be endowed with the topology of weak convergence, then it is a Polish space (i.e., a complete and separable metric space, see Billingsley (2013)).

3.1.3. Action

$\mathcal{A} = \{\mathcal{A}_i, i \in \mathcal{I}\}$ denotes the joint action space. For each *type agent* $i \in \mathcal{S}$, they share the same action set $\mathcal{A}_i = \Delta(\mathbb{E}_s \times \mathbb{C})$. The action set for the attacker is $\mathcal{A}_{N+1} = \Delta(\mathbb{E}_a)$. We denote by $\mathbf{A}_{i,k} \in \mathcal{A}_i$ the action played by player i at stage k: the *type agent* $i \in \{0, \ldots, N\}$ selects the transmission energy $a_{1,k} \in \mathbb{E}_1$ and chooses cheating action $a_k^c \in \mathbb{C}$ simultaneously w.p. $\mathbf{A}_{i,k}(a_{1,k}, a_k^c)$;

[5] This representation refers to the agent-normal form proposed by Selten to cope with the possible information states of the original players (i.e., the senor and the attacker), see Haurie, Krawczyk, and Zaccour (2012) for details.

[6] *Ex ante* means that a player makes a decision before knowing the particular actions of other players. Interested readers are referred to Pearce (1982) for more details.

$$\begin{aligned}\mathbf{B}_k^+(s^+) &\triangleq \mathtt{Pr}(s_k = m^+|\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a}, a_k^e = a^e, \tilde{\eta}_k = \eta) \\ &= \frac{\mathtt{Pr}(s_k = m^+, a_k^e = a^e, \tilde{\eta}_k = \eta|\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})}{\sum_{m=0}^N \mathtt{Pr}(s_k = m, a_k^e = a^e, \tilde{\eta}_k = \eta|\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})} \\ &= \frac{\mathtt{Pr}(a_k^e = a^e, \tilde{\eta}_k = \eta|s_k = m^+, \mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})\mathtt{Pr}(s_k = m^+|\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})}{\sum_{m=0}^N \mathtt{Pr}(a_k^e = a^e, \tilde{\eta}_k = \eta|s_k = m, \mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})\mathtt{Pr}(s_k = m|\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})} \\ &= \frac{\psi_1(a_k^e = a^e, \tilde{\eta}_k = \eta, s_k = m^+, \mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})\mathbf{b}(m^+)}{\sum_{m=0}^N \psi_1(a_k^e = a^e, \tilde{\eta}_k = \eta, s_k = m, \mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})\mathbf{b}(m)},\end{aligned} \tag{9}$$

$$\begin{aligned}\mathbf{Q}(\mathbf{b}'|\mathbf{b}, \mathbf{a}) &= \begin{cases} \mathtt{Pr}(a_k^e = a^e, \tilde{\eta}_k = \eta|\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a}), & \text{if } \mathbf{b}' = \varphi_2(\varphi_1(\mathbf{b}, \mathbf{a}, a^e, \eta), \mathbf{a}, a^e, \eta) \text{ for some } \{a^e, \eta\} \in \mathbb{E}_1 \times \mathbb{E}_2 \times \mathbb{C} \\ 0, & \text{otherwise.} \end{cases} \\ &= \begin{cases} \sum_{m=0}^N \psi_1(a^e, \eta, m, \mathbf{b}, \mathbf{a})\mathbf{b}(m), & \text{if } \mathbf{b}' = \varphi_2(\varphi_1(\mathbf{b}, \mathbf{a}, a^e, \eta), \mathbf{a}, a^e, \eta) \text{ for some } \{a^e, \eta\} \in \mathbb{E}_1 \times \mathbb{E}_2 \times \mathbb{C} \\ 0, & \text{otherwise.} \end{cases}\end{aligned} \tag{10}$$

Box I.

and $\mathbf{A}_{N+1,k}(a_{2,k})$ indicates the probability of the jamming power $a_{2,k} \in \mathbb{E}_2$ taken by the attacker $i = N + 1$. For brevity, the joint action at stage k is denoted by $\mathbf{A}_k \triangleq \{\mathbf{A}_{0,k}, \dots, \mathbf{A}_{N+1,k}\}$. Moreover, we define $\mathbf{a} = \{\mathbf{a}_0, \dots, \mathbf{a}_{N+1}\}$ as the aggregated actions chosen by all the players. As for the belief state space, we let $\mathcal{A}_i$ be endowed with the topology of weak convergence. The metric for action space $\mathcal{A}$ is then defined as $\mathrm{d}(\mathbf{a}, \mathbf{a}') = \max_{i\in\mathcal{I}}\{\mathrm{d}_P(\mathbf{a}_i, \mathbf{a}_i')\}$, where $\mathrm{d}_P(\cdot, \cdot)$ is the Prohorov metric (Billingsley, 2013) that induces the weak convergence topology for $\mathcal{A}_i$.

3.1.4. *Transition probability*

The law of the movement for the belief state is given by a transition function: $\mathbf{Q} : \mathcal{B} \times \mathcal{A} \times \mathcal{B} \Longrightarrow [0, 1]$ with the transition probability: $\mathbf{Q}(\mathbf{b}'|\mathbf{b}, \mathbf{a}) \triangleq \mathtt{Pr}(\mathbf{B}_{k+1} = \mathbf{b}'|\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a})$. That is, $\mathbf{Q}(\cdot|\cdot)$ describes the probability of the next belief state given the current belief state and the joint action.

The update of the belief state, involving the **correction** and **prediction** steps, is given as follows.

Correction: At each stage k, based on the probabilistic action $\mathbf{A}_k$, player i chooses the energy level (with/without the cheating action) randomly. Notice that, the attacker knows that the ACKs may be fake, and the joint energy $a_k^e \triangleq \{a_{1,k}, a_{2,k}\}$ is assumed to be monitored perfectly by each player. Thereafter, conditional on a_k^e and the collected ACK signal $\tilde{\eta}_k$, the attacker is capable to correct its *a priori* probability distribution $\mathbf{B}_k$ following the Bayes' rule. The corrected belief state, denoted by $\mathbf{B}_k^+ \triangleq \varphi_1(\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a}, a_k^e = a^e, \tilde{\eta}_k = \eta)$, is computed in (9) given in Box I.The function $\psi_1(a_k^e, \tilde{\eta}, s_k, \mathbf{B}_k, \mathbf{A}_k)$ in (9) is defined as

$$\begin{aligned}&\psi_1(a^e, \eta, m^+, \mathbf{b}, \mathbf{a}) \\ &\triangleq \mathtt{Pr}(a_k^e = a^e, \tilde{\eta}_k = \eta|s_k = m^+, \mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a}) \\ &= \mathtt{Pr}(a_k^e = a^e|s_k = m^+, \mathbf{A}_k = \mathbf{a})\cdot \\ &\quad \mathtt{Pr}(\tilde{\eta}_k = \eta|a_k^e = a^e, s_k = m^+, \mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a}) \\ &= \mathbf{a}_{N+1}(a_{2,k} = a_2)\sum_{a^c\in\mathbb{C}} \mathbf{a}_{m^+}(a_{1,k} = a_1, a_k^c = a^c)\cdot \\ &\quad \Bigg[\frac{\mathtt{Pr}(\eta_k = \eta|a_k^e = a^e)\mathbf{a}_m(a_{1,k} = a_1, a_k^c = 0)}{\sum_{a^c\in\mathbb{C}} \mathbf{a}_{m^+}(a_{1,k} = a_1, a_k^c = a^c)} \\ &\quad + \frac{\mathtt{Pr}(\eta_k = \eta \oplus 1|a_k^e = a^e)\mathbf{a}_m(a_{1,k} = a_1, a_k^c = 1)}{\sum_{a^c\in\mathbb{C}} \mathbf{a}_{m^+}(a_{1,k} = a_1, a_k^c = a^c)}\Bigg] \\ &= \mathbf{a}_{N+1}(a_{2,k} = a_2)\big[\mathtt{Pr}(\eta_k = \eta|a_k^e = a^e)\mathbf{a}_{m^+}(a_1, 0) \\ &\quad + \mathtt{Pr}(\eta_k = \eta \oplus 1|a_k^e = a^e)\mathbf{a}_{m^+}(a_1, 1)\big],\end{aligned} \tag{11}$$

in which $\mathbf{a}_m(a_{1,k} = a_1, a_k^c = a^c)$, for example, indicates the probability that transmission energy a_1 and cheating action a^c are adopted simultaneously by player m. Obviously, the probability $\mathtt{Pr}(\eta_k = \eta|a_k^e = a^e)$ can be calculated according to the packet-drop rate $q(a_k^e)$ in (4). The third equation is derived based on the assumption that each player takes actions independently.

Prediction: Based on the *a posterior* belief state $\mathbf{b}^+$ and the observations $\{a^e, \eta\}$, we can predict the probability over the original state s_{k+1} for the next stage: $\mathbf{B}_{k+1}(m) \triangleq \mathtt{Pr}(s_{k+1} = m|\mathbf{B}_k^+ = \mathbf{b}^+, \mathbf{A}_k = \mathbf{a}, a_k^e = a^e, \tilde{\eta}_k = \eta)$. As mentioned in Section 2.3, the transition of the original state s_k is deterministic provided the ACK information η_k. With the absence of η_k, the attacker generates a corrected transition probability (or prediction) matrix based on the modified ACK $\tilde{\eta}_k$:

$$\tilde{\mathbb{T}}(\mathbf{b}^+, \mathbf{a}, a^e, \eta) = \begin{pmatrix} \tilde{q}(0) & 1 - \tilde{q}(0) & & \\ \tilde{q}(1) & & 1 - \tilde{q}(1) & \\ \vdots & & & \ddots \end{pmatrix}, \tag{12}$$

in which $\tilde{q}(m) = (1 - \eta)\psi_2(a_k^c = 1, s_k = m, \mathbf{B}_k^+ = \mathbf{b}^+, \mathbf{A}_k = \mathbf{a}, a_k^e = a^e, \tilde{\eta} = \eta) + \eta\psi_2(0, m, \mathbf{b}^+, \mathbf{a}, a^e, \eta)$. Here, the function $\psi_2(a_k^c, s_k, \mathbf{B}_k^+, \mathbf{A}_k, a_k^e, \tilde{\eta})$ is:

$$\begin{aligned}&\psi_2(a^c, m, \mathbf{b}^+, \mathbf{a}, a^e, \eta) \\ &\triangleq \mathtt{Pr}(a_k^c = a^c|s_k = m, \mathbf{B}_k = \mathbf{b}^+, \mathbf{A}_k = \mathbf{a}, a_k^e = a^e, \tilde{\eta} = \eta) \\ &= \frac{\mathtt{Pr}(\eta_k = \eta \oplus a^c|a_k^e = a^e)\mathbf{a}_m(a_{1,k} = a_1, a_k^c = a^c)}{\sum\limits_{\tilde{a}^c\in\mathbb{C}} \mathtt{Pr}(\eta_k = \eta \oplus \tilde{a}^c|a_k^e = a^e)\mathbf{a}_m(a_{1,k} = a_1, a_k^c = \tilde{a}^c)}.\end{aligned}$$

Then, the next belief state is

$$\begin{aligned}\mathbf{B}_{k+1} &\triangleq \varphi_2(\mathbf{B}_k^+ = \mathbf{b}^+, \mathbf{A}_k = \mathbf{a}, a_k^e = a^e, \tilde{\eta} = \eta) \\ &= \mathbf{B}_k^+\tilde{\mathbb{T}}(\mathbf{b}^+, \mathbf{a}, a^e, \eta).\end{aligned} \tag{13}$$

Therefore, the belief state $\mathbf{B}_k$ transits deterministically given the public observations, and the explicit representation of the transition probability is given in (10). Suppose that the initial state m_0 is known by the two original players, therefore the initial belief state is $\mathbf{B}_0(s_1) = \mathbf{1}_{m_0}(s_1)$.

3.1.5. *Payoff*

Let $r_i : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$ denote the one-stage reward function for each player $i \in \mathcal{I}$. The sensor attempts to improve the estimation quality at the remote estimator without wasting energy. Hence, for $i \leq N$, we have:

$$r_i(\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a}) \tag{14}$$
$$\triangleq \sum_{a_1 \in \mathbb{E}_1} \sum_{a_2 \in \mathbb{E}_2} \mathbf{a}_i(a_1)\mathbf{a}_{N+1}(a_2)[u_i(a_1, a_2) - \delta_1 a_1],$$

in which $u_i(a_1, a_2) \triangleq -\mathtt{Tr}[q(a_1, a_2)\overline{P} + (1 - q(\cdot))h^{i+1}(\overline{P})]$ represents the estimation performance, a_1 is the transmission energy consumed by player i, and $\delta_1 \geq 0$ represents the proportion of the energy term in the reward function. Moreover, $\mathbf{a}_i(a_1) \triangleq \sum_{a^c \in \mathbb{C}} \mathbf{a}_i(a_{1,k} = a_1, a^c_k = a^c)$. The one-stage reward function for the attacker is:

$$r_{N+1}(\mathbf{B}_k = \mathbf{b}, \mathbf{A}_k = \mathbf{a}) \tag{15}$$
$$\triangleq \sum_{a_1} \sum_{a_2} \sum_{i=0}^{N} \mathbf{a}_i(a_1)\mathbf{a}_{N+1}(a_2)\mathbf{b}(i)[-u_i(a_1, a_2) - \delta_2 a_2],$$

in which $\delta_2 \geq 0$ is the weight parameter. Notice that the cheating action a^c does not affect the reward functions explicitly; it impacts the reward of the attacker indirectly through tampering the feedback information and further disturbing the belief state developed by the attacker.

Hence, we formulate the interaction between the sensor and the attacker under deception actions as a stochastic game. To solve this coupled optimization problem, we focus on *stationary strategies*, which are defined as time-independent mappings from the belief state space into the players' actions: $\pi : \mathcal{B} \to \mathcal{A}$. Suppose $\mathbf{b}$ is the belief state at time k. When adopting the joint strategy π, we specify the probability $\pi_i(\mathbf{b})_{[a,a^c]}$ of taking energy choice $a \in \mathbb{E}_1$ and deception choice $a^c \in \mathbb{C}$ jointly by *type agent* $i \in \{0, \ldots, N\}$ at state $\mathbf{b}$, and also define by $\pi_i(\mathbf{b})_{[a]}$ the probability of interference energy $a \in \mathbb{E}_2$ played by the attacker $i = N + 1$.

Considering future effects brought by the current actions, the infinite-time discounted payoff for player i under the stationary strategy π is:

$$\mathcal{J}_i(\mathbf{b}_0, \pi) = \sum_{k=0}^{\infty} \delta^k r_i(\mathbf{B}_k = \mathbf{b}, \pi(\mathbf{b})), \ \ i \in \{0, \ldots, N + 1\}, \tag{16}$$

in which $\delta \in [0, 1)$ is the discount factor. The generalized players will take a joint policy π such that a long-term performance objective is maximized for each player.

3.2. Equilibrium analysis

We next study the equilibrium solution of the stochastic game $\mathcal{G}^S$ introduced in Section 3.1. As mentioned previously, we limit our attention to the set of *stationary strategies*. The *stationary Nash equilibrium* (SNE) is defined as follows:

Definition 3.1. For the stochastic game $\mathcal{G}^S$, a stationary policy $\pi^\star_i, \forall i \in \mathcal{I}$ is a stationary Nash equilibrium if and only if no player can unilaterally improve his expected payoff by deviating his equilibrium strategy; namely, for all player $i \in \mathcal{I}$ and any initial state $\mathbf{B}_0 \in \mathcal{B}$,

$$\mathcal{J}^\star_i \triangleq \mathcal{J}_i(\mathbf{B}_0, [\pi^\star_i, \pi^\star_{-i}]) \geq \mathcal{J}_i(\mathbf{B}_0, [\psi(\pi^\star_i), \pi^\star_{-i}]), \tag{17}$$

in which $\psi(\pi^\star_i)$ represents a meta-strategy when player i is suggested to take $\pi^\star_i$ at the equilibrium point and $\mathcal{J}^\star_i$ is the corresponding game value for the ith player. ■

Many previous works have investigated the existence of SNE in stochastic game with a finite number of states and actions. However, our stochastic game $\mathcal{G}^S$ is with continuous state space and action set, of which the existence of SNE is much harder to analyze. Now, we show the main result about SNE existence for $\mathcal{G}^S$ in the following theorem.

Theorem 3.2. *The game $\mathcal{G}^S$ has a stationary Nash equilibrium.*

Proof. By Sobel (1973), in order to prove Theorem 3.2, it is sufficient to verify the following conditions.
C1 (State Space): $\mathcal{B}$ is a compact metric space.
C2 (Action Space): $\mathcal{A}_i$ is a compact metric space for every $i \in \mathcal{I}$.
C3 (Reward Functions): $r_i(\cdot, \cdot)$ is continuous on $\mathcal{B} \times \mathcal{A}$ for every $i \in \mathcal{I}$.
C4 (Transition Probability): $\mathbf{Q}$ is weakly continuous on $\mathcal{B}\times\mathcal{A}$, i.e., if $(\mathbf{b}^n, \mathbf{a}^n) \to (\mathbf{b}, \mathbf{a})$, then $\mathbf{Q}(\cdot|\mathbf{b}^n, \mathbf{a}^n)$ converges weakly[7] to $\mathbf{Q}(\cdot|b, \mathbf{a})$.

We next verify the conditions one by one.
C1 and **C2**: Since $\mathcal{B}$ and $\mathcal{A}_i$ all are probability measure spaces on a finite set, then by Billingsley (2013, Theorem 6.4), they are compact metric spaces.
C3: For probability measures $\mu, \mu^n, n \in \mathbb{N}$, we write $\mu^n \overset{w}{\to} \mu$ if μ^n converges weakly to μ. By the definition of the metric defined for the action space $\mathcal{A}$, one sees that as $\mathbf{a}^n \to \mathbf{a}$, $\mathbf{a}^n_i \overset{w}{\to} \mathbf{a}_i, \forall i \in \mathcal{I}$ holds. Since either $\mathbb{S}, \mathbb{E}_1 \times \mathbb{C}, \mathbb{E}_2$ is a finite set, of which each subset is a continuity set, then by the Portmanteau Theorem in Billingsley (2013), one obtains that for $0 \leq i \leq N$:

$$\begin{aligned} \mathbf{a}^n_i \overset{w}{\to} \mathbf{a}_i &\iff \mathbf{a}^n_i(\alpha) \to \mathbf{a}_i(\alpha), \forall \alpha \in \mathbb{E}_1 \times \mathbb{C}, \\ \mathbf{a}^n_{N+1} \overset{w}{\to} \mathbf{a}_{N+1} &\iff \mathbf{a}^n_{N+1}(\alpha) \to \mathbf{a}_{N+1}(\alpha), \forall \alpha \in \mathbb{E}_2, \\ \mathbf{b}^n \overset{w}{\to} \mathbf{b} &\iff \mathbf{b}^n(i) \to \mathbf{b}(i), \forall 0 \leq i \leq N, \end{aligned}$$

where $\iff$ means equivalence. The dominated convergence theorem yields that, as $(\mathbf{b}^n, \mathbf{a}^n) \to (\mathbf{b}, \mathbf{a})$, $r_i(\mathbf{b}^n, \mathbf{a}^n) \to r_i(\mathbf{b}, \mathbf{a})$ for every $i \in \mathcal{I}$. The continuity of reward functions is thus verified.

C4: Notice that given the current state $\mathbf{b}$ and action $\mathbf{a}$, the possible values of the next state are finite. Again by the Portmanteau Theorem, to verify this condition, it suffices to prove that for any $\{a^e, \eta\} \in \mathbb{E}_1 \times \mathbb{E}_2 \times \mathbb{C}$, if $(\mathbf{b}^n, \mathbf{a}^n) \to (\mathbf{b}, \mathbf{a})$, then

$$\varphi_2(\varphi_1(\mathbf{b}^n, \mathbf{a}^n, a^e, \eta), \mathbf{a}^n, a^e, \eta) \overset{w}{\to} \varphi_2(\varphi_1(\mathbf{b}, \mathbf{a}, a^e, \eta), \mathbf{a}, a^e, \eta),$$
$$\sum_{m=0}^{N} \psi_1(a^e, \eta, m, \mathbf{b}^n, \mathbf{a}^n)\mathbf{b}^n(m) \to \sum_{m=0}^{N} \psi_1(a^e, \eta, m, \mathbf{b}, \mathbf{a})\mathbf{b}(m).$$

This can be done using similar arguments to those employed for the previous **C3** verification. ■

Notice that, the original asymmetric game is transformed to an equivalent symmetric one $\mathcal{G}^S$ and the corresponding stationary Nash equilibrium is also a solution of the original one.

We now outline the computation of the channel power and the estimation performance when the game is in the stationary Nash equilibrium schemes $\pi^\star$. Notice that under the stationary equilibrium, the sequence of random variables $\{\mathbf{B}_k, k \geq 0\}$ establishes a controlled Markov chain with the transition function denoted by $\mathbf{Q}^\pi(\mathbf{b}'|\mathbf{b})$. For a belief state $\mathbf{b}$, we denote by $r^e(\mathbf{b})$ and $r^p(\mathbf{b})$ the expected estimation error covariance and the expected transmission power for the sensor. Hence, we have

$$r^e(\mathbf{b}) \triangleq \sum_{i=0}^{N} \mathbf{b}(i)\mathtt{Tr}[h^i(\overline{P})],$$
$$r^p(\mathbf{b}) \triangleq \sum_{i=0}^{N} \mathbf{b}(i) \sum_{a_i \in \mathbb{E}_1} a_i \sum_{a^c \in \mathbb{C}} \pi(\mathbf{b})_{[a_i,a^c]}.$$

With a slight abuse of notation, we call $\mathbf{Q}^\pi$, r^e and r^p the transition probability matrix, the estimation vector and the power vector

[7] Interested readers are referred to Billingsley (2013) for more details of weak convergence of probability measures.

corresponding to the stationary equilibrium $\pi^\star$. According to Taylor (2012, Theorem 7.1), the expected total discounted estimation error covariance and transmission power of the equilibrium schemes can be calculated using the formulas

$$\mathrm{E}^{\pi^\star}[\sum_{k=0}^{\infty} \delta^k \mathrm{Tr}[P_k]] = (I - \delta \mathbf{Q}^{\pi})^{-1} r^e,$$

$$\mathrm{E}^{\pi^\star}[\sum_{k=0}^{\infty} \delta^k a_{1,k}] = (I - \delta \mathbf{Q}^{\pi})^{-1} r^p.$$

3.3. Stability condition

Owing to the malicious jamming attacks and packet losses, the estimator cannot guarantee a successful transmission within a finite time. In this section, we will study the stability conditions under which the expected error covariance at the remote estimator will converge. According to (11), the transition probabilities of the holding time s_k (equivalently, the error covariance P_k) depend on the transmission power and jamming power, instead of the cheating actions. Denote by $\Theta_1 \triangleq \{\mathbf{Z}_{1,k}, k \geq 0\}$ (or $\Theta_2 \triangleq \{\mathbf{Z}_{2,k}, k \geq 0\}$) the transmission (or jamming) scheme for the sensor (or the attacker), in which $\mathbf{Z}_{1,k} \in \Delta(\mathbb{E}_1)$ and $\mathbf{Z}_{2,k} \in \Delta(\mathbb{E}_2)$. The corresponding equilibrium strategies for the sensor and the attacker are denoted by θ_1^{NE} and θ_2^{NE}, respectively. Consider the averaged expected estimation error covariance denoted by

$$F(\Theta_1, \Theta_2) \triangleq \lim_{T \to +\infty} \frac{1}{T} \sum_{k=0}^{T} \mathrm{E}(P_k | \Theta_1, \Theta_2).$$

Let $\rho(A)$ represent the spectral radius of A. Based on the property of A, Q and the packet-dropout rate, we have the following theorem:

Theorem 3.3. *Under the equilibrium schemes $(\theta_1^{NE}, \theta_2^{NE})$, $F(\theta_1^{NE}, \theta_2^{NE})$ converges if and only if*

$$\rho^2(A)[1 - q(a_{1,k} = o_1^{(m)}, a_{2,k} = o_2^{(n)})] < 1, \tag{18}$$

in which $o_1^{(m)}$ and $o_2^{(n)}$ correspond to the greatest transmission power and jamming power for the sensor and the attacker.

Proof. Define two special schemes (in which the estimation packets are transmitted or jammed by the sensor or the attacker with their highest power levels constantly) as $\theta_1^H \triangleq \{\mathbf{Z}_{1,k}(o_1^{(m)}) = 1, k \geq 0\}$ and $\theta_2^H \triangleq \{\mathbf{Z}_{2,k}(o_2^{(n)}) = 1, k \geq 0\}$.

Sufficiency: If (18) is satisfied, then we have $F(\theta_1^H, \theta_2^H) < \infty$ based on Ren, Cheng, Chen, Shi, and Zhang (2014, Lemma 3). For any given transmission strategies Θ_1, Θ_2^H is the worst jamming attacks and it corresponds to the largest packet-dropout rate, from (8) we have

$$F(\theta_1^H, \theta_2^H) \geq F(\theta_1^H, \theta_2^{NE}).$$

The proof is based on the majorization theory similar to the proof in our previous work (Ding et al., 2016, Theorem 4). Due to the space limitation, we ignore the details here.

Next, we prove $F(\theta_1^{NE}, \theta_2^{NE}) < +\infty$ by contradiction. If $F(\theta_1^{NE}, \theta_2^{NE})$ is unbounded, the equilibrium payoff for the sensor is $-\infty$. Notice that $F(\theta_1^H, \theta_2^{NE}) < +\infty$, and the energy-related term in the payoff function is bounded as the highest energy level $o_1^{(m)}$ is finite. Hence, adopting θ_1^H will improve the performance of the sensor if the attacker keeps its equilibrium strategy θ_2^{NE} unchanged. It contradicts the definition of equilibrium.

Necessity: For unstable systems, if (18) is not satisfied, the attacker can adopt a trivial jamming scheme (i.e., θ_2^H) to obtain an unbounded expected error covariance. Actually, it corresponds to a dominated strategy for the attacker to obtain infinite benefit no matter what defensive strategies are adopted by the sensor. That is, $\theta_2^{NE} = \theta_2^H$ and then $F(\theta_1^{NE}, \theta_2^{NE}) = \infty$ ■

3.4. Practical design

We now present an implementation of the stationary strategy of the stochastic game $\mathcal{G}^S$. To cope with the continuous state space, we first discretize the belief state space and build a look-up table about the pairs of discrete state and optimal strategy. When the actual game is in some continuous-valued state, each player executes the action w.r.t. the discretized state. We sample the state space with a regular grid (for details, see Section 4).

With a slight abuse of notation, the discretized game and its corresponding state space are also denoted by $\mathcal{G}^S$ and $\mathcal{B}$. One inevitable difficulty of the practical implementation stems from the fact that there exist multiple equilibria for the discretized game $\mathcal{G}^S$. An advanced method can be adopted to find all stationary equilibria of this game (Iskhakov, Rust, & Schjerning, 2016). Nevertheless, this approach suffers from a curse of dimensionality that originates both from an exponential increase in the number of directional components of the state space and also from the number of equilibria (which may increase with the total number of states). To circumvent this, among the set of equilibria, we focus on finding only one of them so that the optimal strategies for the sensor and the attacker can be well designed. Numerical algorithms have been proposed to solve stochastic games with finite state space, such as, Shapley value iteration, policy iteration algorithm and Newton-type methods (Haurie et al., 2012). These approaches assume that players have knowledge of the parameters of the game (i.e., the reward and transition probability functions), which unfortunately may not be available in real applications. To overcome this limitation, we present an algorithm to find the NE of $\mathcal{G}^S$ using a reinforcement learning method (Greenwald, Hall, & Serrano, 2003; Hu & Wellman, 2003). The discretized stochastic game $\mathcal{G}^S$ also satisfies the Bellman equation, that is, for a given stationary equilibrium policy $\pi^\star$, the expected payoff value (i.e., game value) for each player has the following recursive property:

$$\begin{aligned} \mathcal{J}_i^\star(\mathbf{b}) &= \mathbf{val}_i\{Q_0^\star(\mathbf{b}, \mathbf{a}), \ldots, Q_{N+1}^\star(\mathbf{b}, \mathbf{a})\}, \\ Q_i^\star(\mathbf{b}, \mathbf{a}) &= r_i(\mathbf{b}, \mathbf{a}) + \delta \sum_{\mathbf{b}' \in \mathcal{B}} \mathbf{Q}(\mathbf{b}' | \mathbf{b}, \mathbf{a}) \mathcal{J}_i^\star(\mathbf{b}'), \end{aligned} \tag{19}$$

in which $\mathbf{Q}(\mathbf{b}'|\mathbf{b}, \mathbf{a})$ is the probability of transition from the current state $\mathbf{b}$ to state $\mathbf{b}'$. The operator $\mathbf{val}_i$ computes the value of states $\mathbf{b}$ for the ith player by solving a one-stage game with parameters $\{Q_0^\star(\mathbf{b}, \mathbf{a}), \ldots, Q_{N+1}^\star(\mathbf{b}, \mathbf{a})\}$. That is, in this game, for the ith player, its payoff with respect to action $\mathbf{a}$ is $Q_i^\star(\mathbf{b}, \mathbf{a})$, and its NE strategy is $\pi^\star(\mathbf{b}) = \arg\max_{\mathbf{a}_i \in \mathcal{A}_i} Q_i^\star(\mathbf{b}, [\mathbf{a}_i, \pi_{-i}^\star(\mathbf{b})])$. Therefore, $\mathbf{val}_i\{Q_0^\star(\mathbf{b}, \mathbf{a}), \ldots, Q_{N+1}^\star(\mathbf{b}, \mathbf{a})\} = Q_i^\star(\mathbf{b}, [\pi^\star(\mathbf{b}), \pi_{-i}^\star(\mathbf{b})])$. The notion of $Q^\star(\mathbf{b}, \mathbf{a})$ represents the expected cumulative discounted reward of action $\mathbf{a}$ taken in state $\mathbf{b}$ and following the optimal policy $\pi^\star$ afterwards. Notice that the number of possible values $Q^\star(\mathbf{b}, \mathbf{a})$ is innumerable since $\mathbf{a} \in \mathcal{A}$. With some abuse of notation $Q_i^\star(\cdot)$, we define the Q-value over the pair of the discretized state and the finite choices (composed of the joint energy action pairs $a_k^e = a^e$ and the cheating action $a_k^c = a^c$) as $Q_i^\star(\mathbf{b}, a^e, a^c)$. Notice

Table 1
Summary for parameters.

System parameters				Channel		Discount
A	Q	C	R	$\mathbb{E}_s$	$\mathbb{E}_a$	δ
1.3	0.6	0.8	0.6	$\{0.5, 0.6\}$	$\{0.2, 0.4\}$	0.96

that $Q_i^\star(\mathbf{b}, \mathbf{a}) = \sum_{a^e, a^c} \mathtt{Pr}(a^e, a^c|\mathbf{b}, \mathbf{a})Q_i^\star(\mathbf{b}, a^e, a^c)$. Based on (19), we have

$$\mathcal{J}_i^\star(\mathbf{b}) = \mathbf{val}_i\{Q_0^\star(\mathbf{b}, a^e, a^c), \ldots, Q_{N+1}^\star(\mathbf{b}, a^e, a^c)\},$$
$$Q_i^\star(\mathbf{b}, a^e, a^c) = r_i(\mathbf{b}, a^e, a^c) + \delta \sum_{\mathbf{b}' \in \mathcal{B}} \overline{\mathbf{Q}}(\mathbf{b}'|\mathbf{b}, a^e, a^c)\mathcal{J}_i^\star(\mathbf{b}'),$$

in which $\overline{\mathbf{Q}}(\mathbf{b}'|\mathbf{b}, a^e, a^c)$ is given in

$$\overline{\mathbf{Q}}(\mathbf{b}'|\mathbf{b}, a^e, a^c) = \begin{cases} \mathtt{Pr}(\tilde{\eta}_k = \eta|\mathbf{b}, a^e, a^c), & \text{if } \mathbf{b}' = \varphi_2(\varphi_1(\mathbf{b}, \mathbf{a}, a^e, \eta), \mathbf{a}, a^e, \eta); \\ 0, & \text{others.} \end{cases}$$

and

$$r_i(\mathbf{b}, a^e, a^c) = \begin{cases} u_i(a_1, a_2) - \delta_1 a_1, & \text{if } i \le N; \\ \sum_{i=0}^{N} \mathbf{b}(i)[-u_i(a_1, a_2) - \delta_2 a_2], & \text{others.} \end{cases}$$

Notice that $r_i(\mathbf{b}, a^e, a^c)$ is indifferent to a^c and for simplicity we use $r_i(\mathbf{b}, a^e)$ instead in what follows.

Here, the operation $\mathbf{val}_i$ is similar to searching NE in a one-stage game, except that the optimal value $Q_i^\star(\mathbf{b}, a)$ is unknown. A reinforcement learning process is proposed to replace the sum over belief state space with a Monte Carlo approximation. To be specific, at stage k, players know the current state $\mathbf{b}$, and each possesses an evaluation function over the state–choice pairs, denoted by $Q_i^k(\mathbf{b}, a^e, a^c)$. The Q-value summarizes the learning result from past experience, and is used to estimate the model as mentioned previously. Specifically, a new random experiment is organized by each player, resulting in a pair of choices, a reward and a new state. Then, the Q-values with the actually visited states are updated via temporal difference methods. The iteration of the Q-value for each player is developed as follows

$$\mathcal{J}_i^k(\mathbf{b}) = \mathbf{val}_i\{Q_0^k(\mathbf{b}, a^e, a^c), \ldots, Q_{N+1}^k(\mathbf{b}, a^e, a^c)\}, \tag{20}$$
$$Q_i^{k+1}(\mathbf{b}, a^e, a^c) = (1 - \gamma_k)Q_i^k(\mathbf{b}, a^e, a^c) + \gamma_k[r(\mathbf{b}, a^e) + \delta \mathcal{J}_i^k(\mathbf{b}')], \tag{21}$$

where γ_k is the learning rate. To guarantee the convergence of the learning algorithm, the learning rate should satisfy two conditions (for details see Ding et al. (2017)). It is sufficient to satisfy Condition 1 by a large number of iterations and adopting random actions in the learning process. The specific design of the learning rate satisfying the decaying condition (Condition 2) is provided in the simulation part. The learning algorithm provably converges to the NE if either every stage game during learning has a globally optimal strategy or a saddle point (Hu & Wellman, 2003), which does not hold in our problem. However, these conditions are not necessary as shown in many experiments on a standard test bed of Markov games (Greenwald et al., 2003). We test this algorithm on $\mathcal{G}^S$ under different tuples of parameters, and the results all show that the Q-value will converge empirically, which previous multiagent reinforcement learning algorithms have not achieved. The algorithm for searching NE of the game $\mathcal{G}^S$ is summarized in **Algorithm 1**. Notice that $\|\cdot\|$ is a matrix norm and ϵ represents the accuracy condition.

4. Examples

In this section, we illustrate the practical implementation outlined in Section 3.4 using some examples. Consider the following scalar system with parameters shown in Table 1. Assume that the channel in Fig. 1 is wireless fast-fading channel with $q(x_1, x_2) = 1 - (\frac{x_1}{0.5x_2+0.1})^{-2}$. To reduce the computational burden, we restrict the possible values of state s_k to be finite: $s_k \in \{0, 1, 2, 3\}$. The belief state space is discretized by a regular grid with resolution rate 0.05. The learning rate is $\gamma_k = \frac{10}{15+o(\mathbf{b}, a^e, a^c)}$, in which $o(\mathbf{b}, a^e, a^c)$ is an occupation counter of the state–choice pair $(\mathbf{b}, a^e, a^c)$ from stage 0 to stage k. Some parameter settings for the (transmission/jamming) power levels and the reward function are shown in Table 1, and the weights are $\delta_1 = \delta_2 = 1$. For each learning stage, there may exist multiple Nash equilibria for the general-sum multi-player games. Our method attempts to find an equilibrium as an example instead of designing an equilibrium selection mechanism. We tested the algorithm under around 15 000 iterations. The performance of the learning algorithm is as described below.

Algorithm 1 Nash Equilibrium Q-learning algorithm

```
Initialization:
$k = 0$ and set the initial state $\mathbf{b} \in \mathcal{B}$
Initialize the Q-value $Q_i^k(\mathbf{b}, a^e, a^c)$ for all states $\mathbf{b}$ and arbitrary joint choices $\{a^e, a^c\}$, where $i \in \mathcal{I}$
While $||Q^{k+1}(\cdot) - Q^k(\cdot)|| < \epsilon$
At stage $k$, find the NE (i.e., optimal mixed strategies $\mathbf{a}$ for the current state $\mathbf{b}_k$ based on (20) through linear programming
Randomly select the energy and cheating choices $\{a^e, a^c\}$ based on the optimal mixed strategy profiles $\mathbf{a}$
Compute the next state $\mathbf{b}'$ based on the observations and update the Q-value for each player according to (21)
Update the state: $\mathbf{b}_k = \mathbf{b}'$ and decay the learning rate $\gamma_k$
$k := k + 1$
End
```

- As depicted in Fig. 3, $\mathcal{J}_i^k(\cdot)$ converges to an expected value $\mathcal{J}_i^\star(\cdot)$ for each player $i \in \{0, \ldots, 4\}$ within around 10 000 iterations. Here, in order to describe the convergence result for the entire belief states, we use $\max_{\mathbf{b} \in \mathcal{B}} \mathcal{J}_i(\mathbf{b})$ instead of considering that value for each state. Also, we observe that the converged value $\max \mathcal{J}_i^\star$ is positive for the attacker ($i = 4$) and negative for the *type agents*[8] of the sensor, which is intuitive.
- A partial iteration of the belief state is represented in Fig. 5, w.r.t. the original state shown in Fig. 4. Notice that the probabilities for $s_k = 0$ and $s_k = 1$ are always greater than the others, which is consistent with the state iteration shown in Fig. 4. When $k = 26$, $s_k = 3$ and $s_k = 1$ are more likely to happen, while actually $s_{26} = 0$. We conclude that the attacker may develop a rough guess about state value s_k based on the historical information. However, due to the cheating actions adopted by the sensor, the guess may not be accurate even as the collected information increases.
- Taking $\mathbf{b} = [1\ 0\ 0\ 0]$ as an example, the optimal strategies are concluded in Table 2. The entries of Table 2 are explained as follows. If the discretized version of the belief state is $\mathbf{b} = [1\ 0\ 0\ 0]$, then the attacker will play according to the last row of Table 2, that is, adopting the jamming power $a_2 = 0.2$ with probability one; as the sensor has the interim status of the game (i.e., $s_k = i$), it will execute the ith *type agent*'s optimal strategy in Table 2. For example, if $i = 0$, then the sensor may select the power $a_1 = 0.6$ for transmission and randomly modify the ACK information

[8] Recall that in the belief-based game $\mathcal{G}^S$, the possible values of private information (that is, $s_k \triangleq \{0, \ldots, 3\}$ in this example) possessed by the sensor is regarded as an individual player.

with probability 0.55. Recall that the attacker will adopt a deterministic policy under this belief state. It follows the intuition that the attacker believes the last data packet to be received successfully (i.e., $s_k = 0$), and the transmitted data at time k contains less valuable information. Hence, in order to save energy, it is not urgent to jam the data with the highest power level. The sensor's optimal strategies are computed as a solution to the game, and do not have an immediate simple intuitive explanation.

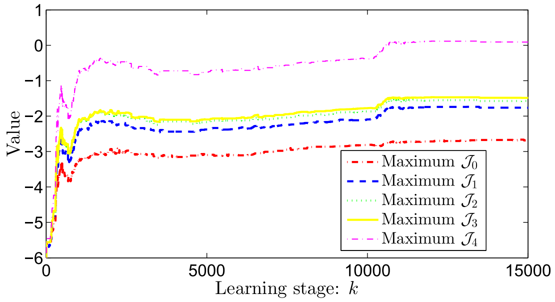


Fig. 3. Converged maximum J-value for each player.

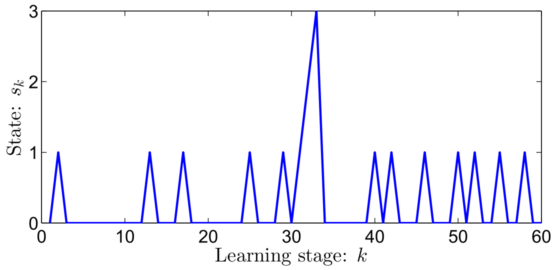


Fig. 4. Iteration of actual state s_k.

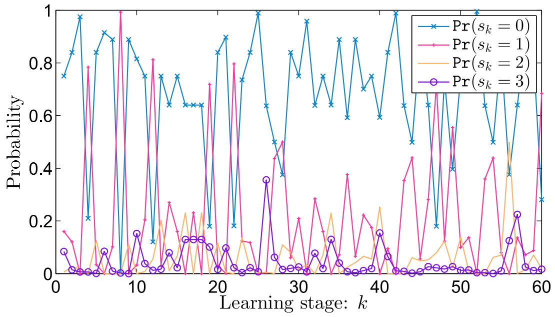


Fig. 5. Iteration of belief state $\mathbf{B}_k$.

Table 2
Optimal mixed strategies for the sensor and the attacker under the belief state $\mathbf{b} = [1\ 0\ 0\ 0]$.

Type agent i		Action choice (a_1, a^c)			
		(0.5, 0)	(0.6,0)	(0.5, 1)	(0.6, 1)
Sensor	0	0	0.43	0	0.57
	1	0	0.45	0	0.55
	2	0	0.18	0.81	0.01
	3	0	0	0	1
Attacker	4	$a_2 = 0.2$ 1		$a_2 = 0.4$ 0	

5. Conclusion

This paper investigated a security issue in remote state estimation, in which a sensor transmits the data packet to the remote estimator through a vulnerable communication channel suffering from jamming attacks. To against them, we considered a new defensive deception mechanism and wanted to foster its emerging promise as a tool for CPS security. Specifically, the sensor will play a deception trick to manipulate the attacker's belief and further alleviates the damage to estimation performance. As deceptive interactions are strategic confrontations between the tactical sensor and attacker, a general asymmetric-information stochastic game model is utilized to analyze their strategic interaction. To solve it, this game was converted into an equivalent symmetric-information one using a partially observable Markov decision process (POMDP) approach.

References

Anderson, B. D., & Moore, J. B. (2012). *Optimal filtering*. Courier Corporation.

Befekadu, G. K., Gupta, V., & Antsaklis, P. J. (2015). Risk-sensitive control under markov modulated denial-of-service (dos) attack strategies. *IEEE Transactions on Automatic Control*, *60*(12), 3299–3304.

Billingsley, P. (2013). *Convergence of probability measures*. New York: John Wiley & Sons.

Cardenas, A. A., Amin, S., & Sastry, S. (2008). Secure control: Towards survivable cyber-physical systems. In *Proc. IEEE 28th int. conf. distributed computing system workshops* (pp. 495–500).

Carroll, T. E., & Grosu, D. (2011). A game theoretic investigation of deception in network security. *Security and Communication Networks*, *4*(10), 1162–1172.

De Persis, C., & Tesi, P. (2015). Input-to-state stabilizing control under denial-of-service. *IEEE Transactions on Automatic Control*, *60*(11), 2930–2944.

Ding, K., Li, Y., Quevedo, D. E., Dey, S., & Shi, L. (2017). A multi-channel transmission schedule for remote state estimation under DoS attacks. *Automatica*, *78*, 194–201.

Ding, K., Ren, X., & Shi, L. (2016). Deception-based sensor scheduling for remote estimation under DoS attacks. *IFAC-PapersOnLine*, *49*(22), 169–174.

Feng, S., & Tesi, P. (2017). Resilient control under denial-of-service: Robust design. *Automatica*, *79*, 42–51.

Feng, X., Zheng, Z., Mohapatra, P., & Cansever, D. (2017). A stackelberg game and markov modeling of moving target defense. In *International conf. on decision and game theory for security* (pp. 315–335). Springer.

Fudenberg, D., & Tirole, J. (1991). *Game theory*. MIT Press.

Greenwald, A., Hall, K., & Serrano, R. (2003). Correlated Q-learning. *ICML*, *3*, 242–249.

Grover, K., Lim, A., & Yang, Q. (2014). Jamming and anti-jamming techniques in wireless networks: a survey. *International Journal of Ad Hoc and Ubiquitous Computing*, *17*(4), 197–215.

Gupta, V., Hassibi, B., & Murray, R. M. (2007). Optimal lqg control across packet-dropping links. *Systems & Control Letters*, *56*(6), 439–446.

Haurie, A., Krawczyk, J. B., & Zaccour, G. (2012). *Games and dynamic games*. World Scientific Publishing Co. Pte. Ltd.

Horák, K., Zhu, Q., & Bošanskỳ, B. (2017). Manipulating adversary's belief: A dynamic game approach to deception by design for proactive network security. In *International conf. on decision and game theory for security* (pp. 273–294). Springer.

Hovareshti, P., Gupta, V., & Baras, J. S. (2007). Sensor scheduling using smart sensors. In *Proc. IEEE 46th annu. conf. decision and control* (pp. 494–499).

Hu, J., & Wellman, M. P. (2003). Nash q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, *4*(Nov), 1039–1069.

Iskhakov, F., Rust, J., & Schjerning, B. (2016). Recursive lexicographical search: Finding all markov perfect equilibria of finite state directional dynamic games. *The Review of Economic Studies*, *83*(2), 658–703.

Khattab, S., Mosse, D., & Melhem, R. (2008). Modeling of the channel-hopping anti-jamming defense in multi-radio wireless networks. In *Proc. of the 5th annual international conf. on mobile and ubiquitous systems: computing, networking, and services* (p. 25).

Kim, K.-D., & Kumar, P. (2012). Cyber-physical systems: A perspective at the centennial. *Proceedings of the IEEE*.

Kurose, J. F., & Ross, K. W. (2012). *Computer networking: A top-down approach* (6th edition). Pearson.

Lazos, L., Liu, S., & Krunz, M. (2009). Mitigating control-channel jamming attacks in multi-channel ad hoc networks. In *Proc. of the second ACM conf. on wireless network security* (pp. 169–180).

Li, Y., Shi, L., Cheng, P., Chen, J., & Quevedo, D. (2015). Jamming attacks on remote state estimation in cyber-physical systems: A game-theoretic approach. *IEEE Transactions on Automatic Control*, *60*(10), 2831–2836.

Madani, O., Hanks, S., & Condon, A. (1999). On the undecidability of probabilistic planning and infinite-horizon partially observable markov decision problems. In *AAAI/IAAI* (pp. 541–548).

Mo, Y., Kim, T. H.-J., Brancik, K., Dickinson, D., Lee, H., Perrig, A., & Sinopoli, B. (2012). Cyber–physical security of a smart grid infrastructure. *Proceedings of the IEEE*, *100*(1), 195–209.

Navda, V., Bohra, A., Ganguly, S., & Rubenstein, D. (2007). Using channel hopping to increase 802.11 resilience to jamming attacks. In *The 26th IEEE international conf. on computer communications* (pp. 2526–2530).

Noubir, G. (2004). On connectivity in ad hoc networks under jamming using directional antennas and mobility. In *International conf. on wired/wireless internet communications* (pp. 186–200). Springer.

Pajic, M., & Mangharam, R. (2009). Anti-jamming for embedded wireless networks. In *IEEE international conf. on information processing in sensor networks* (pp. 301–312).

Pawlick, J., Colbert, E., & Zhu, Q. (2017). A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. arXiv preprint arXiv:1712.05441.

Pawlick, J., & Zhu, Q. (2015). Deception by design: evidence-based signaling games for network defense. arXiv preprint arXiv:1503.05458.

Pearce, D. (1982). Ex ante equilibrium: Strategic behaviour and the problem of perfection, Econometric Research Program Research Memorandum 301.

Pickholtz, R., Schilling, D., & Milstein, L. (1982). Theory of spread-spectrum communications–a tutorial. *IEEE Transactions on Communications*, *30*(5), 855–884.

Qin, J., Li, M., Shi, L., & Yu, X. (2018). Optimal denial-of-service attack scheduling with energy constraint over packet-dropping networks. *IEEE Transactions on Automatic Control*, *63*(6), 1648–1663.

Ren, Z., Cheng, P., Chen, J., Shi, L., & Zhang, H. (2014). Dynamic sensor transmission power scheduling for remote state estimation. *Automatica*, *50*(4), 1235–1242.

Rifa-Pous, H., & Herrera-Joancomartí, J. (2011). Computational and energy costs of cryptographic algorithms on handheld devices. *Future Internet*, *3*(1), 31–48.

SANS, I. C. S. (2016). Confirmation of a coordinated attack on the ukrainian power grid.

Sobel, M. J. (1973). Continuous stochastic games. *Journal of Applied Probability*, 597–604.

Taylor, J. (2012). Markov decision processes: Lecture notes for stp 425. *Stochastic Processes*.

Tse, D., & Viswanath, P. (2005). *Fundamentals of wireless communication*. Cambridge University Press.

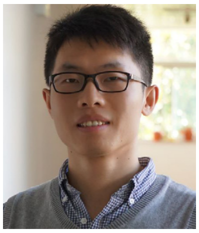
Wang, C., & Lu, Z. (2018). Cyber deception: overview and the road ahead. *IEEE Security & Privacy*, *16*(2), 80–85.

Xu, W., Wood, T., Trappe, W., & Zhang, Y. (2004). Channel surfing and spatial retreats: defenses against wireless denial of service. In *Proceedings of the 3rd ACM workshop on Wireless security* (pp. 80–89).

Zhang, H., Cheng, P., Shi, L., & Chen, J. (2015). Optimal denial-of-service attack scheduling with energy constraint. *IEEE Transactions on Automatic Control*, *60*(11), 3023–3028.

Zhu, Q., & Başar, T. (2013). Game-theoretic approach to feedback-driven multi-stage moving target defense. In *International conf. on decision and game theory for security* (pp. 246–263). Springer.

Kemi Ding received the B.S. degree in Electronic and Information Engineering from Huazhong University of Science and Technology, Wuhan, China, in 2014 and the Ph.D. degree in the Department of Electronic and Computer Engineering from Hong Kong University of Science and Technology, Kowloon, Hong Kong, China, in 2018. She is currently a postdoctoral researcher at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Prior to this, she was a postdoctoral researcher in the School of Electrical, Computer and Energy Engineering, Arizona State University from September 2018 to August 2019. Her current research interests include cyber–physical system security/privacy, networked state estimation, game theory and social networks.



Xiaoqiang Ren received the B.E. degree in the Department of Control Science and Engineering from Zhejiang University, Hangzhou, China, in 2012 and the Ph.D. degree in the Department of Electronic and Computer Engineering from Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2016. He is currently a professor at the School of Mechatronic Engineering and Automation, Shanghai University, China. Prior to this, he was a postdoctoral researcher in the Hong Kong University of Science and Technology in 2016, Nanyang Technological University from 2016 to 2018, and KTH Royal Institute of Technology from 2018 to 2019. His research interests include security of cyber–physical systems, sequential decision, and networked estimation and control.

Daniel E. Quevedo (S'97–M'05–SM'14) is Head of the Chair of Automatic Control (*Regelungs- und Automatisierungstechnik*) at Paderborn University, Germany. He received Ingeniero Civil Electrónico and M.Sc. degrees from the Universidad Técnica Federico Santa María, Chile, in 2000. In 2005, he was awarded the Ph.D. degree from the University of Newcastle in Australia.

Dr. Quevedo was supported by a full scholarship from the alumni association during his time at the Universidad Técnica Federico Santa María and received several university-wide prizes upon graduating. He received the IEEE Conference on Decision and Control Best Student Paper Award in 2003 and was also a finalist in 2002. In 2009 he was awarded a five-year Research Fellowship from the Australian Research Council. He is co-recipient of the 2018 IEEE Transactions on Automatic Control George S. Axelby Outstanding Paper Award.

Prof. Quevedo is Associate Editor of the *IEEE Control Systems Magazine*, Editor of the *International Journal of Robust and Nonlinear Control*, and past Chair of the IEEE Control Systems Society *Technical Committee on Networks & Communication Systems*. His research interests are in control of networked systems and of power converters.

Subhrakanti Dey received the Bachelor in Technology and Master in Technology degrees from the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, in 1991 and 1993, respectively, and the Ph.D. degree from the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, in 1996.

He is currently a Professor with the Hamilton Institute, National University of Ireland, Maynooth, Ireland. He is also a Professor with the Dept. of Engineering Sciences in Uppsala University, Sweden. Prior to this, he was a Professor with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Australia, from 2000 until early 2013, and a Professor of Telecommunications at University of South Australia during 2017–2018. From September 1995 to September 1997, and September 1998 to February 2000, he was a Postdoctoral Research Fellow with the Department of Systems Engineering, Australian National University. From September 1997 to September 1998, he was a Postdoctoral Research Associate with the Institute for Systems Research, University of Maryland, College Park. His current research interests include wireless communications and networks, signal processing for sensor networks, networked control systems, and molecular communication systems.

Professor Dey currently serves on the Editorial Board of IEEE Control Systems Letters, IEEE Transactions on Control of Network Systems, and IEEE Transactions on Wireless Communications. He was also an Associate Editor for IEEE Transactions on Signal Processing, (2007–2010, 2014–2018), IEEE Transactions on Automatic Control, (2004–2007) and Elsevier Systems and Control Letters (2003–2013).

Ling Shi received the B.S. degree in electrical and electronic engineering from Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2002 and the Ph.D. degree in Control and Dynamical Systems from California Institute of Technology, Pasadena, CA, USA, in 2008. He is currently an associate professor at the Department of Electronic and Computer Engineering, and the associate director of the Robotics Institute, both at the Hong Kong University of Science and Technology. His research interests include cyber–physical systems security, networked control systems, sensor scheduling, event-based state estimation and exoskeleton robots. He is a senior member of IEEE. He served as an editorial board member for The European Control Conference 2013–2016. He was a subject editor for International Journal of Robust and Nonlinear Control (2015–2017). He has been serving as an associate editor for IEEE Transactions on Control of Network Systems from July 2016, and an associate editor for IEEE Control Systems Letters from Feb 2017. He also served as an associate editor for a special issue on Secure Control of Cyber–Physical Systems in the IEEE Transactions on Control of Network Systems in 2015–2017. He served as the General Chair of the 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS 2018).