Abstract—We consider the joint design of control and scheduling under stochastic Denial-of-Service (DoS) attacks in the context of networked control systems. A sensor takes measurements of the system output and forwards its dynamic state estimates to a remote controller over a packet-dropping link. The controller determines the optimal control law for the process using the estimates it receives. An attacker aims at degrading the control performance by increasing the packet-dropout rate with a DoS attack towards the sensor-controller channel. Assume both the controller and the attacker are rational in a game-theoretic sense. We establish a partially observable stochastic game to derive the optimal joint design of scheduling and control. Using dynamic programming we prove that the control and scheduling policies can be designed separately without sacrificing optimality, making the problem equivalent to a complete information game. We employ Nash Q-learning to solve the problem and prove that the solution is guaranteed to constitute an \( \epsilon \)-Nash equilibrium. Numerical examples are provided to illustrate the tradeoffs between control performance and communication cost.

I. INTRODUCTION

Cyber-physical systems (CPSs) have successfully integrated computing devices and networking infrastructure to remotely sense and control the physical world. This enables the development of exciting applications in manufacturing, transportation, and smart grid [1]. However, the wireless communication among components of CPSs introduces vulnerability against malicious adversaries [2]. For example, data integrity and availability can be easily deteriorated by replay attacks [3] and denial-of-service (DoS) attacks [4]. This weakness brings a great threat to the safety and security of CPSs with real-time operations. This raises the issue of how to design systematic prevention mechanisms to enhance system security.

Game theory is often employed as a tool for resilient design and analysis in the presence of adversaries [5]. Assuming the rationality of the adversary, the game-theoretic framework can model the interactive decision-making process and yield non-conservative strategies. Game theory has been widely applied to the secure design of control and estimation. For example, Zhu and Başar [6] formulated an infinite time horizon zero-sum game to derive a secure control policy that is resilient against various types of cyberattacks. Li et al. [7] presented a power control strategy for remote state estimation in the presence of DoS attacks through a stochastic game. The idea has been recently extended to address linear quadratic Gaussian control problems in [8]. Miao et al. [9] proposed a stochastic game with hybrid states for designing optimal switching control policies to counteract data injection attacks.

In this work, we employ the game-theoretic framework to study a secure co-design of control and scheduling to save communication costs and enhance CPS security. As shown in Fig. 1, we consider a remote control problem where the system state estimate is sent from the sensor to the controller across a TCP link. The communication channel suffers from potential DoS attacks launched by an adversary who intends to degrade the control performance and increase the controller’s communication costs. Since the controller and the adversary have opposite interests, we model this interactive decision-making process by formulating a zero-sum stochastic game with the controller and the adversary taken as two players.

To date, control-scheduling co-design has been extensively studied without considering security issues and proved to be an efficient approach to reduce communication cost in networked control systems [10]–[13]. However, due to the coupling effects of control and scheduling, co-design methods will commonly lead to optimization problems of increased computational complexity. Similar issues arise when the co-design is conducted in a game-theoretic framework. Specifically, to derive the jointly optimal strategy for the control and scheduling, both the system input and scheduling
command have to be taken as actors in the stochastic game. Since the controller output is commonly a continuous-valued variable and the scheduling command is a discrete variable, the resulting game constitutes a game with hybrid states. An additional difficulty arises in the “natural” situation where the cost function is related to the system state $x_k$, which is unobserved to both of the players, leading to a partially observable stochastic game.

To tackle the challenges in computation, in the present work we explore the possibility of separating the control and scheduling. With the help of dynamic programming, we establish that when the controller works as the scheduler, the optimal feedback gain is independent of packet arrival distributions and realizations and the optimal scheduling command is independent of the system state estimates. Based on these properties, we can simplify the partially observable stochastic game with hybrid states into a complete information stochastic game with countably infinite discrete states. This allows us to use Nash Q-learning, where we prove that it provides an $\epsilon$-Nash equilibrium.

We emphasize that, in contrast to existing co-design works, including [8], [14], [15], where the scheduling variable is a priori assumed to be independent of the system state, in the present work we do not impose any assumptions on the structure of the solution. Therefore, the current separate design procedure does not sacrifice optimality.

The contribution of this work is three-fold: (1) we establish a partially observable stochastic game with hybrid states to derive the jointly optimal scheduling and control strategy under a DoS attack in the context of the remote control; (2) we prove that the design of control and scheduling can be separated without sacrificing optimality and show that the game can be reduced to a complete information stochastic game with countably finite discrete states; (3) we show that Nash Q-learning can efficiently solve such problems and guarantee an $\epsilon$-equilibrium.

The remaining parts of this note are organized as follows: Section II describes the model of the networked control system and the problem setup; Section III presents the formulation of the partially observable stochastic game and analyzes the optimality of the separate design, as well as proves the $\epsilon$-equilibrium of the resulting solution; Section IV provides a numerical example to verify the theoretical results and illustrate the tradeoffs between the communication cost and the control performance. Finally, Section V draws conclusions.

II. PROBLEM FORMULATION

A. System model

Consider a linear dynamic system

\[ x_{k+1} = Ax_k + Bu_k + w_k, \]
\[ y_k = Cx_k + v_k, \]

where $x_k \in \mathbb{R}^{n_x}$, $u_k \in \mathbb{R}^{n_u}$ and $y_k \in \mathbb{R}^{n_y}$ denote the system state, input and output respectively; $w_k \in \mathbb{R}^{n_w}$ and $v_k \in \mathbb{R}^{n_v}$ are i.i.d. Gaussian noises with zero mean, i.e., $\mathbb{E}(w_k) = 0$, $\mathbb{E}(v_k) = 0$, and $\mathbb{E}(w_i w_j^\top) = \mathbb{E}(v_i v_j^\top) = 0$ if $i \neq j$. Denote $\mathbb{E}(w_k w_k^\top) = Q$ and $\mathbb{E}(v_k v_k^\top) = R$, with $Q > 0$ and $R > 0$. Assume the pair $(A, B)$ is controllable and $(A, C)$ is detectable.

As depicted in Fig. 1 a local sensor takes measurements of the system output $y_k$ and generates a state estimate which is denoted as $\hat{x}_k^s$. This estimate is wrapped as a packet and forwarded to the remote controller over a TCP link; $\nu_k \in \{0, 1\}$ denotes the transmission command that the sensor commits to. Assume the link has i.i.d. packet dropouts and the probability of a successful transmission is $\lambda$. Define $\gamma_k^c$ as an indicator variable of successful transmission. We define

\[ \gamma_k^c \triangleq \begin{cases} 0 & \text{if a packet dropout occurs}, \\
1 & \text{if } \hat{x}_k^s \text{ is successfully received} \end{cases} \]

and note that

\[ \mathbb{P}(\gamma_k^c = 0 | \nu_k = 0) = 1, \]
\[ \mathbb{P}(\gamma_k^c = 0 | \nu_k = 1) = 1 - \lambda. \]

The remote controller determines the input $u_k$ and the scheduling command $\nu_k$ based on the received information. If the packet is successfully received, then the TCP-link will send an acknowledgment back to the sensor, from which the sensor can infer the value of $\gamma_k^c$.

B. State estimation

Denote the information set at the sensor and at the controller at time $k$ as $I_k^s$ and $I_k^c$, respectively. According to the problem setup described above, we have

\[ I_k^s = \{y_0, \ldots, y_k, \hat{x}_0^s, \ldots, \hat{x}_k^s, v_0, \ldots, v_k, \gamma_0^s, \ldots, \gamma_k^s\}, \]
\[ I_k^c = \{\gamma_0^c, \ldots, \gamma_k^c, \nu_0, \ldots, \nu_k, \gamma_0^c, \ldots, \gamma_k^c\}, \]

so that $I_k^c \subset I_k^s$. Assume the sensor knows the control policy. Then, it can infer $u_k$ from $\gamma_k^c$ for each $k$. Note that the separation principle of estimation and control holds when both $\{u_0, \ldots, u_k\}$ and $\{y_0, \ldots, y_k\}$ are available to the controller. Thus, despite of the control strategy, the sensor can run a Kalman filter to obtain the optimal estimate of the state. Specifically, we have

\[ \hat{x}_k = A\hat{x}_{k-1} + Bu_{k-1} + K_k \left( y_k - C(\hat{x}_{k-1} + Bu_{k-1}) \right) \]
\[ \tilde{x}_k = (I_{n_x} - K_k C)A\hat{x}_{k-1} + (I - K_k)w_{k-1} - v_{k-1} \]
\[ P_k^s = (I_{n_x} - K_k C)(AP_{k-1}^s A^\top + Q), \]

where $\hat{x}_k = x_k - \tilde{x}_k^s$ denotes the state estimation error at the sensor. In (7) $P_k^s$ denotes the estimation error covariance defined as $P_k^s \triangleq \mathbb{E}(\tilde{x}_k\tilde{x}_k^\top)$ and $K_k \triangleq (AP_{k-1}^s A^\top + Q)C^\top \left( C(\tilde{x}_{k-1} + Bu_{k-1})^\top + R \right)^{-1}$.

Let $\tilde{P}$ denote the steady-state error covariance [16]. Assume that the Kalman filter has reached the steady state, i.e., $P_k^s = \tilde{P}$.
Taking into account the packet dropouts, the optimal state estimate at the controller, denoted as \( \hat{x}_k \), follows the recursion

\[
\hat{x}_k = \begin{cases} 
\hat{x}_{k}^c & \nu_k \gamma_k^c = 1, \\
\hat{x}_k & \nu_k \gamma_k^c = 0,
\end{cases}
\]  

(8)

where

\[
\hat{x}_k = A\hat{x}_{k-1} + Bu_{k-1}.
\]  

(9)

Accordingly, the expected error covariance in terms of \( \hat{x}_k \), denoted as \( P_k \), obeys

\[
P_k = \begin{cases} 
\bar{P} & \nu_k \gamma_k^c = 1, \\
\bar{P} & \nu_k \gamma_k^c = 0.
\end{cases}
\]  

(10)

where \( f(X) = AXA^T + Q \). Define \( \tau_k \) as the holding time, which indicates the time steps since the last successful transmission, i.e.

\[
\tau_k = \begin{cases} 
0 & \nu_k \gamma_k^c = 1, \\
\tau_{k-1} + 1 & \nu_k \gamma_k^c = 0.
\end{cases}
\]  

(11)

Given (10) and (11), we can express \( P_k \) in terms of \( \tau_k \) and \( \bar{P} \) as \( P_k = f^{\tau_k}(\bar{P}) \). Our subsequent analysis uses the following properties of the state estimate and the estimation error.

Lemma 1: The following facts are true:

(a) \( \mathbb{E}[\left(\hat{x}_k - \bar{x}_k\right)\bar{x}_k^\top | \bar{x}_k] = 0; \)
(b) \( \mathbb{E}[\left(\hat{x}_k^c - \bar{x}_k\right)\bar{x}_k^\top | \bar{x}_k] = 0; \)
(c) \( \mathbb{E}[\left(\hat{x}_k^c - \bar{x}_k\right)\left(\hat{x}_k^c - \bar{x}_k\right)^\top | \bar{x}_k] = f^{\tau_{k-1}+1}(\bar{P}) - \bar{P}. \)

Proof: See Appendix A.

C. Model of the adversary

Consider an adversary that intends to degrade the control performance by deteriorating the availability of the communication channel. In particular, the adversary may launch a DoS attack, e.g., by increasing the signal noise ratio, and decrease the successful transmission probability \([7]\). We model the DoS attack as a binary process where the successful transmission probability is lowered to \( \lambda_s < \lambda \) when an attack is launched and remains unchanged otherwise.

Let \( a_k \) be an indicator variable of the DoS attack, i.e.

\[
a_k = \begin{cases} 
1 & \text{an attack is launched}, \\
0 & \text{no attack is launched}.
\end{cases}
\]

In view of Eq. (5), we have

\[
P\left(\gamma_k^c = 0 | (\nu_k, a_k) = (0, 0)\right) = 1, \\
P\left(\gamma_k^c = 0 | (\nu_k, a_k) = (1, 0)\right) = 1 - \lambda, \\
P\left(\gamma_k^c = 0 | (\nu_k, a_k) = (1, 1)\right) = 1 - \lambda_a.
\]  

(12)

From Eq. (11) and Eq. (12), the evolution of the holding time \( \tau_k \) in the presence of an attacker can be modeled as a Markov decision process (MDP) where \( \tau_k \) is taken as the state, and the pair \( (\nu_k, a_k) \) is taken as the action, as illustrated in Fig. 2.

![Fig. 2. State transition probabilities: the circle denotes the value of the state \( \tau_k \), the square denotes the value of the action pair \( (\nu_k, a_k) \).](image)

State transition probabilities are computed accordingly as

\[
\mathbb{P}(\tau_k = i + 1 | (\tau_k, \nu_k, a_k) = (i, 0, j)) = 1, \quad j \in \{0, 1\}, \\
\mathbb{P}(\tau_k = 0 | (\tau_k, \nu_k, a_k) = (i, 1, 0)) = \lambda, \\
\mathbb{P}(\tau_k = 0 | (\tau_k, \nu_k, a_k) = (i, 1, 1)) = \lambda_a.
\]  

(13)

In addition, assume that the attacker can overhear the acknowledgments sent from the controller to the sensor and make use of them to efficiently schedule the attacks. Denote the information available to the attacker at time \( k \) as \( I_k^a \). We then have

\[
I_k^a = \{\gamma_0, \ldots, \gamma_k\}
\]  

(14)

and, when comparing with [4], note that the attacker has significantly less information than the controller-scheduler.

III. CO-DESIGN OF CONTROL AND SCHEDULING

In this section, we will formulate the interactive decision making process as a partially observable stochastic game and show that, from the controller’s perspective, the design of the control and scheduling can be conducted separately without sacrificing optimality. This property simplifies the game into a complete information stochastic game.

A. Design of the partially observable stochastic game

Consider that the legitimate controller aims at optimizing the control performance by using the information in \( I_k^c \) to adjust the input \( u_k \) and scheduling the transmission command \( \nu_k \). In particular, we evaluate the control performance via a discounted linear combination of a quadratic term of the state \( x_k \) and the input \( u_k \), as well as the controller’s transmission cost and the negative of the adversary’s attacking cost:

\[
J = \frac{1}{K} \sum_{k=0}^{K-1} \eta^k \mathbb{E} \left( x_k^\top W x_k + u_k^\top U u_k + c_s \nu_k - c_a a_k \right).
\]  

(15)

Here \( \eta \in (0, 1) \) denotes the discount factor, \( c_s \) denotes the cost of transmission and \( c_a \) denotes the cost of launching a jamming attack.

Contrarily, the adversary intends to maximize \( J \) by smartly scheduling the attack \( a_k \) according to its observations \( I_k^a \).

Assume the two agents select their actions independently. The interaction can then be formulated as a zero-sum game.
stochastic game. Specifically, we denote the stochastic game \( \mathcal{G} \) as a tuple \( \{ \mathcal{I}, \{ b^0 \}, \mathcal{A}_c, \mathcal{A}_a, \mathcal{C}_c, \mathcal{C}_a, \mathcal{P}, \mathcal{C} \} \), where
- \( \mathcal{I} \) denotes the number of agents. Here we have two agents, i.e., the controller and the attacker.
- \( \mathcal{S} \) denotes the set of states \( s_k \). Here we define
  \[ s_k = [x_k^T, \tau_k-1]^T \]
  where \( x_k \) is the system state vector taking values from the continuous space \( \mathbb{R}^{n_x} \) and the holding time \( \tau_k \) is an integer taking values from the set \( \{0, 1, 2, \ldots \} \).
- \( \{ b^0 \} \) denotes the initial value of the state. Without loss of generality, assume \( x_0 \) is a zero-mean Gaussian noise and \( \tau_0 = 0 \).
- \( \mathcal{A}_c \) denotes the action space of the controller. In our problem, the controller has two actors, i.e., the system input \( u_k \) which is continuous and the scheduling command \( \nu_k \in \{0, 1\} \).
- \( \mathcal{A}_a \) denotes the action space of the controller, which is \( a_k \in \{0, 1\} \).
- \( \mathcal{C}_c \) denotes the set of observations available to the controller. Denote the observation available at time \( k \) as \( o_k^c \). Consider the observation available at time \( k \) as \( o_k^c \). From Eq. (4) and (5), the controller knows \( \hat{x}_k \), which is an unbiased estimate of \( x_k \). Beyond that, the value of \( \tau_k \) can be inferred from \( \{ \gamma_k \} \).
- \( \mathcal{C}_a \) denotes the set of observations available to the attacker. Therefore, we define \( o_k^a \) as follows.
  \[ o_k^a = \{ \tau_k-1, \nu_k \gamma_k \hat{x}_k \} \] \( (16) \)
- \( \mathcal{P} \) is the transition probabilities. From Eq. (1), we may notice that the transition of \( x_k \) depends on \( u_k \) and the distribution of \( w_k \). The transition probability of \( x_k \) follows
  \[ P(x_k | x_{k-1}, u_{k-1}) \sim N(Ax_{k-1} + Bu_{k-1}, Q) \]
  In addition, the evolution of the holding time \( \tau_k \) depends on the actions \( \nu_k \) and \( a_k \) as given in Eq. (13) and Fig. 2
- \( \mathcal{C} \) denotes the cost function. From Eq. (15), we have the immediate cost at time \( k \) given as
  \[ c_k = x_k^T W x_k + u_k^T U u_k + c_a \nu_k - c_a a_k \]
  The controller aims at minimizing the cost while the attacker aims at maximizing the cost.
  In view of Eq. (16) and (17), neither of the two agents has complete information of the state \( s_k \). Therefore, the game is essentially a partially observable stochastic game with a hybrid of discrete and continuous states [9]. In the following section, we will use dynamic programming ideas to show that the design of control and scheduling can be conducted separately without sacrificing optimality. This will enable us to simplify the problem into a complete information game with countably infinite discrete states.

B. Separation of scheduling and control

In this section, we will study the separation of control and scheduling. We will first consider the case where the time horizon \( K \) in (15) is finite, and then extend the conclusions to infinite horizons, \( K \to \infty \). For the zero-sum stochastic game \( \mathcal{G} \) with finite stages, the existence of a stationary Nash equilibrium and the separability is summarized as follows:

**Theorem 1:** For the stochastic game \( \mathcal{G} \), the stationary Nash equilibrium exists and the design of the control and scheduling can be conducted separately. In particular, the optimal scheduling command \( \nu_k \) solely depends on the holding time \( \tau_{k-1} \).

**Proof:** See Appendix B.

**Remark 1:** Theorem [1] established that in the present setting, the optimal scheduling command \( \nu_k \) is independent of the state estimate. This fundamental property is often assumed without detailed justification in co-design works, e.g., [15].

Our result guarantees that the separate design does not sacrifice optimality, when the controller works as the scheduler. However, if the sensor schedules, such as considered in [8], [14], then this separation property does not hold. In fact, consider that \( \nu_k \) is determined by the sensor, so that \( \hat{x}_k^c \) and \( \hat{x}_k \) are available before the decision is made. The expectations in Eq. (16) (in the appendix) should be computed conditioned on both \( \hat{x}_k \) and \( \hat{x}_k^c \), in which case both \( \hat{x}_k \) and \( \hat{x}_k^c \) will appear in the coefficient terms of \( \nu_k \). Therefore, the optimal schedule \( \nu_k \) will in general depend on \( \hat{x}_k \) and \( \hat{x}_k^c \). Since \( \nu_k \) contains information of \( \hat{x}_k^c \), the optimal estimate at the controller does not follow Eq. (8), but becomes a nonlinear estimation problem, as analyzed in [17], [18], and will further complicate the design of the feedback gain.

□

C. Equivalence to a complete information game

We will next show that the partially observable game \( \mathcal{G} \) is equivalent to a stochastic game with complete information.

**Corollary 1:** The optimal control policy is in the form of
  \[ u_k^* = -\eta (B^T \eta S_{k+1} B + U)^{-1} B^T S_{k+1} A \hat{x}_k \] \( (18) \)
  and the optimal scheduling policy is equivalent to the solution to the minimax optimization problem
  \[ \min_{\{\nu_k\}} \max_{\{a_k\}} \sum_{k=0}^{K-1} \eta^k \left( f_k ((A^T \eta S_{k+1} A + W - S_k) f_k (P)) + c_a \nu_k - c_a a_k \right) \] \( (19) \)
  where \( S_{K+1} = W \) and \( S_k = A^T \eta S_{k+1} A + W - A^T \eta^2 S_{k+1} B (\eta B^T S_{k+1} B + U)^{-1} B^T S_{k+1} A \) with \( k = K, \ldots, 0 \).

**Proof:** See Appendix C.
Corollary 1 simplifies the game into a complete information stochastic game, where the state \( s_k \) is reduced to \( \tau_{k-1} \) and the immediate cost becomes
\[
\hat{c}_k = \text{tr} \left( (\eta A^T S_{k+1} A + W - S_k) f^{\tau_k}(P) \right) + c_a \nu_k - c_a a_k.
\]
Since both the controller and the attacker can observe the holding time \( \{\tau_k\} \), the game is of complete information and can be efficiently solved with dynamic programming for this finite horizon scenario.

Next, by taking limits as \( K \to \infty \), we can derive similar results for the infinite horizon scenario.

**Corollary 2:** Suppose that
\[
\lambda_a > 1 - \frac{\eta}{\max_i |\sigma_i(A)|^2}. \tag{20}
\]
Then the infinite horizon stochastic game \( \mathcal{G} \) has a Nash equilibrium. The optimal control policy is given as
\[
a_k^* = -\eta (B^T \eta S_\infty B + U)^{-1} B^T S_\infty A \hat{x}_k, \tag{21}
\]
where
\[
S_\infty = \eta A^T S_\infty A + W - A^T \eta^2 S_\infty B (\eta B^T S_\infty B + U)^{-1} B^T S_\infty A.
\]
The optimal scheduling variables \( \{\nu_k\} \) and \( \{a_k\} \) can be derived by solving
\[
\min_{\{\nu_k\} \{a_k\}} \lim_{K \to \infty} \frac{1}{K} \sum_{k=0}^{K-1} \eta^k \left( \text{tr} \left( (A^T \eta S_\infty A + W - S_\infty) f^{\tau_k}(P) \right) + c_a \nu_k - c_a a_k \right). \tag{22}
\]
We note that (20) gives a sufficient condition to ensure the boundedness of (22), such that even if the attacker launches the DoS attack at all time instances, the expected estimation error covariances are ensured to be bounded.

**D. Nash Q-learning and State truncation**

The state \( \tau_k \) takes values from a set with countably infinite elements. Since Nash Q-learning can only be applied to MDPs with finite states, we truncate the MDP in Fig. 2 in aggregating the states with \( \tau_k \geq N - 1 \). Specifically, we modify the transition probability for \( \tau_k \geq N - 1 \) as
\[
\mathbb{P}(\tau_k = N - 1 | (\tau_{k-1}, \nu_k, a_k) = (N - 1, 0, j)) = 1
\]
with \( j = 0, 1 \). In this way, we keep \( N \) states in the MDP and therefore facilitate the application of Nash Q-learning. We learn the \( Q \)-function, denoted as \( Q(\tau, \nu_k, a_k) \), for \( \tau \in [0, N - 1] \) according to the Nash Q-learning algorithm presented in [7]. Thereafter, the transmission and attack policies for \( \tau \in [0, N - 2] \) are derived from (23):
\[
\pi^*_N(\tau) = \arg \min_{\pi} \arg \max_{\nu_k, a_k} \sum_{\nu_k, a_k} Q(\tau, \nu_k, a_k) \mathbb{P}(\nu_k | \pi^*_N) \mathbb{P}(a_k | \pi^*_N), \tag{23}
\]
where \( \pi^*_N(\tau) \) and \( \pi^*_N(\nu_k) \) give the probabilities of \( \mathbb{P}(\nu_k) \) and \( \mathbb{P}(a_k) \) for each \( \tau \in [0, N - 2] \) and \( \nu_k, a_k \in \{0, 1\} \).

For \( \tau \geq N - 1 \), we set \( \mathbb{P}(\nu_k = 1) = 1 \) and \( \mathbb{P}(a_k = 1) = 1 \). Next, we will show that this truncated policy gives an \( \epsilon \)-Nash equilibrium when \( N \) is large.

**Theorem 2:** Denote the truncated policy in Eq. (23) for the stochastic game \( \mathcal{G} \) as \( (\pi^*_N, \pi^*_N) \). Denote the value of Problem (22) as \( J \). Then, the pair \( (\pi^*_N, \pi^*_N) \) is an \( \epsilon \)-Nash equilibrium of the game \( \mathcal{G} \), i.e.
\[
J(\pi^*_N, \pi^*_N) \leq \min_{\pi} J(\pi, \pi_N^*) + \epsilon_N, \tag{24}
\]
\[
J(\pi^*_N, \pi^*_N) \geq \max_{\pi} J(\pi_N^*, \pi^*_N) - \epsilon_N, \tag{25}
\]
where \( \lim_{N \to \infty} \epsilon_N = 0 \).

**Proof:** See Appendix D.

**IV. SIMULATION**

Consider a linear system with
\[
A = \begin{bmatrix} 1.25 & 0.1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 1 \end{bmatrix}^T.
\]
The noise covariances are set as \( Q = I \) and \( R = I \). When there is no attack, the successful transmission rate is \( \lambda = 0.9 \). At instances when an attack is launched, the rate is reduced to \( \lambda_a = 0.6 \).
We consider an infinite time horizon co-design problem with weight matrices $U = I$ and $W = I$, and a discount factor $\gamma = 0.999$. The transmission cost is set to $c_s = 300$ and the attack cost is $c_a = 200$.

The optimal control law is computed from (21) as $L = \begin{bmatrix} -1.09 & -0.1 \end{bmatrix}$. Based on that, we employ the Nash Q-learning in [7] to solve the complete information stochastic game in (19) with the truncation horizon $N$ increased from 3 to 15. The corresponding optimal value of the game is plotted in Fig. 3, showing the convergence of the value as $N$ is increased.

We next study the tradeoff between the transmission cost and the empirical control performance, evaluated with

$$
\log \left( \frac{1}{\text{Iter}} \sum_{i=0}^{\text{Iter}} x_i^T W x_i + u_i^T U u_i \right),
$$

where Iter is the length of the simulation. To this end, we run Monte Carlo simulations with the cost $c_s$ and $c_a$ screening from 0 to 2000 with an interval 100. Iter is set to be $10^5$ for each group of $(c_s, c_a)$. Fig. 4 illustrates that the control performance deteriorates when the attack frequency increases and the transmission frequency decreases.

V. CONCLUSIONS

In the present work, we have studied the design of control and scheduling in the presence of stochastic DoS attacks. We have designed a zero-sum partially observable stochastic to jointly optimize the control and scheduling strategies. We have established that the optimal transmission scheduling command is independent of the state estimate and that the design of control and scheduling can be conducted separately without sacrificing optimality. Based on this property, the partially observable game is simplified as a complete information game. Moreover, we have proved that by applying a Nash Q-learning to the stochastic game with appropriately truncated states, an $\epsilon$-equilibrium can be derived. Future works may consider employing the sensor as the scheduler, in which case more information can be exploited for scheduling and enhanced security can be expected.

APPENDIX

A. Proof of Lemma 1

Proof: Item (a) is the same as Lemma 4.1 in [19]. We will prove Item (b) and (c). Assume that the latest successful transmission occurs at time $k_0$ where $k_0 \leq k$. From Eq. (8), we have

$$\bar{x}_k = A^{k-k_0} \hat{x}_{k_0} + \sum_{i=k_0}^{k-1} A^{k-1-i} B u_i.$$  \hspace{1cm} (26)

From Eq. (5), we have

$$\hat{x}_k^{\pi} = A \hat{x}_{k-1} + B u_{k-1} + K C A \hat{x}_{k-1} + K C w_{k-1} + v_{k-1},$$

$$\hspace{1cm} = A^{k-k_0} \hat{x}_{k_0} + \sum_{i=k_0}^{k-1} A^{k-1-i} B u_i + \sum_{i=k_0}^{k-1} A^{k-1-i} r_i.$$  \hspace{1cm} (27)

Eq. (26) and (27) together give that

$$\hat{x}_k - \bar{x}_k = \sum_{i=k_0}^{k-1} A^{k-1-i} r_i.$$  

Since $r_i$ is independent of $\hat{x}_{k_0}$ and $u_i$ for $i \geq k_0$, $\hat{x}_k - \bar{x}_k$ is independent of $\bar{x}_k$. Therefore, we have

$$E[ (\hat{x}_k - \bar{x}_k) x_k^T | \bar{x}_k ] = 0.$$  \hspace{1cm} (28)

Moreover, from Eq. (1), we have

$$x_k = A^{k-k_0} x_{k_0} + \sum_{i=k_0}^{k-1} A^{k-1-i} B u_i + \sum_{i=k_0}^{k-1} A^{k-1-i} w_i,$$

$$\bar{x}_k = A^{k-k_0} \bar{x}_{k_0} + \sum_{i=k_0}^{k-1} A^{k-1-i} (r_i - u_i).$$  \hspace{1cm} (28)

Since $\bar{x}_{k_0}$, $w_i$ and $v_i$ are i.i.d and mutually independent when $i \geq k_0$, we have $E(\bar{x}_k, \bar{x}_k) = 0$, which further gives that

$$E((x_k - \hat{x}_k)(x_k - \hat{x}_k)^T | \bar{x}_k) = E((x_k - \hat{x}_k)(x_k - \hat{x}_k)^T) = P,$$

and

$$E((\hat{x}_k - \bar{x}_k)(\hat{x}_k - \bar{x}_k)^T | \bar{x}_k) = E((\hat{x}_k - x_k)(\hat{x}_k - x_k)^T | \bar{x}_k) = E((\hat{x}_k - x_k)(\hat{x}_k - x_k)^T + (x_k - \hat{x}_k)(x_k - \hat{x}_k)^T + 2(x_k - \hat{x}_k)(x_k - \hat{x}_k)^T | \bar{x}_k) = x_k - \bar{x}_k + \hat{x}_k - x_k = E((x_k - \hat{x}_k)(x_k - \hat{x}_k)^T | \bar{x}_k) = f^{\pi_{k+1}}(P) - P.$$  \hspace{1cm} (29)

B. Proof of Theorem 7

Proof: Denote the strategy adopted by the controller at stage $k$ as $\pi_k(o_k^c)$ and the strategy adopted by the attacker at stage $k$ as $\pi_k(o_k^a)$. From the controller’s perspective, the game can be formulated in a recursive form as

$$J_{k+1}^c = E( x_{K+1}^c S_{K+1} x_{K+1}^c ),$$

$$\vdots$$

$$J_0^c = \inf_{\pi_k(o_k^c)} \sup_{\pi_k(o_k^a)} E(c_k + \gamma J_{k+1}^c),$$  \hspace{1cm} (29)

with $k = K, \ldots, 0$. Similarly, from the adversary’s perspective, the game can be expressed as

$$J_{K+1}^a = E( x_{K+1}^a S_{K+1} x_{K+1}^a ),$$

$$\vdots$$

$$J_0^a = \sup_{\pi_k(o_k^a)} \inf_{\pi_k(o_k^c)} E(c_k + \gamma J_{k+1}^a),$$  \hspace{1cm} (30)

with $k = K, \ldots, 0$. Next, we will prove the existence of a stationary Nash equilibrium by showing that $J_k^c = J_k^a$ for all $k \in [0, K]$. 

We will start with $k = K$ and follow the principle of dynamic programming. From Eq. (31), we have

$$J_{K}^{c} = \inf_{\eta K} \sup_{\{v_k(x_k)\}_{k=1}^{K}} \mathbb{E}(x_{K}^{T}Wx_{K} + u_{K}^{T}Uu_{K} + c_{a}u_{K} - c_{a}a_{K} + \eta \mathbb{E}(x_{K}^{T}SK_{K+1}x_{K+1})).$$

If $\gamma_{K} = 0$, from Eq. (8), we obtain

$$J_{K}^{c,\gamma_{K}=0} = \min \mathbb{E}(x_{K}^{T}(W + \eta A^{T}SK_{K+1}A)x_{K} + u_{K}^{T}(U + \eta B^{T}SK_{K+1}B)u_{K} + 2\eta \mathbb{E}B^{T}SK_{K+1}Ax_{K} | x_{K}) + \eta \mathbb{E}(S_{K+1}Q) + \eta \mathbb{E}(\nu_{K}u_{K} - c_{a}a_{K}).$$

According to [20], the optimal input is given as

$$u_{K} = -\eta \mathbb{E}(B^{T}SK_{K+1} + U)^{-1}B^{T}S_{K}A \hat{x}_{K}.$$  \hspace{1cm} (31)

Substitution of $J_{K}^{c}$ into $J_{K}^{c,\gamma_{K}=0}$ provides

$$J_{K}^{c,\gamma_{K}=0} = \mathbb{E}(x_{K}^{T}SK_{K}x_{K} + \text{tr}(W + \eta A^{T}SK_{K+1}A)f^{0}(\tilde{P}) + \eta \mathbb{E}(S_{K+1}Q) + \eta \mathbb{E}(\nu_{K}u_{K} - c_{a}a_{K}).$$

If $\gamma_{K} = 1$, then $\hat{x}_{K}$ is available to the controller. Similar to Eq. (31), the optimal control law follows

$$u_{K} = -\eta \mathbb{E}(B^{T}SK_{K+1} + U)^{-1}B^{T}S_{K}A \hat{x}_{K}.$$  \hspace{1cm} (32)

The corresponding value is

$$J_{K}^{c,\gamma_{K}=1} = \mathbb{E}(x_{K}^{T}SK_{K}x_{K} + \text{tr}(W + \eta A^{T}SK_{K+1}A)f^{0}(\tilde{P}) + \eta \mathbb{E}(S_{K+1}Q) + \eta \mathbb{E}(\nu_{K}u_{K} - c_{a}a_{K}),$$

and

$$\mathbb{E}(J_{K}^{c,\gamma_{K}=1} | \tilde{x}_{K}) = \mathbb{E}(\tilde{x}_{K}^{T}SK_{K}\tilde{x}_{K} | \tilde{x}_{K}) + \text{tr}(W + \eta A^{T}SK_{K+1}A)f^{0}(\tilde{P}) + \eta \mathbb{E}(S_{K+1}Q) + \eta \mathbb{E}(\nu_{K}u_{K} - c_{a}a_{K}).$$  \hspace{1cm} (33)

Eqs. (31) and (33) indicate that the optimal feedback gain is independent of the network properties. In addition, $J_{K}^{c}$ can be expressed as

$$J_{K}^{c} = \min_{\nu_{K}(x_{K},\tau_{K-1})} \max_{a_{K}(\tau_{K-1})} \mathbb{E}

(\langle 1 - \gamma_{K} \rangle J_{K,\gamma_{K}=0}^{c} + \gamma_{K} J_{K,\gamma_{K}=1}^{c}).$$

From Lemma I (b) and (c), we have

$$\mathbb{E}

(\langle 1 - \gamma_{K} \rangle J_{K,\gamma_{K}=0}^{c} + \gamma_{K} J_{K,\gamma_{K}=1}^{c} | \tilde{x}_{K}, \tau_{K-1}) = \mathbb{E}(\tilde{x}_{K}^{T}SK_{K}\tilde{x}_{K} + (1 - \nu_{K}(1 - a_{K})\lambda - \nu_{K}a_{K}A\lambda)\text{tr}(W + \eta A^{T}SK_{K+1}A)f^{0}(\tilde{P}) + \nu_{K}(1 - a_{K})\lambda + \nu_{K}a_{K}A\lambda)\text{tr}(S_{K+1}f^{0}(\tilde{P}) + (W + \eta A^{T}SK_{K+1} - S_{K})\tilde{P}) + \nu_{K}(1 - a_{K})\lambda - \nu_{K}a_{K}A\lambda + \eta \mathbb{E}(S_{K+1}Q),$$

where the coefficients of $\nu_{K}$ are all functions of $\tau_{K-1}$, showing that the optimal $\nu_{K}$ is independent of the network state estimate $\tilde{x}_{K}$. In addition, we may notice that $\mathbb{E}(\langle 1 - \gamma_{K} \rangle J_{K,\gamma_{K}=0}^{c} + \gamma_{K} J_{K,\gamma_{K}=1}^{c} | \tilde{x}_{K}, \tau_{K-1})$ is a linear function of both $\nu_{K}$ and $a_{K}$. According to the minimax theorem, we have

$$\min_{\nu_{K}(\tau_{K-1})} \max_{a_{K}(\tau_{K-1})} \mathbb{E}(\langle 1 - \gamma_{K} \rangle J_{K,\gamma_{K}=0}^{c} + \gamma_{K} J_{K,\gamma_{K}=1}^{c}) = \max_{\nu_{K}(\tau_{K-1})} \min_{a_{K}(\tau_{K-1})} \mathbb{E}(\langle 1 - \gamma_{K} \rangle J_{K,\gamma_{K}=0}^{c} + \gamma_{K} J_{K,\gamma_{K}=1}^{c}),$$

which is equivalent to $J_{K}^{c} = J_{K}^{a}$, verifying the existence of the stationary Nash equilibrium.

In a similar way, we can prove that the conclusion holds for any $k = 0, \ldots, K$.

C. Proof of Corollary [7]

Proof: Since both $\{v_{k}\}$ and $\{a_{k}\}$ solely depend on $\{\gamma_{k}\}$, the indicator variable $\gamma_{k}$ is independent of the system state. From [15, Theorem 1] and [21], the original minimax optimization problem is equivalent to

$$\min \max \min_{\{a_{k}\}} \sum_{k=0}^{K} \eta_{k} c_{k},$$

and the optimal solution to $\min_{\{a_{k}\}} \sum_{k=0}^{K} \eta_{k} c_{k}$ is in the form of Eq. (18). Accordingly, the optimal cost is in the form

$$\text{tr}(S_{K}P_{K}) + \sum_{k=0}^{K-1} \eta_{k+1} \text{tr}(S_{k+1}Q) + \sum_{k=0}^{K-1} \eta_{k} \text{tr}(\langle \eta A^{T}S_{k+1}A + W - S_{k} \rangle f^{0}(\tilde{P}) + \eta \mathbb{E}(S_{K+1}Q) + \eta \mathbb{E}(\nu_{K}u_{K} - c_{a}a_{K}).$$

Since the first two terms are irrelevant to $\{v_{k}\}$ and $\{a_{k}\}$, the solution to (19) is the same as that to (37).

D. Proof of Theorem 2

Proof: Define $J_{N}$ as the value of the stochastic game with truncated states, i.e.

$$J_{N} = \lim_{K \to \infty} \frac{1}{K} \sum_{k=0}^{K-1} \eta_{k} \left( \text{tr}(M_{K}\mathbb{E}(\min_{\{a_{k}\}} (\eta) \right) + \eta \mathbb{E}(S_{K+1}Q) + \eta \mathbb{E}(\nu_{K}u_{K} - c_{a}a_{K}),$$

where $M_{\infty} = A^{T}\eta S_{K}A + W_{\infty}$. Since $\pi_{N}^{c}$ and $\pi_{N}^{a}$ give the stationary Nash equilibrium, we have

$$\max_{\pi_{N}^{c}} J_{N}(\pi_{N}^{c}, \pi_{N}^{a}) \leq J_{N}(\pi_{N}^{c}, \pi_{N}^{a}) \leq \min_{\pi_{N}^{c}} J_{N}(\pi_{N}^{c}, \pi_{N}^{a}).$$

Next, we will prove Eq. (34) and (25) by showing that

$$|J_{N}(\pi_{N}^{c}, \pi_{N}^{a}) - J(\pi_{N}^{c}, \pi_{N}^{a})| < \frac{1}{2} \epsilon_{N},$$

$$|J_{N}(\pi_{N}^{c}, \pi_{N}^{a}) - J(\pi_{N}^{c}, \pi_{N}^{a})| < \frac{1}{2} \epsilon_{N},$$

$$|\max_{\pi_{N}^{c}} J_{N}(\pi_{N}^{c}, \pi_{N}^{a}) - \max_{\pi_{N}^{c}} J(\pi_{N}^{c}, \pi_{N}^{a})| < \frac{1}{2} \epsilon_{N}.$$
Here $\mathcal{P}_{(\nu,a)\rightarrow}$ is a short form for the transition probability
$\mathbb{P}(s_{k+1} = \tau' \mid s_k = \tau, \nu_k = \nu, a_k = a)$. Transmission and attack policies can be derived from the occupation measures according to

$$
\mathbb{P}(r_k, a_k \mid \tau) = \frac{\omega_N^\tau(\tau, (\nu, a))}{\sum_{\tau', \nu, a \in \{0,1\}} \omega_N^\tau(\tau, (\nu, a))},
$$

(43)

$$
J_N(\pi_N^\tau, \pi_N^\nu) = \sum_{0 \leq \tau \leq N-1} \sum_{\nu, a \in \{0,1\}} \omega_N^\tau(\tau, (\nu, a)) f^\tau(\tilde{P}).
$$

(44)

Similarly, we define $\omega(\tau, (\nu, a))$ as an occupation measure corresponding to the policy $\pi_N^\tau, \pi_N^\nu$ for the non-truncated stochastic game. Then, we have

$$
J(\pi_N^\tau, \pi_N^\nu) = \sum_{\tau \in \{0, \infty\}} \sum_{\nu, a \in \{0,1\}} \omega(\tau, (\nu, a)) f^\tau(\tilde{P}).
$$

(45)

Since the policies corresponding to $\{\omega_N^\tau(\tau, (\nu, a))\}$ and $\{\omega(\tau, (\nu, a))\}$ are the same and the number of states are different, from [17], we have

$$
\omega(\tau, (\nu, a)) = \omega_N^\tau(\tau, (\nu, a)), \quad \tau \in [0, N),
$$

$$
\omega(i, (\nu, a)) = \omega_N^\tau(\tau, (\nu, a)), \quad \tau = N.
$$

According to Eq. (20), $J(\pi_N^\tau, \pi_N^\nu)$ is bounded, indicating that

$$
\lim_{N \to \infty} \sum_{\tau \in \{0, \infty\}} \sum_{\nu, a \in \{0,1\}} \omega(\tau, (\nu, a)) f^\tau(\tilde{P}) = 0.
$$

Compare Eq. (43) and Eq. (45). We can prove that for any scalar $\frac{1}{2}c_N > 0$, there exists an integer $N$ such that Eq. (39) holds.

Similarly, we fix the attack policy as $\pi_N^\nu$. Define the occupation measure corresponding to the policy $\pi_N^\tau$ as $\omega_N^\tau(\tau, \nu)$ for the truncated game and the occupation measure corresponding to the policy $\pi^\nu$ as $\omega^\nu(\tau, \nu)$ for the non-truncated game. The value of $\min_{\pi_N^\tau} J_N(\pi_N^\tau, \pi_N^\nu)$ is equivalent to the value of

$$
\tilde{J} = \min_{\pi^\nu} \sum_{0 \leq \tau \leq N-1} \sum_{\nu, a \in \{0,1\}} \omega_N^\tau(\tau, \nu) f^\tau(\tilde{P}),
$$

(46)

and the value of $\min_{\pi^\nu} J(\pi^\tau, \pi^\nu)$ is equivalent to the value of

$$
\tilde{J} = \min_{\pi^\nu} \sum_{\tau \in \{0, \infty\}} \sum_{\nu, a \in \{0,1\}} \omega^\nu(\tau, \nu) f^\tau(\tilde{P}).
$$

(47)

Denote the optimal solution to the problem in [47] as $\omega^\nu(\tau, \nu)$, from which we can construct a group of feasible solutions to the problem in [46] as

$$
\omega_N^\tau(\tau, \nu) = \omega^\nu(\tau, \nu), \quad \tau \in [0, N),
$$

$$
\omega_N^\tau(\tau, \nu) = \sum_{i=0}^{\infty} \omega^\nu(i, \nu), \quad \tau = N.
$$

(48)

Since the value of

$$
\sum_{0 \leq \tau \leq N-1} \sum_{\nu, a \in \{0,1\}} \omega_N(\tau, \nu) f^\tau(\tilde{P})
$$

corresponding to the feasible solution in [48] is smaller than $\tilde{J}$, according to the optimality of $\tilde{J}$, we have $\tilde{J} \leq J$. In the meantime, from the optimal policy derived from [46], we can construct a feasible solution to [45] using [43]. Denote the corresponding value as $\tilde{J}_f$. The optimality of $\tilde{J}$ gives that $\tilde{J} \leq \tilde{J}_f$. Similar to the proof of [39], we can prove that $\tilde{J}_N - \tilde{J}_f \leq \frac{1}{2}c_N$, which further establishes [40]. The bound in [41] can be proved similarly.

REFERENCES

[1] K.-D. Kim and P. R. Kumar, “Cyber–physical systems: A perspective,” in the centennial issue, IEEE transactions on Automatic Control, vol. 100, no. Special Centennial Issue, pp. 1287–1308, 2012.

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

[3] Y. Mo and B. Sinopoli, “Secure control against replay attacks,” in 2009 47th annual Allerton conference on communication, control, and computing (Allerton). IEEE, 2009, pp. 911–918.

[4] H. Zhang, P. Cheng, L. Shi, and J. Chen, “Optimal dos attack scheduling in wireless networked control system,” IEEE Transactions on Control Systems Technology, vol. 24, no. 1, pp. 843–852, 2015.

[5] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, “Game theory meets network security and privacy,” ACM Computing Surveys (CSUR), vol. 45, no. 3, pp. 1–39, 2013.

[6] Q. Zhu and T. Basar, “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 Systems Magazine, vol. 35, no. 1, pp. 46–65, 2015.

[7] Y. Li, D. E. Quevedo, S. Dey, and L. Shi, “Snr-based dos attack on remote state estimation: A game-theoretic approach,” IEEE Transactions on Control Systems Technology, vol. 4, no. 3, pp. 632–642, 2017.

[8] J. Zhang, J. Sun, and C. Zhang, “Stochastic game in linear quadratic gaussian control for wireless networked control systems under dos attacks,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020.

[9] F. Miao, Q. Zhu, M. Pajic, and J. G. Pappas, “A hybrid stochastic game for secure control of cyber-physical systems,” Automatica, vol. 93, pp. 55–63, 2018.

[10] D. Antunes, W. Heemels, and P. Tabuada, “Dynamic programming formulation of periodic event-triggered control: Performance guarantees and co-design,” in 2012 IEEE 51st IEEE conference on decision and control (CDC). IEEE, 2012, pp. 7212–7217.

[11] E. G. Peters, D. E. Quevedo, and M. Fu, “Controller and scheduler co-design for feedback control over ieee 802.15.4 networks,” IEEE Transactions on Control Systems Technology, vol. 24, no. 6, 2016.

[12] M. Kögel, D. Quevedo, and R. Findeisen, “Combined control and communication scheduling for constrained system using robust output feedback mpc,” in 2019 18th European Control Conference (ECC). IEEE, 2019, pp. 1778–1783.

[13] S. Wölfhagen and F. Allgöwer, “Scheduling and control over networks using mpc with time-varying terminal ingredients,” in 2020 American Control Conference (ACC). IEEE, 2020, pp. 1913–1918.

[14] S. Kuhn and S. Dey, “Optimal energy allocation for linear control with packet loss under energy harvesting constraints,” Automatica, vol. 77, pp. 259–267, 2017.

[15] A. S. Leong, D. E. Quevedo, T. Tanaka, S. Dey, and A. Ahlén, “Event-based transmission scheduling and lag control over a packet dropping link,” IFAC-PapersOnLine, vol. 50, no. 1, pp. 8945–8950, 2017.

[16] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, “Kalman filtering with intermittent observations,” IEEE transactions on Automatic Control, vol. 49, no. 9, pp. 1453–1464, 2004.

[17] W. Li, Z. Wang, Q. Liu, and L. Guo, “An information aware event-triggered scheme for particle filter based remote state estimation,” Automatica, vol. 103, pp. 151–158, 2019.
[18] E. Kung, J. Wang, J. Wu, D. Shi, and L. Shi, “On the nonexistence of event triggers that preserve gaussian state in presence of packet-drop,” IEEE Transactions on Automatic Control, vol. 65, no. 10, pp. 4302–4307, 2019.

[19] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, “Foundations of control and estimation over losoy networks,” Proceedings of the IEEE, vol. 95, no. 1, pp. 163–187, 2007.

[20] V. Gupta, B. Hassibi, and R. M. Murray, “Optimal lqg control across packet-dropping links,” Systems & Control Letters, vol. 56, no. 6, pp. 439–446, 2007.

[21] D. P. Bertsekas, “Dynamic programming and optimal control 3rd edition, volume ii,” Belmont, MA: Athena Scientific, 2011.

[22] E. Altman, Constrained Markov decision processes. CRC Press, 1999, vol. 7.