

# A Game Theoretic Analysis of LQG Control under Adversarial Attack

Zuxing Li, György Dán, and Dong Liu

**Abstract**—Motivated by recent works addressing adversarial attacks on deep reinforcement learning, a deception attack on linear quadratic Gaussian control is studied in this paper. In the considered attack model, the adversary can manipulate the observation of the agent subject to a mutual information constraint. The adversarial problem is formulated as a novel dynamic cheap talk game to capture the strategic interaction between the adversary and the agent, the asymmetry of information availability, and the system dynamics. Necessary and sufficient conditions are provided for subgame perfect equilibria to exist in pure strategies and in behavioral strategies; and characteristics of the equilibria and the resulting control rewards are given. The results show that pure strategy equilibria are informative, while only babbling equilibria exist in behavioral strategies. Numerical results are shown to illustrate the impact of strategic adversarial interaction.

## I. INTRODUCTION

Deep reinforcement learning (DRL) has recently emerged as a promising solution for solving large Markov decision processes (MDPs) and partially observable MDPs (POMDPs), thanks to deep neural networks used as policy approximators [2]. DRL has, however, been shown to be vulnerable to small perturbations of the state observation, called adversarial examples, which were found to mislead the control agent to take suboptimal control actions [3]. While there has been a significant recent interest in the design of adversarial examples against DRL [3], [4], [5], [6], [9], there has been little work on characterizing the ability of agents to adapt to those.

Recent work proposed to use adversarial examples for making DRL agents more robust to perturbations, by letting the adversary and the agent play against each other, and formulating the interaction as a stochastic game (SG) [8]. Nonetheless, in the case of adversarial examples the agent cannot observe the system state directly, neither can the adversary affect the state transition probabilities directly, only through the actions taken by the agent. Hence, the SG model does not capture the information structure of the problem. Effectively, in the presence of adversarial examples the agent has to solve a POMDP, where the observations are subverted by the adversary so as to mislead the agent.

As a model of this interaction, in this paper we propose a game theoretical model to study the strategic interaction between an agent that has to solve a linear quadratic Gaussian (LQG) control problem, and an adversary that can manipulate the agent’s observations by a randomly

chosen affine transformation subject to a mutual information constraint, and aims at minimizing the control reward. The resulting problem is formulated as a dynamic cheap talk game, which captures information asymmetry, the beliefs of the adversary and the agent, and the undetectability constraint imposed on the adversarial attacks.

Our paper contributes to the solution of the formulated game theoretical problem in two ways. First, we address necessary and sufficient conditions for the existence of subgame perfect equilibria (SPEs) in pure strategies and in behavioral strategies, and we characterize the equilibrium strategies. Second, we characterize the rewards achievable in equilibria, and relate them to the rewards of a naive agent and an alert agent under attack. The key novelty of our contribution is that we characterize the strategies to be followed by the agent and by the adversary under strategic interaction, which has not been addressed by the existing literature.

The rest of the paper is organized as follows. In Section II we review related work. In Section III we present the system model and problem formulation. In Section IV we provide analytical results. In Section V we provide numerical results. Section VI concludes the paper.

**Notation:** Unless otherwise specified, we denote a random variable by a capital letter and its realization by the corresponding lower-case letter. We denote by  $\mathcal{N}(\cdot, \cdot)$  the Gaussian distribution, by  $\mathbb{S}(\cdot)$  the support set, by  $I(\cdot; \cdot)$  the mutual information, and by  $|\cdot|$  the cardinality of a set.

## II. RELATED WORK

Related to our work are previous researches on robust POMDP under uncertainty of the system dynamics [7]. In [7] the control action was optimized under the worst-case assumption of the system dynamics in each stage, i.e., the agent plays as the leader and the dynamic system plays as the follower in a Stackelberg game. SG and partially observable SG (POSG) were used to model the strategic interaction of players in a dynamic system, and have been employed in robust and adversarial problems [8], [9], [10]. But unlike in the case of learning under adversarial attacks, in SG and in POSG the players interact with each other through the impact of their actions on the state transitions, not on the state observations. Our work is related to the cheap talk game [19], where a sender with private information sends a message to a receiver and the receiver takes an action based on the received message and based on its belief on the inaccessible private information. Closest to our model are [11], [12], [13]. In [11] a dynamic cheap talk game was proposed to study a deception attack on a Markovian system, where the actions

This work was partly supported by the Swedish Foundation for Strategic Research (SSF) through the CLAS project and by MSB through the CERCES project.

Z. Li, G. Dán, and D. Liu are with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden {zuxing;gyuri;doli}@kth.se

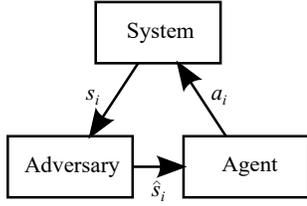


Fig. 1. Considered adversarial attack on dynamic system control. In the  $i$ -th stage, the adversary observes the true system state  $s_i$  and presents the manipulated state  $\hat{s}_i$  to the agent. The agent does not have access to the true state  $s_i$  but takes an action  $a_i$  upon observing  $\hat{s}_i$ .

do not affect the state transitions. In [12] authors developed a dynamic game model of the attacker-defender interaction, and characterized the optimal attack strategy as a function of the defense strategy, allowing for a static optimal defense strategy. In our preliminary work [13] we proposed a dynamic cheap talk framework to model deception attacks on a general MDP, and addressed computational issues.

Adversarial variants of LQG control were considered in a number of recent works. A Stackelberg game was formulated in [14], where the dynamic system is the leader, while the agent is the follower and may be an adversary. The authors formulated a finite horizon hierarchical signaling game between the sender and the receiver in a dynamic environment and showed that linear sender and receiver strategies can yield the equilibrium [15]. In [16], the opposite problem was studied but without considering the complete strategic interaction, where the adversary optimally manipulates the control actions instead of the system states. The optimal attack on both the system state and the control action in LQG control was studied in [17]. In [18], a targeted attack strategy was studied to mislead the LQG system to a particular state while evading detection.

### III. ADVERSARIAL LQG CONTROL PROBLEM

We consider an  $N$ -stage LQG control problem under adversarial attack, as illustrated in Fig. 1. The system states  $\{s_i\}_{i=1}^N$ , the manipulated states  $\{\hat{s}_i\}_{i=1}^N$ , the actions  $\{a_i\}_{i=1}^N$ , and the instantaneous rewards  $\{r_i\}_{i=1}^N$  for  $1 \leq i \leq N$  are described by

$$s_{i+1} = \alpha_i s_i + \beta_i a_i + z_i, \text{ given } \alpha_i \neq 0, \beta_i \neq 0; \quad (1)$$

$$\hat{s}_i = \pi_i s_i + c_i; \quad (2)$$

$$a_i = \kappa_i \hat{s}_i + \rho_i; \quad (3)$$

$$r_i = R_i(s_i, a_i) = -\theta_i s_i^2 - \phi_i a_i^2, \text{ given } \theta_i > 0, \phi_i > 0; \quad (4)$$

$$S_1 \sim b_1 \triangleq \mathcal{N}(\mu_1, \sigma_1^2), \text{ given } \mu_1, \sigma_1^2 > 0; \quad (5)$$

$$Z_i \sim \mathcal{N}(0, \omega_i^2), \text{ given } \omega_i^2 > 0; \quad (6)$$

$$C_i \sim \mathcal{N}(0, \delta_i^2). \quad (7)$$

#### A. LQG Recapitulation

If  $\{\pi_i\}_{i=1}^N$  and  $\{\delta_i^2\}_{i=1}^N$  are known by the agent, the above problem is a standard LQG control. In the  $i$ -th stage, the agent observes  $\hat{s}_i$  but not  $s_i$ , and determines the action  $a_i$  with the aim to maximize its expected accumulated reward.

Note that it is sufficient to consider an affine function of  $\hat{s}_i$  for  $a_i$  in the standard LQG problem as the optimal action  $a_i^*$  is a linear function of the mean of the Gaussian posterior distribution of  $S_i$  for the agent after observing  $\{\hat{s}_k\}_{k=1}^i$  and  $\{a_k\}_{k=1}^{i-1}$  [21]. To compute the optimal coefficients  $\kappa_i^*$  and  $\rho_i^*$ , we first define  $\tilde{\theta}_{N+1} = 0$ , and for  $1 \leq i \leq N$  define  $\tilde{\theta}_i$  as

$$\tilde{\theta}_i = \theta_i + \tilde{\theta}_{i+1} \alpha_i^2 - \frac{\tilde{\theta}_{i+1}^2 \alpha_i^2 \beta_i^2}{\phi_i + \tilde{\theta}_{i+1} \beta_i^2}. \quad (8)$$

Furthermore, we denote by  $b_i \triangleq \mathcal{N}(\mu_i, \sigma_i^2)$  the belief of the agent about  $S_i$ , which is the Gaussian posterior distribution of  $S_i$  for the agent after observing  $\{\hat{s}_k\}_{k=1}^{i-1}$  and  $\{a_k\}_{k=1}^{i-1}$ . Then, given the manipulated state  $\hat{s}_i$ , the optimal action can be expressed as

$$\kappa_i^* = -\frac{\tilde{\theta}_{i+1} \alpha_i \beta_i \pi_i \sigma_i^2}{(\phi_i + \tilde{\theta}_{i+1} \beta_i^2)(\pi_i^2 \sigma_i^2 + \delta_i^2)}, \quad (9)$$

$$\rho_i^* = -\frac{\tilde{\theta}_{i+1} \alpha_i \beta_i \mu_i \delta_i^2}{(\phi_i + \tilde{\theta}_{i+1} \beta_i^2)(\pi_i^2 \sigma_i^2 + \delta_i^2)}, \quad (10)$$

$$a_i^* = \kappa_i^* \hat{s}_i + \rho_i^* = -\frac{\tilde{\theta}_{i+1} \alpha_i \beta_i}{\phi_i + \tilde{\theta}_{i+1} \beta_i^2} \frac{\pi_i \sigma_i^2 \hat{s}_i + \mu_i \delta_i^2}{\pi_i^2 \sigma_i^2 + \delta_i^2}, \quad (11)$$

where the coefficients  $\kappa_i^*$  and  $\rho_i^*$  depend on  $b_i$ ;  $\frac{\pi_i \sigma_i^2 \hat{s}_i + \mu_i \delta_i^2}{\pi_i^2 \sigma_i^2 + \delta_i^2}$  is the mean of the Gaussian posterior distribution of  $S_i$  for the agent after observing  $\{\hat{s}_k\}_{k=1}^i$  and  $\{a_k\}_{k=1}^{i-1}$ . Note that when  $\pi_i \equiv 1$  and  $\delta_i^2 \equiv 0$  the LQG strategy reduces to the linear quadratic regulator (LQR) strategy

$$(\kappa_i^*, \rho_i^*) = \left( -\frac{\tilde{\theta}_{i+1} \alpha_i \beta_i}{\phi_i + \tilde{\theta}_{i+1} \beta_i^2}, 0 \right). \quad (12)$$

#### B. Adversarial Model

The adversary can manipulate the observation of the agent and its objective is to minimize the agent's expected accumulated reward, similar to [3], [6]. In the  $i$ -th stage, the adversary chooses the manipulation parameters  $\pi_i$ ,  $\delta_i^2$ , manipulates the state  $s_i$  to  $\hat{s}_i$ , and then reports the manipulated state  $\hat{s}_i$  to the agent. We consider that the adversarial manipulation is "small", since a large manipulation may be easily detected and may also involve a high manipulation cost. Given the agent's belief  $b_i \triangleq \mathcal{N}(\mu_i, \sigma_i^2)$ , we impose the following constraints on the manipulation:

$$-\infty < \varepsilon' \leq \pi_i \leq \varepsilon < \infty; \quad (13)$$

$$I(\hat{S}_i; S_i) = \frac{1}{2} \log \frac{\pi_i^2 \sigma_i^2 + \delta_i^2}{\delta_i^2} \geq \frac{1}{2} \log \lambda > 0, \forall \pi_i, \delta_i^2, \quad (14)$$

i.e.,  $\frac{\pi_i^2 \sigma_i^2 + \delta_i^2}{\delta_i^2} \geq \lambda > 1$ . The mutual information constraint (14) implies that the manipulated state conveys at least a certain amount of information about the system state to the agent. A larger value of  $\lambda$  means a weaker adversary, and vice versa. Note that in order to satisfy the mutual information constraint, the adversary *cannot* use  $\pi_i = 0$ . We denote by  $\mathcal{A}_i(b_i, \varepsilon', \varepsilon, \lambda)$  the set of feasible adversarial actions  $(\pi_i, \delta_i^2)$  in the  $i$ -th stage subject to (13)-(14). Finally, in the end of



Thus the SPE has to satisfy (19), which is the basis for the analysis we present in the following.

### B. Pure Strategy Equilibria

We start the analysis considering pure strategy equilibria. With slight abuse of notation, we denote by  $(\pi_i, \delta_i^2) = g_i(b_i)$  a pure strategy of the adversary as a function of the belief  $b_i$ .

We first consider the case  $N = 1$ .

*Proposition 1:* Let  $N = 1$ . An SPE consists of  $(f_1^*, g_1^*)$ , where  $(\kappa_1^*, \rho_1^*) = f_1^*(b_1) = (0, 0)$  for any belief  $b_1$ ; and  $g_1^*$  can be any adversarial strategy defined on  $\mathcal{A}_1(b_1, \varepsilon', \varepsilon, \lambda)$ .

*Proof:* Observe that  $(\kappa_1^*, \rho_1^*) = f_1^*(b_1) = (0, 0)$  is a dominant strategy for the agent for any belief  $b_1$ . Under this strategy, the adversarial strategy has no impact on the agent's reward. This proves the result. ■

The existence of a pure strategy SPE for  $N = 1$  is encouraging, even if the equilibrium is degenerate. Unfortunately, for  $N \geq 2$  an SPE may not exist as shown in the following theorem.

*Theorem 1:* Let  $N \geq 2$ . If  $\varepsilon' \neq \varepsilon$  or if  $\varepsilon' = \varepsilon = 0$ , then there is no pure strategy SPE for the ALQG game. If  $\varepsilon' = \varepsilon \neq 0$ , then there is a unique pure strategy SPE. The SPE strategies for  $1 \leq i \leq N$  are given by

$$\tilde{\theta}_{N+1} = \hat{\theta}_{N+1} = 0; \quad (20)$$

$$\tilde{\theta}_i = \theta_i + \tilde{\theta}_{i+1} \alpha_i^2 - \frac{\tilde{\theta}_{i+1}^2 \alpha_i^2 \beta_i^2}{\phi_i + \tilde{\theta}_{i+1} \beta_i^2}; \quad (21)$$

$$\hat{\theta}_i = \theta_i + \hat{\theta}_{i+1} \alpha_i^2 - \left( \frac{\tilde{\theta}_{i+1}^2 \alpha_i^2 \beta_i^2}{\phi_i + \tilde{\theta}_{i+1} \beta_i^2} + (\hat{\theta}_{i+1} - \tilde{\theta}_{i+1}) \alpha_i^2 \right) \frac{\lambda - 1}{\lambda}; \quad (22)$$

$$\pi_i^* = g_i^*(b_i) = \varepsilon' = \varepsilon; \quad (23)$$

$$\delta_i^{2*} = g_i^*(b_i) = \frac{\varepsilon^2 \sigma_i^2}{\lambda - 1}; \quad (24)$$

$$\kappa_i^* = f_i^*(b_i) = -\frac{\tilde{\theta}_{i+1} \alpha_i \beta_i (\lambda - 1)}{(\phi_i + \tilde{\theta}_{i+1} \beta_i^2) \lambda \varepsilon}; \quad (25)$$

$$\rho_i^* = f_i^*(b_i) = -\frac{\tilde{\theta}_{i+1} \alpha_i \beta_i \mu_i}{(\phi_i + \tilde{\theta}_{i+1} \beta_i^2) \lambda}. \quad (26)$$

*Corollary 1:* If  $\varepsilon' = \varepsilon \neq 0$ , the value function induced by the unique pure strategy SPE is

$$V_i^N(b_i) = -\tilde{\theta}_i \mu_i^2 - \hat{\theta}_i \sigma_i^2 - \sum_{j=i+1}^N \hat{\theta}_j \omega_{j-1}^2. \quad (27)$$

The proofs of Theorem 1 and Corollary 1 are provided in the appendix.

*Property 2:* It follows from (20)-(22) that for  $1 \leq i \leq N$ ,

$$\hat{\theta}_i \geq \tilde{\theta}_i > 0. \quad (28)$$

*Remark 2:* We can make the following observations based on Theorem 1 and Corollary 1.

- Since the LQG control can be seen as the best response of any given (adversarial manipulated) observation model, it is sufficient to consider a pure agent strategy in the form of an affine function for an SPE of the ALQG game with a pure adversarial strategy.

- It follows from (24) that a rational adversary will always apply a manipulation with the largest variance.
- The value function  $V_i^N$  consists of a constant term and two separable terms depending on the mean  $\mu_i$  and the variance  $\sigma_i^2$  of the belief, which allows a closed form solution for arbitrary  $N$ .

**Time-invariant system:** We now turn to the asymptotic analysis of a time-invariant system, i.e.,  $\alpha_i = \alpha \neq 0$ ,  $\beta_i = \beta \neq 0$ ,  $\omega_i^2 = \omega^2 > 0$ ,  $\theta_i = \theta > 0$ , and  $\phi_i = \phi > 0$  for  $i \geq 1$ , and we let  $N \rightarrow \infty$ . Let us define the mapping  $L: \mathbb{R}_{\geq 0}^2 \rightarrow \mathbb{R}_{\geq 0}^2$  as

$$L(x, y) = \left( \theta + \frac{\phi \alpha^2 x}{\phi + \beta^2 x}, \theta + \frac{\phi \alpha^2 x}{\phi + \beta^2 x} \frac{\lambda - 1}{\lambda} + \alpha^2 y \frac{1}{\lambda} \right). \quad (29)$$

Observe that  $L$  is effectively the coefficient update (20)-(22) for the time-invariant model. In what follows, we first characterize  $L$  and furthermore the pure strategy SPE of the ALQG game in the asymptotic regime.

*Proposition 2:* Let  $\lambda > \alpha^2$ . Then the mapping  $L$  admits a least fixed point  $(\tilde{\theta}, \hat{\theta}) \in \mathbb{R}_{\geq 0}^2$ , for which

$$\lim_{n \rightarrow \infty} L^n(0, 0) = L(\tilde{\theta}, \hat{\theta}) = (\tilde{\theta}, \hat{\theta}), \quad (30)$$

with  $L^n(0, 0) \triangleq \underbrace{L(L(\dots(L(L(0, 0))))}_{n \text{ L-mappings}} \dots$ .

*Proof:* We start the proof by observing that the mapping  $L$  is order-preserving. That is, for all  $(x, y), (x', y') \in \mathbb{R}_{\geq 0}^2$  satisfying  $(x, y) \preceq (x', y')$ , i.e.,  $x \leq x'$  and  $y \leq y'$ , we have  $L(x, y) \preceq L(x', y')$ . This can be easily shown by analyzing (29).

Furthermore, the fixed point equation  $L(\tilde{\theta}, \hat{\theta}) = (\tilde{\theta}, \hat{\theta})$  has a unique solution on  $\mathbb{R}_{\geq 0}^2$  if and only if  $\lambda > \alpha^2$ . Since  $L$  is order-preserving, the convergence result  $\lim_{n \rightarrow \infty} L^n(0, 0) = (\tilde{\theta}, \hat{\theta})$  follows from Kleene's fixed point theorem [22]. ■

Analytical expressions for  $\tilde{\theta}$  and  $\hat{\theta}$  can be obtained by solving the fixed point equation  $L(\tilde{\theta}, \hat{\theta}) = (\tilde{\theta}, \hat{\theta})$ , and can be used for characterizing the SPE in pure strategies, using Theorem 1 and Proposition 2, as follows.

*Theorem 2:* Let  $\lambda > \alpha^2$ ,  $\varepsilon' = \varepsilon \neq 0$ , and  $N \rightarrow \infty$ . Then the ALQG game of the time-invariant model has a stationary SPE in pure strategies as: For  $i \geq 1$ ,

$$\pi_i^* = g_i^*(b_i) = \varepsilon; \quad (31)$$

$$\delta_i^{2*} = g_i^*(b_i) = \frac{\varepsilon^2 \sigma_i^2}{\lambda - 1}; \quad (32)$$

$$\kappa_i^* = f_i^*(b_i) = -\frac{\tilde{\theta} \alpha \beta (\lambda - 1)}{(\phi + \tilde{\theta} \beta^2) \lambda \varepsilon}; \quad (33)$$

$$\rho_i^* = f_i^*(b_i) = -\frac{\tilde{\theta} \alpha \beta \mu_i}{(\phi + \tilde{\theta} \beta^2) \lambda}. \quad (34)$$

*Proof:* Since  $\varepsilon' = \varepsilon \neq 0$ , a unique SPE in pure strategies exists and the SPE strategies are given in Theorem 1. Since  $\lambda > \alpha^2$  and  $N \rightarrow \infty$ , it follows from Proposition 2 that  $\tilde{\theta}_i, \hat{\theta}_i$  converge to  $\tilde{\theta}, \hat{\theta}$ , respectively. This leads to the stationary SPE in Theorem 2. ■

Interestingly, for this stationary SPE in pure strategies we can obtain the expected average reward per stage in steady state in closed form.

*Corollary 2:* Let  $b_1 \triangleq \mathcal{N}(\mu_1, \sigma_1^2)$  with *bounded* mean and variance. For the stationary SPE in pure strategies in Theorem 2, the expected average reward per stage in steady state is independent of the initial belief and is given by

$$\lim_{N \rightarrow \infty} \frac{V_1^N(b_1)}{N} = -\hat{\theta} \omega^2. \quad (35)$$

*Proof:* This result follows from Corollary 1, Proposition 2, and Theorem 2. ■

### C. Equilibria in Behavioral Strategies

The previous results show that a pure strategy SPE does not exist if there are multiple choices of the coefficient  $\pi_i$  for the adversary. We thus turn to the analysis of SPE in behavioral strategies.

*Theorem 3:* Let  $N \geq 2$  and  $\varepsilon' < 0 < \varepsilon$ . Then for  $1 \leq i \leq N$ ,  $\check{\theta}_i$  and  $\check{\theta}_i$  are as given by (21), and

$$\check{\theta}_{N+1} = \check{\theta}_{N+1} = 0, \quad (36)$$

$$\check{\theta}_i = \theta_i + \check{\theta}_{i+1} \alpha_i^2 - (\check{\theta}_{i+1} - \check{\theta}_{i+1}) \alpha_i^2 \frac{\lambda - 1}{\lambda}. \quad (37)$$

Furthermore, there is a continuum of SPEs in behavioral strategies. Each SPE in the  $i$ -th stage consists of a behavioral strategy of the adversary and a pure strategy of the agent that satisfies

$$\mathbb{S}(g_i^* | b_i) \triangleq \left\{ (\pi_i, \delta_i^2) : \delta_i^2 = \frac{\pi_i^2 \sigma_i^2}{\lambda - 1} \right\} \subseteq \mathcal{A}_i(b_i, \varepsilon', \varepsilon, \lambda); \quad (38)$$

$$\|\mathbb{S}(g_i^* | b_i)\| \geq 2; \quad (39)$$

$$E_{g_i^*}(\Pi_i) = 0; \quad (40)$$

$$(\kappa_i^*, \rho_i^*) = f_i^*(b_i) = \left( 0, -\frac{\check{\theta}_{i+1} \alpha_i \beta_i \mu_i}{\phi_i + \check{\theta}_{i+1} \beta_i^2} \right). \quad (41)$$

Interestingly, this SPE is a babbling equilibrium in which the agent's action is based on its belief, not on the manipulated state. Nonetheless, the value of the game depends on the adversarial manipulation, as shown in the following corollary.

*Corollary 3:* Let  $\varepsilon' < 0 < \varepsilon$ . For any SPE in behavioral strategies, we have

$$V_i^N(b_i) = -\check{\theta}_i \mu_i^2 - \check{\theta}_i \sigma_i^2 - \sum_{j=i+1}^N \check{\theta}_j \omega_{j-1}^2. \quad (42)$$

The proof ideas of Theorem 3 and Corollary 3 are to justify that the strategies in Theorem 3 satisfy the dynamic programming (19) and to show that these strategies lead to the value function in Corollary 3. The complete proofs can be found in the appendix of the full version [1].

*Property 3:* It follows from the update rules (20)-(22) and (36)-(37) that for all  $1 \leq i \leq N$ ,

$$\check{\theta}_i \geq \hat{\theta}_i \geq \check{\theta}_i > 0. \quad (43)$$

*Remark 3:* We can make the following observations based on Theorem 3 and Corollary 3.

- It is sufficient for the agent to use a pure affine strategy against a behavioral strategy of the adversary.
- Although the adversary cannot use  $\pi_i = 0$ , the behavioral strategy  $g_i^*$  needs to achieve zero-mean of the random coefficient  $\Pi_i$ .
- A rational adversary will always use a manipulation with the largest variance.
- From Property 3, the value (42) of an SPE in behavioral strategies is always less than or equal to the value (27) of a pure strategy SPE.

**Time-invariant system:** We again turn to the time-invariant system for  $N \rightarrow \infty$ . Let us define the mapping  $J: \mathbb{R}_{\geq 0}^2 \rightarrow \mathbb{R}_{\geq 0}^2$  as

$$J(x, y) = \left( \theta + \frac{\phi \alpha^2 x}{\phi + \beta^2 x}, \theta + \alpha^2 x \frac{\lambda - 1}{\lambda} + \alpha^2 y \frac{1}{\lambda} \right). \quad (44)$$

Observe that  $J$  is effectively the coefficient update rule (21), (36), (37) for the time-invariant system. In what follows, we characterize  $J$  and the stationary SPEs in behavioral strategies for the ALQG game.

*Proposition 3:* Let  $\lambda > \alpha^2$ . Then the mapping  $J$  admits a least fixed point  $(\check{\theta}, \check{\theta}) \in \mathbb{R}_{\geq 0}^2$ , for which

$$\lim_{n \rightarrow \infty} J^n(0, 0) = J(\check{\theta}, \check{\theta}) = (\check{\theta}, \check{\theta}). \quad (45)$$

The proof of Proposition 3 follows using the arguments in the proof of Proposition 2.

*Theorem 4:* Let  $\lambda > \alpha^2$ ,  $\varepsilon' < 0 < \varepsilon$ , and  $N \rightarrow \infty$ . Then the ALQG game of the time-invariant model has a stationary SPE in behavioral strategies as: For  $i \geq 1$ ,

$$g_i^* \left( \pi_i = \varepsilon', \delta_i^2 = \frac{\varepsilon'^2 \sigma_i^2}{\lambda - 1} \middle| b_i \right) = \frac{\varepsilon}{\varepsilon - \varepsilon'}; \quad (46)$$

$$g_i^* \left( \pi_i = \varepsilon, \delta_i^2 = \frac{\varepsilon^2 \sigma_i^2}{\lambda - 1} \middle| b_i \right) = -\frac{\varepsilon'}{\varepsilon - \varepsilon'}; \quad (47)$$

$$(\kappa_i^*, \rho_i^*) = f_i^*(b_i) = \left( 0, -\frac{\check{\theta} \alpha \beta \mu_i}{\phi + \check{\theta} \beta^2} \right). \quad (48)$$

*Corollary 4:* Let  $b_1 \triangleq \mathcal{N}(\mu_1, \sigma_1^2)$  with *bounded* mean and variance. For the stationary SPE in behavioral strategies in Theorem 4, the expected average reward per stage in steady state is independent of the initial belief and is given by

$$\lim_{N \rightarrow \infty} \frac{V_1^N(b_1)}{N} = -\check{\theta} \omega^2. \quad (49)$$

The proofs of Theorem 4 and Corollary 4 are based on Theorem 3, Corollary 3, and Proposition 3, and follow from similar arguments as used in the proofs of Theorem 2 and Corollary 2. Observe that Corollary 2, Corollary 4, and Property 3 jointly imply that the expected average agent reward per stage in steady state is higher when considering pure strategies, as behavioral strategies allow for more uncertainty about the attack and thus make the adversary stronger.

The behavioral strategy of the adversary in Theorem 3 needs to achieve zero-mean of the random coefficient  $\Pi_i$ , which *cannot* be satisfied if  $0 \leq \varepsilon' < \varepsilon$  or  $\varepsilon' < \varepsilon \leq 0$ . In the following, we study SPE under these conditions.

*Theorem 5:* Let  $N \geq 2$ . If  $0 = \varepsilon' < \varepsilon$  or if  $\varepsilon' < \varepsilon = 0$ , there is no SPE for the ALQG game.

*Theorem 6:* Let  $N = 2$ . If  $0 < \varepsilon' < \varepsilon$  or if  $\varepsilon' < \varepsilon < 0$ , there is a unique SPE in behavioral strategies for the ALQG game: For any belief  $b_1 \triangleq \mathcal{N}(\mu_1, \sigma_1^2)$ ,

$$g_1^* \left( \pi_1 = \varepsilon', \delta_1^2 = \frac{\varepsilon'^2 \sigma_1^2}{\lambda - 1} \middle| b_1 \right) = \frac{\varepsilon}{\varepsilon' + \varepsilon}; \quad (50)$$

$$g_1^* \left( \pi_1 = \varepsilon, \delta_1^2 = \frac{\varepsilon^2 \sigma_1^2}{\lambda - 1} \middle| b_1 \right) = \frac{\varepsilon'}{\varepsilon' + \varepsilon}; \quad (51)$$

$$\kappa_1^* = f_1^*(b_1) \quad (52)$$

$$= \frac{-\theta_2 \alpha_1 \beta_1 E_{g_1^*}(\Pi_1) \sigma_1^2}{(\phi_1 + \theta_2 \beta_1^2) \left( E_{g_1^*}(\Pi_1^2) \left( \mu_1^2 + \frac{\lambda}{\lambda-1} \sigma_1^2 \right) - E_{g_1^*}^2(\Pi_1) \mu_1^2 \right)}; \quad (53)$$

$$\rho_1^* = f_1^*(b_1) = -E_{g_1^*}(\Pi_1) \mu_1 \kappa_1^* - \frac{\theta_2 \alpha_1 \beta_1 \mu_1}{\phi_1 + \theta_2 \beta_1^2}; \quad (54)$$

for any belief  $b_2 \triangleq \mathcal{N}(\mu_2, \sigma_2^2)$ ,  $g_2^* = g_1^*$ ;  $(\kappa_2^*, \rho_2^*) = f_2^*(b_2) = (0, 0)$ ; and

$$\begin{aligned} & V_1^2(b_1) \\ &= - \left( \theta_1 + \theta_2 \alpha_1^2 - \frac{\theta_2^2 \alpha_1^2 \beta_1^2}{\phi_1 + \theta_2 \beta_1^2} \right) \mu_1^2 - \theta_2 \omega_1^2 - (\theta_1 + \theta_2 \alpha_1^2) \sigma_1^2 \\ &+ \frac{\theta_2^2 \alpha_1^2 \beta_1^2 E_{g_1^*}^2(\Pi_1) \sigma_1^4}{(\phi_1 + \theta_2 \beta_1^2) \left( E_{g_1^*}(\Pi_1^2) \left( \mu_1^2 + \frac{\lambda}{\lambda-1} \sigma_1^2 \right) - E_{g_1^*}^2(\Pi_1) \mu_1^2 \right)}. \end{aligned} \quad (55)$$

The proof ideas of Theorems 5 and 6 are to show that for an SPE of a two-stage ALQG game, the behavioral strategy of the adversary has the binary support set  $\left\{ \left( \pi_1 = \varepsilon', \delta_1^2 = \frac{\varepsilon'^2 \sigma_1^2}{\lambda-1} \right), \left( \pi_1 = \varepsilon, \delta_1^2 = \frac{\varepsilon^2 \sigma_1^2}{\lambda-1} \right) \right\}$ , to prove Theorem 5 from the contradiction between the binary support set condition and the adversarial constraint  $\pi_1 \neq 0$ , and to prove Theorem 6 from the SPE definition. The complete proofs can be found in the appendix of the full version [1].

*Remark 4:* Different from the value functions (27) and (42), the value function (55) *cannot* be decomposed into separable terms of mean and variance, which makes the extension of Theorem 6 to  $N \geq 3$  difficult.

## V. NUMERICAL RESULTS

We illustrate the impact of strategic interaction on a time-invariant adversarial LQG problem. The parameters used for the evaluation are shown in Table I.

We start with illustrating the convergence of the mappings in Propositions 2 and 3. Fig. 3 shows the mappings  $L^n(0, 0)$  and  $J^n(0, 0)$  as functions of the iteration  $n$  for  $\lambda = 1.5$  and  $\lambda = 2$ . Both mappings increase monotonically starting from  $(0, 0)$  and converge to their least fixed points  $(\tilde{\theta}, \tilde{\hat{\theta}})$  and  $(\hat{\theta}, \hat{\theta})$ , respectively, which confirms these propositions.

We continue with the evaluation of stationary SPEs for the time-invariant system. Fig. 4 shows the expected average reward per stage for a stationary SPE in pure strategies and that for a stationary SPE in behavioral strategies. Observe that the adversarial manipulation capability decreases as the mutual information lower bound  $\lambda$  increases, and hence the expected average rewards increase. The results also confirm Property 3, i.e., the expected average reward per stage for a stationary SPE in behavioral strategies *cannot* be higher than that for a stationary SPE in pure strategies.

Next, we assess the importance of strategic interaction on the agent's performance, as compared to an agent that is unaware of the attack [3], [6]. Fig. 5 shows the expected average reward per stage for a stationary SPE in behavioral strategies, as per Corollary 4, that for a naive agent under an optimal adversarial attack, and that for an alert agent under SPE adversarial behavioral strategy  $g_i^*$  in Corollary 4, for  $-\varepsilon' = \varepsilon > 0$ . The naive agent is unaware of the adversarial manipulation, i.e., it uses the optimal LQR strategy (12). The corresponding optimal stationary adversarial strategy can be obtained through dynamic programming, and is the pure strategy:

$$(\pi_i, \delta_i^2) = g_i^A(b_i) = \left( -\varepsilon, \frac{\varepsilon^2 \sigma_i^2}{\lambda - 1} \right).$$

The alert agent suspects an adversary but does not act strategically despite the presence of an adversary. The alert agent assumes  $\hat{\pi}_i = \hat{\pi} \neq 0$ ,  $\hat{\delta}_i^2 = \frac{\hat{\pi}^2 \sigma_i^2}{\lambda - 1}$ , and uses the corresponding best response strategy

$$(\kappa_i, \rho_i) = f_i^C(b_i) = \left( -\frac{\tilde{\theta} \alpha \beta (\lambda - 1)}{(\phi + \tilde{\theta} \beta^2) \lambda \hat{\pi}}, -\frac{\tilde{\theta} \alpha \beta \mu_i}{(\phi + \tilde{\theta} \beta^2) \lambda} \right).$$

The figure shows that the expected average reward per stage of the naive agent is always lower than that for the SPE. Clearly, as the mutual information lower bound  $\lambda$  increases, the adversarial manipulation capability becomes weaker and the expected average rewards per stage increase. At the same time, we can observe that if the bound  $\varepsilon$  of the manipulation coefficient is higher then the adversarial manipulation capability becomes stronger and therefore the expected average reward per stage of the naive agent decreases. Note that the limit of the expected average reward per stage of the naive agent does not exist when  $\varepsilon$  is larger than a threshold due to the resulting instability of the control system. The poor performance of the naive agent is consistent with recent works on adversarial DRL [3], [6], where the naive DRL agents were found to perform poorly against strategic adversaries. The results for the SPE show, however, that an agent that is aware of the adversary can adjust its strategy to be resilient to adversarial attack. The figure also shows that the expected

TABLE I  
LQG MODEL DEFAULT PARAMETERS

Parameter	$\mu_1$	$\sigma_1^2$	$\alpha$	$\beta$	$\omega^2$	$\theta$	$\phi$
Value	0	1	-0.5	-1.5	1	2	1

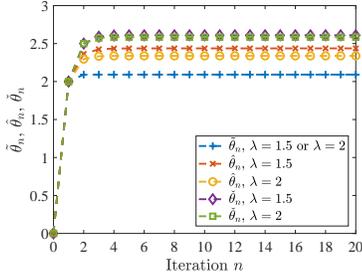


Fig. 3.  $\tilde{\theta}_n$ ,  $\hat{\theta}_n$ ,  $\check{\theta}_n$  computed as  $L^n(0,0)$  and  $J^n(0,0)$  v.s. the number of iterations  $n$ , for  $\lambda = 1.5$  and  $\lambda = 2$  ( $\lambda > \alpha^2$ ).

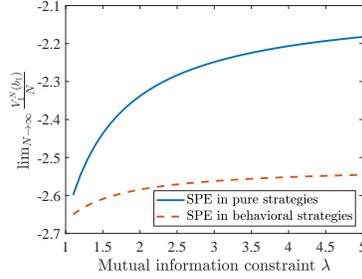


Fig. 4. Expected average reward per stage v.s. mutual information constraint  $\lambda$ , for stationary SPEs in pure strategies and behavioral strategies.

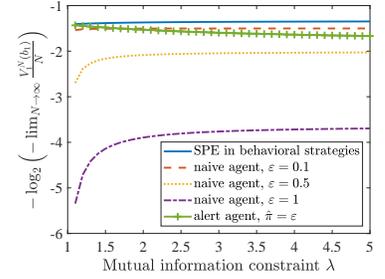


Fig. 5. Expected average reward per stage for stationary SPE in behavioral strategies, that for a naive agent, and that for an alert agent v.s. mutual information constraint  $\lambda$ .

average reward per stage of the alert agent is always lower than that for the SPE since the alert agent does not adjust its best response strategically to the SPE adversarial strategy. Different from the SPE and the naive agent, it is interesting to observe that the alert agent's performance deteriorates as the adversarial constraint  $\lambda$  increases. This is due to that the alert agent's strategy  $f_i^C$  deviates more from the SPE strategy  $f_i^*$ , as shown in Corollary 4, as  $\lambda$  increases.

## VI. CONCLUSION

We proposed a game theoretic model to capture the strategic interaction, information asymmetry and system dynamics for LQG control under adversarial input subject to a mutual information constraint. We characterized the subgame perfect equilibria in pure strategies and in behavioral strategies, including stationary equilibria for time-invariant systems. Our results show that if an equilibrium exists then the agent can use an affine, pure strategy, but randomization enables the adversary to construct more powerful attacks, under a wider range of parameters, and forces the agent into a babbling equilibrium. Our numerical results show the importance of strategic interaction for LQG control, and highlight that an agent that is aware of an adversarial attack can be designed resilient. Our work could be extended in a number of interesting directions, including considering a non-scalar state dynamic system, and relaxing the assumption that the adversarial strategy is revealed to the agent after each stage.

## APPENDIX

### A. Proofs of Theorem 1 and Corollary 1

*Proof:* We prove the result using backward dynamic programming. Recall that there is no feasible adversarial strategy for  $\varepsilon' = \varepsilon = 0$ . Thus, it is sufficient to consider the cases  $\varepsilon' \neq \varepsilon$  or  $\varepsilon' = \varepsilon \neq 0$ . In stage  $N$  the value function for  $b_N$  can be expressed as

$$V_N^N(b_N) = -\tilde{\theta}_N \mu_N^2 - \hat{\theta}_N \sigma_N^2, \quad (56)$$

where  $\tilde{\theta}_N = \hat{\theta}_N = \theta_N$  from the update rules (20)-(22). The pure strategies, which form an SPE and achieve the value function,

consist of any pure strategy  $g_N^*$  satisfying the adversarial constraints (13)-(14), and the dominant pure strategy

$$(\kappa_N^*, \rho_N^*) = f_N^*(b_N) = (0, 0). \quad (57)$$

In stage  $N-1$ , the Q-function of using  $\pi_{N-1} \neq 0$ ,  $\delta_{N-1}^2$ ,  $\kappa_{N-1}$ , and  $\rho_{N-1}$  given a belief  $b_{N-1}$  is

$$\begin{aligned} Q_{N-1}^N(b_{N-1}, \pi_{N-1}, \delta_{N-1}^2, \kappa_{N-1}, \rho_{N-1}) &= -\theta_{N-1} E(S_{N-1}^2) - \phi_{N-1} E(A_{N-1}^2) \\ &\quad - \tilde{\theta}_N E(\Lambda_\mu(b_{N-1}, \pi_{N-1}, \delta_{N-1}^2, \hat{S}_{N-1}, A_{N-1}))^2 \\ &\quad - \hat{\theta}_N \Lambda_V(b_{N-1}, \pi_{N-1}, \delta_{N-1}^2) \\ &= -(\theta_{N-1} + \tilde{\theta}_N \alpha_{N-1}^2) \mu_{N-1}^2 - \hat{\theta}_N \omega_{N-1}^2 \\ &\quad - (\theta_{N-1} + \hat{\theta}_N \alpha_{N-1}^2) \sigma_{N-1}^2 \\ &\quad - (\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2) (\pi_{N-1} \kappa_{N-1} \mu_{N-1} + \rho_{N-1})^2 \\ &\quad - 2\tilde{\theta}_N \alpha_{N-1} \beta_{N-1} \mu_{N-1} (\pi_{N-1} \kappa_{N-1} \mu_{N-1} + \rho_{N-1}) \\ &\quad + (\hat{\theta}_N - \tilde{\theta}_N) \alpha_{N-1}^2 \frac{\pi_{N-1}^2 \sigma_{N-1}^2}{\pi_{N-1}^2 \sigma_{N-1}^2 + \delta_{N-1}^2} \sigma_{N-1}^2 \\ &\quad - (\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2) \kappa_{N-1}^2 (\pi_{N-1}^2 \sigma_{N-1}^2 + \delta_{N-1}^2) \\ &\quad - 2\tilde{\theta}_N \alpha_{N-1} \beta_{N-1} \pi_{N-1} \kappa_{N-1} \sigma_{N-1}^2, \end{aligned} \quad (58)$$

where the expectations are induced by the given  $b_{N-1}$ ,  $\pi_{N-1}$ ,  $\delta_{N-1}^2$ ,  $\kappa_{N-1}$ , and  $\rho_{N-1}$ .

Given  $b_{N-1}$ ,  $\pi_{N-1} \neq 0$ ,  $\delta_{N-1}^2$ , and  $\kappa_{N-1}$ , the Q-function  $Q_{N-1}^N$  is a concave quadratic function of  $\rho_{N-1}$ . As the best response to maximize the agent reward, we can substitute  $\rho_{N-1}$  in terms of  $b_{N-1}$ ,  $\pi_{N-1}$ , and  $\kappa_{N-1}$  as

$$\rho_{N-1} = -\pi_{N-1} \kappa_{N-1} \mu_{N-1} - \frac{\tilde{\theta}_N \alpha_{N-1} \beta_{N-1}}{\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2} \mu_{N-1}. \quad (59)$$

Thus, it is sufficient to consider the Q-function

$$\begin{aligned}
& \underline{Q}_{N-1}^N(b_{N-1}, \pi_{N-1}, \delta_{N-1}^2, \kappa_{N-1}) \\
&= - \left( \theta_{N-1} + \tilde{\theta}_N \alpha_{N-1}^2 - \frac{\tilde{\theta}_N^2 \alpha_{N-1}^2 \beta_{N-1}^2}{\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2} \right) \mu_{N-1}^2 \\
&\quad - (\theta_{N-1} + \hat{\theta}_N \alpha_{N-1}^2) \sigma_{N-1}^2 - \hat{\theta}_N \omega_{N-1}^2 \\
&\quad + (\hat{\theta}_N - \tilde{\theta}_N) \alpha_{N-1}^2 \frac{\pi_{N-1}^2 \sigma_{N-1}^2}{\pi_{N-1}^2 \sigma_{N-1}^2 + \delta_{N-1}^2} \sigma_{N-1}^2 \\
&\quad - (\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2) \kappa_{N-1}^2 (\pi_{N-1}^2 \sigma_{N-1}^2 + \delta_{N-1}^2) \\
&\quad - 2\tilde{\theta}_N \alpha_{N-1} \beta_{N-1} \pi_{N-1} \kappa_{N-1} \sigma_{N-1}^2. \tag{60}
\end{aligned}$$

As shown in Property 1, the belief variance is always positive. Therefore, given  $b_{N-1}$ ,  $\pi_{N-1} \neq 0$ , and  $\delta_{N-1}^2$ , the Q-function  $\underline{Q}_{N-1}^N$  is a concave quadratic function of  $\kappa_{N-1}$ , and the best response of the agent in terms of  $\kappa_{N-1}$  for  $b_{N-1}$ ,  $\pi_{N-1}$ , and  $\delta_{N-1}^2$  can be expressed as

$$\kappa_{N-1} = - \frac{\tilde{\theta}_N \alpha_{N-1} \beta_{N-1} \pi_{N-1} \sigma_{N-1}^2}{(\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2)(\pi_{N-1}^2 \sigma_{N-1}^2 + \delta_{N-1}^2)} \neq 0. \tag{61}$$

Given  $b_{N-1}$ ,  $\pi_{N-1} \neq 0$ , and  $\kappa_{N-1} \neq 0$ , the Q-function  $\underline{Q}_{N-1}^N$  is a decreasing function of  $\delta_{N-1}^2$ . The best response of the adversary in terms of  $\delta_{N-1}^2$  for  $b_{N-1}$  and  $\pi_{N-1}$  is thus

$$\delta_{N-1}^2 = \frac{\pi_{N-1}^2 \sigma_{N-1}^2}{\lambda - 1}. \tag{62}$$

Consequently, it is sufficient to consider the Q-function

$$\begin{aligned}
& \underline{Q}_{N-1}^N(b_{N-1}, \pi_{N-1}, \kappa_{N-1}) \\
&= - \left( \theta_{N-1} + \tilde{\theta}_N \alpha_{N-1}^2 - \frac{\tilde{\theta}_N^2 \alpha_{N-1}^2 \beta_{N-1}^2}{\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2} \right) \mu_{N-1}^2 \\
&\quad - (\theta_{N-1} + \hat{\theta}_N \alpha_{N-1}^2) \sigma_{N-1}^2 - \hat{\theta}_N \omega_{N-1}^2 \\
&\quad + (\hat{\theta}_N - \tilde{\theta}_N) \alpha_{N-1}^2 \frac{\lambda - 1}{\lambda} \sigma_{N-1}^2 \\
&\quad - (\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2) \kappa_{N-1}^2 \frac{\lambda}{\lambda - 1} \pi_{N-1}^2 \sigma_{N-1}^2 \\
&\quad - 2\tilde{\theta}_N \alpha_{N-1} \beta_{N-1} \pi_{N-1} \kappa_{N-1} \sigma_{N-1}^2. \tag{63}
\end{aligned}$$

The pure strategies  $(g_{N-1}^*, f_{N-1}^*)$  form an SPE if  $\pi_{N-1}^* = g_{N-1}^*(b_{N-1})$  and  $\kappa_{N-1}^* = f_{N-1}^*(b_{N-1})$  satisfy

$$\pi_{N-1}^* = \arg \min_{\varepsilon' \leq \pi_{N-1} \leq \varepsilon, \pi_{N-1} \neq 0} \underline{Q}_{N-1}^N(b_{N-1}, \pi_{N-1}, \kappa_{N-1}^*), \tag{64}$$

$$\kappa_{N-1}^* = \arg \max_{\kappa_{N-1} \in \mathbb{R}} \underline{Q}_{N-1}^N(b_{N-1}, \pi_{N-1}^*, \kappa_{N-1}). \tag{65}$$

If  $\varepsilon' = \varepsilon \neq 0$ ,  $\pi_{N-1}^* = \varepsilon = \varepsilon' = g_{N-1}^*(b_{N-1})$  is a dominant adversarial strategy. Therefore, the SPE must exist. The pure strategies  $(g_{N-1}^*, f_{N-1}^*)$  can be obtained by substituting  $\pi_{N-1}^* = \varepsilon$  into (59), (61), and (62) as

$$\begin{aligned}
\delta_{N-1}^{2*} &= g_{N-1}^*(b_{N-1}) = \frac{\varepsilon^2 \sigma_{N-1}^2}{\lambda - 1}; \\
\kappa_{N-1}^* &= f_{N-1}^*(b_{N-1}) = - \frac{\tilde{\theta}_N \alpha_{N-1} \beta_{N-1} (\lambda - 1)}{(\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2) \lambda \varepsilon}; \\
\rho_{N-1}^* &= f_{N-1}^*(b_{N-1}) = - \frac{\tilde{\theta}_N \alpha_{N-1} \beta_{N-1} \mu_{N-1}}{(\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2) \lambda}.
\end{aligned}$$

The value function  $V_{N-1}^N(b_{N-1})$  can be obtained by substituting  $\pi_{N-1}^*$  and  $\kappa_{N-1}^*$  into the Q-function (63) as

$$V_{N-1}^N(b_{N-1}) = -\tilde{\theta}_{N-1} \mu_{N-1}^2 - \hat{\theta}_{N-1} \sigma_{N-1}^2 - \hat{\theta}_N \omega_{N-1}^2.$$

Let us now consider the case  $\varepsilon' \neq \varepsilon$ . Assume that there exists an SPE with  $\pi_{N-1}^* = g_{N-1}^*(b_{N-1}) \neq 0$ . As the best response, solving (65) leads to  $\kappa_{N-1}^* = - \frac{\tilde{\theta}_N \alpha_{N-1} \beta_{N-1} (\lambda - 1)}{(\phi_{N-1} + \tilde{\theta}_N \beta_{N-1}^2) \lambda \pi_{N-1}^*}$ . For all  $\pi_{N-1} \neq \pi_{N-1}^*$  and  $\pi_{N-1} \neq 0$  we have

$$\underline{Q}_{N-1}^N(b_{N-1}, \pi_{N-1}^*, \kappa_{N-1}^*) > \underline{Q}_{N-1}^N(b_{N-1}, \pi_{N-1}, \kappa_{N-1}^*).$$

Thus, condition (64) cannot hold and hence the assumption is not true, i.e., there is no pure strategy SPE in this case.

In the case of  $\varepsilon' = \varepsilon \neq 0$ , Theorem 1 and Corollary 1 can be justified in the remaining stages of the backward dynamic programming by using the same analysis. ■

## REFERENCES

- [1] Available: <https://arxiv.org/abs/2009.02877>
- [2] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-533, 2015.
- [3] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," arXiv:1702.02284.
- [4] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and S. Min, "Tactics of adversarial attack on deep reinforcement learning agents," in *Proc. of IJCAI*, 2017.
- [5] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," in *Proc. of MLDM*, 2017, pp. 262-275.
- [6] A. Russo and A. Proutiere, "Optimal attacks on reinforcement learning policies," arXiv:1907.13548.
- [7] T. Osogami, "Robust partially observable Markov decision process," in *Proc. of ICML*, 2015.
- [8] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *Proc. of ICML*, 2017.
- [9] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," arXiv:1905.10615.
- [10] K. Horak, Q. Zhu, and B. Bosansky, "Manipulating adversary's belief: A dynamic game approach to deception by design for proactive network security," in *Proc. of GameSec*, 2017.
- [11] S. Saritas, S. Yuksel, and S. Gezici, "Nash and Stackelberg equilibria for dynamic cheap talk and signaling games," in *Proc. of ACC*, 2017.
- [12] S. Saritas, E. Shereen, H. Sandberg, and G. Dán, "Adversarial attacks on continuous authentication security: A dynamic game approach," in *Proc. of GameSec*, 2019.
- [13] Z. Li and G. Dán, "Dynamic cheap talk for robust adversarial learning," in *Proc. of GameSec*, 2019.
- [14] M.O. Sayin and T. Basar, "Secure sensor design for cyber-physical systems against advanced persistent threats," in *Proc. of GameSec*, 2017.
- [15] M.O. Sayin, E. Akyol, and T. Basar, "Hierarchical multistage Gaussian signaling games in noncooperative communication and control systems," *Automatica*, vol. 107, pp. 9-20, 2019.
- [16] R. Zhang and P. Venkatasubramanian, "Stealthy control signal attacks in linear quadratic Gaussian control systems: Detectability reward tradeoff," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1555-1570, 2017.
- [17] Q. Zhang, K. Liu, Y. Xia, and A. Ma, "Optimal stealthy deception attack against cyber-physical systems," *IEEE Transactions on Cybernetics*, 2020.
- [18] Y. Chen, S. Kar, and J.M.F. Moura, "Cyber physical attacks constrained by control objectives," in *Proc. of ACC*, 2016, pp. 1185-1190.
- [19] V.P. Crawford and J. Sobel, "Strategic information transmission," *Econometrica*, vol. 50, no. 6, pp. 1431-1451, 1982.
- [20] L. Shapley, "Stochastic games," *Proc. of the National Academy of Sciences*, vol. 39, no. 10, pp. 1095-1100, 1953.
- [21] T. Soderstrom, *Discrete-Time Stochastic Systems*, Springer, 2002.
- [22] A. Baranga, "The contraction principle as a particular case of Kleene's fixed point theorem," *Discrete Mathematics*, vol. 98, pp. 75-79, 1991.