Learning-Based Control Policy and Regret Analysis for Online Quadratic Optimization With Asymmetric Information Structure

Cheng Tan, *Member, IEEE*, Lin Yang, and Wing Shing Wong, *Life Fellow, IEEE*

***Abstract*—In this article, we propose a learning approach to analyze dynamic systems with an asymmetric information structure. Instead of adopting a game-theoretic setting, we investigate an online quadratic optimization problem driven by system noises with unknown statistics. Due to information asymmetry, it is infeasible to use the classic Kalman filter nor optimal control strategies for such systems. It is necessary and beneficial to develop an admissible approach that learns the probability statistics as time goes forward. Motivated by the online convex optimization (OCO) theory, we introduce the notion of regret, which is defined as the cumulative performance loss difference between the optimal offline-known statistics cost and the optimal online-unknown statistics cost. By utilizing dynamic programming and linear minimum mean square biased estimate (LMMSUE), we propose a new type of online state-feedback control policy and characterize the behavior of regret in a finite-time regime. The regret is shown to be sublinear and bounded by $O(\ln T)$. Moreover, we address an online optimization problem with output-feedback control policy and propose a heuristic online control policy.**

***Index Terms*—Asymmetric information, learning-based control policy, linear minimum mean square unbiased estimation (LMMSUE), online quadratic optimization, regret analysis.**

I. Introduction

MANY previously reported works on dynamic systems assume the classic information structure that postulates all agents have equal access to the available system information. Such a symmetric information structure is encountered in a host of application scenarios, such as pursuit-evasion games [1], [2]; networked control systems [3]–[5]; multiagent systems [6], [7]; and seller-buyer supply chain models [8]. In different game settings, it is common to assume that the opposing parties have peering information in regard to location, velocity, player utility functions, and control policies. While such a symmetric information assumption is satisfied in many applications, from a general application perspective, it is of interest to investigate models with an asymmetric information structure. Moreover, early pioneering work in [9] and [10] has pointed out the important role played by the information structure on the decision and control strategy and, thus, offers theoretical motivation to study systems with a nonclassic information structure. There are a number of works analyzing models with asymmetric information in dynamic games [11], [12]; pursuit-evasion problem [13]; and economic theory [14].

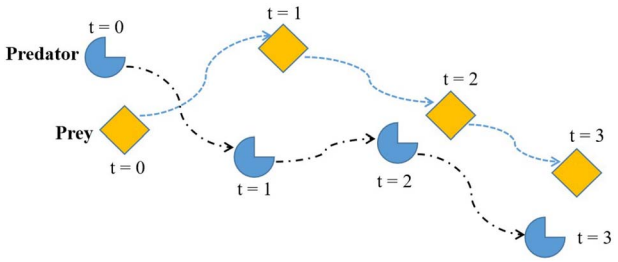


Fig. 1. Movement trajectories of predator and prey.

In this article, we aim to analyze two-player systems in which a single agent with rich input information, the predator, is pitted against the other agent with limited input information, the prey. The motivation of the model comes from application scenarios that include pursuit-evasion and product pricing. Below, we use two simple examples to illustrate the types of online quadratic optimization we focus on in this article.

The first example is motivated by the pursuit-evasion model in [15] and the Mission 7 of the International Aerial Robotics Competition in [16], consisting of a single predator and a single prey (a simple illustration of the dynamic game is depicted in Fig. 1). It is assumed that the predator has access to location information of both players and based on that, selects a predation policy at each decision instant (e.g., whether bait or camouflage is used). The prey has no access to location information of the predator. Hence, it adopts a simple randomized evading policy for each predation policy. To be specific, the dynamic of the predator and the prey is described as

$$x_p(t+1) = x_p(t) + u(t) \tag{1}$$

Manuscript received September 21, 2018; revised October 10, 2019 and September 24, 2020; accepted January 2, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61803224 and Grant U1806204; in part by the Research Grants Council of the Hong Kong Special Administrative Region under Project GRF 14630915; and in part by the Natural Science Foundation of Shandong Province under Grant ZR2019QF005 and Grant ZR201702170323. This article was recommended by Associate Editor D. Liu. *(Corresponding author: Lin Yang.)*

Cheng Tan is with the School of Engineering, Qufu Normal University, Rizhao 276800, China (e-mail: tancheng1987love@163.com).

Lin Yang is with the Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA (e-mail: yanglin.cuhk@gmail.com).

Wing Shing Wong is with the Department of Information Engineering, Chinese University of Hong Kong, Hong Kong (e-mail: wswong@ie.cuhk.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3049357.

Digital Object Identifier 10.1109/TCYB.2021.3049357

$$x_e(t+1) = x_e(t) + v(t) \tag{2}$$

where $x_p(t)$ and $x_e(t)$ are the respective positions of the predator, and prey and $u(t)$ is the predation policy. Assume the evading policies are defined by nonzero-mean random variables, $v(t)$'s, which take value in an admissible bounded set $\{v_1, \ldots, v_M\}$ with $\text{prob}(v(t) = v_i) = p_i,\ \sum_{i=1}^{M} p_i = 1$. The objective of the predator is to minimize both control cost and distance, which is captured by the following quadratic index function:

$$W_T = \sum_{t=0}^{T} \mathbf{E}\Big[\big\|x_e(t) - x_p(t)\big\|^2 + \|u(t)\|^2\Big]. \tag{3}$$

We emphasize that the evading policy distributions are *a priori* unknown to the predator, which leads to an asymmetric information structure.

The second example is related to product pricing [17]. Consider a product pricing that is determined by a single producer, which has absolute control over the pricing and the producing rate. The market demand rate $d(t)$ satisfies

$$d(t+1) = \zeta(t) - \alpha p(t) \geq 0,\ d(0) = d_0 \tag{4}$$

where $\alpha > 0$ and $p(t)$ is the pricing set by the producer. $\zeta(t)$ is the positive utility value that satisfies

$$\zeta(t) = b + e(t) \tag{5}$$

where $e(t)$'s are independent and identically distributed (i.i.d.) random variables with zero mean and unknown variance v_e. On the other hand, the production process is modeled by

$$z(t+1) = z(t) + u(t) \tag{6}$$

where $z(t)$ is the production rate and $u(t)$ is the rate control. For a given optimization horizon of T periods, we define the following objective function:

$$J_T = \sum_{t=0}^{T} \mathbf{E}\Big[c_1(z(t)-d(t))^2 + c_2 u^2(t) - c_3(p(t)-C)d(t+1)\Big]$$
$$C > 0,\ c_i > 0,\ i = 1, 2, 3. \tag{7}$$

The first component in J_T measures how the production process tracks the demands, the second term is a measure of the production rate changes, and the last component represents the total profit assuming the demands are met. The objective of the producer is to minimize J_T via the control variables $(u(t), p(t))$, which are assumed to be measurable with respect to (w.r.t.) the σ-algebra $\mathcal{F}_{z,d} \triangleq \{z(s), d(s), s = 0, 1, \ldots, t\}$.

In the two simple examples above, we formulate the problem as a quadratic optimization instead of a game theoretic setting. Due to its asymmetric nature, the probability statistics of $v(t)$ in (2) and $e(t)$ in (5) are *a priori* unknown to the predator and the producer, respectively. As a result, it is infeasible to use the classic dynamic programming approach [18] nor the maximum principle [19], which defines a challenging task. It is necessary as well as beneficial to develop an admissible approach that learns as time goes forward.

The recent emergence of online convex optimization (OCO) holds promises for solving optimization problems with the asymmetric information structure [21]–[25]. The framework of OCO was first defined in the machine-learning literature, which is closely tied to the statistical learning theory and convex optimization. A popular performance metric for online algorithms is *regret*. In principle, the regret analysis aims to study how far an online algorithm deviates from the optimum [26]. An important property is that the regret of an online algorithm grows at a sublinear rate, which means the time average of the index function converges to the optimal value as T approaches infinity. In [15], we reformulated the first predator–prey model as a multiarmed bandit problem. Although the proposed heuristic algorithm outperforms a random decision policy, its regret is proved to be linear. In the OCO framework, various cutting-edge online algorithms have been proposed to attain the sublinear regret of $O(\sqrt{T})$, such as the online gradient decent method [21], the stochastic gradient decent method [22] and the online Newton step method [23]. In [24], when the cost function is strictly convex, the regret can be improved to $O(\ln T)$. However, for the considered quadratic optimization problem with an asymmetric information structure, how to derive an online strategy to ensure the sublinear regret of $O(\ln T)$ is challenging and remains an open question, which motivates us to undertake an in-depth study.

In this article, we focus on two-player systems in which the players have asymmetric ability to information as motivated by the above examples. Instead of adopting a game-theoretic setting, we formulate the problem as an online quadratic optimization driven by system noises with unknown statistics. Our research methodology contains three powerful techniques, namely, dynamic programming, linear minimum mean square biased estimate (LMMSUE), and regret analysis.

Specifically, for the state-feedback case, if the mean and variance of the system noises are known *a priori*, the optimal offline control policy is derived based on the dynamic programming approach. The optimal state feedback gains, independent of the unknown statistics, are uniquely determined by solving a standard Riccati equation. However, in the current model, since the probability statistics of the system noises are unknown, it is infeasible to apply the optimal offline control strategies. To address this, we introduce an admissible approach that learns the probability statistics of the system noises with the LMMSUE. Based on that, we propose a learning-based optimal control policy. Moreover, under some basic assumptions, the regret of the proposed online control policy grows at a sublinear rate, which is shown to be bounded by $O(\ln T)$. Simulation results show the performance of the developed control policy. On the other hand, we try to address the online quadratic optimization problem with output feedback control. Due to information asymmetry, the classic Kalman filter cannot be applied directly. With the LMMSUE, we propose a heuristic online control policy. The regret between the online known statistics cost and the proposed heuristic offline unknown statistics cost is sublinear, which is shown to be bounded by $O(\ln T)$.

Notation: Let $\mathbb{R}^n$ denote the n-dimensional real Euclidean space and $\mathbb{R}^{m\times n}$ be the space formed by all $m \times n$ real matrices with the usual 2-norm $\|\cdot\|$. The superscript $'$ represents matrix transpose. $\mathbf{Tr}(A)$ represents the trace of a square

matrix A and $\text{diag}\{a_1\ a_2\ \cdots\ a_n\}$ denotes a diagonal matrix. $A \geq 0$ (> 0) represents that A is a positive-semi definite (positive-definite) matrix and $A \geq B$ ($A > B$) means that $A - B \geq 0$ ($A - B > 0$). $\{w(t), t = 0, 1, \ldots\}$ denotes a sequence of real random variables defined on the complete filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t)$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_t = \sigma\{w(s)|s = 0, 1, 2, \ldots, t\}$. Moreover, prob$(A)$ denotes the probability if the event A occurs.

II. State Feedback Control With Learning

A. Problem Formulation

Consider the following discrete time dynamic system:

$$x(t+1) = \text{A}x(t) + \text{B}u(t) + w(t) \tag{8}$$

where $x(t) \in \mathbb{R}^n$ is the state and $u(t) \in \mathbb{R}^m$ is the input control. A and B are the known system parameters with the compatible dimensions and $x(0) = x_0 \in \mathbb{R}^n$ is the given initial state. We assume that $w(t)$'s, are bounded and i.i.d. stochastic process with

$$\text{prob}(w(t) = w_i) = p_i,\ \ i = 1, 2, \ldots, M \tag{9}$$

$$\max_i \|w_i\| \leq w_b < \infty. \tag{10}$$

Define $\mathbf{p}_w = [p_1\ p_2\ \cdots\ p_M]'$, $\mathbf{P}_w = \text{diag}\{p_1\ p_2\ \cdots\ p_M\}$, and $\mathbf{W} = [w_1\ w_2\ \cdots\ w_M]$. It follows that:

$$\mu_w = \mathbf{E}[w(t)] = \sum_{i=1}^{M} p_i w_i = \mathbf{W}\mathbf{p}_w \tag{11}$$

$$Q_w = \mathbf{E}\big[w(t)w(t)'\big] = \sum_{i=1}^{M} p_i w_i w_i' = \mathbf{W}\mathbf{P}_w\mathbf{W}'. \tag{12}$$

Generally, $w(t)$ is nonzero mean, that is, $\mu \neq 0$. Moreover, the covariance of $w(t)$ is

$$C_w = \mathbf{E}\big[(w(t) - \mu_w)(w(t) - \mu_w)'\big] = Q_w - \mu_w\mu_w'. \tag{13}$$

Therefore, the probability statistic of $w(t)$ depends on $\mathbf{p}_w$. We emphasize that $\mathbf{p}_w$ is *a priori* unknown to the decision maker, which leads to the asymmetric information structure.

Without loss of generality, the index function is defined as the general quadratic form

$$J_T(u(t)) = \sum_{t=0}^{T} \mathbf{E}\big[x'(t)Q(t)x(t) + u'(t)R(t)u(t)\big] + \mathbf{E}\big[x'(T+1)P_{T+1}x(T+1)\big] \tag{14}$$

where $Q(t) \geq 0$, $R(t) > 0$, and $P_{T+1} \geq 0$. The goal of the decision maker is to minimize the index function (14) by an online algorithm.

Assume that the probability $\mathbf{p}_w$ is known *a priori*. The finite horizon quadratic optimization problem (14) subject to (8) is fairly standard, which can be solved by utilizing the classic dynamic programming approach (see Theorem 1 hereinafter). Unfortunately, in the current model, $\mathbf{p}_w$ is unknown and the optimal known statistics control strategies cannot be applied directly for asymmetric information case. How to address this unknown statistics problem?

Motivated by the OCO theory, we introduce the regret function as follows:

$$\text{Reg}_T(u(t)) = J_T(u(t)) - J_T^*. \tag{15}$$

The regret measures the cumulative performance loss between the optimal offline case with known statistics cost J_T^* and the online case with unknown statistics cost $J_T(u(t))$. Generally, we say an online policy performs well if its regret is sublinear, that is, $o(T)$, which implies the instantaneous online performance can converge asymptotically to that of the offline performance. Our goal in this article is to develop an admissible approach to estimate the probability $\mathbf{p}_w$ based on the observed state trajectory and then propose a learning-based control policy to reach a better sublinear regret, $O(\ln T)$.

Remark 1: In the first predator–prey model, if we set $x(t) = x_p(t) - x_e(t)$, the first example can be equivalently reformulated as

$$\text{minimize}\ \ J_T = \sum_{t=0}^{T} \mathbf{E}\big[x'(t)x(t) + u'(t)u(t)\big]$$
$$\text{subject to}\ \ x(t+1) = x(t) + u(t) + w(t)$$

where $w(t) = -v(t)$ takes value in an admissible bounded set $\{-v_1, \ldots, -v_M\}$ with $\text{prob}(w(t) = -v_i) = p_i$. In the second product pricing example, if we set

$$v(t) = -p(t) + \frac{1}{2}\left(C + \frac{b}{\alpha}\right),\ \ y(t) = \frac{d(t)}{\alpha} \tag{16}$$

the objective function J_T in (7) can be reformulated as

$$J_T = \bar{J}_T + \sum_{t=0}^{T} \mathbf{E}\left[bC - \frac{1}{4}\left(C + \frac{\zeta(t)}{\alpha}\right)^2\right] \tag{17}$$

where

$$\bar{J}_T = \sum_{t=0}^{T} \mathbf{E}\Big[c_1(z(t) - \alpha y(t))^2 + c_2u^2(t) + c_3\alpha v^2(t)\Big] \tag{18}$$

$$y(t+1) = v(t) + w(t) \tag{19}$$

$$z(t+1) = z(t) + u(t) \tag{20}$$

$$w(t) = \frac{e(t)}{\alpha} - \frac{1}{2}\left(C - \frac{b}{\alpha}\right). \tag{21}$$

In this case, the control variables $(u(t), v(t))$ are measurable w.r.t. the σ-algebra $\mathcal{F}_{z,y}$ generated by $\{z(s),\ y(s),\ s = 0, 1, \ldots, t\}$. Since the difference between the two objective functions J_T and $\bar{J}_T$ is independent of the control policy, the original problem can be reduced to minimize $\bar{J}_T$. If we further set $X(t) = [y(t)\ z(t)]'$, $U(t) = [v(t)\ u(t)]'$ and $W(t) = [w(t)\ 0]'$, the optimization problem can be rewritten as

$$\text{minimize}\ \ \bar{J}_T = \sum_{t=0}^{T} \mathbf{E}\big[X'(t)\text{Q}\text{X}(t) + U'(t)\text{R}\text{U}(t)\big]$$
$$\text{subject to}\ \ Z(t+1) = \text{A}\text{X}(t) + \text{B}\text{U}(t) + W(t)$$

where

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},\ \ B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$Q = c_1\begin{bmatrix} \alpha^2 & -\alpha \\ -\alpha & 1 \end{bmatrix} \geq 0,\ \ R = \begin{bmatrix} c_3 & 0 \\ 0 & c_2 \end{bmatrix} > 0.$$

Therefore, both examples are the special cases of the online quadratic optimization problem with asymmetric information structure.

Remark 2: In this article, we extend the classic quadratic optimization model to involve asymmetric information nature, with that, a traditional controller fails to guarantee the performance when using the classic dynamic programming approach nor the maximum principle. The main challenges of the novel model include the so-called exploitation and exploration dilemma. The recent emergence of machine-learning techniques holds promises for solving such issues and the performance of an online learning algorithm can be measured by comparing its gains with those obtained like knowing the inputs in hindsight. Thus, we adopt *regret* as the performance metric for our algorithm, which has been commonly used in the online learning context. Specifically, a sublinear regret of a learning algorithm implies that the time average of the index function is guaranteed to converge to the optimum as time goes on.

B. Preparatory Results

To begin with, we derive the optimal control strategy $u^*(t)$ and the known statistics optimum J_T^* based on perfect information of the probability $\mathbf{p}_w$.

Theorem 1: Suppose $\mathbf{p}_w$ is known *a priori*. The optimal offline control policy of the optimization problem (14) is

$$u^*(t) = -\Upsilon_T(t)^{-1}\big(B'P_T(t+1)Ax(t) + B'P_T(t+1)\mu_w + B'L_T(t+1)\mu_w\big) \tag{22}$$

while the optimal offline index value of (14) is

$$J_T^* = x_0'P_T(0)x_0 + 2x_0'L_T(0)\mu_w + \sum_{t=0}^{T} H_T(t) \tag{23}$$

where $\Upsilon_T(t)$, $P_T(t)$, $L_T(t)$, and $H_T(t)$ satisfy the following iterative equations:

$$\Upsilon_T(t) = R(t) + B'P_T(t+1)B \tag{24}$$

$$P_T(t) = A'P_T(t+1)A + Q(t) - A'P_T(t+1)B\Upsilon_T(t)^{-1} \times B'P_T(t+1)A \tag{25}$$

$$L_T(t) = \Big(A' - A'P_T(t+1)B\Upsilon_T(t)^{-1}B'\Big) \times (P_T(t+1) + L_T(t+1)) \tag{26}$$

$$H_T(t) = -\mu_w'(P_T(t+1) + L_T(t+1))'B\Upsilon_T(t)^{-1}B' \times (P_T(t+1) + L_T(t+1))\mu_w + 2\mu_w'L_T(t+1)\mu_w + \mathbf{Tr}(P_T(t+1)Q_w) \tag{27}$$

with the terminal condition $P_T(T+1) = P_{T+1}$ and $L_T(T+1) = 0$.

Proof: See Appendix A. ∎

Since the probability $\mathbf{p}_w$ in the optimal offline control strategy is unknown *a priori*, the exact values of μ_w and Q_w are unavailable. Moreover, the optimum J_T^* in Theorem 1 is unavailable and can only be viewed as the optimal known statistics (offline) cost.

Note that $\Upsilon_T(t)$, $P_T(t)$, and $L_T(t)$ are independent of $\mathbf{p}_w$ and thus can be computed offline at the initial time. Therefore, the information of $\Upsilon_T(t+1)$, $P_T(t+1)$, and $L_T(t+1)$ is available to the decision maker at time t.

If we set $K_P(t) = -\Upsilon_T(t)^{-1}B'P_T(t+1)A$, the iterative Riccati (25) is reduced to

$$P_T(t) = (A + BK_P(t))'P_T(t+1)(A + BK_P(t)) + Q(t) + K_P(t)'R(t)K_P(t). \tag{28}$$

Since $Q(t) \geq 0$ and $R(t) > 0$, it follows from (28) that for any terminal condition $P_T(T+1) = P_{T+1} \geq 0$, $P_T(t) \geq 0$ is unique and bounded. Denote

$$\Omega_T(t) = A' - A'P_T(t+1)B\Upsilon_T(t)^{-1}B'. \tag{29}$$

With the terminal condition $L_T(T+1) = 0$, the adjoint (26) can be rewritten as

$$L_T(t) = \sum_{i=t+1}^{T+1}\left(\prod_{j=t}^{i-1}\Omega_T(j)\right)P_T(i) \tag{30}$$

which indicates that the adjoint parameter $L_T(t)$ is uniquely determined by $P_T(s)$, $s = t+1, \ldots, T+1$ and is thus bounded.

Next, we evaluate the cost value in (14) associated with any available control policy.

Proposition 1: For any admissible control policy $u(t)$, the cost of the index function (14) is

$$J_T(u(t)) = J_T^* + \sum_{t=0}^{T}\mathbf{E}\big((u(t) - u^*(t))'\Upsilon_T(t)(u(t) - u^*(t))\big) \tag{31}$$

where $u^*(t)$ and J_T^* are given in (22) and (23).

Proof: See Appendix B. ∎

By Theorem 1 and Proposition 1, we have

$$J_T^* - J_T(u(t)) = \sum_{t=0}^{T}\mathbf{E}\Big[\big(u(t) - u^*(t)\big)'\Upsilon_T(t)\big(u(t) - u^*(t)\big)\Big].$$

For any admissible control policy $u(t)$, it follows from (15) that the regret in this model can be rewritten as:

$$\mathrm{Reg}_T(u(t)) = \sum_{t=0}^{T}\mathbf{E}\Big[\big(u(t) - u^*(t)\big)'\Upsilon_T(t)\big(u(t) - u^*(t)\big)\Big]. \tag{32}$$

For each time $t = 1, 2, \ldots, T$, and any admissible control policy $u(t)$, we define the one-step regret

$$\mathrm{reg}_T(t, u(t)) = \mathbf{E}\Big[\big(u(t) - u^*(t)\big)'\Upsilon_T(t)\big(u(t) - u^*(t)\big)\Big]. \tag{33}$$

It follows that $\mathrm{Reg}_T(u(t)) = \sum_{t=0}^{T}\mathrm{reg}_T(t, u(t))$. The original optimization problem (14) can be reduced to a minimization of (32) with some admissible online control policy.

C. Learning-Based Control Policy and Regret Analysis

First, we focus on a simple but powerful learning tool of estimating an unknown parameter in the statistical learning theory, that is, LMMSUE.

Denote $\widehat{\mathbf{p}}_w(t) = [\hat{p}_1(t)\ \hat{p}_2(t)\ \cdots\ \hat{p}_M(t)]'$ to be the linear unbiased estimate of the probability $\mathbf{p}_w$. With the initial estimate $\widehat{\mathbf{p}}_w(0) = [0\ 0\ \cdots\ 0]'$, it follows from [15] that $\widehat{\mathbf{p}}_w(t)$ satisfies:

$$\widehat{\mathbf{p}}_w(t) = \sum_{i=0}^{t-1} c_i(t)\xi(i),\ \ t = 1, 2, \ldots, T \tag{34}$$

where $\sum_{i=0}^{t-1} c_i(t) = 1$ and $\xi(i)$ is an i.i.d. stochastic process with $\text{prob}(\xi(i) = \xi_j) = p_j,\ j = 1, 2, \ldots, M$, and

$$\xi_1 = [1\ 0\ \cdots\ 0]', \ldots, \xi_M = [0\ 0\ \cdots\ 1]'.$$

Actually, $\xi(t)$ defines the random observation that $w(t)$ takes the value of w_i with the probability $p_i,\ i = 1, 2, \ldots, M$. In this case, we obtain that

$$\mathbf{E}[\xi(i)] = \mathbf{p}_w,\ \ \mathbf{E}\big[\xi(i)\xi(i)'\big] = \mathbf{P}_w. \tag{35}$$

With the linear unbiased estimate, we define the following admissible control policy set by:

$$\mathcal{U}_{\text{ad}} \triangleq \Big\{u(t) = -\Upsilon_T(t)^{-1}B'P_T(t+1)\text{Ax}(t) + l_w(t)\Big\} \tag{36}$$

where

$$l_w(t) = -\Upsilon_T(t)^{-1}B'\big(P_T(t+1) + B'L_T(t+1)\big)\mathbf{W}\widehat{\mathbf{p}}_w(t).$$

To begin with, we propose the LMMSUE $\widehat{\mathbf{p}}_{\min}(t)$ to minimize $\mathbf{E}\|\widehat{\mathbf{p}}_w(t) - \mathbf{p}_w\|^2$.

Lemma 1 [15]: The linear minimum mean square unbiased estimate of $\mathbf{p}_w$ is

$$\widehat{\mathbf{p}}_{\min}(t) = \frac{1}{t}\sum_{i=0}^{t-1}\xi(i). \tag{37}$$

Note that the LMMSUE is the sample mean of the random observation $\xi(t)$. It follows from (37) that:

$$\widehat{\mathbf{p}}_{\min}(t+1) = \frac{1}{t+1}(t\widehat{\mathbf{p}}_{\min}(t) + \xi(t)). \tag{38}$$

Utilizing the Kolmogorov strong law of large numbers [20], we obtain

$$\lim_{t\to\infty}\widehat{\mathbf{p}}_{\min}(t) = \mathbf{p}_w,\ \ a.s. \tag{39}$$

where "a.s." refers to "almost surely."

Remark 3: In principle, at each time $t = 1, 2, \ldots, T$, since $x(t)$, $x(t-1)$, and $u(t-1)$ are known to the decision maker, it is feasible to reach $w(t-1)$ with

$$w(t-1) = x(t) - \text{Ax}(t-1) - \text{Bu}(t-1). \tag{40}$$

Observe that $w(t-1) = w_{h(t-1)}$, $h(t-1) \in \mathbb{M} \triangleq \{1, 2, \ldots, M\}$. We update the LMMSUE $\widehat{\mathbf{p}}_{\min}(t) = (\hat{p}_1(t)\ \hat{p}_2(t)\ \cdots\ \hat{p}_M(t))'$ with

$$\hat{p}_i(t) = \begin{cases} \frac{(t-1)\hat{p}_i(t-1)+1}{t}, & i = h(t-1) \\ \frac{(t-1)\hat{p}_i(t-1)}{t}, & i \neq h(t-1). \end{cases} \tag{41}$$

Define $\hat{\mu}_w(t) = \mathbf{W}\widehat{\mathbf{p}}_{\min}$. Then, we have

$$\mathbf{E}[\hat{\mu}_w(t)] = \mu_w,\ \ \mathbf{E}\big[(\hat{\mu}_w(t) - \mu_w)(\hat{\mu}_w(t) - \mu_w)'\big] = \frac{1}{t}C_w.$$

In this case, $\hat{\mu}_w(t)$ is the LMMSUE of μ_w. Next, based on the LMMSUE, we derive a learning-based $\mathcal{U}_{\text{ad}}$ admissible control policy that is optimal for the unknown statistics case.

Theorem 2: For the online quadratic optimization problem with asymmetric information structure, the optimal online control policy in $\mathcal{U}_{\text{ad}}$ is designed as

$$\begin{aligned}\hat{u}(t) = -\Upsilon_T(t)^{-1}\big(&B'P_T(t+1)\text{Ax}(t) + B'P_T(t+1)\hat{\mu}_w(t) \\ &+ B'L_T(t+1)\hat{\mu}_w(t)\big)\end{aligned} \tag{42}$$

where $\widehat{\mathbf{p}}_{\min}(t)$ is the LMMSUE (37) and $\hat{\mu}_w(t) = \mathbf{W}\widehat{\mathbf{p}}_{\min}$. Moreover, the optimal online index value in (14) is

$$J_T\big(\hat{u}(t)\big) = J_T^* + \mathbf{Tr}\big(D_T(0)\mu_w\mu_w'\big) + \sum_{t=1}^{T}\frac{1}{t}\mathbf{Tr}(D_T(t)C_w) \tag{43}$$

where $\Upsilon_T(t)$, $P_T(t)$, $L_T(t)$, and $H_T(t)$ satisfy the iterative (24)–(27) and

$$\begin{aligned}\mathcal{D}_T(t) = &(P_T(t+1) + L_T(t+1))'B\Upsilon_T(t)^{-1}B' \\ &\times (P_T(t+1) + L_T(t+1)) \geq 0.\end{aligned} \tag{44}$$

Proof: See Appendix C. ∎

To better understand the performance of the proposed online policy, we need to carry out a detailed regret analysis. For convenience, we state the following hypotheses:

H1: $Q(t) = Q \geq 0$, $R(t) = R > 0$ and $P_{T+1} = 0$;

H2: (A, B) is stabilizable and $(A, Q^{[1/2]})$ is observable.

Lemma 2: Suppose $P_T(t)$ is the unique positive-semidefinite solution to the Riccati equation (25). Under hypotheses H1 and H2, $P_T(t)$ is bounded and monotonically nondecreasing as time decreases. Moreover, when $t \to -\infty$, $P_T(t)$ converges to the unique solution $\hat{P} > 0$ to the following algebraic Riccati equation (ARE):

$$\hat{P} = A'\hat{P}A + Q - A'\hat{P}B\Big(R + B'\hat{P}B\Big)^{-1}B'\hat{P}A. \tag{45}$$

Proof: See Appendix D. ∎

Theorem 3: Under hypotheses H1 and H2, the regret $\text{Reg}_T(\hat{u}(t))$ satisfies

$$\text{Reg}_T\big(\hat{u}(t)\big) \leq O(\ln(T)). \tag{46}$$

Proof: By Lemma 2, $P_T(t)$ is uniformly bounded by $0 \leq P_T(t) \leq \hat{P}$, where $\hat{P}$ is the unique positive-definite solution satisfying the ARE (45). By (30), $L_T(t)$ is uniquely determined by $P_T(s)$, $s = t+1, \ldots, T$, and thus bounded. Moreover, by (44), $\mathcal{D}_T(t) \geq 0$ is determined by $P_T(s)$, $s = t+1, \ldots, T$ and also bounded. For $\mathcal{D}_T(t) \geq 0$ and $C_w \geq 0$, there exists a constant $\hat{c} > 0$ such that

$$\mathbf{Tr}(\mathcal{D}_T(t)C_w) \leq \hat{c}. \tag{47}$$

By Theorem 2, the regret satisfies

$$\begin{aligned}\text{Reg}_T(u(t)) &\leq \mathbf{Tr}\big(D_T(0)\mu\mu'\big) + \sum_{t=1}^{T}\frac{1}{t}\hat{c} \\ &= \mathbf{Tr}\big(D_T(0)\mu\mu'\big) + \ln(T + r_T)\hat{c}\end{aligned} \tag{48}$$

where $\lim_{T\to\infty} r_T = r$ and $r > 0$ is the Euler constant. It follows that $\text{Reg}_T(u(t)) \leq O(\ln T)$. ∎

Assume that $T > 0$ is sufficiently large. Next, we analyze the efficiency of the proposed online control policy $\hat{u}(t)$ compared with the other type of admissible control policies.

Case 1: Consider the following admissible control policy based on the linear biased estimation defined as follows:

$$u_1(t) = -\Upsilon_T(t)^{-1}\big(B'P_T(t+1)\mathrm{Ax}(t)+B' \times (P_T(t+1)+L_T(t+1))\tilde{\mu}_w(t)\big),\ t=1,2,\ldots,T \quad (49)$$

where $\tilde{\mu}_w(t) = \mathbf{W}\tilde{\mathbf{p}}(t)$, $\tilde{\mathbf{p}}(t)$ is a linear biased estimate, that is

$$\tilde{\mathbf{p}}(t) = \sum_{i=0}^{t-1} \tilde{c}_i(t)\xi(i),\ t = 1,2,\ldots,T. \quad (50)$$

In this case, the one-step regret satisfies

$$\mathrm{reg}_T(t,u_1(t)) = \sum_{i=0}^{t-1} \tilde{c}_i^2(t)\mathbf{Tr}(D_T(t)C_w) + \left(\sum_{i=0}^{t-1}\tilde{c}_i(t)-1\right)^2 \mathbf{Tr}\big(D_T(t)\mu_w\mu_w'\big). \quad (51)$$

The minimum regret value of (51) achieved at

$$\tilde{c}_i^*(t) = \frac{\mathbf{Tr}\big(D_T(t)\mu_w\mu_w'\big)}{\mathbf{Tr}\big(tD_T(t)\mu_w\mu_w'\big)+\mathbf{Tr}(D_T(t)C_w)}.$$

However, the exact values of μ_w and C_w are unknown to the decision maker, it is infeasible to apply the proposed LMMSUE.

Case 2: Suppose that the decision maker will terminate updating the estimate after a critical time $\bar{t}$, $1 \le \bar{t} < T$. That is to say, for $0 \le t \le \bar{t}$, $u_2(t) = \hat{u}(t)$, and for $\bar{t} < t \le T$

$$u_2(t) = -\Upsilon_T(t)^{-1}\big(B'P_T(t+1)\mathrm{Ax}(t) + B'P_T(t+1)\hat{\mu}_w(\bar{t}) + B'L_T(t+1)\hat{\mu}_w(\bar{t})\big). \quad (52)$$

From the proof of Theorem 2, the regret satisfies

$$\mathrm{Reg}_T(u_2(t)) = \mathbf{Tr}\big(D_T(0)\mu_w\mu_w'\big) + \sum_{t=1}^{\bar{t}}\frac{1}{t}\mathbf{Tr}(D_T(t)C_w) + \sum_{t=\bar{t}+1}^{T}\frac{1}{\bar{t}}\mathbf{Tr}(D_T(t)C_w)$$

which implies that $\mathrm{Reg}_T(\hat{u}(t)) \le \mathrm{Reg}_T(u_2(t))$. The online control policy $\hat{u}(t)$ offers a better performance than $u_2(t)$.

Case 3: Consider the decision maker only utilizes the state-feedback control policy $u(t) = Kx(t)$. In this case, the optimal feedback control policy is derived as

$$u_3(t) = -\Upsilon_T(t)^{-1}B'P_T(t+1)\mathrm{Ax}(t). \quad (53)$$

It follows that:

$$u_3(t) - u^*(t) = \Upsilon_T(t)^{-1}B'(P_T(t+1)+L_T(t+1))\mu_w$$

and

$$\mathrm{reg}_T(t,u_3(t)) = \mathbf{Tr}\big(D_T(t)\mu_w\mu_w'\big). \quad (54)$$

If $T > 0$ is sufficiently large, there exists a critical time $1 \le t_c < T$ such that

$$\mathrm{reg}_T\big(t,\hat{u}(t)\big) \le \mathrm{reg}_T(t,u_3(t)),\ t_c \le t \le T. \quad (55)$$

Moreover, under hypotheses H1 and H2, the regret of $u_3(t)$ is shown to be linear, which indicates that our policy $\hat{u}(t)$ in Theorem 2 offers a better performance than $u_3(t)$.

Remark 4: In the machine-learning area, various learning algorithms, which can be either model based or model free with different application scenarios, have been proposed to solve the quadratic optimization problem with unknown statistics. Typical policies dealing with the asymmetric information dilemma in the online quadratic optimization problem include adaptive dynamic programming (ADP), optimism-in-face-of-uncertainty (OFU), and Thompson sampling (TS). Specifically, [27] gave an OFU-based algorithm that suffers a $O(\sqrt{T})$ regret for continuous-time dynamics with quadratic cost. This result is better than the $O(T^{2/3})$ regret developed in [28] by utilizing TS. Then, an improved TS algorithm was proposed in [29], whose regret is shown to be cumulatively bounded as $O(\sqrt{T})$. Moreover, ADP has been applied to handle the model-free case with unknown linear or nonlinear dynamics [30]–[32]. It was proved in [31] that the convergence of the developed algorithm is equivalent to the classic Newton step method [23]. However, these results were obtained for zero-mean random variables, which appear to be decreasingly effective in the first predator–prey model with $\mathbf{E}[v(t)] \neq 0$. Compared with these results, our innovative contributions are two-fold: 1) different from [27]–[29], our model involves a more general nonzero noise and 2) without knowing the statistics of the noise, we developed a novel strategy, called LMMSUE, which guarantees a much better regret of $O(\ln T)$ than existing policies as TS etc.

Besides, we remark that the alterative (41) of the LMMSUE is simple and thus the developed method has low computational complexity.

Remark 5: Note that our proposed strategy relies on the exact information of system matrices and a linear dependency on the dimension of the system matrices or noise vector is attainable. As a result, the proposed method maintains good performance with massive unknown parameters in a complex industrial system. Despite of the above fact, we highlight that it is a promising research direction to develop more effective learning policies tailored for complex systems and attain a better dimension dependency.

III. Output Feedback Control With Learning

A. Problem Formulation

Consider the following discrete time dynamic system:

$$x(t+1) = \mathrm{Ax}(t) + \mathrm{Bu}(t) + w(t) \quad (56)$$

$$y(t) = \mathrm{Cx}(t) + v(t) \quad (57)$$

where $y(t) \in \mathbb{R}^n$ is the measurement and $C \in \mathbb{R}^{n\times n}$ is nonsingular with the compatible dimension. The initial state $x_0 \in \mathbb{R}^n$ is a Gaussian random vector with

$$\mu_0 = \mathbf{E}[x_0],\ C_0 = \mathbf{E}\big[(x_0-\mu_0)(x_0-\mu_0)'\big]. \quad (58)$$

The measurement noise $v(t)$ is assumed to be a bounded and i.i.d. stochastic process [33] with

$$\max_t \|v(t)\| \le v_b < \infty \quad (59)$$

$$0 = \mathbf{E}[v(t)],\ Q_v = \mathbf{E}\big[v(t)v(t)'\big]. \quad (60)$$

We assume that $w(t)$'s are bounded and form an i.i.d. stochastic process satisfying (9) and (10). The random variables x_0, $w(t)$, and $v(t)$ are assumed to be mutually independent. Moreover, we emphasize that the probability $\mathbf{p}_w$ is *a priori* unknown to the decision maker. The objective is to minimize the index function (14) with asymmetric information structure.

Generally speaking, to solve the quadratic optimization problem (14) subject to (56) and (57), one could apply the well known Kalman filter to estimate the value of the state $x(t)$ and based on that design the optimal offline control policy to minimize the index function. To be specific, denote $Y(t)$ to be the observation set $\{y(0), y(0), \ldots, y(t)\}$. Define $\hat{x}_{t|t-1} = \mathbf{E}[x(t)|Y(t-1)]$, $\hat{x}_{t|t} = \mathbf{E}[x(t)|Y(t)]$, and

$$\Lambda_{t|t-1} = \mathbf{E}\Big[\big(x(t)-\hat{x}_{t|t-1}\big)\big(x(t)-\hat{x}_{t|t-1}\big)'|Y(t-1)\Big] \quad (61)$$

$$\Lambda_{t|t} = \mathbf{E}\Big[\big(x(t)-\hat{x}_{t|t}\big)\big(x(t)-\hat{x}_{t|t}\big)'|Y(t)\Big] \quad (62)$$

where $\mathbf{E}[x|Y]$ defines the conditional expectation of the random variable x w.r.t. Y. Applying the standard Kalman filtering [34] yields that

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + \Lambda_{t|t-1}C'\big(C\Lambda_{t|t-1}C' + Q_v\big)^{-1}\big(y(t) - C\hat{x}_{t|t-1}\big) \quad (63)$$

$$\hat{x}_{t+1|t} = A\hat{x}_{t|t} + \mathrm{B}u(t) + \mu_w \quad (64)$$

where

$$\Lambda_{t|t} = \Lambda_{t|t-1} - \Lambda_{t|t-1}C'\big(C\Lambda_{t|t-1}C' + Q_v\big)^{-1}C\Lambda_{t|t-1}$$
$$\Lambda_{t+1|t} = A\Lambda_{t|t}A' + C_w. \quad (65)$$

The initial conditions are

$$\hat{x}_{0|-1} = \mathbf{E}[x_0] = \mu_0,\ \ \Lambda_{0|-1} = \mathbf{E}\big[(x_0-\mu_0)(x_0-\mu_0)'\big]. \quad (66)$$

By utilizing the classic separation principle, the optimal offline control policy is

$$u^*(t) = -\Upsilon_T(t)^{-1}\big(B'P_T(t+1)A\hat{x}_{t|t} + B'P_T(t+1)\mu_w + B'L_T(t+1)\mu_w\big). \quad (67)$$

However, in the current model, since the exact values of μ_w and C_w are unknown, the classic Kalman filter and the optimal offline control strategy cannot be applied for the asymmetric information case. Instead, we introduce an one-step state estimation based on the observation $y(t)$ at each time $t = 1, 2, \ldots, T$ and then the original problem is reduced to a quadratic optimization problem with a nonwhite system noise [35]. This modified optimization problem is challenging. In this study, we propose a suboptimal offline control policy conditioned on the assumption that the one-step state estimation is applied and the probability statistics of the system are known. Based on the LMMSUE, we propose a learning-based online control policy. The quasiregret between the online known statistics cost and the heuristic offline unknown statistics suboptimal cost is shown to be to be sublinear and bounded by $O(\ln T)$.

B. Learning-Based Control Policy and Regret Analysis

With the output dynamic (57), we introduce a simple one-step state estimate

$$\hat{x}(t) = \mathbf{E}\big[x(t)|y(t)\big] = C^{-1}y(t),\ \ t = 0, 1, \ldots, T \quad (68)$$

which implies that

$$\hat{x}(t+1) = A\hat{x}(t) + \mathrm{B}u(t) + s(t) \quad (69)$$

$$s(t) = w(t) - AC^{-1}v(t) + C^{-1}v(t+1). \quad (70)$$

In this case, $s(t)$ is a colored noise with $\mu_s = \mathbf{E}[s(t)] = \mu_w$

$$Q_s = \mathbf{E}\big[s(t)s'(t)\big] = Q_w + C^{-1}Q_vC^{-1} + AC^{-1}Q_vC^{-1}A'.$$

Moreover, the error covariance is

$$\Lambda(t) = \mathbf{E}\Big[\big(x(t)-\hat{x}(t)\big)\big(x(t)-\hat{x}(t)\big)'\Big] = C^{-1}Q_vC^{-1}.$$

The index function (14) can be rewritten as

$$J_T(u(t)) = \sum_{t=0}^{T}\mathbf{E}\big[\hat{x}'(t)Q(t)\hat{x}(t) + u'(t)R(t)u(t)\big] + \mathbf{E}\big[\hat{x}'(T+1)P_{T+1}\hat{x}(T+1)\big] - D_T \quad (71)$$

where

$$D_T = \sum_{t=0}^{T}\mathbf{Tr}\big(Q(t)\bar{Q}_v\big) + \mathbf{Tr}\big(P_{T+1}\bar{Q}_v\big) \quad (72)$$

$$\bar{Q}_v = C^{-1}Q_vC^{-1}. \quad (73)$$

The original quadratic optimization problem (14) is reduced to minimizing (71) w.r.t. (69), (70). By utilizing the one-step state estimate, we derive a heuristic suboptimal offline result.

Theorem 4: Assume that the probability $\mathbf{p}_w$ is known *a priori*. A suboptimal offline control policy of the quadratic optimization problem (14) is given by

$$u_a(t) = -\Upsilon_T(t)^{-1}\big(B'P_T(t+1)A\hat{x}(t) + B'P_T(t+1)\mu_w + B'L_T(t+1)\mu_w\big) \quad (74)$$

while the index value of (14) is

$$J_T(u_a(t)) = \hat{x}'(0)P_T(t)\hat{x}(0) + 2\hat{x}'(0)L_T(0)\mu_w + H_T \quad (75)$$

where $\Upsilon_T(t)$, $P_T(t)$, and $L_T(t)$ satisfy (24)–(26) and

$$H_T = \sum_{t=0}^{T}\Big\{-\mu_w'(P_T(t+1)+L_T(t+1))'B\Upsilon_T(t)^{-1}B' \times(P_T(t+1)+L_T(t+1))\mu_w + 2\mu_w'L_T(t+1)\mu_w + \mathbf{Tr}\Big(A'P_T(t+1)B\Upsilon_T(t)^{-1}B'P_T(t+1)A\bar{Q}_v\Big) + \mathbf{Tr}(P_T(t+1)Q_w)\Big\} - \mathbf{Tr}\big(P_T(0)\bar{Q}_v\big). \quad (76)$$

Proof: See Appendix E. ∎

Next, we study the LMMSUE of $\mathbf{p}_w$. Due to the presence of the measurement noise $v(t)$, at each time $t = 1, 2, \ldots, T$, it is difficult to reach the exact value of $w(t-1)$. To guarantee the exact observation of $w(t-1)$, we state the following hypothesis

H3: For each $i, j = 1, 2, \ldots, M$ and $i \neq j$

$$\big\|w_i - w_j\big\| > \frac{2(1+\|A\|)v_b}{\|C\|}. \quad (77)$$

At each time $t = 1, 2, \ldots, T$, define $\hat{w}(t-1) = \hat{x}(t) - A\hat{x}(t-1) - \mathrm{Bu}(t-1)$. It follows that:

$$\begin{aligned}\left\|\hat{w}(t-1) - w(t-1)\right\| &= \left\|C^{-1}v(t) - AC^{-1}v(t-1)\right\| \\ &\le \frac{(1+\|A\|)v_b}{\|C\|}. \end{aligned} \tag{78}$$

Suppose that $w(t-1) = w_{h(t-1)}$. For $i \neq h(t-1)$, we obtain

$$\begin{aligned}\frac{2(1+\|A\|)v_b}{\|C\|} &< \|w_i - w_h\| \\ &\le \left\|w_i - \hat{w}(t-1)\right\| + \left\|w_h - \hat{w}(t-1)\right\| \\ &\le \left\|w_i - \hat{w}(t-1)\right\| + \frac{(1+\|A\|)v_b}{\|C\|}\end{aligned}$$

which implies that

$$\left\|\hat{w}(t-1) - w_i\right\| > \frac{(1+\|A\|)v_b}{\|C\|} \ge \left\|\hat{w}(t-1) - w_h\right\|. \tag{79}$$

Therefore, at each time $t = 0, 1, \ldots, T-1$, we have $w(t) = w_{h(t)}$, where $h(t)$ is determined by

$$h(t) = \arg\min_{i=1,\ldots,M} \left\|\hat{w}(t) - w_i\right\|. \tag{80}$$

By Remark 3, we have $\xi(t) = \xi_{h(t)}$ and update $\widehat{\mathbf{p}}_{\min}(t+1) = [\hat{p}_1(t+1)\ \hat{p}_2(t+1)\ \cdots\ \hat{p}_M(t+1)]'$ with

$$\hat{p}_i(t+1) = \begin{cases} \frac{t\hat{p}_i(t)+1}{t+1}, & i = h(t) \\ \frac{t\hat{p}_i(t)}{t+1}, & i \neq h(t). \end{cases} \tag{81}$$

Based on the LMMSUE, we are in a position to present a learning-based online control policy as follows.

Theorem 5: Suppose the probability $\mathbf{p}_w$ is unknown. Under hypothesis H3, an admissible online control policy is

$$\begin{aligned}\hat{u}_a(t) = -\Upsilon_T(t)^{-1}\big(&B'P_T(t+1)A\hat{x}(t) + B'P_T(t+1)\hat{\mu}_w(t) \\ &+B'L_T(t+1)\hat{\mu}_w(t)\big)\end{aligned} \tag{82}$$

while the index value in (14) is

$$\begin{aligned}J_T\big(\hat{u}_a(t)\big) &= \hat{x}'(0)P_T(t)\hat{x}(0) + 2\hat{x}'(0)L_T(0)\mu_w + H_T \\ &+ \mathbf{Tr}\big(D_T(0)\mu_w\mu_w'\big) + \sum_{t=1}^{T}\frac{1}{t}\mathbf{Tr}(D_T(t)C_w)\end{aligned} \tag{83}$$

where $\Upsilon_T(t)$, $P_T(t)$, and $L_T(t)$ satisfy (24) and (25), $\mathcal{D}_T(t)$ satisfies (44), and H_T satisfies (76). Moreover, under hypotheses H1)-H2), the quasiregret between the online cost $J_T(\hat{u}_a(t))$ and the offline cost $J_T(u_a(t))$ satisfies

$$\bar{R}eg_T\big(\hat{u}_a(t)\big) \le O(\ln T). \tag{84}$$

Proof: See Appendix F. ■

IV. Illustrative Examples

In this section, we present two numerical examples to illustrate the effectiveness of our theoretical results.

Example 1: Consider the predator–prey model in (1)–(3). For convenience, we simply set $\beta = 1$ and $\mathbb{R}^n = \mathbb{R}^2$. Assume the initial positions are $x_p = [1\ 0]'$ and $x_e = [0\ 0]'$. The prey has the following four evading policies:

$$v_1 = [1\ 0]',\ v_2 = [-1\ 0]',\ v_3 = [0\ 1]',\ v_4 = [0\ -1]'$$

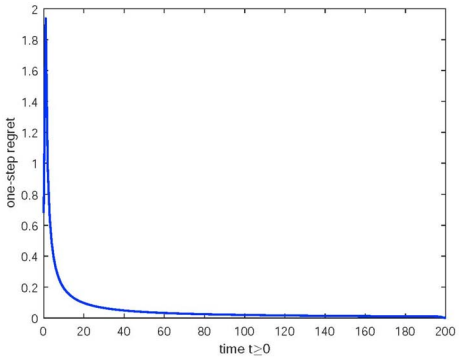


Fig. 2. Trajectories of one-step regret.

with the evading probability distribution

$$\mathbf{p}_v = [0.2\ 0.1\ 0.6\ 0.1]'.$$

In this case, we obtain

$$\mu_v = \begin{bmatrix}0.1\\0.5\end{bmatrix},\ Q_v = \begin{bmatrix}0.3 & 0\\0 & 0.7\end{bmatrix}.$$

If we set $x(t) = x_p(t) - x_e(t)$ and $T = 200$, the first predator–prey problem in (1)–(3) can be reformulated as the state-feedback case (14) with

$$A_1 = B_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix},\ Q_1 = R_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix},\ x(0) = \begin{bmatrix}1\\0\end{bmatrix}.$$

It follows that (A_1, B_1) is stabilizable and $(A_1, Q_1^{(1/2)}$ is observable. Suppose the evading probability distribution $\mathbf{p}_v$ is known to the predator. By Theorem 1, we obtain the optimal offline control policy $u^*(t)$ in (22) which minimizes the index function (14) with $J_T^* = 292.1660$.

By utilizing the proposed admissible control policy $\hat{u}(t)$ in (42) with the LMMSUE $\widehat{\mathbf{p}}_{\min}(t)$, we obtain the index cost with $J_T(\hat{u}(t)) = 304.2107$. Thus, the regret is

$$\mathrm{Reg}_T\big(\hat{u}(t)\big) = J_T\big(\hat{u}(t)\big) - J_T^* = 12.0446.$$

We propose the trajectories of the one-step regret $\mathrm{reg}_T(\hat{u}(t))$ as shown in Fig. 2, where $\mathrm{reg}_T(\hat{u}(T)) = 0$ due to the terminal conditions $P_{T+1} = L_{T+1} = 0$.

Define the regret percentage to be

$$c_T\big(\hat{u}(t)\big) = \frac{\mathrm{Reg}_T\big(\hat{u}(t)\big)}{T} \times 100\%. \tag{85}$$

For different terminal time $T > 0$, the optimal offline index value J_T^*, the optimal online index value $J_T(\hat{u}(t))$, the regret $\mathrm{Reg}_T(\hat{u}(t))$, and the percentage $c_T(\hat{u}(t))$ can be summarized in Table I. It can be concluded that the regret of the proposed online control policy grows at a sublinear rate.

Moreover, to illustrate the effectiveness of our theoretical results, we introduce the Gittins Index-based algorithm to solve the predator–prey model numerically (see the detail in [15, Fig. 2]). The starting point of this benchmark algorithm is to minimize an one-step utility function for each time $t \in [1, T]$ as a surrogate cost function, which provides an upper bound of the index function. As shown in Table II, the developed learning-based control policy in Theorem 2 outperforms the Gittins index-based algorithm.

TABLE I
REGRET ANALYSIS IN EXAMPLE 1

T	J_T^*	$J_T(\hat{u})$	$Reg_T(\hat{u})$	$c_T(\hat{u})$
20	29.8439	37.2439	7.4000	37.0000%
50	73.5643	82.8646	9.3004	18.6008%
100	146.4315	157.1141	10.6825	10.6825%
200	292.1660	304.2107	12.0446	6.0223%
500	729.3696	743.2008	13.8312	2.7662%
1000	1458.0422	1473.2200	15.1778	1.5178%
2000	2915.3873	2931.9100	16.5227	0.8261%

TABLE II
REGRET OF LEARNING-BASED POLICY $\hat{u}$ AND GITTINS INDEX-BASED POLICY $\tilde{u}$

T	J_T^*	$Reg_T(\hat{u})$	$Reg_T(\tilde{u})$
20	29.8439	7.4000	7.6215
50	73.5643	9.3004	17.0375
100	146.4315	10.6825	31.0157
200	292.1660	12.0446	57.2150
500	729.3696	13.8312	141.5870

TABLE III
REGRET ANALYSIS IN EXAMPLE II

T	$\bar{J}_T^*$	$\bar{J}_T(\hat{u})$	$Reg_T(\hat{u})$	$c_T(\hat{u})$
20	1.3786	2.3309	0.9523	4.7620%
50	2.5936	3.5845	0.9909	1.9818%
100	4.6186	5.6380	1.0194	1.0194%
200	8.6686	9.7163	1.0477	0.5239%
500	20.8186	21.9036	1.0850	0.2170%
1000	41.0686	42.1817	1.1131	0.1113%
2000	81.5686	82.7098	1.1412	0.0571%

TABLE IV
REGRET ANALYSIS IN EXAMPLE II

T	$J_T(u_a(t))$	$J_T(\hat{u}_a(t))$	$\bar{Reg}_T(\hat{u}_a(t))$
10	9.3972	9.7077	0.3105
20	19.7015	20.0455	0.3440
50	50.6143	51.0004	0.3861
100	102.1356	102.5530	0.4174
200	205.1782	205.6267	0.4485
500	514.3061	514.7954	0.4894

Example 2: Consider the modified product pricing model in (19) and (20). We assume that $\alpha = (1/4)$, $b = 2$, and $e(t)$'s are bounded and i.i.d. stochastic process with

$$e_1 = 0,\ e_2 = 0.1,\ e_3 = -0.1,\ e_4 = 0.2,\ e_5 = -0.2$$
$$e_6 = 0.3,\ e_7 = -0.3,\ e_8 = 0.4,\ e_9 = -0.4.$$

The probability distribution $\mathbf{p}_e$ is assume to be

$$\mathbf{p}_e = [0.25\ 0.15\ 0.15\ 0.1\ 0.1\ 0.075\ 0.075\ 0.05\ 0.05]'.$$

Moreover, we assume that $c_1 = c_2 = c_3 = 1$.

If we set $X(t) = [y(t)\ z(t)]'$, $U(t) = [v(t)\ u(t)]'$, and $W(t) = [w(t)\ 0]'$, the second product pricing problem can be reformulated as the state-feedback case (14) with

$$A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},\ B_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$Q_2 = \begin{bmatrix} \frac{1}{16} & -\frac{1}{4} \\ -\frac{1}{4} & 1 \end{bmatrix} \geq 0,\ R_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} > 0.$$

It follows that (A_2, B_2) is stabilizable and $(A_2, Q_2^{[1/2]})$ is observable. Moreover, since $w(t) = [e(t)/\alpha] - (1/2)(C - [b/\alpha])$, the probability statistics of $W(t) = [w(t)\ 0]'$ is

$$\mu_W = \begin{bmatrix} 3.6 \\ 0 \end{bmatrix},\ Q_W = \begin{bmatrix} 13.6080 & 0 \\ 0 & 0 \end{bmatrix}.$$

Assume $X(0) = [1\ 1]'$. For different terminal time $T > 0$, it follows from Theorems 1–3 that the optimal index value $\bar{J}_T^*$, the index value $\bar{J}_T(\hat{u}(t))$, the regret $\mathrm{Reg}_T(\hat{u}(t))$, and the percentage $c_T(\hat{u}(t))$ can be summarized in Table III, where the regret of the proposed online control policy grows at a sublinear rate.

Example 3: Consider the output-feedback control model in (56) and (57) where $A_3 = 1.2,\ B_3 = 2.75,\ C_3 = 1,\ Q_3 = 0.5$, and $R_3 = 0.78$. Assume that $w(t)$'s are bounded and i.i.d. stochastic process with

$$w_1 = 0,\ w_2 = 0.3,\ w_3 = 0.5,\ w_4 = 0.7,\ w_5 = 0.9$$
$$\mathbf{p}_w = [0.1\ 0.2\ 0.4\ 0.2\ 0.1]'.$$

It follows that:

$$\mu_w = \mathbf{E}[w(t)] = \sum_{i=1}^{5} p_i w_i = 0.49$$
$$Q_w = \mathbf{E}\big[w(t)w(t)'\big] = \sum_{i=1}^{5} p_i w_i w_i' = 0.297.$$

Moreover, the measurement noise $v(t)$ is assumed to satisfy

$$0 = \mathbf{E}[v(t)],\ Q_v = \mathbf{E}\big[v(t)v(t)'\big] = 1.25.$$

For different terminal time $T > 0$ and a given initial state $x_0 = 0.35$, it follows from Theorems 4 and 5 that the quasiregret $\bar{Reg}_T(\hat{u}_a(t))$ between the online cost $J_T(\hat{u}_a(t))$ and the offline cost $J_T(u_a(t))$ is summarized in Table IV.

V. CONCLUSION

In this article, we focused on an online quadratic optimization problem with an asymmetric information structure. We assumed that a single predator with rich information input is pitted against a single prey with limited input information. Motivated by the OCO methodology, we developed an admissible approach that enables the predictor-agent learn the probability statistics of the system with the LMMSUE. Based on that, we proposed a learning-based optimal online control policy. Its regret grows at a sublinear rate, and is shown to be bounded by $O(\ln T)$, which implies the online performance can converge asymptotically to that of the offline optimal performance.

As future work, there are two promising research directions. The first research direction is to figure out more optimal online control strategies and analysis framework for the existing online quadratic optimization problems. The other direction is to extend the two-player models to more complicated models, such as multiagent systems. With unknown statistics of multiplicative noise or network topology, it is infeasible to utilize the classic distributed control strategies. The online optimization approach can offer a promising but challenging new direction.

APPENDIX A
PROOF OF THEOREM 1

Proof: The proof is based on the dynamic programming approach. For each time $t = 0, 1, \ldots, T$, define the following cost-to-go function:

$$\mathcal{G}(t) = \min_{u(t)} G(t) \tag{86}$$

where

$$G(t) = \mathbf{E}\big[x'(t)Q(t)x(t) + u'(t)R(t)u(t) + \mathcal{G}(t+1)\big] \tag{87}$$

and

$$\mathcal{G}(T+1) = \mathbf{E}\big[x'(T+1)P_{T+1}x(T+1)\big]. \tag{88}$$

Next, we show that

$$\mathcal{G}(t) = \mathbf{E}\big[x'(t)P_T(t)x(t) + 2x'(t)L_T(t)\mu_w\big] + \sum_{j=t}^{T} H_T(j) \tag{89}$$

where $P_T(t)$, $L_T(t)$, and $H_T(t)$ satisfy (25)–(27).

For $t = T$, it follows from (87) that:

$$\begin{aligned}
G(T) = \mathbf{E}\big[&x'(T)Q(T)x(T)+u'(T)R(T)u(T)\\
&+(\mathrm{Ax}(T)+\mathrm{Bu}(T)+w(T))'P_{T+1}\\
&\times(\mathrm{Ax}(T)+\mathrm{Bu}(T)+w(T))\big]\\
= \mathbf{E}\Big[&\Big(u(T)+\Upsilon_T(T)^{-1}\big(B'P_{T+1}\mathrm{Ax}(T)+B'P_{T+1}\mu_w\big)\Big)'\\
&\times\Upsilon_T(T)\Big(u(T)+\Upsilon_T(T)^{-1}\\
&\quad\times\big(B'P_{T+1}\mathrm{Ax}(T)+B'P_{T+1}\mu_w\big)\Big)\\
&-\big(B'P_{T+1}\mathrm{Ax}(T)+B'P_{T+1}\mu_w\big)'\Upsilon_T(T)^{-1}\\
&\times\big(B'P_{T+1}\mathrm{Ax}(T)+B'P_{T+1}\mu_w\big)+x'(T)\\
&\times\big(Q(T)+A'P_{T+1}A\big)x(T)+2x'(T)A'P_{T+1}\mu_w\Big]\\
&+\mathbf{Tr}(P_{T+1}Q_w)
\end{aligned}$$

where $\Upsilon_T(T) = R(T) + B'P_{T+1}B$. At time T, the optimal control policy $u^*(T)$ is

$$u^*(T) = -\Upsilon_T(T)^{-1}\big(B'P_{T+1}\mathrm{Ax}(T) + B'P_{T+1}\mu_w\big) \tag{90}$$

while the cost-to-go function $\mathcal{G}(T)$ satisfies

$$\begin{aligned}
\mathcal{G}(T) &= \min_{u(T)} G(T)\\
&= \mathbf{E}\Big[x'(T)\Big(Q(T)+A'P_{T+1}A-A'P_{T+1}B\Upsilon_T(T)^{-1}B'P_{T+1}A\Big)\\
&\quad\times x(T)+2x'(T)\Big(A'-A'P_{T+1}B\Upsilon_T(T)^{-1}B'\Big)\\
&\quad\times P_{T+1}\mu_w\Big]-\mu_w'P_{T+1}B\Upsilon_T(T)^{-1}B'P_{T+1}\mu_w\\
&\quad+\mathbf{Tr}(P_{T+1}Q_w)\\
&= \mathbf{E}\big[x'(T)P_T(T)x(T)+2x'(T)L_T(T)\mu_w\big]+H_T(T)
\end{aligned}$$

where $P_T(T)$, $L_T(T)$, and $H_T(T)$ satisfy (25)–(27) with $t = T$.

For each $t = 0, 1, \ldots, T-1$, suppose

$$\begin{aligned}
\mathcal{G}(t+1) = \mathbf{E}\big[&x'(t+1)P_T(t+1)x(t+1)\\
&+2x'(t+1)L_T(t+1)\mu_w\big]+\sum_{j=t+1}^{T}H_T(j).
\end{aligned}$$

It follows that:

$$\begin{aligned}
G(t) = \mathbf{E}\big[&x'(t)Q(t)x(t)+u'(t)R(t)u(t)+x'(t+1)P_T(t+1)\\
&\times x(t+1)+2x'(t+1)L_T(t+1)\mu_w\big]+\sum_{j=t+1}^{T}H_T(j)\\
= \mathbf{E}\Big[&\Big(u(t)+\Upsilon_T(t)^{-1}\big(B'P_T(t+1)\mathrm{Ax}(T)\\
&+B'(P_T(t+1)+L_T(t+1))\mu_w\big)\Big)'\\
&\times\Upsilon_T(t)\Big(u(t)+\Upsilon_T(t)^{-1}\\
&\quad\times\big(B'P_T(t+1)\mathrm{Ax}(t)\\
&\quad+B'(P_T(t+1)+L_T(t+1))\mu_w\big)\Big)\\
&-\big(B'P_T(t+1)\mathrm{Ax}(t)+B'(P_T(t+1)+L_T(t+1))\mu_w\big)'\\
&\times\Upsilon_T(t)^{-1}\big(B'P_T(t+1)\mathrm{Ax}(t)\\
&\quad+B'(P_T(t+1)+L_T(t+1))\mu_w\big)\\
&+x'(t)\big(Q(t)+A'P_T(t+1)A\big)x(t)\\
&+2x'(t)A'P_T(t+1)\mu_w+2x'(t)A'L_T(t+1)\mu_w\big]\\
&+2\mu_w'L_T(t+1)\mu_w\\
&+\mathbf{Tr}(P_T(t+1)Q_w)+\sum_{j=t+1}^{T}H_T(j).
\end{aligned}$$

Then, the optimal control policy is $u^*(t)$ in (22), which implies that

$$\begin{aligned}
\mathcal{G}(t) &= \min_{u(t)} G(t)\\
&= \mathbf{E}\Big[x'(t)\Big(Q(t)+A'P_T(t+1)A-A'P_T(t+1)B\Upsilon_T(t)^{-1}\\
&\quad\times B'P_T(t+1)A\Big)x(t)\\
&\quad+2x'(t)\Big(A'-A'P_T(t+1)B\Upsilon_T(t)^{-1}B'\Big)\\
&\quad\times(P_T(t+1)+L_T(t+1))\mu_w]\\
&\quad-\mu_w'(P_T(t+1)+L_T(t+1))'B\Upsilon_T(t)^{-1}\\
&\quad\times B'(P_T(t+1)+L_T(t+1))\mu_w+2\mu'L_T(t+1)\mu_w\\
&\quad+\mathbf{Tr}(P_T(t+1)Q_w)+\sum_{j=t+1}^{T}H_T(j)\\
&= \mathbf{E}\big[x'(t)P_T(t)x(t)+2x'(t)L_T(t)\mu_w\big]+\sum_{j=t}^{T}H_T(j).
\end{aligned}$$

Utilize the dynamic programming with the cost-to-go function (86) yields the optimal index value satisfies (23), which completes this proof. ■

APPENDIX B
PROOF OF PROPOSITION 1

Proof: For each time $t = 0, 1, \ldots, T$, define the following Lyapunov function:

$$V(t) = \mathbf{E}\big[x'(t)P_T(t)x(t) + 2x'(t)L_T(t)\mu_w\big] + \sum_{i=t}^{T} H_T(i).$$

It follows that:

$$\begin{aligned}
&V(t) - V(t+1) \\
&\quad = \mathbf{E}\big[x'(t)P_T(t)x(t) + 2x'(t)L_T(t)\mu_w\big] + \sum_{i=t}^{T} H_T(i) \\
&\qquad - \mathbf{E}\big[(\mathrm{A}x(t) + \mathrm{B}u(t) + w(t))'P_T(t+1) \\
&\qquad\qquad \times (\mathrm{A}x(t) + \mathrm{B}u(t) + w(t)) \\
&\qquad\qquad + 2(\mathrm{A}x(t) + \mathrm{B}u(t) + w(t))'L_T(t+1)\mu_w\big] \\
&\qquad - \sum_{i=t+1}^{T} H_T(i) \\
&\quad = \mathbf{E}\big[x'(t)Q(t)x(t) + u'(t)R(t)u(t)\big] \\
&\qquad - \mathbf{E}\big[(u(t) - u^*(t))'\Upsilon_T(t)(u(t) - u^*(t))\big]
\end{aligned}$$

which implies that

$$\begin{aligned}
V(0) - V(T+1) &= \sum_{t=0}^{T} V(t) - V(t+1) \\
&= x_0'P_T(0)x_0 + 2x_0'L_T(0)\mu + \sum_{t=0}^{T} H_T(t) \\
&\quad - \mathbf{E}\big[x'(T+1)P_T(T+1)x(T+1)\big] \\
&= \sum_{t=0}^{T} \mathbf{E}\big[x'(t)Q(t)x(t) + u'(t)R(t)u(t)\big] - \sum_{t=0}^{T} \\
&\quad \times \mathbf{E}\Big[\big(u(t) - u^*(t)\big)'\Upsilon_T(t)\big(u(t) - u^*(t)\big)\Big].
\end{aligned}$$

In this case, we obtain

$$\begin{aligned}
J_T(u(t)) &= x_0'P_T(0)x_0 + 2x_0'L_T(0)\mu + \sum_{t=0}^{T} H_T(t) \\
&\quad + \sum_{t=0}^{T} \mathbf{E}\Big[\big(u(t) - u^*(t)\big)'\Upsilon_T(t)\big(u(t) - u^*(t)\big)\Big]
\end{aligned}$$

which completes this proof. ∎

Appendix C
Proof of Theorem 2

Proof: By Proposition 1, we first show that the regret for the developed control policy $\hat{u}(t)$ satisfies

$$\mathrm{Reg}_T(\hat{u}(t)) = \mathbf{Tr}\big(D_T(0)\mu_w\mu_w'\big) + \sum_{t=1}^{T} \frac{1}{t}\mathbf{Tr}(D_T(t)C_w). \tag{91}$$

For $t_0 = 0$, the decision maker has no observation. With the initial estimate $\widehat{\mathbf{p}}_{\min}(0) = [0\ 0\ \cdots\ 0]'$, the control policy is designed to be

$$\hat{u}(0) = -\Upsilon_T(0)^{-1}B'P_T(1)\mathrm{A}x(0) \tag{92}$$

which is the feedback of the initial state $x(0) = x_0$. In this case, we have

$$\hat{u}(0) - u^*(0) = \Upsilon_T(0)^{-1}B'(P_T(1) + L_T(1))\mu_w \tag{93}$$

which implies that

$$\begin{aligned}
\mathrm{reg}_T\big(0, \hat{u}(0)\big) &= \mathbf{E}\Big[\big(\hat{u}(0) - u^*(0)\big)'\Upsilon_T(0)\big(\hat{u}(0) - u^*(0)\big)\Big] \\
&= \mathbf{Tr}\big(D_T(0)\mu_w\mu_w'\big)
\end{aligned} \tag{94}$$

where $\mathcal{D}_T(t) \geq 0$ is given in (44). For each time $t = 1, 2, \ldots, T$, the decision maker observes the exact value of $\xi(i)$, $i = 0, 1, \ldots, t-1$. In this case, we obtain

$$\begin{aligned}
&\hat{u}(t) - u^*(t) \\
&\quad = -\Upsilon_T(t)^{-1}B'(P_T(t+1) + L_T(t+1))\big(\hat{\mu}_w(t) - \mu_w\big)
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathrm{reg}_T\big(t, \hat{u}(t)\big) &= \mathbf{E}\Big[\big(\hat{\mu}_w(t) - \mu_w\big)'\mathcal{D}_T(t)(\hat{\mu}_w(t) - \mu_w)\Big] \\
&= \frac{1}{t}\mathbf{Tr}(D_T(t)C_w).
\end{aligned} \tag{95}$$

With the updated estimate $\widehat{\mathbf{p}}_{\min}(t)$ and the control policy $\hat{u}(t)$, the regret satisfies (91). It follows from Proposition 1 that the index value in (14) is:

$$\begin{aligned}
J_T\big(\hat{u}(t)\big) &= x_0'P_T(0)x_0 + 2x_0'L_T(0)\mu_w + \sum_{t=0}^{T} H_T(t) \\
&\quad + \mathbf{Tr}\big(D_T(0)\mu_w\mu_w'\big) + \sum_{t=1}^{T} \frac{1}{t}\mathbf{Tr}(D_T(t)C_w).
\end{aligned}$$

For each online control policy $u_1(t) \in \mathcal{U}_{\mathrm{ad}}$ satisfying

$$\begin{aligned}
u_1(t) &= -\Upsilon_T(t)^{-1}\big(B'P_T(t+1)\mathrm{A}x(t) + B' \\
&\qquad \times (P_T(t+1) + L_T(t+1))\check{\mu}_w(t)\big) \\
t &= 1, 2, \ldots, T
\end{aligned} \tag{96}$$

where $\check{\mu}_w(t) = \mathbf{W}\check{\mathbf{p}}(t)$ and $\check{\mathbf{p}}(t)$ is a linear unbiased estimate satisfying

$$\check{\mathbf{p}}(t) = \sum_{i=0}^{t-1} \check{c}_i(t)\xi(i), \ \ t = 1, 2, \ldots, T$$

with $\check{c}_i(t) \neq (1/t)$, $i = 0, 1, \ldots, t-1$, and $\sum_{i=0}^{t-1} \check{c}_i(t) = 1$. In this case, the regret of $u_1(t)$ satisfies

$$\begin{aligned}
\mathrm{Reg}_T(u_1(t)) &= \sum_{t=0}^{T} \mathbf{E}\Big[\big(u_1(t) - u^*(t)\big)'\Upsilon_T(t)\big(u_1(t) - u^*(t)\big)\Big] \\
&= \mathbf{Tr}\big(D_T(0)\mu_w\mu_w'\big) + \sum_{t=1}^{T}\sum_{i=0}^{t-1} \check{c}_i^2(t)\mathbf{Tr}(D_T(t)C_w).
\end{aligned}$$

Define

$$f\big(\check{c}\big) = \sum_{i=0}^{t-1} \check{c}_i^2(t). \tag{97}$$

Applying $\check{c}_0(t) = 1 - \sum_{i=1}^{t-1} \check{c}_i(t)$ to (97), we obtain

$$f\big(\check{c}\big) = \sum_{i=1}^{t-1} \check{c}_i^2(t) + \left(1 - \sum_{i=1}^{t-1} \check{c}_i(t)\right)^2. \tag{98}$$

For each $j = 1, 2, \ldots, t-1$, we have

$$f_{\check{c}_j(t)}\big(\check{c}\big) = 2\left(\check{c}_j(t) + \sum_{i=1}^{t-1} \check{c}_i(t) - 1\right). \tag{99}$$

Suppose $f_{\check{c}_j(t)}(\check{c}) = 0$ holds, we have the minimum point is $\check{c}_i(t) = (1/t),\ \ i = 0, 1, \ldots, t-1$. It follows that:

$$\mathrm{Reg}_T\big(\hat{u}(t)\big) \leq \mathrm{Reg}_T(u_1(t))$$

which yields that the online control policy $\hat{u}(t)$ serves as a better performer than $u_1(t)$. ■

APPENDIX D
PROOF OF LEMMA 2

Proof: Consider the following quadratic optimization problem:

$$\begin{aligned} \text{minimize} \;\; & W_T = \sum_{t=0}^{T} z'(t)\mathrm{Q}z(t) + v'(t)Rv(t) \\ \text{subject to} \;\; & z(t+1) = Az(t) + Bv(t). \end{aligned} \tag{100}$$

It follows from Theorem 1 with $w(t) = 0$ that the optimal index value of (100) is:

$$W_T^* = z(0)'P_T(0)z(0). \tag{101}$$

Due to the time invariance of the Riccati equation (25), for any $0 \le t \le T$, we have $P_T(t) = P_{T-t}(0)$. For any $z(0)$ and $0 \le t_1 < t_2 \le T$, it follows that:

$$z(0)'P_T(t_1)z(0) = W_{T-t_1}^* \ge W_{T-t_2}^* = z(0)'P_T(t_2)z(0)$$

which indicates that $P_T(t_1) \ge P_T(t_2)$. Since (A, B) is stabilizable, there exists a stabilizing control policy $v_s(t) = K_s z(t)$ such that $\lim_{t\to\infty} \|z_s(t)\|^2 = 0$ and

$$\sum_{t=0}^{\infty} \|z_s(t)\|^2 \le c_1 \|z(0)\|^2.$$

In this case, we have

$$\begin{aligned} z(0)'P_T(t)z(0) = W_{T-t}^* &\le \sum_{i=0}^{T-t} z'(i)\mathrm{Q}z(i) + v_s'(i)Rv_s(i) \\ &\le \sum_{t=0}^{\infty} z'(t)\mathrm{Q}z(t) + v_s'(t)Rv_s(t) \\ &\le \sum_{t=0}^{\infty} z'(t)\big(Q + K_s'\mathrm{R}\mathrm{K}_s\big)z(t) \\ &\le c\|z(0)\|^2 \end{aligned}$$

where $c = \lambda_{\max}(Q + K_s'\mathrm{R}\mathrm{K}_s)c_1$. Thus, $P_T(t)$ is bounded. Moreover, we have

$$\lim_{t\to-\infty} P_T(t) = \lim_{T\to\infty} P_T(0) = \hat{P}$$

where $\hat{P}$ satisfies the ARE (45). Moreover, since (A, B) is stabilizable and $(A, Q^{\frac{1}{2}})$ is observable, it follows from [36, Th. 1] that the ARE (45) has a unique positive-definite solution $\hat{P} > 0$. The proof is completed. ■

APPENDIX E
PROOF OF THEOREM 4

Proof: Since D_T is independent with the control policy $u(t)$, we only need to consider the following optimization problem:

$$\begin{aligned} \text{minimize} \;\; & J_1(u(t)) = \sum_{t=0}^{T} \mathbf{E}\big[\hat{x}'(t)Q(t)\hat{x}(t) + u'(t)R(t)u(t)\big] \\ & + \mathbf{E}\big[\hat{x}'(T+1)P_{T+1}\hat{x}(T+1)\big] \\ \text{subject to} \;\; & \hat{x}(t+1) = A\hat{x}(t) + \mathrm{B}u(t) + s(t). \end{aligned} \tag{102}$$

For each time $t = 0, 1, \ldots, T$, define $\mathcal{G}(t) = G(t, u_a(t))$ with

$$G(t, u(t)) = \mathbf{E}\big[\hat{x}'(t)Q(t)\hat{x}(t) + u'(t)R(t)u(t) + \mathcal{G}(t+1)\big].$$

The terminal condition is given as

$$\mathcal{G}(T+1) = \mathbf{E}\big[\hat{x}'(T+1)P_{T+1}\hat{x}(T+1)\big]. \tag{103}$$

Next, we show that

$$\mathcal{G}(t) = \mathbf{E}\big[\hat{x}'(t)P_T(t)\hat{x}(t) + 2\hat{x}'(t)L_T(t)\mu_w\big] + \sum_{i=t}^{T} M(i)$$

where

$$\begin{aligned} M(t) = & -\mu_w'(P_T(t+1) + L_T(t+1))'B\Upsilon_T(t)^{-1}B' \\ & \times (P_T(t+1) + L_T(t+1))\mu_w + 2\mu_w'L_T(t+1)\mu_w \\ & + 2\mathbf{Tr}\Big(A'P_T(t+1)B\Upsilon_T(t)^{-1}B'P_T(t+1)A\bar{Q}_v\Big) \\ & + \mathbf{Tr}(P_T(t+1)Q_w) + \mathbf{Tr}\big(P_T(t+1)\bar{Q}_v\big) \\ & - \mathbf{Tr}\big(A'P_T(t+1)A\bar{Q}_v\big). \end{aligned} \tag{104}$$

For each time $t = 0, 1, \ldots, T$, it follows that:

$$\begin{aligned} G(t, u(t)) = & \mathbf{E}\Big[\hat{x}'(t)Q(t)\hat{x}(t) + u'(t)R(t)u(t) \\ & + \big(A\hat{x}(t) + \mathrm{B}u(t) + s(t)\big)'P_T(t+1) \\ & \times \big(A\hat{x}(t) + \mathrm{B}u(t) + s(t)\big) \\ & + 2\big(A\hat{x}(t) + \mathrm{B}u(t) + s(t)\big)'L_T(t+1)\mu_w\Big] \\ & + \sum_{i=t+1}^{T} M(i) \\ = & \mathbf{E}\Big[\Big(u(t) + \Upsilon_T(t)^{-1}h\big(B'P_{t+1}A\hat{x}(t) + B'P_T(t+1)\mu_w \\ & \qquad + B'L_T(t+1)\mu_w\big)\Big)' \\ & \times \Upsilon_T(t)\Big(u(t) + \Upsilon_T(t)^{-1} \\ & \qquad \times \big(B'P_T(t+1)A\hat{x}(t) \\ & \qquad + B'(P_T(t+1) + L_T(t+1))\mu_w\big)\Big) \\ & - \big(B'P_T(t+1)A\hat{x}(t) \\ & \quad + B'(P_T(t+1) + L_T(t+1))\mu_w\big)' \\ & \times \Upsilon_T(t)^{-1}\big(B'P_T(t+1)A\hat{x}(t) + B'P_T(t+1)\mu_w \\ & \qquad + B'L_T(t+1)\mu_w\big) \\ & + \hat{x}'(t)\big(Q + A'P_T(t+1)A\big)\hat{x}(t) \\ & + 2\hat{x}'(T)A'P_T(t+1)\mu_w + 2\hat{x}'(t)A'L_T(t+1)\mu_w \\ & - 2u'(T)B'P_T(t+1)AC^{-1}v(t)\Big] \\ & + 2\mu_w'L_T(t+1)\mu_w \\ & + \mathbf{Tr}(P_T(t+1)Q_w) + \mathbf{Tr}\big(P_T(t+1)\bar{Q}_v\big) \\ & - \mathbf{Tr}\big(A'P_T(t+1)A\bar{Q}_v\big) + \sum_{i=t+1}^{T} M(i). \end{aligned}$$

By utilizing the control policy $u_a(t)$ in (74), we have

$$\mathcal{G}(t) = \mathbf{E}\big[\hat{x}'(t)P_T(t)\hat{x}(t) + 2\hat{x}'(t)L_T(t)\mu_w\big] + \sum_{i=t}^{T} M(i).$$

If we set $M_T = \sum_{t=0}^{T} M(t)$, it follows that:

$$\begin{aligned} J_1(u_a(t)) &= \sum_{t=0}^{T} \mathbf{E}\big[\hat{x}'(t)Q(t)\hat{x}(t) + u_a'(t)R(t)u_a(t)\big] \\ &\quad + \mathbf{E}\big[\hat{x}'(T+1)P_{T+1}\hat{x}(T+1)\big] \\ &= \hat{x}'(0)P_T(t)\hat{x}(0) + 2\hat{x}'(0)L_T(0)\mu_w \\ &\quad + M_T - D_T \end{aligned}$$

where

$$\begin{aligned} M_T - D_T = \sum_{t=0}^{T}\Big\{ & -\mu_w'(P_T(t+1) + L_T(t+1))'B\Upsilon_T(t)^{-1}B' \\ & \times (P_T(t+1) + L_T(t+1))\mu_w \\ & + 2\mu_w' L_T(t+1)\mu_w \\ & + \mathbf{Tr}\Big(A'P_T(t+1)B\Upsilon_T(t)^{-1}B'P_T(t+1)A\bar{Q}_v\Big) \\ & + \mathbf{Tr}(P_T(t+1)Q_w)\Big\} - \mathbf{Tr}\big(P_T(0)\bar{Q}_v\big). \end{aligned}$$

This proof is completed. ∎

APPENDIX F
PROOF OF THEOREM 5

Proof: For the optimization problem (102), define the following Lyapunov function:

$$W(t) = \mathbf{E}\big[\hat{x}'(t)P_T(t)\hat{x}(t) + 2\hat{x}'(t)L_T(t)\mu_w\big] + \sum_{i=t}^{T} M(i)$$

where $M(t)$ satisfies (104). It follows that:

$$\begin{aligned} W(t) - W(t+1) = {} & \mathbf{E}\big[\hat{x}'(t)Q(t)\hat{x}(t) + u'(t)R(t)u(t)\big] \\ & - \mathbf{E}\big[(u(t) - u_a(t))'\Upsilon_T(t)(u(t) - u_a(t))\big] \\ & + 2\mathbf{E}\Big[u'(t)B'P_T(t+1)AC^{-1}v(t)\Big] + 2\mathbf{Tr} \\ & \times \Big(A'P_T(t+1)B\Upsilon_T(t)^{-1}B'P_T(t+1)A\bar{Q}_v\Big) \end{aligned} \tag{105}$$

where $u_a(t)$ is given in (74). Summarizing (105) from $t = 0$ to $t = T$ yields that

$$\begin{aligned} W(0) - W(T+1) &= \sum_{t=0}^{T} W(t) - W(t+1) \\ &= \hat{x}'(0)P_T(0)\hat{x}(0) + 2\hat{x}'(0)L_T(0)\mu_w + \sum_{t=0}^{T} M(t) \\ &\quad - \mathbf{E}\big[\hat{x}'(T+1)P_T(T+1)\hat{x}(T+1)\big] \\ &= \sum_{t=0}^{T} \mathbf{E}\big[\hat{x}'(t)Q(t)\hat{x}(t) + u'(t)R(t)u(t)\big] \\ &\quad - \sum_{t=0}^{T} \mathbf{E}\big[(u(t) - u_a(t))'\Upsilon_T(t)(u(t) - u_a(t))\big] \\ &\quad + \sum_{t=0}^{T} 2\mathbf{E}\Big[u'(t)B'P_T(t+1)AC^{-1}v(t)\Big] \\ &\quad + \sum_{t=0}^{T} 2\mathbf{Tr}\Big(A'P_T(t+1)B\Upsilon_T(t)^{-1} \\ &\qquad \times B'P_T(t+1)A\bar{Q}_v\Big) \end{aligned}$$

which implies that

$$\begin{aligned} J_1(u(t)) = {} & \hat{x}'(0)P_T(0)\hat{x}(0) + 2\hat{x}'(0)L_T(0)\mu_w + \sum_{t=0}^{T} M(t) \\ & + \sum_{t=0}^{T} \mathbf{E}\big[(u(t) - u_a(t))'\Upsilon_T(t)(u(t) - u_a(t))\big] \\ & - \sum_{t=0}^{T} 2\mathbf{Tr}\Big(A'P_T(t+1)B\Upsilon_T(t)^{-1}B'P_T(t+1)A\bar{Q}_v\Big) \\ & - \sum_{t=0}^{T} 2\mathbf{E}\Big[u'(t)B'P_T(t+1)AC^{-1}v(t)\Big]. \end{aligned} \tag{106}$$

By utilizing the admissible control policy $\hat{u}_a(t)$ in (82), we obtain

$$\begin{aligned} J_T\big(\hat{u}_a(t)\big) &= J_1\big(\hat{u}_a(t)\big) - D_T \\ &= \hat{x}'(0)P_T(t)\hat{x}(0) + 2\hat{x}'(0)L_T(0)\mu_w + M_T - D_T \\ &\quad + \sum_{t=0}^{T} \mathbf{E}\Big[\big(\hat{u}_a(t) - u_a(t)\big)'\Upsilon_T(t)\big(\hat{u}_a(t) - u_a(t)\big)\Big] \\ &\quad - \sum_{t=0}^{T} 2\mathbf{E}\Big[\hat{u}_a(t)B'P_T(t+1)AC^{-1}v(t)\Big] \\ &\quad - \sum_{t=0}^{T} 2\mathbf{Tr}\Big(A'P_T(t+1)B\Upsilon_T(t)^{-1}B'P_T(t+1)A\bar{Q}_v\Big) \\ &= \hat{x}'(0)P_T(t)\hat{x}(0) + 2\hat{x}'(0)L_T(0)\mu_w + H_T \\ &\quad + \mathbf{Tr}\big(D_T(0)\mu_w\mu_w'\big) + \sum_{t=1}^{T} \frac{1}{t}\mathbf{Tr}(D_T(t)C_w). \end{aligned}$$

Moreover, by Theorem 4, the regret satisfies

$$\bar{Reg}_T\big(\hat{u}_a(t)\big) = \sum_{t=1}^{T} \frac{1}{t}\mathbf{Tr}(D_T(t)C_w).$$

Under hypotheses H1 and H2, we obtain that $\bar{Reg}_T(\hat{u}_a(t)) \leq O(\ln T)$. ∎

REFERENCES

[1] Y. Ho, A. Bryson, and S. Baron, "Differential games and optimal pursuit-evasion strategies," *IEEE Trans. Autom. Control*, vol. 10, no. 4, pp. 385–389, Oct. 1965.

[2] W. Li, "A dynamics perspective of pursuit-evasion: Capturing and escaping when the pursuer runs faster than the agile evader," *IEEE Trans. Autom. Control*, vol. 62, no. 1, pp. 451–457, Jan. 2017.

[3] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints II: Stabilization with limited information feedback," *IEEE Trans. Autom. Control*, vol. 44, no. 5, pp. 1049–1053, May 1999.

[4] C. Tan and H. Zhang, "Necessary and sufficient stabilizing conditions for networked control systems with simultaneous transmission delay and packet dropout," *IEEE Trans. Autom. Control*, vol. 62, no. 8, pp. 4011–4016, Aug. 2017.

[5] C. Tan, H. Zhang, and W. S. Wong, "Delay-dependent algebraic Riccati equation to stabilization of networked control systems: Continuous-time case," *IEEE Trans. Cyber.*, vol. 48, no. 10, pp. 2783–2794, Oct. 2018.

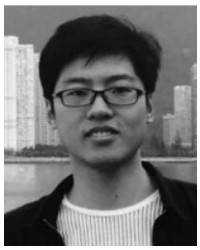
[6] K.-K. Oh, M.-C. Park, and H.-S. Ahn, "A survey of multi-agent formation control," *Automatica*, vol. 53, pp. 424–440, Mar. 2015.

[7] C. Tan, Z. Liu, and W. S. Wong, "Formation control for a string of interconnected second-order systems via target feedback," *J. Franklin Inst.*, vol. 356, no. 15, pp. 8521–8541, 209.

[8] A. H. L. Lau and H.-S. Lau, "Some two-echelon supply-chain games: Improving from deterministic-symmetric-information to stochastic-asymmetric-information models," *Eur. J. Oper. Res.*, vol. 161, no. 1, pp. 203–223, Feb. 2005.
[9] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM J. Control*, vol. 6, no. 1, pp. 131–147, Jan. 1968.
[10] Y.-C. Ho and K.-C. Chu, "Team decision theory and information structures in optimal control problems—Part I," *IEEE Trans. Autom. Control*, vol. AC-17, no. 1, pp. 15–22, Feb. 1972.
[11] P. Cardaliaguet, "Differential games with asymmetric information," *SIAM J. Control Optim.*, vol. 46, no. 3, pp. 816–838, 2007.
[12] A. Gupta, A. Nayyar, C. Langbort, and M. T. Basar, "Common information based Markov perfect equilibria for linear-Gaussian games with asymmetric information," *SIAM J. Control Optim.*, vol. 52, no. 5, pp. 3228–3560, 2014.
[13] K. Sugihara and I. Suzuki, "Optimal algorithms for a pursuit-evasion problem in grids," *SIAM J. Discrete Math.*, vol. 2, no. 1, pp. 126–143, 1989.
[14] M. Esmaeili and P. Zeephongsekul, "Seller-buyer models of supply chain management with an asymmetric information structure," *Int. J. Product. Econ.*, vol. 123, no. 1, pp. 146–154, Jan. 2010.
[15] C. Tan, C. Xu, L, Yang, and W. S. Wong, "Gittins index based control policy for a class of pursuit-evasion problems," *IET Control Theory Appl.*, vol. 12, no. 1, pp. 110–118, Jan. 2018.
[16] J. L. Sanchez-Lopez, J. Pestana, J.-F. Collumeau, R. Suarez-Fernandez, P. Campoy, and M. Molina, "A vision based aerial robot solution for the mission 7 of the international aerial robotics competition," in *Proc Int. Conf. Unmanned Aircraft Syst.*, Denver, CO, USA, 2015, pp. 1391–1400.
[17] C.-L. Chen and W.-C. Lee, "Multi-objective optimization of multi-echelon supply chain networks with uncertain product demands and prices," *Comput. Chem. Eng.*, vol. 28, nos. 6–7, pp. 1131–1144, Jun. 2004.
[18] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA, USA: Athena Sci., 1995.
[19] A. E. Bryson, *Applied Optimal Control: Optimization, Estimation and Control*. New York, NY, USA: Halsted, 1975.
[20] K. L. Judd, "The law of large numbers with a continuum of IID random variables," *J. Econ. Theory*, vol. 35, no. 1, pp. 19–25, Feb. 1985.
[21] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. 20th Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2003, pp. 928–936.
[22] E. Hazan and S. Kale, "Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2489–2512, 2014.
[23] Y. Wang and S. Boyd, "Fast model predictive control using online optimization," *IEEE Trans. Control Syst. Technol.*, vol. 18, no. 2, pp. 267–278, Mar. 2010.
[24] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Mach. Learn.*, vol. 69, no. 2, pp. 169–192, 2007.
[25] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 714–725, Mar. 2018.
[26] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012.
[27] Y. Abbasi-Yadkori and C. Szepesvari, "Regret bounds for the adaptive control of linear quadratic systems," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 1–26.
[28] M. Abeille and A. Lazaric, "Thompson sampling for linear-quadratic control problems," in *Proc. Artif. Intell. Stat.*, 2017, pp. 1246–1254.
[29] M. Abeille and A. Lazaric, "Improved regret bounds for thompson sampling in linear quadratic control problems," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–9.
[30] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.
[31] H. Li, D. Liu, and D. Wang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 3, pp. 706–714, Jul. 2014.
[32] D. Liu, Q. Wei, D. Wang, and H. Li, *Adaptive Dynamic Programming With Applications in Optimal Control*. Berlin, Germany: Springer, 2017.
[33] L. El Ghaoui and G. Calafiore, "Robust filtering for discrete-time systems with bounded noise and parametric uncertainty," *IEEE Trans. Autom. Control*, vol. 46, no. 7, pp. 1084–1089, Jul. 2001.
[34] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.
[35] M. J. Vrhel and H. J. Trussell, "Optimal color filters in the presence of noise," *IEEE Trans. Image Process.*, vol. 4, no. 6, pp. 814–823, Jun. 1995.
[36] Y. Huang, W. Zhang, and H. Zhang, "Infinite horizon linear quadratic optimal control for discrete-time stochastic systems," *Asian J. Control*, vol. 10, no. 5, pp. 608–615, Oct. 2008.

Cheng Tan (Member, IEEE) received the B.S. and M.S. degrees from the School of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, China, in 2010 and 2012, respectively, and the Ph.D. degree from the School of Control Science and Engineering, Shandong University, Jinan, China in 2016.

He was a Postdoctoral Fellow with the Department of Information Engineering, Chinese University of Hong Kong, Hong Kong, from 2016 to 2019. He is currently an Associate Professor with the School of Engineering, Qufu Normal University, Rizhao, China. His research interests include networked control systems, stochastic control, time-delay systems, and optimization control.



Lin Yang received the B.Eng. and M.Sc. degrees from the University of Science and Technology of China, Hefei, China, in 2012 and 2015, respectively, and the Ph.D. degree from the Chinese University of Hong Kong, Hong Kong, in 2018.

He is a Postdoctoral Research Fellow with the Department of Computer Science, University of Massachusetts, Amherst, MA, USA. His research interests include machine learning and online optimization.

Wing Shing Wong (Life Fellow, IEEE) received the combined bachelor's and master's degrees from Yale University, New Haven, CT, USA, in 1976, and the M.S. and Ph.D. degrees from Harvard University, Cambridge, MA, USA, in 1978 and 1980, respectively.

He worked with AT&T Bell Laboratories, Murray Hill, NJ, USA, from 1982 to 1992. In 1992, he joined the Chinese University of Hong Kong, Hong Kong, where he is currently a Choh-Ming Li Research Professor of Information Engineering. He was the Chairman of the Department of Information Engineering from 1995 to 2003 and the Dean of Graduate School from 2005 to 2014. He served as a Science Advisor with the Innovation and Technology Commission, Hong Kong Government, Hong Kong, from 2003 to 2005. He has participated in a variety of research projects on topics ranging from mobile communication and networked control to network control.