Learning in Hide-and-Seek

Qingsi Wang, *Member, IEEE,* and Mingyan Liu, *Fellow, IEEE, Member, ACM*

***Abstract*—Existing work on pursuit-evasion problems typically either assumes stationary or heuristic behavior of one side and examines countermeasures of the other, or assumes both sides to be strategic which leads to a game theoretical framework. Results from the former often lack robustness against changes in the adversarial behavior, while those from the second category, typically as equilibrium solution concepts, may be difficult to justify: either due to the implied knowledge of other players' actions/beliefs and knowledge of their knowledge, or due to a lack of efficient dynamics to achieve such equilibria. In this paper, we take a different approach by assuming an intelligent pursuer/evader that is *adaptive* to the information available to it and is capable of learning over time with performance guarantee. Within this context we investigate two cases. In the first case we assume either the evader or the pursuer is aware of the type of learning algorithm used by the opponent, while in the second case neither side has such information and thus must try to learn. We show that the optimal policies in the first case have a greedy nature. This result is then used to assess the performance of the learning algorithms that both sides employ in the second case, which is shown to be mutually optimal and there is no loss for either side compared to the case when it knows perfectly the adaptive pattern used by the adversary and responses optimally. We further extend our model to study the application of jamming defense.**

***Index Terms*—pursuit and evasion, adversarial learning, two-player game, jamming defense**

I. Introduction

THE pursuit-evasion (or hide-and-seek) problem arises in a variety of applications and has been extensively studied. For instance, it models the pursuit of a moving target by a radar or an unmanned vehicle [2], or a radio performing channel switching in an attempt to hide from a jammer [3].

Existing work in this area typically falls into two categories. The first considers stationary or heuristic behavior of one side and examines corresponding countermeasures of the other. Examples include [4]–[7] and the references therein, which assume a stationary target (the evader) hiding in any of a set of locations with known prior probabilities. Variants of this model include, e.g., [8] that uses a random prior probability of hiding in a given location, and [9] where the detection probability is random with known distribution. Search problems with a moving evader have also been extensively studied. However, the evasion is typically either independent of the pursuer's activity, or heuristically given without clearly defined rationale or performance guarantee, see e.g. [10], where the evader's motion is given by a discrete-time Markov chain independent of the pursuer's activity, and [11] for a similar, continuous-time formulation. The second category assumes both sides to be strategic, leading to a game theoretical framework. A typical method is to use differential games [12] to capture the continuous evolution; in fact, the pursuit-evasion problem bears the genesis of differential games. See also [13]–[15] for texts and examples of differential games and their application in the pursuit-evasion problem. We note that results from the first category often lack robustness against changes in the adversarial behavior, while those from the second category, typically in the form of equilibrium solution concepts, may be difficult to justify: either due to the implied knowledge of other players' actions or beliefs, and knowledge of other players' knowledge, both of which may be limited in practice, or due to a lack of efficient dynamics to achieve such equilibria.

In this paper, we take a different approach by assuming an adaptive pursuer or evader that is simply capable of learning over time, and investigate the resulting decision problem. In other words we assume the pursuer is able to adapt over time using observations of the evader's behavior; it need not possess all the information available to the evader nor does it presume that the evader is rational. The same applies to the evader.

To model the adaptive behavior of the pursuer or the evader, we shall employ online learning algorithms developed for the class of adversarial or non-stochastic multi-armed bandit problems [24], [25], which provide robust and considerable performance guarantee, without assuming any probabilistic model of the underlying reward process. We then investigate two cases. In the first case we assume either the evader or the pursuer is aware of the type of learning algorithm used by the opponent, while in the second case we consider the more realistic scenario when neither side has such information and thus both must try to learn. We show that the optimal policies in the first case have a greedy nature, hiding/seeking in the location least/most likely searched/used by the opponent. We also examine the use of a decoy by the evader to sufficiently mislead the pursuer's learning process. These results are then used to assess the performance of the learning algorithms that both sides employ in the second case, which is shown to be mutually optimal. Furthermore, we show that in this case there is no loss for either side compared to when it knows the adaptive pattern of the adversary and responses optimally.

While the above pursuit-evasion model applies to a variety of scenarios as mentioned, in this study we will primarily focus on the application of jamming defense. Existing literature on jamming tends to heavily focus on specific attack and defense mechanisms. For instance, [16] and [17] introduce a collection of jamming attacks and anti-jamming measures; examples also include using stronger error detection, correction, and spreading codes at the physical layer [18]–[21], exploring the vulnerability in the rate adaptation mechanism of

The work is partially supported by the NSF under grants CIF-0910765 and CNS-1217689, and by the NASA grant NNX09AE91G. This paper is an extended version of the conference paper [1].

Q. Wang is with Qualcomm Research, San Diego, California, USA, and M. Liu is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA (e-mail: {qingsi, mingyan}@umich.edu).

IEEE 802.11 [22], and multi-channel jamming using a single cognitive ratio [23]. The approach presented in this paper explores an alternate dimension and thus complements existing methods. By putting the jamming defense in a pursuit-evasion framework with adaptive players, our approach sheds light on the optimal decision a legitimate user should adopt to counter a strong learning attacker with a known learning rationale, and shows how the learning technique can itself be considered a countermeasure when there is no such prior information. To better model jamming defense, we extend the basic pursuit-evasion model to allow heterogeneous payoffs associated with different pursuit/evasion actions, which captures the diversity in wireless spectrum quality and transmission conditions. We note that the model adopted in this paper is limited to the medium access control of the user while assuming constant packet arrival from upper layers. As a result, it does not capture the impact of jamming on the user's upper-layer control mechanisms such as TCP; this remains an interesting direction of future studies.

Our main contributions are summarized as follows:

- We formulate the pursuit-evasion problem from an on-line learning perspective, and show the optimal evading and pursuing polices with respect to different levels of knowledge of the opponent's behavioral pattern.
- We show the effects of information asymmetry in the pursuit-evasion problem: when the asymmetry is given by different levels of knowledge of the opponent, the side with more information does not necessarily hold advantage in the long run.
- On the other hand, when the information asymmetry stems from one's ability to distinguish between the actual opponent and a *decoy* device, it can be pivotal in deciding the long-term outcome of the interactions. In particular, it leads to an interesting decoy lemma inherent in all no-regret learning algorithms.
- We generalize the pursuit-evasion model to the application of jamming defense, by explicitly modeling the spectrum diversity.
- We also formulate two related families of problems within the same framework, namely the rendezvous and the collision problems, and discuss to what extent our results can be generalized to these two problems as well as open questions.

The remainder of the paper is organized as follows. Section II describes the system model and the problem formulation, followed by detailed analysis in Sections III-V, and application to jamming defense in Section VI. Section VII discusses the rendezvous and the collision problems and Section VIII concludes the paper. All missing proofs of our results can be found in the appendix unless otherwise noted.

II. System Model and Problem Formulation

A. System model

Consider the repeated hide-and-seek interaction between a pursuer and an evader in discrete time. At each time step t, the evader selects one of m locations, indexed by the set $\mathcal{C} = \{1, 2, \ldots, m\}$, to hide in, while the pursuer searches possibly multiple locations simultaneously. The evader's and the pursuer's behaviors are generally described by their respective sets of marginal probabilities $\tau(t) = (\tau_k(t))_{k\in\mathcal{C}}$ and $\alpha(t) = (\alpha_k(t))_{k\in\mathcal{C}}$, where $\tau_k(t)$ and $\alpha_k(t)$ are the respective probabilities that the k-th location is chosen by the evader and the pursuer at time t; we shall also call $\tau(t)$ and $\alpha(t)$ the adversarial behavior with respect to one's opponent at time t. There are two interpretations of $\tau(t)$ and $\alpha(t)$: they can describe randomized strategies of the players, or a probabilistic belief held by one side about the likelihood of an action by the other side. In the following, we shall also introduce a number of other notation for formal presentation. A summary of our main notation can be found in Table I at the end of this section.

The evader's objective is to maximize its total number of successful evasion, while the pursuer aims to maximize its total number of successful pursuits. Within this context we investigate two case. In the first case, we assume either the evader or the pursuer knows the type of learning algorithm or decision process used by its opponent (Section III and IV), while in the second case both sides have no such information (Section V). This leads to different perceptions one side has on the other as we elaborate below.

We define two sets of variables $z_k(t)$ and $x_k(t)$ such that $z_k(t) = 1$ if the pursuer does not search location k at time t, and $z_k(t) = 0$ otherwise, while $x_k(t) = 1$ if the evader hides at location k at time t, and $x_k(t) = 0$ otherwise. When the evader (or the pursuer) knows the type of algorithm/reasoning the pursuer (resp. the evader) uses, it may regard $z_k(t)$ (resp. $x_k(t)$) as stochastic, i.e., assuming its opponent behaves probabilistically according to $\mathbb{P}(z_k(t) = 0) = \alpha_k(t)$ (resp. $\mathbb{P}(x_k(t) = 1) = \tau_k(t)$), though the value of this probability may be unknown to the evader (resp. the pursuer). When the evader (or the pursuer) has no such information, it may regard $z_k(t)$ (or $x_k(t)$) as a predetermined but unknown number.

B. Formulation: against known adaptive search/evasion

In Sections III and IV, we assume either the evader or the pursuer knows the type of adaptive algorithm used by the other, and seeks to make optimal location selections so as to maximally evade/discover the opponent in repeated interaction. For simplicity of presentation, in the following we assume the evader is the party with the knowledge as in Section III; the other case can be formulated similarly. Specifically, the evader assumes the pursuer behaves probabilistically as the latter indeed does, and knows the value of the adversarial behavior $\alpha(t)$ at the beginning of the time step t. $\alpha(t)$ as a vector of probability distribution will be referred to as the state of the system at t and may be random itself. We describe the pursuit pattern in detail in Section III-A. Thus, the evader perceives the pursuer activity $z_k(t)$ as stochastic. Results obtained in this section are then used as benchmarks when we examine the more realistic situation where both sides do not presume to know the other's adaptive behavior.

We assume that the evader has perfect recall of all past states and control actions, though later (c.f. the remarks after Theorem 3) it is shown that this assumption can be significantly weakened. At time t, the evader decides the

control action $\pi(t) \in \mathcal{C}$, i.e., the location to hide in, as a function of the history of system states, past control actions, and a private randomization device that is independent from any activity of the pursuer (to allow randomized strategies):

$$\pi(t) = \gamma_t(\alpha^{[t]}, \pi^{[t-1]}, \omega(t)),$$

where $\alpha^{[t]} := (\alpha(1), \ldots, \alpha(t))$ with $\pi^{[t-1]}$ similarly defined, and $(\omega(t), t = 1, 2, \ldots)$ denotes the private randomization device. The control policy is given by $\gamma = (\gamma_t, t = 1, 2, \ldots)$ and Γ denotes the policy space. Given a location selection sequence $\pi = (\pi(1), \pi(2), \ldots)$ under policy γ, the evader receives an expected reward $r^\pi(t) = 1 - \alpha_{\pi(t)}(t)$ at time t, which is the mean number of successful evasions at the chosen location. Note that in this setup the reward is independent of the location in the sense that a successful evasion in any location amounts to one unit of reward. We also consider the case with location-dependent reward for the application of jamming defense in Section VI. The evader then considers the following two reward maximization problems,

$$\underset{\gamma\in\Gamma}{\text{maximize}} \quad \mathbb{E}\left\{\sum_{t=1}^{T} r^\pi(t)\right\}, \tag{1}$$

and

$$\underset{\gamma\in\Gamma}{\text{maximize}} \quad \liminf_{T\to\infty} \mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} r^\pi(t)\right\}, \tag{2}$$

where the expectation is with respect to (w.r.t.) the randomness of system states and the private randomization device.

For the case when the pursuer holds the knowledge of the evader, we shall denote the pursuer's control rule and control policy by λ_t and λ, respectively, with Λ being the policy space, and $\theta(t)$ its private randomization device. We also denote by $\xi = (\xi(1), \xi(2), \ldots)$ the induced location selection sequence, and by $b^\xi(t) = \tau_{\xi(t)}(t)$ the expected reward of the pursuer at time t. Similar problems can then be formulated in parallel:

$$\max_{\lambda\in\Lambda} \quad \mathbb{E}\left\{\sum_{t=1}^{T} b^\xi(t)\right\}, \text{ and } \max_{\lambda\in\Lambda} \quad \liminf_{T\to\infty} \mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} b^\xi(t)\right\}.$$

C. Formulation: against unknown adversarial behavior

In Section V, we consider the more realistic scenario where neither side has information on the adaptive behavior of the opponent. Both sides hence regard $z_k(t)$ and $x_k(t)$ as predetermined but unknown numbers, respectively. We assume the evader can observe the value of $z_k(t)$ of the selected location after the action at time t, so can the pursuer for the value of $x_k(t)$. We also assume both sides have perfect recall of past observations and control actions, and the resulting control actions are given by

$$\pi(t) = \gamma_t(z_\pi^{[t-1]}, \pi^{[t-1]}, \omega(t)),$$

and

$$\xi(t) = \lambda_t(x_\xi^{[t-1]}, \xi^{[t-1]}, \theta(t)),$$

where $z_\pi^{[t-1]} := (z_{\pi(1)}(1), \ldots, z_{\pi(t-1)}(t-1))$ with $x_\xi^{[t-1]}$ similarly defined. We define the control policies γ and λ, and the policy spaces Γ and Λ in parallel. The evader receives a reward $r^\pi(t) = z_{\pi(t)}(t)$ at each time t; the pursuer receives $b^\xi(t) = x_{\xi(t)}(t)$.

In principle, both sides can consider the same reward maximization problems as in the previous case, by imposing an *arbitrary* belief on the adversarial behavior (i.e., associating with $z_k(t)$ and $x_k(t)$ probabilistic models), which however renders nonsensical notion of optimality. The optimal control in this setting is then typically addressed in the framework of non-stochastic online learning, where existing literature focuses on minimizing the (weak) regret of a strategy compared to a best single-action strategy for any given realization (sample path) of the adversarial behavior. These online learning techniques are employed as our main model for the adaptive behavior of either side.

TABLE I
SUMMARY OF MAIN NOTATION.

$\tau_k(t)/\alpha_k(t)$	(marginal) probability that the evader hides in/the pursuer searches location k at time t
$x_k(t)/z_k(t)$	indicator variable of whether the evader hides in/the pursuer searches location k at time t
$\pi(t)/\xi(t)$	index of the location where the evader hides/the pursuer searches at time t
$r(t)/b(t)$	single-step reward of the evader/pursuer at time t
$\omega(t)/\theta(t)$	private randomization device of the evader/pursuer
γ_t/λ_t	control policy of the evader/pursuer
T	finite time horizon
m	total number of locations
$\mathcal{C}$	index set of locations
M	number of locations that can be simultaneously searched by the pursuer

III. Optimal Evasion Against Adaptive Pursuit

A. Against single-location pursuit

We start by considering a pursuer who is only capable of searching one location at a time. Both sides decide which location to use (for hiding or searching) at the beginning of a time step and cannot change their mind till the next step. Both sides also receive feedback by the end of a step: the evader finds out whether it has been discovered by the pursuer, while the pursuer finds out which location the evader has been hiding. In other words, we assume the pursuer could scan through the locations to find out *after the fact* the evader's action, although it needs to make the right decision a priori in order to make the pursuit effective (e.g., to have the right resources in place).

The pursuer is not assumed to know the evader's decision making rationale, and thus regards the evader activity variable $x_k(t)$ as deterministic but unknown. Given the full information on past activity in all locations to the pursuer, we assume it adopts the Hedge algorithm presented in [24] by Auer et al.; this is a variant of the original Hedge algorithm (or exponential weights algorithm) introduced by Freund and Schapire [26], within the line of work on multiplicative weights learning [27] (see [28] for an in-depth survey). Hedge is an online learning algorithm in the adversarial multi-arm bandit setting [24], [25], which presumes no probabilistic behavior of the opponent (in our case, the evader). It is shown to guarantee an order-optimal sublinear weak regret[1], which in our context translates into

[1]An algorithm with a sublinear weak regret is also often called no-regret.

sublinear "missing" of discovery opportunities compared to always searching the most active/used location (in hindsight) under an arbitrary evasion policy.

Formally, let $x(t) := (x_k(t), \forall k \in \mathcal{C})$ for $t = 1, \cdots, T$ over a finite horizon T. For any search sequence $\xi = (\xi(1), \xi(2), \ldots)$ and a fixed sequence of evasion $(x(1), x(2), \ldots)$, the total reward of the pursuer at T, denoted by $G_\xi(T)$, is given by

$$G_\xi(T) = \sum_{t=1}^{T} b^\xi(t) = \sum_{t=1}^{T} x_{\xi(t)}(t),$$

while the maximum reward from consistently searching the most evader-active location is

$$G_{\max}(T) = \max_{k\in\mathcal{C}} \sum_{t=1}^{T} x_k(t).$$

Hedge aims to minimize the gap (i.e., regret) between its total reward G_{Hedge} and $G_{\max}$, by selecting locations randomly using an adaptive probability distribution based on past evader activities: it selects the most rewarding (evader-active) location seen in the past with the highest probability. The algorithm is shown below.

Hedge

Parameter: A real number $a > 1$.

Initialization: Set $G_k(0) := 0$ for all $k \in \mathcal{C}$.

Repeat for $t = 1, 2, \ldots, T$

1) Choose location k_t according to the distribution $\alpha(t) = (\alpha_1(t), \alpha_2(t), \ldots, \alpha_m(t))$ on $\mathcal{C}$, where
$$\alpha_k(t) = \frac{a^{G_k(t-1)}}{\sum_{j=1}^{m} a^{G_j(t-1)}}$$
2) Observe (reward) vector $(x_1(t), x_2(t), \ldots, x_m(t))$.
3) Set $G_k(t) = G_k(t-1) + x_k(t)$ for all $k \in \mathcal{C}$.

The performance of Hedge is formally characterized by the following theorem from [24].

Theorem 1: If $a = 1 + \sqrt{2\ln(m)/T}$, then $\mathbb{E}G_{\text{Hedge}}(T) \geq G_{\max}(T) - \sqrt{2T\ln m}$, where the expectation is w.r.t. the randomness in the actions taken by Hedge.

Under our assumption, the evader knows the fact that the pursuer is using Hedge and its initial condition[2]. Due to its perfect recall of past actions, it maintains the correct belief about the evolution of the adversarial behavior $\alpha^\pi(t)$ determined by Hedge. In principle, the finite-horizon problem (1) can be solved backwards using dynamic programming. However, we shall first try to argue intuitively what the optimal policy should behave like. Since Hedge has a sublinear regret for the pursuer, if the evader favors one location, the pursuer will eventually identify this most evader-active location and search it at a rate linear in T and miss it at a rate no more than sublinear in T. It follows that the best strategy for the evader is to use each location equally, either deterministically or stochastically. This intuition indeed provides the precise solution to the infinite-horizon problem (2) as shown below. Let $\overline{r}_\infty := \liminf_{T\to\infty} \mathbb{E}\{\frac{1}{T}\sum_{t=1}^{T} r^\pi(t)\}$. Denote by g the location selection sequence of the greedy policy γ_{greedy}, where $g(t) \in \arg\min_{k\in\mathcal{C}} \alpha_k^g(t)$ for all t. Note that the greedy policy can be deterministic, i.e., independent of the private randomization device $\omega(t)$ or in the case of $\omega(t)$ being a constant.

Theorem 2: $\overline{r}_\infty \leq \frac{m-1}{m}$ for any policy γ, and the greedy policy achieves this upper bound.

Proof: Note that

$$\begin{aligned}\mathbb{E}G^\pi_{\text{Hedge}}(T) &= \mathbb{E}\left\{\sum_{t=1}^{T} x^\pi_{\xi(t)}(t)\right\} = \sum_{t=1}^{T}\sum_{k=1}^{m} x_k^\pi(t)\alpha_k^\pi(t) \\ &= \sum_{t=1}^{T} \alpha^\pi_{\pi(t)}(t) = T - \sum_{t=1}^{T} r^\pi(t)\end{aligned}$$

for any realization of π. Therefore,

$$\begin{aligned}\overline{r}_\infty &= 1 - \limsup_{T\to\infty} \mathbb{E}\left\{\frac{1}{T}\mathbb{E}G^\pi_{\text{Hedge}}(T)\right\} \\ &\leq 1 - \limsup_{T\to\infty} \mathbb{E}\left\{\frac{1}{T}(G^\pi_{\max}(T) - \sqrt{2T\ln m})\right\} \\ &= 1 - \limsup_{T\to\infty} \mathbb{E}\left\{\frac{1}{T}G^\pi_{\max}(T)\right\} \leq \frac{m-1}{m},\end{aligned}$$

for all γ, where the outer expectation is over the randomness of the private randomization device, and the last inequality is due to the fact $G^\pi_{\max}(T) \geq \frac{T}{m}$ for any π.

Under the greedy policy we have $\alpha^g_{g(t)}(t) \leq \frac{1}{m}$ and hence $r^g(t) \geq \frac{m-1}{m}$ for any t, which implies that using γ_{greedy}, $\overline{r}_\infty \geq \frac{m-1}{m}$, i.e., the greedy policy is optimal. ■

Note that the above argument does not invoke any property of Hedge other than the sublinear-regret guarantee, thus the optimality of the greedy policy holds as a countermeasure against the entire family of no-regret algorithms. In particular, there exist no-regret algorithms with less stringent assumptions on the feedback to the pursuer than the assumed perfect posterior observation of the evader's action. For example, only partial observation of the location(s) searched would be assumed if the Exp3 algorithm [25] is used. No-regret algorithms for noisy feedback given by a probabilistic model can also be adapted based on Hedge or other similar algorithms using the one-sample estimate of reward, which is the technique Exp3 utilizes for partial observation, see e.g. [29]. Learning with sublinear regret for delayed feedback has also been recently proposed [30].

The same argument also suggests that some other evading strategies that possess the equal-occupancy property, i.e., long-term, equal amount of presence in each location, can be optimal for the infinite-horizon problem; examples include uniformly and randomly choosing a location in each time step. The same argument, however, may not hold for a finite-horizon problem. This is because the no-regret property is only achieved asymptotically; as we elaborate at the end of this subsection, an equal-occupancy policy is not necessarily optimal for the finite-horizon problem posed in (1). However, the greedy policy is optimal in the finite-horizon case as we show next. It should be noted that this is not an entirely

[2]This is to simplify the presentation; it is possible for the evader to estimate the initial condition of Hedge. The resulting policy however is much more complex than the greedy one derived here. The impact of acquiring estimates can become non-negligible for a finite-horizon problem.

intuitive result. This is because given the limited horizon, the evader could decide to first manipulate the location weights in the pursuer's learning process (by persistent presence in a few selected locations and sacrificing payoff in the short term) and then take advantage of the skewed weights by hiding in other locations. When the weights regain balance the evader can repeat this process. It is not immediately clear whether the greedy policy is necessarily better than this policy. Below we examine this in detail.

Without loss of generality, we shall assume under the greedy policy ties are broken in favor of the lowest-indexed location. Note that since the greedy policy always selects the location least likely to be searched, it eventually (in finite time) leads to equal weights over all locations even if the initial weights under Hedge is unequal. Once the weights are equal, the evader's action is a simple round robin, using locations in the order $1, 2, \cdots, m$. Below we prove the finite-horizon optimality of the greedy policy for a two-location scenario so as to avoid letting technicalities obscure the main idea. The general case is stated in a theorem. For simplicity we drop the superscript π when this dependence is clear from the context.

Lemma 1: In a two-location scenario, the optimal finite-horizon policy yields $\pi(t) \in \arg\min_{k=1,2} \alpha_k(t)$.

Proof: For any policy, let $\Delta(t) := |G_1(t) - G_2(t)|$; this is the difference between the number of times that locations 1 and 2 have been used by the end of step t. Thus $|\Delta(t+1) - \Delta(t)| = 1$ for all t. An example of $\Delta(t)$ up to T is shown in Figure 1: an edge connecting two adjacent time points represents a particular location selection, a down edge indicating the selection of a currently under-utilized location. At t we have

$$r(t) = \begin{cases} \frac{a^{\Delta(t-1)}}{1+a^{\Delta(t-1)}}, & \Delta(t) < \Delta(t-1) \\ \frac{1}{1+a^{\Delta(t-1)}}, & \Delta(t) > \Delta(t-1) \end{cases}.$$

Suppose along any trajectory of $\Delta(t)$ there exists a point $\Delta(t) = d \geq 2$ such that either of the following cases is true: (C1) $d-1 = \Delta(t-1) = \Delta(t+1) < \Delta(t)$, $t < T$; or (C2) $\Delta(T-1) < \Delta(T)$. Then consider a change of policy by "folding" the point at t down in (C1) and the point at T in (C2), as shown by the dashed line in the figure. Clearly, we would only change the reward collected at time t and $t+1$ for the case (C1) and the reward at time T for (C2). Let r' denote the reward of this alternate policy. For (C1) we have

$$\begin{aligned} & r'(t) + r'(t+1) - r(t) - r(t+1) \\ =& \frac{a^{d-1}}{1+a^{d-1}} + \frac{1}{1+a^{d-2}} - \frac{1}{1+a^{d-1}} - \frac{a^d}{1+a^d} \\ =& \frac{1}{1+a^d} + \frac{1}{1+a^{d-2}} - \frac{2}{1+a^{d-1}} > 0 \end{aligned}$$

as $\frac{1}{1+a^x}$ is strictly convex in x for $x > 0$. It is clear the reward also increases in (C2) with this change. Thus the reward can always be increased by folding down such "peaks" if they exist. This eventually leads us to the greedy policy where $\Delta(t) \leq 1$ at all times. ■

Theorem 3: The greedy policy is optimal for the finite-horizon problem (1).

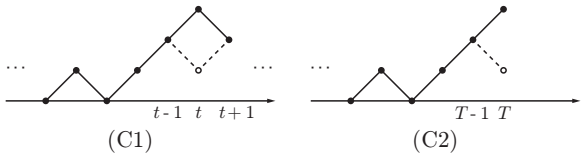

Fig. 1. The change of policy in two cases.

Note that $\alpha(t)$ can be recursively updated as follows:

$$\alpha_k^\pi(t+1) = \frac{\alpha_k^\pi(t) a^{\mathbb{I}(\pi(t)=k)}}{\sum_{j\in\mathcal{C}} \alpha_j^\pi(t) a^{\mathbb{I}(\pi(t)=j)}},$$

with $\mathbb{I}(\cdot)$ being the indicator function. It is therefore only necessary for the evader to recall/store the last control action and the last system state. Note also that the above sequential change-of-policy argument essentially shows the uniqueness of the greedy policy as the optimal policy; thus the policy given earlier for illustration purposes, whereby the evader intentionally skews the weights of locations to take advantage later, is strictly suboptimal. The same result can also be extended to the case where the evader is able to hide and perform its operation in multiple locations simultaneously.

In Figure 2 we plot the finite-horizon (expected) average reward for the greedy and a randomized uniform policy that selects either location with equal probability in a two-location scenario. Our infinite-horizon proof suggests that this latter policy is asymptotically optimal; it is however clearly not optimal for the finite-horizon problem. Based on the proof of Theorem 2, analytically the finite-horizon average reward $\overline{r}_T := \frac{1}{T}\sum_{t=1}^T r(t)$ of the greedy policy is given by

$$\overline{r}_T = \frac{1}{T}\left(\lfloor T/m \rfloor \sum_{j=1}^{m} r(j) + \sum_{j=1}^{(T \bmod m)} r(j) \right)$$

where $r(j) = 1 - \frac{1}{ja+(m-j)}$, while the expected average reward of the uniform policy is simply $\frac{m-1}{m}$. Note that in this two-location example, the zigzag in the reward of the greedy policy when T is small is due to the fact that the single-step reward at an even step is higher than an odd step.

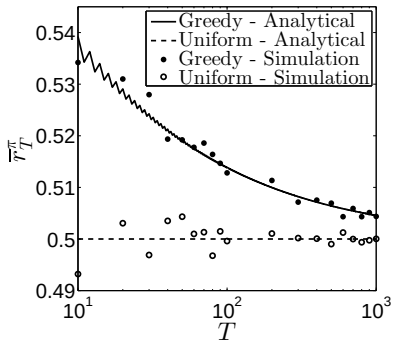

Fig. 2. The finite-horizon (expected) average reward of the greed policy and the uniform policy in a two-location example.

We conclude this part by noting that our formulation implicitly assumes zero detection error when the pursuer selects the right location; similar results can be obtained for the more general case of positive detection error.

B. Against multi-location pursuit

We next consider a pursuer capable of searching $M > 1$ locations simultaneously, with all other assumptions being the same. Accordingly, we assume the pursuer employs the following multiple-play (search) extension of the Hedge algorithm called Hedge-M [3].

Hedge-M

Parameter: A real number $a > 1$.

Initialization: Set $w_k(1) := 1$ for all $k \in \mathcal{C}$.

Repeat for $t = 1, 2, \ldots, T$

1) If $\max_{k\in\mathcal{C}} \frac{w_k(t)}{\sum_{j=1}^m w_j(t)} > \frac{1}{M}$, compute $v(t)$ such that
$$\frac{v(t)}{\sum_{k:w_k(t)\geq v(t)} v(t) + \sum_{k:w_k(t)<v(t)} w_k(t)} = \frac{1}{M},$$
and set $\mathcal{C}_0(t) := \{k : w_k(t) \geq v(t)\}$. Otherwise, set $\mathcal{C}_0(t) := \emptyset$.
2) Set
$$w'_k(t) = \begin{cases} v(t), & k \in \mathcal{C}_0(t) \\ w_k(t), & k \in \mathcal{C} \setminus \mathcal{C}_0(t) \end{cases}.$$
3) Let $\alpha(t) = (\alpha_1(t), \alpha_2(t), \ldots, \alpha_m(t))$ where
$$\alpha_k(t) = M \frac{w'_k(t)}{\sum_{j=1}^m w'_j(t)},$$
and choose M locations with the marginal distribution α, using a subroutine **Dependent Rounding** that returns the set $\mathcal{C}_1(t)$ of locations selected.
4) Observe (reward) vector $(x_1(t), x_2(t), \ldots, x_m(t))$.
5) Set
$$w_k(t+1) = \begin{cases} w_k(t), & k \in \mathcal{C}_0(t) \\ w_k(t)a^{x_k(t)}, & k \in \mathcal{C} \setminus \mathcal{C}_0(t) \end{cases}.$$

Note that $\alpha_k(t)$ is the marginal probability that location k is searched at time t in this case for each k, and their sum is the total number of locations that can be searched, i.e. M. For this reason, Hedge-M generates $\alpha_k(t)$ as in Step 3, after it re-scales the ratios of weights in Step 1 to maintain a probability measure. To compute $v(t)$ in Step 1, one can perform a line search by starting from $v(t) = \max_{k\in\mathcal{C}} w_k(t)$, and decreasing the value of $v(t)$ until the equality is achieved. The subroutine Dependent Rounding [32] draws M out of m items with the given marginal distribution, and can be found in the appendix. For any arbitrary searching strategy $A = (\mathcal{C}_M(1), \mathcal{C}_M(2), \ldots)$, where $\mathcal{C}_M(t)$ is the set of M locations searched at time t, the total reward of the pursuer is given by $G_A(T) = \sum_{t=1}^T \sum_{k\in\mathcal{C}_M(t)} x_k(t)$. The maximum reward $G_{\max}$ of searching the M most evader-active locations is similarly re-defined. The following result shows that Hedge-M also has a sublinear regret w.r.t. consistently searching the M most evader-active locations (in hindsight); the proof is based on that of Hedge [24] and Exp3.M [31].

Theorem 4: If $a = 1 + \sqrt{2\ln(m/M)/(MT)}$, then $\mathbb{E}G_{\text{Hedge-M}}(T) \geq G_{\max}(T) - \sqrt{2\ln(m/M)MT}$, where the expectation is w.r.t. the randomness in the actions taken by Hedge-M.

We first show the optimality of the greedy policy for the infinite-horizon problem. Using the same argument as in the proof of Theorem 2, we have

$$\overline{r}_\infty \leq 1 - \limsup_{T\to\infty} \mathbb{E}\left\{\frac{1}{T} G^\pi_{\max}(T)\right\} \leq \frac{m-M}{m}$$

for any policy, since $G^\pi_{\max}(T) > \frac{TM}{m}$ for any π in the multiple-search case. On the other hand, the greedy policy yields $\alpha^g_{g(t)} \leq \frac{M}{m}$ and hence $r^g(t) \geq \frac{m-M}{m}$ for any t. Therefore, using γ_{greedy}, we have $\overline{r}_\infty \geq \frac{m-M}{m}$, which shows the optimality of the greedy policy. With a bit more effort compared to the single-location pursuit case, we can also obtain the optimality result for the finite-horizon problem. The proof is based on reducing this case to that proved in Theorem 3, and is omitted for brevity.

Theorem 5: The greedy policy is optimal for both the finite- and infinite-horizon problems under the multi-location pursuit.

C. Using a decoy

We now consider the effect of using a *decoy* by the evader, a device capable of performing similar operations as the evader, and indistinguishable to the pursuer (i.e., a double)[4]. Intuitively, the introduction of a decoy can artificially create the impression of a "most evader-active" location so as to attract a majority of the searches, thereby allowing the evader to perform "under the radar" in a location less likely to be searched.

Indeed, this idea can be immediately verified in the infinite-horizon problem, assuming the pursuer is only capable of single-location pursuit. Define a greedy decoy (GD) policy by letting the decoy and the evader respectively select the locations with the highest and the lowest probabilities (the worst and the best locations) to be searched. This policy causes the decoy to persistently transmit in one location, and the evader to use other locations in a round-robin fashion. With a similar argument:

$$r(t) \geq 1 - \frac{a^{\lceil t/(m-1)\rceil}}{a^t + (m-1)a^{\lfloor t/(m-1)\rfloor}} \to 1$$

as $t \to \infty$. Hence,

$$\overline{r}_\infty = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T r^g(t) = \lim_{t\to\infty} r^g(t) = 1.$$

This asymptotic performance is asymptotically optimal and less careful schemes can result in much inferior gain. For example, if the evader and the decoy respectively select the best and the second best locations in each time step (referred to as the doubly greedy (G2) policy), we have

$$\overline{r}_\infty = \lim_{T\to\infty} \frac{2}{m} \sum_{j=0}^{m/2-1} \frac{m-2j-1+2ja}{m-2j+2ja} = \frac{m-1}{m},$$

[3]Hedge-M is reverse-engineered from the algorithm Exp3.M [31], which is a multiple-play algorithm with partial information (the pursuer only observes activities in locations it searched).

[4]In the jamming application, the decoy can be a regular but much cheaper transceiver, one without the ability to receive or perform channel switching.

assuming m even for simplicity. In Figure 3, we plot the finite-horizon average reward for the greedy decoy (GD) policy, the doubly greedy (G2) policy, and the original greedy policy without a decoy (GwoD) as a baseline. As can be seen, GD significantly outperforms the others.

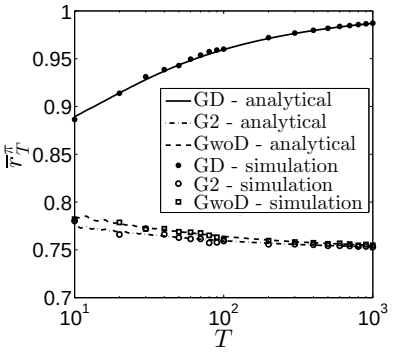

Fig. 3. The finite-horizon average reward of the greedy decoy (GD) policy, the doubly greedy (G2) policy, and the greedy policy without the decoy (GwoD) in a system of four locations.

We now show that GD is also optimal for the finite-horizon problem (1). Note that Hedge can start from any (non-zero) initial condition without affecting the scaling of the regret w.r.t. the horizon. Given any set of the exponents of weights at t, i.e., $(G_k(t-1))_{k\in\mathcal{C}}$, let $\mathcal{L}(t) = \arg\max_{k\in\mathcal{C}} G_k(t-1)$. The optimality result is then established using the following two lemmas.

Lemma 2: For any given horizon T and any initial condition, an optimal policy is such that the decoy always uses a location from $\mathcal{L}(t)$ before the horizon and the evader from $\mathcal{C}\setminus\mathcal{L}(t)$.

Lemma 3: Given the decoy always uses the worst location, it is optimal for the evader to select the best location.

Combining these lemmas we have the following result.

Theorem 6: The greedy decoy policy is optimal for the finite-horizon problem, i.e., it is optimal to let the decoy and the evader respectively select the worst and the best locations in each time step.

The above result can be readily extended to the case when the pursuer is capable of searching multiple locations simultaneously, with the evader deploying multiple decoys at or exceeding the number of locations the pursuer is capable of searching.

We can obtain the same asymptotic performance as using a single decoy against single-location pursuit. In essence, the use of decoys *cancels out* or neutralizes the adversarial effect[5]. Conversely, the pursuer can increase the number of locations it searches (if it has the resources) to counter the effect of decoys. However, the mere possibility of using a decoy can create interesting and difficult dilemmas for the pursuer as we elaborate in Section V-B.

IV. Optimal Pursuit Against Adaptive Evasion

We next consider the parallel problem for the pursuer when the evader hides adaptively. We now have the opposite situation: the evader does not know the decision process of the pursuer, and regards its action $z_k(t)$ as a deterministic but unknown value. Both sides receives feedback after a decision: the pursuer on whether the search is successful, and the evader on which location is searched regardless of its success. The evader adopts the Hedge algorithm given its full information on the pursuer's action after the fact, and the pursuer is aware of the evader's using Hedge.

Due to the symmetry between this and the previous sections, most results can be readily obtained along similar reasoning. For this reason we only highlight the main difference and will limit our attention to the single-location pursuit. To avoid ambiguity, we separately introduce the notation for the evader's version of Hedge. Denote by $R_k(t)$ the exponent of the weight assigned to location k at time t, and $R_k(t) = R_k(t-1)+z_k(t)$. The probability that the evader chooses location k is then given by $\tau_k(t) = \frac{a^{R_k(t-1)}}{\sum_{j\in\mathcal{C}} a^{R_j(t-1)}}$. Denote by $R_{\text{Hedge}}(T)$ the total reward of the evader at a horizon T under Hedge and by $R_{\max}(T)$ the total reward from consistently hiding in the least searched location in hindsight. Recall that $\xi = (\xi(1), \xi(2), \ldots)$ denotes the search sequence of a policy λ by the pursuer, and $b^\xi(t)$ its expected reward at time t. Observe that

$$\mathbb{E}R^\xi_{\text{Hedge}} = \mathbb{E}\left\{\sum_{t=1}^T z^\xi_{\pi(t)}(t)\right\} = \sum_{t=1}^T\sum_{k=1}^m z^\xi_k(t)\tau^\xi_k(t)$$
$$= \sum_{t=1}^T\sum_{k\neq\xi(t)} \tau^\xi_k(t) = T - \sum_{t=1}^T b^\xi(t)$$

Let $\bar{b}_\infty := \liminf_{T\to\infty}\mathbb{E}\{\frac{1}{T}\sum_{t=1}^T b^\xi(t)\}$. Using a similar argument as for the evader, we can obtain

$$\bar{b}_\infty \le 1 - \limsup_{T\to\infty}\mathbb{E}\left\{\frac{1}{T}R_{\max}(T)\right\} \le \frac{1}{m}$$

since $R_{\max}(T) \ge T\frac{m-1}{m}$ for any ξ. Define a greedy policy λ_{greedy}, of which the search sequence is given by $\tilde{g}(t) \in \arg\max_{k\in\mathcal{C}}\tau^{\tilde{g}}_k(t)$. It is clear that $b^{\tilde{g}}(t) \ge \frac{1}{m}$, implying the optimality of λ_{greedy} for the infinite-horizon problem. The same can be established for the finite-horizon problem. Consider the two-location scenario in Section III as an example, and define $\tilde{\Delta}(t) := |R_1(t) - R_2(t)|$. One can similarly find that

$$b(t) = \begin{cases} \frac{a^{\tilde{\Delta}(t-1)}}{1+a^{\tilde{\Delta}(t-1)}}, & \tilde{\Delta}(t) < \tilde{\Delta}(t-1) \\ \frac{1}{1+a^{\tilde{\Delta}(t-1)}}, & \tilde{\Delta}(t) > \tilde{\Delta}(t-1) \end{cases}.$$

Hence using the same argument, the optimality of λ_{greedy} can be shown.

Theorem 7: The greedy policy is optimal for the pursuer in both the infinite- and finite-horizon problems when the evader adopts Hedge.

V. Against Unknown Adversarial Behavior

We now turn to the more realistic case where both sides presume no knowledge of the reasoning used by the opponent, and accordingly employ their respective learning techniques.

[5]This greedy decoy policy can also be shown to be optimal over a finite horizon against multi-location pursuit; the technical detail is omitted for brevity.

A. Hiding versus multi-location seeking

We first consider the case when each side has full posterior information on its adversary's action, and thus respectively adopts Hedge and Hedge-M as the hiding and seeking strategies, though this fact is unknown to the other side. Note that our pursuit-evasion game is constant-sum with a location-independent reward (cf. Section VI where the reward is location-dependent). From known results on the convergence of no-regret learning algorithms to Nash equilibrium (NE) in constant-sum games [33, Chapter 4], it immediately follows that Hedge and Hedge-M are *mutually best responses* for the infinite-horizon problem, up to a diminishing term over a finite horizon.[6] Also note that the above results suggest that Hedge results in the same average reward for the evader compared to the case when it knows that the pursuer is using Hedge-M and responds optimally (Section III-B). This shows that there is no loss of optimality when using online learning techniques against an unknown pursuer who is also an online learner with sublinear regret guarantee. On the other hand, this result also indicates that the information asymmetry between the two interacting players, which is given by the different levels of knowledge on the learning rationale of the opponent, does not yield advantage to the one that possesses additional information (which is the evader in this setup). Nevertheless, as we shall see in the next subsection, the information asymmetry caused by different abilities to distinguish between the actual opponent and a decoy device can be a determining factor of the long term outcome.

Moreover, the above conclusion holds when the evader only gets to find out whether a search is conducted in the location it happens to be hiding, but not otherwise (as opposed to finding out after the fact the set of locations searched, as we have previously assumed). This results in partial information for the evader (or called the bandit setup for the evader), and for this reason it can no longer use Hedge. In this case its partial information counterpart Exp3 [24], [25] can be used to update its probability $\tau_k(t)$ of choosing location k at t. Then, the mutual optimality between Exp3 and Hedge-M is also implied by the aforementioned general learning convergence result, and for the same reason the mutual optimality in fact holds for *any* pair of no-regret algorithms for our hide-and-seek problem.

B. Using a decoy

We re-examine the idea where the evader employs a decoy but assumes no knowledge on the pursuer, which makes using the decoy as a camouflage more difficult. Toward this end we make the important observation that if the most evader-active location is unique and *dominant*, that is, there exists a unique location k, such that for any subsequence of time $\{t_i\}_{i=1}^{I(T)} \subseteq \{1, 2, \ldots, T\}$ with $I(T)$ of the order of $\Theta(T)$,

$$\liminf_{T\to\infty} \frac{1}{I(T)} \sum_{i=1}^{I(T)} x_k(t_i) > \limsup_{T\to\infty} \frac{1}{I(T)} \sum_{i=1}^{I(T)} x_j(t_i)$$

for any $j \neq k$, then the pursuer can guarantee sublinear weak regret (uniformly or asymptotically) if and only if all suboptimal locations are searched with time sublinear in T asymptotically. In other words, a strategy that guarantees sublinear weak regret for the pursuer must ultimately identify and aim for the dominantly evader-active location if any. Therefore, the evader can always use the decoy to "create" this dominant location while performing operations in a virtually search-free environment, by letting the decoy reside in one location and using an algorithm like Exp3 on the rest $m-1$ locations. This will result in an asymptotic average reward of 1, the same as in the case when the adversarial behavior is known.

Embedded in this observation is an interesting dilemma that the pursuer faces in the presence of the *possibility* of a decoy that it cannot distinguish. On one hand, if the pursuer adopts a no-regret algorithm like Hedge (or Hedge-M), arguably the best class of algorithms to use under uncertainty, then it is setting itself up for a very effective decoy defense by the evader, so much so that its search is rendered useless (asymptotically). This is the point illustrated above. On the other hand, if for this reason the pursuer decides not to use such algorithms, then it may face a worse outcome as the alternative algorithm may provide no performance/regret guarantee. In this sense the mere possibility or threat of using a decoy may be viewed as effective defense.

VI. Application to Jamming Defense: Against Adaptive Adversary with Heterogeneous Rewards

In the context of jamming defense in a multi-channel communication system, the evader and the pursuer respectively model a legitimate user that attempts data transmission and a jamming attacker, and each location represents one channel. In this section, we consider the case when the reward associated with each successful evasion (data transmission) or pursuit (jamming attack) is *location-dependent*, which models the variable amount of deliverable data of a transmission as the result of the dynamic channel bandwidth (data rates) with spectral diversity. We denote the bandwidth by μ_k for each channel $k \in \mathcal{C}$, the value of which we assume is known to both the user and the attacker, and we accordingly re-define the reward that the user obtains from using channel $\pi(t)$ as $r^\pi(t) = \mu_{\pi(t)}(1 - \alpha_{\pi(t)}(t))$ or $\mu_{\pi(t)} z_{\pi(t)}(t)$, depending on the knowledge of the user on the attacker's behavior. We also re-define the reward of attacking channel $\xi(t)$ as $b^\xi(t) = \mu_{\xi(t)} x_{\xi(t)}(t)$ when the attacker has no information on the reasoning used by the user, which is assumed throughout this section. We will focus on the single-attack case (i.e., single-location pursuit), and we show optimality results for the infinite-horizon problem. Also, we change our terminology in accordance to the context of application.

A. Against known attack pattern

In this part, we consider the problem in parallel to that presented in Section III. We assume the attacker adopts the Hedge algorithm, which is known to the user, and the step of updating weights in Hedge (step 3)) is accordingly $G_k(t) =$

[6] In our setup of the pursuit-evasion problem, the underlying constant-sum game is of the type of the matching-pennies games, and has a unique NE.

$G_k(t-1)+\mu_k x_k(t)$ for all $k \in \mathcal{C}$. Without loss of generality, we assume that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_m$. Given any sequence of channel selection π, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mu_{\pi(t)} = \frac{1}{T}\sum_{k=1}^{m}\ell_k^{\pi}(T)\mu_k = \sum_{k=1}^{m}a_k^{\pi}(T)\mu_k$$

where $\ell_k^{\pi}(T) = |\{t \leq T : \pi(t) = k\}|$ and $a_k^{\pi}(T) = \ell_k^{\pi}(T)/T$. Using a similar argument as before, the average gain of the attacker when using Hedge is given by

$$\mathbb{E}G_{\text{Hedge}}^{\pi}(T) = \sum_{t=1}^{T}\mu_{\pi(t)} - \sum_{t=1}^{T}r^{\pi}(t)$$

for any realization of π, and thus,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}r^{\pi}(t) &= \frac{1}{T}\sum_{t=1}^{T}\mu_{\pi(t)} - \frac{1}{T}\mathbb{E}G_{\text{Hedge}}^{\pi}(T)\\ &\leq \sum_{k=1}^{m}a_k^{\pi}(T)\mu_k - \frac{1}{T}(G_{\max}^{\pi}(T) - \sqrt{2T\ln m})\\ &= \sum_{k=1}^{m}a_k^{\pi}(T)\mu_k - \max_{k\in\mathcal{C}}a_k^{\pi}(T)\mu_k + \sqrt{2\ln m/T},\end{aligned}$$

where $G_{\max}^{\pi}(T) = \max_{k\in\mathcal{C}}\ell_k^{\pi}(T)\mu_k$ by definition. Hence,

$$\overline{r}_{\infty} \leq \liminf_{T\to\infty}\mathbb{E}\left\{\sum_{k=1}^{m}a_k^{\pi}(T)\mu_k - \max_{k\in\mathcal{C}}a_k^{\pi}(T)\mu_k\right\}$$

for any γ. Consider the following optimization problem

$$\underset{a\in\Delta_m}{\text{maximize}} \quad \sum_{k=1}^{m}a_k\mu_k - \max_{k\in\mathcal{C}}a_k\mu_k, \tag{3}$$

where Δ_m is the set of distributions over $\mathcal{C}$ and $a = (a_k, k \in \mathcal{C})$. We denote an optimal solution by $a^{\star}$, and we then have

$$\overline{r}_{\infty} \leq \sum_{k=1}^{m}a_k^{\star}\mu_k - \max_{k\in\mathcal{C}}a_k^{\star}\mu_k,$$

for any policy γ. Let $\text{supp}(a) = \{k \in \mathcal{C} : a_k > 0\}$ for any feasible solution a, and let $K^{\star} = |\text{supp}(a^{\star})|$.

Lemma 4: For any optimal solution $a^{\star}$, 1) $a_k^{\star}\mu_k = a_j^{\star}\mu_j$ for any $k, j \in \text{supp}(a^{\star})$, and 2) $\text{supp}(a^{\star})$ consists of the indices of channels with the $K^{\star}$ highest bandwidth.

Without loss of generality, we assume that $\text{supp}(a^{\star}) = \{1, 2, \ldots, K^{\star}\}$. Hence, $a_k^{\star} = \frac{1/\mu_k}{\sum_{j=1}^{K^{\star}}1/\mu_j}$ for all $k \leq K^{\star}$. The optimal value of the problem (3) is then given by $(K^{\star}-1)/\sum_{k=1}^{K^{\star}}1/\mu_k$. Note that $\Gamma(K) := (K-1)/\sum_{k=1}^{K}1/\mu_k$ is an increasing function of K for $K = 1, 2, \ldots, m$. Hence, the optimal value as well as $K^{\star}$ can also be readily obtained without solving (3). Using the above lemma, we obtain

$$\overline{r}_{\infty} \leq \frac{K^{\star}-1}{\sum_{k=1}^{K^{\star}}1/\mu_k},$$

for any policy in Γ. Given the value of $K^{\star}$ as in Lemma 4, consider now the greedy policy γ_{greedy} with the channel selection sequence g, where $g(t) \in \arg\min_{k\leq K^{\star}}\alpha_k^g(t) = \arg\min_{k\leq K^{\star}}G_k(t)$, and we have the following result.

Theorem 8: The greedy policy is optimal.

Proof: Using the greedy policy, we have $\alpha_{g(t)}^g(t) \leq \frac{1}{K^{\star}}$ for all t and thus

$$r^g(t) \geq \mu_{g(t)}\left(1 - \frac{1}{K^{\star}}\right) = \frac{K^{\star}-1}{K^{\star}}\mu_{g(t)}.$$

Therefore,

$$\begin{aligned}\overline{r}_{\infty} &\geq \frac{K^{\star}-1}{K^{\star}}\liminf_{T\to\infty}\mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T}\mu_{g(t)}\right\}\\ &= \frac{K^{\star}-1}{K^{\star}}\liminf_{T\to\infty}\mathbb{E}\left\{\sum_{k=1}^{m}a_k^g(T)\mu_k\right\}\\ &\geq \frac{K^{\star}-1}{K^{\star}}\mathbb{E}\left\{\liminf_{T\to\infty}\sum_{k=1}^{m}a_k^g(T)\mu_k\right\},\end{aligned}$$

where the expectation is taken w.r.t. the private randomization device, and the last inequality is due to Fatou's lemma. Fix any realization of g. Since

$$|G_k^g(t) - G_j^g(t)| \leq \max_{l\leq K^{\star}}\mu_l = \mu_1$$

for any $k, j \leq K^{\star}$ due to the greedy nature of the policy, and $a_k(T)\mu_k = G_k(T)/T$ for any k, we have

$$\lim_{T\to\infty}|a_k^g(T)\mu_k - a_j^g(T)\mu_j| = 0,$$

for any $k, j \leq K^{\star}$, which implies

$$\liminf_{T\to\infty}a_k^g(T)\mu_k = \liminf_{T\to\infty}a_j^g(T)\mu_j,$$

for any $k, j \leq K^{\star}$, and that if $\{T_l\}_{l=1}^{\infty}$ is a subsequence of time that achieves the limit inferior of $a_k^g(T_l)\mu_k$, it is also a subsequence that achieves the limit inferior of $a_j^g(T_l)\mu_j$. Combining the above implications, we have

$$\liminf_{T\to\infty}\sum_{k=1}^{m}a_k^g(T)\mu_k = K^{\star}\liminf_{T\to\infty}a_k^g(T)\mu_k = K^{\star}a_k^g\mu_k,$$

for some $a_k^g > 0$ for each $k \leq K^{\star}$, and moreover $\sum_{k=1}^{K^{\star}}a_k^g = 1$. Finally, $a_k^g\mu_k = \frac{1}{\sum_{j=1}^{K^{\star}}1/\mu_j}$ for all $k \in \text{supp}$, and we have

$$\overline{r}_{\infty} \geq \frac{K^{\star}-1}{\sum_{k=1}^{K^{\star}}1/\mu_k},$$

which establishes the optimality of the greedy policy. ■

B. Against unknown attack pattern

Given the location-dependent reward, the underlying pursuit-evasion game is no longer constant-sum, and the known learning convergence result is not directly applicable. Nonetheless, we provide a simple proof below to show the mutual optimality between a pair of no-regret learning algorithms similarly holds. We assume the attacker and the user respectively adopt randomized learning algorithms λ and γ with sublinear weak regret, which is unknown to the other side. Given the no-regret feature of λ, the average reward of the user is upper bounded by the optimal solution to (3) as shown before. Let the total reward of the user be R_{γ}, and then

$$\mathbb{E}R_{\gamma}(T) \geq R_{\max}(T) - o(T),$$

where the expectation is taken w.r.t. the private randomness in γ, which is independent from any attacker's behavior. Let $\xi = (\xi(1), \xi(2), \ldots, \xi(T))$ be an arbitrary jamming sequence of the attacker, where $\xi(t) \in \mathcal{C}$, and set $\ell_k^\xi(T) = |\{t \le T : \xi(t) = k\}|$ and $c_k(T) = \ell_k^\xi(T)/T$. Since $R_{\max}(T) = \max_{k\in\mathcal{C}}(T - \ell_k^\xi(T))\mu_k$, we have the average reward of the user using a no-regret γ is lower bounded as

$$\overline{r}_\infty \ge \liminf_{T\to\infty} \mathbb{E}\left\{\max_{k\in\mathcal{C}}(1-c_k(T))\mu_k\right\}.$$

Consider the optimization problem

$$\underset{c\in\Delta_m}{\text{minimize}} \quad \max_{k\in\mathcal{C}}(1-c_k)\mu_k, \tag{4}$$

and denote its optimal $\overline{r}_l$, which is thus a lower bound of $\overline{r}_\infty$. In fact, we have $\overline{r}_l = \frac{K^\star - 1}{\sum_{k=1}^{K^\star} 1/\mu_k}$.

Lemma 5: Problem (4) has the same optimal value as (3).

Proof: Reformulate the problem (3) and problem (4) respectively as follows:

$$\max \sum_{k=1}^m a_k\mu_k - b \quad \text{s.t. } a_k\mu_k \le b, a_k \ge 0, \forall k, \sum_{k=1}^m a_k = 1, \tag{5}$$

and

$$\min d \quad \text{s.t. } (1-c_k)\mu_k \le d, c_k \ge 0, \forall k, \sum_{k=1}^m c_k = 1. \tag{6}$$

It can be shown that the problem (5) is equivalent to the dual problem of (6), and the result then follows. ∎

The above results show that any pair of policies γ and λ with sublinear weak regret are mutually best responses, as in the formulation with homogeneous reward for both sides.

VII. Rendezvous and Collision Problems

In this section, we present a unified framework that extends our formulation of the pursuit-evasion problem to two other important families of problems, namely the rendezvous and the collision problems. The rendezvous problem arises in communication systems using dynamic spectrum access [34] where two radio transceivers attempt to meet in a common channel to communicate, or autonomous robotic systems [35] where two robots attempt to come within the range of each other. The collision problem arises in various resource sharing scenarios in wireless networks, where different radio transceivers try to switch to different channels in order to minimize interference, see e.g., [36]. We start by describing the unified formulation, and show to what extent the results obtained under the pursuit-evasion model extends to these two problems. We then discuss in detail the associated open problems.

A. A unified framework for the three problems

Given two interacting players, each player takes an action $a_i(t)$ at time t from the action space $\mathcal{A}_i$ where $i = 1$ or 2, and respectively receives a reward $r_i(t) := r_i(a_1(t), a_2(t))$ as a function of their actions. Denote by $\mathcal{I}_i(t)$ the informational state of player i at time t, which consists of all information available to the player for decision-making, and by g_t^i the decision rule at time t, that is, $a_i(t) = g_t^i(\mathcal{I}_i(t))$. Denote by $g^i = (g_1^i, g_2^i, \ldots)$ the decision policy, which is the collection of decision rules, and let the space of all policies be $\mathcal{G}$. Given a probabilistic belief on the other player's strategy, each player can then consider the optimization problems

$$\max_{g\in\mathcal{G}} \mathbb{E}\left\{\sum_{t=1}^T r_i(t)\right\}, \text{ and } \max_{g\in\mathcal{G}} \liminf_{T\to\infty} \mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^T r_i(t)\right\},$$

where T is a finite time horizon, and the expectation is taken w.r.t. any randomness involved in the evaluation of the reward. Depending on the nature of the payoff, three families of problems are given as follows:

1) Pursuit-Evasion problem. Assume that player 1 is the pursuer and player 2 is the evader. We have $r_1(a_1, a_2) = h_1(a_2)\cdot\mathbb{I}(a_1 = a_2)$ and $r_2(a_1, a_2) = h_2(a_2)\cdot\mathbb{I}(a_1 \ne a_2)$, where h_t^i is some bounded mapping from $\mathcal{A}_i$ to $\mathbb{R}_+$. In our previous study, we had $h_(a_2) = 1$ or μ_{a_2}.
2) Rendezvous problem. $r_i(a_1, a_2) = h_i(a_i)\cdot\mathbb{I}(a_1 = a_2)$ for $i = 1, 2$.
3) Collision problem. $r_i(a_1, a_2) = h_i(a_i)\cdot\mathbb{I}(a_1 \ne a_2)$ for $i = 1, 2$.

B. Generalization

Following our framework for the pursuit-evasion problem, we discuss the rendezvous and the collision problems using a similar structure. We assume player 2 has no knowledge of the (algorithmic) reasoning of player 1 as we did for the pursuer in the pursuit-evasion problem. We again consider two versions of either problem, depending on player 1's knowledge, which we shall refer to as the "known" case and the "unknown" case with a bit abuse of language. In the following, we briefly formulate the rendezvous problem, while the collision problem can be understood from the context (c.f. Section VII-A). Our notation is reproduced (and re-defined when necessary) from previous sections. Let $x_k(t)$ and $z_k(t)$ be the variables indicating respectively the activities of players 1 and 2, with $x_k(t) = 1$ ($z_k(t) = 1$) if player 1 (resp. 2) is at location k at time t and $x_k(t) = 0$ ($z_k(t = 0)$) otherwise.

Given a selection sequence π under a policy γ, of which the re-definition is standard and thus omitted, when player 1 regards $z_k(t)$ as stochastic with $P(z_k(t) = 1) = \alpha(t)$ (i.e., the known case), it has a mean reward $r^\pi(t) = \alpha_{\pi(t)}(t)$ at time t; when $z_k(t)$ is considered non-stochastic (i.e., the unknown case), we have $r^\pi(t) = z_{\pi(t)}(t)$. As for player 2, it regards $x_k(t)$ as non-stochastic and receives $b^\xi(t) = x_{\xi(t)}(t)$ when using the search sequence ξ. We assume both sides have feedback on the other's action at the end of an interacting round, and player 2 employs online technique as its strategy for location selection, in particular the Hedge algorithm.

For the known case, the analysis of the optimal policy is nothing but the parallel adaption of what we have developed for the pursuit-evasion problem, and the optimality of the greedy policy follows. The greedy policy for the rendezvous and the collision problem will be respectively to choose the location with the maximum and the minimum probability $\alpha_k(t)$ of player 2's occurrence; in words, it is simply to stay

in the same location for both problem, e.g., "be there or be square" for rendezvous.

C. Open problems

Interestingly, for the unknown case when both sides adopt no-regret learning, the previous analysis from the pursuit-evasion problem fails to provide a sharp characterization of the performance (e.g. the asymptotic mutual optimality of strategies). Given the constant-sum nature of the pursuit-evasion problem, the no-regret property of learning techniques provides either player tight lower (performance guarantee for one player) and upper (performance guarantee for its opponent) reward bounds, thus establishing the convergence of learning limits, as we have seen from the existing analysis on the connection between no-regret learning and minimax theorem in constant-sum games [33]. Applying the same argument in the rendezvous and the collision problems, however, one only obtains lower bounds for both players, which is distinct from the optimal solution for both sides. For example, in a two-location scenario of seeking rendezvous, the no-regret property only provides a lower bound 1/2 on the average reward for each player, while the optimal reward would be 1 for both players when they manage to only disagree on the location a sublinear number of times. We note that the rendezvous and collision problems have pure NEs (in a game-theoretical sense) as optimal solutions for both players, while the pursuit-evasion problem possesses no pure NE, and hence results on the convergence of learning limits for general-sum games (that are not necessarily of constant sums, including the rendezvous and the collision problem) and their relation to game-theoretical solution concepts, if exist, will be a key to solving the unknown case.

This subject has been an active research field, and it has been shown no-regret dynamics may not converge to NE in general games [37]. The generic characterization of the learning limit using no-regret algorithms is concerned with weaker notions of equilibria than NE [38] [39]. As to the convergence of learning to NEs, there are a few affirmative results in special cases [40], [41], while none of them addresses the rendezvous and the collision problems that we have posed in this section.

VIII. Concluding Remark

Modeling individual behavior from a learning perspective as shown in this paper typically requires weaker knowledge assumptions than a game theoretical framework does. Interestingly, the convergence of these learning algorithms has been shown to be closely related to game theoretical solution concepts. The learning perspective thus provides a different and possibly more natural angle in interpreting certain game-theoretic results. Extending the "two-player" scenario investigated in this paper to groups of evaders and pursuers is an interesting direction of future research. From an application viewpoint, incorporating temporal variation in the reward process and the impact of attacks on the upper-layer control mechanism, e.g. TCP, are also important open issues.

Appendix A
Proofs

Proof of Theorem 3: Define $\Delta_{ij}(t) := G_i(t) - G_j(t)$. Then,

$$\alpha_k(t) = \frac{1}{\sum_{j=1}^{m} a^{\Delta_{jk}(t-1)}},$$

and

$$r^{\pi}(t) = \frac{\sum_{j\neq\pi(t)} a^{\Delta_{j\pi(t)}(t-1)}}{1+\sum_{j\neq\pi(t)} a^{\Delta_{j\pi(t)}(t-1)}}.$$

Let $\mathcal{K}(t) = \arg\min_{k\in\mathcal{C}} G_k(t)$, and define $\mathcal{T} = \{t \leq T : \max_{k\notin\mathcal{K}(t)} \Delta_{k,j}(t) \geq 2, j \in \mathcal{K}(t)\}$. Suppose that $\mathcal{T} \neq \emptyset$, and let $t_0 = \min \mathcal{T}$. Then, either (C1) there exists some time t_1 with $t_0 < t_1 \leq T$ when some location $j \in \mathcal{K}(t_0)$ is selected for the first time after t_0 by the evader or (C2) any location $j \in \mathcal{K}(t_0)$ is never selected by the horizon T.

Consider first the case (C1). Without loss of generality, assume that the location selected at t_1-1 is 2 and 1 is chosen at t_1. Let $\Delta_{ij}(t_1-1) = d_{ij}$. Then,

- $\Delta_{ij}(t_1) = \Delta_{ij}(t_1+1) = d_{ij}$ for all $i,j \geq 3$;
- $\Delta_{1j}(t_1) = d_{1j}$ for all $j \geq 3$, $\Delta_{12}(t_1) = d_{12}-1$, $\Delta_{1j}(t_1+1) = d_{1j}+1$ for all $j \geq 3$, and $\Delta_{12}(t_1) = d_{12}$;
- $\Delta_{2j}(t_1) = d_{2j}+1$ for all $j \neq 2$, $\Delta_{2j}(t_1+1) = d_{2j}+1$ for all $j \geq 3$, and $\Delta_{21}(t_1+1) = d_{21}$.

Consider now a change of policy by selecting location 1 at t_1-1 and location 2 at t_1. Denote Δ under this new policy by Δ'. Then,

- $\Delta'_{ij}(t_1) = \Delta'_{ij}(t_1+1) = d_{ij}$ for all $i,j \geq 3$.
- $\Delta'_{1j}(t_1) = d_{1j}+1$ for all $j \geq 2$, $\Delta'_{1j}(t_1+1) = d_{1j}+1$ for all $j \geq 3$, and $\Delta'_{12}(t_1) = d_{12}$;
- $\Delta'_{2j}(t_1) = d_{2j}$ for all $j \geq 3$, $\Delta'_{21}(t_1) = d_{21}-1$, $\Delta'_{2j}(t_1+1) = d_{2j}+1$ for all $j \geq 3$, and $\Delta'_{21}(t_1+1) = d_{21}$.

Hence, this change of policy only affects the reward of the evader collected at t_1-1 and t_1. Denote by r' the reward under this alternative policy, and we have

$$\begin{aligned}
&r'(t_1-1)+r'(t_1)-r(t_1-1)-r(t_1)\\
&= \frac{\sum_{k\geq3} a^{d_{k1}} + a^{d_{21}}}{1+\sum_{k\geq3} a^{d_{k1}} + a^{d_{21}}} + \frac{\sum_{k\geq3} a^{d_{k2}} + a^{d_{12}+1}}{1+\sum_{k\geq3} a^{d_{k2}} + a^{d_{12}+1}}\\
&\quad - \frac{\sum_{k\geq3} a^{d_{k2}} + a^{d_{12}}}{1+\sum_{k\geq3} a^{d_{k2}} + a^{d_{12}}} - \frac{\sum_{k\geq3} a^{d_{k1}} + a^{d_{21}+1}}{1+\sum_{k\geq3} a^{d_{k1}} + a^{d_{21}+1}}\\
&= \frac{1}{1+C+a^{d_{21}+1}} + \frac{1}{1+D+a^{d_{12}}}\\
&\quad - \frac{1}{1+C+a^{d_{21}}} - \frac{1}{1+D+a^{d_{12}+1}},
\end{aligned}$$

where $C = \sum_{k\geq3} a^{d_{k1}}$ and $D = \sum_{k\geq3} a^{d_{k2}}$. Note that $C = Da^{d_{21}}$ and $d_{12} = -d_{21}$. Set $d = d_{21}$, and we obtain

$$\begin{aligned}
&r'(t_1-1)+r'(t_1)-r(t_1-1)-r(t_1)\\
&= \frac{1}{1+Da^d+a^{d+1}} + \frac{1}{1+D+a^{-d}} - \frac{1}{1+Da^d+a^d} -\\
&\quad - \frac{1}{1+D+a^{-d+1}}\\
&= \frac{(a^{2d-1}-a^{d-1})(a-1)^2}{(1+Da^d+a^{d+1})(1+Da^d+a^d)(1+Da^{d-1}+a^{d-1})}\\
&> 0.
\end{aligned}$$

For (C2), it is clear that alternatively selecting location 1 at T results in a higher reward.

Therefore, the optimal policy would never allow the difference between the times that any two locations are selected to be greater than 2. In other word, the optimal policy always selects the most under-utilized location. When there are multiple locations with the same lowest number of times of the evader's presence, the evader would be indifferent in selecting any location between/among them, since locations are symmetric (and the reward is only related to the relative difference between the numbers of location usage). ■

Proof of Theorem 4: Let $W_t := \sum_{k=1}^m w_k(t)$ and $W'_t := \sum_{k=1}^m w'_k(t)$, and let $a = 1+\theta$ for some $\theta > 0$. Denote $\mathcal{C} \setminus \mathcal{C}_0(t)$ by $\mathcal{C}_0^c$. Then, for any $t \le T$,

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{k\in\mathcal{C}_0^c(t)} \frac{w_k(t+1)}{W_t} + \sum_{k\in\mathcal{C}_0(t)} \frac{w_k(t+1)}{W_t} \\
&= \sum_{k\in\mathcal{C}_0^c(t)} \frac{w_k(t)}{W_t}(1+\theta)^{x_k(t)} + \sum_{k\in\mathcal{C}_0(t)} \frac{w_k(t)}{W_t} \\
&\le \sum_{k\in\mathcal{C}_0^c(t)} \frac{w_k(t)}{W_t}(1+\theta x_k(t)) + \sum_{k\in\mathcal{C}_0(t)} \frac{w_k(t)}{W_t} \\
&= 1+\theta \sum_{k\in\mathcal{C}_0^c(t)} \frac{w_k(t)}{W_t} x_k(t) = 1+\theta \frac{W'_t}{W_t} \sum_{k\in\mathcal{C}_0^c(t)} \frac{w'_k(t)}{W'_t} x_k(t) \\
&\le 1+\theta \sum_{k\in\mathcal{C}_0^c(t)} \alpha_k(t) x_k(t),
\end{aligned}$$

where the first inequality is due to the fact that $x_k(t) \in \{0,1\}$. Therefore,

$$\begin{aligned}
\ln \frac{W_{T+1}}{W_1} &= \sum_{t=1}^T \ln \frac{W_{t+1}}{W_t} \le \sum_{t=1}^T \ln \left(1+\theta \sum_{k\in\mathcal{C}_0^c(t)} \alpha_k(t) x_k(t)\right) \\
&\le \theta \sum_{t=1}^T \sum_{k\in\mathcal{C}_0^c(t)} \alpha_k(t) x_k(t) \qquad (7)
\end{aligned}$$

where the last inequality is due to $\ln(1+x) \ge x$. On the other hand, let $A^* \subset \mathcal{C}$ be the set of locations with the top M highest total rewards, and then we have

$$\begin{aligned}
\ln \frac{W_{T+1}}{W_1} &\ge \ln \frac{\sum_{k\in A^*} w_k(T+1)}{W_1} \\
&\ge \frac{\sum_{k\in A^*} \ln w_k(T+1)}{M} - \ln \frac{m}{M} \\
&= \ln(1+\theta) \sum_{k\in A^*} \sum_{t:k\in\mathcal{C}_0^c(t)} x_k(t) - \ln \frac{m}{M} \qquad (8)
\end{aligned}$$

where the second inequality is due to the inequality of arithmetic and geometric means, $\frac{1}{M}\sum_{j=1}^M a_j \ge \left(\prod_{j=1}^M a_j\right)^{\frac{1}{M}}$. Note that

$$\begin{aligned}
\sum_{k\in A^*} \sum_{t:k\in\mathcal{C}_0(t)} x_k(t) &\le \sum_{t=1} \sum_{k\in\mathcal{C}_0(t)} x_k(t) \\
&= \sum_{t=1} \sum_{k\in\mathcal{C}_0(t)} \alpha_k(t) x_k(t). \qquad (9)
\end{aligned}$$

Combining (7) (8) and (9), we obtain

$$\begin{aligned}
\mathbb{E}G_{\text{Hedge-M}} &= \sum_{t=1}^T \sum_{k\in\mathcal{C}} \alpha_k(t) x_k(t) \\
&\ge \frac{\ln(1+\theta)}{\theta} \sum_{k\in A^*} \sum_{t=1}^T x_k(t) - \frac{\ln(m/M)}{\theta} \\
&= \frac{\ln(1+\theta)}{\theta} G_{\max} - \frac{\ln(m/M)}{\theta} \\
&\ge G_{\max} - \frac{\theta}{2} G_{\max} - \frac{\ln(m/M)}{\theta} \ge G_{\max} - \sqrt{2MT \ln \frac{m}{M}}
\end{aligned}$$

when $\theta = \sqrt{2\ln(m/M)/(MT)}$, where the third inequality is due to $\ln(1+x) \ge x(1-x/2)$, and the last inequality is due to the fact that $G_{\max} \le MT$. ■

Proof of Theorem 5: Note that for an optimal policy of the evader, any location in $\mathcal{C}_0(t)$ is never selected. Consider the set Γ_0 of policies that never choose from $\mathcal{C}_0(t)$ at each time. For any $\gamma \in \Gamma_0$, we have $w'_{\pi(t)}(t) = w_{\pi(t)}(t)$. Let

$$\beta_k(t) := M \frac{w_k(t)}{\sum_{j=1}^m w_j(t)}$$

and let $\tilde{r}^\pi(t) := 1 - \beta_{\pi(t)}(t)$. We then have

$$\begin{aligned}
r^\pi(t) = 1 - \alpha_{\pi(t)}(t) &= 1 - M \frac{w_k(t)}{\sum_{j=1}^m w'_j(t)} \\
&\le 1 - M \frac{w_k(t)}{\sum_{j=1}^m w_j(t)} = 1 - \beta_{\pi(t)}(t) = \tilde{r}^\pi(t).
\end{aligned}$$

Hence, $\sum_{t=1}^T r^\pi(t) \le \sum_{t=1}^T \tilde{r}^\pi(t)$ for any $\gamma \in \Gamma_0$.

Consider now the finite-horizon reward maximization problem with the reward function given by $\tilde{r}$ within the policy space Γ_0. Note also that for any $\gamma \in \Gamma_0$, the exponent of $w_k(t)$ always is given by $\sum_{s=1}^t x_k(s)$. Using then the same argument as the single-play case, it can be shown that the greedy policy maximizes $\mathbb{E}\{\sum_{t=1}^T \tilde{r}^\pi(t)\}$. On the other hand, $\sum_{t=1}^T r^\pi(t) = \sum_{t=1}^T \tilde{r}^\pi(t)$ for the greedy policy since $\mathcal{C}_0(t) = \emptyset$ for all t. [7] Therefore, the greedy policy is also optimal for the original finite-horizon reward maximization problem. ■

Proof of Lemma 2: Given any initial condition $(G_k(0))_{k\in\mathcal{C}}$, we can relabel locations so that $1 \in \arg\max_{k\in\mathcal{C}} G_k(0)$. Since the choice of the decoy at T does not affect the reward of the evader, we assume it always selects from $\mathcal{L}(T)$ for simplicity. We then prove by induction. For $T = 1$, the claim is clearly true. Assume that the claim holds for $T = 1, 2, \ldots, t'$. For $T = t'+1$. At the first time step, suppose that using an optimal policy the decoy node selects some location i such that $G_i(0) < G_1(0)$, and the evader selects location j. If $G_j(0) > G_i(0)$, we can always swap the choice of the decoy and the evader to obtain a higher reward of the evader, and hence $G_j(0) \le G_i(0)$. Thus, $1 \in \arg\max_{k\in\mathcal{C}} G_k(1)$. Then, the rest t' steps until reaching the horizon can be thought as using Hedge with the initial condition $(G_k(1))_{k\in\mathcal{C}}$. Hence, by the induction hypothesis, the

[7] To be rigorous, this holds when $T \ge 2$. To avoid triviality, we assume so in this paper.

decoy always selects a location from $\mathcal{L}(t)$ from $t=2$. It can be easily seen that some location in $\mathcal{L}(t)$ is then always selected by the decoy until the horizon. Without loss of generality, we assume that the decoy always selects location 1. We also denote the location chosen by the evader at time t by k_t. Set $d_{ij}(t) := G_i(t-1) - G_j(t-1)$ for this optimal policy. At each time $t>1$, we have

$$r(t) = \frac{\sum_{l\neq k_t,1,i} a^{d_{lk_t}(t)} + a^{d_{1k_t}(t)} + a^{d_{ik_t}(t)}}{1+\sum_{l\neq k_t,1,i} a^{d_{lk_t}(t)} + a^{d_{1k_t}(t)} + a^{d_{ik_t}(t)}}.$$

Consider now a change of policy by letting the decoy select location 1 at the first step, and keeping the choice of the evader unchanged. The reward of the evader at each time $t>1$ becomes

$$\begin{aligned} r'(t) &= \frac{\sum_{l\neq k_t,1,i} a^{d_{lk_t}(t)} + a^{d_{1k_t}(t)+1} + a^{d_{ik_t}(t)-1}}{1+\sum_{l\neq k_t,1,i} a^{d_{lk_t}(t)} + a^{d_{1k_t}(t)+1} + a^{d_{ik_t}(t)-1}} \\ &> r(t), \end{aligned}$$

since $d_{1k_t}(t) \geq d_{ik_t}(t)$ for all t and $a>1$, which is a contradiction of the optimality, and the proof is then complete. ∎

Proof of Lemma 3: The proof is similar to that of Theorem 3, and we use the same notation without repeated definition whenever there is no ambiguity. Consider the case (C1), and as in the proof of Theorem 3 we assume without loss of generality that $1\in\mathcal{K}(t_0)$ is chosen at t_1 and the location selected by the evader at t_1-1 is 2. Furthermore, suppose that the decoy node selects channel 3 at t_1-1, and hence the decoy node also selects channel 3 at t_1. Consider a change of policy of the user by selecting channel 1 at t_1-1 and channel 2 at t_1. We then have

$$\begin{aligned} &r'(t_1-1)+r'(t_1)-r(t_1-1)-r(t_1) \\ &= \frac{\sum_{k\geq 4} a^{d_{k1}} + a^{d_{21}} + a^{d_{31}}}{1+\sum_{k\geq 4} a^{d_{k1}} + a^{d_{21}} + a^{d_{31}}} + \\ &\quad + \frac{\sum_{k\geq 4} a^{d_{k2}} + a^{d_{12}+1} + a^{d_{32}+1}}{1+\sum_{k\geq 4} a^{d_{k2}} + a^{d_{12}+1} + a^{d_{32}+1}} - \\ &\quad - \frac{\sum_{k\geq 4} a^{d_{k2}} + a^{d_{12}} + a^{d_{32}}}{1+\sum_{k\geq 4} a^{d_{k2}} + a^{d_{12}} + a^{d_{32}}} - \\ &\quad - \frac{\sum_{k\geq 4} a^{d_{k1}} + a^{d_{21}+1} + a^{d_{31}+1}}{1+\sum_{k\geq 4} a^{d_{k1}} + a^{d_{21}+1} + a^{d_{31}+1}} \\ &= \frac{1}{1+C+a^{d_{21}+1}+a^{d_{31}+1}} + \frac{1}{1+D+a^{d_{12}}+a^{d_{32}}} - \\ &\quad - \frac{1}{1+C+a^{d_{21}}+a^{d_{31}}} - \frac{1}{1+D+a^{d_{12}+1}+a^{d_{32}+1}}, \end{aligned}$$

where $C=\sum_{k\geq 4} a^{d_{k1}}$ and $D=\sum_{k\geq 4} a^{d_{k2}}$. Set $d=d_{12}$ and $d'=d_{31}$, and we obtain

$$\begin{aligned} &r'(t_1-1)+r'(t_1)-r(t_1-1)-r(t_1) \\ &= \frac{1}{1+Da^d+a^{d+1}+a^{d'+1}} + \frac{1}{1+D+a^{-d}+a^{d'-d}} - \\ &\quad - \frac{1}{1+Da^d+a^d+a^{d'}} - \frac{1}{1+D+a^{-d+1}+a^{d'-d+1}}, \\ &= ((a^{2d-1}-a^{d-1})(a-1)^2 + (a^{d'}-a^{d'-1})(Da^{2d}+ \\ &\quad + a^{2d+1}+a^{d'+d+1}-Da^d-a-a^{d'+1}))/\mathrm{Den} > 0, \end{aligned}$$

where $\mathrm{Den} = (1+Da^d+a^{d+1}+a^{d'+1})(1+Da^d+a^d+a^{d'})(1+Da^{d-1}+a^{d-1}+a^{d'})$. For (C2), it is clear that alternatively selecting channel results in a higher reward for the user. With the same conclusion as the proof of Theorem 3, the result follows. ∎

Proof of Lemma 4: 1) Note that problem (3) is equivalent to

$$\max_{a\in\Delta_m} \min_{b\in\Delta_m} \quad (\sum_{k=1}^{m} a_k\mu_k - a_j\mu_j)b_j, \tag{10}$$

or compactly,

$$\max_{a\in\Delta_m} \min_{b\in\Delta_m} \quad a^\top H b, \tag{11}$$

where $^\top$ denotes the transpose, and

$$H = \begin{bmatrix} 0 & \mu_1 & \cdots & \mu_1 \\ \mu_2 & 0 & \cdots & \mu_2 \\ \vdots & & \ddots & \vdots \\ \mu_m & \mu_m & \cdots & 0 \end{bmatrix}.$$

Consider now a zero-sum game with the payoff matrices for the row and the column players being H and $-H$, who choose a and b, respectively. Any optimal solution $a^\star$ to problem (3) is a Nash equilibrium strategy for the row player, and by the indifference condition, we obtain for any $j\in\mathrm{supp}(a^\star)$,

$$\sum_{\substack{k\neq j \\ k\in\mathrm{supp}(a^\star)}} a_k^\star\mu_k = \mathrm{Const.},$$

which implies $a_k^\star\mu_k = a_j^\star\mu_j$ for any $k,j\in\mathrm{supp}(a^\star)$.

2) Assume for contradiction that there exist $i\in\mathrm{supp}(a^\star)$ and $j\in\mathcal{C}\setminus\mathrm{supp}(a^\star)$ such that $\mu_j>\mu_i$. Let c be the constant such that $c=a_k^\star\mu_k$ for any $k\in\mathrm{supp}(a^\star)$. Consider then a feasible solution a, where $a_k=0$ for all $k\in((\mathcal{C}\setminus\mathrm{supp}(a^\star))\setminus\{j\})\cup\{i\}$, and $a_k=c+\epsilon\mu_k$ for all $k\in(\mathrm{supp}(a^\star)\setminus\{i\})\cup\{j\}$, with $\epsilon=a_i^\star(1-\mu_i/\mu_j)/K^\star$, which yields a higher objective value. ∎

Appendix B
The Dependent Rounding Algorithm

Dependent Rounding
Input: A marginal distribution $(\alpha_k, k\in\mathcal{C})$ and a natural number $M<|\mathcal{C}|$ such that $\sum_{k\in\mathcal{C}}\alpha_k = M$.
Output: A subset $\mathcal{C}_1$ of $\mathcal{C}$ such that $|\mathcal{C}_1|=M$.
Initialization: $p_k=\alpha_k$ for all $k\in\mathcal{C}$.
While $\{k\in\mathcal{C}: 0<p_k<1\}\neq\emptyset$ **do**

1) Choose distinct i and j with $0<p_i<1$ and $0<p_j<1$.
2) Set $a=\min\{1-p_i,p_j\}$ and $b=\min\{p_i,1-p_j\}$.
3) Update p_i and p_j as

$$(p_i,p_j) = \begin{cases} (p_i+a, p_j-a), & \text{w.p. } \frac{b}{a+b} \\ (p_i-b, p_j+b), & \text{w.p. } \frac{a}{a+b} \end{cases}$$

Return $\{k\in\mathcal{C}: p_k=1\}$.

REFERENCES

[1] Q. Wang and M. LIu, "Learning in Hide-and-Seek," in *INFOCOM '14*, 2014.

[2] R. Vidal, O. Shakernia, H. Kim, D. Shim, and S. Sastry, "Probabilistic Pursuit-Evasion Games: Theory, Implementation, and Experimental Evaluation," *Robotics and Automation, IEEE Transactions on*, vol. 18, no. 5, pp. 662–669, 2002.

[3] V. Navda, A. Bohra, S. Ganguly, and D. Rubenstein, "Using Channel Hopping to Increase 802.11 Resilience to Jamming Attacks," in *INFOCOM '07, Mini-Conference*, 2007, pp. 2526–2530.

[4] D. Matula, "A Periodic Optimal Search," *The American Mathematical Monthly*, vol. 71, no. 1, pp. 15–21, 1964.

[5] W. Black, "Discrete Sequential Search," *Information and Control*, vol. 8, pp. 159–162, 1965.

[6] J. Milton C. Chew, "A Sequential Search Procedure," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 494–502, 1967.

[7] R. Ahlswede and I. Wegener, *Search Problems*. John Wiley & Sons, 1987.

[8] D. Assaf and S. Zamir, "Optimal Sequential Search: A Bayesian Approach," *The Annals of Statistics*, vol. 13, no. 3, pp. 1213–1221, 1985.

[9] F. Kelly, "On Optimal Search with Unknown Detection Probabilities," *Journal of Mathematical Analysis and Applications*, vol. 88, no. 2, pp. 422–432, 1982.

[10] S. M. Pollock, "A Simple Model of Search for a Moving Target," *Operations Research*, vol. 18, no. 5, pp. 883–903, 1970.

[11] R. R. Weber, "Optimal Search for a Randomly Moving Object," *Journal of Applied Probability*, vol. 23, no. 3, pp. 708–717, 1986.

[12] R. Isaacs, *Differential Games*. Wiley, 1965.

[13] J. D. Grote, Ed., *The Theory and Application of Differential Games*. D. Reidel Publishing Company, 1975.

[14] Y. Yavin and M. Pachter, Eds., *Pursuit-Evasion Differential Games*. Pergamon Press, 1987.

[15] T. Başar and G. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed. Society for Industrial and Applied Mathematics, 1998.

[16] W. Xu, W. Trappe, Y. Zhang, and T. Wood, "The Feasibility of Launching and Detecting Jamming Attacks in Wireless Networks," in *MobiHoc '05*, 2005, pp. 46–57.

[17] A. Wood, J. Stankovic, and G. Zhou, "DEEJAM: Defeating Energy-Efficient Jamming in IEEE 802.15.4-based Wireless Networks," in *SECON '07*, 2007, pp. 60 –69.

[18] G. Noubir and G. Lin, "Low-power DoS Attacks in Data Wireless LANs and Countermeasures," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 7, no. 3, pp. 29–30, 2003.

[19] E. Kehdi and B. Li, "Null Keys: Limiting Malicious Attacks Via Null Space Properties of Network Coding," in *INFOCOM '09*, april 2009, pp. 1224 –1232.

[20] J. Chiang and Y.-C. Hu, "Cross-Layer Jamming Detection and Mitigation in Wireless Broadcast Networks," *Networking, IEEE/ACM Transactions on*, vol. 19, no. 1, pp. 286–298, 2011.

[21] C. Popper, M. Strasser, and S. Capkun, "Anti-jamming Broadcast Communication Using Uncoordinated Spread Spectrum Techniques," *Selected Areas in Communications, IEEE Journal on*, vol. 28, no. 5, pp. 703–715, 2010.

[22] G. Noubir, R. Rajaraman, B. Sheng, and B. Thapa, "On the Robustness of IEEE 802.11 Rate Adaptation Algorithms Against Smart Jamming," in *WiSec '11*, ser. WiSec '11. New York, NY, USA: ACM, 2011, pp. 97–108.

[23] A. Sampath, H. Dai, H. Zheng, and B. Zhao, "Multi-channel Jamming Attacks using Cognitive Radios," in *ICCCN '07*, 2007, pp. 352–357.

[24] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "Gambling in a Rigged Casino: The Adversarial Multi-armed Bandit Problem," in *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, 1995, pp. 322–331.

[25] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The Non-stochastic Multiarmed Bandit Problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2003.

[26] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997.

[27] N. Littlestone and M. K. Warmuth, "The Weighted Majority Algorithm," *Information and Computation*, vol. 108, no. 2, pp. 212–261, 1994.

[28] S. Arora, E. Hazan, and S. Kale, "The Multiplicative Weights Update Method: a Meta-Algorithm and Applications," *Theory of Computing*, vol. 8, no. 6, pp. 121–164, 2012.

[29] J. Xu, Q. Wang, R. Jin, K. Zeng, and M. Liu, "Secondary User Data Capturing for Cognitive Radio Network Forensics under Capturing Uncertainty," in *MILCOM, to appear*, 2014.

[30] P. Joulani, A. György, and C. Szepesvári, "Online Learning under Delayed Feedback," in *International Conference on Machine Learning (ICML-2013)*, 2013.

[31] T. Uchiya, A. Nakamura, and M. Kudo, "Algorithms for Adversarial Bandit Problems with Multiple Plays," in *Proceedings of the 21st international conference on Algorithmic learning theory*. Springer-Verlag, 2010, pp. 375–389.

[32] R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan, "Dependent Rounding and its Applications to Approximation Algorithms," *J. ACM*, vol. 53, no. 3, pp. 324–360, 2006.

[33] E. T. Noam Nisan, Tim Roughgarden and V. V. Vazirani, Eds., *Algorithmic Game Theory*. Cambridge University Press, 2007.

[34] N. Theis, R. Thomas, and L. DaSilva, "Rendezvous for Cognitive Radios," *Mobile Computing, IEEE Transactions on*, vol. 10, no. 2, pp. 216–227, Feb 2011.

[35] N. Roy and G. Dudek, "Collaborative Robot Exploration and Rendezvous: Algorithms, Performance Bounds and Observations," *Auton. Robots*, vol. 11, no. 2, pp. 117–136, Sep. 2001.

[36] Q. Wang and M. Liu, "Throughput Optimal Switching in Multi-channel WLANs," *Mobile Computing, IEEE Transactions on, to appear*, 2012.

[37] C. Daskalakis, R. Frongillo, C. Papadimitriou, G. Pierrakos, and G. Valiant, "On Learning Algorithms for Nash Equilibria," in *Algorithmic Game Theory*, ser. Lecture Notes in Computer Science, S. Kontogiannis, E. Koutsoupias, and P. Spirakis, Eds. Springer Berlin Heidelberg, 2010, vol. 6386, pp. 114–125.

[38] H. P. Young, *Strategic Learning and its Limits*. Oxford University Press, 2004.

[39] A. Blum and Y. Mansour, "From External to Internal Regret," *The Journal of Machine Learning Research*, vol. 8, pp. 1307–1324, 2007.

[40] R. Kleinberg, G. Piliouras, and E. Tardos, "Multiplicative Updates Outperform Generic No-regret Learning in Congestion Games: Extended Abstract," in *STOC '09*, 2009.

[41] G. Kasbekar and A. Proutiere, "Opportunistic Medium Access in Multi-channel Wireless Systems: A Learning Approach," in *Allerton '10*, 2010, pp. 1288–1294.

PLACE PHOTO HERE

Qingsi Wang (S'13, M'14) received his B.E. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009, and the M.S. degrees in electrical engineering: systems and applied mathematics and the Ph.D. degree in electrical engineering: systems from the University of Michigan, Ann Arbor, Michigan, in 2011, 2014 and 2014, respectively. His research interests are broadly in performance modeling and analysis, and optimal resource allocation in networked systems, with the application of stochastic control, game theory and learning theory. He is currently affiliated with Qualcomm Research, San Diego, California.

PLACE PHOTO HERE

Mingyan Liu (M'00, SM'11, F'14) received her Ph.D. Degree in electrical engineering from the University of Maryland, College Park, in 2000, and has since been with the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, where she is currently a Professor. Her research interests are in optimal resource allocation, sequential decision theory, incentive design, and performance modeling and analysis, all within the context of communication networks. Her most recent research activities involve online learning, modeling and mining of large scale Internet measurement data concerning cyber security, and incentive mechanisms for inter-dependent security games.