Learning in Hide-and-Seek

Qingsi Wang and Mingyan Liu
EECS Department, University of Michigan, Ann Arbor
{qingsi, mingyan}@umich.edu

***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 may lack robustness against changes in the adversarial behavior, while those from the latter are often difficult to justify due to the implied full information (either as realizations or as distributions) and rationality, both of which may be limited in practice. 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, hiding/seeking in the location that the opponent is the least/most likely to appear. 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 completely knows the adaptive pattern used by the adversary and responses optimally.**

I. Introduction

The pursuit-evasion (or hide-and-seek) problem models a variety of applications and has been extensively studied. For instance, it can be used to model the pursuit of a moving target by a radar or an unmanned vehicle [1], or a radio performing channel switching in an attempt to hide from a jammer [2].

Existing work typically falls into two categories. The first considers stationary or heuristic behavior of one side and examines corresponding countermeasures of the other. Examples include [3], [4], [5], [6] and the references therein, that 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., [7] that uses a random prior probability of hiding in a given location, and [8] 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., [9], where the evader's motion is given by a discrete-time Markov chain independent of the pursuer's activity, and [10] 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 [11] to capture the continuous evolution; in fact, the pursuit-evasion problem bears the genesis of differential games. See also [12], [13], [14] for texts and examples of differential games and their application in the pursuit-evasion problem. We note that results from the first category may lack robustness against changes in the adversarial behavior, while those from the second category are often difficult to justify due to the implied full information (either as realizations or as distributions) and rationality, both of which may be limited in practice.

The work is partially supported by the NSF under grants CIF-0910765 and CNS-1217689.

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 problems. In other words we assume the pursuer is able to adapt over time using its 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 will employ online learning algorithms developed for the class of adversarial or non-stochastic multi-armed bandit problems [15], [16], 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, there is no loss for either side compared to the case when it knows the adaptive pattern of the adversary and responses optimally.

The remainder of the paper is organized as follows. Section II describes the system model and the problem formulation, followed by the two cases in Sections III, IV and Section V, respectively. Section VI concludes the paper. All 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 behavior 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 kth location is chosen by the evader and the pursuer at time t; we will also call $\tau(t)$ and $\alpha(t)$ the 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 possessed by one side about the likelihood of an action by the other side.

The evader's objective is to maximize its total number of successful evasion, while the pursuer aims to maximize its total number of successful pursuit. Within this context we investigate two cases. 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 $P(z_k(t) = 0) = \alpha_k(t)$ (resp. $P(x_k(t) = 1) = \tau_k(t)$), though the value of this probability may be unknown to the evader (resp. the pursuer). Accordingly, if the evader knows the behavior pattern of the pursuer, the expected utility it derives from using location k, denoted by U_k, is given by $U_k(t) = 1 - \alpha_k(t)$. Symmetrically, if the pursuer is the side with such knowledge, its expected utility from searching location k, denoted by V_k, is given by $V_k(t) = \tau_k(t)$. Note that U_k and V_k are essentially the average numbers of successful evasion and pursuit at this location if chosen.

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. Accordingly, the evader's utility of choosing location k in this case is given by $U_k(t) = z_k(t)$, while for the pursuer's utility, $V_k(t) = x_k(t)$.

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 slot t. $\alpha(t)$ is a vector of probability distribution and 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 $\gamma = (\gamma_t, t \geq 1)$ 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 $r^\pi(t) = U_{\pi(t)}(t) = 1 - \alpha_{\pi(t)}(t)$ at time t, and 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 w.r.t. the randomness of system states and the private randomization device.

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

C. Formulation: against unknown adversarial behavior

In Section V, we consider the more practical scenario where neither side has the 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, and 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 also define the control policies γ and λ, and the policy spaces Γ and Λ in parallel. The evader receives a reward $r^\pi(t) = U_{\pi(t)}(t) = z_{\pi(t)}(t)$ at each time t; the pursuer receives $b^\xi(t) = V_{\xi(t)}(t) = x_{\xi(t)}(t)$. Both sides consider the similar reward maximization problems as in the previous case, with the expectation in the objective taken w.r.t. the randomness of private randomization devices.

Note that the objectives typically cannot be directly evaluated by either side in this case, given the unknown and non-stochastic nature of the opponent. The optimal control in this setting is typically addressed in the framework of non-stochastic online learning, where the existing literature focuses on minimizing the (weak) regret of a strategy compared to a best single-action strategy. These online learning techniques are employed as our main model for the adaptive behavior of either side.

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 slot and cannot change their mind till the next slot. Both sides also receive feedback by the end of a slot: 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 introduced by Auer et al. [15]; this is a variant of the original Hedge algorithm introduced by Freund and Schapire [17], within the line of work on multiplicative weights learning [18] (see [19] for an in-depth survey and references therein). Hedge is an online learning algorithm in the adversarial multi-arm bandit setting [15], [16], which presumes no probabilistic behavior of the opponent (in our case, the evader). It is shown to guarantee an order-optimal sublinear weak regret, which in our context translates into sublinear "missing" of discovery opportunities compared to always searching the (hindsight) most active/used location 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

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

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

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

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 $\alpha(t) = (\alpha_1(t), \alpha_t(t), \ldots, \alpha_m(t))$ 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 [15].

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[1]. 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 will 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 $g(t) \in \arg\min_{k\in C} \alpha_k^g(t)$ 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.

[1]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 policy derived here.

Proof: Note that

$$\mathbb{E}G^{\pi}_{\text{Hedge}}(T) = \mathbb{E}\left\{\sum_{t=1}^{T} x^{\pi}_{\xi(t)}(t)\right\} = \sum_{t=1}^{T}\sum_{k=1}^{m} x^{\pi}_{k}(t)\alpha^{\pi}_{k}(t)$$
$$= \sum_{t=1}^{T} \alpha^{\pi}_{\pi(t)}(t) = T - \sum_{t=1}^{T} r^{\pi}(t)$$

for any realization of π. Therefore,

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

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. ∎

Without loss of generality, we will assume under the greedy policy ties are broken in favor of the lowest-indexed location. Note that since γ_{greedy} 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 $1, 2, \cdots, m$. The above proof also suggests that any policy that results in an equal frequency of presence on each location has the same infinite-horizon average reward, thus asymptotically optimal. It should be noted that these equi-occupancy polices are not necessarily optimal for the finite-horizon problem posed in (1) as we elaborate at the end of this subsection. The greedy policy, however, is in fact also optimal over the finite horizon. Below we prove this result 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 $\pi(t) = k$ if $\alpha_k(t) < 1/2$, $k = 1, 2$, and $\pi(t)$ can be either 1 or 2 when $\alpha_1(t) = \alpha_2(t) = 1/2$.

Proof: For any policy, let $\Delta(t) := |G_1(t) - G_2(t)|$; this is the difference between the number of times locations 1 and 2 have been used by the end of slot 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

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

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. ∎

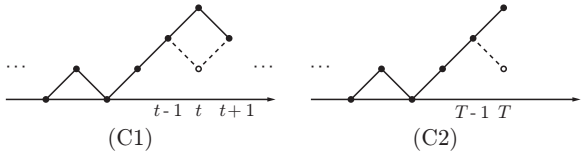

Fig. 1. The change of policy in two cases.

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

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

$$\alpha^{\pi}_{k}(t+1) = \frac{\alpha^{\pi}_{k}(t)a^{\mathbf{1}_{\pi(t)=k}}}{\sum_{j\in\mathcal{C}} \alpha^{\pi}_{j}(t)a^{\mathbf{1}_{\pi(t)=j}}},$$

with $\mathbf{1}_{\{\cdot\}}$ being the indicator function. It is therefore only necessary for the evader to recall/store the last control action and the last system state. 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.

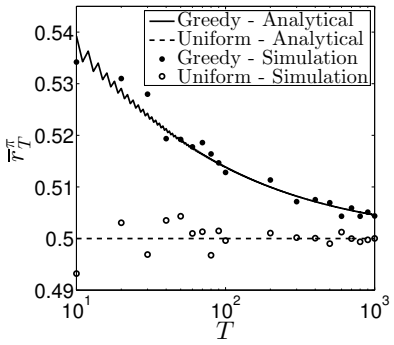

Fig. 2. The finite-horizon (expected) average reward of the greed policy and the uniform policy in a two-location example.

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 [2].

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 $\max_{k\in\mathcal{C}} \frac{w_k(t)}{\sum_{j=1}^m w_j(t)} > \frac{1}{M}$, then 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 $\alpha(t) = (\alpha_1(t), \alpha_t(t), \ldots, \alpha_m(t))$ 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

$$w_k(t+1) = \begin{cases} w_k(t), & k \in \mathcal{C}_0(t) \\ w_k(t) a^{x_k(t)}, & k \in \mathcal{C} \setminus \mathcal{C}_0(t) \end{cases}.$$

---

[2]Hedge-M is reverse-engineered from the algorithm Exp3.M [20], which is a multiple-play algorithm with partial information (the pursuer only observes activities in locations it searched).

The subroutine Dependent Rounding [21] 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 (hindsight) M most evader-active locations; the proof is based on that of Hedge [15] and Exp3.M [20].

Theorem 4: If $a = 1 + \sqrt{2\ln(m/M)/(MT)}$, then $\mathbb{E}G_{\text{Hedge-M}}(T) \geq G_{\max}(T) - \sqrt{2\ln(m/M)MT}$, the expectation over 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)[3]. 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,

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

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

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 slot (referred to as the doubly greedy (G2) policy), we have

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

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.

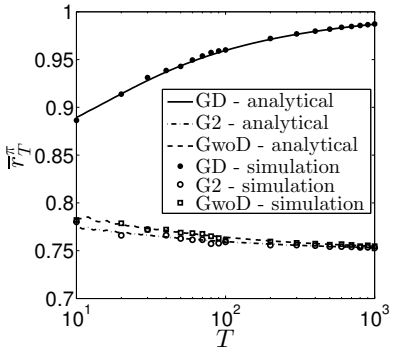

Fig. 3. The finite-horizon average reward of the greedy decoy (GD) policy, the doubly greedy (G2) policy, and the greedy policy without the decoy (GwoD) in a system of four locations.

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. The proof of Lemma 3 is similar to that of Theorem 3 and is thus omitted for brevity.

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 slot.

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[4]. 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 similarly. 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

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

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

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

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

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

[4]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.

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

Theorem 7: The greedy policy is optimal for the pursuer for 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 on the reasoning used by the opponent, and accordingly employ their respective learning techniques.

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. We have seen from the weak regret results that $\frac{1}{T}\mathbb{E}R_{\text{Hedge}} \geq \frac{m-M}{m}+o(1)$ and $\frac{1}{T}\mathbb{E}G_{\text{Hedge-M}} \geq \frac{M}{m}+o(1)$ when the pursuer can search M locations simultaneously and the evader hides in one location, where the $o(1)$ terms are w.r.t. the growth of T. Hence,

$$\overline{r}(\text{Hedge};\lambda) \geq \frac{m-M}{m}$$

when the evader uses Hedge and the pursuer uses a policy λ, and

$$\overline{b}(\text{Hedge-M};\gamma) \geq \frac{M}{m},$$

when the pursuer uses Hedge-M and the evader uses a policy γ, where we explicitly denote the average reward as a function of a chosen pair of policies. Note that $\overline{r}(\gamma;\lambda)+\overline{b}(\lambda;\gamma)=1$ for any γ and λ. Therefore, the above inequalities become equalities when Hedge and Hedge-M are respectively used. That is, Hedge and Hedge-M are *mutually best responses* for the infinite-horizon problem, and up to a diminishing term over a finite horizon. 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 happens to use a similar algorithm.

Moreover, the above conclusion also 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, and for this reason it can no longer use Hedge. In this case a partial information counterpart Exp3 [15], [16] can be used to update its probability $\tau_k(t)$ of choosing location k in at t. Following the same line of argument, we can show that Exp3 and Hedge-M are also mutually best responses. As might have been realized, the mutual optimality is the result of the sublinear-regret performance of these non-stochastic online learning algorithms. For our hide-and-seek problem, the mutual optimality holds for *any* pair of sublinear-regret algorithms.

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 there is a single most evader-active location, then the pursuer can guarantee sublinear weak regret if and only if all suboptimal locations are searched with time sublinear in T. In other words, a strategy that guarantees sublinear weak regret for the pursuer must ultimately identify and aim for the most evader-active location. Therefore, the evader can always use the decoy to "create" this most evader-active 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 sublinear-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. 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 [22]. The learning perspective thus provides a different and possibly more natural angle to interpret 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.

References

[1] 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.

[2] 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*, pp. 2526–2530, 2007.

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

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

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

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

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

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

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

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

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

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

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

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

[15] 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*, pp. 322–331, 1995.

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

[17] 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.

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

[19] 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.

[20] 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*, pp. 375–389, Springer-Verlag, 2010.

[21] 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.

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

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

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

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 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 $\mathcal{C}\backslash\mathcal{C}_0(t)$ by $\mathcal{C}_0^c$. Then, for any $t \leq T$,

$$\frac{W_{t+1}}{W_t} = \sum_{k\in\mathcal{C}_0^c(t)} \frac{w_k(t+1)}{W_t} + \sum_{k\in\mathcal{C}_0(t)} \frac{w_k(t+1)}{W_t}$$

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

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

$$
\begin{aligned}
\ln \frac{W_{T+1}}{W_1} = \sum_{t=1}^{T} \ln \frac{W_{t+1}}{W_t} &\le \sum_{t=1}^{T} \ln \left(1+\theta \sum_{k\in\mathcal{C}_0^c(t)} \alpha_k(t) x_k(t) \right) \\
&\le \theta \sum_{t=1}^{T} \sum_{k\in\mathcal{C}_0^c(t)} \alpha_k(t) x_k(t) \qquad (3)
\end{aligned}
$$

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

$$
\begin{aligned}
\ln \frac{W_{T+1}}{W_1} &\ge \ln \frac{\sum_{k\in A^*} w_k(T+1)}{W_1} \\
&\ge \frac{\sum_{k\in A^*} \ln w_k(T+1)}{M} - \ln \frac{m}{M} \\
&= \ln(1+\theta) \sum_{k\in A^*} \sum_{t:k\in\mathcal{C}_0^c(t)} x_k(t) - \ln \frac{m}{M} \qquad (4)
\end{aligned}
$$

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

$$
\begin{aligned}
\sum_{k\in A^*} \sum_{t:k\in\mathcal{C}_0(t)} x_k(t) &\le \sum_{t=1} \sum_{k\in\mathcal{C}_0(t)} x_k(t) \\
&= \sum_{t=1} \sum_{k\in\mathcal{C}_0(t)} \alpha_k(t) x_k(t). \qquad (5)
\end{aligned}
$$

Combining (3) (4) and (5), we obtain

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

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 Lemma 2: Given any initial condition $(G_k(0))_{k\in\mathcal{C}}$, we can relabel locations such 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 slot, 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 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 slot, 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 $d_{1k_t}(t) \ge d_{ik_t}(t)$ for all t and $a > 1$, which is a contradiction of the optimality, and the proof is then complete. ∎

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