An Online Learning Framework for UAV Target Search Missions in Non-Stationary Environments

Khial, Noor; Mhaisen, Naram; Mabrok, Mohamed; Mohamed, Amr

DOI
10.1109/CCECE59415.2024.10667171

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the 2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)

2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) | 979-8-3503-7162-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CCECE59415.2024.10667171

An Online Learning Framework for UAV Target Search Missions in Non-Stationary Environments

Noor Khial[1], Naram Mhaisen[2], Mohamed Mabrok[3], Amr Mohamed[1]
[1] College of Engineering, Qatar University, Qatar
[2] College of Electrical Engineering, Mathematics, and Computer Science, TU Delft, Netherlands
[3] College of Arts and Sciences, Qatar University, Qatar
Email: {noor.khial, m.a.mabrok, amrm}@qu.edu.qa, N.Mhaisen@tudelft.nl

***Abstract*—The rapid evolution of Unmanned Aerial Vehicles (UAVs) has revolutionized target search operations in various fields, including military applications, search and rescue missions, and post-disaster management. In this paper, we propose the use of a multi-armed bandit algorithm for a UAV's search mission in an unknown and adversarial setting. The UAV's objective is to locate a mobile target formation, assuming that their mobility resembles an adversarial behavior. To achieve this, we formulate an optimization problem and leverage the Exp3 (exponential-weighted exploration and exploitation) algorithm to solve it. The targets are assumed to be moving under the assumption of an unknown and potentially non-stationary probability distribution. To enhance the learning process, we integrate environmental observations as contextual information, resulting in a variant called C-Exp3, which optimizes the search process. Finally, we evaluate the performance of C-Exp3 in UAV search missions, focusing on adversarial environments. The primary objective for the UAV is to converge towards an optimal policy as time t approaches the horizon $\mathcal{T}$, reflecting the UAV's capacity to learn the formation's strategy.**

***Index Terms*—UAV, Search Mission, Online Learning, Multi-Armed Bandits.**

I. Introduction

Unmanned aerial vehicles (UAVs) have become essential tools in various civilian sectors, such as military applications, post-disaster wireless service restoration, and search and rescue (SAR) operations [1], [2]. Equipped with state-of-the-art sensors and advanced imaging technologies, UAVs provide precise and real-time information regarding the locations and conditions of both individuals and infrastructure [3]. Search and rescue missions, as well as surveillance tasks, often entail the challenge of finding targets in vast areas. In military operations, this can involve finding enemy positions or tracking potential threats. During these operations, sensors and imaging technologies are systematically employed to explore unknown areas and allocate targets. UAVs strategically determine flight paths to observe and locate these targets. The primary objective is to identify the formation of targets while learning about their strategies. Consequently, the UAV dynamically fine-tunes its flight trajectory based on real-time observations, optimizing coverage, and adapting to changes influenced by targets that modify their strategies.

In [4]–[6], new techniques were introduced that utilize reinforcement learning (RL) to optimize the search process executed by UAVs. RL agents gain the ability to make decisions (actions) based on their present state and the anticipated outcomes (rewards) of those actions. Notably, RL necessitates certain assumptions concerning targets probability distribution, particularly in the context of Markov Decision Processes (MDPs) and stochastic environments [7]. In these scenarios, RL algorithms model the agent-environment interaction as a Markovian process. This means that the agent's actions and the environment's responses follow underlying probability distributions, which are assumed to be stationary. However, it's crucial to acknowledge that many real-world situations encompass non-stationary environments, where the parameters of the MDP may vary over time. In such instances, learning algorithms must adapt to this non-stationary.

We focus on the investigation of the location of dynamic targets in environments characterized by uncertainty, where no prior knowledge is available, and assumptions about the probability distribution governing target formations are not made. Our approach involves the deployment of a single agent. In this paper, we focus on exploring online learning (OL) algorithms, which are known to provide performance guarantees. Notably, our approach stands apart from RL methods, as we do not incorporate stochastic assumptions regarding the probability distributions of target formations. This paper highlights an unexplored aspect by considering online learning as a potential solution for UAV searching mission, specifically addressing challenges arising from adversarial environments, in which the agent might operate. These environments involve interactions with targets showing varying levels of randomness in their mobility patterns, often including intentional efforts to mislead the UAV, especially in military applications. Consequently, this situation adds complexity to the search mission. We address these challenges by employing an online learning algorithm for UAV search mission in an adversarial environment.

We leverage the multi-armed bandits (MAB) technique, as an effective approach to address the challenges inherent in navigating through adversarial conditions within unknown environments. Our approach involves the utilization of MAB algorithms, specifically the Exp3 algorithm for exponential-weighted exploration and exploitation [8]. The bandit problem is modeled as a sequential game between the learner (agent) and the environment, spanning multiple rounds represented as the time horizon $\mathcal{T}$. In each round t, the learner selects an action k from a predefined set of actions denoted as $\mathcal{K}$, next receiving feedback from the environment

Research reported in this publication was supported by the Qatar Research Development and Innovation Council ARG01-0527-230356. The content is solely the responsibility of the authors and does not necessarily represent the official views of Qatar Research Development and Innovation Council.

in the form of a penalty associated with the chosen action k. The main objective of the learner is to continually enhance their performance over time by learning from the feedback to reach the optimal policy (best action) through ongoing interaction with the environment. This adaptive learning process empowers the agent to efficiently learn the strategy of the targets, assuming they are moving with adversarial behavior.

We review existing techniques used in the search mission problem involving a single UAV or groups of UAVs where finding an optimal search path is NP-hard [9]. Researchers have employed classical techniques, heuristics, meta-heuristics, and machine learning methods to enhance the search process, accounting for efficiency, obstacle avoidance, and mission objectives [10]–[12]. This paper focuses on real-time learning of the strategic approach of a target formation.

Optimization-oriented methods have been used to maximize target detection probability [13]–[15], but they may face challenges in unknown environments with real-time changes in the target formation movement. Adversarial environments are addressed in the context of target discovery using RL with adversarial training. The application of RL with adversarial training techniques enhances the model's robustness to adversarial environments [16]–[18]. In this paper, we employ an adversarial multi-armed bandit framework that is designed to adapt online in adversarial environments without the need for adversarial training. This design is based on the assumption of a non-stationary probability distribution, enabling effective performance in such challenging conditions.

The main contributions of this paper are as follows:

1) Formulate the problem of the search mission of UAV as an optimization problem. The objective is to maximize the overall performance of the agent by converging towards the optimal policy that reflects the ability of the UAV to learn the formation's strategy.
2) Apply the contextual Exp3 algorithm to address the challenge of a single UAV in an unknown and adversarial environment, which converges to the optimal policy.
3) Evaluate the performance of the Exp3 in the context of a search mission using a single UAV. By conducting a series of experiments to comprehensively assess their capabilities and show their convergence of reaching the optimal policy.

The paper is structured as follows: section II, we establish the system model and we articulate the problem formulation as an optimization problem in section III. Moving forward, section IV introduce our proposed methodologies and outlines how we integrate the problem into an online learning framework under the assumption of a fixed target strategy. To validate the effectiveness of the algorithm, Section V assesses the performance of the MAB algorithm across various scenarios.

II. SYSTEM MODEL

In this section, we present the system model for a mobile UAV(agent) with a search mission in an unknown and adversarial environment. We focus on highlighting the key components of the environment that facilitate solving the problem using bandit algorithms.

We define the coverage area and the movement of the agent within this area. The coverage area is represented as a grid consisting of $|\mathcal{N}|$ cells, each labeled with $n \in \{1, 2, \ldots, |\mathcal{N}|\}$. These cells correspond to potential actions k within the action space $\mathcal{K}$, as shown in Fig. 1. Furthermore, each cell represents a unique location that can either be empty or contain one or more targets. This grid structure serves as a framework for the agent to systematically explore the area.

Agent's Movement. The agent's movement is defined by discrete steps, ensuring that the UAV's trajectory across the predefined grid-based coverage area is expressed in terms of discrete actions. At any given time slot t, the agent selects an action k from its available action space $\mathcal{K}$, determining its transition to a neighboring cell within the grid. Alternatively, the agent can choose to remain in the current cell. The time slot t represents the time needed by the agent to move from one cell to another. This duration depends on both the size of the cell and the speed of the UAV. It is important to note that the non-stationary nature, reflecting the movements of the targets, occurs between time slots. It is important to note that all algorithms discussed in this paper are agnostic to the actual duration of the abstract concept of the *time slot*, and are still directly applicable to different configurations with the same guarantees.

Exploration Strategy. The primary objective of the agent is to search for targets within the fixed coverage area. By systematically moving from cell to cell, the agent continually observes each cell to gather information about target presence. This process enables the agent to learn the strategy adopted by the target formation across the grid, assuming the movement of the formation of targets resembles an adversarial behavior. The targets can observe the actions taken by the agent, and as a result, they have some information about the policy adopted by the agent. Consequently, the targets may attempt to deceive the UAV.

Observation and Decision. The agent's information is

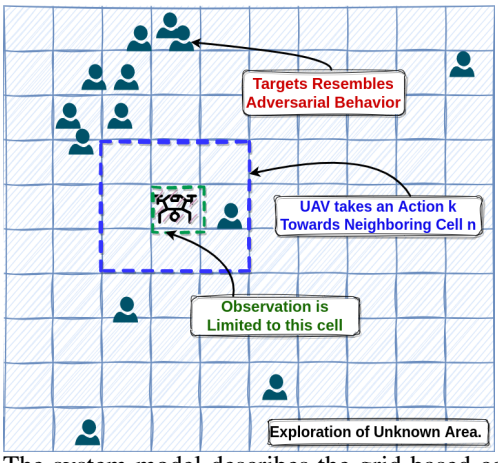


Fig. 1: The system model describes the grid-based coverage area and the discrete movement of the agent. Each cell n corresponds to a possible action k in the agent's action space $\mathcal{K}$. The objective is to learn the strategy of the targets.

constrained to its current cell n. It has the capability to observe the presence or absence of targets within the cell it currently occupies, leading to a loss $\mathbf{L}_t$ associated with the observation of targets at time slot t. Based on the accumulated observations or losses, the agent then decides the next movement (action). If the agent fails to detect targets after taking an action k_t, a loss of $\mathbf{L}_t^{k_t}$ is assigned as 1; otherwise, the loss is set to 0. Moreover, the optimal action k^* is estimated based on the full knowledge of the future losses. The benchmark for k^* is the cumulative loss incurred by the agent, assuming it knows the presence of targets across all $t \in \mathcal{T}$. Therefore, the optimal action is unknown to the agent since the agent is not capable of observing the time-varying cost associated with action k_t unless the agent samples k_t at first, then observes the cost only at time slot t for k_t.

III. Problem Statement

Our main objective is to maximize the agent's performance by formulating the problem to minimize the cumulative losses, thus, minimizing the cumulative regret $\mathcal{R}_\mathcal{T}$, with the ultimate goal of identifying the optimal action k^*. The optimal policy k^* describes the underlying strategy of the target formation. However, the policy k^* is unknown to the agent. Therefore, our aim is to find a policy k^* that minimizes the cumulative regret $\mathcal{R}_\mathcal{T}$. This can be formulated as the following optimization problem:

$$\mathbf{P:} \qquad \min_k \sum_{t=1}^{\mathcal{T}} \mathbf{L}_t^k \tag{1}$$

Within this optimization problem, the agent selects action k to be sampled at each time slot t within the time horizon $\mathcal{T}$. The primary challenge arises at each time slot t when the agent must decide on the next move. At this point, the loss $\mathbf{L}_{t+1}^k$ linked with the subsequent move is inaccessible, complicating the resolution of the optimization problem. Nevertheless, the loss will be revealed in the subsequent time slot once the agent decides on an action for the next destination cell $nt+1$ to be visited and subsequently observes the existing targets in cell n_{t+1}. Subsequently, $\mathbf{L}_{t+1}^k$ becomes known to the agent. Hence, we employ MAB algorithms [8], which learn from continuous interaction with the environment and converge towards the optimal policy k^* as $t \to \mathcal{T}$.

A. Optimal Policy and Cumulative Regret.

In this section, we introduce the concept of **regret** within the framework of online convex optimization. Regret serves as a measure for evaluating the UAV's decision-making performance throughout the mission, measuring how effectively the agent's decisions align with the optimal action in hindsight. In online convex optimization, the player, in our case, the UAV, iteratively makes decisions without knowledge of future outcomes, incurring costs based on its selected actions, all with the ultimate objective of attaining the optimal policy k^*. This concept holds particular significance in our UAV search mission scenario, where the UAV endeavors to optimize its actions while adapting to the non-stationary behavior of the target formation.

Regret. To achieve this, we first establish the components of the system. We adopt the standard online setup, wherein at each time slot $t \in 1, 2, ..., \mathcal{T}$, the agent selects an action k_t from a set $\mathcal{K}$ at time slot t. This set is characterized as closed and bounded. The consequence of action k_t is reflected in the loss $\mathbf{L}_t^{k_t}$ associated with action k at time t. The regret, with respect to the best fixed action k^*, is defined as the sequence of actions k_t every time slot t in terms of their cumulative losses:

$$\mathcal{R}_\mathcal{T} = \sum_{t=1}^{\mathcal{T}} (\mathbf{L}_t^{k_t} - \mathbf{L}_t^{k^*}) \tag{2}$$

Optimal Policy. Here, $k^* \in \arg\min_{k \in \mathcal{K}} \sum_{t=1}^{\mathcal{T}} \mathbf{L}_t^k$ represents the action k^* that minimizes the accumulated loss between all actions in the action space $\mathcal{K}$. In essence, k^* serves as a benchmark for comparison, representing the optimal action based on perfect knowledge of the outcomes. Over time, we expect to observe a saturation in regret $R_\mathcal{T}$ as t approaches $\mathcal{T}$, or in mathematical terms, $\lim_{t \to \infty} \frac{\mathcal{R}_\mathcal{T}}{\mathcal{T}} = 0$. Specifically, we aim to achieve $\mathcal{R}_\mathcal{T} = O(\sqrt{\mathcal{T}})$, demonstrating that the agent achieves the optimal policy as t approaches the time horizon $\mathcal{T}$.

IV. Contextual Exp3

The Exp3 algorithm is a widely recognized method used to address the multi-armed bandit problem. Its primary objective is to strike a balance between exploration and exploitation in an online manner. Exploration entails visiting new destination cells or sampling new actions, while exploitation involves leveraging actions with the aim of minimizing cumulative regret $\mathcal{R}_\mathcal{T}$ [8]. The algorithm begins by initializing the weights $\hat{S}_{0,k}$ for each action k as a uniform distribution, which defines the probability of the possible destination cells for the agent within the grid. In each time slot t, a sampling distribution P_k^t is calculated to determine the probability of selecting each action k_t. Then, an action is sampled according to this distribution, and the associated loss $\mathbf{L}_t^{k_t}$ is observed, which represents the observation of targets in the visited cell n. Based on the observed loss, the estimated weight $\hat{S}_{t,k}$ for the action k_t is updated. This update process ensures that actions with higher weights are assigned higher probabilities in subsequent time slots. The algorithm iterates through this process over the $\mathcal{T}$. By adjusting the learning rate parameter η, the Exp3 algorithm controls the trade-off between exploring other actions and exploiting the action with higher estimated weight. The optimal learning rate η_{opt} can be estimated using the equation in Theorem (11.1) from [8]. Algorithm 1 shows the steps of contextual Exp3.

Contextual Bandits. In many bandit problems, the agent has access to additional information (context) that could aid in predicting the quality of actions [8]. In our scenario, we incorporate context into the Exp3 algorithm (C-Exp3). Here, we utilize the UAV's current location, which can be any cell $n \in \mathcal{N}$, as a context c where $c \in \mathcal{C}$ with $\mathcal{C}$ representing the set of available contexts. The action space for each context c is denoted as $\mathcal{K}$ and remains consistent across all contexts. It is defined as $\mathcal{K}$ = Current, North (N), South (S), East (E), West (W), Northeast (NE),

Algorithm 1 Contextual Exp3 (C-Exp3)

Input: $\mathcal{T}$, $\mathcal{K}$, η, $\mathcal{C}$
Create the action set $\mathcal{K}$ for each context $c \in C$.
Set $\hat{S}_{0,k} = 0$ for all $k \in \mathcal{K}$ ▷ % Initial weights for actions.
for $t = 1$ to $\mathcal{T}$ **do**
  Observe context $c_t \in \mathcal{C}$
  Calculate the sampling distribution $P_{t,k}$:
  $= \frac{\exp(\eta \hat{S}_{t-1,k})}{\sum_{k=1}^{\mathcal{K}} \exp(\eta \hat{S}_{t-1,k})}$ for all $\mathcal{K}$.
  Sample $k_t \sim P_{t,k}$ and observe loss $\mathbf{L}_{c_t}^{k_t}$.
  for each $k \in \mathcal{K}$ **do**
    Calculate $\hat{S}_{t,k} = \hat{S}_{t-1,k} + \frac{1-\mathbb{I}\{k=k_t\}(\mathbf{L}_{c_t}^{k_t})}{\mathcal{P}_{t,k}}$ ▷ % weights are updated.
  end for
end for

Northwest (NW), Southeast (SE), Southwest (SW). This action space offers the agent a choice among eight possible directions for movement within a 2D grid at each time slot, corresponding to the agent's current position and its neighboring cells. Additionally, this definition of the action space imposes a constraint on the agent's movement, allowing only one step in any direction. Consequently, the agent moves to one of the adjacent cells during each time slot, promoting smaller and more localized actions. Collectively, the action space for all contexts $\mathcal{C}$ defines the entire grid. In each context, each cell n signifies a potential direction relative to the current context c_t. For instance, cell $n = 6$ can be associated with context $c = 5$ when moving in the *E* direction or context $c = 7$ when moving in the *W* direction.

Regret. To assess the agent's performance, we employ a regret measure for each context. This measure quantifies the agent's accumulated loss relative to an ideal context-dependent policy, denoted as k_c^* in hindsight. It's expressed as follows in reference to Eq. (18.1) in [8]:

$$\mathcal{R}_{\mathcal{T},c} = \sum_{t=1}^{\mathcal{T}} \left[(\mathbf{L}_{c_t}^{k_t} - \mathbf{L}_{c_t}^{k_c^*}) \right] \mathbb{I}\{c = c_t\} \tag{3}$$

Hence, the regret $\mathcal{R}_{\mathcal{T}}$ for agent is:

$$\mathcal{R}_{\mathcal{T}} = \sum_{c \in \mathcal{C}} \mathcal{R}_{\mathcal{T},c} \tag{4}$$

Here, $k_c^* \in \arg\min_{k \in \mathcal{K}} \sum_{t=1}^{\mathcal{T}} \mathbf{L}_{c_t}^{k} \mathbb{I}\{c = c_t\}$. We measure the difference in loss between the agent's decision k_t and the optimal context-dependent best action k_c^* for a given context c at each time slot t. By summing these differences across the time horizon $\mathcal{T}$ and potential contexts from the set $\mathcal{C}$, we derive the total regret $\mathcal{R}_{\mathcal{T}}$. The identity function $\mathbb{I}\{c = c_t\}$ acts as a filter. It equals 1 when the condition $c = c_t$ is true, indicating that the regret calculation only applies to instances when c matches c_t (i.e., we're interested in the regret for c at time t). If the condition is false, the identity function equals 0, effectively excluding those instances from the sum. Furthermore, we define the optimal policy to be a possible cell n if it reflects a k_c^* for multiple contexts that reflect the strategy of the formation. Our goal is to learn the mapping between contexts and optimal actions. The Exp3 algorithm provides an **upper-bound** for worst-case scenarios, accounting for the observed context distribution. This bound is valid in adversarial environments for the entire system, as detailed in Eq. (18.3) in [8]:

$$\mathcal{R}_{\mathcal{T}} \leq \sum_{c \in \mathcal{C}} \mathcal{R}_{\mathcal{T},c} \leq 2 \sum_{c \in \mathcal{C}} \sqrt{\log(|\mathcal{K}|) \sum_{t=1}^{\mathcal{T}} |\mathcal{K}| \mathbb{I}\{c = c_t\}} \tag{5}$$

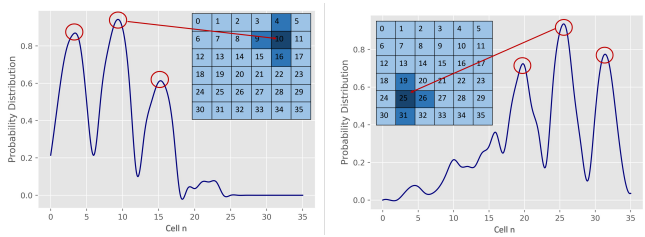
V. Performance evaluation

In this section, we evaluate the performance of contextual Exp3 in a search mission involving a single UAV agent. The context space $\mathcal{C}$ is defined as a grid of size 36 (6x6 grid). Our environment assumes no obstacles and focuses solely on the mobility of targets, disregarding external factors that may influence their movement. The environment itself is unknown, and the targets move according to a mobility pattern within a time horizon T that describes the adversarial behavior.

Environment: The mobility pattern of the formation of targets follows the reference point group mobility pattern (**RPG**). The RPG model serves as a potent tool for emulating the collective dynamics of target formations. The RPG model is used in [19] for the problem of low complexity target tracking to cover and follow moving targets using UAV. Within this model, each target is affiliated with a logical center, known as the group leader, which governs the collective motion characteristics of the group. The targets comprising a group are distributed in an adversarial manner around the reference point. By employing their distinct mobility models, these targets are moving with random magnitude ν and angle direction θ assimilated into the reference point, which steers their trajectories in accordance with the group's direction.

At any given moment in time T, each target possesses unique values of θ and ν, which deviate randomly from those of the group leader. However, the target remains within the leader's boundary. In our RPG model, the movement of the group leader involves selecting a destination point within the deployment region in a stochastic manner. The leader then moves towards this destination with corresponding values of θ^* and ν^*. This motion profile establishes the leader's distinct trajectory and sets the overall motion trend for the entire group. As a result, each group member exhibits variations from this predominant motion vector, introducing individualistic dynamics into the collective formation. Furthermore, in our scenario, we assume the agent operates at a fixed altitude and is equipped with imaging sensors that enable it to observe the targets in the environment.

Model Parameters: We use the optimal learning rate η_{opt} as discussed in Sec. IV for all the experiments. The action space $\mathcal{K}$ and the context space $\mathcal{C}$ are predefined. In this paper, we specifically focus on analyzing the algorithm's performance, which is why we opted for a relatively small context space. Furthermore, the time horizon $\mathcal{T}$ is predefined and varies according to each experiment. We consider two distinct RPG-based mobility patterns denoted as $p \in \mathcal{P}$, each associated with a unique probability distribution for the leader. Notably, each mobility pattern corresponds to its specific context-dependent best action $k_{c,p}^*$.



(a) p_1 (b) p_2

Fig. 2: Probability Distribution of $p \in \mathcal{P}$ of the Leader. Fig 2a describes that the leader is most likely to be located around cell 25. While Fig. 2b describes another policy where the leader is located near cell 13.

Performance Metric: We evaluate the algorithm's performance using cumulative regret and we show the utility and action probabilities. The goal is to minimize cumulative regret, which should steadily converge over time. The utility measures the algorithm's success in accumulating rewards relative to the maximum potential rewards achievable with the optimal policy. The utility at each time step is expressed as a percentage of the utility obtained by the optimal policy. Furthermore, to enhance performance stability, we average each data point with the preceding 100 points. Additionally, action probabilities indicate the assigned weights to actions, with the highest weight signifying the agent's effective policy to minimize losses. In this section, we evaluate the C-Exp3 algorithm in environments where the optimal policy is fixed, and if the optimal policy is changing across $\mathcal{T}$.

For experiments 1 and 2, we employ the C-Exp3 algorithm outlined in Algorithm 1. The regret is computed using Eq. 4, and the upper bound is determined using the approach explained in Sec. IV through Eq. 5. We adopt the optimal learning rate η_{opt} to be 0.0018, along with the mobility pattern p_1 and p_2 described in Fig. 2.

Fixed Optimal Policy: Fig. 3 illustrates the agent's performance in an adversarial environment with a fixed optimal policy. The leader follows the mobility pattern $p = 1$ described in Fig. 2a, with the optimal policy cell $n = 10$. To identify the optimal policy, we use the probability distribution of $p = 1$ of the leader during $\mathcal{T} = [0, 100] \times 10^3$.

Discussion: In Fig. 3a, the agent's cumulative regret consistently remains lower than the upper bound, clearly indicating convergence. This convergence pattern suggests a steady relationship between the cumulative regret of the optimal policy $n = 10$ and the agent's cumulative regret, confirming the successful adoption of an optimal strategy ($n = 10$). The utility in Fig. 3b further emphasizes this trend by displaying a continuous increase over time, ultimately reaching the performance of the optimal policy $n = 10$ and outperforming the baseline which is based on a random action selection. Additionally, the alignment seen in Fig. 3c of the highest probability of the cell $n = 10$ adds further support to these observations. Taken together, these combined findings provide compelling evidence of the C-Exp3 algorithm's effectiveness in acquiring and maintaining an optimal strategy within an environment where the strategy of the formation remains constant but resembles an adversarial behavior.

Changing Optimal Policy: Fig. 4 depicts the agent's performance in an adversarial environment with a changing of optimal policy. Initially, the leader follows the mobility pattern $p = 1$ described in Fig. 2a, where the optimal policy is $n = 10$, during the time interval $t \in \mathcal{T}_1 = [0, 40] \times 10^3$. Subsequently, the leader switches to the mobility pattern $p = 2$ described in Fig. 2b, where the optimal policy becomes $n = 25$, during the time interval $t \in \mathcal{T}_2 = [40, 200] \times 10^3$. The optimal policies $n = 10$ and $n = 25$ are identified using the probability distribution of $p = 1$ and $p = 2$ of the leader, respectively.

Discussion: The cumulative regret exhibited by the agent in Fig. 4a shows a notable pattern. Throughout the initial interval $\mathcal{T}_1$, we observe convergence in cumulative regret, with the agent's performance consistently remaining below the upper bound. However, a distinctive shift occurs during the subsequent interval $\mathcal{T}_2$, marked by a sharp increase in the agent's regret. This transition aligns with changes in the mobility pattern of target formations to $p = 2$, prompting a shift in optimal policy to $n = 25$. During $\mathcal{T}_2$, the regret experiences a period of ascent, nearing the upper bound, before ultimately converging. The convergence signifies the agent's ability to adapt and refine its strategy to minimize loss.

The difference between the agent's performance and the optimal policy is further highlighted through the utility in Fig. 4b. Initially, the agent's performance mirrors that of the optimal policy $n = 10$, with a subsequent sharp decline coinciding with the change in the optimal policy, followed by a subsequent recovery. Similarly, the action probability in Fig. 4c accentuates this phenomenon. During $\mathcal{T}_1$, the action with the highest probability consistently corresponds to $n = 10$. As the target formation's strategy undergoes changes, the agent requires time to unlearn its previous experiences from $\mathcal{T}_1$ and adapt to the new strategy, which is $n = 25$ during $\mathcal{T}_2$.

VI. Conclusion

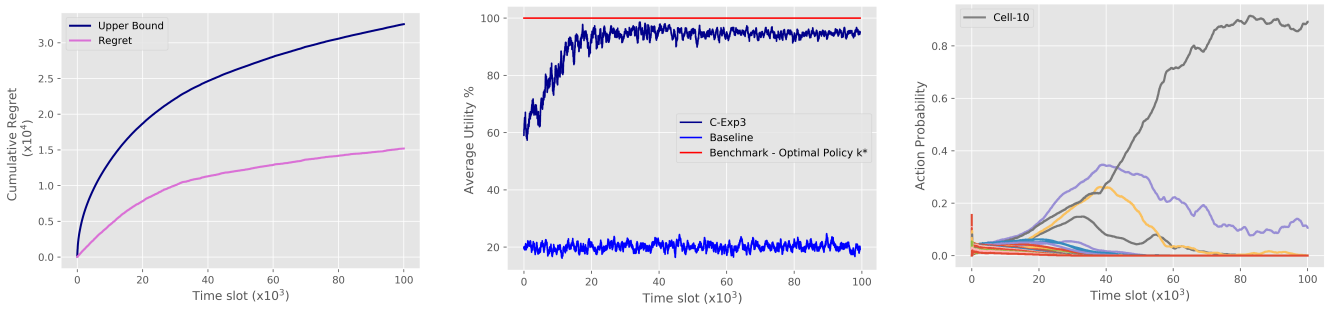
In conclusion, this paper has proposed the utilization of the Exp3 algorithm to optimize UAV search missions in unknown and adversarial environments. The UAV's objective is to locate mobile target formations that exhibit behavior resembling an adversarial behavior. This is achieved through learning their strategic approach to navigating the environment. The performance of C-Exp3 has been evaluated through a series of experiments in adversarial environments. These experiments have primarily focused on assessing the UAV's capacity to converge towards an optimal policy, which, in turn, reflects its ability to effectively learn and adapt to the strategies employed by the target formation.

References

[1] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on uavs for wireless networks: Applications, challenges, and open problems," *IEEE communications surveys & tutorials*, vol. 21, no. 3, pp. 2334–2360, 2019.

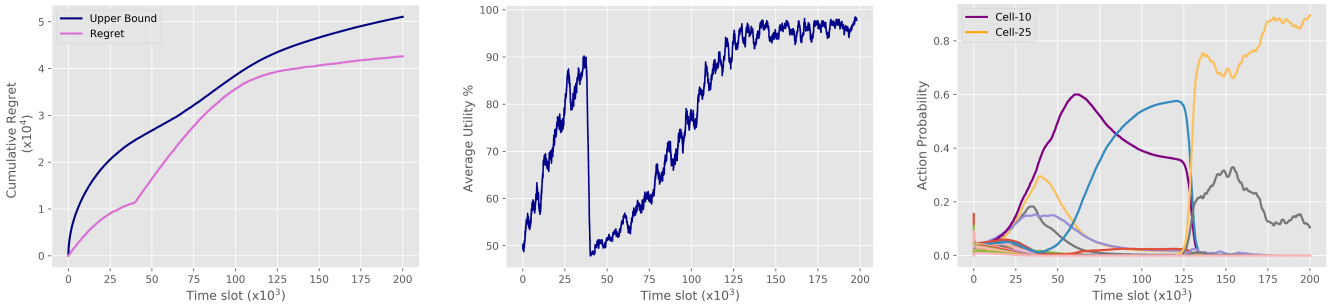
[2] Z. Xiaoning, "Analysis of military application of uav swarm technology," in *2020 3rd International Conference on Unmanned Systems (ICUS)*. IEEE, 2020, pp. 1200–1204.

[3] J. Gu, T. Su, Q. Wang, X. Du, and M. Guizani, "Multiple moving targets surveillance based on a cooperative network for multi-uav," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 82–89, 2018.



(a) Cumulative Regret (b) Average Utility (c) Cell Probability

Fig. 3: Performance of Contextual Exp3 in Adversarial Environment with Fixed Best Action.



(a) Cumulative Regret (b) Average Utility (c) Cell Probability

Fig. 4: Performance of Contextual Exp3 in Adversarial Environment with Changing Best Action.

[4] W. Yue, X. Guan, and L. Wang, "A novel searching method using reinforcement learning scheme for multi-uavs in unknown environments," *Applied Sciences*, vol. 9, no. 22, p. 4964, 2019.

[5] X. L. Wei, X. L. Huang, T. Lu, and G. G. Song, "An improved method based on deep reinforcement learning for target searching," in *2019 4th international conference on robotics and automation engineering (icrae)*. IEEE, 2019, pp. 130–134.

[6] A. Soliman, A. Al-Ali, A. Mohamed, H. Gedawy, D. Izham, M. Bahri, A. Erbad, and M. Guizani, "Ai-based uav navigation framework with digital twin technology for mobile target visitation," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106318, 2023.

[7] J. Kwon, Y. Efroni, C. Caramanis, and S. Mannor, "Rl for latent mdps: Regret guarantees and a lower bound," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 523–24 534, 2021.

[8] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[9] L. Lin and M. A. Goodrich, "Hierarchical heuristic search using a gaussian mixture model for uav coverage planning," *IEEE transactions on cybernetics*, vol. 44, no. 12, pp. 2532–2544, 2014.

[10] A. A. R. Newaz, F. A. Pratama, and N. Y. Chong, "Exploration priority based heuristic approach to uav path planning," in *2013 Ieee Ro-Man*. IEEE, 2013, pp. 521–526.

[11] X. Zhang, J. Chen, B. Xin, and H. Fang, "Online path planning for uav using an improved differential evolution algorithm," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 6349–6354, 2011.

[12] I. Oz, H. R. Topcuoglu, and M. Ermis, "A meta-heuristic based three-dimensional path planning environment for unmanned aerial vehicles," *Simulation*, vol. 89, no. 8, pp. 903–920, 2013.

[13] M. D. Phung and Q. P. Ha, "Motion-encoded particle swarm optimization for moving target search using uavs," *Applied Soft Computing*, vol. 97, p. 106705, 2020.

[14] M. A. Alanezi, H. R. Bouchekara, M. S. Shahriar, Y. A. Sha'aban, M. S. Javaid, and M. Khodja, "Motion-encoded electric charged particles optimization for moving target search using unmanned aerial vehicles," *Sensors*, vol. 21, no. 19, p. 6568, 2021.

[15] S. K. Gan and S. Sukkarieh, "Multi-uav target search using explicit decentralized gradient-based negotiation," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 751–756.

[16] X. Bai, J. Guan, and H. Wang, "A model-based reinforcement learning with adversarial training for online recommendation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] M. Fischer, M. Mirman, S. Stalder, and M. Vechev, "Online robustness training for deep reinforcement learning," *arXiv preprint arXiv:1911.00887*, 2019.

[18] Z. Yijing, Z. Zheng, Z. Xiaoyi, and L. Yang, "Q learning algorithm based uav path learning and obstacle avoidence approach," in *2017 36th Chinese control conference (CCC)*. IEEE, 2017, pp. 3397–3402.

[19] M. Khan, K. Heurtefeux, A. Mohamed, K. A. Harras, and M. M. Hassan, "Mobile target coverage and tracking on drone-be-gone uav cyber-physical testbed," *IEEE Systems Journal*, vol. 12, no. 4, pp. 3485–3496, 2017.