Theoretical Computer Science 1079 (2026) 116034

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

A multi-armed bandit approach to UAV sensor fusion for target tracking ⋆

Yang Lv [a], Guochao Fan [b], Mengzhen Li [a], Xiongjun Liu [b], Pengqing Liu [a], Yapu Zhang [a,*]

[a] Institute of Operations Research and Information Engineering, Beijing University of Technology, Beijing, 100124, PR China
[b] Beijing Jinghang Computation and Communication Research Institute, Beijing, 100074, PR China

ARTICLE INFO

Section Editor: Prof. Seth Pettie
Handling Editor: Prof. Katie Harris

Keywords:
Adversarial multi-armed bandit
Data fusion
Partially observable target tracking

ABSTRACT

Unmanned Aerial Vehicles, known for their low cost and high maneuverability, have become a promising tool for target tracking tasks. This paper addresses the challenge of tracking targets in partially observable and adversarial environments. We propose a data fusion strategy based on the Exp4-IX algorithm, which integrates multiple expert predictions to enhance tracking accuracy under adversarial conditions. To assist a single UAV in monitoring targets that are difficult to detect due to limited sensing capabilities, we construct a model of the target environment and design a corresponding reward function. Our algorithm adopts the Exp4-IX method, which is derived from the adversarial multi-armed bandit framework with expert advice and guarantees a sub-linear regret bound, implying that the UAV's average regret asymptotically approaches zero over time. In simulation experiments, the algorithm integrates the advantages of Previous Position, Particle Filtering, and Trajectory Fitting methods, and is compared against the Average Fusion baseline. Results show that this fusion method performs well in both smooth and adversarial trajectory scenarios. By leveraging the Exp4-IX algorithm, the UAV is able to anticipate the target's next position and achieve superior tracking performance.

1. Introduction

Unmanned Aerial Vehicles (UAVs) are increasingly valued for their low cost, compact size, and excellent maneuverability, offering wide-ranging applications. Equipped with sensors, IoT modules, and processors, UAVs can execute tasks involving sensing, communication, and computation. Their versatility makes them useful in numerous fields, including maritime missions [2], emergency response [3], precision agriculture [4,5], network coverage [6], and data collection [7], target tracking [8] and sweep coverage [9]. Among these, tracking moving targets is particularly compelling, as UAVs can navigate expansive or hard-to-reach areas inaccessible to humans.

Extensive literature exists on the problem of tracking a single moving target using a single UAV. For example, Lai and Huang [10] employed YOLO for real-time object detection in scenarios with stationary UAVs, while Bhagat and Sujit [11] demonstrated that Deep Q-Networks (DQN) perform well under conditions of partial visibility. Building on this, Bhagat and Sujit [12] proposed the TF-DQN framework and further extended it to TF-DDQN, incorporating Double DQN to reduce value estimation bias. This approach enabled

⋆ This is an enhanced and extended version of a paper presented at the 17th International Conference on Combinatorial Optimization and Applications (COCOA 2024) [1] .

* Corresponding author.
E-mail address: zhangyapu@bjut.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.tcs.2026.116034
Received 5 August 2025; Received in revised form 27 October 2025; Accepted 12 May 2026
Available online 15 May 2026

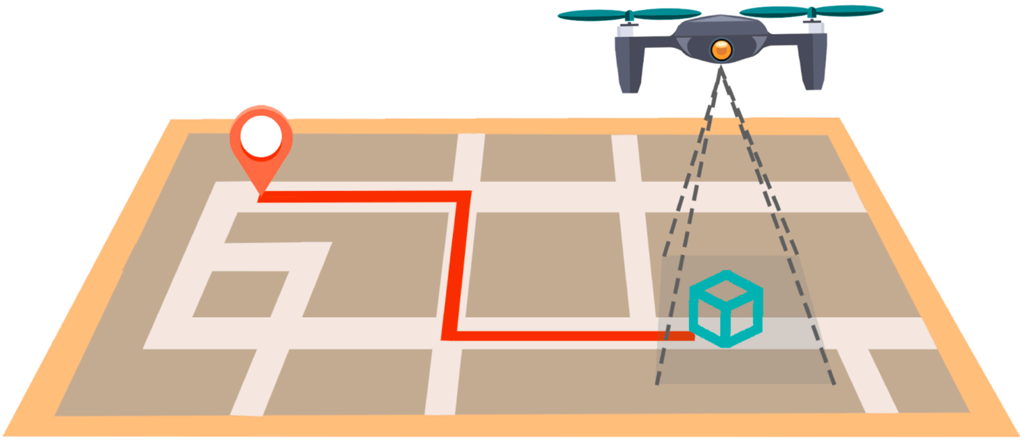

Fig. 1. Illustration of a UAV tracking an adversarial target.

stable tracking performance even in complex urban environments with frequent occlusions and randomly moving targets. In a related effort, Zhao et al. [13] developed a true end-to-end Soft Actor-Critic (SAC) architecture, which directly takes UAV downward-facing images as input and outputs velocity control commands for dynamic target tracking. When transmitting tracking data to a fusion center, it is crucial to evaluate the performance of each method and select the most reliable expert advice to ensure continuous target coverage.

In cases where multiple expert strategies are available, tracking can be formalized using a decision-making model. Here, the UAV selects one of k discrete actions—in our case, four: forward, backward, left, and right. Let x_t denote the reward at time t, determined by the UAV's selected action. The goal is to identify the most reliable expert by solving the following optimization:

$$\min R_n = \min \mathbb{E}\left[\max_{m\in[M]} \sum_{t=1}^{n} E_m^{(t)} x_t - \sum_{t=1}^{n} X_t\right] \tag{1}$$

Here, n is the number of decision steps, R_n quantifies regret (the difference between the cumulative reward of the best expert and the chosen actions), $E_m^{(t)}$ is the probability distribution over actions from expert m at time t, and X_t is the reward received from the actual decision. If the optimal expert were chosen from the start, the regret would be zero.

Autonomous target tracking using UAVs has emerged as a critical capability in various applications. Tracking a moving target using a UAV involves two key challenges: partial observability and adversarial dynamics.

a) Partial Observability: Despite UAVs being equipped with long-range sensors, their sensing region is limited—a 60 km forward range and 80 km width. When the target is outside this area, it becomes invisible. To handle this, we employ online learning techniques that assess the quality of actions and refine the UAV's decisions over time. Balancing exploration and exploitation, characteristic of bandit problems, is key. Poor expert advice leads to losing track of the target, yielding zero reward and increasing regret quickly.

b) Adversarial Dynamics: The target may deliberately move in the opposite direction of the UAV, presenting adversarial behavior [14]. Expert strategies must not only predict target behavior but also respond effectively under such hostile conditions.

As shown in Fig. 1, a single UAV is tasked with tracking a dynamic target in an adversarial setting. Due to partial observability, the UAV can only monitor a square region, as illustrated in the figure. If the target exits this region, it becomes undetectable. Adversarial conditions arise when the target moves in a direction opposite to the UAV, requiring the UAV to rapidly adapt its strategy. To address this challenge, multiple expert strategies are employed to predict the target's next location more effectively.

While some prior works [15] examine adversarial bandits, approaches like Fixed-Share [16] result in higher regret bounds. Other studies [17–19] assume random or independent motion models, which fail to reflect the often adversarial nature of real-world target movements.

Our contribution is to propose a UAV target tracking algorithm based on the bandits with expert advice framework, integrating multiple prediction algorithms to improve performance. A preliminary version of this work appeared in COCOA 2024 [1]. Leveraging the Exp4-IX method by Neu [20], we establish that under the condition $\delta < O(1/M)$, our method achieves a strong high-probability regret bound (as shown in Theorem 1 of Neu [20]). This constraint is not strict, as δ is typically small. Our empirical evaluations demonstrate that the proposed algorithm excels in predicting target motion when integrating three expert strategies. The results confirm the UAV's capacity to integrate expert advice and accurately predict the most probable target trajectory, even in challenging scenarios.

2. Related work

The multi-armed bandit (MAB) framework under adversarial settings was fundamentally addressed by Auer et al. [21], who proposed the Exp3 and Exp4 algorithms. Unlike classical stochastic bandit models, their approach makes no probabilistic assumptions about reward generation. In particular, Exp4 incorporates expert advice into the decision-making process and guarantees a sublinear regret bound of order $\mathcal{O}(\sqrt{KT\ln N})$, making it especially suitable for structured decision-making under uncertainty. Seldin and Lugosi [22] later established a matching lower bound for the adversarial MAB problem with expert advice. They proved that any algorithm must suffer a regret of at least $\Omega\left(\sqrt{\frac{KT\ln N}{\ln K}}\right)$, thereby resolving a long-standing open question and confirming the near-optimality of Exp4 up to a logarithmic factor. In recent years, there has been increasing attention to UAV-based target search and tracking in non-stationary and adversarial environments. Khiala et al. [23] proposed an online learning strategy based on the MAB framework to enhance the target search capabilities of a single UAV under unknown and non-stationary target distributions. Their work formulates the task as a game-theoretic optimization problem and introduces a Contextual Exp3 (C-Exp3) algorithm, along with its adaptive variant AC-Exp3. By incorporating a human-in-the-loop mechanism, their system is able to detect distributional drift and adapt to changes in target behavior in real time.

In terms of state estimation and sensor fusion, Baidya et al. [24] proposed a unified EKF/UKF-based fusion scheme for optical, radar, and RF inputs, demonstrating superior robustness and accuracy compared to single-sensor and naive averaging approaches, especially in the presence of sensor noise or partial failure. Masalskyi et al. [25] introduced a hybrid filtering architecture combining particle filters with UKF using adaptive weighting, effectively handling unstable or intermittent observations in micro-scale UAV positioning. Dabbabi et al. [26] modeled multi-sensor inputs (e.g., RGB, depth, radar) as graph-structured data and employed a graph neural network to automatically learn fusion strategies that adapt to varying targets and environmental conditions.

3. Bandit with expert advice

This section introduces the background of the *Bandits with Expert Advice* problem in the context of UAV-based target tracking. Consider an interaction process lasting n rounds between a UAV and a moving target. In each round t, the UAV selects one of k possible movement directions: forward, right, downward, or left. Simultaneously, a set of M experts provide movement suggestions based on prior observations. Each expert's recommendation at round t is denoted by $E_m^{(t)} \in [0,1]^k$, satisfying the constraint $\sum_{i=1}^k E_{m,i}^{(t)} = 1$ for all $m \in \{1,\dots,M\}$. The UAV receives a reward vector $x_t \in [0,1]^k$, where each component reflects the reward corresponding to a particular action.

Given the partially observable nature of the UAV's environment, let (x, y) denote its current position. The UAV can observe within a sensing window that extends x_s units forward and y_s units laterally, forming a rectangular detection zone. This region is defined as:

$$\begin{cases} [x-y_s, x+y_s]\times[y, y+x_s] & \text{if moving forward;} \\ [x, x+x_s]\times[y-y_s, y+y_s] & \text{if moving right;} \\ [x-y_s, x+y_s]\times[y-x_s, y] & \text{if moving downward;} \\ [x-x_s, x]\times[y-y_s, y+y_s] & \text{if moving left.} \end{cases} \tag{2}$$

Rather than assigning binary rewards (e.g., 1 if the target is visible, 0 otherwise), we define the reward as a function of the distance between the UAV and the target. The closer the UAV is to the target, the higher the reward. Specifically, let d_t^i represent the distance after taking action i at time t, then the reward is defined as:

$$x_t^{(i)} = \begin{cases} \min\{\frac{1}{d_t^i}, 1\} & \text{if the target is within the sensing range;} \\ 0 & \text{if the target lies outside the sensing region.} \end{cases} \tag{3}$$

This formulation motivates the UAV to stay close to the target and maintain visibility. In adversarial scenarios, our proposed strategy guarantees sublinear regret, indicating that the UAV's long-term performance approaches that of the best expert in hindsight.

4. Exp4-IX and regret analysis

Exp4-IX is a widely recognized algorithm designed for adversarial multi-armed bandit problems with expert input. The name stands for *Exponential-weight algorithm for Exploration and Exploitation with Expert advice—Implicit eXploration*. It begins by assigning uniform weights to each expert, i.e., $Q_1 = (1/M,\dots,1/M) \in [0,1]^M$, and updates these weights based on received rewards through the following two main steps in each round:

1. Based on expert recommendations, calculate the action distribution for the UAV via a linear combination.
2. Observe the actual reward after action execution and adjust expert weights accordingly.

The step-by-step process is formalized in Algorithm 1. For notational simplicity, we define $y_{ti} = 1 - x_{ti}$ in place of the raw reward.

It is important to note that when $\gamma = 0$, the term $\frac{1}{P_{ti}}$ can lead to high variance when P_{ti} is close to zero. To mitigate this, Exp4-IX introduces an implicit exploration term γ to ensure adequate randomness in the action selection process. This mechanism helps stabilize updates and promotes more reliable learning by forcing occasional exploration [27]. With the inclusion of γ, the regret can be bounded with high probability.

The theoretical guarantee is given in Theorem 1, which establishes a high-probability upper bound for the cumulative regret. Compared to the result in Neu [20], our bound is tighter when $\delta < O(1/M)$, a condition that is generally easy to satisfy.

Algorithm 1: Exp4-IX Algorithm.

Input: Number of rounds n, number of actions k, number of experts M, expert advice at different times $E^{(t)} \in [0,1]^{k\times M}$, algorithm parameters η, γ

Initialize $Q_1 = (1/M, \dots, 1/M)^T \in [0,1]^{M\times 1}$, $\tilde{S}_0 = (0,\dots,0)^T_{M\times 1}$

for $t = 1, \dots, n$ **do**

```
// Receive advice from M experts and give flight suggestions to the UAV
```

Receive advice $E^{(t)}$

Choose the action $A_t \sim P_t$, where $P_t = E^{(t)} Q_t$

```
// Receive reward based on UAV flight state and update expert weights
```

Receive the reward $Y_t = 1 - x_{tA_t}$

Estimate the action rewards: $\hat{Y}_{ti} = \frac{\mathbb{I}\{A_t = i\} Y_{ti}}{P_{ti} + \gamma}$

Propagate the rewards to the experts: $\tilde{Y}_t = E^{(t)T} \hat{Y}_t$

Compute experts' cumulative rewards:

$$\tilde{S}_{t,i} = \tilde{S}_{t-1,i} + \tilde{Y}_{ti} \quad \text{for all } i \in [M]$$

Update the distribution Q_t using exponential weighting:

$$Q_{t+1,i} = \frac{\exp(-\eta \tilde{S}_{ti})}{\sum_j \exp(-\eta \tilde{S}_{tj})} \quad \text{for all } i \in [M]$$

Theorem 1. *Let* $\delta \in (0,1)$ *be a confidence parameter, and define the learning rates as*

$$\eta = \sqrt{\frac{\log M + \log(k+1) - \log\delta}{nk}}, \quad \gamma = \frac{\eta}{2}.$$

Then, with probability at least $1-\delta$*, the regret satisfies:*

$$R_n \le \frac{k}{2} \log\left(\frac{k+1}{\delta}\right) + 2\sqrt{nk(\log M + \log(k+1) - \log\delta)}. \tag{4}$$

To facilitate the proof of Theorem 1, we decompose it into three components. Each of the following lemmas establishes an upper bound for one of these components. Finally, we will combine the results to complete the proof of Theorem 1.

Let the reward obtained by the m-th expert at round t be defined as $Y_{tm} = \sum_{i=1}^k E_{mi}^{(t)} y_{ti}$. The total reward of expert m over n rounds is then given by $\sum_{t=1}^n Y_{tm}$. Let the optimal expert be $m^* = \arg\min_{m\in[M]} \sum_{t=1}^n Y_{tm}$. We aim to compute the high-probability upper bound on the regret R_n, where $R_n = \sum_{t=1}^n Y_t - \sum_{t=1}^n Y_{tm^*}$.

Lemma 1. *The following inequality holds:*

$$\sum_{t=1}^n \sum_{m=1}^M Q_{tm}\tilde{Y}_{tm} - \sum_{t=1}^n \tilde{Y}_{tm} \le \frac{\log M}{\eta} + \frac{\eta n k}{2}.$$

Proof. Let $W_n = \sum_{m=1}^M \exp(-\eta \tilde{S}_{nm})$, it follows that $W_0 = M$,

$$\exp(-\eta\tilde{S}_{nm}) \le \sum_{m=1}^M \exp(-\eta\tilde{S}_{nm}) = W_0 \frac{W_1}{W_0} \cdots \frac{W_n}{W_{n-1}} = M \prod_{t=1}^n \frac{W_t}{W_{t-1}}.$$

Simplifying each term within the product, we obtain

$$\begin{aligned}\frac{W_t}{W_{t-1}} &= \sum_{m=1}^M \frac{\exp(-\eta\tilde{S}_{tm})}{\sum_{m=1}^M \exp(-\eta\tilde{S}_{t-1,m})} = \sum_{m=1}^M \frac{\exp(-\eta\tilde{S}_{t-1,m})\exp(-\eta\tilde{Y}_{tm})}{\sum_{m=1}^M \exp(-\eta\tilde{S}_{t-1,m})} \\ &= \sum_{m=1}^M Q_{tm}\exp(-\eta\tilde{Y}_{tm}).\end{aligned}$$

$$\begin{aligned}\sum_{m=1}^M Q_{tm}\exp(-\eta\tilde{Y}_{tm}) &\le \sum_{m=1}^M Q_{tm}\left(1 - \eta\sum_{m=1}^M \tilde{Y}_{tm} + \frac{\eta^2}{2}\sum_{m=1}^M \tilde{Y}_{tm}^2\right) \\ &= 1 - \eta\sum_{m=1}^M Q_{tm}\tilde{Y}_{tm} + \frac{\eta^2}{2}\sum_{m=1}^M Q_{tm}\tilde{Y}_{tm}^2 \\ &\le \exp\left(-\eta\sum_{m=1}^M Q_{tm}\tilde{Y}_{tm} + \frac{\eta^2}{2}\sum_{m=1}^M Q_{tm}\tilde{Y}_{tm}^2\right).\end{aligned}$$

Thus, we have $\sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm} - \sum_{t=1}^{n}\tilde{Y}_{tm} \le \frac{\log M}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm}^2$.

$$\mathbb{E}_t\left[\tilde{Y}_{tm}^2\right] = \mathbb{E}_t\left[\left(\frac{E_{mA_t}^{(t)} y_{tA_t}}{P_{tA_t}+\gamma}\right)^2\right] \le \mathbb{E}_t\left[\left(\frac{E_{mA_t}^{(t)} y_{tA_t}}{P_{tA_t}}\right)^2\right]$$
$$= \sum_{i=1}^{k} P_{ti}\left(\frac{E_{mi}^{(t)} y_{ti}}{P_{ti}}\right)^2 \le \sum_{i=1}^{k}\frac{E_{mi}^{(t)}}{P_{ti}}.$$

The second inequality follows from $0 \le E_{mi}^{(t)} \le 1$ and $0 \le y_{ti} \le 1$.

$$\mathbb{E}\left[\sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm}^2\right] \le \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k}\sum_{m=1}^{M} Q_{tm}\frac{E_{mi}^{(t)}}{P_{ti}}\right] = nk.$$

Thus,

$$\sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm} - \sum_{t=1}^{n}\tilde{Y}_{tm} \le \frac{\log M}{\eta} + \frac{\eta nk}{2}.$$

□

To present the lemma, we introduce two key concepts. Let $\mathbb{F} = (\mathcal{F}_t)_{t=0}^{n}$ denote a filtration. A process $(Z_t)_{t=1}^{n}$ is said to be $\mathbb{F}$-adapted if, for every $t \in [n]$, the random variable Z_t is measurable with respect to $\mathcal{F}_t$. Moreover, $(Z_t)_{t=1}^{n}$ is called $\mathbb{F}$-predictable if, for each $t \in [n]$, the variable Z_t is measurable with respect to $\mathcal{F}_{t-1}$.

Lemma 2. *Let $\mathcal{F} = (\mathcal{F}_t)_{t=0}^{n}$ be a filtration and for $i \in [k]$, let $(\tilde{Y}_{ti})_t$ be $\mathcal{F}$-adapted, such that: 1. For $S \in [k]$ with $|S| \le 1$, $\mathbb{E}\left[\prod_{i\in S}\tilde{Y}_{ti}|\mathcal{F}_{t-1}\right] \le 0$. 2. For all $t \in [n]$ and $i \in [k]$, $\mathbb{E}\left[\tilde{Y}_{ti}|\mathcal{F}_{t-1}\right] = y_{ti}$.*
Let $(\alpha_{ti})_{ti}$ and $(\lambda_{ti})_{ti}$ be $\mathbb{F}$-predictable stochastic processes, satisfying for all t and i, $0 \le \alpha_{ti}\tilde{Y}_{ti} \le 2\lambda_{ti}$. Then, for any δ',

$$\mathbb{P}\left(\sum_{t=1}^{n}\sum_{i=1}^{k}\alpha_{ti}\left(\frac{\tilde{Y}_{ti}}{1+\lambda_{ti}} - y_{ti}\right) \le \log\left(\frac{1}{\delta'}\right)\right) \ge 1-\delta'.$$

Proof. For a detailed discussion and formal proof, refer to Lemma 12.2 in Bandit Algorithms by Lattimore and Szepesvári [27]. □

Define $Y_{tm} = \sum_{i=1}^{k} E_{mi}^{(t)} y_{ti}$, according to the above lemma, for the optimal expert m^*, we let $\alpha_{ti} = 2\gamma E_{m^*i}^{(t)}$, $\lambda_{ti} = \frac{\gamma}{P_{ti}}$, $\tilde{Y}_{ti} = \frac{A_{ti}y_{ti}}{P_{ti}}$,

$$\mathbb{P}\left(\sum_{t=1}^{n}\sum_{i=1}^{k} 2\gamma E_{m^*i}^{(t)}\left(\frac{1}{1+\frac{\gamma}{P_{ti}}}\frac{A_{ti}y_{ti}}{P_{ti}} - y_{ti}\right) \le \log\left(\frac{1}{\delta'}\right)\right) \ge 1-\delta'.$$

Therefore, with probability at least $1-\delta'$,

$$\mathbb{P}\left(\sum_{t=1}^{n}\tilde{Y}_{tm^*} - \sum_{t=1}^{n}Y_{tm^*} \le \frac{1}{2\gamma}\log\frac{1}{\delta'}\right) \ge 1-\delta'. \tag{5}$$

Similarly, for any $j \in [k]$, let $\alpha_{ti} = 2\gamma\mathbb{I}\{i=j\}$, $\lambda_{ti} = \frac{\gamma}{P_{ti}}$, $\tilde{Y}_{ti} = \frac{A_{ti}y_{ti}}{P_{ti}}$,

$$\mathbb{P}\left(\sum_{t=1}^{n}\sum_{i=1}^{k} 2\gamma\mathbb{I}\{i=j\}\left(\frac{1}{1+\frac{\gamma}{P_{ti}}}\frac{A_{ti}y_{ti}}{P_{ti}} - y_{ti}\right) \le \log\left(\frac{1}{\delta'}\right)\right) \ge 1-\delta'.$$

Simplifying, we obtain

$$\mathbb{P}\left(\sum_{t=1}^{n}\hat{Y}_{tj} - \sum_{t=1}^{n}y_{tj} \le \frac{1}{2\gamma}\log\frac{1}{\delta'}\right) \ge 1-\delta'. \tag{6}$$

By choosing $\delta' = \frac{\delta}{1+k}$, the probability that both Eqs. (5) and (6) hold simultaneously is greater than $1-\delta$.

Lemma 3. *Under the condition $1-\delta$,*

$$\sum_{t=1}^{n}Y_t - \sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm} \le \frac{k}{2}\log\frac{1}{\delta'} + \gamma nk.$$

Proof. Simplifying the left-hand side of the above expression:

$$
\begin{aligned}
\sum_{t=1}^{n} Y_t - \sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm} &= \sum_{t=1}^{n}\sum_{j=1}^{k} A_{tj}y_{tj} - \sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\sum_{j=1}^{k} E_{mj}^{(t)}\hat{Y}_{tj} \\
&= \sum_{t=1}^{n}\sum_{j=1}^{k} A_{tj}y_{tj} - \sum_{t=1}^{n}\sum_{j=1}^{k} \hat{Y}_{tj}\sum_{m=1}^{M} Q_{tm}E_{mj}^{(t)} \\
&= \sum_{t=1}^{n}\sum_{j=1}^{k} A_{tj}y_{tj} - \sum_{t=1}^{n}\sum_{j=1}^{k} P_{tj}\hat{Y}_{tj} \\
&= \sum_{t=1}^{n}\sum_{j=1}^{k} \left(A_{tj}y_{tj} - \frac{P_{tj}A_{tj}y_{tj}}{P_{tj}+\gamma} \right) \\
&= \gamma \sum_{t=1}^{n}\sum_{j=1}^{k} \hat{Y}_{tj}.
\end{aligned}
$$

Since under the condition $1-\delta$, for any $j \in [k]$, $\sum_{t=1}^{n} \hat{Y}_{tj} - \sum_{t=1}^{n} y_{tj} \le \frac{1}{2\gamma}\log\frac{1}{\delta'}$, we have

$$
\begin{aligned}
\gamma \sum_{t=1}^{n}\sum_{j=1}^{k} \hat{Y}_{tj} &\le \frac{k}{2}\log\frac{1}{\delta'} + \gamma\sum_{t=1}^{n}\sum_{j=1}^{k} y_{tj} \\
&\le \frac{k}{2}\log\frac{1}{\delta'} + \gamma nk.
\end{aligned}
$$

□

Finally, we prove Theorem 1,

Proof. Decompose R_n into three components and bound them separately,

$$
\begin{aligned}
R_n &= \sum_{t=1}^{n} Y_t - \sum_{t=1}^{n} Y_{tm^*} \\
&= \left(\sum_{t=1}^{n} Y_t - \sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm} \right) + \left(\sum_{t=1}^{n}\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm} - \sum_{t=1}^{n} \tilde{Y}_{tm^*} \right) + \left(\sum_{t=1}^{n} \tilde{Y}_{tm^*} - \sum_{t=1}^{n} Y_{tm^*} \right) \\
&\le \frac{k}{2}\log\frac{k+1}{\delta} + \gamma nk + \frac{\log M}{\eta} + \frac{\eta nk}{2} + \frac{1}{2\gamma}\log\frac{k+1}{\delta} \\
&= \frac{k}{2}\log\frac{k+1}{\delta} + \frac{\eta nk}{2} + \frac{\log M}{\eta} + \frac{\eta nk}{2} + \frac{1}{\eta}\log\frac{k+1}{\delta} \\
&= \frac{k}{2}\log\frac{k+1}{\delta} + \eta nk + \frac{1}{\eta}\left(\log M + \log\frac{k+1}{\delta} \right) \\
&= \frac{k}{2}\log\frac{k+1}{\delta} + 2\sqrt{nk(\log M + \log(k+1) - \log\delta)}.
\end{aligned}
$$

□

5. Numerical evaluation for Exp4-IX

This section presents the evaluation of the Exp4-IX algorithm applied to a UAV-target tracking task involving one UAV and a single dynamic target. We consider diverse trajectory patterns for the target and multiple expert-based flight strategies. The target's motion is examined under both smooth and adversarial conditions. Expert strategies include the Previous Position Algorithm, Particle Filter, Trajectory Fitting, and the Exp4-IX method itself. We assess their effectiveness in different tracking scenarios.

The simulation is conducted in a 100×100 km square area. Each time step corresponds to one second, and the full simulation spans 2000 seconds. The UAV moves in one of four cardinal directions at a constant speed of 280 m/s. Its sensing area is a rectangle extending 60 km ahead and 80 km in width. The target's speed varies between 100 and 280 m/s, and its trajectory is generated under two contrasting schemes:

- **Smooth Trajectory**: A smooth path is generated by fitting a curve through 20 planar points. The curve is further segmented to ensure inter-point distances remain within the target's velocity constraint. If the number of resulting points exceeds 2000, the excess is trimmed.
- **Adversarial Trajectory**: The target moves in one of four directions (forward, backward, left, or right), with adaptive changes designed to counteract the UAV's tracking path.

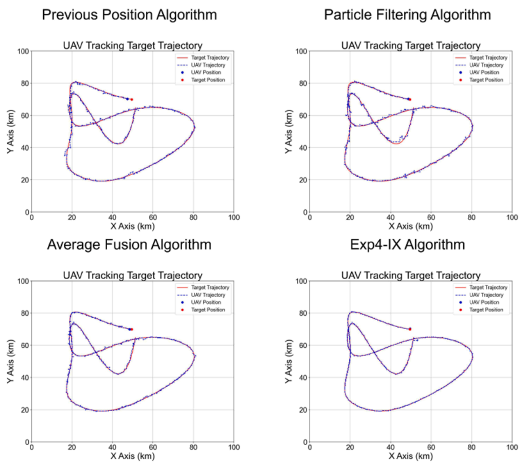

Fig. 2. Comparison of UAV tracking performance using different algorithms.

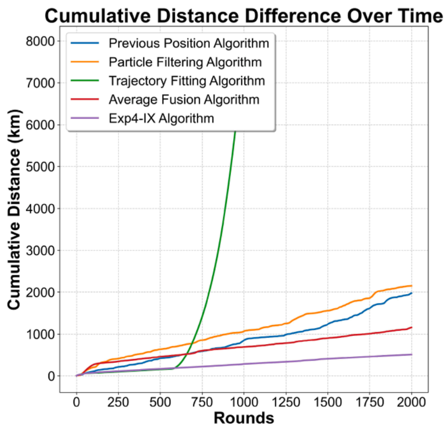

Fig. 3. Cumulative distances of UAV to the target using different tracking algorithms.

The following expert strategies are evaluated:

- **Previous Position Algorithm**: When the UAV observes the target in the previous step, it estimates the direction of movement based on their relative positions. Otherwise, it assigns equal probabilities to each direction to increase the likelihood of regaining visibility.
- **Particle Filtering Algorithm**: This method initializes particles representing target state hypotheses. It adds noise to propagate particles to the next timestep, determines action direction based on the predicted location, and updates particle weights based on new observations. Resampling is performed to maintain diversity.
- **Trajectory Fitting Algorithm**: The UAV uses past observations to model the target's trajectory and estimate its next position for guidance.
- **Average Fusion Algorithm**: This baseline integrates the three algorithms above by averaging their probability distributions.
- **Exp4-IX Algorithm**: An online fusion approach that dynamically reweights expert advice to prioritize the most effective strategy over time.

The implementation code is publicly available on the GitHub Repository.

All simulations were executed on a Windows laptop featuring an Intel Core i7-12700H CPU (2.30 GHz) and 32 GB of RAM, using Python 3.9.

5.1. Performance in smooth scenarios

In scenarios where the target moves smoothly and independently of the UAV's actions, tracking remains relatively stable with minor reward fluctuations. As shown in Fig. 2, the trajectory generated by the UAV using the Exp4-IX algorithm outperforms those of other methods. The Exp4-IX trajectory is noticeably smoother, as the UAV is able to more accurately predict the target's next position. As shown in Fig. 3, Exp4-IX achieves the highest accuracy, followed by the Average Fusion method, suggesting that intelligent fusion

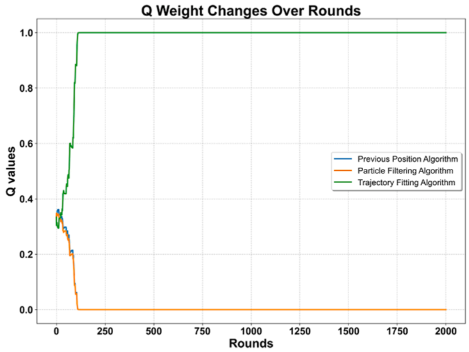

Fig. 4. Weight evolution of individual strategies within Exp4-IX.

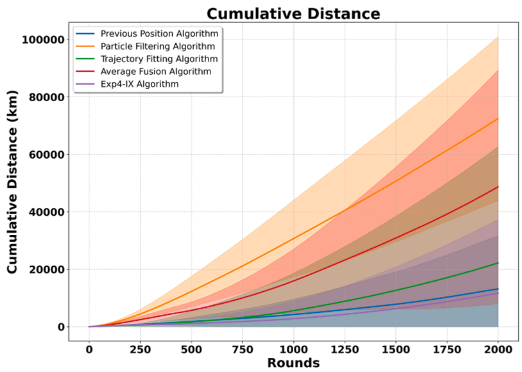

Fig. 5. Average tracking error across algorithms in adversarial scenarios.

outperforms individual strategies. Among the standalone approaches, the Previous Position algorithm yields better results than Particle Filtering, but still exhibits more oscillation than Exp4-IX. The Trajectory Fitting method performs the worst; when the target is lost, it predicts based solely on earlier positions, leading to cumulative drift.

While the Previous Position and Particle Filtering methods are more effective at keeping the target visible, the Trajectory Fitting algorithm excels in proximity. In Fig. 4, the rising weight of the Trajectory Fitting method implies it performs well when visibility is maintained. When the target is lost, Exp4-IX leverages other experts to recover tracking.

5.2. Performance in adversarial scenarios

In an adversarial setting, the target deliberately maneuvers to evade detection, such as by making sharp turns when approached. We conduct 100 runs under these conditions and average the cumulative distances to the target. As illustrated in Fig. 5, Exp4-IX, Previous Position, and Trajectory Fitting algorithms exhibit comparable performance, with Exp4-IX slightly outperforming the rest due to its adaptive fusion mechanism. The Average Fusion baseline, however, lags behind, demonstrating that static averaging is insufficient in such dynamic conditions. The Particle Filtering method also suffers due to sampling variability. Notably, the average tracking error exceeds 9 km, a significant jump from the 0.5 km error in smooth scenarios, highlighting the difficulty introduced by adversarial movement.

6. Conclusion

This study explores UAV-based target tracking under adversarial and partially observable conditions, and introduces the Exp4-IX algorithm for adaptive online fusion. The method offers both strong theoretical guarantees and empirical performance benefits. By aggregating multiple expert strategies, Exp4-IX enhances the UAV's ability to consistently maintain visual contact with dynamic targets.

Nevertheless, certain limitations persist. First, the reward model is based on rectangular sensing zones; alternative sensor geometries would require model adaptation. Second, in cases where an expert strategy is near-optimal, Exp4-IX may not outperform it, due to its inherent fusion overhead. Future work will focus on expanding this framework to collaborative multi-UAV systems with limited communication, aiming to track multiple targets more effectively.

CRediT authorship contribution statement

Yang Lv: Writing – original draft, Methodology; **Guochao Fan:** Validation, Supervision, Conceptualization; **Mengzhen Li:** Methodology, Formal analysis; **Xiongjun Liu:** Supervision, Conceptualization; **Pengqing Liu:** Methodology; **Yapu Zhang:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Beijing Natural Science Foundation (No. Z220004).

References

[1] Y. Lv, G. Fan, M. Li, X. Liu, P. Liu, Y. Zhang, UAV target tracking with bandit-based data fusion, in: International Conference on Combinatorial Optimization and Applications, 2024, pp. 278–286.
[2] G. Wang, F. Wang, J. Wang, M. Li, L. Gai, D. Xu, Collaborative target assignment problem for large-scale UAV swarm based on two-stage greedy auction algorithm, Aerosp. Sci. Technol. 149 (2024) 1–11.
[3] W. Xu, W. Liang, Z. Xu, J. Peng, D. Peng, T. Liu, X. Jia, S.K. Das, Approximation algorithms for the generalized team orienteering problem and its applications, IEEE/ACM Transact. Network. 29 (1) (2020) 176–189.
[4] Y. Khosiawan, I. Nielsen, A system of UAV application in indoor environment, Product. Manufact. Res. 4 (1) (2016) 2–22.
[5] M.H.M. Saad, N.M. Hamdan, M.R. Sarker, State of the art of urban smart vertical farming automation system: advanced topologies, issues and recommendations, Electronics 10 (2021) 1–40.
[6] Y. Lv, C. Wu, D. Xu, R. Yang, H-hop independently submodular maximization problem with curvature, High-Confiden. Comput. 4 (2024) 1–7.
[7] C. Luo, N. Liu, Y. Hou, Y. Hong, Z. Chen, D. Li, Trajectory optimization of laser-charged UAV to minimize the average age of information for wireless rechargeable sensor network, Theoret. Comput. Sci. 945 (2023) 113680.
[8] S. Ojha, S. Sakhare, Image processing techniques for object tracking in video surveillance-a survey, in: International Conference on Pervasive Computing, IEEE, St Louis, 2015, pp. 1–6.
[9] H. Wang, H. Du, Time sensitive sweep coverage with minimum UAVs, Theoret. Comput. Sci., 928, 2022 197–209.
[10] Y.C. Lai, Z.Y. Huang, Detection of a moving UAV based on deep learning-based distance estimation, Remote Sens. 12 (18) (2020) 1–28.
[11] S. Bhagat, P.B. Sujit, UAV target tracking in urban environments using deep reinforcement learning, in: International Conference on Unmanned Aircraft Systems, IEEE, Athens, 2020, pp. 694–701.
[12] S. Bhagat, P.B. Sujit, Target-Following Double Deep Q-Networks for UAVs, 2021, arXiv preprint arXiv:2105.05464.
[13] J. Zhao, H. Liu, J. Sun, K. Wu, Z. Cai, Y. Ma, Y. Wang, Deep reinforcement learning-based end-to-end control for UAV dynamic target tracking, Biomimetics 7 (4) (2022) 197.
[14] C. Jin, T. Jin, H. Luo, S. Sra, T. Yu, Learning adversarial markov decision processes with bandit feedback and unknown transition, in: International Conference on Machine Learning, PMLR, 2020, pp. 4860–4869.
[15] Z. Xu, X. Lin, V. Tzoumas, Bandit submodular maximization for multi-robot coordination in unpredictable and partially observable environments, 2023, arXiv preprint arXiv:2305.12795.
[16] M. Herbster, M.K. Warmuth, Tracking the best expert, Mach. Learn. 32 (2) (1998) 151–178.
[17] P. Landgren, V. Srivastava, N.E. Leonard, Distributed cooperative decision making in multi-agent multi-armed bandits, Automatica 125 (2021) 1–13.
[18] K.M.B. Lee, F. Kong, R. Cannizzaro, J.L. Palmer, D. Johnson, C. Yoo, R. Fitch, An upper confidence bound for simultaneous exploration and exploitation in heterogeneous multi-robot systems, in: International Conference on Robotics and Automation, IEEE, 2021, pp. 8685–8691.
[19] C. Zhang, S.C. Hoi, Partially observable multi-sensor sequential change detection: a combinatorial multi-armed bandit approach, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 5733–5740.
[20] G. Neu, Explore no more: improved high-probability regret bounds for non-stochastic bandits, in: Advances in Neural Information Processing Systems, NIPS, Montréal, 2015, pp. 1–9.
[21] P. Auer, Cesa-Bianchi, N. Freund, Y. Schapire, E. Robert, The nonstochastic multiarmed bandit problem, SIAM J. Comput. 2002, pp. 48–77.
[22] Y. Seldin, G. Lugosi, A lower bound for multi-armed bandits with expert advice, 2016, arXiv preprint arXiv:1602.02553.
[23] N. Khial, N. Mhaisen, M. Mabrok, A. Mohamed, An online learning framework for UAV search mission in adversarial environments, Expert System. Applic. 267 (1) (2025) 126136.
[24] B. Baidya, A. Mondal, S. Manna, G. Das, A. Santra, A. Chakraborty, Enhanced UAV tracking through multi-Sensor fusion and extended kalman filtering, in: Sixth Doctoral Symposium on Intelligence Enabled Research, Sukanta Mahavidyalaya, Dhupguri, India, 2024.
[25] V. Masalskyi, A. Dzedzickis, I. Korobiichuk, V. Bucinskas, Hybrid mode sensor fusion for accurate robot positioning, Sensors 25 (10) (2025) 3008.
[26] K. Dabbabi, T. Delleji, Graph neural network-tracker: a graph neural network-based multi-sensor fusion framework for robust unmanned aerial vehicle tracking. visual computing for industry, Biomed. Art 8 (1) (2025) 18.
[27] T. Lattimore, C. Szepesvári, Bandit Algorithms, Cambridge University Press, 2020.