Neurocomputing 544 (2023) 126281

Contents lists available at ScienceDirect

Neurocomputing

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

AID-RL: Active information-directed reinforcement learning for autonomous source seeking and estimation ☆

Zhongguo Li [a], Wen-Hua Chen [b,*], Jun Yang [b], Yunda Yan [c]

[a] Department of Computer Science, University College London, London WC1E 6BT, UK
[b] Department of Aeronautical and Automotive Engineering, Loughborough University, Loughborough LE11 3TU, UK
[c] School of Engineering and Sustainable Development, De Montfort University, LeicesterLE1 9BH, UK

ARTICLE INFO

Article history:
Received 3 January 2023
Revised 13 February 2023
Accepted 22 April 2023
Available online 3 May 2023

Keyword:
Autonomous search
Reinforcement learning
Dual control
Active learning
Exploration
Exploitation

ABSTRACT

This paper proposes an active information-directed reinforcement learning (AID-RL) framework for autonomous source seeking and estimation problem. Source seeking requires the search agent to move towards the true source, and source estimation demands the agent to maintain and update its knowledge regarding the source properties such as release rate and source position. These two objectives give rise to the newly developed framework, namely, dual control for exploration and exploitation. In this paper, the greedy RL forms an exploitation search strategy that navigates the agent to the source position, while the information-directed search commands the agent to explore most informative positions to reduce belief uncertainty. Extensive results are presented using a high-fidelity dataset for autonomous search, which validates the effectiveness of the proposed AID-RL and highlights the importance of active exploration in improving sampling efficiency and search performance.

1. Introduction

Autonomous search is to use mobile platforms equipped with physical sensors to localise and estimate a possible release of chemical, biological or radioactive substance. The mission of autonomous search for an airborne release comprises dual objectives: moving the search agent towards the source location (source seeking) and estimating the source properties (source term estimation). The importance of autonomous search has been manifested in its wide applications, including search and rescue, safety monitoring and emergency responses [1]. To realise fully autonomous operation, extensive path planning and estimation approaches have been established in recent years, which can be roughly classified into three categories: informative path planning (IPP) [2–4], bio-inspired mechanisms [5,6] and dual control methods [7,8]. The aforementioned approaches are non-episodic, i.e., the search process is not repetitive.

Benefiting from extensive interactions with the unknown environment, reinforcement learning has achieved remarkable success in playing complex games [9] and virtual robotic systems [10]. However, its applicability in real world applications remains quite limited, mainly owing to its poor sampling efficiency. Recently, reinforcement learning based approaches have been introduced to deal with source search and estimation [11,12]. A deep Q network with particle filter assisted source term estimation approach is developed in [12]. Deep deterministic policy gradient (DDPG) is used to train the optimal policy together with particle filter and Gaussian mixture model for source term approximation in [11]. Extensive simulation results have been reported in comparison with several benchmark algorithms including Entrotaxis [13] and Infotaxis [3]. In both studies, random exploration mechanisms are employed for probing the spaces regardless of the search and estimation performance, where standard ϵ-greedy algorithm is utilised in [12] and random noise perturbation is added to the policy output in [11]. Despite their exploratory efforts – inefficient random exploration, it should be noted that the source term estimation is passively updated in the sense that the estimation performance is not integrated in the decision-making processes of the RL algorithms.

Balancing between exploitation and exploration has been a long-standing issue in RL [14,15], particularly when observations obtained from the environment are uncertain and noisy. This is

☆ This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Established Career Fellowship "Goal-Oriented Control Systems: Disturbance, Uncertainty and Constraints" under the grant number EP/T005734/1. Corresponding author: Wen-Hua Chen.

* Corresponding author.
E-mail addresses: zhongguo.li@ucl.ac.uk (Z. Li), w.chen@lboro.ac.uk (W.-H. Chen), j.yang3@lboro.ac.uk (J. Yang), yunda.yan@dmu.ac.uk (Y. Yan).

https://doi.org/10.1016/j.neucom.2023.126281
0925-2312/

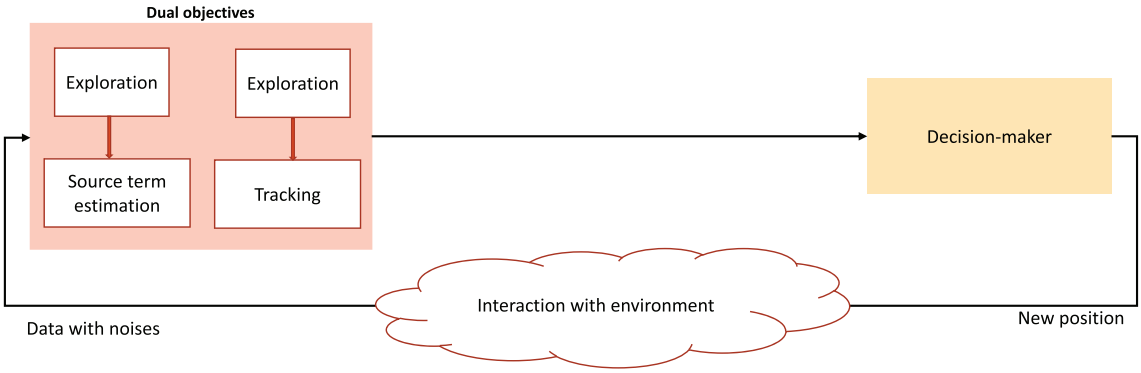


Fig. 1. Dual objectives in autonomous search and estimation.

the classic dilemma of RL algorithms: should the search agent maximise its reward based on its current knowledge or explore poorly understood states and actions to potentially improve future performance [16]. In autonomous search problem, the search agent inherently suffers heavily from both sensor noises and local turbulence disturbances such that concentration information collected is often sparse and noisy. In this challenging scenario, effective exploration mechanism is of vital importance to ensure mission success [15]. The agent using greedy RL algorithms is in fact driven to move towards the source location, which only accounts for one of the dual objectives in an autonomous search mission [7,17,18]. In order to improve the sampling efficiency and enhance the estimation performance, more delicate exploratory approaches are required.

In the last few years, active reinforcement learning, also termed as Bayesian reinforcement learning, has emerged as one of the hottest research areas in machine learning community driven by its prominent capability in improving data efficiency and enhancing learning performance [14,19–21]. To incentivise directed exploration over poorly understood states and actions, it is important to quantify the uncertainty of the agent's belief about its operational environment. Houthooft et al. [22] suggest Bayesian neural networks for uncertainty measure and develop a variational information maximising exploration strategy. Shyam et al. [23] advocate an ensemble-based approach where uncertainty is measured by the amount of conflict among the predictions of the constituent models, which is proved to be more efficient and has been widely used in recent studies [8,20]. In the control system community, Chen [24] establishes a similar paradigm from control-theoretic perspective, namely Dual Control for Exploitation and Exploration (DCEE), which holds high learning efficiency for autonomous control in an uncertain and unknown environment. Both approaches are targeted to provide a solution framework to deal with autonomous decision-making problems in an uncertain environment, and rather coincidentally they both emphasise the importance of active exploration for constructing knowledge regarding the operational environment.

Motivated by the above observations, this paper develops an active information-directed reinforcement learning (AID-RL) to solve the autonomous search problem. To our knowledge, this is the first attempt to deploy reinforcement learning with active exploration for autonomous search and estimation problem. While there has been a significant amount of research on active RL algorithms, such as [23,22,20,15], they are not directly applicable to autonomous seeking and estimation. Most of them utilise neural networks or ensembles of randomly generated dynamic models to acquire information about the environment, which makes it challenging to extract physically meaningful parameters for source term estimation. The exploration mechanism in this paper is originated from classic information-theoretic path planning methods where Infotaxis [3,25] and Entrotaxis [13] are regarded as two benchmark representatives. Essentially, the conventional approach of exploitative RL drives the search agent towards the source location by rewarding it for collecting high concentration values (reward-driven exploitation), while the information-directed exploration aims to lead the search agent to positions that can reduce its belief uncertainty regarding the environment by maximising the information gain. The dual objectives of autonomous search are depicted in Fig. 1, where their connections with exploration and exploitation are also outlined. This information-directed exploratory RL method is partially inspired by the dual control algorithm [7,26]. Compared with the benchmark ϵ-greedy RL algorithm, we demonstrate that the proposed AID-RL not only produces better search and estimation performance but also maintains high sampling efficiency using less training episodes.

The key contributions of this paper are summarised as follows.

1. This paper provides a unified formulation for autonomous search problem using information-directed RL. The proposed framework, AID-RL, unifies greedy reward-driven exploitation and information-directed exploration in autonomous search.
2. An active reinforcement learning algorithm is developed, which combines a reformulated greedy Q learning algorithm and an Entrotaxis based exploration strategy. Such a new exploration mechanism, originated from traditional IPP, can significantly improve the sampling efficiency and search performance. Different from existing studies in active RL [22,23,20] where Bayesian neural networks or ensembles of dynamic models are employed to formulate information gain, the proposed AID-RL captures belief uncertainty using particle filter, from which meaningful source parameters can be extracted.
3. The proposed algorithm is implemented on a real dataset, consisting of sparse and uncertain measurements. Currently, the existing algorithms using RL only validated their effectiveness through numerical simulations, which did not take measurement uncertainty and reward sparseness into consideration. In our high-fidelity studies, efficacy of RL for autonomous search is manifested and the advantages of the proposed AID-RL are demonstrated by comparing with classic ϵ-greedy RL algorithms [12].

The rest of this paper is organised as follows. Section 2 presents the formulation of autonomous search and estimation. In Section 3, an active reinforcement learning algorithm is developed combining reward-driven exploitation and information-directed exploration. Section 4 presents the experimental results using high-fidelity dataset, and provides detailed discussions and comparisons

with random-exploration RL algorithm. Conclusions are drawn in Section 5.

2. Problem formulation

In this section, we elaborate the key functionalities in a source search and estimation problem including the search agent and dispersion modelling, reward function formulation and Bayesian inference for source term estimation. In particular, an estimation algorithm using particle filter will be implemented to achieve source estimation, based on which information-directed reinforcement learning will be designed later in Section 3.

2.1. Agent and dispersion modelling

Autonomous search algorithm is to direct a robotic searcher, equipped with onboard sensors, to locate and estimate an airborne source from a point release. The source release is usually characterised by $\Theta_s = [q_s, p_s]^\mathsf{T}$, where $q_s > 0$ denotes a positive release rate and $p_s = [x_s, y_s]^\mathsf{T}$ represents the source position in a search domain $\Omega \subset \mathbb{R}^2$. A search agent, located at $p_k = [x_k, y_k]^\mathsf{T}$, navigates its search path by choosing actions from an admissible set $\mathscr{A} = \{\rightarrow, \leftarrow, \uparrow, \downarrow\}$ according to pre-loaded search algorithm with information collected from chemical/biological sensors. Autonomous search is mainly concerned with the high-level decision-making for path planning. It is assumed that the search agent has been programmed with low-level controller that can achieve the given actions in $\mathscr{A}$.

Dispersion models are used to describe the airborne transport and diffusion of released materials. In this paper, the convection–diffusion plume is adopted to reconstruct the source dispersion, which has been widely utilised in related studies [3,13,11,12,7]. The expected concentration at agent position p_k is given by

$$z_k(p_k|\Theta_s) = \frac{q_s}{4\pi D|p_k - p_s|} \exp\left(\frac{-\Delta p_k V}{2D}\right) \exp\left(\frac{-|p_k - p_s|}{\mu}\right) \tag{1}$$

where $\Delta p_k = -(x_k - x_s)\sin(\psi) + (y_k - y_s)\cos(\psi)$, $\mu = \sqrt{\frac{D\tau}{1+\frac{V^2\tau}{4D}}}$, V is the wind speed and ψ is the wind direction, D represents the isotropic diffusivity and τ denotes the particle lifetime.

It is crystal clear that the collected concentration information is highly uncertain due to local turbulence and sensor noises, and consequently there is usually high discrepancy between the sensor readings and the modelled output from $z_k(p_k|\Theta_s)$. Nevertheless, it has been proven that such a dispersion model is of great importance to achieve fast localisation and efficient source term acquisition [8,7,13]. The dispersion model used in this paper has physically meaningful parameters that can be interpreted in a practical way. Learning these parameters is one of the key objectives in autonomous search. Existing active RL algorithms [22,23,20] using probabilistic predictive models or dynamic ensemble models cannot be utilised for source term estimation because they lack physical meanings.

2.2. Reward function

From the dispersion model, the expected sensor reading will be higher when the search agent moves closer to the source location. Therefore, the concentration value from sensor can naturally be used to formulate the reward function, which serves as the incentives to promote the search agent moving towards the source. It should be noted that the expected concentration value is exponentially decaying with respect to the distance between the search agent and the source location. In this paper, we take the logarithmic value of the concentration value as the reward function, i.e.,

$$R(p_k, a_k) = \log(z_{k+1}(p_{k+1}) + \varepsilon) \tag{2}$$

where $z_{k+1}(p_{k+1})$ denotes the sensor reading at future agent position p_{k+1}, and $\varepsilon > 0$ is a small positive number to avoid ill-conditioned logarithms. Future reward is determined by the current location p_k and current action $a_k \in \mathscr{A}$. Because sensor reading is usually sparse, i.e., no meaningful detection, ε is included in (2) to deal with zero concentration measurements. As having been widely discussed in related works [27,7,13], non-detection events occur quite often due to several reasons, for example, local turbulence, sensor sensitivity and failure. This type of information sparseness is reflected by a lower reward in reinforcement learning. Note that other reward designs, such as the stage-wise function used in [11], exist as well.

The objective of the search agent is to maximise its cumulative reward function over an infinite horizon, denoted by

$$J^\infty = \sum_{i=1}^{\infty} \lambda^i R(p_i, a_i) \tag{3}$$

where $0 < \lambda \leqslant 1$ is the discount factor.

2.3. Source Term Estimation

Given a set of historical sensor readings $\mathscr{Z}_{k-1} := \{z_1(p_1), z_2(p_2), \ldots, z_{k-1}(p_{k-1})\}$, the posterior probability of the environment parameter Θ_k can be approximated by recursive Bayesian estimation

$$P(\Theta_k|\mathscr{Z}_k) = \frac{P(z_k|\Theta_k)P(\Theta_k|\mathscr{Z}_{k-1})}{P(z_k|\mathscr{Z}_{k-1})} \tag{4}$$

where

$$P(z_k|\mathscr{Z}_{k-1}) = \int P(z_k|\Theta_k)P(\Theta_k|\mathscr{Z}_{k-1})d\Theta_k \tag{5}$$

and initial condition is specified as $P(\Theta_0|\mathscr{Z}_0) = P(\Theta_0)$. In the Bayesian estimation process, $P(\Theta_k|\mathscr{Z}_{k-1})$represents the prior distribution and the likelihood function $P(z_k|\Theta_k)$ is determined by the dispersion model (1).

A variety of source term estimation techniques have been developed in the literature, including gradient based methods [8], Gaussian mixture model [11] and particle filters [13]. A comprehensive review on source term estimation algorithms can be found in [28]. Among them, particle filters have been increasingly popular due to its flexibility and effectiveness in handling high uncertain and nonlinear estimation problems as in the case of autonomous search. As a result, they are widely employed to construct the nonlinear inference engine [7,13,2,27]. The posterior distribution of the environment parameter can be approximated by a set of N weighted samples $\{\Theta_k^{(i)}, \omega_k^{(i)}\}_{i=1}^N$ such that

$$P(\Theta_k|\mathscr{Z}_k) \approx \sum_{i=1}^{N} \omega_k^{(i)} \delta\left(\Theta_k - \Theta_k^{(i)}\right) \tag{6}$$

where $\delta(\cdot)$ denotes Dirac delta function, $\Theta_k^{(i)}$ is a potential realisation of environment parameter at the kth step, and $\omega_k^{(i)}$ represents the corresponding normalised weight of the particles with $\sum_{i=1}^N \omega_k^{(i)} = 1$. The implementation structure of particle filter for environment estimation is summarised in Algorithm 1, and more detailed elaborations can be found in [27,7].

Algorithm 1: Particle filter for environment parameter estimation.

Require: prior samples $\{\Theta_{k-1}^{(i)}, \omega_{k-1}^{(i)}\}_{i=1}^{N}$

1. **for** $i = 1, 2, \ldots, N$, **do**
2. draw sample $\Theta_k^{(i)} \sim q(\Theta_{k-1}^{(i)})$
3. assign weight $\bar{\omega}_k^{(i)} = \omega_{k-1}^{(i)} \cdot \frac{P(z_k|\Theta_k^{(i)}) P(\Theta_k^{(i)}|\Theta_{k-1}^{(i)})}{q(\Theta_k^{(i)}|\Theta_{k-1}^{(i)}, \mathscr{Z}_k)}$
4. **end for**
5. normalise sample weights $\omega_k^{(i)} = \frac{\bar{\omega}_k^{(i)}}{\sum_{i=1}^{N} \bar{\omega}_k^{(i)}}$
6. calculate effective sample size $N_{eff} = 1 / \sum_{i=1}^{N} \left(w_k^{(i)}\right)^2$
7. **if** N_{eff} is less than a threshold N_T **then**
8. resample $\{\Theta_k^{(i)}, \omega_k^{(i)}\}_{i=1}^{N}$
9. apply a Markov chain Monte Carlo move
10. **end if**

Ensure: posterior samples $\{\Theta_k^{(i)}, \omega_k^{(i)}\}_{i=1}^{N}$

3. Active reinforcement learning

In Section 2.2, the reward function is formulated according to the sensor reading collected from the environment. Targeting to collect high reward, the search agent will be directed towards the source position where higher concentration is expected. This corresponds to one of the dual objectives, i.e., source seeking. On the other hand, active exploration can be integrated into this process to improve the source estimation performance, which is another critical objective in autonomous search and estimation. The dual objectives will then motivate our proposed framework of AID-RL in Section 3.3.

3.1. Reward-driven exploitation

In this paper, we implement the state-action iteration based Q-learning algorithm to solve the autonomous search problem. The Q value function is denoted by

$$Q(p_k, a_k) = R(p_k, a_k) + \max_{a_{k+1} \in \mathscr{A}} \lambda J^{\infty}(p_{k+1}, a_{k+1}) \tag{7}$$

and the Bellman optimality condition is given by

$$Q^*(p_k, a_k) = R(p_k, a_k) + \max_{a_{k+1} \in \mathscr{A}} \lambda Q(p_{k+1}, a_{k+1}). \tag{8}$$

The value iteration algorithm can be designed as

$$Q'(p_k, a_k) = Q(p_k, a_k) + \alpha \left[R(p_k, a_k) + \lambda \max_{a_{k+1} \in \mathscr{A}} Q(p_{k+1}, a_{k+1}) - Q(p_k, a_k) \right] \tag{9}$$

where $0 < \alpha \leqslant 1$ is the learning rate. In viewing of the above value iteration algorithm, the control action a_k is chosen to maximise the cumulative reward function J^{∞}. If we have an ideal Q table, pure exploitative actions will generate maximal cumulative reward. However, in real situation, the Q values are usually randomly initialised and the collected rewards during the search process are highly uncertain due to local turbulence and noises. As a result, exploiting untrustworthy Q table will deteriorate the overall search performance. Therefore, exploration efforts should be included while choosing the control action, and one of the most commonly-used mechanisms is so-called ϵ-greedy algorithm [11,12] that randomly chooses an action from the admissible set with probability $1 - \epsilon$. This type of random perturbation based RL algorithm is termed as undirected exploration.

The essence of the iteration algorithm is to update the Q table using information collected from interactions between the agent and its operational environment. The reward-driven exploitation is a model-free algorithm as in (9). Despite its wide success in robotic control without models, it is prohibitively expensive in sampling complexity.

3.2. Information-directed exploration

Reinforcement learning has been widely criticised in control and robotic society due to its poor sampling efficiency. It is undeniably true that real robotic systems cannot be implemented for hundreds and thousands trial-and-error attempts due to physical constraints, time and energy consumption and other safety issues. In recent years, significant research effort has been dedicated to improving data efficiency of RL using active learning and model based approaches [20,14]. Instead of random exploration, it has been demonstrated in recent works such as [20,14] that active exploration based RL yields outstanding performance compared with passive learning in the machine learning community. Similarly, active learning based control approaches have also emerged as a promising paradigm in control community [24,29].

In autonomous search, it is required that the search agent reconstructs source parameters. From the perspective of dual control [7], the exploration strategy is to reduce knowledge uncertainty by directing the search agent to probe the most informative positions. Informative path planning (including Infotaxis [3] and Entrotaxis [13]) is one of the mainstreams, which has proven to be very robust and effective in localising source position and reconstructing source parameters. In this paper, we deploy Entrotaxis as the exploration algorithm, which will guide the searcher to where there is the most uncertainty in the next measurement.

The information measure is defined according to Shannon entropy

$$I(a_k) = -\int P(\hat{z}_{k+1}(\hat{p}_{k+1})|\mathscr{Z}_k) \log P(\hat{z}_{k+1}(\hat{p}_{k+1})|\mathscr{Z}_k) d\hat{z}_{k+1} \tag{10}$$

where $\hat{z}_{k+1}(\hat{p}_{k+1})$ represents the predicted measurement at future agent position $\hat{p}_{k+1}$ if action a_k is taken. Given the current estimation of the source parameters, the probability density function of the future measurement can be obtained by

$$P(\hat{z}_{k+1}(\hat{p}_{k+1})|\mathscr{Z}_k) = \int P(\hat{z}_{k+1}(\hat{p}_{k+1})|\Theta_k) P(\Theta_k|\mathscr{Z}_k) d\Theta_k. \tag{11}$$

The approximation strategy using particle filter in Algorithm 1 has been well-documented in [13,30]. The exploration strategy is to maximise the information gain by choosing a control action from $a_k \in \mathscr{A}$, i.e.,

$$a_k^* = \arg\max_{a_k \in \mathscr{A}} \{I(a_k)\}. \tag{12}$$

The information gain is approximated by

$$I(a_k) \approx \sum_{\hat{z}_{k+1}=0}^{\hat{z}^{\max}} P(\hat{z}_{k+1}(\hat{p}_{k+1})|\mathscr{Z}_k) \log P(\hat{z}_{k+1}(\hat{p}_{k+1})|\mathscr{Z}_k) \tag{13}$$

where $\hat{z}_{k+1} = \{0, 1, 2, \ldots, \hat{z}^{\max}\}$ denotes the potential future measurements. Note that the expected information gain $I(a_k)$ is calculated for all possible actions in the action set $\mathscr{A}$. Then, (12) is solved by selecting a control action that yields the maximum information gain.

There are a variety of informative measures to quantify knowledge uncertainties, for example, variance, mutual information and Kullback–Leibler divergence [31]. By integrating knowledge uncertainty into the decision process, an information-directed mechanism for active exploration is achieved such that the probing actions are inserted to explore the most promising and informative direction instead of random search. Evidently, the inference engine using Bayesian approximation requires a dispersion model as in (1)

and (4). Hence, the formulation of information-directed exploration is a model-based estimation approach. It is argued that encapsulating model-based knowledge in RL can greatly accelerate the learning speed and achieve high sample efficiency [32].

3.3. AID-RL: Active information-directed reinforcement learning

From the perspective of dual control [17], the formulation of exploitative RL is in fact a control-driven algorithm, which aims to navigate the search agent to the believed source location that yields higher reward. The current belief of the possible source location is encapsulated implicitly in the state-action Q values. The success of RL algorithm heavily relies on the exploration mechanism, and all current studies in autonomous search deploy random exploration strategy, i.e., ϵ-greedy RL algorithm [11,12]. Considering that the efficiency of random exploration is usually quite poor in many real applications, we introduce the classic informative path planning methods to improve sampling efficiency. This new algorithm, namely active information-directed reinforcement learning, AID-RL, is partially motivated by the paradigm of recently developed DCEE in autonomous search [24].

The overall implementation structure of AID-RL is summarised in Algorithm 2. It is composed of an initialisation procedure and an episodic learning process. Each learning episode should be viewed as a trial of search mission, and thus the search agent is required to make sequential decisions from a randomly-initialised start positions. During each trial, the decision-making process balances between exploration and exploitation by using either information-directed search or reward-driven control. In existing benchmark algorithms, e.g., IPP [13,3] and DCEE [7], the search space is represented by grid map, which is also well-fitted to the classic Q learning. To achieve fair comparison, we will keep this classic setting for autonomous search. It is worth mentioning that the proposed AID-RL is ready to be tailored to deal with large-scale search problem by using neural network approximation [12,11].

Algorithm 2: AID-RL: Active information-directed reinforcement learning for autonomous search.

1. Q values of state-action $Q(p_0, a_0)$
2. prior knowledge of source $\{\Theta_0^{(i)}, \omega_0^{(i)}\}_{i=1}^N$
3. initialise learning hyper-parameters

Episodic learning:

4. **For** episode = $1 : M$ **do**
5. randomly initialise agent location p_0
6. **For** $k = 1 :$ MaxIt **do**
7. **If** random value $< \epsilon$, then choose a control action by information-directed exploration: $a_k^* = \arg\max_{a_k \in \mathscr{A}} \{I(a_k)\}$
8. **Else** choose a control action by reward-driven exploitation: $a_k^* = \arg\max_{a_k \in \mathscr{A}} \{Q(p_k, a_k)\}$
9. **End if**
10. execute action a_k^*
11. collect reward $R_k(p_k, a_k)$ at p_{k+1}
12. update the particle filter using Algorithm 1
13. update Q table by value iteration: $Q'(p_k, a_k) = Q(p_k, a_k) + \alpha[R(p_k, a_k) + \lambda \max_{a_{k+1} \in \mathscr{A}} Q(p_{k+1}, a_{k+1}) - Q(p_k, a_k)]$
14. **If** terminal condition is satisfied, **break**
15. **End for**
16. **End for**

Compared with the classic ϵ-greedy RL algorithm, the proposed AID-RL is of similar learning structure, except that AID-RL replaces the random exploration mechanism with an active information-directed search algorithm. It is noticed that the random exploration mechanism usually leads to low sampling efficiency and thus consumes a large amount of training episodes. Recently, significant research efforts have been dedicated to developing efficient exploration strategy, yet there is currently no consensus in the design of exploration techniques [15,20,22,23]. In this paper, the exploration search strategy is based on maximum entropy sampling principle by selecting the manoeuvre actions that navigate the agent to the most uncertain positions to reduce its knowledge uncertainty aggregated by particle filters. None of the aforementioned works in active RL have implemented such particle filter assisted exploration. It is necessary to use dispersion models that have physical meanings in order to meet the practical requirements of autonomous search, i.e., source term estimation.

In Algorithm 2, the hyper-parameter $\epsilon \in (0, 1)$ plays a central role in balancing between exploration and exploitation. In essence, by varying the value of ϵ, the search agent will alter between information-directed exploration and reward-driven exploitation with a probability determined by ϵ. A large value of ϵ means the search agent spends more efforts in probing the search space in order to reduce its belief uncertainty regarding the environment, while a small value of ϵ emphasises on the exploitation of its current Q table to move closer to its believed source position. It is worth mentioning that the tuning principle in ϵ-greedy RL can be applied to our proposed AID-RL. For example, the value of ϵ can be set as a large value initially to enable the search agent have more opportunity to probe the environment since the Q table is not trustworthy at the early stage of training. Then, after the initial training period, the value of ϵ can be decreased to let the agent make use of its belief.

AID-RL is a hybrid approach that combines model-free greedy RL and model-based source term estimation. Model-free RL provides an effective way to capture to the features of autonomous search by using state-action Q table. On the other hand, the model-based estimation technique enables fast adaptation of the source knowledge and provides an uncertainty measure to develop our information-aware RL algorithm. Recently, remarkable empirical results have been reported in many studies to demonstrate the combined strength of model-based and model-free (MB-MF) algorithms, for example, [10,14,20].

Table 1 summarises key features and differences of AID-RL compared with existing autonomous search algorithms. RL-based algorithms rely on episodic path sampling from randomly-initialised start points, and thereby are fundamentally different from those classic methods, such as bio-inspired search, IPP and DCEE. The concept of DCEE provides a new perspective to elucidate the dual objectives in source tracking and estimation: one is related to exploitation and another is linked with exploration [7,17]. Information-directed exploration dictates the search agent to probe most uncertain locations, and consequently improves the estimation performance due to accelerated information acquisition. Incorporating this information-aware exploration mechanism into the greedy RL (essentially, greedy RL aims to seek the source position where there is maximum reward), a balanced trade-off between exploration and exploitation is achieved. Additionally, it is important to mention that if the RL component is removed, the proposed AID-RL approach reverts back to Entrotaxis, which is a well-known IPP method.

4. Experimental results and discussions

In this section, we validate the effectiveness of the proposed AID-RL using a challenging experimental dataset, which was col-

Table 1
Characteristic comparison of different autonomous search algorithms.

Algorithm	Reference	Objective	Exploration mechanism	Sampling mechanism
Bio-inspired methods	[6,5]	tracking	none	non-episodic
IPP	[25,13,3]	estimation	information-directed exploration	non-episodic
DCEE	[7,8]	tracking and estimation	information-directed exploration	non-episodic
ϵ-greedy RL	[11,12]	tracking and estimation	random exploration	episodic
AID-RL	this paper	tracking and estimation	information-directed exploration	episodic

Table 2
Key parameters for AID-RL.

Parameter	Value	Description
Episodes	5,000	M, number of total trials
Maximum path	1,000	MaxIt, maximum path length
Discount	0.9	λ, discount for future reward
Learning rate	0.01	α, learning step per iteration
Exploration	0.2	ϵ, rate of exploration
Particle filter	1,000	N, number of particles

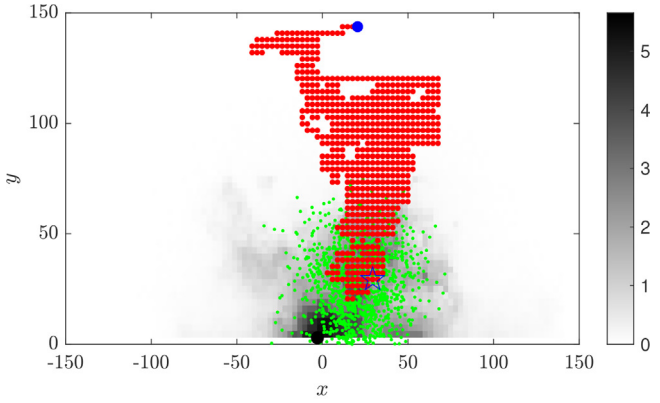
lected by COANDA Research and Development Corporation, and supplied by the DST group [25]. To demonstrate the advantages of AID-RL, we compare the proposed algorithm with the random exploration RL algorithm, i.e., ϵ-greedy learning.

The dataset contains a total number of 340 sequential frames, and each of them is composed of 49×98 point-wise measurements over the entire search space. Therefore, the search space is represented by a map of 49 rows and 98 columns, and each cell corresponds to the square area of 2.935×2.935 mm^2. Although the frames are sampled sequentially with a sampling time $t = 0.435s$, the dispersion filed changes significantly from one sample to another. In real airborne source search, it is exactly the case that the dispersion and measurements change rapidly due to local turbulence and sensor noises. Key parameters of the proposed AID-RL are summarised in Table 2.

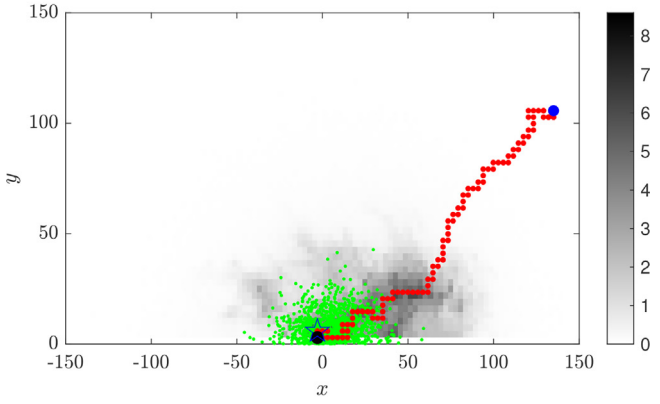
To show the search behaviour of the agent, we present two representative patterns at the beginning and at the end of training process, as shown in Fig. 2. The example path shown in Fig. 2a is sampled at the first episode. This search trial is classified as a failure search because the search agent fails to locate the true source after reaching the maximum path length of 1,000. Although this search path is taken at the earliest stage of training, it is clear that the search agent gradually approaches to the source location due to the deployment of information-directed exploration. The second path in Fig. 2b is sampled at the last episode, which successfully leads the agent to the true source position with a reasonably short path. The green dots represent the agent's belief of source location, assembled from particle filters. It is observed that the proposed AID-RL algorithm achieves efficient source localisation and simultaneously acquires meaningful source parameters.

RL algorithms are distinct from other classic path planning methods. They aim to approximate the optimal solution over the entire search space, while traditional path planning methods, like Entrotaxis and DCEE, solve the optimisation problem from the current state. This fundamental difference gives rise to the nature of RL, i.e., episodic training over all possible actions and initial states. Random exploration based RL [12], ϵ-greedy, achieves this by adding random perturbation to its greedy actions.

While keeping all parameters of ϵ-greedy RL the same as AID-RL, we depict the moving average of the path length in Fig. 3. It is clear that the influence of information-directed exploration is more significant at the early stage of training, during which AID-RL demonstrates fast adaptation because of the deployment of



(a) Sampled path at the first episode.



(b) Sampled path at the last episode.

Fig. 2. Representative search paths using AID-RL at the beginning and end of the training process, respectively. Red line represents the search trajectory; green dots denote the estimated source location in the particle filter; blue dot denotes the agent start point (randomly initialised at each episode); blue star marks the end point of the search agent and back dot indicates the location of the true source; the greyscale shading delineates the instantaneous dispersion field at the current step.

active exploration. From extensive simulation and experimental results of the classic IPP methods (see e.g., [25,13,33]), the search agent will gradually approach to the source while seeking to explore the most uncertain positions in the domain. This reveals the fundamental reason for the success of IPP as most informative locations are usually adjacent to the source position. Comparing the average path length in Fig. 3, AID-RL outperforms random exploration based RL at all stages of training, even though the performance margin of AID-RL is narrowed after a significant amount of episodic training. Encompassing information-directed exploration into RL algorithm greatly accelerates the process of finding source location.

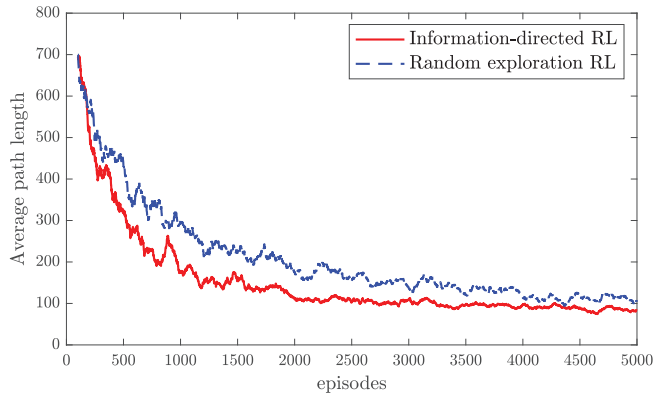
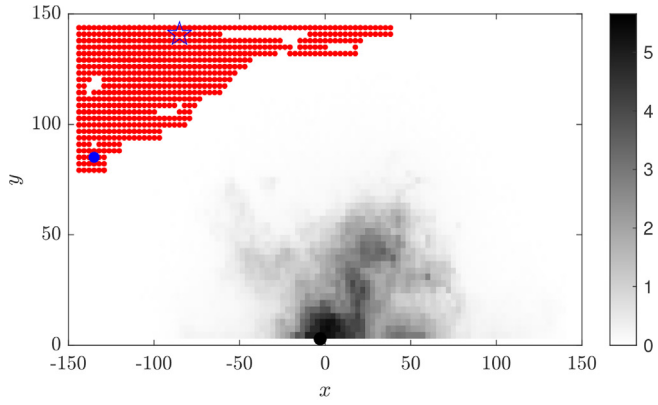
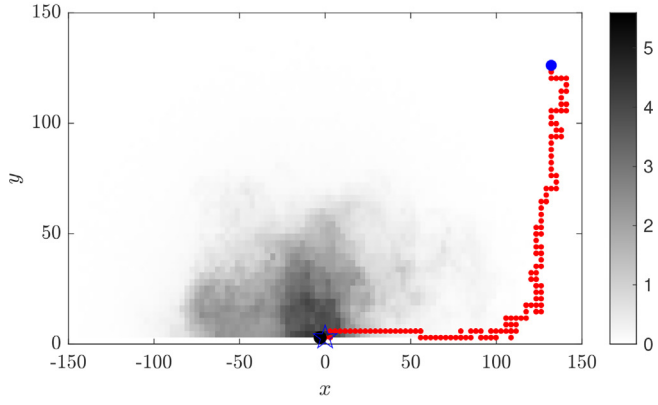


Fig. 3. Moving average of the path length under information-directed and random exploration RL, respectively.

Fig. 4 shows two representative trajectories of the search agent at the beginning and the end of the training process using ϵ-greedy RL. It is obvious that the search agent tends to move close to the source by using the information-directed exploration mechanism (Fig. 2a), while the search pattern using ϵ-greedy RL algorithm is more random without any clear direction (Fig. 4a). This clearly manifests that information/uncertainty awareness plays a central role in directing the agent to conduct more effective and efficient exploration. At the end of training process, both algorithms can successfully lead the search agent to the source location but AID-RL requires less steps in average to approach the source as shown in Fig. 3.



(a) Sampled path at the first episode.



(b) Sampled path at the last episode.

Fig. 4. Representative search paths using ϵ-greedy RL at the beginning and end of the training process, respectively.

5. Conclusions

In this paper, an active exploration autonomous search framework has been established based on greedy reinforcement learning. Inspired by the pioneering work in dual control [24], we propose an information-directed reinforcement learning to enable a balanced trade-off between exploration and exploitation. The greedy RL essentially implements an exploitation strategy that navigates the search agent to collect maximum reward (concentration), and the Entrotaxis search enables the agent to explore the most uncertain areas to improve the level of belief confidence. Such a model-based and model-free (MB-MF) paradigm shares the strengths from both sides. From the experiment results using high-fidelity dataset, the proposed AID-RL greatly improves the search and estimation performance and consumes less training episodes compared with traditional ϵ-greedy algorithms.

Recently, multi-agent systems have shown great potential in solving complex problems using collaborative swarm robots [34,35]. One of the key advantages of using multi-agent systems in autonomous search problems is that it enables collective intelligence, where the collaboration of multiple agents can lead to improved performance and faster problem-solving. Therefore, our future research efforts will focus on developing a distributed framework for AID-RL. This framework will incorporate multiple agents working together to achieve the source search goal, and our aim is to further enhance the search performance and learning speed by allowing the agents to collaborate and share information.

CRediT authorship contribution statement

Zhongguo Li: Conceptualization, Methodology, Software, Writing - original draft. **Wen-Hua Chen:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition. **Jun Yang:** Methodology, Supervision, Writing - review & editing. **Yunda Yan:** Conceptualization, Methodology, Writing - review & editing.

Data availability

Data will be made available on request.

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.

References

[1] T.F. Villa, F. Gonzalez, B. Miljievic, Z.D. Ristovski, L. Morawska, An overview of small unmanned aerial vehicles for air quality measurements: Present applications and future prospectives, Sensors 16 (7) (2016) 1072.
[2] M. Hutchinson, C. Liu, W.-H. Chen, Information-based search for an atmospheric release using a mobile robot: Algorithm and experiments, IEEE Trans. Control Syst. Technol. 27 (6) (2018) 2388–2402.
[3] M. Vergassola, E. Villermaux, B.I. Shraiman, 'Infotaxis' as a strategy for searching without gradients, Nature 445 (7126) (2007) 406–409.
[4] B. Ristic, A. Gunatilaka, Information driven localisation of a radiological point source, Inform. Fusion 9 (2) (2008) 317–326.
[5] X. Jiang, S. Li, B. Luo, Q. Meng, Source exploration for an under-actuated system: A control-theoretic paradigm, IEEE Trans. Control Syst. Technol. 28 (3) (2019) 1100–1107.
[6] J.B. Stock, M. Baker, Chemotaxis Encyclopedia of Microbiology, Elsevier Inc, 2009, pp. 71–78.
[7] W.-H. Chen, C. Rhodes, C. Liu, Dual control for exploitation and exploration (DCEE) in autonomous search, Automatica 133 (2021).

[8] Z. Li, W.-H. Chen, J. Yang, Concurrent active learning in autonomous airborne source search: Dual control for exploration and exploitation, IEEE Trans. Autom. Control 68 (2023) 3123–3130.
[9] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.
[10] A. Nagabandi, G. Kahn, R.S. Fearing, S. Levine, Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 7559–7566.
[11] M. Park, P. Ladosz, H. Oh, Source term estimation using deep reinforcement learning with Gaussian mixture model feature extraction for mobile sensors, IEEE Robot. Autom. Lett. 7 (3) (2022) 8323–8330.
[12] Y. Zhao, B. Chen, X. Wang, Z. Zhu, Y. Wang, G. Cheng, R. Wang, R. Wang, M. He, Y. Liu, A deep reinforcement learning based searching method for source localization, Inf. Sci. 588 (2022) 67–81.
[13] M. Hutchinson, H. Oh, W.-H. Chen, Entrotaxis as a strategy for autonomous search and source reconstruction in turbulent conditions, Inform. Fusion 42 (2018) 179–189.
[14] G. Ostrovski, P.S. Castro, W. Dabney, The difficulty of passive learning in deep reinforcement learning, Adv. Neural Inform. Process. Syst. 34 (2021) 23283–23295.
[15] P. Ladosz, L. Weng, M. Kim, H. Oh, Exploration in deep reinforcement learning: A survey, Inform. Fusion (2022).
[16] N. Nikolov, J. Kirschner, F. Berkenkamp, A. Krause, Information-directed exploration for deep reinforcement learning, 2018, arXiv preprint arXiv:1812.07544.
[17] Z. Li, W.-H. Chen, J. Yang, A dual control perspective for exploration and exploitation in autonomous search, in: 2022 European Control Conference (ECC) IEEE, 2022, pp. 1876–1881.
[18] C. Rhodes, C. Liu, W.-H. Chen, Autonomous source term estimation in unknown environments: From a dual control concept to UAV deployment, IEEE Robot. Autom. Lett. 7 (2) (2021) 2274–2281.
[19] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, et al., Foundations and Trends® in Machine Learning, Mach. Learn. 8 (5–6) (2015) 359–483.
[20] K. Chua, R. Calandra, R. McAllister, S. Levine, Deep reinforcement learning in a handful of trials using probabilistic dynamics models, Advances in Neural Information Processing Systems 31 (NIPS) vol. 31 (2018) 2018.
[21] J.Y. Shin, C. Kim, H.J. Hwang, Prior preference learning from experts: Designing a reward with active inference, Neurocomputing 492 (2022) 508–515.
[22] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, VIME: Variational information maximizing exploration, Advances in Neural Information Processing Systems, vol. 29, 2016.
[23] P. Shyam, W. Jaśkowski, and F. Gomez, Model-based active exploration, in International Conference on Machine Learning, pp. 5779–5788, PMLR, 2019.
[24] W.-H. Chen, Perspective view of autonomous control in unknown environment: Dual control for exploitation and exploration vs reinforcement learning, Neurocomputing (2022).
[25] B. Ristic, A. Skvortsov, A. Gunatilaka, A study of cognitive strategies for an autonomous search, Inform. Fusion 28 (2016) 1–9.
[26] Z. Li, W.-H. Chen, J. Yang, and Y. Yan, Dual control of exploration and exploitation for self-optimisation control in uncertain environments, arXiv preprint arXiv:2301.11984, 2023.
[27] M. Hutchinson, C. Liu, W. Chen, Source term estimation of a hazardous airborne release using an unmanned aerial vehicle, J. Field Robot. 36 (4) (2019) 797–817.
[28] M. Hutchinson, H. Oh, W.-H. Chen, A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors, Inform. Fusion 36 (2017) 130–148.
[29] A. Mesbah, Stochastic model predictive control with active uncertainty learning: A survey on dual control, Annu. Rev. Control 45 (2018) 107–117.
[30] Y. Zhao, B. Chen, Z. Zhu, F. Chen, Y. Wang, D. Ma, Entrotaxis-jump as a hybrid search algorithm for seeking an unknown emission source in a large-scale area with road network constraint, Expert Syst. Appl. 157 (2020).
[31] T. Alpcan, I. Shames, An information-based learning approach to dual control, IEEE Trans. Neural Networks Learn. Syst. 26 (11) (2015) 2736–2748.
[32] T.M. Moerland, J. Broekens, A. Plaat, C.M. Jonker, et al., Model-based reinforcement learning: A survey, Foundations and Trends, Mach. Learn. 16 (1) (2023) 1–118.
[33] P. Ojeda, J. Monroy, J. Gonzalez-Jimenez, Information-driven gas source localization exploiting gas and wind local measurements for autonomous mobile robots, IEEE Robot. Autom. Lett. 6 (2) (2021) 1320–1326.
[34] Y. Dong, Z. Li, X. Zhao, Z. Ding, X. Huang, Decentralised and cooperative control of multi-robot systems through distributed optimisation, in: The 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2023, 2023.
[35] U. Dah-Achinanon, S.E. Marjani Bajestani, P.-Y. Lajoie, G. Beltrame, Search and rescue with sparsely connected swarms, Autonomous Robots (2023) 1–15.

Zhongguo Li received the B.Eng. and Ph.D. degrees in electrical and electronic engineering from the University of Manchester, Manchester, U.K., in 2017 and 2021, respectively. He was a Research Associate with the Department of Aeronautical and Automotive Engineering, Loughborough University, Loughborough, U.K., from 2020 to 2022. He is currently a Lecturer in Robotics and AI with Department of Computer Science, University College London, London, U.K. His research interests include optimisation and decision-making for advanced control, game theory and learning in multi-agent systems, and their applications in autonomous vehicles and robotics.

Wen-Hua Chen holds Chair in Autonomous Vehicles with the Department of Aeronautical and Automotive Engineering, Loughborough University, U.K. He is the founder and the Head of the Loughborough University Centre of Autonomous Systems. He is interested in control, signal processing and artificial intelligence and their applications in robots, aerospace, and automotive systems. Dr. Chen is a Chartered Engineer, and a Fellow of IEEE, the Institution of Mechanical Engineers and the Institution of Engineering and Technology, U.K. He has authored or coauthored near 300 papers and 2 books. Currently he holds the UK Engineering and Physical Sciences Research Council (EPSRC) Established Career Fellowship in developing new control theory for robotics and autonomous systems.

Jun Yang is a Senior Lecturer in Autonomous and Electric Vehicles at Loughborough University. His research interests include disturbance observer, motion control, visual servoing, nonlinear control and autonomous systems. He serves as Associate Editor or Technical Editor of IEEE Transactions on Industrial Electronics, IEEE-ASME Transactions on Mechatronics, IEEE Open Journal of Industrial Electronics Society, etc. He was the recipient of the EPSRC New Investigator Award. He is a Fellow of IEEE and IET.

Yunda Yan received the B.Sc. degree in automation and the Ph.D. degree in control theory and control engineering from the School of Automation in Southeast University, Nanjing, China, in 2013 and 2019, respectively. He was a Research Associate with the Department of Aeronautical and Automotive Engineering, Loughborough University, from 2020 to 2022. He joined the School of Engineering and Sustainable Development, De Montfort University from Dec. 2022 as a Lecturer in Control Engineering. His current research interest focuses on the safety–critical control design for autonomous systems, especially related with optimisation and learning-based methods.