MANCHESTER
1824
The University of Manchester

The University of Manchester Research

Cooperative Active Learning based Dual Control for Exploration and Exploitation in Autonomous Search

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Li, Z., Chen, W. H., Yang, J., & Liu, C. (in press). Cooperative Active Learning based Dual Control for Exploration and Exploitation in Autonomous Search. *IEEE Transactions on NEural Networks and Learning Systems*.

Published in:
IEEE Transactions on NEural Networks and Learning Systems

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:18. Jan. 2024

Cooperative Active Learning based Dual Control for Exploration and Exploitation in Autonomous Search

Zhongguo Li, *Member, IEEE*, Wen-Hua Chen, *Fellow, IEEE*
Jun Yang, *Fellow, IEEE*, and Cunjia Liu, *Member, IEEE*

***Abstract*—In this paper, a multi-estimator based computationally efficient algorithm is developed for autonomous search in an unknown environment with an unknown source. Different from the existing approaches that require massive computational power to support nonlinear Bayesian estimation and complex decision-making process, an efficient cooperative active learning based dual control for exploration and exploitation (COAL-DCEE) is developed for source estimation and path planning. Multiple cooperative estimators are deployed for environment learning process, which is helpful to improving the search performance and robustness against noisy measurements. The number of estimators used in COAL-DCEE is much smaller than that of particles required for Bayesian estimation in information-theoretic approaches. Consequently, the computational load is significantly reduced. As an important feature of this study, the convergence and performance of COAL-DCEE are established in relation to the characteristics of sensor noises and turbulence disturbances. Numerical and experimental studies have been carried out to verify the effectiveness of the proposed framework. Compared with existing approaches, COAL-DCEE not only provides convergence guarantee, but also yields comparable search performance using much less computational power.**

***Index Terms*—Autonomous search, active learning, dual control, exploration and exploitation, goal-oriented control systems.**

I. Introduction

SEEKING a release of hazardous materials (including chemical, biological, radiological and nuclear materials) is of great importance in many applications such as disaster management and environment protection [1]. It is undeniably true that information plays a central role in source term estimation and tracking. Roughly speaking, information collection is mainly achieved by two type of approaches: static sensors pre-deployed onsite and dynamic sensors equipped on mobile platforms [1–3]. The former approach is costly and only practicable for some industries of potential risks, e.g., nuclear power plant. Owing to the rapid development of control and automation technologies, autonomous search using unmanned vehicles has attracted considerable research interest, and great successes have been demonstrated through experimental studies, e.g. [4–7].

Most of autonomous search algorithms can be classified into two categorises: control-driven and information-driven approaches. The former aims to reduce the distance between estimated source position and search agent, which corresponds to classic control problems, while the latter intends to explore the environment to acquire better source estimation by optimising some information gains, such as entropy, Kullback-Leibler divergence and variance [5, 7, 8]. Typical control approaches include extremum seeking [9] and model predictive control (MPC) [7]. Information-driven approaches have been extensively studied in recent years owing to their strong robustness against measurement noise and sparsity, such as Entrotaxis [10] and Infotaxis [3, 11].

More recently, a control-theoretic approach has been formulated in [7] to balance two the existing approaches, named as *dual control for exploration and exploitation* (DCEE). In that framework, the objective is two-fold: reconstructing more accurate knowledge of the operational environment (information objective) and navigating the search agent towards the source (control objective). However, similar to all existing information-theoretic approaches, the proposed DCEE framework still requires intensive computational resources and lacks stability and convergence analysis. The computational load is mainly caused by the optimisation loop and its entangled relationship with nonlinear particle filters, since at each iteration the optimisation-based path planner is required to interact with the inference engine in order to calculate the predicted posterior of the source and environment parameters for all possible actions. Currently, a limited set of moving directions and fixed length of step sizes are utilised to ensure the computational feasibility of those algorithms [4–7]. Nonetheless, the implementation is still carried out by remote computation centres rather than onboard processors [4–7]. In [12], a sampled-based path generation method using RRT* path planner has been integrated with dual control, which can effectively expand the action set but still suffers from intensive computational load.

Although extensive experimental results have manifested the effectiveness of those algorithms (see [7, 12]), theoretical properties such as convergence and performance have not

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. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising. Corresponding author: Wen-Hua Chen.

Z. Li is with Department of Electrical and Electronic Engineering, University of Manchester, Manchester, M13 9PL, U.K. (email: zhongguo.li@manchester.ac.uk).

W. Chen, J. Yang and C. Liu are with Department of Aeronautical and Automotive Engineering, Loughborough University, Loughborough, LE11 3TU, U.K. (emails: w.chen@lboro.ac.uk; j.yang3@lboro.ac.uk; c.liu5@lboro.ac.uk).

yet been systematically investigated. Actually, it has been a common issue along with IPP and DCEE. The coupling between path planning and environment learning, together with noisy measurements, turbulence disturbances and their inherent stochastic nature, significantly compounds the theoretical analysis. The reformulation in DCEE establishes a useful link between learning and control, which provides an access to extensive analytical tools in control and learning theory.

In fact, there is a significant difference between DCEE and traditional control. Conventionally, a control strategy requires a predefined trajectory or setpoint, for example, output regulation [13], path following [14, 15] and model predictive control [14, 16]. In autonomous search, the source position is unknown and consequently there is no direct path that can lead the agent to the unknown source. To solve such a problem, it involves dual objectives: one is to learn the source parameters by actively probing the operational environment (exploration), and another is to plan the search path leading the agent to the believed source location (exploitation). From this perspective, IPP can be viewed as a pure exploration strategy aiming at learning more accurate source position, while control-driven approaches use a pure exploitation strategy targeting at moving close to current estimation of the source [17, 18]. Passive learning methods, that is, estimating the source properties by accidentally collected data points without intention to improve the the estimation performance, are very unreliable and inefficient due to the presence of disturbances and uncertainties in autonomous search of airborne hazardous materials (e.g., intermittent sensor measurements caused by local turbulence and sensory noises). On the other hand, the active learning mechanism intentionally collects those data that are mostly conducive to improving the estimation performance, and thus it can potentially enhance learning outcomes in uncertain environments. DCEE achieves a natural balance between active exploration and efficient exploitation, as it is derived from a physically meaningful objective. Noticeable performance improvement has been demonstrated by extensive simulations and experimental results compared with the other approaches like IPP and MPC [7].

Introducing active learning requires an information measure for uncertainty quantification. In general, the performance of single estimator (such as observer and learning machine) is often severely influenced by the initialisation and setting of the individual estimator, for example, state estimation [19], disturbance observer [20] and parameter adaptation [21]. Population-based methods are often employed in large-scale optimisation and learning problems such as particle swarm optimisation [22], evolutionary algorithm [23] and multi-agent based learning [24]. Surprisingly, very few studies have attempted to use multiple estimators/observers for control problems. In autonomous search, deploying multiple parallel estimators can eliminate undesirable behaviours caused by improper random initialisation of an individual, and also it allows us to make full use of prior probability density function (PDF) of source parameters, e.g. the range of wind speed and direction from local weather forecast. More importantly, it provides a means for quantifying *uncertainty* associated with source estimators, which is of great importance to empower the search agent with *dual capability of exploration and exploitation*. In the machine learning community, methods such as hyperspherical energy minimisation [25] and distribution-shattering strategy [26] have been designed for active learning. Those dual effects become increasingly important as modern control systems are required to accomplish high-level goals subject to environmental uncertainties [27]. We name this type of intelligent system as *goal-oriented control system* (GOCS), and regard it as the key for improving the level of autonomy. Recent success of active learning in robotics and control applications has been reported in many related works [28–30].

Inspired by the DCEE framework [7], our previous study [31] has formulated the autonomous search as a learning based control problem, namely concurrent learning for exploration and exploitation (CLEE). The learning process, supported by multiple estimators, is intended to establish the source and environment knowledge based on all available measurements and prior information. The control action is empowered with dual capability that *exploits* the acquired knowledge from the estimators and in the meanwhile *explores* the environment to reduce *predicted* future uncertainty. Our main motivation of devising multiple learners in CLEE [31] was to provide a means to quantify the estimation uncertainty without using computationally intensive particle filters in IPP and DCEE. From the simulation results, we noticed that CLEE can provide satisfactory performance under good sensor conditions, while its resilience to intermittent sensor dropouts and unknown environment parameters is quite limited. In the *concurrent learning* approach, some estimators may be occasionally confined to local optimum, leading to degraded estimation/search performance and even failed search.

Motivated by the above observations, this paper further develops a *cooperative active learning* based dual control, named as COAL-DCEE for short, which not only holds considerable computational efficiency as the CLEE approach, but also shares strong robustness against measurement noises as IPP and DCEE methods. More importantly, rigorous convergence and performance analysis of the proposed COAL-DCEE will be analysed under mild assumptions. To our knowledge, there is no existing study that has endeavoured to examine convergence properties for dual control or IPP, except for our previous attempt in [31]. The introduction of multiple cooperative estimators enables us to make use of the wide analytical tools in learning and control for establishing the convergence of COAL-DCEE. The ensemble based estimation method advocated in this paper is distinct from those probabilistic or dynamic ensemble estimation approaches dedicated for machine learning problems [32, 33]. Existing active learning based algorithms (e.g., [32, 33]) 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 the autonomous search problem. The proposed multi-estimator based ensemble approach makes use of the environment model and the learned parameters are physically meaningful. In addition, COAL-DCEE demonstrates its robustness particularly in incorporating with intermittent and noisy sensor measurements arising during airborne source search. We summarise the key contributions of this paper as follows.

1) COAL-DCEE provides a computationally efficient algorithm for autonomous search problems with dual effects of exploration and exploitation, which renders a unified paradigm for control-driven and information-driven approaches. It embeds an active learning effect allowing the system to actively explore the unknown environment to reduce the level of uncertainty.
2) Instead of using computationally expensive particle filters as in information-driven methods, this paper develops an efficient multi-estimator based ensemble approach to quantify the estimation uncertainty online and allow the system to actively explore the unknown environment with a much reduced computational effort. It overcomes a major obstacle in a wide application of the dual control concept in DCEE. It is shown that we managed to speed up the DCEE implementation by about 100 times.
3) The convergence guarantee of COAL-DCEE is established using control and optimisation theories, and algorithm performance is rigorously analysed under mild assumptions on sensor noises and turbulence disturbances.
4) The proposed COAL-DCEE demonstrates superior search performance with high *computational efficiency*. Both simulation and experimental studies have been provided to illustrate the advantages of the proposed algorithm by extensive comparison with existing strategies such as Entrotaxis, MPC and the original version of DCEE.

The rest of this paper is organised as follows. In Section II, autonomous source seeking is formulated as a goal-oriented control problem, and a position-driven dual control framework is developed by introducing multiple dynamic estimators. Section III presents the learning-based control algorithm, and establishes the convergence of COAL-DCEE. In Section IV, the proposed algorithm is validated using both numerical and experimental datasets with comparison to existing methods. Conclusion is drawn in Section V.

II. Modelling and Formulation

A. Agent Modelling

The focus of this paper is on the high-level decision-making for autonomous search. It is assumed that the search agent, such as an unmanned ground robot or aerial vehicle, is devised with low-level controller that can achieve movement instructions directed by the high-level decision maker. Hence, the agent's dynamics can be represented by

$$\boldsymbol{p}_{k+1} = \boldsymbol{p}_k + \boldsymbol{u}_k + w_k \tag{1}$$

where $\boldsymbol{p}_k = [p_{k,x}, p_{k,y}, p_{k,z}]^\mathsf{T} \in \Omega \subseteq \mathbb{R}^3$ denotes the position of the agent at step k, Ω is a compact set of searching space, $\boldsymbol{u}_k \in \mathcal{U} \subseteq \mathbb{R}^3$ is the control action with $\mathcal{U}$ being the admissible set of actions, and w_k is the control error. The control error w_k quantifies the position disturbances. For example, it may be caused by mapping error in the positioning system and/or the lower level controller mismatch caused by local disturbance or turbulences. Different from the existing studies for information-theoretic approaches [4, 7, 10] that select the actions from a small set of movement directions with fixed length, the admissible actions $\mathcal{U}$ in this paper can be continuous with arbitrary direction and length.

B. Dispersion and Environment Modelling

Dispersion models are used to calculate the expected concentrations at different locations given a set of source and environment parameters. There have been various dispersion models proposed for different applications, as summarised in [34]. In this paper, the isotropic dispersion model will be utilised for source term estimation, which has demonstrated great flexibility and efficiency in dynamic source tracking in many recent studies [4–7]. Given a source $\boldsymbol{\Theta}_s := [\boldsymbol{s}^\mathsf{T}, q]^\mathsf{T} \in \mathbb{R}^4$ at position $\boldsymbol{s} = [s_x, s_y, s_z]^\mathsf{T} \in \mathbb{R}^3$ with a positive release rate $q \in \mathbb{R}^+$, the expected concentration at agent's position $\boldsymbol{p}_k \in \mathbb{R}^3$ can be obtained by

$$\begin{aligned}\mathcal{M}\left(\boldsymbol{p}_k, \boldsymbol{\Theta}_s\right) = & \frac{q}{4\pi\zeta_{s1}\|\boldsymbol{p}_k - \boldsymbol{s}\|} \exp\left[\frac{-\left\|\boldsymbol{p}_k - \boldsymbol{s}\right\|}{\lambda}\right] \\ & \times \exp\left[\frac{-\left(p_{k,x} - s_x\right) u_s \cos\phi_s}{2\zeta_{s1}}\right] \\ & \times \exp\left[\frac{-\left(p_{k,y} - s_y\right) u_s \sin\phi_s}{2\zeta_{s1}}\right]\end{aligned} \tag{2}$$

where environmental parameters are composed of the wind speed u_s, wind direction ϕ_s, diffusivity ζ_{s1}, the particle lifetime ζ_{s2}, and a composite coefficient $\lambda = \sqrt{\frac{\zeta_{s1}\zeta_{s2}}{1+(u_s^2\zeta_{s2})/(4\zeta_{s1})}}$.

C. Sensor Modelling

Information collection is the key to the decision-making of source estimation and path planning. In the field of source seeking, chemical concentration is one of the essential measurement required for most of the algorithms [4, 7, 10, 35]. In some works, the search agent is also required to measure additional source and environmental parameters, for example, chemical gradients [36], wind speed and direction [37]. In this paper, we assume that the agent is equipped with onboard chemical/biological sensors, which can measure the local concentration values. Due to the physical nature of these sensors, the search agent needs to stay at the current location for a short period to obtain reliable readings, i.e., the sampling time. Nevertheless, as a result of sensor noises and local turbulence disturbances, information collected is often noisy and sparse. The sensory measurement can be modelled as

$$z(\boldsymbol{p}_k) = \begin{cases} \mathcal{M}\left(\boldsymbol{p}_k, \boldsymbol{\Theta}_s\right) + v_k, & D = 1 \\ \bar{v}_k, & D = 0 \end{cases} \tag{3}$$

where $\mathcal{M}$ is the true chemical concentration from the dispersion model, D represents either a detection event for $D = 1$ or a non-detection event for $D = 0$, and v_k and $\bar{v}_k$ are additive noises. For ease of theoretical analysis, we assume that the sensor can always receive concentration with additive noises, i.e., $D = 1$, but in the simulation and experiment studies we will examine how the proposed algorithm will behaviour under non-detection events.

D. Multi-Estimator Cooperative Learning Based Dual Control

The concentration information collected up to time step k is denoted by $\boldsymbol{\mathcal{Z}}_k := \{z(\boldsymbol{p}_1), z(\boldsymbol{p}_2), \ldots, z(\boldsymbol{p}_k)\}$. The posterior distribution of source estimation can be represented by $\rho_{k|k} := p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{Z}}_k)$ at time k, where $p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{Z}}_k)$ is a probability density function denoting the belief of $\boldsymbol{\Theta}_s$ conditional on $\boldsymbol{\mathcal{Z}}_k$. When the search agent moves to a new position directed by the control input $\boldsymbol{u}_k$, the one-step-ahead hypothetical posterior distribution of source estimation will be updated as $\hat{\rho}_{k+1|k} := p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{Z}}_{k+1|k})$ where $\boldsymbol{\mathcal{Z}}_{k+1|k} = \{\boldsymbol{\mathcal{Z}}_k, \hat{z}_{k+1|k}\}$ with $\hat{z}_{k+1|k}$ being the one-step-ahead possible measurement. As a result, the control input $\boldsymbol{u}_k$ will not only affect the future agent position but also affect the future belief of source location.

Motivated by the above discussion, the control input $\boldsymbol{u}_k$ should be designed to navigate the agent move closer to the predicted posterior estimation of source location. Therefore, the goal of COAL-DCEE for autonomous search can be formulated as the following optimisation problem

$$\begin{aligned}\min_{\boldsymbol{u}_k \in \mathcal{U}_k} J(\boldsymbol{u}_k) &= \min_{\boldsymbol{u}_k \in \mathcal{U}_k} \mathbb{E}_{\boldsymbol{\Theta}} \left[\mathbb{E}_{\hat{z}_{k+1|k}} \left[\left\|\boldsymbol{p}_{k+1|k} - \boldsymbol{s}\right\|^2 |\boldsymbol{\mathcal{Z}}_{k+1|k}\right]\right] \\ \text{subject to } & \ \boldsymbol{p}_{k+1|k} = \boldsymbol{p}_k + \boldsymbol{u}_k + w_k.\end{aligned} \tag{4}$$

By minimising the cost function defined in (4), the control action $\boldsymbol{u}_k$ at current step k will not only change future agent position $\boldsymbol{p}_{k+1|k}$, but also result in different concentration measurements $\hat{z}_{k+1|k}(\boldsymbol{p}_{k+1|k})$. Consequently, the future belief of the source location and release rate, $\hat{\rho}_{k+1|k} := p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{Z}}_{k+1|k})$, is influenced by the available information $\boldsymbol{\mathcal{Z}}_k$ as well as predicted future measurement $\hat{z}_{k+1|k}(\boldsymbol{p}_{k+1|k})$. We can split the cost function (4) into two terms to reflect the exploration and exploitation effects, which has been shown in [7].

First, we define the mean of the predicted posterior source location conditional on $\boldsymbol{\mathcal{Z}}_{k+1|k}$ as

$$\bar{\boldsymbol{s}}_{k+1|k} := \mathbb{E}\left[\mathbf{s}_{k+1|k}\right] = \mathbb{E}\left[\boldsymbol{s}|\boldsymbol{\mathcal{Z}}_{k+1|k}\right] \tag{5}$$

based on which the estimation error with respect to the mean can be denoted as

$$\tilde{\boldsymbol{s}}_{k+1|k} = \boldsymbol{s} - \bar{\boldsymbol{s}}_{k+1|k}. \tag{6}$$

Lemma 1 ([7]): For the autonomous search problem with an unknown release location, the objective function defined in (4) implicitly contains dual effects for exploration and exploitation, as reflected by the following equivalent formulation

$$\begin{aligned}J(\boldsymbol{u}_k) =& \mathbb{E}\left[\|\boldsymbol{p}_{k+1|k} - \bar{\boldsymbol{s}}_{k+1|k}\|^2|\boldsymbol{\mathcal{Z}}_{k+1|k}\right] \\ &+ \mathbb{E}\left[\|\tilde{\boldsymbol{s}}_{k+1|k}\|^2|\boldsymbol{\mathcal{Z}}_{k+1|k}\right].\end{aligned} \tag{7}$$

In previous works [7, 12], Bayesian frameworks are used to approximate the probability distribution of the source parameters, which are usually computationally expensive. In this paper, we resort to an efficient ensemble based approximation method using a group of dynamic estimators. The deployment of ensemble approximation is motivated by its outstanding success in machine learning, e.g., [32, 33]. More justifications on ensemble-based approximation will be detailed in Remark 7. When we have a set of N estimators, the dual control problem for autonomous search can be written as

$$\min_{\boldsymbol{u}_k \in \mathcal{U}} J(\boldsymbol{u}_k) = \min_{\boldsymbol{u}_k \in \mathcal{U}} \left[\|\boldsymbol{p}_{k+1|k} - \bar{\boldsymbol{s}}_{k+1|k}\|^2 + \mathcal{P}_{k+1|k}\right] \tag{8}$$

where

$$\bar{\boldsymbol{s}}_{k+1|k} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{s}^i_{k+1|k} \tag{9}$$

$$\mathcal{P}_{k+1|k} := \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{s}^i_{k+1|k} - \bar{\boldsymbol{s}}_{k+1|k})^{\mathsf{T}}(\boldsymbol{s}^i_{k+1|k} - \bar{\boldsymbol{s}}_{k+1|k}) \tag{10}$$

with $\boldsymbol{s}^i_{k+1|k}$ being the predicted source position of the ith estimator at time k, for $i = 1, 2, \ldots, N$. In this paper, superscript i denotes the index of the estimators and subscript k represents the number of time step.

To acquire the source parameters, we follow a similar idea as in our previous work [31] by resorting to the least square method. The optimal source term can be obtained by minimising the difference between evaluated concentration from the dispersion model and the collected concentration from sensors, given by

$$f(\boldsymbol{\Theta}^i_k, \boldsymbol{p}_k) = \left[\mathcal{M}(\boldsymbol{p}_k, \boldsymbol{\Theta}^i_k) - z(\boldsymbol{p}_k)\right]^2 \tag{11}$$

By this formulation, existing techniques in system identification and adaptive control can be leveraged [38]. In this paper, we will use a memory regressor extension based adaptation method for parameter acquisition.

Remark 1: From the above formulation, it is clear that the control action $\boldsymbol{u}_k$ obtained from the optimisation problem in (8) has dual effects. Optimising the first term in (8) navigates the search agent to the believed source location (the mean), which is related to exploitation. On the other hand, minimising the second term steers the agent to search over some positions that can reduce the estimation uncertainty (the variance) of the estimators since $\mathcal{P}_{k+1|k}$ is the predicted covariance of the estimation uncertainty.

Remark 2: It should be noted that the value function defined in (4) is different from [31]. The formulation (4) is a position-driven optimisation problem, whereas [31] uses a concentration-driven mechanism. It is observed from previous studies [6, 7, 12] that the position-driven formulation is more robust compared with [31]. In Section III, a cooperative active learning approach will be developed to enhance algorithm resilience against noises and disturbances. In summary, two new features, i.e. formulation and algorithm, are introduced to achieve stronger robustness compared with our previous study [31], and both of them are validated through empirical results in Section IV.

Remark 3: In a control system, the level of autonomy can be measured in terms of the set of *goals* that the system is able to accomplish subject to a set of *uncertainties* [27]. In order to improve system's autonomy, it is required that the system can *exploit* its available knowledge to accomplish the goals, and at the same time it should be able to *explore* the operational environment to reduce knowledge uncertainty. In some recent works (see, e.g. [14, 39–41]), explicit trade-off coefficients are introduced on purpose to incorporate the exploration terms into model predictive control problems. This inevitably incurs additional efforts in tuning the coefficients to balance exploration and exploitation, which, as we all know, is never a trivial task due to deep and complex interaction between the system and the uncertain environment. From Lemma 1, it is clear that the dual effects in COAL-DCEE are *naturally embedded*, since they are derived from a physically meaningful value function in (4).

III. Cooperative Learning based Dual Control for Exploration and Exploitation

In this section, we will propose a cooperative learning based dual control algorithm for autonomous search problems. Then, we analyse the convergence and steady-state performance of the cooperative source estimators where noises and disturbances will be taken into consideration. Those results will then be used to establish the convergence of the overall algorithm.

A. Algorithm Development

Now, we present the gradient-based learning algorithm for the source term estimation and path planning. Inspired by the memory based adaptation [38, 42] and the concurrent learning [31], we propose a cooperative ensemble estimation method, given by

$$\begin{aligned}\mathbf{\Theta}_k^i =&\mathbf{\Theta}_{k-1}^i - \eta_{k-1}^i \sum_{t=k-q}^{k-1} \tilde{\nabla}_{\mathbf{\Theta}} f(\mathbf{\Theta}_{k-1}^i, \boldsymbol{p}_t) \\ &- \tau_{k-1}^i(\mathbf{\Theta}_{k-1}^i - \bar{\mathbf{\Theta}}_{k-1}), \quad \forall i = 1, 2, \ldots, N\end{aligned} \tag{12}$$

where $\tilde{\nabla} f(\mathbf{\Theta}_{k-1}^i, \boldsymbol{p}_t)$ denotes the perturbed gradients of the least square function under sensor noises, q is the number of past measurements used in memory extension, and $\bar{\mathbf{\Theta}}_{k-1}$ is the nominal estimation, defined as

$$\bar{\mathbf{\Theta}}_{k-1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{\Theta}_{k-1}^i. \tag{13}$$

The path planning is given by

$$\begin{aligned}\boldsymbol{p}_{k+1} &= \boldsymbol{p}_k + \boldsymbol{u}_k + w_k \\ \boldsymbol{u}_k &= -\delta_k \left[\nabla_{\boldsymbol{p}} \mathcal{C}_{k+1|k} + \nabla_{\boldsymbol{p}} \mathcal{P}_{k+1|k}\right]\end{aligned} \tag{14}$$

where $\mathcal{C}_{k+1|k} = ||\boldsymbol{p} - \bar{\boldsymbol{s}}_{k+1|k}||^2$ denotes the exploitation term in the dual objective (8), and $\eta_{k-1}^i, \tau_{k-1}^i, \delta_k \in \mathbb{R}^+$ are constant step sizes to be designed. It should be noted that $\nabla_{\boldsymbol{p}} \mathcal{C}_{k+1|k}$ and $\nabla_{\boldsymbol{p}} \mathcal{P}_{k+1|k}$ are pure prediction based on current estimation and model without noises. In essence, algorithms (12), (13) and (14) use gradient descent methods by which the source estimators converge to the true parameters that minimise the least square function in (11) and the agent moves towards the believed position of a release.

Remark 4: When updating the source parameters using (12) and (13), a coupling term $\bar{\mathbf{\Theta}}_k$ is introduced. This global coupling between the individual estimator and the ensemble average is conducive to improving the estimation performance by avoiding undesirable contractions to local optimal solutions [1]. However, the cooperation among estimators significantly complicates the theoretical analysis on algorithm convergence and performance. The cooperative learning framework, in conjunction with the analytical tools developed later, forms a new contribution of this study, as compared with our previous work in [31].

The overall implementation structure of COAL-DCEE has been summarised in Algorithm 1. It consists of an initialisation process and an iteration loop. From step 1 to 3, the source estimators and search agent are initialised according to available prior knowledge. During the iteration process, the agent will first collect concentration measurement at the current position, and then use new information to update N cooperative estimators as in step 6. To obtain the covariance $P_{k+1|k}$, we leverage the classical principle of predicting covariance estimation in extended Kalman filters, which has been elaborated in [31]. Based on the predicted future covariance, the search agent plans its next movement using gradient descent algorithm in step 9. The search process is terminated if the source is successfully identified or the budget is approached.

Algorithm 1 Overall structure of COAL-DCEE.

```
Initialisation:
 1. allocate the number of estimators $N$
 2. conduct a number of $q$ initial samples $(\boldsymbol{p}_i, z(\boldsymbol{p}_i))$ indexed by $i = -q+1, -q+2, \ldots, 0$
 3. initialise agent's position $\boldsymbol{p}_0$ and prior knowledge of the source parameter $\mathbf{\Theta}_0^i$ for all $i = 1, 2, \ldots, N$
Iteration:
 4. set $k := k + 1$
 5. collect the concentration reading $z_k(\boldsymbol{p}_k)$ from the sensor at position $\boldsymbol{p}_k$
 6. for $i = 1 : N$
      update the estimated source terms by
      $\mathbf{\Theta}_k^i = \mathbf{\Theta}_{k-1}^i - \eta_{k-1}^i \sum_{t=k-q}^{k-1} \tilde{\nabla}_{\mathbf{\Theta}} f(\mathbf{\Theta}_{k-1}^i, \boldsymbol{p}_t) - \tau_{k-1}^i(\mathbf{\Theta}_{k-1}^i - \bar{\mathbf{\Theta}}_{k-1})$
      $\bar{\mathbf{\Theta}}_{k-1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{\Theta}_{k-1}^i$
    end for
 7. calculate current estimation covariance
      $\mathcal{P}_{k|k} = \mathrm{diag}\left((s_k^i - \bar{s}_k)(s_k^i - \bar{s}_k)^{\mathsf{T}}\right)$
 8. predict future covariance as a function of $\boldsymbol{p}_{k+1|k}$
      $\boldsymbol{F}_{k+1} = \mathrm{col}\left( \frac{\partial f}{\partial \mathcal{M}} \frac{\partial \mathcal{M}}{\partial \boldsymbol{p}} \Big|_{\boldsymbol{p}_{k+1|k}, \mathbf{\Theta}_k^i} \right)$
      $\mathcal{P}_{k+1|k} = \mathrm{trace}(\boldsymbol{F}_{k+1}^{\mathsf{T}} \mathcal{P}_{k|k} \boldsymbol{F}_{k+1})$
 9. update the next movement for the agent by
      $\boldsymbol{u}_k = -\delta_k \left[\nabla_{\boldsymbol{p}} \mathcal{C}_{k+1|k} + \nabla_{\boldsymbol{p}} \mathcal{P}_{k+1|k}\right]$
      $\boldsymbol{p}_{k+1} = \boldsymbol{p}_k + \boldsymbol{u}_k + w_k$
End if termination condition is satisfied or time budget is approached.
```

The system-environment interaction is depicted in Fig 1, comprising two main components: an environment estimator and a path planner. The environment estimator utilises cooperative ensemble learning method with memory-based adaptation. The input of the environment estimator includes agent's positions and sensory measurements which are subject to noises and disturbances in an uncertain environment. The output is an ensemble of estimated source parameters that represent the agent's current belief about the unknown environment. The path planner determines the optimal action that

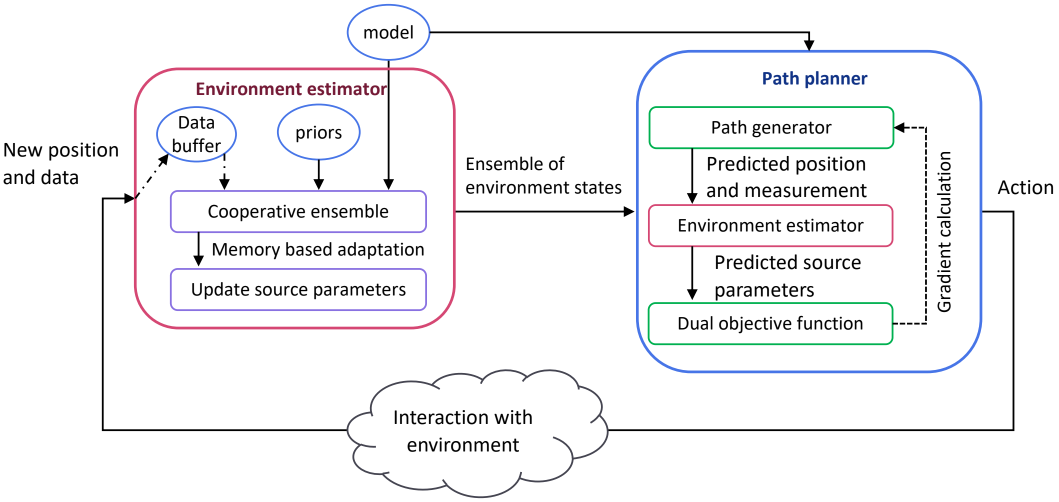


Fig. 1: System-environment interaction in autonomous search using COAL-DCEE.

aligns with the dual objective function. In order to construct the dual objective function, the planner requires the current belief of the source parameters and calculates the predicted source parameters by running the environment estimator with hypothetical positions and measurements.

B. Convergence Analysis

Before analysing the overall algorithm of COAL-DCEE, we first establish the convergence of the cooperative estimators. The underlying principle is inspired by the observation that increasing the number of samples (unbiased measurements with bounded variance) will always contribute to the convergence of the estimators. Before proceeding, we first introduce the following assumption on the noises.

Assumption 1: The measurement noise of the onboard chemical sensor v_k in (3) satisfies the following properties:

$$\mathbb{E}\left[v_k\right] = 0 \tag{15}$$

$$\mathbb{E}\left[\|v_k\|^2\right] \leq \varrho^2 \tag{16}$$

where ϱ is a positive constant. The position control error has similar properties:

$$\mathbb{E}\left[w_k\right] = 0 \tag{17}$$

$$\mathbb{E}\left[\|w_k\|^2\right] \leq \rho^2 \tag{18}$$

where ρ is a positive constant.

Remark 5: Assumption 1 indicates that the noises have zero mean and bounded variance. Zero mean implies that the noises are unbiased, which is usually satisfied, for example, by means of calibration and compensation. Any physical systems inherently suffer from noises, caused by sensors, turbulence disturbances and control errors.

We are now in position to give the convergence of the cooperative estimators.

Lemma 2: Consider an autonomous search problem with dispersion model (2) and noises satisfying Assumption 1. Design the learning rates η^i_{k-1} and τ^i_{k-1} such that all eigenvalues of $A_{k-1} = I_{4N} - \Lambda_{k-1}\operatorname{diag}(\mathcal{G}^i_{k-1}) - \Upsilon_{k-1}B$ are within a unit circle, where

$$\begin{aligned}
\mathcal{G}^i_{k-1} &= \sum_{t=k-q}^{k-1} \int_0^1 \nabla^2_{\mathbf{\Theta}} f(\mathbf{\Theta}_s + \tau\tilde{\mathbf{\Theta}}^i_{k-1}, \boldsymbol{p}_t)d\tau \in \mathbb{R}^{4\times4} \\
B &= \left(I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\mathsf{T}\right) \otimes I_4 \in \mathbb{R}^{4N\times4N} \\
\Lambda_{k-1} &= \operatorname{diag}(\eta^1_{k-1}, \dots, \eta^N_{k-1}) \otimes I_4 \in \mathbb{R}^{4N\times4N} \\
\Upsilon_{k-1} &= \operatorname{diag}(\tau^1_{k-1}, \dots, \tau^N_{k-1}) \otimes I_4 \in \mathbb{R}^{4N\times4N}
\end{aligned} \tag{19}$$

with the symbol $\otimes$ denoting Kronecker product. Then, the cooperative estimators converge to a neighbourhood of the true source parameters, given by

$$\lim_{k\to\infty} \mathbb{E}\left\|\mathbf{\Theta}_k - \mathbf{1}_N \otimes \mathbf{\Theta}_s\right\|^2 \leq \frac{\max_{j\in\{1,\dots,k-1\}} \sum_{i=1}^N (\eta^i_j)^2 q^2 L^2 \varrho^2}{1 - \max_{j\in\{1,\dots,k-1\}} \rho(A_j)} \tag{20}$$

where $\mathbf{\Theta}_k = \operatorname{col}\left(\mathbf{\Theta}^1_k, \mathbf{\Theta}^2_k, \dots, \mathbf{\Theta}^N_k\right) \in \mathbb{R}^{4N}$, $\operatorname{col}(\cdot)$ denotes a column vector formed by stacking the elements on top of each other, and $0 < \rho(A_j) < 1$ is the spectral radius of the transition matrix A_j for $j = 1, 2, \dots, k-1$.

Proof: Since all estimators are coupled by $\bar{\mathbf{\Theta}}_{k-1}$ in (12) and (13), in the following we will analyse their joint performance by collecting the parameters of all estimators into an augmented vector $\mathbf{\Theta}_{k-1} = \operatorname{col}\left(\mathbf{\Theta}^1_{k-1}, \mathbf{\Theta}^2_{k-1}, \dots, \mathbf{\Theta}^N_{k-1}\right)$. According to Assumption 1, the gradient term $\tilde{\nabla}_{\mathbf{\Theta}} f(\mathbf{\Theta}^i_{k-1}, \boldsymbol{p}_{k-1})$ can be written as

$$\tilde{\nabla}_{\mathbf{\Theta}} f(\mathbf{\Theta}^i_{k-1}, \boldsymbol{p}_{k-1}) = \nabla_{\mathbf{\Theta}} f(\mathbf{\Theta}^i_{k-1}, \boldsymbol{p}_{k-1}) + \mu_{k-1} \tag{21}$$

where

$$\mathbb{E}\left[\mu_{k-1}\right] = 0 \tag{22}$$

$$\mathbb{E}\left[\|\mu_{k-1}\|^2\right] \leq L^2\varrho^2 \tag{23}$$

with L being the Lipschitz constant of the least square function (11). We can then rewrite (12) in a compact form as

$$\mathbf{\Theta}_k = \mathbf{\Theta}_{k-1} - \Lambda_{k-1}[\mathbf{\Psi}(\mathbf{\Theta}_{k-1}) + q(\mathbf{1}_N \otimes \mu_{k-1})] - \Upsilon_{k-1}B\mathbf{\Theta}_{k-1} \tag{24}$$

where

$$\begin{aligned}\boldsymbol{\Psi}(\boldsymbol{\Theta}_{k-1}) = \mathrm{col}\Bigg(&\sum_{t=k-q}^{k-1} \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}_{k-1}^1, \boldsymbol{p}_t), \ldots, \\ &\sum_{t=k-q}^{k-1} \nabla_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}_{k-1}^N, \boldsymbol{p}_t)\Bigg) \in \mathbb{R}^{4N}\end{aligned} \tag{25}$$

In (24), we represent the approximated gradient as the true gradient with additive noises, i.e. $\boldsymbol{\Psi}(\boldsymbol{\Theta}_{k-1})$ and $\mathbf{1}_N \otimes \mu_{k-1}$. Define the estimation error as $\tilde{\boldsymbol{\Theta}}_k = \boldsymbol{\Theta}_k - \mathbf{1}_N \otimes \boldsymbol{\Theta}_s$, with $\tilde{\boldsymbol{\Theta}}_k^i := \boldsymbol{\Theta}_k^i - \boldsymbol{\Theta}_s$. From (24), the estimation error recursion can be written as

$$\tilde{\boldsymbol{\Theta}}_k = \tilde{\boldsymbol{\Theta}}_{k-1} - \Lambda_{k-1}[\boldsymbol{\Psi}(\boldsymbol{\Theta}_{k-1}) + q(\mathbf{1}_N \otimes \mu_{k-1})] - \Upsilon_{k-1} B \tilde{\boldsymbol{\Theta}}_{k-1} \tag{26}$$

where $B(\mathbf{1}_N \otimes \boldsymbol{\Theta}_s) = \mathbf{0}$ has been used to derive the above equation.

We apply the mean value theorem [43] to relate the gradient term $\boldsymbol{\Psi}(\boldsymbol{\Theta}_{k-1})$ with $\tilde{\boldsymbol{\Theta}}_{k-1}$, that is, for a twice-differentiable function $h(x) : \mathbb{R}^m \to \mathbb{R}$, we have

$$\begin{aligned}\nabla h(y) =& \nabla h(x) \\ &+ \left[\int_0^1 \nabla^2 h[x + \tau(y-x)] d\tau\right] (y-x), \forall x, y \in \mathbb{R}^m .\end{aligned} \tag{27}$$

Applying the above theorem leads to

$$\begin{aligned}\nabla_{\boldsymbol{\Theta}} f\left(\boldsymbol{\Theta}_{k-1}^i, \boldsymbol{p}_t\right) =& \nabla_{\boldsymbol{\Theta}} f\left(\boldsymbol{\Theta}_s, \boldsymbol{p}_t\right) \\ &+ \left[\int_0^1 \nabla_{\boldsymbol{\Theta}}^2 f(\boldsymbol{\Theta}_s + \tau \tilde{\boldsymbol{\Theta}}_{k-1}^i, \boldsymbol{p}_t) d\tau\right] \tilde{\boldsymbol{\Theta}}_{k-1}^i \\ =& \left[\int_0^1 \nabla_{\boldsymbol{\Theta}}^2 f(\boldsymbol{\Theta}_s + \tau \tilde{\boldsymbol{\Theta}}_{k-1}^i, \boldsymbol{p}_t) d\tau\right] \tilde{\boldsymbol{\Theta}}_{k-1}^i\end{aligned} \tag{28}$$

where $\nabla_{\boldsymbol{\Theta}} f\left(\boldsymbol{\Theta}_s, \boldsymbol{p}_t\right) = \mathbf{0}$ has been used.

Therefore, we have

$$\tilde{\boldsymbol{\Theta}}_k = (I_{4N} - \Lambda_{k-1}\mathcal{G}_{k-1} - \Upsilon_{k-1}B)\tilde{\boldsymbol{\Theta}}_{k-1} - q\Lambda_{k-1}(\mathbf{1}_N \otimes \mu_{k-1}) \tag{29}$$

where $\mathcal{G}_{k-1} := \mathrm{diag}(\mathcal{G}_{k-1}^1, \mathcal{G}_{k-1}^2, \ldots, \mathcal{G}_{k-1}^N)$. Now, taking the squared Euclidean norm of (29) leads to

$$\begin{aligned}\|\tilde{\boldsymbol{\Theta}}_k\|^2 =& \|(I_{4N} - \Lambda_{k-1}\mathcal{G}_{k-1} - \Upsilon_{k-1}B)\tilde{\boldsymbol{\Theta}}_{k-1}\|^2 \\ &+ \|q\Lambda_{k-1}(\mathbf{1}_N \otimes \mu_{k-1})\|^2 \\ &- 2q[(I_{4N} - \Lambda_{k-1}\mathcal{G}_{k-1} - \Upsilon_{k-1}B)\tilde{\boldsymbol{\Theta}}_{k-1}]^{\mathsf{T}} \\ &\times \Lambda_{k-1}(\mathbf{1}_N \otimes \mu_{k-1}).\end{aligned} \tag{30}$$

Applying the expectation operator and using (22) and (23), we have

$$\begin{aligned}\mathbb{E}\,\|\tilde{\boldsymbol{\Theta}}_k\|^2 =& \mathbb{E}\,\|(I_{4N} - \Lambda_{k-1}\mathcal{G}_{k-1} - \Upsilon_{k-1}B)\tilde{\boldsymbol{\Theta}}_{k-1}\|^2 \\ &+ \mathbb{E}\,\|q\Lambda_{k-1}(\mathbf{1}_N \otimes \mu_{k-1})\|^2.\end{aligned} \tag{31}$$

To obtain (31), we have used $\mathbb{E}[\Lambda_{k-1}(\mathbf{1}_N \otimes \mu_{k-1})] = 0$ from (22), which is independent of $(I_{4N} - \Lambda_{k-1}\mathcal{G}_{k-1} - \Upsilon_{k-1}B)\tilde{\boldsymbol{\Theta}}_{k-1}$. We further resort to (23) to bound the second term in (31), which yields

$$\mathbb{E}\,\|q\Lambda_{k-1}(\mathbf{1}_N \otimes \mu_{k-1})\|^2 \le \sum_{i=1}^N (\eta_{k-1}^i)^2 q^2 L \varrho^2. \tag{32}$$

Substituting (32) into (31) results in

$$\mathbb{E}\,\|\tilde{\boldsymbol{\Theta}}_k\|^2 \le \mathbb{E}\,\|\tilde{\boldsymbol{\Theta}}_{k-1}\|_{A_{k-1}}^2 + \sum_{i=1}^N (\eta_{k-1}^i)^2 q^2 L^2 \varrho^2 \tag{33}$$

where $A_{k-1} = [I_{4N} - \Lambda_{k-1}\mathcal{G}_{k-1} - \Upsilon_{k-1}B]^T[I_{4N} - \Lambda_{k-1}\mathcal{G}_{k-1} - \Upsilon_{k-1}B]$. If we select the learning rate η_{k-1}^i such that the eigenvalues of A are within unit circle, the convergence of (33) is guaranteed. Moreover, the estimation mean-square-error is given by

$$\lim_{k\to\infty} \mathbb{E}\,\|\tilde{\boldsymbol{\Theta}}_k\|^2 \le \frac{\max_{j\in\{1,\ldots,k-1\}} \sum_{i=1}^N (\eta_j^i)^2 q^2 L^2 \varrho^2}{1 - \max_{j\in\{1,\ldots,k-1\}} \rho(A_j)} \tag{34}$$

This completes the proof. ■

Remark 6: In parameter adaptation literature, how to guarantee persistent excitation has been a long-lasting research issue [38, 42]. It is often assumed that the control input can meet the excitation condition even though there is no such a stimulating effort added [44]. In this paper, a probing effort is inherently embraced in the dual controller, which will be conducive to environment acquisition. Further resorting to the memory-based regression method, it is shown that the convergence of the cooperative estimators is guaranteed under properly designed learning rates.

Remark 7: The cooperative estimator-based ensemble approach is of great importance in realising the dual control framework online with affordable computational costs. In fact, it is a hybrid approach that combines both model-based and model-free techniques in the estimation process. On the one hand, the source parameters are from the isotropic dispersion model, and the estimators are updated by minimising the observed data and the model outputs. On the other hand, a model-free ensemble is developed to evaluate the mean and variance of the estimator distribution that are used to formulate the exploration and exploitation terms in the dual controller. In machine learning community, this hybrid approach has been proven to be very successful and promising, mainly demonstrated by extensive simulation and experimental studies [32, 33, 45]. Inspired by its great success in machine learning, we further propose a cooperative ensemble method, and we establish its convergence properties by leveraging classic parameter adaptation tools in the control society. Nevertheless, the ensemble based estimation method advocated in this paper is distinct from those probabilistic or dynamic ensemble estimation approaches dedicated for machine learning problems [32, 33]. Existing active learning based algorithms 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 the autonomous search problem. The proposed multi-estimator based ensemble approach makes use of the environment model and the learned parameters are physically meaningful.

In conjunction with the path update law (14), the proposed COAL-DCEE will be able to steer the agent to the true source, as demonstrated in the following theorem.

Theorem 1: Consider an autonomous search problem with dispersion model (2) and noise conditions in Assumption 1.

Let the conditions in Lemma 2 hold. Suppose the learning rate δ_k is properly designed such that

$$0 < \alpha_k := 2\|I_3 - \delta_k \boldsymbol{\mathcal{L}}_k\|^2 < 1 \tag{35}$$

where $\boldsymbol{\mathcal{L}}_k = \int_0^1 \nabla^2_{\boldsymbol{p}} \mathcal{C}_{k+1|k}(\boldsymbol{s}+\tau\tilde{\boldsymbol{p}}_k)d\tau$. Then, the search agent converges to a bounded neighbourhood of the true source location using COAL-DCEE in Algorithm 1. Furthermore, the steady-state mean-square-error between agent and true source is given by

$$\lim_{k\to\infty} \mathbb{E}\,\|\boldsymbol{p}_k - \boldsymbol{s}\|^2 \leq \frac{\gamma+\rho^2}{1-\bar{\alpha}} \tag{36}$$

where $\gamma > 0$ denotes an upper bound of $2\,\mathbb{E}[\delta^2\|\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}\|^2]$, and $\bar{\alpha} := \max_{j\in\{1,\ldots,k\}} \alpha_k$.

Proof: To show the convergence of agent position, we first introduce an error variable $\tilde{\boldsymbol{p}}_k = \boldsymbol{p}_k - \boldsymbol{s}$, which is the difference between agent position and source position. Recalling the dual controller in (14), we can obtain

$$\tilde{\boldsymbol{p}}_{k+1} = \tilde{\boldsymbol{p}}_k - \delta_k \nabla_{\boldsymbol{p}}\mathcal{C}_{k+1|k} - \delta_k \nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k} + w_k. \tag{37}$$

Appealing to the mean value theorem in (27), it follows that

$$\begin{aligned}\nabla_{\boldsymbol{p}}\mathcal{C}_{k+1|k}(\boldsymbol{p}_k) =& \nabla_{\boldsymbol{p}}\mathcal{C}_{k+1|k}(\boldsymbol{s}) \\ &+ \left[\int_0^1 \nabla^2_{\boldsymbol{p}}\mathcal{C}_{k+1|k}(\boldsymbol{s}+\tau\tilde{\boldsymbol{p}}_k)d\tau\right]\tilde{\boldsymbol{p}}_k.\end{aligned} \tag{38}$$

Denoting $\boldsymbol{\mathcal{L}}_k = \int_0^1 \nabla^2_{\boldsymbol{p}}\mathcal{C}_{k+1|k}(\boldsymbol{s}+\tau\tilde{\boldsymbol{p}}_k)d\tau$ and applying $\nabla_{\boldsymbol{p}}\mathcal{C}_{k+1|k}(\boldsymbol{s}) = \boldsymbol{0}$, we have

$$\nabla_{\boldsymbol{p}}\mathcal{C}_{k+1|k}(\boldsymbol{p}_k) = \boldsymbol{\mathcal{L}}_k\tilde{\boldsymbol{p}}_k. \tag{39}$$

Substituting (39) into (37) results in

$$\tilde{\boldsymbol{p}}_{k+1} = (I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k - \delta_k\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k} + w_k. \tag{40}$$

Then, taking the squared Euclidean norm for both sides of the error dynamics (40) leads to

$$\begin{aligned}\|\tilde{\boldsymbol{p}}_{k+1}\|^2 =& \|(I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k - \delta_k\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k} + w_k\|^2 \\ =& \|(I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k\|^2 + \|w_k\|^2 + \delta_k^2\|\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}\|^2 \\ &+ 2[(I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k]^{\mathsf{T}}w_k - 2\delta_k\nabla^{\mathsf{T}}_{\boldsymbol{p}}\mathcal{P}_{k+1|k}w_k \\ &- 2\delta_k[(I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k]^{\mathsf{T}}\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}.\end{aligned} \tag{41}$$

Taking the expectation of both sides of (41) yields

$$\begin{aligned}\mathbb{E}\,\|\tilde{\boldsymbol{p}}_{k+1}\|^2 \leq& \|I_3 - \delta_k\boldsymbol{\mathcal{L}}_k\|^2\,\mathbb{E}\,\|\tilde{\boldsymbol{p}}_k\|^2 + \mathbb{E}\,\|w_k\|^2 \\ &+ \delta_k^2\,\mathbb{E}\,\|\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}\|^2 \\ &+ 2\,\mathbb{E}[[(I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k]^{\mathsf{T}}w_k] \\ &- 2\delta_k\,\mathbb{E}[\nabla^{\mathsf{T}}_{\boldsymbol{p}}\mathcal{P}_{k+1|k}w_k] \\ &- 2\delta_k\,\mathbb{E}[[(I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k]^{\mathsf{T}}\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}].\end{aligned} \tag{42}$$

It follows from Assumption 1 that

$$\begin{aligned}&\mathbb{E}[[(I_3 - \delta\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k]^{\mathsf{T}}w_k] = 0 \\ &\mathbb{E}[\nabla^{\mathsf{T}}_{\boldsymbol{p}}\mathcal{P}_{k+1|k}w_k] = 0.\end{aligned} \tag{43}$$

Moreover, for the last term in (42), we have

$$\begin{aligned}&\mathbb{E}[-2\delta_k[(I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k]^{\mathsf{T}}\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}] \\ &\quad\leq \mathbb{E}\,\|\delta_k\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}\|^2 + \mathbb{E}\,\|(I_3 - \delta_k\boldsymbol{\mathcal{L}}_k)\tilde{\boldsymbol{p}}_k\|^2 \\ &\quad= \mathbb{E}[\delta_k^2\|\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}\|^2] + \|I_3 - \delta_k\boldsymbol{\mathcal{L}}_k\|^2\,\mathbb{E}\,\|\tilde{\boldsymbol{p}}_k\|^2.\end{aligned} \tag{44}$$

For notational convenience, we denote $\mathcal{X}_k := \mathbb{E}\,\|\tilde{\boldsymbol{\Theta}}_k\|^2$ and $\mathcal{Y}_k := \mathbb{E}\,\|\tilde{\boldsymbol{p}}_k\|^2$. Therefore, combining (42), (43) and (44) gives

$$\mathcal{Y}_{k+1} \leq 2\|I_3 - \delta_k\boldsymbol{\mathcal{L}}_k\|^2\mathcal{Y}_k + 2\,\mathbb{E}[\delta^2\|\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}\|^2] + \rho^2 \tag{45}$$

According to the definition of $\mathcal{P}_{k+1|k}$, the term $2\,\mathbb{E}[\delta_k^2\|\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}\|^2]$ is determined by the error variance of estimators, which is upper bounded by

$$\mathbb{E}\,\|\tilde{\boldsymbol{\Theta}}_k\|^2 \leq \max\left\{\|\tilde{\boldsymbol{\Theta}}_0\|^2, \frac{\max_{j\in\{1,\ldots,k\}}\sum_{i=1}^N(\eta_j^i)^2q^2L^2\varrho^2}{1-\max_{j\in\{1,\ldots,k\}}\rho(A_j)}\right\} \tag{46}$$

where $\|\tilde{\boldsymbol{\Theta}}_0\|^2$ is the initial estimation error of the estimators, under the conditions specified in Lemma 2. Consequently, there always exists a bounded function $0 \leq U(\mathcal{X}_k) \leq \gamma$ such that $U(\mathcal{X}_k) = 2\,\mathbb{E}[\delta^2\|\nabla_{\boldsymbol{p}}\mathcal{P}_{k+1|k}\|^2]$. Then, (45) becomes

$$\mathcal{Y}_{k+1} \leq 2\|I_3 - \delta_k\boldsymbol{\mathcal{L}}_k\|^2\mathcal{Y}_k + U(\mathcal{X}_k) + \rho^2. \tag{47}$$

As having been proved in Lemma 2, the convergence of the estimators can be decoupled from the path planning, while it can be observed from (47) that the agent movement is related to the estimators via the term $U(\mathcal{X}_k)$. When $2\|I_3 - \delta_k\boldsymbol{\mathcal{L}}_k\|^2$ and $\rho(A)$ are within $[0,1)$, the convergence of (47) is guaranteed. By recursively iterating (47), we obtain

$$\mathcal{Y}_k \leq \bar{\alpha}^k\mathcal{Y}_0 + \sum_{j=0}^{k-1}\bar{\alpha}^j(\gamma+\rho^2). \tag{48}$$

where $\bar{\alpha} := \max_{j\in\{1,\ldots,k\}}\alpha_j$ with $\alpha_k := 2\|I_3 - \delta_j\boldsymbol{\mathcal{L}}_j\|^2$. Therefore, recalling $\mathcal{Y}_k = \mathbb{E}\,\|\tilde{\boldsymbol{p}}_k\|^2$ and $\tilde{\boldsymbol{p}}_k = \boldsymbol{p}_k - \boldsymbol{s}$, the steady-state search performance is can be obtained by

$$\lim_{k\to\infty}\mathbb{E}\,\|\boldsymbol{p}_k - \boldsymbol{s}\|^2 \leq \frac{\gamma+\rho^2}{1-\bar{\alpha}}. \tag{49}$$

This completes the proof. ■

Remark 8: The main stream of adaptive control methods can be regarded as passive learning. For example, MPC in autonomous search is targeted at navigating the agent to the source position, whereas during this pure exploitation process the estimators are updated passively by accidentally collected information from the environment. Recently, there are a wide range of engineering problems involved in balancing between exploration and exploitation, e.g. machine learning, control and decision-making with uncertain information [46–50]. In control society, related works are usually focused on stochastic model predictive control with active learning [40]. A similar concept is referred to as active reinforcement learning in artificial intelligence [50, 51]. Nevertheless, there is a critical distinction between previous works and the proposed COAL-DCEE framework. In existing dual control formulation, the probing effect is introduced to learn the *system* states or parameters (see, e.g. MPC with active learning [14, 52] and active adaptive control [53, 54]), while in our formulation the probing effect is used to actively explore the operational *environment*. We believe that future autonomous control should be able to deal with not only system uncertainty but also environment uncertainty.

TABLE I: Operational parameters and environmental knowledge.

	Search agent	Source	Prior
Measurement budget	250	-	-
Flight budget	$3,000$s	-	-
Sampling time	10s	-	-
Velocity	2m/s	-	-
Maximum step size	4m	-	-
Start position	$[2,2]$	-	-
x position	-	80m	$U(x_{min}, x_{max})$
y position	-	80m	$U(y_{min}, y_{max})$
Release rate	-	10g/s	$N(11, 2)$

IV. Empirical Results and Discussions

In this section, we will first implement the proposed COAL-DCEE algorithm using simulated data to validate the effectiveness of estimation and path planning. In particular, the existing methods, including CLEE [31], Entrotaxis [10], MPC [17] and DCEE [7], will be employed to demonstrate the advantages of COAL-DCEE. Then, an experimental dataset will be utilised to test the algorithm feasibility in real search problems. The dataset was collected by COANDA Research and Development Corporation, and supplied by the DST group [55].

A. Numerical Study

1) Simulation setup: To achieve fair comparisons using different algorithms, we keep agent parameters and source settings the same, which are summarised in Table I. The search area is confined in 100m × 100m. The rest of the parameters in the isotropic model (2) are set as follows: the wind speed $u_s = 4\text{m/s}$, wind direction $\phi_s = 1.5\pi$ rad, diffusivity $\zeta_{s1} = 1$, the particle lifetime $\zeta_{s2} = 20$. The number of measurement budget is set as 250 times, and the sampling time for taking one measurement is 10s to ensure stable sensor readings. The search agent is initialised at position $[2, 2]$m while the source is located at $[80, 80]$m, which is unknown to the search agent. The memory integer in parameter adaptation (12) is set as $q = 1$ to reduce computational cost. We also assume that there is no prior information regarding the source position, i.e. uniformly distributed in the area of interest.

2) Comparison on estimation and tracking performance: We have implemented six different algorithms: concentration-driven and position-driven CLEE [31], Entrotaxis [10], MPC [17], DCEE [7] and COAL-DCEE proposed in this paper. For clarity, key features of those algorithms are summarised in Table II. Entrotaxis and DCEE utilise Bayesian inference engine for source term estimation, which is computationally expensive due to the interaction between high-level path planner and large-scale particle filters. Currently, a fixed length of movement with a limited set of directions is deployed to maintain computational tractability. Our previous work on CLEE introduces multiple estimators for source acquisition, where each of them is independent. We observed that some of the estimators may be trapped into local optimal estimation, which consequently deteriorates the estimation performance. Note that initially CLEE was proposed using a concentration-driven formulation, which is quite sensitive to sensor noises, and thereby we further introduce a position-driven CLEE for fair comparison with our new algorithm in this paper (recalling discussions in Remark 2). The proposed framework COAL-DCEE not only holds high computational efficiency as CLEE, but also shares superior resilience to sensor noises as Entrotaxis and DCEE.

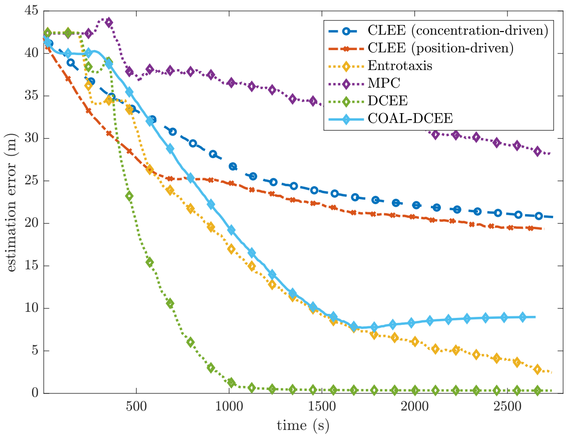


Fig. 2: Estimation performance using different algorithms.

In practical situation, the sensor readings are often intermittent due to turbulence disturbance and sensor characteristics. We assume that there is a 40% chance that the sensor encounters a non-detection event, i.e. fails to return any meaningful measurement. To achieve reliable comparison, each algorithm has been repeated for 200 times with the same settings. Source estimation performance and tracking performance are respectively shown in Figs. 2 and 3 using different algorithms. In general, all algorithms can gradually acquire the source terms and approach the source location. Due to high rate of sensor dropouts, the performance of CLEE has been significantly deteriorated. This observation coincides with our previous findings in [31]. While Entrotaxis demonstrates decent estimation performance, it falls short in terms of tracking performance. Conversely, the exploitative MPC method, which relies on passive learning, produces inadequate estimation results. Notably, COAL-DCEE exhibits superior steady-state tracking performance compared to CLEE, MPC, and Entrotaxis, as shown in Fig. 3.

3) Comparison on computational efficiency: It is worth highlighting that COAL-DCEE is significantly more efficient than DCEE and Entrotaxis, as reflected by the number of estimators/particles needed in Table II. To quantitatively demonstrate this important aspect, the total computational time for conducting 200 trials is plotted in Fig. 4. In this simulation study, COAL-DCEE only requires 29.15s, which is considerably less than the time 2697.66s and 2827.96s taken by Entrotaxis and DCEE, respectively. Computational time should be distinguished from the time consumed in a search mission: the former one is a reflection of the computational burden of a search algorithm, while the latter is determined by the travel distance/speed of the search agent and sampling time of the sensors.

TABLE II: Features of different algorithms.

	Algorithm	Estimators/Particles	Movement size	Movement direction	Exploitation/Exploration	Learning mechanism
1	CLEE (concentration-driven)	100	arbitrary	arbitrary	dual	independent estimators
2	CLEE (position-driven)	100	arbitrary	arbitrary	dual	independent estimators
3	Entrotaxis	$10,000$	2m	$[0°, 45°, \ldots, 315°]$	exploration	Bayesian filters
4	MPC	$10,000$	2m	$[0°, 45°, \ldots, 315°]$	exploitation	Bayesian filters
5	DCEE	$10,000$	2m	$[0°, 45°, \ldots, 315°]$	dual	Bayesian filters
6	COAL-DCEE	100	arbitrary	arbitrary	dual	cooperative estimators

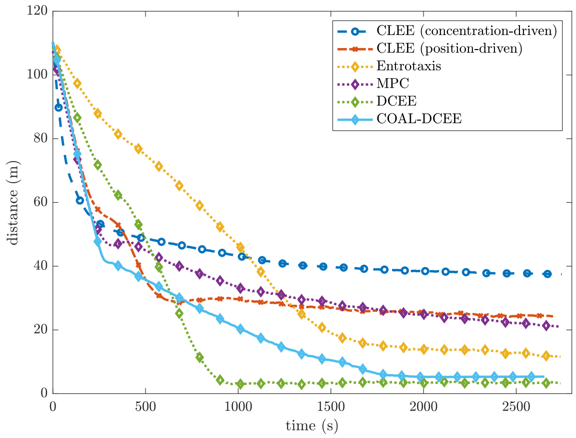


Fig. 3: Tracking performance using different algorithms.

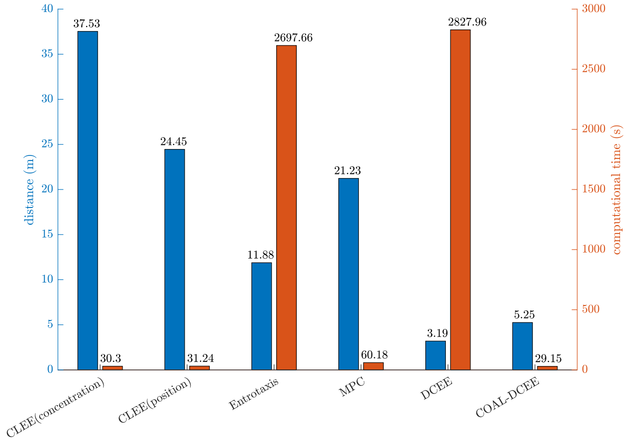


Fig. 4: Comparison of steady-state tracking performance and computational time over 200 trials.

B. Experimental Study

1) Experimental setup: In this subsection, we test the proposed COAL-DCEE algorithm using a real experimental dataset from DST group [55]. This is a quite challenging dataset for autonomous search due to rapid changes of the dispersion filed. More detailed descriptions of the experiment settings can be found in [10, 55]. The dataset is composed of a total number of 340 sequential frames, where each of them consists of 49×98 pixels. As shown in previous subsection, DCEE demonstrates superior performance among the existing solutions, and thus is deployed in this experimental study for comparison.

2) Comparison on search behaviour and performance: Figs. 5 and 6 show representative search paths using COAL-DCEE and DCEE algorithms, respectively. Although their search paths are different, both algorithms can gradually acquire the source position and navigate the search agent move towards the source. Apparently, COAL-DCEE adopts much less estimators (100) to learn the source terms compared with particles used in DCEE $(10,000)$, as demonstrated by the green dots in Figs. 5 and 6 [1].

During the experimental studies, we notice that both algorithms might fail to complete the search task. A mission is classified as a success search if both the estimated source location and agent's position are less than 10 unit length to the true source before exhausting 340 samples; otherwise it is categorised as a fail search. In Table III, we have listed the success rates, where COAL-DCEE obtains 98% and DCEE achieves 99% success rate. In Figs. 7 and 8, we provide two failure examples using COAL-DCEE and DCEE, respectively. Both trials are classified as search failure due to exhaustion of measurement budget, that is, they fail to accomplish the search task within the measurement budget (340 sequential samples). However, their searching behaviours are quite different. In Fig. 7, COAL-DCEE converges to a local optimal solution (near to the true source) while DCEE exhibits a random search path over the space until sample budget is approached as shown in Fig. 8.

3) Comparison on search time and computational efficiency: It is worth mentioning that COAL-DCEE only uses 84.80 samples (frames) in average to complete the search task, while DCEE requires 119.26 samples (averaged over 200 trials). As can be seen from the search paths in Fig. 6, the movement of DCEE is quite random due to the use of stochastic particle filters, which may consequently lead to much wasted effort during the searching process. In order to compare the search time, i.e., time required for sensor sampling and agent movement, we set the sampling time as 10s and the agent's speed as 2 cells/s. The average mission times using COAL-DCEE and DCEE are 1328.7s and 1560.9s, respectively. Therefore, in this scenario, COAL-DCEE completes the search mission faster in average. An important feature of COAL-DCEE is its computational efficiency. We compare the time consumed by running both algorithms in Matlab using a processor of 2.8 GHz Quad-Core Intel Core i7. Table III shows that COAL-DCEE only needs 15.26s to finish 200 searches, whereas DCEE requires 1.5587×10^3s.

[1] Video clips of the experimental study have been included in supplementary materials.

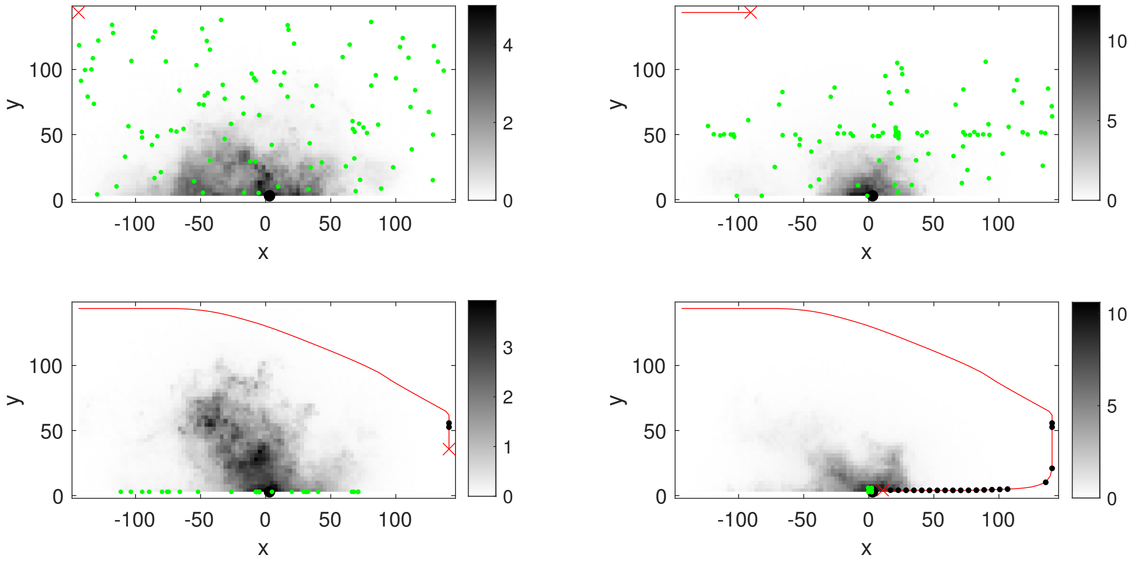


Fig. 5: Representative search path of COAL-DCEE on real dataset. The sub-figures are taken at 1, 10, 60 and 86 sample instances, respectively. The grey-scale shade depicts the instantaneous concentration field at the current time step. Red lines are the paths of the search agent, the green dots represent the estimated source position, and the black dots represent non-zero measurements.

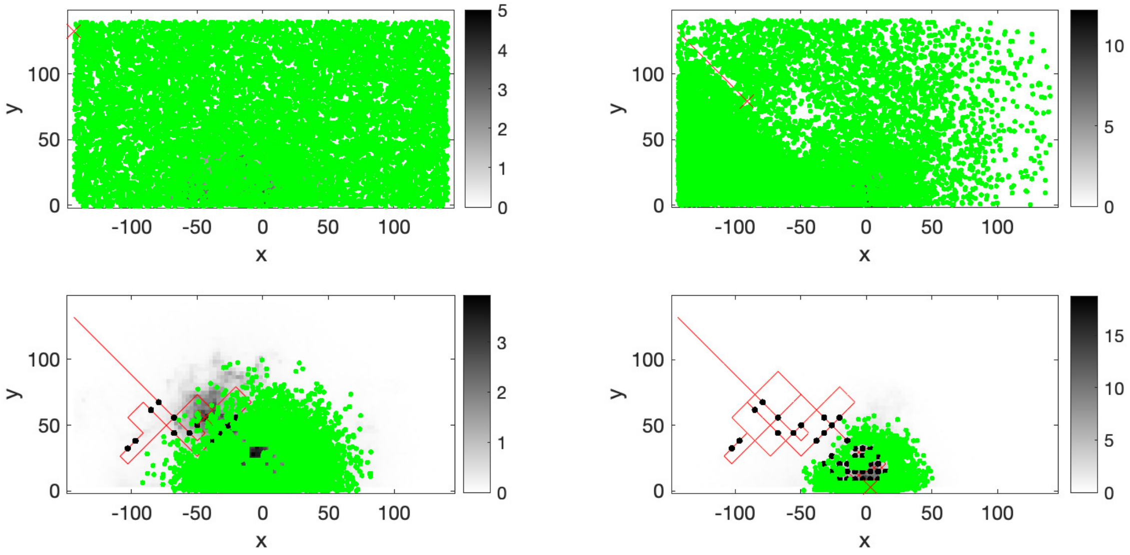


Fig. 6: Representative search path of DCEE on real dataset. The sub-figures are taken at 1, 10, 60 and 115 sample instances, respectively.

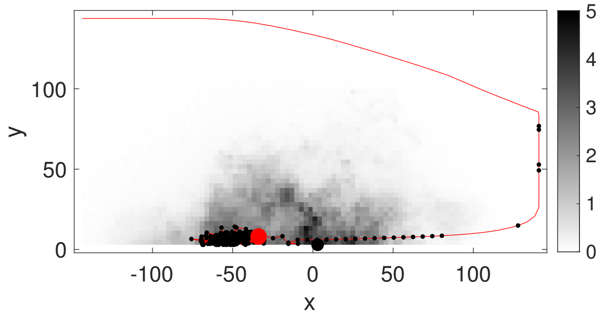


Fig. 7: An illustrative example of search fail using COAL-DCEE.

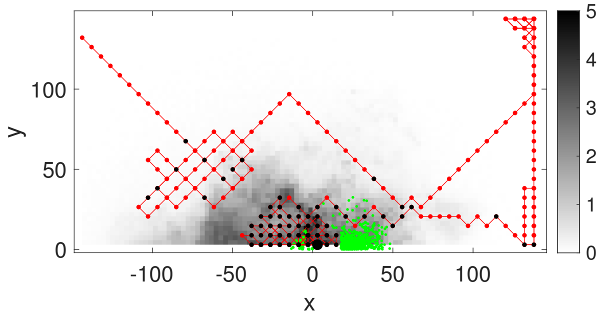


Fig. 8: An illustrative example of search fail using DCEE.

COAL-DCEE consumes less than 1% of the time used for DCEE.

C. Discussions and Reflections

Entrotaxis and other informative path planning methods direct the agent to the most informative positions, maximising the information gain. However, as these information-theoretic approaches do not include any tracking index in their objective functions, they often result in poor tracking performance, known as the pure exploration strategy. Although the pure exploitative MPC is generally efficient due to the use of passive learning methods, it yields unsatisfactory results in terms of learning and tracking performance. Passive learning does not need to quantify the information gain for all possible actions, which can be computationally burdensome for active learning based approaches. Due to the high uncertainties and disturbances in autonomous search problems, active learning has been considered an effective way to mitigate these influences [7].

TABLE III: Performance comparison between COAL-DCEE and DCEE over 200 trials.

	COAL-DCEE	DCEE
Success rate	98%	99%
Average number of samples	84.80	119.26
Average mission time	1328.7s	1560.9s
Computational time	15.26s	1.5587×10^3s

DCEE balances the control efforts between active exploration and exploitation, resulting in superior estimation and tracking performance, albeit at the expense of computational burden. In contrast, the proposed COAL-DCEE achieves balanced performance in terms of source estimation and tracking while maintaining high computational efficiency. COAL-DCEE only requires about 1% of the computational time used by DCEE, which is achieved by adopting significantly fewer estimators (100) compared to the particles used in DCEE (10,000). This computational efficiency, coupled with resilience to sensor noises and turbulence disturbances, makes it an ideal solution for future development of autonomous search using portable processors on mobile platforms, and facilitates a wider application of this advanced learning and control concept in challenging environments, where computational resources and sensor reliability may be limited.

V. Conclusion

In this paper, we have developed a computationally efficient algorithm for unknown source seeking using autonomous vehicles equipped with onboard sensors and processors. A cooperative learning based dual control is proposed using multiple dynamic estimators for source term estimation and path planning. In our formulation, the control effort achieves a balanced trade-off between reducing estimation uncertainty (exploration) and making use of the current belief to navigate the agent towards the source (exploitation). Under reasonable assumptions on sensor noises and turbulences, the convergence and steady-state performance of the proposed COAL-DCEE have been rigorously analysed. Extensive simulation and experimental results have demonstrated that COAL-DCEE not only achieves comparable performance as existing Bayesian filtering based methods, but also holds high computational efficiency. The computational tractability of the proposed COAL-DCEE is crucial for developing multi-stage dual control. In our opinion, such a far-sighted control will greatly promote the advancement of intelligent goal-oriented systems under uncertainties and constraints, and catalyse more practical applications in the near future.

References

[1] M. Hutchinson, H. Oh, and W.-H. Chen, “A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors,” *Information Fusion*, vol. 36, pp. 130–148, 2017.

[2] B. K. Patle, G. Babu L, A. Pandey, D. R. K. Parhi, and A. Jagadeesh, “A review: On path planning strategies for navigation of mobile robot,” *Defence Technology*, vol. 15, no. 4, pp. 582–606, 2019.

[3] M. Vergassola, E. Villermaux, and B. I. Shraiman, “Infotaxis as a strategy for searching without gradients,” *Nature*, vol. 445, no. 7126, pp. 406–409, 2007.

[4] M. Hutchinson, C. Liu, and W.-H. Chen, “Information-based search for an atmospheric release using a mobile robot: Algorithm and experiments,” *IEEE Transactions on Control Systems Technology*, vol. 27, no. 6, pp. 2388–2402, 2018.

[5] M. Hutchinson, C. Liu, and W.-H. Chen, “Source term estimation of a hazardous airborne release using an unmanned aerial vehicle,” *Journal of Field Robotics*, vol. 36, no. 4, pp. 797–817, 2019.

[6] M. Hutchinson, C. Liu, P. Thomas, and W.-H. Chen, “Unmanned aerial vehicle-based hazardous materials response: Information-theoretic hazardous source search and reconstruction,” *IEEE Robotics & Automation Magazine*, vol. 27, no. 3, pp. 108–119, 2019.

[7] W.-H. Chen, C. Rhodes, and C. Liu, "Dual control for exploitation and exploration (DCEE) in autonomous search," *Automatica*, vol. 133, no. 109851, 2021.

[8] C. Kreucher, A. O. Hero, and K. Kastella, "A comparison of task driven and information driven sensor management for target tracking," in *Proceedings of the 44th IEEE Conference on Decision and Control*. IEEE, 2005, Conference Proceedings, pp. 4004–4009.

[9] S.-J. Liu and M. Krstic, "Stochastic source seeking for nonholonomic unicycle," *Automatica*, vol. 46, no. 9, pp. 1443–1453, 2010.

[10] M. Hutchinson, H. Oh, and W.-H. Chen, "Entrotaxis as a strategy for autonomous search and source reconstruction in turbulent conditions," *Information Fusion*, vol. 42, pp. 179–189, 2018.

[11] A. Loisy and C. Eloy, "Searching for a source without gradients: how good is infotaxis and how to beat it," *Proceedings of the Royal Society A*, vol. 478, no. 2262, pp. 1–28, 2022.

[12] C. Rhodes, C. Liu, and W.-H. Chen, "Autonomous source term estimation in unknown environments: From a dual control concept to UAV deployment," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2274–2281, 2022.

[13] Z. Ding, "Adaptive consensus output regulation of a class of nonlinear systems with unknown high-frequency gain," *Automatica*, vol. 51, pp. 348–355, 2015.

[14] K. Zhang, Q. Sun, and Y. Shi, "Trajectory tracking control of autonomous ground vehicles using adaptive learning mpc," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 12, pp. 5554–5564, 2021.

[15] J. Yang, C. Liu, M. Coombes, Y. Yan, and W.-H. Chen, "Optimal path following for small fixed-wing UAVs under wind disturbances," *IEEE Transactions on Control Systems Technology*, vol. 29, no. 3, pp. 996–1008, 2020.

[16] W.-H. Chen, D. J. Ballance, and J. O'Reilly, "Model predictive control of nonlinear systems: computational burden and stability," *IEE Proceedings-Control Theory and Applications*, vol. 147, no. 4, pp. 387–394, 2000.

[17] Z. Li, W.-H. Chen, and J. Yang, "A dual control perspective for exploration and exploitation in autonomous search," in *2022 European Control Conference (ECC)*, 2022, Conference Proceedings, pp. 1876–1881.

[18] 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.

[19] S. Li, J. Yang, W.-H. Chen, and X. Chen, "Generalized extended state observer based control for systems with mismatched uncertainties," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 12, pp. 4792–4802, 2011.

[20] W.-H. Chen, J. Yang, L. Guo, and S. Li, "Disturbance-observer-based control and related methods: An overview," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 2, pp. 1083–1095, 2015.

[21] G. C. Goodwin and D. Q. Mayne, "A parameter estimation perspective of continuous time model reference adaptive control," *Automatica*, vol. 23, no. 1, pp. 57–70, 1987.

[22] G. Wu, R. Mallipeddi, and P. N. Suganthan, "Ensemble strategies for population-based optimization algorithms – A survey," *Swarm and Evolutionary Computation*, vol. 44, pp. 695–711, 2019.

[23] Y. Jin, M. Olhofer, and B. Sendhoff, "A framework for evolutionary optimization with approximate fitness functions," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 5, pp. 481–494, 2002.

[24] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5872–5881.

[25] X. Cao, W. Liu, and I. W. Tsang, "Data-efficient learning via minimizing hyperspherical energy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 422–13 437, 2023.

[26] X. Cao and I. W. Tsang, "Shattering distribution for active learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 1, pp. 215–228, 2022.

[27] P. Antsaklis, "Autonomy and metrics of autonomy," *Annual Reviews in Control*, 2020.

[28] X. Nie, Z. Deng, M. He, M. Fan, and Z. Tang, "Online active continual learning for robotic lifelong object recognition," *IEEE Transactions on Neural Networks and Learning Systems*, early access, 2023, doi: 10.1109/TNNLS.2023.3308900.

[29] D. Yuan, X. Chang, Q. Liu, Y. Yang, D. Wang, M. Shu, Z. He, and G. Shi, "Active learning for deep visual tracking," *IEEE Transactions on Neural Networks and Learning Systems*, early access, 2023, doi: 10.1109/TNNLS.2023.3266837.

[30] J. Li, T. Le, and E. Shlizerman, "AL-SAR: Active learning for skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, early access, 2023, doi: 10.1109/TNNLS.2023.3297853.

[31] Z. Li, W.-H. Chen, and J. Yang, "Concurrent active learning in autonomous airborne source search: Dual control for exploration and exploitation," *IEEE Transactions on Automatic Control*, vol. 68, pp. 3123–3130, 2023.

[32] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 4759–4770.

[33] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[34] N. S. Holmes and L. Morawska, "A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available," *Atmospheric Environment*, vol. 40, no. 30, pp. 5902–5928, 2006.

[35] G. Kowadlo and R. A. Russell, "Robot odor localization: A taxonomy and survey," *The International Journal of Robotics Research*, vol. 27, no. 8, pp. 869–894, 2008.

[36] A. Dhariwal, G. S. Sukhatme, and A. A. Requicha, "Bacterium-inspired robots for environmental monitoring," in *IEEE International Conference on Robotics and Automation*. IEEE, 2004, Conference Proceedings, pp. 1436–1443.

[37] P. Ojeda, J. Monroy, and J. Gonzalez-Jimenez, "Information-driven gas source localization exploiting gas and wind local measurements for autonomous mobile robots," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1320–1326, 2021.

[38] R. Ortega, V. Nikiforov, and D. Gerasimov, "On modified parameter estimators for identification and adaptive control. a unified framework and some new schemes," *Annual Reviews in Control*, vol. 50, pp. 278–293, 2020.

[39] T. A. N. Heirung, B. E. Ydstie, and B. Foss, "Dual adaptive model predictive control," *Automatica*, vol. 80, pp. 340–348, 2017.

[40] A. Mesbah, "Stochastic model predictive control with active uncertainty learning: A survey on dual control," *Annual Reviews in Control*, vol. 45, pp. 107–117, 2018.

[41] B. Wittenmark, "Adaptive dual control methods: An overview," *Adaptive Systems in Control and Signal Processing*, pp. 67–72, 1995.

[42] F. Ding and T. Chen, "Performance analysis of multi-innovation gradient type identification methods," *Automatica*, vol. 43, no. 1, pp. 1–14, 2007.

[43] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York, NY, USA: McGraw-hill, 1976.

[44] M. Guay and T. Zhang, "Adaptive extremum seeking control of nonlinear dynamic systems with parametric uncertainties," *Automatica*, vol. 39, no. 7, pp. 1283–1293, 2003.

[45] Z. Li, W.-H. Chen, J. Yang, and Y. Yan, "AID-RL: Active information-directed reinforcement learning for autonomous source seeking and estimation," *Neurocomputing*, vol. 544, p. 126281, 2023.

[46] Y. Cui, W. Yao, Q. Li, A. B. Chan, and C. J. Xue, "Accelerating monte carlo bayesian prediction via approximating predictive uncertainty over the simplex," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 4, pp. 1492–1506, 2020.

[47] Y. Bar-Shalom and E. Tse, "Dual effect, certainty equivalence, and separation in stochastic control," *IEEE Transactions on Automatic Control*, vol. 19, no. 5, pp. 494–500, 1974.

[48] E. Tse and Y. Bar-Shalom, "An actively adaptive control for linear systems with random parameters via the dual control approach," *IEEE Transactions on Automatic Control*, vol. 18, no. 2, pp. 109–117, 1973.

[49] A. I. Cowen-Rivers, D. Palenicek, V. Moens, M. A. Abdullah, A. Sootla, J. Wang, and H. Bou-Ammar, "Samba: Safe model-based & active reinforcement learning," *Machine Learning*, vol. 111, no. 1, pp. 173–203, 2022.

[50] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar, "Bayesian reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 8, no. 5-6, pp. 359–483, 2015.

[51] H. Jeong, B. Schlotfeldt, H. Hassani, M. Morari, D. D. Lee, and G. J. Pappas, "Learning Q-network for active information acquisition," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6822–6827.

[52] A. Mesbah, "Stochastic model predictive control: An overview and perspectives for future research," *IEEE Control Systems Magazine*, vol. 36, no. 6, pp. 30–44, 2016.

[53] M. K. Bugeja, S. G. Fabri, and L. Camilleri, "Dual adaptive dynamic control of mobile robots using neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp.

129–141, 2008.

[54] T. Alpcan and I. Shames, “An information-based learning approach to dual control,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 2736–2748, 2015.

[55] B. Ristic, A. Skvortsov, and A. Gunatilaka, “A study of cognitive strategies for an autonomous search,” *Information Fusion*, vol. 28, pp. 1–9, 2016.

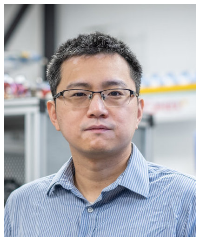
Zhongguo Li (Member, IEEE) received the B.Eng. and Ph.D. degrees in electrical and electronic engineering from the University of Manchester, Manchester, U.K., in 2017 and 2021, respectively.

He is currently a Lecturer in Robotics, Control, Communication and AI at the University of Manchester. Before joining Manchester, he was a Lecturer at University College London and a Research Associate at Loughborough University. His research interests include optimisation and decision-making for advanced control, distributed algorithm development for game and learning in network connected multi-agent systems, and their applications in robotics and autonomous vehicles.

Wen-Hua Chen (Fellow, IEEE) 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 (Fellow, IEEE) received the B.Sc. degree in automation from the Department of Automatic Control, Northeastern University, Shenyang, China, in 2006, and the Ph.D. degree in control theory and control engineering from the School of Automation, Southeast University, Nanjing, China, in 2011.

He joined the Department of Aeronautical and Automotive Engineering at Loughborough University in 2020 as a Senior Lecturer and was promoted to a Reader in 2023. His research interests include disturbance estimation and compensation, and advanced control theory and its application to mechatronic control systems 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, IET and AAIA.



Cunjia Liu (Member, IEEE) received the Ph.D. degree in autonomous vehicle control from Loughborough University in 2011.

He became an academic with Loughborough University in 2013 and is currently a Professor of Robotics and Autonomous Systems. His research cuts across the boundary between information fusion and control engineering with a focus on robotics and autonomous systems and their applications in defence, security, environment monitoring, and precision agriculture. He has published more than 100 peer reviewed papers and currently serves as an Associate Editor for IEEE Robotics and Automation and IEEE Robotics & Automation Magazine.