Dual Control of Exploration and Exploitation for Auto-Optimisation Control with Active Learning

Dual Control of Exploration and Exploitation for Auto-Optimisation Control with Active Learning

Zhongguo Li, *Member, IEEE*, Wen-Hua Chen, *Fellow, IEEE*
Jun Yang, *Fellow, IEEE*, Yunda Yan, *Member, IEEE*

***Abstract*—The quest for optimal operation in environments with unknowns and uncertainties is highly desirable but critically challenging across numerous fields. This paper develops a dual control framework for exploration and exploitation (DCEE) to solve an auto-optimisation problem in such complex settings. In general, there is a fundamental conflict between tracking an unknown optimal operational condition and parameter identification. The DCEE framework stands out by eliminating the need for additional perturbation signals, a common requirement in existing adaptive control methods. Instead, it inherently incorporates an exploration mechanism, actively probing the uncertain environment to diminish belief uncertainty. An ensemble based multi-estimator approach is developed to learn the environmental parameters and in the meanwhile quantify the estimation uncertainty in real time. The control action is devised with dual effects, which not only minimises the tracking error between the current state and the believed unknown optimal operational condition but also reduces belief uncertainty by proactively exploring the environment. Formal properties of the proposed DCEE framework like convergence are established. A numerical example is used to validate the effectiveness of the proposed DCEE. Simulation results for maximum power point tracking are provided to further demonstrate the potential of this new framework in real world applications.**

***Note to Practitioners*—In numerous engineering applications, it is highly desirable to operate a system to improve the efficiency, enhance performance or save energy. However, attaining this optimal control is a challenging task, due to the presence of unknown system and/or environment parameters. We develop a principled approach to balance between exploration and exploitation, involving active learning to estimate unknown parameters and tracking the optimal operational condition based on current estimation. This paper provides a unified framework to solve general auto-optimisation control problems. The simulation results demonstrate that the proposed method outperforms existing methods in terms of efficiency and optimality for maximum power point tracking problem, and it can be readily implemented for many other engineering problems. Future research include generalising the proposed method to nonlinear systems, as well as exploring novel applications to facilitate the widespread adoption of our method.**

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.

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.-H. Chen and J. Yang 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).

Y. Yan is with Department of Computer Science, University College London, London, WC1E 6BT, U.K. (email: yunda.yan@ucl.ac.uk).

***Index Terms*—Adaptation and control, dual control, auto-optimisation control, active learning, exploration and exploitation.**

I. Introduction

Traditionally, adaptive control algorithms are mostly designed for either regulation problems with known setpoints or tracking problems with known reference trajectories [1]. In many applications, setpoints or references are usually dependent on unknown or changing environment parameters, and thus cannot be pre-specified in advance. Operating a system at optimal condition is strongly desirable for best profit, productivity or efficiency, but it can be particularly challenging in an unknown or changing environment due to the presence of uncertainties, disturbances and noises. Typical examples include anti-lock braking systems to maintain maximal friction under various unknown road surfaces and vehicle conditions [2], maximum power point tracking to continuously deliver the highest possible power to the load in presence of variations in environments [3], [4]. In this paper, these examples are referred to as auto-optimisation control problems.

As a classic control problem with a wide range of applications, early solution for static optimal operation can be traced as far back as 1922 [5]. It was popular in 1950s and 1960s, and regained significant attention since 2000s due to a solid theoretical foundation established for the stability and performance in [6], [7]. Several approaches have been proposed under different names including extremum seeking control [6], [8] and hill-climbing systems [9]. In the literature, the concept of self-optimisation control is introduced as a distinct control design methodology [10]. It centers on identifying a physical or virtual variable that remains constant despite disturbances or uncertainties when the system reaches its optimal state. This concept slightly differs from auto-optimisation control, which aims to maintain the system at an optimal setpoint. This setpoint optimises a performance function that may vary with unknown or changing environmental parameters, effectively adapting to uncertainties, disturbances and noises. Since the optimal operation is unknown and possibly changes during the operation, a control system must be able to adapt to unknown or changing environments, for example, by means of learning, adaptation and action through limited interactions between the system and its operational environment. Then, the control system devises possible strategies to track the estimated setpoints or references based on its perceived environment knowledge and the level of confidence. This type of extremum seeking and learning problem has been identified one of the key emerging methodologies in “Control for societal-scale challenges: Road map 2030” in the field of control systems [11]. Additionally, an optimal balance between exploration and exploitation is deemed as a crucial aspect for learning and control problems, which we will discussed in the sequel.

Generally speaking, there are dual objectives in an auto-optimisation control problem in an unknown and uncertain environment: *parameter identification* and *optimality tracking*. Quite often, the dual objectives are conflicting in the sense that new observations do not provide sufficient information for identifying the unknown parameters when the system state settles to some local optimal solutions. This phenomenon widely exists in adaptive extremum seeking when an extreme searching algorithm converges to its local optimal solution, the identifiability will naturally loss due to the lack of persistent excitation (PE). As a trade-off, dither perturbations are introduced on purpose to sustain the identifiability, but such dithers inevitably deteriorate the tracking performance. Various approaches have been proposed to design the dither signals, e.g., sinusoidal perturbations [6], [12], stochastic perturbations [13], [14] and decaying perturbations [15]. However, they are usually pre-specified, and thereby cannot make online adjustments according to real-time inference performance. In other words, active learning cannot be embedded, that is, actively generate data for the purpose of learning.

This paper proposes a new approach to auto-optimisation control by embedding *active learning* from a new perspective: dual control of exploration and exploitation (DCEE). DCEE was originally proposed in [16] for autonomous search of sources of atmospheric release where the source location and other environmental factors are unknown. To realise autonomous search, it proposes each move of the robotic agent shall have dual effects: driving the agent towards the believed location of the source (exploitation) and probing the environment to reduce the level of uncertainty of the current belief (exploration). An optimal autonomous search strategy is realised by optimally trading-off these two effects. We argue in this paper that DCEE is actually applicable to a much wider range of systems that operate in an unknown or uncertain environment without well-defined control specifications, e.g., the reward or cost functions are unknown. We present a new auto-optimisation control framework by extending DCEE from a specific autonomous search application to a general design approach for achieving or maintaining optimal operation in an unknown environment. Furthermore, the DCEE scheme developed in this paper is different from the classic dual control in handling the two intricate coupling elements of the *system* and the *environment*. Existing dual control approaches impose a probing effect on the *system* itself, for example, state estimation in stochastic control [17], [18] and parameter estimation in adaptive control [19], [20]. On the other hand, the dual effect introduced in our formulation is used to explore the operational *environment*, as our objective is to acquire a better understanding of the unknown environment such that the agent is able to approach the true optimal operational condition. This exploration approach bears resemblance to the mechanisms encountered in reinforcement learning, where the focus is also on the interactive interplay between the system and its environment [21], [22].

The contribution of this paper is of twofold. On one side, for auto-optimisation control problems, we propose a new and systematic framework which is able to actively probe the environment to reduce the level of uncertainty through active learning. There is no need to artificially introduce perturbation as in the current extremum seeking control. It also provides an optimal transition from any initial operation condition to acquire the unknown optimal operation condition in terms of a reformulated objective conditional upon current knowledge and future predicted information. By formulating the auto-optimisation control in this framework, it enables to establish proven properties by getting access to a wide range of theoretic tools in control theory such as parameter adaptation and optimal control. On the other side, we generalise and extend the DCEE concept from a specific application, where specific system dynamics, reward function and properties are considered, to a general control system problem. A systematic design procedure for general descriptions of the system and control objectives is presented. We show that DCEE provides a powerful and promising framework to design control systems operating in an uncertain environment, which is an important feature of autonomous systems.

Compared with all the existing schemes for auto-optimisation control, our approach is most related to the work where the model based approach is adopted and the uncertainty of the objective or system dynamics are parameterised by uncertain parameters [6], [8], [10], [17], [23]. There are three main features in the new DCEE based auto-optimisation control framework, detailed as follows.

1) The proposed method embeds an active learning effect allowing the system to actively explore the unknown environment to reduce the level of uncertainty. Instead of using computationally expensive particle filters in information-driven methods, this paper develops an efficient multi-estimator based ensemble approach to quantify the estimation uncertainty online, based on which the controller effectively trades off between *exploration and exploitation* to balance the dual objectives of identification and tracking.
2) The ensemble based estimation method advocated in this paper is distinct from those probabilistic or dynamic ensemble estimation approaches dedicated for machine learning problems. Existing active learning based algorithms [24]–[26] 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 auto-optimisation control problem. The proposed multi-estimator based ensemble approach makes use of the environment model and the learned parameters are physically meaningful in control problems.
3) Different from all the existing schemes where probing effect is artificially introduced or inserted (usually by means of dithers and perturbations), the probing effect *naturally* occurs depending on the confidence of the estimation by assembling the outcomes of these individual estimators.

In order to guide the reader through the content, we provide a brief summary outlining the structure of this paper. In Section II, we formulate the auto-optimisation control problem and demonstrate the dual effects embedded in the new formulation. In Section III, an active learning based ensemble approach is developed for unknown environment acquisition and then a dual controller for exploration and exploitation is designed to achieve optimal trade-off between parameter identification and optimality tracking for a special single integrator system. Section IV extends DCEE to general linear systems and formal properties of the proposed auto-optimisation control method are established. Section V demonstrates the effectiveness of the proposed algorithm using a numerical example. Section VI formulates maximum power point tracking (MPPT) problem as an auto-optimisation control problem and compares the proposed algorithm with other existing approaches. Section VII concludes this paper.

II. Problem Statement

In this section, we elaborate the dual effects embedded in the reformulated auto-optimisation control problem. Then, an ensemble active learning based approach is introduced to realise efficient parameter adaptation and assess the estimation performance.

A. Dual Control Reformulation

Consider a reward function for a system operating in an unknown environment

$$J(\theta, y) = \phi^{\mathsf{T}}(y)\theta \tag{1}$$

where $\theta \subset \mathbb{R}^m$ is unknown, depending on the operational environment, $y \in \mathbb{R}^q$ is the system output, and $\phi(y) \in \mathbb{R}^m$ is the basis function of the reward function. In other words, the reward function is parameterised by unknown θ^*. The reward function (1) accommodates a broad spectrum of functions through either first-principle modelling or function approximation methods. Specifically, first-principle modeling can be used to construct various engineering objectives, such as quadratic or polynomial functions with unknown coefficients. Moreover, it can also encompass neural network approximations. In these cases, the neural network's activation functions, like radial basis function (RBF) networks or Gaussian Kernels, act as the basis $\phi(y)$, and the network's weights correspond to the unknown parameters θ. Therefore, the considered reward function demonstrates wide applicability in a range of scenarios.

Without loss of generality, it is assumed the the optimal condition is achieved at the maximum of J. An auto-optimisation control is designed to automatically drive the system to the unknown operational condition, maintain there despite disturbances and automatically adjust the optimal operation condition accordingly when the operational environment changes.

The system dynamics under consideration are described by

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) \end{aligned} \tag{2}$$

where $x(k) \in \mathbb{R}^n$, $u(k) \in \mathbb{R}^p$ and $y(k) \in \mathbb{R}^q$ are system state, control input and output, respectively, and $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times p}, C \in \mathbb{R}^{q\times n}$ are constant matrices. Suppose that at each time, the system output and the reward $J(k)$ can be measured or derived subject to measurement noise $v(k)$. The measurement information at the kth step is given by

$$z(k) = [x(k); y(k); J(k) + v(k)] \tag{3}$$

and the information state is denoted as

$$I_k = [u(k-1); z(k)] \tag{4}$$

All the measurement up to the current time k is given by

$$\mathbf{I}_k = [I_0, I_1, \ldots, I_k] \tag{5}$$

with $I_0 = [z(0)]$.

There are two ways to formulate this problem using the dual control for exploration and exploitation (DCEE) concept. The first approach is similar to extremum seeking control [8], [23] aiming to select the control such that the reward function is maximised with all the information up to now including the prior and all the measurements

$$\max_{u(k)\in\mathbb{R}^p} \mathbb{E}_{\theta, I_{k+1|k}}\{J(\theta, y(k+1|k))|\mathbf{I}_{k+1|k}\} \tag{6}$$

subject to the system dynamics (2), where $\mathbf{I}_{k+1|k} = [\mathbf{I}_k, I_{k+1|k}]$ with $I_{k+1|k} = [u(k), z(k+1|k)]$. $z(k+1|k)$ consists of the predicted output $y(k+1|k)$ and the predicted reward function under the control $u(k)$.

Another approach is to drive the system output to the unknown optimal condition directly. Since the optimal operation condition is unknown, the best one can do is to drive the system to the best estimation of the optimal operation condition with all the information up to now. This can be formulated as

$$\min_{u(k)\in\mathbb{R}^p} \mathbb{E}\{(y(k+1|k) - r^*)^{\mathsf{T}}(y(k+1|k) - r^*)|\mathbf{I}_{k+1|k}\} \tag{7}$$

where $r^* = l(\theta^*)$ denotes the predicted optimal operational condition conditional upon $\mathbf{I}_{k+1|k}$, which is a function of the environment parameter θ^*. In realm of auto-optimisation control, it is often required that the mapping $l(\theta)$ is a smooth function of θ and $r^* = l(\theta^*)$ is a unique optimum of the objective function [6].

These two problems have been solved previously in autonomous search [16], [27]. The research question is how to extend these results from this specific application to general auto-optimisation control problems. In this paper, we will focus our attention on the latter formulation in (7), which is related to the operational condition determined by unknown environment parameters.

Before proceeding further, we demonstrate that the control input $u(k)$ obtained by minimising (7) *naturally* carries dual effects corresponding to exploration and exploitation, respectively. Intuitively, the control input $u(k)$ will influence the future system state $y(k+1|k)$ via the system dynamics (2), and at the same time affect the future information to be collected $I_{k+1|k}$ via the reward function in (1) from the environment subject to uncertainties.

We define the predicted nominal operational condition as

$$\bar{r}(k+1|k) = \mathbb{E}\left[r(k+1|k)|\mathbf{I}_{k+1|k}\right] \tag{8}$$

based on which the prediction error conditional on $\mathbf{I}_{k+1|k}$ can be written as

$$\tilde{r}(k+1|k) = r^* - \bar{r}(k+1|k). \tag{9}$$

Expanding (7) and substituting (8) and (9) into (7), we have

$$\begin{aligned} &\mathbb{E}\left[\|y(k+1|k) - \bar{r}(k+1|k) - \tilde{r}(k+1|k)\|^2|\mathbf{I}_{k+1|k}\right] \\ =&\,\mathbb{E}\left[\|y(k+1|k) - \bar{r}(k+1|k)\|^2|\mathbf{I}_{k+1|k}\right] \\ &- 2\,\mathbb{E}\left[(y(k+1|k) - \bar{r}(k+1|k))^{\mathsf{T}}\tilde{r}(k+1|k)|\mathbf{I}_{k+1|k}\right] \\ &+ \mathbb{E}\left[\|\tilde{r}(k+1|k)\|^2|\mathbf{I}_{k+1|k}\right]. \end{aligned} \tag{10}$$

It follows from the definition of $\tilde{r}(k+1|k)$ in (9) that $\mathbb{E}\left[\tilde{r}(k+1|k)|\mathbf{I}_{k+1|k}\right] = 0$. Thus, by further noting that $y(k+1|k)$ and $\bar{r}(k+1|k)$ are deterministic, the cross term in (10) equals to zero, yielding

$$\begin{aligned} D(u(k)) := &\mathbb{E}\left[\|y(k+1|k)-\bar{r}(k+1|k)\|^2|\mathbf{I}_{k+1|k}\right] \\ &+\mathbb{E}\left[\|\tilde{r}(k+1|k)\|^2|\mathbf{I}_{k+1|k}\right]. \end{aligned} \tag{11}$$

Remark 1: The objective function in (11) exhibits dual effects. Minimising the first term in (11) drives the system state to estimated nominal value, which corresponds to the exploitation effect. In control terminology, it can be understood as tracking a nominal reference, thus also referred to as optimality tracking. The second term characterises the level of uncertainty (variance in this case) associated with the predicted optimal operational condition, which is related to the exploration effect. According to the classic dual control concept [28], [29], a control input is said to have dual effects if it can affect at least one jth-order central moment of a state variable ($j > 1$), in addition to its effect on the state. In fact, the dual control framework developed in this paper is a generalisation of the classic one [28] in the sense that our formulation deals with not only system uncertainty but also environment uncertainty (the operational condition $r^* = l(\theta^*)$ is determined by the environment parameters θ^*). This subtle difference endows the system with capability of exploring the operational environment and in the meanwhile exploiting its current belief. Moreover, this paper provides a solution framework to auto-optimisation control problems with complete theoretical analysis, whereas [28] considers adaptive control problems without formal analysis. Recently, DCEE has demonstrated superior and promising performance in autonomous search [16], [30].

Remark 2: According to [31], 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. As a result, 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 actively explore the operational environment to reduce knowledge uncertainty. Effective trading-off between exploration and exploitation has been a long standing issue, particularly in artificial intelligence, control and decision-making in complex and uncertain environment. In control society, some recent works explicitly introduce trade-off coefficients to incorporate the exploration terms into model predictive control problems, e.g., [17], [32]. This inevitably incurs tedious efforts in tuning the coefficients to balance exploration and exploitation. In view of the derivation of (11), it is clear that the dual effects in DCEE are naturally embedded, since they are derived from a physically meaningful value function in (7).

B. Ensemble based Active Learning

Efficient gradient descent algorithms can be used to estimate the unknown parameters. The performance of single estimator based optimisation algorithm is quite poor, due to noisy measurement and nonlinear modelling (see examples in autonomous search [27], [33]). Recently, the ensemble-based approximation in machine learning community has demonstrated great success with tractable computational load [25], [34]. In this paper, we develop a multi-estimator based learning method for parameter adaptation, which shows comparable performance as particle filter using much less computational resources in autonomous search application [27].

Considering an ensemble of N estimators, the dual formulation in (11) becomes

$$\begin{aligned} \min_{u(k)\in\mathbb{R}^p} \; & D(u) = \|y(k+1|k)-\bar{r}(k+1|k)\|^2 + \mathcal{P}(k+1|k) \\ \text{subject to } & x(k+1|k) = Ax(k)+Bu(k) \\ & y(k+1|k) = Cx(k+1|k) \end{aligned} \tag{12}$$

where the nominal estimate and variance of the estimated optimal condition are drawn from the ensemble, i.e.,

$$\bar{r}(k+1|k) = \frac{1}{N}\sum_{i=1}^{N} r_i(k+1|k) = \frac{1}{N}\sum_{i=1}^{N} l(\theta_i(k+1|k)) \tag{13}$$

$$\begin{aligned} \mathcal{P}(k+1|k) = &\frac{1}{N}\sum_{i=1}^{N}(r_i(k+1|k)-\bar{r}(k+1|k))^{\mathsf{T}} \\ &\times (r_i(k+1|k)-\bar{r}(k+1|k)) \end{aligned} \tag{14}$$

where the subscript $i \in \mathcal{N}$ denotes the index of the estimators, with $\mathcal{N}$ representing the set of the ensemble. Note that the relationship between the predicted optimal condition and the unknown parameter, i.e., $r_i(k+1|k) = l(\theta_i(k+1|k))$, is usually known. For example, in autonomous search application, θ^* is composed of the unknown source location and other environment parameters, like wind direction and wind speed. The optimal operation condition r^* in autonomous search is the source location, i.e., part of θ^*, which serves as a tracking reference for the search agent.

In order to estimate the unknown parameter θ^*, we apply a gradient-descent regression method [35], designed as

$$\begin{aligned} \theta_i(k) = &\theta_i(k-1) - \eta_i\phi(y(k-1)) \\ &\times \left[\phi(y(k-1))^{\mathsf{T}}\theta_i(k-1) - J(k-1)\right], \; \forall i \in \mathcal{N}. \end{aligned} \tag{15}$$

where $\eta_i > 0$ is the learning rate of the ith estimator; $J(k-1)$ denotes the observed reward with measurement noise in (3) at $y(k-1)$; and $\theta(k)$ denotes the estimate of unknown reward parameter θ^*. The estimators are randomly initialised or they can be initialised according to *a priori* pdfs of the unknown parameters if available. Denote the estimation error as $\tilde{\theta}_i(k) = \theta_i(k) - \theta^*$. Then, by noting $J(k-1) = \phi(y(k-1))^{\mathsf{T}}\theta^* + v(k-1)$, we have

$$\begin{aligned} \tilde{\theta}_i(k) = &\left[I_m - \eta_i\phi(y(k-1))\phi(y(k-1))^{\mathsf{T}}\right]\tilde{\theta}_i(k-1) \\ &- \eta_i\phi(y(k-1))v(k), \; \forall i \in \mathcal{N}. \end{aligned} \tag{16}$$

Denoting the extended parameter error as $\tilde{\Theta}(k) = \text{col}\{\tilde{\theta}_1(k), \ldots, \tilde{\theta}_N(k)\}$, where $\text{col}\{\cdot\}$ denotes a column vector formed by stacking the elements on top of each other, (16) can be written in a compact form as

$$\begin{aligned} \tilde{\Theta}(k) = &\left[I_N \otimes \left(I_m - \eta_i\phi(y(k-1))\phi(y(k-1))^{\mathsf{T}}\right)\right]\tilde{\Theta}(k-1) \\ &- \left[I_N \otimes \eta_i\phi(y(k-1))\right](1_N \otimes v(k-1)). \end{aligned} \tag{17}$$

In an ensemble-based adaptation, we take their average as the current estimation of the unknown parameters. Thus, averaging (17), we have

$$\begin{aligned}\tilde{\Theta}_{av}(k) =& \frac{1}{N}(1_N^{\mathsf{T}} \otimes I_m)\tilde{\Theta}(k)\\ =& \frac{1}{N}(1_N^{\mathsf{T}} \otimes I_m)\left[I_N \otimes (I_m - \eta\phi\phi^{\mathsf{T}})\right]\tilde{\Theta}(k-1)\\ &- \frac{1}{N}(1_N^{\mathsf{T}} \otimes I_m)\left[I_N \otimes \eta\phi\right](1_N \otimes v(k-1))\\ =& \frac{1}{N}(1_N^{\mathsf{T}} \otimes I_m)\tilde{\Theta}(k-1) - \frac{1}{N}(1_N^{\mathsf{T}} \otimes \eta\phi\phi^{\mathsf{T}})\tilde{\Theta}(k-1)\\ &- \eta\phi v(k-1).\end{aligned} \tag{18}$$

Remark 3: An important observation is that even though we have the same regressor ϕ at one time instant, its excitation impact to each estimator will be different since $\phi\phi^{\mathsf{T}}\tilde{\theta}_i \neq \phi\phi^{\mathsf{T}}\tilde{\theta}_j$, $\forall i \neq j$, almost surely. Due to the introduction of parameter extension by multiple estimators, at any time instant the average estimation can always be excited when there are sufficient estimators. In addition, by introducing a group of estimators, it is possible to evaluate and make full use of the estimation uncertainty by sampling the outcomes of the ensemble in an online manner, which is proved to be crucial in DCEE [16] as we will discuss in the sequel. Another desirable feature of the ensemble approach is its resilience to measurement noises. In view of the last term in (18), instantaneous noises will be averaged out under multiple estimators such that the overall performance of the ensemble can be improved.

III. DCEE for Single Integrator

A. Algorithm Development

In high-level decision-making, system behaviours are usually simplified as single integrators by ignoring low-level dynamics. In this paper, we begin with DCEE for this special case

$$y(k+1) = y(k) + u(k). \tag{19}$$

For general linear systems, we will use this as an internal reference generator, as will be shown later in Section IV.

With the estimated environment parameter in (15), the dual controller can be designed as

$$\begin{aligned}y(k+1) &= y(k) + u(k)\\ u(k) &= -\delta_k\left[\nabla_y \mathcal{C}(k+1|k) + \nabla_y \mathcal{P}(k+1|k)\right]\end{aligned} \tag{20}$$

where $\mathcal{C}(k+1|k) = \|y(k) - \bar{r}(k+1|k)\|^2$ denotes the exploitation term, and $\mathcal{P}(k+1|k)$ is the exploration term in the dual objective (12). According to the gradient-descent regression in (15), the predicted mean of the N ensemble $\theta_i(k+1|k)$, denoted as $\bar{\theta}(k+1|k)$, is given by

$$\begin{aligned}\bar{\theta}(k+1|k) =& \frac{1}{N}\sum_{i=1}^{N}\theta_i(k+1|k)\\ =& \frac{1}{N}\sum_{i=1}^{N}(\theta_i(k) - \eta_i F_i(k+1|k))\end{aligned} \tag{21}$$

where

$$F_i(k+1|k) = [J(\theta_i(k), y(k)) - J(k+1|k)]\phi(y(k)) \tag{22}$$

with $J(k+1|k)$ being the predicted future reward based on current belief $\{\theta_i(k), \forall i \in \mathcal{N}\}$. Note that the predicted future reward is noise-free as there is no influence from sensory devices in prediction. In this paper, we use the average of $\theta_i(k), \forall i \in \mathcal{N}$ to evaluate the predicted future reward, i.e., $J(k+1|k) = J(\bar{\theta}(k), y(k))$. Similarly, the predicted variance of the ensemble is given by

$$\mathcal{P}(k+1|k) = \text{trace}(\boldsymbol{F}(k+1|k)^{\mathsf{T}}\mathcal{P}(k|k)\boldsymbol{F}(k+1|k)) \tag{23}$$

where

$$\begin{aligned}\boldsymbol{F}(k+1|k) =& \text{col}\{F_1(k+1|k), \dots, F_N(k+1|k)\}\\ \mathcal{P}(k|k) =& \text{cov}\{\theta_i(k), \forall i \in \mathcal{N}\}\\ =& \text{diag}\{(\theta_1(k) - \bar{\theta}(k))(\theta_1(k) - \bar{\theta}(k))^{\mathsf{T}}, \dots,\\ & (\theta_N(k) - \bar{\theta}(k))(\theta_N(k) - \bar{\theta}(k))^{\mathsf{T}}\}\end{aligned} \tag{24}$$

where $\text{cov}\{\cdot\}$ is a covariance operator evaluating the covariance matrix of the ensemble, and $\text{diag}\{\cdot\}$ denotes a block-diagonal matrix by putting the elements on its main diagonal. Using the predicted mean $\bar{\theta}(k+1|k)$ and the predicted covariance $\mathcal{P}(k+1|k)$ of the unknown environmental parameter, the dual control terms in (20) can be obtained.

B. Convergence Analysis

In this section, we will examine the convergence of the proposed dual control algorithm, by leveraging parameter adaptation and optimisation techniques. To this end, we introduce some fundamental assumptions that will be used to facilitate the convergence analysis of the proposed dual control algorithm.

Assumption 1: There exist positive constants $T \in \mathbb{Z}^+$ and $\gamma \geq \beta > 0$ such that

$$\gamma I_m \geq \sum_{k=t}^{t+T}[\phi(y(k))][\phi(y(k))]^{\mathsf{T}} \geq \beta I_m > 0,\ \forall t > 0. \tag{25}$$

Assumption 2: The measurement noise $v(k)$ is independent and identically distributed with bounded variance, i.e.,

$$\begin{aligned}\mathbb{E}\left[v(k)\right] &= 0\\ \mathbb{E}\left[\|v(k)\|^2\right] &\leq \varrho^2.\end{aligned} \tag{26}$$

Assumption 3: The reward function $J(\theta, y)$ is twice differentiable and strictly concave on y for any $\theta \in \mathbb{R}^m$, that is,

$$\frac{\partial^2 J(\theta, y)}{\partial y^2} > 0. \tag{27}$$

Remark 4: Assumption 1 is a standard persistent excitation (PE) condition to ensure the identifiability of the unknown environmental parameter θ. Extensive techniques on parameter adaptation have been reported in the past few decades aiming at relaxing or fulfilling the conditions of PE [8], [35]. If we introduce a memory-based regressor extension to the parameter adaptation algorithm in (15), the PE condition can be relaxed to interval excitation [35]. Assumption 2 implies that the noises imposed on sensory information are unbiased with bounded variances [27]. Assumption 3 guarantees the existence and uniqueness of the optimal operational condition, i.e., $r^* = l(\theta^*)$, which is widely used in adaptive control and extremum seeking control [8], [36]. Note that the mapping between the optimal operational condition and parameter θ can be obtained by solving $\frac{\partial J(\theta,y)}{\partial y} = 0$.

First, we examine the convergence of the gradient-descent regression method in (15).

Theorem 1: Under Assumptions 1 and 2, there exists a constant $\eta^* > 0$ such that, for any $0 < \eta_i < \eta^*$, the estimates, $\hat{\theta}_i(k),\ \forall i \in \mathcal{N}$, converge to a bounded neighbourhood of the true environmental parameter θ^*. Moreover, the mean-square-error of the estimator is convergent and bounded by

$$\mathbb{E}\,\|\tilde{\theta}_i(k)\|^2 \leq \frac{\eta_i^2 L^2 \varrho^2}{1 - \max_{j\in\{1,\ldots,k-1\}} \|A_i(j)\|} \tag{28}$$

where $A_i(j) = [I_m - \eta_i[\phi(y(j))][\phi(y(j))]^{\mathsf{T}}]^{\mathsf{T}}[I_m - \eta_i[\phi(y(j))][\phi(y(j))]^{\mathsf{T}}]$ and L denotes the bound of the regressor ϕ. Moreover, in absence of measurement noises, $\lim_{k\to\infty} \mathbb{E}\,\|\tilde{\theta}_i(k)\|^2 = 0$.

Proof: In view of (16) and Assumption 2, the expectation of the estimate is given by

$$\mathbb{E}[\tilde{\theta}_i(k)] = \big[I_m - \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}}\big]\, \mathbb{E}[\tilde{\theta}_i(k-1)] \quad \forall i \in \mathcal{N}. \tag{29}$$

According to Assumption 1, there exists a constant η^* such that, for any $0 < \eta_i < \eta^*$, $0 < \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}} < I_m$. Consequently, for any $0 < \eta_i < \eta^*$, we have

$$0 < I_m - \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}} < I_m. \tag{30}$$

It follows from (29) that

$$\|\,\mathbb{E}[\tilde{\theta}_i(k)]\| \leq \big\|I_m - \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}}\big\| \times \|\,\mathbb{E}[\tilde{\theta}_i(k-1)]\|,\ \forall i \in \mathcal{N}. \tag{31}$$

Therefore,

$$\|\,\mathbb{E}[\tilde{\theta}_i(k)]\| \leq \prod_{j=1}^{k} \big\|I_m - \eta_i[\phi(y(j-1))][\phi(y(j-1))]^{\mathsf{T}}\big\| \times \|\,\mathbb{E}[\tilde{\theta}_i(0)]\|,\ \forall i \in \mathcal{N}. \tag{32}$$

For any bounded error $\tilde{\theta}_i(0)$, the expectation of the estimator converge to zero.

Moreover, the variance of the estimators can be bounded under Assumption 2. Taking the squared Euclidean norm of (16) yields

$$\begin{aligned}\|\tilde{\theta}_i(k)\|^2 =& \big\|\big[I_m - \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}}\big]\tilde{\theta}_i(k-1)\big\|^2 \\ &+ \|\eta_i \phi(y(k-1))v(k)\|^2 \\ &- 2\big[[I_m - \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}}] \\ &\times \tilde{\theta}_i(k-1)\big]\,[\eta_i\phi(y(k-1))v(k)]\,,\ \forall i \in \mathcal{N}.\end{aligned} \tag{33}$$

Applying expectation operation to (33) leads to

$$\begin{aligned}\mathbb{E}\,\|\tilde{\theta}_i(k)\|^2 =& \mathbb{E}\,\big\|\,\big[I_m - \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}}\big] \\ &\times \tilde{\theta}_i(k-1)\big\|^2 \\ &+ \mathbb{E}\,\|\eta_i\phi(y(k-1))v(k)\|^2,\ \forall i \in \mathcal{N}.\end{aligned} \tag{34}$$

where $\mathbb{E}\,[v(k)] = 0$ has been used to eliminate the cross term. Denoting $A_i(k-1) = [I_m - \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}}]^{\mathsf{T}}[I_m - \eta_i[\phi(y(k-1))][\phi(y(k-1))]^{\mathsf{T}}]$ and applying the variance bound in (26), we have

$$\mathbb{E}\,\|\tilde{\theta}_i(k)\|^2 \leq \mathbb{E}\,\big\|\tilde{\theta}_i(k-1)\big\|^2_{A_i(k-1)} + \eta_i^2 L^2 \varrho^2. \tag{35}$$

For any $0 < \eta_i < \eta^*$, the mean-square-error of the estimator is convergent and bounded by

$$\lim_{k\to\infty} \mathbb{E}\,\|\tilde{\theta}_i(k)\|^2 \leq \frac{\eta_i^2 L^2 \varrho^2}{1 - \max_{j\in\{1,\ldots,k-1\}} \|A_i(j)\|}. \tag{36}$$

In absence of measurement noise $v(k) = 0$, $\lim_{k\to\infty} \mathbb{E}\,\|\tilde{\theta}_i(k)\|^2 = 0$. This completes the proof. ■

Remark 5: Theorem 1 establishes the convergence of the estimators. The parameter adaptation algorithm together with its convergence analysis under measurement noises forms a new feature of this paper since existing studies mainly focus on noise-free scenarios [35], [37]. As having been discussed in Remark 4, PE is a standard and commonly-used condition to guarantee the convergence of parameter estimators. Despite significant research efforts have been dedicated to explore weak/alternative assumptions, very few result has been obtained (see recent survey in [35]). In the proposed dual controller (20), a probing effort is inherently embedded aiming to reduce the estimation uncertainty. Such an exploration effect from active learning is beneficial to environment acquisition, which has been validated in autonomous search application [16], [27].

Remark 6: The proposed multi-estimator assisted ensemble method for environment adaptation is a hybrid approach that combines both model-based and model-free techniques. The model-based estimators are trained according to the model structures of the reward function in (1). A model-free ensemble approximation is used to estimate the mean and variance of the unknown environmental parameters in an online manner. It is widely perceived in machine learning community that model-based approach benefits from high learning efficiency due to the utilisation of model knowledge but inevitably inherits model biased errors; on the other hand, model-free approach provides a reliable way to quantify the level of estimation uncertainty but may incur additional computational burden. Recently, the hybrid method has demonstrated superior performance in simulation and experiment in machine learning due to its combined strength from both model-based and model-free learning [25], [34], [38]. Theoretical guarantee on convergence and performance of the hybrid approach has not been well-established but mainly verified by extensive simulation and experimental results. Inspired by its recent success, we develop a concurrent active learning based ensemble algorithm and establish its formal properties in this paper. Additionally, different from existing studies in active control [24], [25] where Bayesian neural networks or ensembles of dynamic models are employed to formulate information gain, the proposed DCEE method captures belief uncertainty using multi-estimator ensemble, from which physically meaningful parameters can be extracted.

Denote the tracking error between current state and unknown optimal condition r^* as $\tilde{y}(k) = y(k) - r^*$. Then, it follows from (20) that

$$\tilde{y}(k+1) = \tilde{y}(k) - \delta_k\big[\nabla_y \mathcal{C}(k+1|k) + \nabla_y \mathcal{P}(k+1|k)\big]. \tag{37}$$

Now, we analyse the convergence to the optimal operational condition.

Theorem 2: Under Assumptions 1-3, for any $0 < \eta_i < \eta^*$, y converges to a bounded neighbourhood of the optimal operational condition $r^* = l(\theta^*)$ if there exists a step size

δ_k such that $0 < 2\|[I_n - \delta_k \mathcal{L}(k)]\|^2 < 1$ with $\mathcal{L}(k) = \int_0^1 \nabla_y^2 \mathcal{C}(r^* + \tau\tilde{y}(k), \bar{r}(k+1|k))d\tau$.

Proof: To relate the gradient term $\nabla_y \mathcal{C}(k+1|k)$ with $\tilde{y}(k)$, we recall the mean value theorem [39], that is, for a twice-differentiable function $h(y) : \mathbb{R}^m \to \mathbb{R}$,

$$\nabla h(y_1) = \nabla h(y_2) + \left[\int_0^1 \nabla^2 h[y_2 + \tau(y_1 - y_2)]d\tau\right](y_1 - y_2), \quad \forall y_1, y_2 \in \mathbb{R}^m. \tag{38}$$

Thus, we have

$$\begin{aligned}\nabla_y \mathcal{C}(y(k), \bar{r}(k+1|k)) &= \nabla_y \mathcal{C}(r^*, \bar{r}(k+1|k)) \\ &+ \left[\int_0^1 \nabla_y^2 \mathcal{C}(r^* + \tau\tilde{y}(k), \bar{r}(k+1|k))d\tau\right]\tilde{y}(k)\end{aligned} \tag{39}$$

where we have expanded the notation $\mathcal{C}(k+1|k)$ for clarity. Denoting $\mathcal{L}(k) = \int_0^1 \nabla_y^2 \mathcal{C}(r^* + \tau\tilde{y}(k), \bar{r}(k+1|k))d\tau$ and applying $\nabla_y \mathcal{C}(r^*, \bar{r}(k+1|k)) = \mathbf{0}$, we have

$$\nabla_y \mathcal{C}(y(k), \bar{r}(k+1|k)) = \mathcal{L}(k)\tilde{y}(k). \tag{40}$$

Applying (40) to (37) results in

$$\tilde{y}(k+1) = [I_n - \delta_k \mathcal{L}(k)]\tilde{y}(k) - \delta_k \nabla_y \mathcal{P}(k+1|k). \tag{41}$$

To examine the boundedness of the tracking error, we take the Euclidean norm for both sides of (41), yielding

$$\begin{aligned}\|\tilde{y}(k+1)\|^2 =& \|[I_n - \delta_k \mathcal{L}(k)]\tilde{y}(k)\|^2 + \|\delta_k \nabla_y \mathcal{P}(k+1|k)\|^2 \\ &- 2\delta_k[(I_n - \delta_k \mathcal{L}(k))\tilde{y}(k)]^\mathsf{T} \nabla_y \mathcal{P}(k+1|k).\end{aligned} \tag{42}$$

Taking the expectation of (42) leads to

$$\begin{aligned}\mathbb{E}\,\|\tilde{y}(k+1)\|^2 \le& \|[I_n - \delta_k \mathcal{L}(k)]\|^2\, \mathbb{E}\,\|\tilde{y}(k)\|^2 \\ &+ \mathbb{E}\,\|\delta_k \nabla_y \mathcal{P}(k+1|k)\|^2 \\ &+ \mathbb{E}[-2\delta_k \nabla_y^\mathsf{T} \mathcal{P}(k+1|k)[(I_n - \delta_k \mathcal{L}(k))\tilde{y}(k)]].\end{aligned} \tag{43}$$

The last term in (43) can be written as

$$\begin{aligned}&\mathbb{E}[-2\delta_k \nabla_y^\mathsf{T} \mathcal{P}(k+1|k)[(I_n - \delta_k \mathcal{L}(k))\tilde{y}(k)]] \\ &\le \|[I_n - \delta_k \mathcal{L}(k)]\|^2\, \mathbb{E}\,\|\tilde{y}(k)\|^2 + \mathbb{E}\,\|\delta_k \nabla_y \mathcal{P}(k+1|k)\|^2.\end{aligned} \tag{44}$$

Therefore, substituting (44) into (43) results in

$$\begin{aligned}\mathbb{E}\,\|\tilde{y}(k+1)\|^2 \le& 2\|[I_n - \delta_k \mathcal{L}(k)]\|^2\, \mathbb{E}\,\|\tilde{y}(k)\|^2 \\ &+ 2\,\mathbb{E}\,\|\delta_k \nabla_y \mathcal{P}(k+1|k)\|^2.\end{aligned} \tag{45}$$

From Theorem 1, the estimation errors are bounded within

$$\mathbb{E}\,\|\tilde{\theta}_i(k)\|^2 \le \max\left\{\|\tilde{\theta}_i(0)\|^2, \frac{\eta_i^2 L^2 \varrho^2}{1 - \max_{j\in\{1,\ldots,k-1\}} \rho(A_i(j))}\right\}. \tag{46}$$

As a result, $0 \le \mathbb{E}\,\|\delta_k \nabla_y \mathcal{P}(k+1|k)\|^2 \le \mu$ is upper bounded, since it is a measure of covariance of the bounded estimators. Consequently, we have

$$\mathbb{E}\,\|\tilde{y}(k+1)\|^2 \le 2\|[I_n - \delta_k \mathcal{L}(k)]\|^2\, \mathbb{E}\,\|\tilde{y}(k)\|^2 + \mu. \tag{47}$$

If there exists a step size δ_k such that $0 < 2\|[I_n - \delta_k \mathcal{L}(k)]\|^2 < 1$, then the expected mean square of the tracking error is convergent. Recursively iterating (47) gives

$$\mathbb{E}\,\|\tilde{y}(k+1)\|^2 \le \bar{\alpha}^k\, \mathbb{E}\,\|\tilde{y}(0)\|^2 + \sum_{j=0}^{k-1} \bar{\alpha}^j \mu \tag{48}$$

where $\bar{\alpha} := \max_{j\in\{1,\ldots,k\}} \alpha_j$ with $0 < \alpha_k := 2\|[I_n - \delta_j \mathcal{L}(j)]\|^2 < 1$. Since $\lim_{k\to\infty} \bar{\alpha}^k\, \mathbb{E}\,\|\tilde{y}(0)\|^2 \to 0$, we have

$$\lim_{k\to\infty} \mathbb{E}\,\|y(k) - r^*\|^2 \le \frac{\mu}{1 - \bar{\alpha}}. \tag{49}$$

This completes the proof. ■

Remark 7: In general, traditional adaptive control can be regarded as passive learning [8], [17], [40] where parameter estimators are updated by accidentally collected data samples. 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 concentration measurements from the environment [16], [41]. Recently, there are a wide range of engineering problems involved in balancing between exploration and exploitation, e.g., machine learning, control and decision-making in uncertain environment [29], [42]–[44]. In control society, related works are usually focused on stochastic model predictive control with active learning [17]. A similar concept is referred to as active reinforcement learning in artificial intelligence [44], [45]. Nevertheless, there is a critical distinction between previous works and the proposed DCEE framework for auto-optimisation control. In existing dual control formulation, the probing effect is introduced to learn the *system* states or parameters (e.g. MPC with active learning [46] and active adaptive control [19], [47]), 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 [16], [31].

IV. DCEE for Linear Systems

In this section, we deal with general linear systems. As the environment estimators are designed by information measurements, the parameter adaptation algorithm in (15) can be used and Theorem 1 remains valid. Now, we design a dual controller that regulates the system output $y(k)$ to minimise the reformulated objective function defined in (12).

The dual controller is proposed as

$$u(k) = -Kx(k) + (G + K\Psi)\xi(k) \tag{50}$$

where the reference $\xi(k)$ is generated by

$$\begin{aligned}\xi(k) &= \xi(k-1) + \psi(k-1) \\ \psi(k) &= -\delta_k\left[\nabla_\xi \mathcal{C}(k+1|k) + \nabla_\xi \mathcal{P}(k+1|k)\right]\end{aligned} \tag{51}$$

where G and Ψ are gain matrices obtained by solving

$$\begin{aligned}(A - I)\Psi + BG &= 0 \\ C\Psi - I &= 0.\end{aligned} \tag{52}$$

and K is chosen such that $A - BK$ is Schur stable as (A, B) is controllable. Note that $\psi(k)$ is exactly the dual gradient term used in the integrator dynamics in Section III. For linear systems, the control input $u(k)$ not only needs to have dual effects for exploration and exploitation but additionally requires control effort to stabilise the system dynamics as in (50).

Assumption 4: The pair (A, B) is controllable, and

$$\text{rank}\begin{bmatrix} A - I & B \\ C & 0 \end{bmatrix} = n + q. \tag{53}$$

Remark 8: The dual control design in (50)-(52) is partly inspired by conventional internal model approaches [48]. The solvability of (52) is guaranteed by (53), which is widely known as regulation equations [48]. The existence of Ψ ensures the existence of optimal state $x^* = \Psi r^*$ such that $Cx^* = r^*$.

Define state transformations $x_s(k) = \Psi\xi(k)$, $u_s(k) = G\xi(k)$. Let $\bar{x}(k) = x(k) - x_s(k)$ and $\bar{u}(k) = u(k) - u_s(k)$. Applying the transformation to the system dynamics (2) leads to

$$\begin{aligned} \bar{x}(k+1) &= x(k+1) - x_s(k+1) \\ &= Ax(k) + Bu(k) - \Psi(\xi(k) + \psi(k)) \\ &= A\bar{x}(k) + B\bar{u}(k) - \Psi\psi(k) \\ e(k) &= C\bar{x}(k) \end{aligned} \tag{54}$$

where (52) has been used to derive above dynamics. Applying the control input (50), we have the closed loop dynamics

$$\begin{aligned} \bar{x}(k+1) &= (A - BK)\bar{x}(k) - \Psi\psi(k) \\ e(k) &= C\bar{x}(k). \end{aligned} \tag{55}$$

The following lemma can be regarded as input-to-output stability of the transformed dynamics (55) by viewing $\psi(k)$ and $e(k)$ as the input and output, respectively.

Lemma 1: Let Assumptions 1–4 hold. Suppose the conditions specified in Theorems 1–2 hold. If the gain matrices G and Ψ are designed according to (52) and K is chosen such that $(A - BK)$ is Schur stable, then

$$\limsup_{k\to\infty} \|e(k)\| \leq \frac{\|C\|\|\Psi\|}{1 - \|A - BK\|} \limsup_{k\to\infty} \|\psi(k)\|. \tag{56}$$

Proof: Putting (52) into a matrix form leads to

$$\begin{bmatrix} A - I & B \\ C & 0 \end{bmatrix} \begin{bmatrix} \Psi \\ G \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix} \tag{57}$$

of which the solvability is guaranteed under (53) in Assumption 4 by transforming the matrix equation (57) to standard linear algebraic equations. For notational convenience, we denote $A_c = A - BK$ and $B_c = -\Psi$. Then, we have

$$\bar{x}(k+1) = A_c\bar{x}(k) + B_c\psi(k). \tag{58}$$

Recursively iterating (58) results in

$$\bar{x}(k) = A_c^k\bar{x}(0) + \sum_{j=0}^{k-1} A_c^{k-j-1}B_c\psi(j). \tag{59}$$

Hence, we have

$$e(k) = C\bar{x}(k) = CA_c^k\bar{x}(0) + C\sum_{j=0}^{k-1} A_c^{k-j-1}B_c\psi(j). \tag{60}$$

Because A_c is Schur, we have $\lim_{k\to\infty} CA_c^k\bar{x}(0) = 0$.

The convergence of reference generator (51) has been established in Theorem 2, and thereby $\psi(k)$, i.e., the gradient of the dual controller, is bounded and converges to zero as $k \to \infty$. Denoting $\varpi := \limsup_{k\to\infty} \|\psi(k)\|$, it can be obtained that, for any small constant $\epsilon > 0$, there exists a positive time index $\zeta > 0$ such that

$$\|\psi(k)\| < \varpi + \epsilon, \ \forall k > \zeta. \tag{61}$$

Now, the second term in (60) can be separated into two parts, written as

$$\begin{aligned} C\sum_{j=0}^{k-1} A_c^{k-j-1}B_c\psi(j) =& C\sum_{j=0}^{\zeta} A_c^{k-j-1}B_c\psi(j) \\ &+ C\sum_{j=\zeta+1}^{k-1} A_c^{k-j-1}B_c\psi(j). \end{aligned} \tag{62}$$

Taking the Euclidean norm of (62) and invoking (61), we obtain

$$\begin{aligned} \left\|C\sum_{j=0}^{k-1} A_c^{k-j-1}B_c\psi(j)\right\| &= \|C\|\|B_c\|\left\|\sum_{j=0}^{\zeta} A_c^{k-j-1}\psi(j) \right. \\ &\quad \left. + \sum_{j=\zeta+1}^{k-1} A_c^{k-j-1}\psi(j)\right\| \\ \leq& \left\|A_c^{k-\zeta-1}\right\|\|C\|\|B_c\|\left\|\sum_{j=0}^{\zeta} A_c^{\zeta-j}\psi(j)\right\| \\ &+ (\varpi + \epsilon)\|C\|\|B_c\|\left\|\sum_{j=\zeta+1}^{k-1} A_c^{k-j-1}\right\|. \end{aligned} \tag{63}$$

Since A_c is Schur stable (i.e., eigenvalues of A_c are of absolute value less than one), we have

$$\lim_{k\to\infty} \left\|A_c^{k-\zeta-1}\right\| = 0. \tag{64}$$

For the sum of a geometric series, we have

$$\sum_{j=\zeta+1}^{k-1} \|A_c\|^{k-1-j} = \frac{1 - \|A_c\|^{k-\zeta}}{1 - \|A_c\|} < \frac{1}{1 - \|A_c\|}. \tag{65}$$

Therefore, combining (60) and (63) leads to

$$\limsup_{k\to\infty} \|e(k)\| \leq \frac{\|C\|\|B_c\|}{1 - \|A_c\|} (\varpi + \epsilon). \tag{66}$$

As ϵ can be set arbitrarily small, it follows from (66) that

$$\limsup_{k\to\infty} \|e(k)\| \leq \frac{\|C\|\|B_c\|}{1 - \|A_c\|} \limsup_{k\to\infty} \|\psi(k)\|. \tag{67}$$

This completes the proof. ■

Now, combining the results in Theorems 1–2 and Lemma 1, we are ready to establish the convergence of the auto-optimisation control for linear systems.

Theorem 3: Let Assumptions 1–4 hold. Suppose the conditions specified in Theorems 1–2 and Lemma 1 hold. The output $y(k)$ of the linear system (2) converges to the neighbourhood of the optimum r^*, using control input (50) together with reference generator (51).

Proof: Denoting $\tilde{x}(k) = x(k) - \Psi r^*$, we have

$$\begin{aligned} \tilde{x}(k+1) &= Ax(k) + B[-Kx(k) + (G + K\Psi)\xi(k)] - \Psi r^* \\ &= (A - BK)\tilde{x}(k) + B(G + K\Psi)(\xi(k) - r^*) \end{aligned} \tag{68}$$

It follows from Theorems 1–2 that $\xi(k)$ converges to the neighbourhood of r^* with bounded error. Thus, the result can be concluded by treating $B(G + K\Psi)(\xi(k) - r^*)$ as $\psi(k)$ in Lemma 1. ■

Remark 9: The auto-optimisation control in this paper is similar to the classic formulation of reinforcement learning in the sense that both of them are targeted to operate a system in an unknown and uncertain environment. There are two bottlenecks in widely applying reinforcement learning, particularly deep RL: one is a large number of trials are required to achieve a satisfactory performance (big data) and the other is its performance could significantly degrade if the real operational environment is different from the training environment (poor adaptiveness) [49]. DCEE establishes a new control framework to provide a promising and complementary method to reinforcement learning in control and robotics society, which can deal with uncertain environment without repetitive training. In fact, active learning for exploration and exploitation in machine intelligence can find strong evidence in human intelligence, which is supported by the biological principles in functional integration in the human brain and neuronal interactions (known as free-energy principle and active inference in neuroscience [50]). Interested readers are referred to [49] for detailed discussions.

V. Numerical Example

In this section, we verify the effectiveness of the proposed algorithm using a dedicate numerical example. Consider a linear system (2) with

$$A = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; C = \begin{bmatrix} 0 & 1 \end{bmatrix}. \tag{69}$$

The reward function is given by

$$J(\theta^*, y) = 2y - \theta^* y^2 = \begin{bmatrix} 2y & -y^2 \end{bmatrix} \begin{bmatrix} 1 \\ \theta^* \end{bmatrix} \tag{70}$$

where θ^* is affected by the unknown environment. The true value is $\theta^* = 1$ but unavailable *a priori*. The optimal operational condition r^* is determined by θ^*, i.e., $r^* = l(\theta^*) = 1/\theta^* = 1$.

We assume the measurements are subject to Gaussian noise $v(k) \sim \mathcal{N}(0, 2)$, which implies that the observations from environment are $J(k) = J(\theta^*, y(k)) + v(k)$. Decision-making under uncertain environment with noisy measurements is of significant importance to promote the system intelligence. In order to explore the uncertain environment, the first step is to quantify the level of uncertainty. An ensemble based multi-estimator approach has been developed in previous sections. Now, the size of the estimator ensemble is chosen as $N = 100$, and each of them is randomly initialised according to a uniform distribution between -10 and 10, i.e., $\theta_i(0) \sim U(-10, 10), \forall i = 1, 2, \ldots, 100$. Choosing the number of the ensemble necessitates balancing computational efficiency with stochastic accuracy. Depending on the specific problem and available resources, it is common practice to experiment with various configurations until a satisfactory balance is achieved. Indeed, those findings align with recent trends observed in machine learning community using ensemble aggregating approaches where a handful of estimators are enough to generate promising results [25], [34]. The system is controllable and regulation condition in (53) is satisfied such that the gain matrices can be obtained as $\Psi = [\frac{1}{3}, 1]^\mathsf{T}$ and $G = -\frac{2}{3}$. The gain matrix $K = [-1, 1.1]$ is chosen by placing the poles of $(A - BK)$ at $[0.4; 0.5]$.

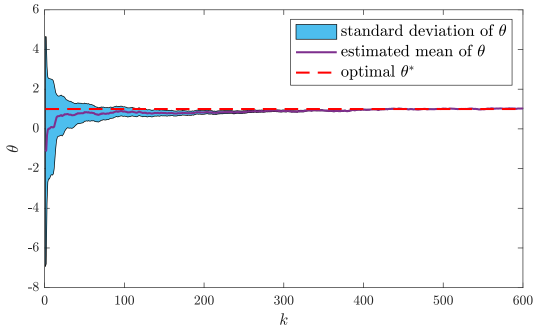
Fig. 1a shows the estimated environmental parameters. Initially, the mean and standard deviation of the ensemble $\{\theta_i, i = 1, \ldots, 100\}$ are -1.137 and 5.788, respectively, randomly initialised using a uniform distribution. The mean of the estimators converges to the true environment parameter $\theta^* = 1$, and the standard deviation among the estimators shrinks quickly, indicating that the estimation uncertainty reduces (quantified by the variance among the estimators in the ensemble). Despite increasing the iteration k significantly, the estimated parameters remain fluctuating within a small neighbourhood of the true value due to the presence of noisy measurements. Fig. 1b displays the observed rewards from the environment. Even though we have imposed quite significant noises to the measurements, the performance of the estimators is fairly satisfactory, which manifests the ensemble based active learning provides superior robustness against noises.

Implementing the dual control in (50) not only contributes to enhanced parameter adaptation performance but also drives the system output to the optimal operational condition, as shown in Fig. 1c. The system output approaches the optimal operational point $r^* = 1$ as shown in Fig. 1c, and the system states are displayed in Fig. 1d. It can be verified that $x^* = \Psi r^* = [\frac{1}{3}, 1]^\mathsf{T}$. The tracking error is determined by the estimation error. In this process, there is no need to tune the weights of exploration and exploitation. As a principled approach, the dual controller in (50) is derived from a physically meaningful objective function, which naturally embeds balanced dual effects for active environment learning and optimality tracking.

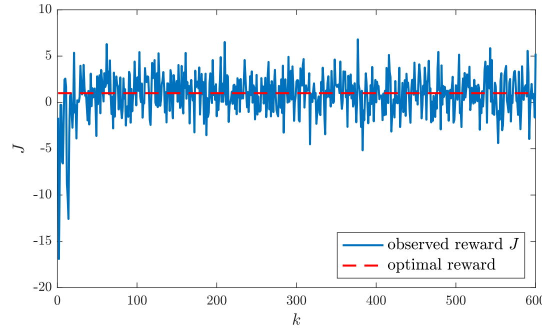
To demonstrate the impact of active learning based exploration mechanism, we have implemented certainty equivalence based auto-optimisation approach using passive learning. For fare comparisons, we have kept all parameters the same but removed the exploration effect from the control input. The obtained results are depicted in Fig. 2. A key observation to highlight is that incorporating active learning into the DCEE framework significantly enhances rapid environmental acquisition and eliminates steady-state discrepancies. These improvements are notable when compared to passive auto-optimisation control approaches, which often suffer from sub-optimal learning performance caused by inaccurate estimation result. Furthermore, our findings challenge the effectiveness of the certainty equivalence principle commonly used in classic adaptive control, particularly in complex and uncertain environments. This validation underscores the need for more advanced control strategies like DCEE in handling uncertainty.

VI. Application for MPPT

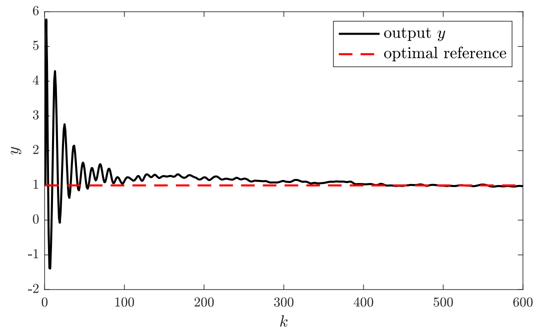
DCEE was originally developed to solve autonomous search problem in [16], which demonstrates outstanding performance compared with other existing approaches. In this section, we take the optimal control for photovoltaic (PV) systems as an example to illustrate that DCEE can be implemented to solve a much wider class of auto-optimisation control problems in real-world applications. Extracting maximum power is a long-lasting pursuit in operating PV systems. Despite significant research efforts made over the past few decades [51]–[53], the energy conversion efficiency of PV systems remains very poor due to high environment uncertainties in temperature, irradiance level, partial shading and other atmospheric conditions. The primary goal in PV operation is simply to extract solar energy as much as possible despite changing



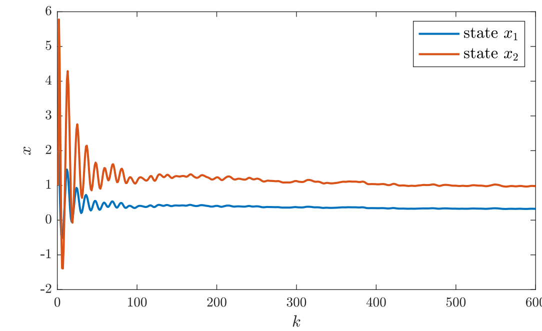
(a) Mean and standard deviation of estimated $\theta(k)$.



(b) Observed reward $J(k)$ with measurement noises $v(t)$.

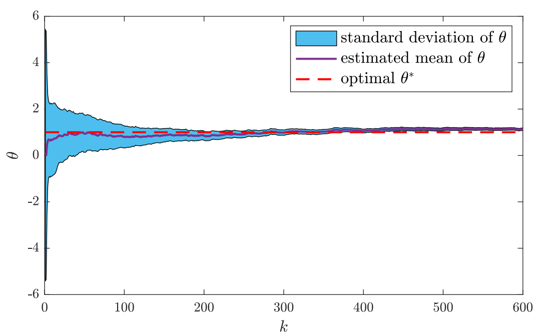


(c) System output $y(k)$.

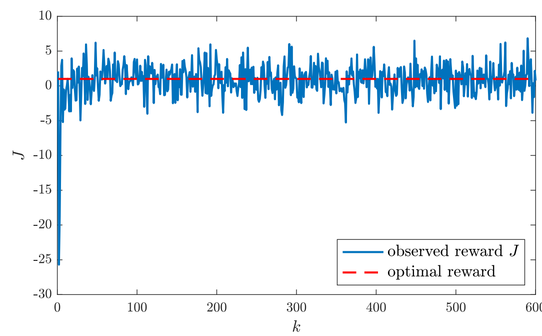


(d) System state $x(k)$.

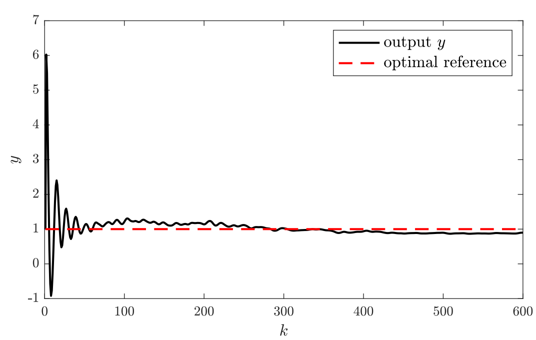
Fig. 1: Simulation results with environment uncertainties using active learning based DCEE.



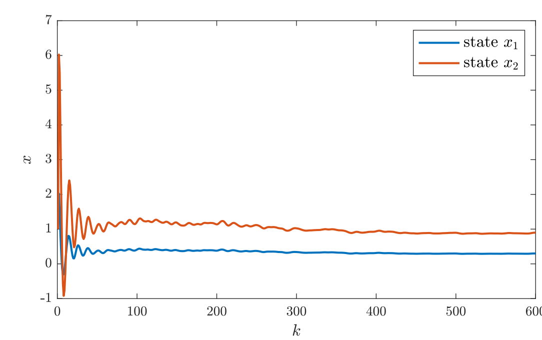
(a) Mean and standard deviation of estimated $\theta(k)$.



(b) Observed reward $J(k)$ with measurement noises $v(t)$.



(c) System output $y(k)$.



(d) System state $x(k)$.

Fig. 2: Simulation results with environment uncertainties using passive learning based auto-optimisation approach.

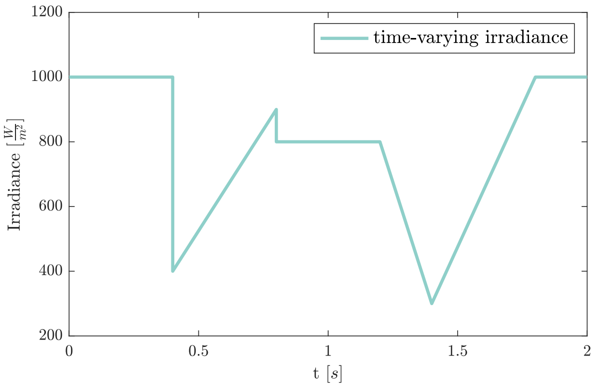


Fig. 3: Time-varying solar irradiance profile.

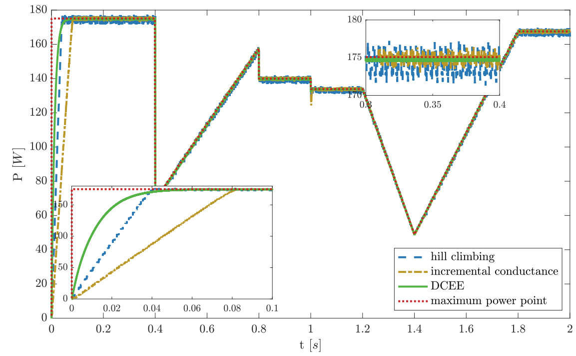


Fig. 4: Power profile using different algorithms.

operational environment, termed as maximum power point tracking (MPPT). There have been a wide variety of methods targeting to solve this problem, which can be roughly classified into three categories: offline methods, online methods, and other methods. Detailed comparisons and classifications can be found in comprehensive survey papers, e.g., [51], [52].

In this section, the proposed DCEE is implemented as an alternative approach to achieve MPPT, and two representative approaches, hill climbing method (HC) and incremental conductance method (IC), are deployed for comparison. It is worth noting that all the three algorithms can be classified as online methods. It has been widely perceived that online methods usually outperform offline counterparts in terms of conversion efficiency due to their inherent adaptiveness to changing environment. According to the curve-fitting based MPPT [51], the power and voltage (P-V) characteristics can be modelled by

$$P = \phi^{\mathsf{T}}(V)\theta \tag{71}$$

where $\phi(V)$ is the polynomial regressor $[1, V, V^2, \ldots, V^n]^{\mathsf{T}}$ and $\theta \in \mathbb{R}^{n+1}$ is the polynomial coefficient. To solve the maximum problem of (71), we need to estimate the unknown parameters θ and then maximise the power output by regulating the voltage V according to

$$V(k+1) = V(k) + u(k). \tag{72}$$

We use solar panel A10Green Technology model number A10J-S72-175 for this simulation [54]. To mimic the real operational environment of PV systems, a time-varying solar irradiance profile is stimulated as shown in Fig. 3, and the temperature is initially set as 25°C and then jumps to 35°C at $t = 1$s. It should be noted that the unknown environment parameter θ changes as the operational condition varies. Although the proposed algorithm is theoretically analysed for static parameters identification, the use of constant learning rate η_i renders the adaptation algorithm in (15) with the capability of tracking drifting parameters.

Simulation results using different algorithms (DCEE, HC and IC) are shown in Fig. 4, 5 and 6. To illustrate more detailed features of different algorithms, enlarged sub-figures are displayed for the time intervals $t \in [0, 0.1]$, and $t \in [0.3, 0.4]$. The power losses, as displayed in Fig. 7, are calculated by integrating the differences between the maximum power point and real power outputs stimulated using different algorithms. Convergence speed, sensed signals, algorithm complexity and conversion efficiency are four commonly-used criteria to assess the characteristics of MPPT techniques. According to the simulation results, we summarise and compare the features of different approaches in Table I. Conversion efficiency directly influences the energy extracted from the PV systems, which is ratio between real generated energy and maximum energy (accumulated over the simulation time interval $[0, 2]$). DCEE produces quite high efficiency (99.1%). Due to the use of perturbed signals in hill climbing method, there are very large voltage and current fluctuations in steady state. This undesirable property not only causes low conversion efficiency but also leads to fast degradation in low level electronic devices. The oscillations are partially solved by incremental conductance method, which measures incremental current and voltage changes to predict the effect of voltage change.

Different from HC, incremental inductance method is able to maintain at MPP without oscillations when there is no change in operational environment. From the simulation results using HC and IC, there is a trade-off between transient convergence speed and steady-state oscillations. The steady-state oscillation of IC is reduced at the cost of slow tracking performance, leading to larger power loss with a conversion efficiency 97.2%. It is argued that DCEE as a balanced approach is able to optimally trade-off between exploitation and exploration: when there is large uncertainty in estimated MPP, it will explore quickly to gain information to construct more accurate estimate of MPP; and when there is less change in operational environment, it will maintain at the current belief of MPP without causing large oscillations. All three algorithms need to measure voltage and current: DCEE requires voltage and power (calculated by the product of current and voltage) to construct P-V curve in (71) (i.e., reward-state mapping), while HC and IC use incremental power to decide the direction of voltage regulation. As mature MPPT techniques, both HC and IC are simple to implement using dedicated hardware devices. Since efficient ensemble approximation and gradient based control are developed in this new approach, DCEE is ready to be implemented in real PV platforms without incurring heavy computational load.

VII. Conclusion

This paper has proposed a dual control framework for exploration and exploitation, designed to address auto-optimisation

TABLE I: Features of different MPPT techniques.

	Methods	Convergence speed	Sensed variables	Algorithm complexity	Conversion efficiency
1	DCEE	Fast	Voltage and current	Medium	99.1%
2	Hill climbing	Fast	Voltage and current	Simple	98.3%
3	Incremental conductance	Medium	Voltage and current	Simple	97.2%

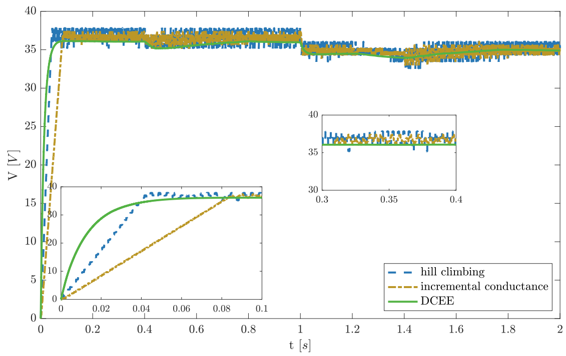


Fig. 5: Voltage profile using different algorithms.

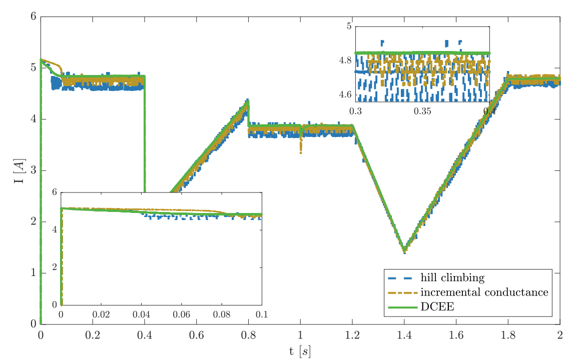


Fig. 6: Current profile using different algorithms.

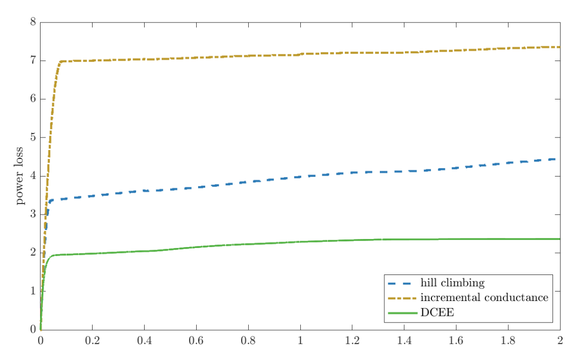


Fig. 7: Power losses using different algorithms.

control challenges in uncertain environments. The DCEE algorithm is uniquely structured to balance exploration and exploitation, resolving the inherent conflict between parameter identification and the pursuit of optimal operation. We have rigorously demonstrated that our approach ensures convergence and maintains performance, considering both the objective function and the characteristics of environmental noise. The effectiveness of the DCEE framework is further evidenced through a numerical example and its successful application in MPPT, underscoring its substantial potential for diverse real-world scenarios.

Acknowledgment

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

References

[1] H. Zheng, J. Wang, D. Shi, D. Liu, and S. Wang, "Quasi time-fuel optimal control strategy for dynamic target tracking," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 416–427, 2024.

[2] C. Zhang and R. Ordonez, "Numerical optimization-based extremum seeking control with application to ABS design," *IEEE Transactions on Automatic Control*, vol. 52, no. 3, pp. 454–467, 2007.

[3] R. Leyva, C. Alonso, I. Queinnec, A. Cid-Pastor, D. Lagrange, and L. Martinez-Salamero, "MPPT of photovoltaic systems using extremum-seeking control," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 1, pp. 249–258, 2006.

[4] J. Lekube, A. J. Garrido, and I. Garrido, "Rotational speed optimization in oscillating water column wave power plants based on maximum power point tracking," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 2, pp. 681–691, 2016.

[5] Y. Tan, W. H. Moase, C. Manzie, D. Nešić, and I. M. Mareels, "Extremum seeking from 1922 to 2010," in *Proceedings of the 29th Chinese Control Conference*. IEEE, 2010, pp. 14–26.

[6] M. Krstić and H.-H. Wang, "Stability of extremum seeking feedback for general nonlinear dynamic systems," *Automatica*, vol. 36, no. 4, pp. 595–601, 2000.

[7] M. Krstić, "Performance improvement and limitations in extremum seeking control," *Systems & Control Letters*, vol. 39, no. 5, pp. 313–326, 2000.

[8] 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.

[9] K. A. Sullivan and S. H. Jacobson, "A convergence analysis of generalized hill climbing algorithms," *IEEE Transactions on Automatic Control*, vol. 46, no. 8, pp. 1288–1293, 2001.

[10] S. Skogestad, "Plantwide control: The search for the self-optimizing control structure," *Journal of Process Control*, vol. 10, no. 5, pp. 487–507, 2000.

[11] A. Alleyne, F. Allgöwer, A. Ames, S. Amin, J. Anderson, A. Annaswamy, P. Antsaklis, N. Bagheri, H. Balakrishnan, B. Bamieh *et al.*, "Control for societal-scale challenges: Road map 2030," in *2022 IEEE CSS Workshop on Control for Societal-Scale Challenges*. IEEE Control Systems Society, 2023.

[12] Y. Tan, D. Nešić, and I. Mareels, "On non-local stability properties of extremum seeking control," *Automatica*, vol. 42, no. 6, pp. 889–903, 2006.

[13] C. Manzie and M. Krstic, "Extremum seeking with stochastic perturbations," *IEEE Transactions on Automatic Control*, vol. 54, no. 3, pp. 580–585, 2009.

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

[15] S. Xie and L. Y. Wang, "Adaptive optimization with decaying periodic dither signals," *IEEE Transactions on Automatic Control*, vol. 68, no. 2, pp. 1208–1214, 2023.

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

[17] 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.

[18] S. Chen, K. Saulnier, N. Atanasov, D. D. Lee, V. Kumar, G. J. Pappas, and M. Morari, "Approximating explicit model predictive control using constrained neural networks," in *Annual American Control Conference (ACC)*. IEEE, 2018, Conference Proceedings, pp. 1520–1527.

[19] 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.

[20] N. M. Filatov and H. Unbehauen, "Survey of adaptive dual control methods," *IEE Proceedings-Control Theory and Applications*, vol. 147, no. 1, pp. 118–128, 2000.

[21] 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.

[22] R. Chai, D. Liu, T. Liu, A. Tsourdos, Y. Xia, and S. Chai, "Deep learning-based trajectory planning and control for autonomous ground vehicle parking maneuver," *IEEE Transactions on Automation Science and Engineering*, vol. 20, no. 3, pp. 1633–1647, 2023.

[23] M. Guay and D. J. Burns, "A proportional integral extremum-seeking control approach for discrete-time nonlinear systems," *International Journal of Control*, vol. 90, no. 8, pp. 1543–1554, 2017.

[24] 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.

[25] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, vol. 31, 2018.

[26] 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, 2022.

[27] 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, no. 5, pp. 3123–3130, 2023.

[28] A. A. Feldbaum, "Dual control theory I," *Avtomatika i Telemekhanika*, vol. 21, no. 9, pp. 1240–1249, 1960.

[29] 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.

[30] 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, 2021.

[31] P. Antsaklis, "Autonomy and metrics of autonomy," *Annual Reviews in Control*, vol. 49, pp. 15–26, 2020.

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

[33] 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.

[34] 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.

[35] 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.

[36] V. Adetola and M. Guay, "Parameter convergence in adaptive extremum-seeking control," *Automatica*, vol. 43, no. 1, pp. 105–110, 2007.

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

[38] X. Yue, Y. Wen, J. H. Hunt, and J. Shi, "Active learning for gaussian process considering uncertainties with application to shape control of composite fuselage," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 1, pp. 36–46, 2020.

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

[40] R. Chai, A. Tsourdos, H. Gao, Y. Xia, and S. Chai, "Dual-loop tube-based robust model predictive attitude tracking control for spacecraft with system constraints and additive disturbances," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 4, pp. 4022–4033, 2021.

[41] 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)*. IEEE, 2022, pp. 1876–1881.

[42] 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.

[43] A. Tariverdi, U. Côté-Allard, K. Mathiassen, O. J. Elle, H. Kalvoy, Ø. G. Martinsen, and J. Torresen, "Reinforcement learning-based switching controller for a milliscale robot in a constrained environment," *IEEE Transactions on Automation Science and Engineering*, early access, 2023, doi: 10.1109/TASE.2023.3259905.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] J. Huang, *Nonlinear Output Regulation: Theory and Applications*. SIAM, 2004.

[49] W.-H. Chen, "Perspective view of autonomous control in unknown environment: Dual control for exploitation and exploration vs reinforcement learning," *Neurocomputing*, vol. 497, pp. 50–63, 2022.

[50] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[51] P. Bhatnagar and R. Nema, "Maximum power point tracking control techniques: State-of-the-art in photovoltaic applications," *Renewable and Sustainable Energy Reviews*, vol. 23, pp. 224–241, 2013.

[52] A. R. Reisi, M. H. Moradi, and S. Jamasb, "Classification and comparison of maximum power point tracking techniques for photovoltaic system: A review," *Renewable and Sustainable Energy Reviews*, vol. 19, pp. 433–443, 2013.

[53] T. Esram and P. L. Chapman, "Comparison of photovoltaic array maximum power point tracking techniques," *IEEE Transactions on Energy Conversion*, vol. 22, no. 2, pp. 439–449, 2007.

[54] I. Shams, S. Mekhilef, and K. S. Tey, "Maximum power point tracking using modified butterfly optimization algorithm for partial shading, uniform shading, and fast varying load conditions," *IEEE Transactions on Power Electronics*, vol. 36, no. 5, pp. 5569–5581, 2020.

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.

Yunda Yan (Member, IEEE) 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.

From 2020 to 2022, he was a Research Associate with the Department of Aeronautical and Automotive Engineering, Loughborough University, U.K. From 2022 to 2023, he was with the School of Engineering and Sustainable Development, De Montfort University, U.K as a Lecturer in Control Engineering and was later promoted to a Senior Lecturer. In Sep. 2023, he joined the Department of Computer Science, University College London, U.K, as a Lecturer in Robotics and AI. His current research interest focuses on the safety-critical control design for autonomous systems, especially related with optimisation and learning-based methods.