Optimal Control of Nonlinear Systems With Unsymmetrical Input Constraints and Its Application to the UAV Circumnavigation Problem

Yangguang Yu®, Xiangke Wang®, Senior Member, IEEE, Zhiyong Sun®, Member, IEEE, and Lincheng Shen®

Abstract—In this article, a novel design scheme is introduced to solve the optimal control problem for nonlinear systems with unsymmetrical and state-dependent input constraints. By introducing an initial stabilizing control policy as the baseline of the constructed optimal control policy, we remove the assumption in the current study for the adaptive optimal control, that is, the internal dynamics should hold zero when the state of the system is in the origin. Particularly, nonlinear control systems with partially unknown dynamics are investigated and the procedure to acquire the corresponding optimal control policy is presented. The stability for the closed-loop dynamics and the optimality of the obtained control policy are both proved. Besides, we apply the proposed control design framework to solve the optimal circumnavigation problem based on the accumulative Fisher information for a fixed-wing unmanned aerial vehicle (UAV). The control performance of our algorithm is compared with that of the existing circumnavigation control policy in a numerical simulation.

Index Terms—Actuator saturation, fisher information, optimal control, unmanned aerial vehicle (UAV) circumnavigation, unsymmetrical constrained input system.

I. INTRODUCTION

A. Literature Review

In the domain of automatic control, the basic requirement for the controller is to stabilize the system and drive the interested state to an equilibrium state. But when the related resource is limited or the system is required to compete for a specific performance index, the optimal behavior of the system with respect to specified long-term goals is desired. Therefore, optimal control for nonlinear systems has been the focus of the research since last century [1] as it can help improve the system performance effectively. There have been numerous successful applications of nonlinear optimal control to different fields, such as spacecraft attitude control [2] and underwater vehicle control [3]. In general, the optimal control problem for nonlinear systems involves the solving of an underlying Hamilton–Jacobi–Bellman (HJB) equation [4], which is usually very difficult to solve and almost impossible to get an analytical solution directly [5]. To solve the HJB equation, Abu-Khalaf and Lewis [6] proposed an offline policy iteration (PI) strategy, in which a sequence of cost functions were approximated. The methods proposed in [6] require that the dynamics of the system is completely known. However, the dynamics of the nonlinear system are usually complex and even time variant in some situations. As a consequence, nonlinear systems are rather difficult to be modeled accurately. For traditional model-based control design methods, the performance degradation caused by the model inaccuracy may be catastrophic.

To overcome the difficulties mentioned above, the adaptive dynamic programming (ADP) [7], [8], [9] method was developed. Different from the traditional model-based control design methods, the ADP method approximates the solution of the HJB equation using online data and further constructs the optimal control law adaptively with the system’s dynamics being partially or completely unknown [10], [11], [12]. It was proved in previous studies [11], [13] that the ADP method can guarantee the ultimate uniform boundedness (UUBs) of the system and thus the risk of system instability caused by the model inaccuracy can be avoided. For systems with partially unknown dynamics, [14], [15], [16] proposed the integral reinforcement learning (IRL) method to approximate the solution of the HJB equation. When the dynamics of the system are completely unknown, an identifier-critic-actor-based structure is usually used. A neural network (NN) [17] or recurrent NN [18] was utilized to fully identify the unknown system dynamics. Recently, the work [19] proposed a deterministic policy gradient ADP algorithm for solving model-free optimal control problems. Although the preconditions and the proposed methods in the studies mentioned above are different, there exists one hidden assumption in common among these works, that is, the internal dynamics of the system should be zero when the state of the system is in the origin [11], [12], [13], [14], [15], [16]. The system that satisfies this assumption is termed as the standard form (SF) system in this article. However, this assumption is not satisfied in many nonlinear systems, which are termed as nonstandard...
form (NF) systems in this article. For example, the control problems of target tracking [11] or unmanned aerial vehicle (UAV) circumnavigation [20] involve with the NF system. It is still an unsolved problem on how to tackle the optimal control problem for the NF system.

Another important issue that is worth considering is the amplitude limitation on the control input. There often exists an unsymmetrical and state-dependent saturation zone for the input of the system’s actuator in reality. Taking the attitude control of a vehicle or aircraft for example, the steering mechanism of a ground or aerial vehicle may partially loss effectiveness due to motor fault [21]. As a consequence, the vehicle’s maximum steering capacity for the left direction and the right direction may be different. Meanwhile, to avoid the risk of rollover, the maximum angular velocity of the vehicle is usually required to decrease with the increment of the linear velocity. To confront the optimal control problem with symmetrical input constraints, a nonquadratic cost function was proposed in [22] and a smooth saturated controller was further constructed. Similar studies are reported in the literature therein [11], [15], [23], [24]. In these studies, the input $u$ is constrained in a symmetrical and fixed set, i.e., $u$ is constrained by $|u| < \lambda$ with $\lambda$ being a positive constant. For the optimal control problem with unsymmetrical input constraint, Zhou et al. [25] proposed an ADP-based neuro-optimal controller for discrete-time nonlinear systems with asymmetric input saturation. More recently, the work [21] proposed an adaptive optimal control law by introducing a switching function. We notice that the switching function in [21] should be carefully selected to guarantee the stability of the closed-loop system. The event-triggered adaptive optimal control problem was studied in [26] and [27] for a class of asymmetrically input-constrained nonlinear systems. The work [28] constructed an optimal neurocontroller under the framework of RL and the work [29] presented an event-driven $H_{\infty}$ controller for continuous nonlinear systems with asymmetric input saturation. However, the work [26], [27], [28], [29] can only guarantee the UUB of the closed-loop system theoretically. In summary, the proposed methods in the existing studies still exist some shortcomings. Meanwhile, none of these works mentioned above considered the adaptive optimal control problem with state-dependent input constraint.

To demonstrate the application value of the method proposed in this article, we apply the proposed algorithm to solve the optimal UAV circumnavigation control problem, which is another main contribution of this article. Although there have been a series of contributions on UAV, the surveillance and tracking of moving ground targets is still one of the most important applications of UAV [30]. To monitor a ground target circumnavigate around the ground target with a preset radius [31]. For robots with single-integrator dynamics, different control algorithms have been proposed to achieve the circumnavigation with distance measurements [32] or bearing measurements [33]. For a nonholonomic agent, Deghat et al. [34] proposed a circumnavigation control method while the position of the target is assumed to be unknown. Assuming the relative position of the target is accessible, Dong et al. [20] designed a guidance law by exploiting the vector fields (VFs) method. However, none of these works have considered the optimality issue of the circumnavigation control.

B. Statement of Contributions

Aimed at solving the optimal control problem for nonlinear systems with state-dependent and unsymmetrical input constraints, this article investigates the partially unknown system whose dynamics is in NF form and an online PI algorithm is presented in this article. The main contributions of this article are summarized as follows.

1) An online PI algorithm is proposed to address the optimal control of nonlinear systems with state-dependent and unsymmetrical input constraints. The stability and convergence of the proposed algorithm are also proved. Compared with the existing studies, our method has the following novelities.

   a) Compared with the existing studies, such as [11], [21], [22], [23], [24], and [28], this article addresses the optimal control problem with the constraint set of the input $u$ being state-dependent rather than being constant.

   b) Compared with [26], [28], and [29], the closed-loop system with the proposed algorithm is theoretically guaranteed to be asymptotically stable rather than being UUB. Moreover, the method proposed in this article does not exist the difficulty of designing switching functions in comparison with [21].

2) The optimal control problem for an NF system is addressed. In the current study of the adaptive optimal control, the internal dynamics are usually required to hold zero when the state of the system is in the origin. To expand the application range of the ADP theory to systems that do not meet this condition, this article designs a special control law which consists of the initial stabilizing control law and the neural-network-based control law.

3) By exploiting the method proposed in this article, an optimal circumnavigation control law with input saturation is proposed. To the best of our knowledge, it is the first time that the UAV optimal circumnavigation problem w.r.t. an infinite-horizon performance index is addressed.

The remainder of this article is organized as follows. Section II develops our adaptive optimal control algorithm for the nonlinear systems with state-dependent and unsymmetrical input constraints. In Section III, we apply our methods to solve the optimal UAV circumnavigation control problem and a comparison with the method proposed in [20] is illustrated. Finally, the concluding remarks are drawn in Section IV.

Notation: The vector $1_{n} \in \mathbb{R}^{n}$ denotes a vector with its elements all being 1 and $0_{n \times m} \in \mathbb{R}^{n \times m}$ is a zero matrix. Here, we define an operator $\text{vec}(\cdot) : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n^2}$. For a vector $a \in \mathbb{R}^{n}$ and a diagonal matrix $M \in \mathbb{R}^{n \times n}$, if $a = \text{vec}(M)$, one has $a_{i} = M_{ii}, i = 1, \ldots, n$, where $a_{i}$ is the $i$th element of the vector $a$ and $M_{ii}$ is the $i$th element on the diagonal of the matrix $M$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
II. OPTIMAL CONTROL PROBLEM FOR NF SYSTEMS WITH UNSYMMETRICAL INPUT CONSTRAINTS

A. Problem Formulation

Consider the following system whose dynamics is:

\[ \dot{x}_1(t) = f_1(x_1, x_2) + g_1(x_1, x_2)u(t) \]  
\[ \dot{x}_2(t) = f_2(x_2) \]  

(1)  

(2)

where \( x_1 \in \mathbb{R}^{n_1} \) is the system state to be stabilized; \( x_2 \in \mathbb{R}^{n_2} \) is the state which is not intended to be controlled and assumed to be bounded; the continuous functions \( f_1(x_1, x_2) \in \mathbb{R}^{n_1} \) and \( f_2(x_2) \in \mathbb{R}^{n_2} \) are the unknown internal dynamics of the system; \( g_1(x_1, x_2) \in \mathbb{R}^{n_1 \times m} \) is the input dynamics of the system; \( u(t) \in \mathbb{R}^m \) is the control input. Note that the function \( f_1(x_1, x_2) \) is not necessary to be zero when \( x_1 = 0 \).

Denote the stack vector \( x \in \mathbb{R}^n \) as \( x = [x_1, x_2]^\top \), where \( n = n_1 + n_2 \). The control input \( u(t) \) is constrained by the condition

\[ d_i(x(t)) \leq u_i(t) \leq \hat{h}_i(x(t)), \quad i = 1, \ldots, m \]  

(3)

where \( u_i(t) \) is the \( i \)th element of \( u(t) \); \( d_i(x(t)) \) and \( \hat{h}_i(x(t)) \) are the known functions that determine the lower bound and the upper bound for the \( i \)th element of \( u(t) \).

Remark 1: The system described by (1) and (2) can be regarded as the general form of many widely studied systems. For example, the state \( x_2 \) can be regarded as a bounded time-varying uncertainty [35] and system (1) is to be stabilized while disturbed by \( x_2 \). Also, the system described by (1) and (2) is commonly used in the target tracking system with \( x_1 \) being the tracking error and \( x_2 \) being the state of the target. There have been many works reported on such systems in the previous literatures, such as [11] and [36].

Remark 2: In this article, the constraint set of \( u \) is unsymmetrical and state dependent as (3) reveals. Meanwhile, the internal dynamics \( f_1(x_1, x_2) \) of system (1) does not satisfy the condition that \( f(x_1, x_2) = 0 \) when \( x_1 = 0 \), which is required in the previous studies of ADP [11], [12], [13], [14], [15], [16], [23]. As a consequence, the methods proposed in the previous studies, such as [15] and [23], are inappropriate.

The aim of this article is to design an optimal policy \( u^*(t) = u^*(x(t)) \) constrained by the unsymmetrical set (3) such that the \( x_1 \)-system is stabilized as well as a performance index \( J(x(0), u) \) defined in the following form is minimized:

\[ J(x(0), u) = \int_0^\infty [Q(x(t)) + U_n(u, x)]dt, \quad x(0) = x^0 \]  

(4)

where \( Q(x_1) \) is a positive semi-definite function related with the state \( x_1 \), \( U_n(u, x) \) is a positive semi-definite function which is to be designed, and \( x^0 \) is the initial state.

Before presenting the solution to the optimal control problem described above, we first introduce a definition of the admissible control.

Definition 1 (Admissible Control) [6]: For a given system described by (1) and (2), a control policy \( u(t) = \mu(x) \) is defined to be admissible with respect to the performance index (4) on a compact set \( \Omega \subseteq \mathbb{R}^n \), written as \( \mu(x) \in \mathcal{A}(\Omega) \), if \( \mu(x) \) is continuous, \( u(t) = \mu(x) \) stabilizes system (1) and \( J(x^0, u) \) is finite for every \( x^0 \in \Omega \).

Similar to the previous works, such as [11] and [14], the following assumption is set in this article.

Assumption 1: There exists a known admissible control policy on a set \( \Omega \subseteq \mathbb{R}^n \) which stabilizes system (1) and satisfies constraint (3).

Remark 3: In many scenarios of nonlinear system controls with input saturation, such as trajectory-tracking control [37] and formation control [38], some Lyapunov-based methods have been proposed to design a stabilizing but nonoptimal control policy with input saturation, and thus an initial admissible control policy can be obtained. In some industrial applications, the initial admissible controller also can be constructed by empirical methods, such as tuning the control parameters of a PID controller.

B. Design Method for NF Systems With Unsymmetrical and State-Dependent Input Constraints

If Assumption 1 holds, the initial admissible control policy is denoted as \( u_i(t) = \mu_i(x(t)) \). Although the initial control policy \( u_i(t) \) is stabilizing, the control performance of \( u_i(t) \) may not be satisfactory. Thus, based on the initial control policy \( u_i(t) \), we design the control policy \( u(t) \) as follows:

\[ u(t) = u_i(t) + \hat{u}(t) \]  

(5)

where \( \hat{u}(t) \) is a to-be-designed virtual input which enables the actual control input \( u(t) \) to achieve the optimal control performance. As \( u(t) \) is constrained by (3), the virtual input \( \hat{u}(t) \) should satisfy

\[ d_i(x(t)) - u_i^0(t) \leq \hat{u}_i(t) \leq \hat{h}_i(x(t)) - u_i^0(t), \quad i = 1, \ldots, m \]  

(6)

where \( \hat{u}_i(t) \) and \( u_i^0(t) \) are the \( i \)th element of \( \hat{u}(t) \) and \( u_i(t) \), respectively. Further by employing (5), system (1) is rewritten as follows:

\[ \dot{x}(t) = F_1(x_1, x_2) + g_1(x_1, x_2)\hat{u}(t) \]  

(7)

where the function \( F_1(x_1, x_2) \) is defined by

\[ F_1(x_1, x_2) = f_1(x_1, x_2) + g_1(x_1, x_2)\mu_i(x). \]

Define two sets of functions \( \tilde{\lambda}_i(u(x)) \) and \( \hat{\lambda}_i(\hat{u}, x) \), \( i = 1, 2, \ldots, m \), as follows:

\[ \tilde{\lambda}_i(u(x)) = \begin{cases} \lambda_i(x) - \mu_i^0(x), & \text{if } u_i - \mu_i^0(x) \geq 0 \\ -d_i(x) + \mu_i^0(x), & \text{if } u_i - \mu_i^0(x) < 0 \end{cases} \]

\[ \hat{\lambda}_i(\hat{u}, x) = \begin{cases} \lambda_i(x) - \mu_i^0(x), & \text{if } \hat{u} \geq 0 \\ -d_i(x) + \mu_i^0(x), & \text{if } \hat{u} < 0. \end{cases} \]

(8)

Note that if Assumption 1 is satisfied, it holds that

\[ d_i(x) < \mu_i^0(x) < \hat{h}_i(x), \quad i = 1, \ldots, m \]

which yields

\[ d_i(x) - \mu_i^0(x) \leq 0 < \hat{h}_i(x) - \mu_i^0(x), \quad i = 1, \ldots, m \]  

(9)

where \( \mu_i^0(x) \) is the \( i \)th element of the function \( \mu_i(x) \). The inequality (9) guarantees that \( \tilde{\lambda}_i(u(x)) = \hat{\lambda}_i(\hat{u}, x) \geq 0 \) if \( \hat{u} = u_i - \mu_i^0(x) \).
Inspired by the previous studies such as [11], we design the control cost function \( U_n(u, x) \) for system (1) as follows:

\[
U_n(u, x) = 2 \sum_{i=1}^{m} \int_{0}^{\mu_i(x)} \bar{\lambda}_i r_i \left( \tanh^{-1} \left( \frac{s}{\sqrt{\lambda_i}} \right) \right) ds \tag{10}
\]

where \( r_i \) is a positive constant and \( \bar{\lambda}_i \triangleq \bar{\lambda}_i(u, x) \).

Remark 4: Note that compared with the design of the previous literatures, such as [11, 15], and [23] the scaling factor \( \bar{\lambda}_i \) is a function related with the system state rather than a constant in order to tackle the unsymmetrical input constraint and the NF form of the system. Meanwhile, the upper bound of the integral in (10) takes the initial admissible control policy \( \mu_i(x) \) into consideration, which is also different from the design of the previous studies.

The issue that remains is to design a virtual input \( \hat{u}(t) \) such that the control policy \( u(t) \) given by (5) is optimal w.r.t. the performance index (4). To achieve this target, we define a performance index \( J_2(x(0), \hat{u}) \) for system (7) as follows:

\[
J_2(x(0), \hat{u}) = \int_{0}^{\infty} \left[ Q(x(t)) + \bar{U}_n(\hat{u}, x) \right] dt, \quad x(0) = x^0 \tag{11}
\]

where the function \( \bar{U}_n(\hat{u}, x) \) is defined by

\[
\bar{U}_n(\hat{u}, x) = 2 \sum_{i=1}^{m} \int_{0}^{\hat{\mu}_i} r_i \hat{\lambda}_i \left( \tanh^{-1} \left( \frac{s}{\sqrt{\hat{\lambda}_i}} \right) \right) ds
\]

\[
= 2 \sum_{i=1}^{m} \hat{\lambda}_i \tanh^{-1} \left( \frac{\hat{\mu}_i}{\sqrt{\hat{\lambda}_i}} \right) r_i \hat{\mu}_i
\]

\[
+ \sum_{i=1}^{m} \hat{\lambda}_i^2 r_i \ln \left( 1 - \left( \frac{\hat{\mu}_i}{\sqrt{\hat{\lambda}_i}} \right)^2 \right) \tag{12}
\]

The function \( F_s(x_1, x_2) \) and the virtual input \( \hat{u}(t) \) can be regarded as the internal dynamics and control input of system (7), respectively. As the initial control policy \( u_0(t) \) is an admissible control of system (1), it holds that \( F_s(0, x_2) = 0 \). Otherwise, when \( u(t) = u_0(t) \) and \( x_1 = 0 \), it holds that \( \dot{x}_1 = F_s(0, x_2) \neq 0 \), which contradicts with the fact that system (1) can be stabilized by the control policy \( u_0(t) \). Moreover, since \( F_s(0, x_2) = 0 \), the virtual input \( \hat{u} \) should hold zero when \( x_1 = 0 \). As a consequence, by using (5), system (1) is transformed into system (7), which meets the requirement for the admissible control defined in Definition 1.

Remark 5: In the previous works, such as [11] and [14], the initial admissible control policy \( u_0(t) \) is only used in the initial PI. However, in this article, it can be observed from (5) that the initial admissible control policy \( u_0(t) \) also acts as a baseline for constructing the optimal control law. Then, the to-be-designed virtual input \( \hat{u}(t) \) is added on the initial control policy \( u_0(t) \) to obtain an optimized control performance. By this method, the internal dynamics \( f_s(x_1, x_2) \) of system (1) is not necessary to be zero when \( x_1 = 0 \).

The following lemma shows that the optimal control problem for system (1) can be transformed into the optimal control problem for system (7).

**Lemma 1:** Given a unique optimal control policy \( \hat{u}^*(t) \) for system (7) w.r.t. the performance index (11), the control policy \( u^*(t) \) given by

\[
u^*(t) = \hat{u}^*(t) + u_0(t) \tag{13}
\]

is the unique optimal control policy for system (1) w.r.t. the performance index (4).

**Proof:** Given two control policies \( u(t) \) and \( \hat{u}(t) \) satisfying (5) with the same initial states, the trajectories of system (1) and system (7) are identical. Meanwhile, from (8), it can be observed that \( \bar{\lambda}_i(u, x) = \hat{\lambda}_i(u, x) \). Further, from (10) and (12), it holds that \( U_n(u, x) = \bar{U}_n(\hat{u}, x) \) if \( u = \mu_i(x) + \hat{u} \), which further implies \( J(x(0), u) = J_2(x(0), \hat{u}) \). Suppose the control policy \( u^*(t) \) given by (13) is not the unique optimal control policy, there always exists a control policy \( \hat{u}'(t) \) such that

\[
J(x(0), u') \leq J(x(0), u^*) \tag{14}
\]

Define a control policy \( \hat{u}'(t) = u'(t) - u_0(t) \). Through the discussion above, one has

\[
J(x(0), u') = J_2(x(0), \hat{u}'). \tag{15}
\]

By using (14) and (15), it yields

\[
J_2(x(0), \hat{u}) = J(x(0), u') \leq J(x(0), u^*) = J_2(x(0), \hat{u}')
\]

which contradicts with the fact that \( \hat{u}^*(t) \) is the optimal control policy. Thus, \( u^*(t) \) given by (13) is the unique optimal control policy for system (1) w.r.t. the performance index (4).

**Lemma 1** shows that the optimal control policy \( u^*(t) \) for system (1) w.r.t. the performance index (4) can be obtained if the optimal control policy \( \hat{u}^*(t) \) is available. Thus, we introduce the procedure of designing \( \hat{u}^*(t) \) in the following part.

Using (2) and (7), the dynamics of the state \( x \) can be rewritten as follows:

\[
\dot{x} = F(x) + G(x)\hat{u} \tag{16}
\]

where

\[
F(x) = \left[ F_s(x_1, x_2) \right]^{\top}, \quad G(x) = \left[ g_1(x_1, x_2), 0^{\top} \right] r_{x \times m}^{\top}
\]

Assume there exists a continuously differentiable value function \( V^*(x) \) defined by

\[
V^*(x(t)) = \min_{\hat{u} \in \mathcal{A}(t)} \int_{t}^{\infty} \left[ Q(x(s)) + \bar{U}_n(\hat{u}, x) \right] ds. \tag{17}
\]

A Hamiltonian function \( H(x, V^*, \hat{u}) \) is defined as follows:

\[
H(x, V^*, \hat{u}) = (V^*_s)^{\top} \left( F + G\hat{u} \right) + Q(x_1) + \bar{U}_n(\hat{u}, x) \tag{18}
\]

where \( F \triangleq F(x), \quad G \triangleq G(x), \quad V^* \triangleq V^*(x), \) and \( V^*_s = \partial V^*(x)/\partial x \in \mathbb{R}^n \). Then, similar to the previous works [11], [15], [23], by using the stationarity condition (see [4]) on the Hamiltonian function \( H(x, V^*, \hat{u}) \) i.e., \( \partial H/\partial \hat{u} = 0 \), we obtain the optimal control policy \( \hat{u}^*(t) \) as follows:

\[
\hat{u}^*(t) = \hat{u}^*(x(t)) = \arg \min_{\hat{u} \in \mathcal{A}(t)} \hat{H}(x, V^*, \hat{u})
\]

\[
= -\lambda \tanh \left( (1/2) \left( \lambda \hat{R} \right)^{-1} G^{\top} V^*_s \right) \tag{19}
\]
where \( \hat{\lambda}, R \in \mathbb{R}^{m \times m} \) are diagonal matrices whose \( \hat{\lambda}_i \) on the diagonal are \( \hat{\lambda}_i \) and \( r_i \), respectively. It can be observed from (19) that \( |\hat{u}_i| \leq \hat{\lambda}_i(\hat{u}_i, x) \), based on which the inequality (6) can be derived. Further, the control policy \( u(t) \) defined by (5) can satisfy the constraint (3).

Next, substituting (19) into (12) results in

\[
\dot{U}_n(\hat{u}^*, x) = (V_n^*)^T \hat{\lambda} \text{tanh}(\hat{D}^*) + \left( \text{vecc}(\hat{\lambda}R\hat{\lambda}) \right)^T \ln \left( 1 - \tanh^2(\hat{D}^*) \right) + Q(x_1) + (V_n^*)^T F(x) = 0.
\]

If the solution \( V_n^* \) of the HJB equation (21) is found, the optimal input \( \hat{u}^* \) for the system (7) can be obtained by (19), and the optimal control policy \( u^*(t) \) for system (1) is further obtained according to (13).

In the following theorem, it is proved that the control law \( u^*(t) \) defined by (13) and (19) is the unique optimal control policy w.r.t. the performance index (4) and stabilizes the \( x_1 \)-system.

**Theorem 1:** Consider the optimal control problem for system (1) w.r.t. the performance index (4). Suppose that \( V^*(x) \) is a positive-definite solution to the HJB equation (21). Then, the control policy \( u^*(t) \) defined by (13) and (19) is the unique optimal control policy such that system (1) is stabilized asymptotically and the performance index (4) is minimized.

**Proof:** First, we will prove that \( u^*(t) \) is the unique optimal control policy that minimizes the performance index (4). It has been proved in Lemma 1 that the optimality of the control policy \( u^*(t) \) for system (1) w.r.t. the performance index (4) is equivalent to the optimality of the control policy \( \hat{u}^*(t) \) w.r.t. the performance index (11). Thus, one only needs to prove that \( \hat{u}^*(t) \) is the unique optimal control policy for system (7) w.r.t. the performance index (11).

Note that for the optimal value function \( V^*(x(t)) \) defined by (17), one has

\[
\int_0^\infty (V^*(x(\tau))) d\tau = -V^*(x(0)).
\]

Then given an admissible control law \( \hat{u}(t) \), the performance index (11) can be rewritten as follows:

\[
\mathcal{J}_2(\hat{u}(t), x) = \int_0^\infty \left[ Q(x_1) + \hat{U}_n(\hat{u}, x) \right] d\tau + V^*(x(0))
\]

\[
= \int_0^\infty \left[ Q(x_1) + \hat{U}_n(\hat{u}, x) \right] d\tau + V^*(x(0))
\]

By adding and subtracting the term \( \hat{U}_n(\hat{u}^*, x) \) in the integral part, (22) becomes

\[
\frac{dV^*(x)}{dt} = (V_n^*)^T \left( F + \hat{G}u^* \right) G(\hat{u}^* - \hat{u}^*) \right) d\tau.
\]

The equality in (28) holds if and only if \( \| x_1 \| = 0 \). As a consequence, the \( x_1 \)-system is asymptotically stable.
C. Online Policy Iteration Algorithm for Solving the HJB Equation

To derive \( \tilde{u}'(t) \) based on (19), the solution \( V^*(x) \) of the HJB equation (21) has to be solved. However, since (21) is usually highly nonlinear, it is quite difficult to get its analytical solution. In the following part, to obtain an equivalent formulation of HJB which does not need the knowledge of the internal dynamics \( f_1(x_1, x_2) \) and \( f_2(x_2) \), the IRL idea introduced in [16] is employed.

Let \( T > 0 \) denote the integral reinforcement interval and it holds that

\[
V(x(t)) = \int_{t}^{t+T} \left[ Q(x_1) + \hat{U}_n(\tilde{u}, x) \right] d\tau + V(x(t + T)).
\]

Based on (29), the following IRL-based PI algorithm is utilized to get the solution of the HJB equation (21) with the internal dynamics \( f_1(x_1, x_2) \) and \( f_2(x_2) \) unknown.

1) Policy Evaluation: Given an admissible control policy \( \tilde{\mu}^{(k)}(x) \), update \( V^{(k)}(x) \) by the Bellman equation

\[
V^{(k)}(x(t)) = \int_{t}^{t+T} \left[ Q(x_1) + \hat{U}_n(\tilde{u}, x) \right] d\tau + V^{(k)}(x(t + T)).
\]

2) Policy Improvement: Update the control policy according to

\[
\tilde{\mu}^{(k+1)}(x) = -\hat{\lambda} \tanh \left( (1/2) \left( 1_{\lambda} \right)^{-1} G(\tau) V^{(k)}(x) \right)
\]

where \( V^{(k)}(x) \triangleq \partial V^{(k)}(x)/\partial x \); the notations \( V^{(k)}(x) \) and \( \tilde{\mu}^{(k+1)}(x) \) represent the value function and the virtual control policy in the kth iteration, respectively.

The following theorem shows that the IRL method introduced above can be employed to improve the control law.

Theorem 2: Let \( \tilde{\mu}^{(k)} = \tilde{\mu}^{(k)}(x) \in A(\Omega) \) and \( V^{(k)}(x) \) satisfy \( H(x, V^{(k)}), \tilde{u}^{(k)} = 0 \) with the boundary condition \( V^{(k)}(0) = 0 \). Then, the control policy \( \tilde{\mu}^{(k+1)}(x) \) defined by (31) is an admissible control for system (7). Moreover, if \( V^{(k+1)}(x) \) is the positive semi-definite function that satisfies\( H(x, V^{(k+1)}), \tilde{u}^{(k+1)} = 0 \) with \( V^{(k+1)}(0) = 0 \), it holds that \( V^{*}(x) \leq V^{(k+1)}(x) \leq V^{(k)}(x) \).

Proof: We first prove that \( \tilde{u}^{(k+1)}(x) \in A(\Omega) \). Taking the derivative of \( V^{(k)}(x) \) along the trajectory of system \( \dot{x} = F(x) + G(x)\tilde{u}^{(k+1)} \), it yields

\[
\dot{V}^{(k)}(x) = \left( V^{(k)}(x) \right)^{\top} F + \left( V^{(k)}(x) \right)^{\top} \tilde{G}(\tilde{u}^{(k+1)}).
\]

Since \( \dot{H}(x, V^{(k)}, \tilde{u}^{(k)}) = 0 \), we get

\[
\left( V^{(k)} \right)^{\top} F = -\left( V^{(k)} \right)^{\top} \tilde{G}(\tilde{u}^{(k)}) - Q(x_1) - \hat{U}_n(\tilde{u}^{(k)}, x).
\]

By substituting the term \( (V^{(k)})^{\top} F \) with (33), (32) becomes

\[
\dot{V}^{(k)}(x) = -Q(x_1) - \hat{U}_n(\tilde{u}^{(k+1)}, x) - M_t(x),
\]

where \( M_t(x) \) is

\[
M_t(x) = \left( V^{(k)} \right)^{\top} G(\tilde{u}^{(k)} - \tilde{u}^{(k+1)}) + \hat{U}_n(\tilde{u}^{(k)}).
\]

It can be deduced from (31) that

\[
\left( V^{(k)} \right)^{\top} G = -2\hat{\lambda} R \tanh^{-1} \left( \left( \hat{\lambda} \right)^{-1} \tilde{u}^{(k+1)} \right).
\]

Combined with (35) and

\[
\hat{U}_n(\tilde{u}^{(k)} - \tilde{u}^{(k+1)} + x) = 2 \sum_{i=1}^{m} r_i \hat{x}_i(t) \left( \tanh^{-1} \left( s/\hat{x}_i \right) \right) ds
\]

the term \( M_t(x) \) can be rewritten as follows:

\[
M_t(x) = 2 \sum_{i=1}^{m} r_i \hat{x}_i m_i
\]

where \( m_i \) is

\[
m_i = \int \hat{x}_i(t) \left( \tanh^{-1} \left( s/\hat{x}_i \right) \right) ds
\]

with further implies \( M_t(x) \geq 0 \).

Since \( V^{(k)}(x) \geq 0 \) and \( V^{(k)}(x) = 0 \) if and only if \( ||x|| = 0 \), the function \( V^{(k)}(x) \) can be treated as a Lyapunov function for \( x_t \). Then from (34), we obtain that \( V^{(k+1)}(x) \leq 0 \) as the functions \( Q(x_1), \hat{U}_n(\tilde{u}^{(k+1)}) \) and \( M_t(x) \) are all positive semi-definite. Hence, the \( x_t \)-system can be stabilized by the control policy \( \tilde{u}^{(k+1)} \). Besides, from (31), it can be observed that \( \dot{u}^{(k+1)} = 0 \) if \( x_1 = 0 \). Thus, \( \tilde{u}^{(k+1)} \) is admissible.

Next, we will prove that \( V^{*}(x) \leq V^{(k+1)}(x) \leq V^{(k)}(x) \). As both \( \tilde{u}^{(k)} \) and \( \tilde{u}^{(k+1)} \) are admissible control policies, we have

\[
V^{(k+1)}(x(t = \infty)) = 0 \quad \text{and} \quad V^{(k+1)}(x(t = \infty)) = 0
\]

Taking the derivative of \( V^{(k+1)}(x) \) and \( V^{(k+1)}(x) \), respectively, along the trajectory of system \( \dot{x} = F(x) + G(x)\tilde{u}^{(k+1)} \), it yields

\[
\dot{V}^{(k+1)}(x(t)) = V^{(k+1)}(x(t))
\]

\[
\dot{V}^{(k+1)}(x(t)) = -\left( V^{(k+1)}(x(t)) \right)^{\top} \tilde{G}(\tilde{u}^{(k+1)}) - Q(x_1) - \hat{U}_n(\tilde{u}^{(k+1)}).
\]

By substituting (33) and (37) into (36), we derive that

\[
V^{(k+1)}(x(t)) - V^{(k)}(x(t))
\]

\[
\dot{V}^{(k+1)}(x(t)) = -\left( V^{(k+1)}(x(t)) \right)^{\top} \tilde{G}(\tilde{u}^{(k+1)}) - Q(x_1) - \hat{U}_n(\tilde{u}^{(k+1)}).
\]

Thus \( V^{(k+1)}(x) \leq V^{(k)}(x) \) for \( \forall x \in \Omega \). Furthermore, by using the contradiction method, it holds that \( V^{*}(x) \leq V^{(k+1)}(x) \leq V^{(k)}(x) \).

Theorem 2 guarantees that the trained control policy is always admissible during the process of PI. Meanwhile, the
updated control policy is always better than its previous one. Then to successively solve (30) and (31), the value function $V(x)$ is approximated by a single-layer NN, which is

$$V(x) = \sum_{j=1}^{M} w_j^{(k)} \sigma_j(x) + \xi(x) = (w^*)^T \sigma_M(x) + \xi(x) \quad (38)$$

where $\sigma_j(x)$ is the activation function and satisfies $\sigma_j(x_1) = 0$; $\xi(x)$ is the approximation residual error; $w_j^*$ represents the ideal weight of the $j$th neuron which minimizes the residual error $\xi(x)$; the vector $\sigma_M(x)$ denotes the vector of activation functions; and $w^*$ denotes the ideal weight vector.

Remark 6: It has been pointed out in [6] that the approximation residual error $\xi(x)$ will converge to zero when the number of neurons $M \to \infty$. Meanwhile, for fixed $M$, the approximation residual error $\xi(x)$ is also bounded [39]. In practical implementation, the approximation residual error is usually reduced by setting the number of neurons as large as possible. But how to eliminate the approximation residual error $\xi(x)$ completely with a limited number of neurons still requires further investigation.

In order to seek the ideal weight vector $w^*$, the value function $V^{(k)}(x)$ in the $k$th iteration is approximated as follows:

$$V^{(k)}_M(x) = \sum_{j=1}^{M} w_j^{(k)} \sigma_j(x) = (w^{(k)}_M)^T \sigma_M(x) \quad (39)$$

where $w_j^{(k)}$ and $w^{(k)}_M$ denote the weight of the $j$th neuron and the weight vector in the $k$th iteration, respectively.

By replacing $V^{(k)}(x)$ in (30) with $V^{(k)}_M(x)$, we have

$$\left(w^{(k)}_M\right)^T \sigma_M(x(t)) = \int_{t}^{t+T} \left[ Q(x_1) + \hat{U}_n(\hat{u}, x) \right] dt + \left(w^{(k)}_M\right)^T \sigma_M(x(t + T)) - e^{(k)}_M \quad (40)$$

where $e^{(k)}_M$ is the residual error defined by

$$e^{(k)}_M = \int_{t}^{t+T} \left[ Q(x_1) + \hat{U}_n(\hat{u}, x) \right] dt + \left(w^{(k)}_M\right)^T \sigma_M(x(t + T)) - \sigma_M(x(t)). \quad (41)$$

Obviously, the parameter $w_M$ should be tuned to reduce the residual error. Here, we define a to-be-minimized index as follows:

$$S = \int_{\Omega} |e^{(k)}_M(x)|^2 dx.$$

To minimize $S$, the weights $w_M$ is determined by

$$\left( \frac{de^{(k)}_M(x)}{dw_M}, e^{(k)}_M(x) \right) = 0 \quad (42)$$

where the notation $(f(x), g(x)) = \int_{\Omega} f(x)g(x) dx$ denotes the Lebesgue integral. Let $g(x, t) = \sigma_M(x(t + T)) - \sigma_M(x(t))$. By substituting (41) into (42), one has

$$\left( g(x, t), \int_{t}^{t+T} \left[ Q(x_1) + \hat{U}_n(\hat{u}, x) \right] dt \right)_{\Omega} w^{(k)}_M = 0. \quad (43)$$

To solve $w^{(k)}_M$, we impose the following assumption in the spirit of persistent excitation (PE) condition.

Assumption 2: For all admissible control policies $\mu(x) \in \mathcal{A}(\Omega)$, there exist constants $m_c > 0$ and $y_c > 0$ such that

$$\frac{1}{m_c} \sum_{i=1}^{m_c} g(x, t_i) g(x, t_i)^T \geq y_c I_M \quad (44)$$

for all $m_c \geq m_c$.

Let $p$ represent the number of points in the sample set $\Omega$. If Assumption 2 holds and $p \geq m_c$, it can be inferred that $\langle g(x, t), g(x, t)^T \rangle_{\Omega}$ is invertible. Thus, based on (43), $w_M^{(k)}$ is updated as follows:

$$w_M^{(k)} = -\left( g(x, t), g(x, t)^T \right)^{-1}_{\Omega} \int_{\Omega} \left[ Q(x_1) + \hat{U}_n(\hat{u}, x) \right] dt. \quad (45)$$

Afterwards, in order to solve (45), an iterative algorithm proposed by [6] is adopted. Given some points over the integration region on $\Omega$, define

$$L = \left[ g(x, t) |_{x_1}, \ldots, g(x, t) |_{x_p} \right],$$

$$Y = \left[ \int_{t}^{t+T} \left( Q(x_1) + \hat{U}_n(\hat{u}, x) \right) dt |_{x_1}, \ldots, \int_{t}^{t+T} \left( Q(x_1) + \hat{U}_n(\hat{u}, x) \right) dt |_{x_p} \right].$$

Then, we get

$$\left( g(x, t), g(x, t)^T \right)_{\Omega} = \lim_{||\delta x|| \to 0} \left( L^T L \right)^{\delta x},$$

$$\left( g(x, t), \int_{t}^{t+T} \left[ Q(x_1) + \hat{U}_n(\hat{u}, x) \right] dt \right)_{\Omega} = \lim_{||\delta x|| \to 0} \left( L^T Y \right)^{\delta x}. \quad (46)$$

By using (46), we rewrite (45) as follows:

$$w_M^{(k)} = -\left( L^T L \right)^{-1} \left( L^T Y \right). \quad (47)$$

Remark 7: Assumption 2 is a common assumption in the current study of ADP [7], [11], [40]. Theoretically, the validity of Assumption 2 is related to the integration time $T$ and the richness of the collected samples. It has been pointed out in [14] that if a proper integral time $T$ is selected, the sample size $p$ just needs to be no smaller than $M$ such that the matrix $\langle g(x, t), g(x, t)^T \rangle_{\Omega}$ is invertible. However, it still remains an unsolved problem on how to select such a proper integral time $T$. As a consequence, in practical implementation, to enrich the diversity of samples so that Assumption 2 holds, the control input in the training phase is expected to be persistently exciting, which is usually guaranteed by adding a small exploration noise to the original control input [14]. After the training is finished, the exploration noise is removed. Meanwhile, the size of the sample set $\Omega$ is usually chosen as large as possible to guarantee $p \geq m_c$.

According to the definition of $\hat{\lambda}_i$ in (8), the sign of each element $\hat{\mu}_i(x)$ of $\hat{\mu}(x)$ should be evaluated in advance. From the structure of $\hat{\mu}_i(x)$ in (31), it can be found that the sign of $\hat{u}_i(t)$ is the same as that of the $i$th element of $-G^T(x) V^{(k)}_i(x)$. Thus, the matrix $\hat{\lambda}$ can be determined by calculating $-G^T(x) V^{(k)}_i(x)$.
in advance. Specifically, the $i$th element of $\hat{\lambda}$ on the diagonal is determined by

$$
\hat{\lambda}_i = \begin{cases} 
\hat{h}_i(x) - \mu_\lambda(x), & \text{if } z_i \geq 0 \\
-\hat{d}_i(x) + \mu_\lambda(x), & \text{if } z_i < 0 
\end{cases}
$$

where $z_i$ is the $i$th element of $-G^T(x)v_k(x,y)$.

The iterations are terminated when the error of the coefficients obtained at two consecutive steps is smaller than a given threshold $\epsilon$. The flow chart of the proposed IRL algorithm is presented in Fig. 1.

Remark 8: As the traditional admissible control policies can stabilize the system, the initial control policy $u(t)$ defined by (5) with the NN weight being zero vector is also admissible. Thus in the PI algorithm described above, the NN weight can be initialized as a zero vector directly.

III. APPLICATION TO THE OPTIMAL UAV CIRCUMNAVIGATION PROBLEM

In this section, we employ the method proposed in this article to solve a practical problem: the optimal UAV circumnavigation control problem. As the target’s position is usually estimated from the on-board sensor measurement, which often contains noise, a filter like an Extended Kalman filter (EKF) is needed. It has been shown in [41] that the performance of the filter is dependent on the UAV’s trajectory. In this section, we intend to design an optimal circumnavigation controller based on the Fisher information, which can quantify the information provided by the sensor measurement [42]. Generally speaking, the more Fisher information the UAV gains, more accurate the estimated target position will be [43]. Specifically, the UAV is controlled to circumnavigate around the target (see Fig. 2), while minimizing an objective function involving Fisher information in the circumnavigation performance of the filter is dependent on the UAV’s trajectory.

$$
\min_{\tilde{u}(t)} J(\tilde{u}(t)) = \int_{t_0}^{t_f} f(x(t), \tilde{u}(t)) dt
$$

where $f(x(t), \tilde{u}(t))$ is the cost function.

In this section, we intend to design an optimal circumnavigation controller based on the Fisher information, which can quantify the information provided by the sensor measurement [42]. Generally speaking, the more Fisher information the UAV gains, more accurate the estimated target position will be [43]. Specifically, the UAV is controlled to circumnavigate around the target (see Fig. 2), while minimizing an objective function involving Fisher information in the circumnavigation performance of the filter is dependent on the UAV’s trajectory. A simulation result is presented in which the performance of the control law designed by our method is compared with the method proposed in [20].

A. Problem Formulation of the Optimal UAV Circumnavigation

Consider a fixed-wing UAV, whose kinematic model is described by

$$
\begin{align*}
\dot{x}_p &= v \cos \theta \\
\dot{y}_p &= v \sin \theta \\
\dot{\theta} &= u_\theta \\
v &= u_v
\end{align*}
$$

(48)

where $(x_p, y_p)$ is the position of the UAV and $\theta$ denotes the heading angle of the UAV; $v$ is the UAV’s linear velocity; and $u_v$ and $u_\theta$ are the control inputs. The height of the UAV is assumed to be held constant. Owing to the roll angle constraint of the fixed-wing UAV, the following unsymmetrical input constraint is enforced on the UAV:

$$
\begin{align*}
\omega_x &= \frac{\omega_{\text{min}}}{1 + 0.02v} \leq u_\theta \leq \frac{\omega_{\text{max}}}{1 + 0.02v}
\end{align*}
$$

(49)

where the constants $\omega_{\text{max}}, \omega_{\text{min}} > 0$ and $\omega_{\text{max}} \neq \omega_{\text{min}}$. It is obvious that the input saturation constraint is unsymmetrical and depends on the UAV’s linear velocity. Note that the linear velocity $v$ is not constant but is dependent on the UAV’s state, which will be illustrated later.

Let $s_t = (x_t, y_t)^T \in \mathbb{R}^2$ represent the position of the moving target. It is assumed that the ground target moves with a constant linear velocity $v_t$ and the dynamics of the target is described by

$$
\begin{align*}
\dot{x}_t &= v_t \cos \theta_t \\
\dot{y}_t &= v_t \sin \theta_t \\
\dot{\theta}_t &= h(\theta_t)
\end{align*}
$$

(50)

where $\theta_t$ is the heading of the target and $h(\theta_t)$ is an unknown function.

As the UAV is expected to hold a constant angular speed with a preset circumnavigation radius, the relative speed $v_r$ of the UAV is also expected to be constant. The relative angle of the UAV w.r.t. the target is denoted by $\theta_r$, which satisfies

$$
\begin{align*}
v_r \cos \theta_r &= v \cos \theta - v_t \cos \theta_t \\
v_r \sin \theta_r &= v \sin \theta - v_t \sin \theta_t
\end{align*}
$$

(51)
The linear speed of the UAV is dependent on the UAV’s state and can be obtained from (50) as follows:

\[ v = v_r \sin \theta + v_r \cos(\theta - \theta_i) + v_t \sin \theta_i. \] (51)

The UAV is assumed to utilize a radar as the measurement sensor. Let \( s_r = (x_r, y_r)^T = (x_o - x_r, y_o - y_r)^T \) denote the relative position between the UAV and the target. Then with the aid of the radar, the UAV can sense the range and bearing information with the observation model described by

\[ \zeta(t) = Z(s_r) + \chi(t) \]

where \( \chi(t) \) is the sensor measurement; \( \chi(t) \) is the measurement noise; and \( Z(s_r) \) is the observation function defined by

\[ Z(s_r) = \left[ \frac{r}{\varphi} \right] = \left[ \sqrt{x_r^2 + y_r^2 + h^2} \right] \]

where \( h \) is the height of the UAV; \( r \) is the distance between the UAV and the target in 3-D space; and \( \varphi \) is the bearing angle between the target and the UAV in the plane.

The circumnavigation radius error \( e_r \) is defined by

\[ e_r = r_h - r_d \]

and a state \( \eta \) is defined by

\[ \eta = \frac{\pi}{2} - (\theta_r - \varphi) \]

where \( r_d \) is the desired circumnavigation radius around the target and \( r_h = \sqrt{x_r^2 + y_r^2} \) represents the current circumnavigation radius. Note that the desired circumnavigation around the target is achieved if the conditions \( e_r = 0 \) and \( \eta = 0 \) are satisfied [44]. Thus, to make the state \( e_r \) and \( \eta \) converge to zero, the dynamics of \( e_r \) and \( \eta \) will be analyzed first. The dynamics of \( e_r \) is

\[ \dot{e}_r = \frac{x_r \dot{x}_r + y_r \dot{y}_r}{r_h} = v_r \cos \varphi \cos(\theta_r - \theta_i) + v_t \sin \varphi \sin \theta_i = v_r \cos(\varphi - \theta) = v_r \sin \eta. \]

Deriving both sides of the (50) by time \( t \), one has

\[ -v_r \sin \theta \dot{\theta}_r = \frac{v_r \cos(\varphi - \theta) - v_t \sin \varphi \sin \theta_i}{r_h} = \frac{v_r \cos \varphi \cos(\theta_r - \theta)}{r_h} \]

Further, the dynamics of \( \varphi \) can be obtained by

\[ \dot{\varphi} = \frac{x_r \dot{y}_r - y_r \dot{x}_r}{r_h} = \frac{v_r \sin(\theta_r - \varphi)}{r_h} = \frac{v_r \cos \eta}{r_d + e_r}. \]

By combining (56) and (57), it yields

\[ \dot{\eta} = \dot{\phi} - \dot{\theta}_r = \frac{v_r \cos \eta}{r_d + e_r} + v_r \cos(\theta_r - \theta) \dot{\theta}_r - \frac{v}{v_r \cos(\theta_r - \theta)} \dot{\theta}. \]

Define a state variable \( x = (e_r, \eta, \theta, \theta_i)^T \). Then, the \( x \)-dynamics is described by

\[ \begin{align*}
\dot{e}_r &= v_r \sin \eta \\
\dot{\eta} &= \frac{v_r \cos \eta}{r_d + e_r} + \frac{v_r \cos(\theta_r - \theta)}{v_r \cos(\theta_r - \theta)} \dot{\theta} - \frac{v}{v_r \cos(\theta_r - \theta)} u_0 \\
\dot{\theta} &= u_0 \\
\dot{\theta}_r &= h(\theta) \end{align*} \]

where the linear speed \( v \) of the UAV is given by (51).

It can be observed that system (58) is an NF system as \( \dot{h} \neq 0 \) when \( e_r, \eta, \) and \( u_0 \) are all 0. By using the method proposed in Section II-B, the control policy \( u_0(t) \) is designed as follows:

\[ u_0(t) = u_s(t) + \hat{u}(t) \]

where \( u_s(t) \) is the initial control policy and \( \hat{u}(t) \) is the virtual input to be designed. Specifically, we adopt the VF method proposed in [20] as the initial admissible control policy \( u_s(t) \), which is described by

\[ u_s = \begin{cases} -k(\theta - \theta_d), & \text{if } k(\theta_d - \theta) \in \left[ -\frac{\omega_{\min}}{1+0.02v}, \frac{\omega_{\max}}{1+0.02v} \right] \\ -\frac{\omega_{\min}}{1+0.02v}, & \text{if } k(\theta_d - \theta) > \frac{\omega_{\max}}{1+0.02v} \\ \frac{\omega_{\min}}{1+0.02v}, & \text{if } k(\theta_d - \theta) < \frac{\omega_{\min}}{1+0.02v} \end{cases} \]

where \( \theta_d \) is determined by

\[ \begin{align*}
\cos(\theta_d) &= -\frac{v_r \sin \eta}{r_h(\frac{r_h^2 + r_d^2}{r_h^2})} + \frac{v_r \sin(\theta_r - \theta)}{r_h}, \\
\sin(\theta_d) &= \frac{v_r \cos(\theta_r - \theta)}{r_h}, \\
\hat{\theta} &= u_s + \hat{u} \\
\hat{\theta}_r &= h(\theta). \end{align*} \]

By substituting (59) into (58), the \( x \)-dynamics becomes

\[ \begin{align*}
\dot{e}_r &= v_r \sin \eta \\
\dot{\eta} &= \frac{v_r \cos(\theta_r - \theta)}{v_r \cos(\theta_r - \theta)} \dot{\theta} - \Lambda u_s - \Lambda \hat{u} \\
\dot{\theta} &= u_s + \hat{u} \end{align*} \]

where \( \Lambda = v/(v_r \cos(\theta_r - \theta)) \) and the virtual input \( \hat{u} \) is constrained by

\[ u_s - \frac{\omega_{\min}}{1+0.02v} \leq \hat{u} \leq u_s + \frac{\omega_{\max}}{1+0.02v}. \]

Up to now, the system dynamics for the UAV circumnavigation problem has been formulated. Then a to-be-minimized optimization criterion is needed. First, using the method proposed in Section II-B, the control cost function \( U_s(\hat{u}, x) \) is defined as follows:

\[ \hat{U}_s(\hat{u}, x) = 2 \int_0^L \frac{\hat{\lambda}}{1+0.02v} \right| \right| \left( s/\hat{\lambda} \right) \right| ds \]

where the constant \( r > 0 \) and \( \hat{\lambda} \in \mathbb{R} \) is defined by

\[ \hat{\lambda} = \begin{cases} u_s(t) + \frac{\omega_{\max}}{1+0.02v}, & \text{if } \hat{u} \geq 0 \\ u_s(t) - \frac{\omega_{\min}}{1+0.02v}, & \text{if } \hat{u} < 0 \end{cases} \]

Next, a function representing the state cost will be constructed based on the so-called accumulative information [41]. To quantify the utilization of the sensor data, we set up the optimization criterion by exploiting the accumulative information \( D \) based on the Fisher information metric, which is

\[ D = \int_{t_0}^\infty \sqrt{I(r, \eta)} dt \]
where \( L(r_h, \eta) \)
\[
L(r_h, \eta) = \frac{v_r^2}{p^2 \sigma_r^2} \sin^2 \eta + \frac{8v_r^2}{r^4} \sin^2 \eta + \frac{v_r^2}{r_h^2 \sigma_r^2} \cos^2 \eta \quad (62)
\]

\( \sigma_r \) and \( \sigma_\eta \) are two constants representing the standard deviations of the rang and bearing measurements, respectively. The detailed derivation for (62) can be found in our previous work [44].

It can be deduced intuitively from (61) and (62) that the UAV will obtain more accumulative information in the unit time if the circumnavigation radius \( r_h \) decreases. In other words, the UAV will fly just above the ground target if one directly takes (61) as a to-be-maximized performance index. However, the UAV is expected to circumnavigate around the target with a given radius, which implies the performance index should reach its extremum at \( r_h = r_d \). To achieve this, similar to our previous work [44], a small variation is made based on the definition of the accumulative information and a to-be-minimized performance index is defined as follows:

\[
\mathcal{J} = \int_0^\infty \left[ \left( Q_{\text{max}} - \hat{Q}(r_h, \eta) \right) / Q_{\text{max}} + \hat{U}_h(\bar{u}, x) \right] d\tau \quad (63)
\]

\( \hat{Q}(r_h, \eta) = \sqrt{L(r_h, \eta)} \tanh(r_h - \kappa) \)

where \( \hat{Q}(r_h, \eta) \) is a function which varies slightly from the function \( \sqrt{L(r_h, \eta)} \); the constant \( Q_{\text{max}} = \hat{Q}(r_d, 0) \) equals the value of \( \hat{Q}(r_h, \eta) \) when \( r_h = r_d \) and \( \eta = 0; \kappa \) is a bias constant determined by

\[
d\hat{Q}(r_h, \eta) / dr_h(r_h = r_d, \eta = 0) = 0. \quad (64)
\]

Specifically, the exact form of (64) can be obtained by

\[
\alpha(\kappa) = -\tanh(r_d - \kappa) + r_d \left( 1 - \tanh^2(r_d - \kappa) \right) = 0.
\]

Note that the function \( \alpha(\kappa) \) is a monotonically decreasing function and \( \kappa < r_d \). Thus the constant \( \kappa \) can be calculated by using the numerical stepwise methods. Affected by the term \( \tanh(r_h - \kappa) \), the function \( \hat{Q}(r_h, \eta) \) reaches its maximum at \( r_h = r_d \) and \( \eta = 0 \). If the UAV is controlled by the optimal control policy \( u^*_h \) w.r.t. the performance index (63), the desired circumnavigation will be achieved while the accumulative information is maximized [44].

Given system (60) and the performance index (63), the optimal virtual input \( \hat{u}(t) \) can be obtained through the methods proposed in this article. Further the optimal control policy \( u^*_h(t) \) is obtained by \( u^*_h(t) = u_s(t) + \hat{u}(t) \).

B. Simulation Result and the Comparison With the Existing Control Law

To validate the performance of the designed circumnavigation control law, a numerical simulation is presented in this section. In the simulation, the UAV is expected to circumnavigate around a ground target moving with a linear speed of 5 m/s. The desired circumnavigation radius and the height of the UAV are set as 50 and 80 m, respectively. The relative linear speed of the UAV w.r.t. the target is 10 m/s. The standard deviation parameters \( \sigma_r \) and \( \sigma_\eta \) are assumed to be \( \sigma_r = 2 \times 10^{-3} \text{m} \) and \( \sigma_\eta = 1.5 \times 10^{-4} \text{rad} \), respectively.

In the simulation, the following NN is utilized to approximate the value function:

\[
V_{350}(e_r, \eta, \theta, \theta_\theta) = \sum_{i=1}^{35} \sum_{j=1}^{10} w_{10(i-1)+j} a_i b_j
\]

where \( a_i \) and \( b_j \) are, respectively, the \( i \)th and \( j \)th elements of the vectors \( \bar{a} \) and \( \bar{b} \) defined as follows:

\[
\bar{a} = \begin{bmatrix}
 e_r^2, e_r \eta, \eta^2, e_r^4, e_r^3 \eta, e_r^2 \eta^2, e_r \eta^3, \eta^4, e_r^6, e_r^5 \eta, e_r^4 \eta^2, e_r^3 \eta^3, e_r^2 \eta^4, e_r \eta^5, e_r \eta^6, e_r \eta^7, \eta^8, e_r^9, e_r^8 \eta, e_r^7 \eta^2, e_r^6 \eta^3, e_r^5 \eta^4, e_r^4 \eta^5, e_r^3 \eta^6, e_r^2 \eta^7, e_r \eta^8, e_r \eta^9, \eta^{10}
\end{bmatrix}^T
\]

\[
\bar{b} = \begin{bmatrix}
 1, \theta, \theta_\theta, \theta_\theta^2, \theta_\theta^3, \theta^2 \theta_\theta, \theta \theta_\theta^2, \theta \theta_\theta^3
\end{bmatrix}^T.
\]

The simulation is conducted with a sample frequency of 200 Hz. For each iteration, 40,000 samples are collected and used to update the control policy. The convergence of eight representative weights of the NN is demonstrated in Fig. 3. It can be observed that after four iterations, the weights are all nearly convergent. Thus, the stableness and convergence of the proposed IRL-based PI algorithm are verified by Fig. 3.

The trajectories of the ground target and the UAVs controlled by our designed control law are demonstrated in Fig. 4. The variation of the circumnavigation radius \( r_h \) controlled by our method is demonstrated in Fig. 6 (red solid line). It is illustrated that the circumnavigation radius \( r_h \) converges to
Fig. 5. Variation of $\eta$ during the circumnavigation.

Fig. 6. Variation of the relative distances $r_h$ during the circumnavigation.

Fig. 7. Variation of the control input.

the preset radius 50 m. Fig. 5 is the illustration of the variation of the state $\eta$ controlled by the proposed algorithm during the simulation, which converges to zero at around 17 s. The control input of the UAV by our method is illustrated in Fig. 7. Note that the upper bound (blue dash line in Fig. 7) and lower bound (black dash line in Fig. 7) for the control input are changing over time as the speed of the UAV varies. Fig. 7 shows that the control input of the UAV is always within the allowed range during the process of the circumnavigation.

To demonstrate the validity of our method, the proposed control law is further compared with the VF guidance law [20]. The comparison of the relative distance $r_h$ between the two methods is demonstrated in Fig. 6. Obviously, the circumnavigation radius controlled by our method converges faster than that of the VF guidance law. Fig. 8 compares the accumulative information of the two methods before the UAV achieves the desired circumnavigation. Note that after achieving the circumnavigation with the desired radius, the UAVs controlled by the two methods gain the same accumulative information in unit time as illustrated in Fig. 9. Therefore, in order to clearly illustrate the difference of the accumulative information obtained by the two methods, Fig. 8 only illustrates the first 25 s of the accumulative information. It can be observed from Fig. 8 that the accumulative information acquired by our method is higher than that of the VF method. In order to show this more clearly, we demonstrate the obtained accumulative information per simulation step, i.e., the value of the function $\hat{Q}(r_h, \eta)$, of the two methods in Fig. 9. Overall, the UAV controlled by our method gains more accumulative information before the UAV achieves the desired circumnavigation (except for the first 3 s). After the UAV achieves the desired circumnavigation, the values of the function $\hat{Q}(r_h, \eta)$ obtained by the two methods are the same.

IV. CONCLUSION

In this article, we have addressed the optimal control problem for NF nonlinear systems with unsymmetrical and state-dependent input constraint. The method proposed in this article relaxes the assumptions on the dynamics and the input constraints of optimal control systems in the existing works. The proposed method is applied to solve an application case: the optimal UAV circumnavigation control problem. The control performance of our algorithm has been compared with the algorithm proposed in [20] by using a numerical simulation.

In the future work, we will extend the proposed optimal control design method to the multiagent systems and further investigate the optimal cooperative circumnavigation control problem of the multi-UAV systems.
Yangguang Yu received the B.S., M.S., and Ph.D. degrees in control science and engineering from the National University of Defense Technology, Changsha, China, in 2015, 2017, and 2022, respectively. He is currently a Lecturer with the College of Intelligence Science and Technology, National University of Defense Technology. His research interests include multi-agent systems, unmanned aerial vehicles, and optimal adaptive control.

Xiangke Wang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control science and engineering from the National University of Defense Technology, Changsha, China, in 2004, 2006, and 2012, respectively. Since 2012, he has been serving as a Lecturer, an Associate Professor, and a Professor with the College of Intelligence Science and Technology, National University of Defense Technology. He was a visiting student with the Research School of Engineering, The Australian National University, Canberra, ACT, Australia, from 2009 to 2011. He has authored or coauthored two books and more than 100 publications in peer reviewed journals and international conferences, including IEEE TRANSACTIONS, International Journal of Robust and Nonlinear Control, CDC, IFAC, and ICRA. His current research interests focus on the control of multi-agent systems and its applications on unmanned aerial vehicles.

Zhiyong Sun (Member, IEEE) received the Ph.D. degree in control science and engineering from The Australian National University (ANU), Canberra, ACT, Australia, in February 2017. He was a Research Fellow/Lecturer with the Research School of Engineering, ANU from 2017 to 2018. From June 2018 to January 2020, he worked as a Postdoctoral Researcher with the Department of Automatic Control, Lund University, Lund, Sweden. Since January 2020, he has been an Assistant Professor with the Eindhoven University of Technology, Eindhoven, The Netherlands. His research interests include multi-robotic systems, control of autonomous formations, and distributed control and optimization.

Lincheng Shen received the B.S., M.S., and Ph.D. degrees in automatic control from the National University of Defense Technology (NUDT), Changsha, China, in 1986, 1989, and 1994, respectively. In 1989, he joined the Department of Automatic Control, NUDT, where he is currently a Full Professor and serves as the Dean of the Graduate School. His research interests include unmanned aerial vehicles, swarm robotics, and artificial intelligence. Dr. Shen has been serving as an Editorial Board Member of the Journal of Bionic Engineering since 2007.