Integral Reinforcement Learning-Based Multi-Robot Minimum Time-Energy Path Planning Subject to Collision Avoidance and Unknown Environmental Disturbances

Chenyuan He, *Graduate Student Member, IEEE*, Yan Wan, *Senior Member, IEEE*, Yixin Gu, *Member, IEEE*, and Frank L. Lewis, *Life Fellow, IEEE*

***Abstract*—In this letter, we study the online multi-robot minimum time-energy path planning problem subject to collision avoidance and input constraints in an unknown environment. We develop an online adaptive solution for the problem using integral reinforcement learning (IRL). This is achieved through transforming the finite-horizon minimum time-energy problem with input constraints to an approximate infinite-horizon optimal control problem. To achieve collision avoidance, we incorporate artificial potential fields into the approximate cost function. We develop an IRL-based optimal control strategy and prove its convergence. The theoretical results are verified through simulation studies.**

***Index Terms*—Robotics, optimal control, constrained control, machine learning, uncertain systems.**

I. Introduction

THE ROBOT technology was developed rapidly during the last decades, with applications that span manufacturing, agriculture, disaster response, transportation and services. To ensure a safe multi-robot system, the online path planning of multiple robots subject to collision avoidance, input constraints and unknown environmental disturbances is crucial.

The multi-robot path planning problem subject to collision avoidance has been considerably studied in literature. The problem can be broadly classified into offline versus online planning, centralized versus decentralized planning, and cooperative versus non-cooperative planning [1]. The objective of multi-robot path planning can be the minimization of time, energy, hybrid time-energy or risks [2]. Among them, minimum time and time-energy problems are non-trivial to solve because the final time to minimize is unknown and standard optimal control solutions do not work [3].

Robots of practical applications often operate in a complex environment, such as wind fields, ocean currents and other weather conditions [4]. With known environmental information, solutions to Dubins path and Zermelo-Markov-Dubins types of problems have been developed [5]. These offline methods are often not effective because precise environmental information is usually unknown in practice. Therefore, online adaptive solutions become valuable. Online multi-robot path planning approaches include estimation-based, vision-based, data-driven, and reinforcement learning (RL)-based methods. In [6], weather conditions are estimated using ocean circulation models and satellite measurements to facilitate path planning. In [7], an omni-directional vision sensing system is adopted to identify the positions and velocities of other robots. In [8], a data-driven framework combines offline query from historical wind scenarios and online tuning to facilitate fast online path planning of multiple unmanned aerial vehicles (UAV). All the above methods rely on additional on-board devices or weather services. RL, on the other hand, allow agents to learn the optimal control solutions adaptively without extensive knowledge of the environment. Some studies have applied RL to the multi-robot path planning problem with collision avoidance in unknown environments [9], [10]. However, these methods solve the problem in discrete state and control input spaces and thus may lose precision.

In this letter, we study the online multi-robot minimum time-energy path planning problem subject to collision avoidance and input constraints in an unknown environment. The minimum time-energy problem is challenging to solve using RL because the final time is not a fixed value. The solutions coming out of the Pontryagin's minimum principle for the problems are discontinuous. Different from existing minimum time-energy studies that are offline without unknown

Manuscript received March 16, 2020; revised June 8, 2020; accepted June 29, 2020. Date of publication July 7, 2020; date of current version July 22, 2020. This work was supported in part by the National Science Foundation for Research Support Germane under Grant 1724248, Grant 1714519, Grant 1714826, and Grant 1839804, in part by the Office of Naval Research under Grant N00014-18-1-2221, and in part by the Army Research Office under Grant W911NF-20-1-0132. Recommended by Senior Editor C. Seatzu. *(Corresponding author: Yan Wan.)*

The authors are with the Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: chenyuan.he@mavs.uta.edu; yan.wan@uta.edu; yixin.gu@uta.edu; lewis@uta.edu).

Digital Object Identifier 10.1109/LCSYS.2020.3007663

environmental conditions and in discrete time, state or control input spaces [3], we here develop an online integral reinforcement learning (IRL) solution in continuous time, state and control input spaces to achieve improved precision subject to unknown environmental conditions. IRL was first developed in [11] to address continuous time RL problems, and was further enhanced in e.g., [12], [13]. The basic idea of IRL is to minimize the integral temporal difference (TD) error for a period of time. The implementation often relies on the least squares and neural network approximation. The contributions of this letter are summarized as follows.

- We develop an IRL-based online adaptive solution to solve the multi-robot minimum time-energy problem subject to collision avoidance, input constraints and unknown environmental disturbances. To the best of our knowledge, this is the first online solution to this free final time path planning problem in continuous time, continuous state and control input spaces subject to unknowns and constraints.
- We provide a novel approximate cost function that is solvable by IRL for this minimum time-energy multi-robot path planning problem. The cost function includes a hyperbolic tangent function to transform the original finite-horizon problem to an infinite-horizon problem so that IRL can be readily applied. To deal with control input constraints and minimum energy consumption, generalized nonquadratic functionals and their estimation procedure are provided.
- To achieve collision avoidance, an artificial potential field is incorporated into the approximate cost function. A special weight matrix is designed to counteract the non-zero tail such that a finite approximate cost function can be obtained.
- We provide a convergence proof for the proposed IRL-based minimum time-energy solution under constraints using a Lyapunov approach.

The reminder of this letter is organized as follows. Section II formulates the multi-robot path planning problem subject to collision avoidance. In Section III, we introduce the approximate cost function and develop an IRL-based policy iteration (PI) solution. Section IV provides a convergence proof. Section V verifies the proposed approach using a simulation study. Section VI concludes this letter.

II. Problem Formulation

We consider the problem of navigating N robots from their initial locations (x_{i0}, y_{i0}) to destinations (x_{if}, y_{if}), $i \in \{1, 2, \ldots, N\}$, in a 2-D plane. Denote the position and velocity of robot i along the X and Y axes at time instant t as $L_i(t) = [x_i(t), y_i(t)]^{\mathrm{T}}$ and $V_i(t) = [v_{ix}(t), v_{iy}(t)]^{\mathrm{T}}$ respectively, where $^{\mathrm{T}}$ denotes the matrix transpose. The initial conditions are $L_i(0) = L_{i0} = [x_{i0}, y_{i0}]^{\mathrm{T}}$ and $V_i(0) = [0, 0]^{\mathrm{T}}$. Let $X_i(t) = [L_i(t)^{\mathrm{T}}, V_i(t)^{\mathrm{T}}]^{\mathrm{T}}$ represent the system state of robot i, $U_i(t) = [u_{ix}(t), u_{iy}(t)]^{\mathrm{T}}$ and $W_i(t) = [w_{ix}(t), w_{iy}(t)]^{\mathrm{T}}$ denote the control inputs and unknown environmental disturbances along the X and Y axes for robot i respectively. We here consider a generic second-order linear dynamics for the robots,

$$\dot{X}_i(t) = \begin{bmatrix} \dot{L}_i(t) \\ \dot{V}_i(t) \end{bmatrix} = \begin{bmatrix} A_s V_i(t) + E_s W_i(t) \\ B_s U_i(t) \end{bmatrix} = AX_i(t) + BU_i(t) + EW_i(t), \quad |U_i(t)| \le U_{iM}, \quad (1)$$

where

$$A = \begin{bmatrix} O_{2\times 2} & A_s \\ O_{2\times 2} & O_{2\times 2} \end{bmatrix}, \quad B = \begin{bmatrix} O_{2\times 2} \\ B_s \end{bmatrix}, E = \begin{bmatrix} E_s \\ O_{2\times 2} \end{bmatrix}. \quad (2)$$

The control inputs are bounded by $U_{iM} = [U_{ixM}, U_{iyM}]^{\mathrm{T}}$, and instantaneous changes are allowed if they are within the bounds. $O_{m\times n}$ is a zero matrix of size $m \times n$. (A, B) is controllable. To avoid collision, the distance $d_{ij}(t) = \|L_i(t) - L_j(t)\|$ between any two robots i and j at time t should be larger than a safety distance r_s, where $\|\cdot\|$ denotes the Euclidean norm. Our goal is to find the minimum time-energy trajectory with collision avoidance and the corresponding optimal control for the N robots such that $L_i(T) = L_{if} = [x_{if}, y_{if}]^{\mathrm{T}}$ and $\dot{L}_i(T) = O_{2\times 1}$ for all $i \in \{1, 2, \ldots, N\}$. We assume that robots are well separated at their destinations, i.e., $\|L_{if} - L_{jf}\| >> r_s$ is satisfied for all $i \neq j$.

Problem 1: Given the initial positions $L_0 = [L_{10}, L_{20}, \ldots, L_{N0}]^{\mathrm{T}}$ and destinations $L_f = [L_{1f}, L_{2f}, \ldots, L_{Nf}]^{\mathrm{T}}$ of the N robots, find the optimal control laws $U = [U_1, U_2, \ldots, U_N]^{\mathrm{T}}$ such that the total travel time and energy for the robots to arrive at their destinations under unknown disturbances $W = [W_1, W_2, \ldots, W_N]^{\mathrm{T}}$ are minimized, and the safety distance constraint r_s is met for all robot pairs at all times. Mathematically,

$$\min_U J = \int_0^T (\rho + U^{\mathrm{T}} P U) dt$$
$$s.t.: \begin{cases} \dot{X}(t) = I_N \otimes AX(t) + I_N \otimes BU(t) + I_N \otimes EW(t), \\ |U(t)| \le U_M, \\ L(0) = L_0, \quad V(0) = O_{2N\times 1}, \\ L(T) = L_f, \quad \dot{L}(T) = O_{2N\times 1}, \\ d_{ij} > r_s \ \forall i, j \in \{1, 2, \ldots, N\} \text{ and } i \neq j, \end{cases} \quad (3)$$

where $X(t) = [X_1(t), X_2(t), \ldots, X_N(t)]^{\mathrm{T}}$, $L(t) = [L_1(t), L_2(t), \ldots, L_N(t)]^{\mathrm{T}}$, $V(t) = [V_1(t), V_2(t), \ldots, V_N(t)]^{\mathrm{T}}$, $U_M = [U_{1M}, U_{2M}, \ldots, U_{NM}]^{\mathrm{T}}$, ρ is a positive constant to account for the importance of travel time, P is a positive diagonal matrix, I_N is an identity matrix of size N, and $\otimes$ is a Kronecker product.

Problem 1 optimizes both travel time and energy in its quadratic form. We note that the upper limit of the integral T, the travel time to minimize is not known and hence the problem is different from the fixed time problem commonly studied in the RL literature. In addition, the unknown environment condition and the existence of control input and collision avoidance constraints further complicate the problem. The problem can not be solved using traditional optimal control methods such as the Pontryagin's minimum principle or the Hamilton–Jacobi–Bellman (HJB) equation.

III. Multi-Robot Path Planning Under Unknown Environmental Disturbances Using IRL Method

In this section, we develop an online solution to adaptively learn the unknown disturbance and solve the minimum time-energy path planning problem subject to collision avoidance in continuous time, continuous state and control input spaces. To do that, we introduce an approximate cost function to transform the problem to a problem that is solvable by IRL and then develop a PI solution.

A. The Approximate Cost Function

We introduce a novel approximate cost function $V(X,0)$ as follows,

$$V(X,0)=\int_0^{\infty}\Big(\rho\tanh\Big((L(t)-L_f)^{\mathrm{T}}(L(t)-L_f)\Big)+\Phi(U(t))+\Omega_R(t)^{\mathrm{T}}f_R(D(t))\Big)dt. \quad (4)$$

The first term under the integral is a hyperbolic tangent function that transforms the finite-horizon minimum time-energy optimal control problem to an infinite-horizon problem. The second term is a generalized nonquadratic functional that deals with the constrained control inputs and energy consumption. The third term under the integral captures the collision avoidance constraint using the concept of artificial potential field. We introduce in details each of these terms and compare them with the terms in (3) to show that $V(X,0)$ well approximates J in the original Problem 1. Hence, the solution to (4) also well approximates the solution to (3).

1) A Hyperbolic Tangent Function to Approximate Minimum Time: The first term $\rho\tanh((L(t)-L_f)^{\mathrm{T}}(L(t)-L_f))$ under the integral in (4) is a hyperbolic tangent function to approximate the minimum time objective [14]. When $L(t)$ is far away from L_f, $\rho\tanh((L(t)-L_f)^{\mathrm{T}}(L(t)-L_f))$ is equal to ρ in (3). When $L(t)$ approaches L_f, $\tanh((L(t)-L_f)^{\mathrm{T}}(L(t)-L_f))$ decreases and approaches 0. Because the tanh function is odd and monotonic, the integral of this term is minimized only when $L(t)=L_f$. The upper limit of the integral is thus extended from T in (3) to ∞, and the state at the fixed final time at ∞ is now incorporated in V instead of the state at an unknown finite time T. The tanh function is also continuous and differentiable. This approximate minimum time problem is thus a form solvable by IRL.

2) Generalized Nonquadratic Functionals to Capture Constrained Inputs and Energy Consumption: The second term $\Phi(U(t))$ under the integral of V is a generalized nonquadratic functional [15] to approximate the minimum energy cost and also capture the input constraints. We first show its expression, then discuss its properties, and in the end show the approximation procedure.

$\Phi(U(t))$ is defined as

$$\Phi(U(t))=2\int_0^{U(t)}\phi^{-1}(\xi)R\,d\xi, \quad (5)$$

where R is a positive diagonal matrix, and $\phi(\xi)=[\phi(\xi_{1x}),\phi(\xi_{1y}),\phi(\xi_{2x}),\ldots,\phi(\xi_{Nx}),\phi(\xi_{Ny})]$ is a monotonic odd function with bounded first derivative. $\phi(\xi_{ix})$ is specifically constructed using *tanh*() [16],

$$\phi(\xi_{ix})=U_{ixM}\tanh\left(\frac{\xi_{ix}}{U_{ixM}}\right). \quad (6)$$

$\phi(\xi_{iy})$ is constructed in a similar way. Note that $d\xi$ is a column vector consisting of $d\xi_{ix}$ and $d\xi_{iy}$, where $i\in\{1,2,\ldots,N\}$, and hence $\Phi(U(t))$ is a scalar.

$\Phi(U(t))$ is a smooth real-valued positive definite performance integrand. Now we show that $\Phi(U(t))$ is a symmetric function with minimum value 0 by examining each of its element. Let $l(u_{ix})=2\int_0^{u_{ix}(t)}\phi^{-1}(\xi_{ix})d\xi_{ix}$. $\Phi(U(t))=\sum_{i=1}^{N}(l(u_{ix})R_{ix}+l(u_{iy})R_{iy})$, where R_{ix} and R_{iy} are the diagonal elements in R corresponding to u_{ix} and u_{iy} respectively. The element $l(u_{ix})R_{ix}$ is calculated as

$$l(u_{ix})R_{ix}=2U_{ixM}R_{ix}u_{ix}(t)\tanh^{-1}\left(\frac{u_{ix}(t)}{U_{ixM}}\right)+U_{ixM}^2R_{ix}\ln\left(1-\frac{u_{ix}^2(t)}{U_{ixM}^2}\right). \quad (7)$$

Because $l(u_{ix})R_{ix}$ is a symmetric function with minimum value 0, $\Phi(U(t))$ is also a symmetric function with minimum value 0. $\Phi(U(t))$ is equal to 0 if and only if $U(t)=O_{2N\times1}$.

To best approximate the energy consumption $U^{\mathrm{T}}PU$ in (3), we calculate the parameters R_{ix} and R_{iy} based on the mean squared error (MSE) such that the difference between the approximate energy and original energy is minimized.

$$R_{ix}=\frac{P_{ix}\int_{-U_{ixM}}^{U_{ixM}}u_{ix}^2l(u_{ix})du_{ix}}{2\int_{-U_{ixM}}^{U_{ixM}}l^2(u_{ix})du_{ix}},\qquad R_{iy}=\frac{P_{iy}\int_{-U_{iyM}}^{U_{iyM}}u_{iy}^2l(u_{iy})du_{iy}}{2\int_{-U_{iyM}}^{U_{iyM}}l^2(u_{iy})du_{iy}}, \quad (8)$$

where P_{ix} and P_{iy} are the diagonal element in P corresponding to u_{ix} and u_{iy} respectively. An illustration of the approximation is shown in Figure 1(a). We see that $\Phi(U(t))$ well approximates $U^{\mathrm{T}}PU$. In addition to approximating energy consumption, the tanh function in $\phi(\cdot)$ also guarantees that the control inputs u_{ix} and u_{iy} are constrained by U_{ixM} and U_{iyM} respectively. This will be explained later in (15) where the explicit form of $U(t)$ is derived.

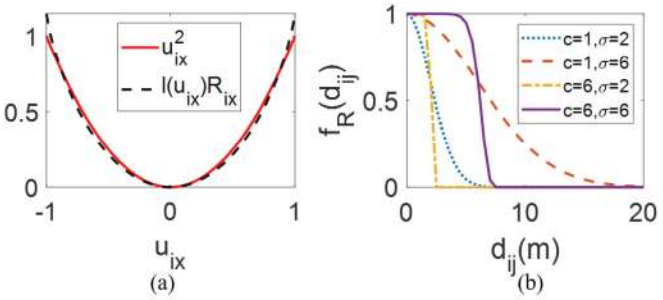

Fig. 1. (a) An illustration of $l(u_{ix})R_{ix}$ to approximate u_{ix}^2 with $R_{ix}=0.83$. (b) An illustration of the Gaussian repulsor function $f_R(d_{ij}(t))$ between two robots i and j.

3) Artificial Potential Fields to Capture the Collision Avoidance Constraint: The main idea is to emanate a repulsive potential field among the robots to force them avoid each other. A two-dimensional Gaussian repulsor function [17] between any two robots i and j at time t is defined as $f_R(d_{ij}(t)) = e^{-0.5(\frac{d_{ij}^2(t)}{\sigma^2})^c}$, where c determines the steepness of the repulsor function and σ determines the repulsive range. In particular, a large c leads to a steep shape, and a large σ leads to a wide repulsive range as shown in Figure 1(b). To capture the safety distance constraint r_s in Problem 1, we compute σ and c by setting the repulsor function value at r_s and $r_s+\Delta$ as

$$f_R(r_s) = \kappa_1,\ f_R(r_s+\Delta) = \kappa_2,\ 0 < \kappa_2 < \kappa_1 < 1, \quad (9)$$

where Δ is a small positive scalar, κ_1 is close to 1 and κ_2 is close to 0. $f_R(d_{ij})$ increases dramatically when the distance d_{ij} decreases to r_s. Solving (9), we obtain a steep repulsor function to avoid collision with parameters,

$$c = \frac{\ln(\log_{\kappa_1}\kappa_2)}{2\ln(1+\frac{\Delta}{r_s})},\ \sigma = e^{\ln r_s - \frac{\ln(-2\ln\kappa_1)}{2c}}. \quad (10)$$

To account for collision avoidance between any pair of robots, we incorporate their Gaussian repulsor functions into the approximate cost function V as shown in the third term under the integral in (4). We define a distance matrix $D(t) = [d_{12}(t), d_{13}(t), \ldots, d_{N-1,N}(t)]^{\mathrm{T}}$ and a corresponding weight matrix $\Omega_R(t) = [\Omega_{12}(t), \Omega_{13}(t), \ldots, \Omega_{N-1,N}(t)]^{\mathrm{T}}$ for the repulsor functions. The Gaussian repulsor function is always larger than 0. In order to have a finite V, we design a special weight matrix $\Omega_{ij}(t)$ to counteract the non-zero tail of the repulsor functions as

$$\Omega_{ij}(t) = \beta \tanh\Big(\big\|L_i(t) - L_{if}\big\|^2 + \big\|L_j(t) - L_{jf}\big\|^2\Big), \quad (11)$$

where β is a positive constant accounting for the importance of collision avoidance. By choosing an appropriately large β, the collision avoidance can be achieved. When robots are far away from their destinations, $\Omega_{ij} = \beta$. When they become closer to their destinations, the value of Ω_{ij} decreases. When both robots i and j arrive at their destinations, $\Omega_{ij} = 0$. Because $\Omega_{ij} > 0$ except at the destinations, the only way to make $V = 0$ is when all the robots arrive at their destinations. There is no local minima issue from the potential field.

B. IRL-Based Policy Iteration Algorithm

In this section, we show the reformulated problem ready to be solved using IRL and also provide a PI solution.

Section III-A transforms Problem 1 to an infinite-time horizon optimization problem (4) with robot dynamics

$$\dot{X}(t) = I_N \otimes AX(t) + I_N \otimes BU(t) + I_N \otimes EW(t). \quad (12)$$

The control and the safety distance constraints are not further imposed as they have been incorporated in V in (4). Let $r(X,U,t) = \rho\tanh((L(t)-L_f)^{\mathrm{T}}(L(t)-L_f)) + \Phi(U(t)) + \Omega_R(t)^{\mathrm{T}} f_R(D(t))$. The minimum of the cost function is denoted by

$$V^* = \min_U \int_0^\infty r(X,U,t)dt. \quad (13)$$

The HJB equation becomes

$$\left(\frac{\partial V^*}{\partial X}\right)^{\mathrm{T}}(I_N \otimes AX + I_N \otimes BU + I_N \otimes EW) + r(X,U,t) = 0. \quad (14)$$

The corresponding optimal control strategy is solved according to the stationary condition

$$U^* = -\phi\left(\frac{1}{2}R^{-1}(I_N \otimes B)^{\mathrm{T}}\frac{\partial V^*}{\partial X}\right). \quad (15)$$

Equation (15) shows that U is always bounded by U_M according to the properties of $\phi(\cdot)$, and hence the control input constraint is satisfied.

Because of the unknown environmental disturbance, the HJB equation can not be solved directly. Therefore, we utilize the Bellman's optimality principle, adopt a value function approximation (VFA), and develop a PI-based IRL method to learn the unknown environment online. The value function written in the IRL form is

$$V(t) = \int_t^{t+T_s} r(X,U,\tau)d\tau + V(t+T_s). \quad (16)$$

According to the Weierstrass approximation theorem [18], a continuous function on a bounded interval can be approximated using polynomials. Hence we can find a VFA such that $V(t) = \Sigma_M^{\mathrm{T}}\psi_M(X)$, where $\psi_M(X)$ denotes a dense basis set and Σ_M denotes the weights.

Two iterative steps are included in the PI algorithm, i.e., policy evaluation and policy improvement. In policy evaluation, we solve $V(t)$ using (17) based on the current control strategy. In policy improvement, the optimal control strategy is updated according to (18). The two steps iterate until a predefined convergence rate is achieved.

Policy Evaluation:

$$V^j(t) = \int_t^{t+T_s} r(X,U^j,\tau)d\tau + V^j(t+T_s). \quad (17)$$

Policy Improvement:

$$U^{j+1} = -\phi\left(\frac{1}{2}R^{-1}(I_N \otimes B)^{\mathrm{T}}\frac{\partial V^j}{\partial X}\right). \quad (18)$$

Combining the VFA, we then write (17) as

$$(\Sigma_M^j)^T[\psi_M(X(t)) - \psi_M(X(t+T_s))] = \int_t^{t+T_s} r(X,U^j,\tau)d\tau. \quad (19)$$

The control policy is updated as

$$U^{j+1} = -\phi\left(\frac{1}{2}R^{-1}(I_N \otimes B)^{\mathrm{T}}\left(\frac{\partial\psi_M(X(t))}{\partial X}\right)^{\mathrm{T}}\Sigma_M^j\right). \quad (20)$$

IV. Convergence Analysis

In this section, we prove the convergence of the proposed IRL-based path planning solution. We show that if starting with an admissible control, the updated control policy at the next iteration is still admissible. In addition, V is a Lyapunov function that decreases over time. Here, a control U is admissible means that it can drive the robot dynamics (12) from the initial positions X_0 to the destinations X_f subject to input and safety constraints and the approximate cost function V in (4) is finite.

Theorem 1: Given an initial admissible control U^0, the policy iteration (17) and (18) converge to the optimal control solution for (4) under an unknown environmental distance $W\forall\ X_0$ and X_f.

Proof: Given an admissible U^j, we show that U^{j+1} is also admissible and $V^* \le V^{j+1} \le V^j$. Taking the derivative of V^j along the $\dot{X} = I_N \otimes AX + I_N \otimes BU^{j+1} + I_N \otimes EW$ trajectory, we have

$$\dot{V}^j(X, U^{j+1}) = \left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes AX + I_N \otimes BU^{j+1} + I_N \otimes EW). \quad (21)$$

According to (14),

$$\left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes EW) = -\left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes AX) - \left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes BU^j) - r(X, U^j). \quad (22)$$

Combining (21) and (22), we have

$$\dot{V}^j(X, U^{j+1}) = \left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes B)(U^{j+1} - U^j) - r(X, U^j). \quad (23)$$

According to (18), we have

$$\left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes B) = -2\phi^{-1}(U^{j+1})R. \quad (24)$$

Substituting (24) into (23), we obtain

$$\begin{aligned}
&\dot{V}^j(X, U^{j+1}) \\
&= -2\left(\phi^{-1}(U^{j+1})R(U^{j+1} - U^j) + \int_0^{U^j} \phi^{-1}(\xi)Rd\xi\right) \\
&\qquad - \Omega_R^{\mathrm{T}} f_R(D) - \rho \tanh\left((L - L_f)^{\mathrm{T}}(L - L_f)\right) \\
&= -2\left(\phi^{-1}(U^{j+1})R(U^{j+1} - U^j) - \int_{U^j}^{U^{j+1}} \phi^{-1}(\xi)Rd\xi \right. \\
&\quad \left. + \int_0^{U^{j+1}} \phi^{-1}(\xi)Rd\xi\right) - \Omega_R^{\mathrm{T}} f_R(D) - \rho \tanh\left((L - L_f)^{\mathrm{T}}(L - L_f)\right). \quad (25)
\end{aligned}$$

Since ϕ^{-1} is monotonic and odd, and R is a positive diagonal matrix, we have

$$\phi^{-1}(U^{j+1})R(U^{j+1} - U^j) - \int_{U^j}^{U^{j+1}} \phi^{-1}(\xi)Rd\xi \ge 0, \quad \int_0^{U^{j+1}} \phi^{-1}(\xi)Rd\xi \ge 0. \quad (26)$$

Therefore, $\dot{V}^j(X, U^{j+1}) < 0$ is always satisfied. Because $V^j(X, U^{j+1})$ is positive definite, continuous and differentiable, we conclude that $V^j(X, U^{j+1})$ is a Lyapunov function for U^{j+1} and U^{j+1} is an admissible control.

The system trajectory of $(I_N \otimes AX + I_N \otimes BU^{j+1} + I_N \otimes EW)\forall\ X_0$ and X_f is shown as follows.

$$\begin{aligned}
&V^{j+1}(X_0, 0) - V^j(X_0, 0) \\
&= -\int_0^\infty \left(\left(\frac{\partial V^{j+1}}{\partial X}\right)^{\mathrm{T}} - \left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}}\right)(I_N \otimes AX + I_N \otimes BU^{j+1} \\
&\quad + I_N \otimes EW)dt. \quad (27)
\end{aligned}$$

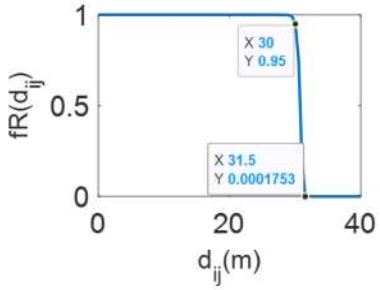

Fig. 2. The repulsor function with $c = 52.5$ and $\sigma = 30.7$.

According to (14),

$$\begin{aligned}
&\left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes EW) \\
&= -\left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes AX) - \left(\frac{\partial V^j}{\partial X}\right)^{\mathrm{T}} (I_N \otimes BU^j) - r(X, U^j), \\
&\left(\frac{\partial V^{j+1}}{\partial X}\right)^{\mathrm{T}} (I_N \otimes EW) \\
&= -\left(\frac{\partial V^{j+1}}{\partial X}\right)^{\mathrm{T}} (I_N \otimes AX) - \left(\frac{\partial V^{j+1}}{\partial X}\right)^{\mathrm{T}} (I_N \otimes BU^{j+1}) - r(X, U^{j+1}). \quad (28)
\end{aligned}$$

Combining (28) and (27), we have

$$V^{j+1}(X_0, 0) - V^j(X_0, 0) = -2\int_0^\infty \left(\phi^{-1}(U^{j+1})R(U^{j+1} - U^j) - \int_{U^j}^{U^{j+1}} \phi^{-1}(\xi)Rd\xi\right)dt. \quad (29)$$

According to (26), $V^{j+1}(X_0, 0) \le V^j(X_0, 0)$. By contradiction, we have $V^* \le V^{j+1}(X_0, 0) \le V^j(X_0, 0)$. ■

V. Simulation Study

We simulate and demonstrate the performance of the proposed multi-robot path planning algorithm in an unknown environment. Consider three robots with initial positions at $(x_{10}, y_{10}) = (185, 160)m$, $(x_{20}, y_{20}) = (0, 160)m$, and $(x_{30}, y_{30}) = (90, 150)m$ respectively. Their destinations are $(x_{1f}, y_{1f}) = (0, 0)m$, $(x_{2f}, y_{2f}) = (185, 0)m$, and $(x_{3f}, y_{3f}) = (0, 60)m$. Here we consider unknown constant environmental disturbances on the X and Y axes as $W_x = W_y = 2m/s$. In the future, we will extend the analysis to spatiotemporal environmental disturbances. Control input constraints are $u_{ixM} = u_{iyM} = 1m/s^2$, where $i \in \{1, 2, 3\}$. In addition, $\rho = 1$, $\beta = 1$, $A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$, $E = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $P = I_6$. Safety distance constraint is set as $r_s = 30m$, $\Delta = 1m$.

We first solve an appropriate repulsor function by setting $f_R(r_s) = 0.95$ and $f_R(r_s + \Delta) = 0.2$, and obtain $c = 52.5$ and $\sigma = 30.7$. The steep repulsor function is shown in Figure 2. Then we solve $R = 0.83I_6$ according to (8). We use a VFA to approximate the cost function V for the three-robot path planning problem. The IRL time interval and the stopping criteria for PI algorithm are set as $T_s = 0.15$s and 10^{-3} respectively.

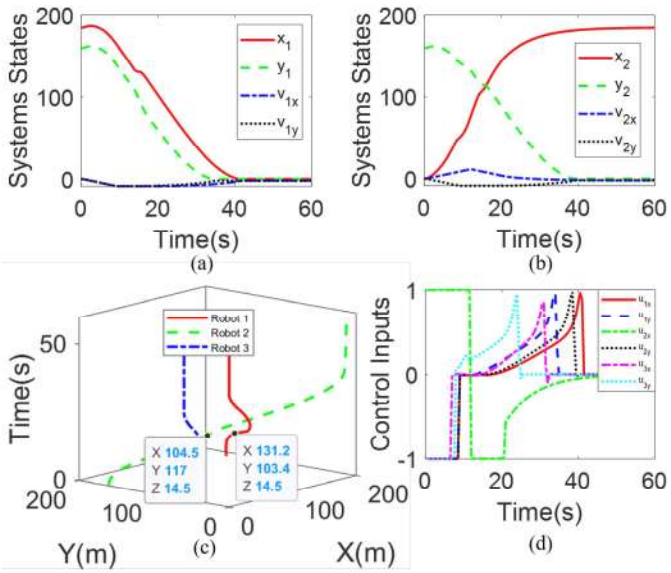

Fig. 3. Simulation results for the proposed IRL-based algorithm of three robots with collision avoidance under unknown environmental disturbances. (a) The system states of robot 1, (b) The system states of robot 2, (c) The robot trajectories, (d) Control inputs.

The simulation results are shown in Figure 3. Figures 3(a) and 3(b) show the states of robots 1 and 2 respectively. They reach their destinations at $T = 40s$. Robot 3's state is omitted here due to space limitation. As shown in Figure 3(c), at time $t = 14.5s$, the distance between the robots 1 and 2 becomes less than r_s and triggers the repulsor function. In particular, it forces robots 1 and 2 to change their directions to avoid collision. When their distance becomes larger than r_s, the repulsor function becomes small and has an infinitesimal effect on the robot directions. Figure 3(d) shows that the control inputs remain within the bounds -1 and 1. The initial admissible controls are required according to Theorem 1. Applying the controls back to the costs in (3) and (4), we find that $J = 190.5$ and $V(X, 0) = 196.4$ for the approximate time-energy consumption. The total approximation gap is less than 3%. The comparison shows that (4) well approximates (3) and our IRL-based path planning solution allows all the robots to reach their destinations with minimum time-energy subject to collision avoidance, control input constraints and unknown environmental disturbances. The total computation time is $315s$, using a Dell XPS 13 laptop with CPU clock time up to 4.9 GHz.

VI. Conclusion

In this letter, we developed an IRL-based online multi-robot minimum time-energy path planning algorithm subject to input constraints and collision avoidance subject to constant or slowly varying unknown disturbances. We introduced an approximate cost function that transforms the problem to a problem solvable using IRL and developed a PI solution in continuous time, state and control input spaces. The approximate cost function contains three items, a hyperbolic tangent function to approximate minimum time, generalized nonquadratic functionals to capture the control constraints and approximate the minimum energy objective, and an artificial potential field for collision avoidance. We proved the convergence for the proposed algorithm and verified its effectiveness using simulation studies. In the future work, we will study algorithms that remove the need of initial admissible controls and extend the algorithms to more complex 3D UAV dynamics and spatiotemporal environmental disturbances. We will also study implementation of the proposed algorithm on a real platform.

References

[1] M. Hoy, A. S. Matveev, and A. V. Savkin, "Algorithms for collision-free navigation of mobile robots in complex cluttered environments: A survey," *Robotica*, vol. 33, no. 3, pp. 463–497, 2015.

[2] L. E. Parker, "Path planning and motion coordination in multiple mobile robot teams," in *Encyclopedia of Complexity and Systems Science*. New York, NY, USA: Springer, 2009, pp. 5783–5800.

[3] O. Wigström, "Energy efficient multi-robot coordination," Ph.D. dissertation, Dept. Signals Syst., Chalmers Univ. Technol., Gothenburg, Sweden, 2016.

[4] N. Dadkhah and B. Mettler, "Survey of motion planning literature in the presence of uncertainty: Considerations for UAV guidance," *J. Intell. Robot. Syst.*, vol. 65, nos. 1–4, pp. 233–246, 2012.

[5] T. McGee, S. Spry, and K. Hedrick, "Optimal path planning in a constant wind with a bounded turning rate," in *Proc. AIAA Guidance Navig. Control Conf. Exhibit*, 2005, p. 6186.

[6] J. C. Rubio, J. Vagners, and R. Rysdyk, "Adaptive path planning for autonomous UAV oceanic search missions," in *Proc. AIAA 1st Intell. Syst. Tech. Conf.*, 2004, p. 6228.

[7] C. Cai, C. Yang, Q. Zhu, and Y. Liang, "Collision avoidance in multi-robot systems," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, 2007, pp. 2795–2800.

[8] C. He, Y. Wan, and J. Xie, "Spatiotemporal scenario data-driven decision for the path planning of multiple uass," in *Proc. 4th Workshop Int. Sci. Smart City Oper. Platforms Eng.*, 2019, pp. 7–12.

[9] H. Bae, G. Kim, J. Kim, D. Qian, and S. Lee, "Multi-robot path planning method using reinforcement learning," *Appl. Sci.*, vol. 9, no. 15, p. 3057, 2019.

[10] W. Luo, Q. Tang, C. Fu, and P. Eberhard, "Deep-sarsa based multi-UAV path planning and obstacle avoidance in a dynamic environment," in *Proc. Int. Conf. Sens. Imag.*, 2018, pp. 102–111.

[11] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[12] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst. Mag.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

[13] S. Li *et al.*, "The design and implementation of aerial communication using directional antennas: Learning control in unknown communication environments," *IET Control Theory Appl.*, vol. 13, no. 17, pp. 2906–2916, Nov. 2019.

[14] M. Abu-Khalaf, J. Huang, and F. L. Lewis, *Nonlinear H2/H-Infinity Constrained Feedback Control: A Practical Design Approach Using Neural Networks*. London, U.K.: Springer, 2006.

[15] S. E. Lyshevski, "Constrained optimization and control of nonlinear systems: New results in optimal control," in *Proc. 35th IEEE Conf. Decis. Control*, vol. 1, 1996, pp. 541–546.

[16] S. E. Lyshevski, "Optimal control of nonlinear continuous-time systems: design of bounded controllers via generalized nonquadratic functionals," in *Proc. IEEE Amer. Control Conf. (ACC)*, vol. 1, 1998, pp. 205–209.

[17] A. Mohamed, J. Ren, A. M. Sharaf, and M. El-Gindy, "Optimal path planning for unmanned ground vehicles using potential field method and optimal control method," *Int. J. Veh. Perform.*, vol. 4, no. 1, pp. 1–14, 2018.

[18] L. N. Trefethen, *Approximation Theory and Approximation Practice*, vol. 164. Philadelphia, PA, USA: Siam, 2019.