Decentralized Adaptive Tracking Control For Large-Scale Multi-Agent Systems Under Unstructured Environment

Shawon Dey
Electrical and Biomedical Engineering
University of Nevada
Reno,NV,USA
shawondey@nevada.unr.edu

Hao Xu
Electrical and Biomedical Engineering
University of Nevada
Reno,NV,USA
haoxu@unr.edu

2022 IEEE Symposium Series on Computational Intelligence (SSCI) | 978-1-6654-8768-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/SSCI51031.2022.10022141

Abstract*—In this paper, a decentralized optimal tracking control problem has been investigated for large scale multi-agent system (LS-MAS) under unstructured environment. Due to the "Curse of Dimensionality" from a large amount of agents and constraints from the unstructured environment, conventional optimal tracking control as well as emerging mean field game and machine learning based design cannot be utilized directly. To overcome those challenges, a novel barrier function has been designed to transform the unstructured environment into a structured environment so that mean field game theory can be used to formulate decentralized optimal tracking control for LS-MAS. Then, the actor-critic-mass reinforcement learning algorithm has been developed to learn the mean field game based optimal solution under structured environment. Specifically, individual agent has three neural networks (NN), i.e., 1) mass NN that learns the behaviors of large population via estimating the solution of Fokker–Planck–Kolmogorov (FPK) equation, 2) critic NN that obtains optimal cost function by learning the solution of the Hamilton–Jacobi–Bellman (HJB) equation, 3) actor NN that solve the decentralized optimal tracking control based on the information provided by the mass and critic NN. Next, the learned decentralized optimal tracking control can be transformed from structured environment back to unstructured environment and implemented in real-time through barrier function. Overall, this developed algorithm is named MFG-based barrier-actor-critic-mass learning. The Lyapunov theorem has been used to prove the stability of the closed-loop system. Eventually, a series of numerical simulation has been conducted to demonstrate the effectiveness of the developed scheme.

I. Introduction

In recent years, large-scale multi-agent systems (LS-MAS) have attracted significant interests from both research societies and industrial communities due to its capability of upgrading traditional multi-agent system performance by using its diversity gain [1]. For instance, [2], [3] have studied the tracking control problem in the LS-MAS. However, It is very difficult to directly utilize conventional control into LS-MAS due to three challenges. The first challenge is the notorious "Curse of Dimensionality" [4]. Since traditional cooperative control needs each agent to know other agents' states, the computational complexity of distributed control will be exponentially increased along with increased number of agents. The second challenge is lacking a realistic reliable communication network that can support information exchange among LS-MAS timely. Due to the limitation of communication capability in practice, traditional distributed cooperative control techniques [5] are very difficult to be applied. The last challenge is the constraints from physical system limitation and practical environment [3], such as obstacles, might cause difficulty in LS-MAS optimal control design.

To address those challenges, a significant number of researches have been conducted. For instance, Liu et al.,[6] developed a cooperative multi-agent traffic signal control system, where the "Curse of Dimensionality" problem has been addressed by integrating Q-learning with function approximation algorithm. In [7], an actor centralized-critic algorithm has been developed to reduce the complexity of cooperation in large scale multi-agent systems. Recently, emerging mean field game (MFG) theory has been widely adopted to solve the decentralized control for LS-MAS due to its capability to handle "Curse of Dimensionality" [8] and [9].

However, using traditional MFG theory, there are two common assumptions need to be ensured, i.e. 1) all agents are located in a structured space, and 2) all agents are homogeneous with respect to their dynamics and environment. When large scale multi-agent system (LS-MAS) has been deployed in the unstructured environment, traditional MFG cannot be directly utilized. To address this issue and develop an efficient decentralized tracking control for LS-MAS under unstructured environment, barrier function [10] based approach has been adopted. Specifically, a novel barrier function is designed to transform the unstructured system space into a structured state space firstly. Then, mean field game theory can be used to formulate the decentralized optimal tracking control problem for LS-MAS under structured space. Next, the optimal solution can be obtained by solving a pair of forward and backward Partial Differential Equation (PDE), called Fokker-Planck-Kolmogorov (FPK) equation and Hamiltonian-Jacobi-Bellman

The support of the National Science Foundation (Grant No. 2144646) is gratefully acknowledged

978-1-6654-8768-9/22/$31.00 ©2022 IEEE

(HJB) equation. It is very difficult and even impossible to directly solve it since those two forward and backward PDEs are closely coupled. To address this difficulty, adaptive dynamic programming and reinforcement learning [11] technique has been adopted. An actor-critic-mass learning algorithm has been developed with mass NN learning the behaviors of large population via estimating the solution of FPK equation with barrier function, critic NN obtaining optimal cost function by learning the solution of the HJB equation with barrier function, and actor NN solving the decentralized optimal tracking control based on the information provided by the mass and critic NN. Eventually, using the inverse barrier function, the practical decentralized optimal tracking control can be obtained and implemented for LS-MAS under unstructured environment in real-time. Overall, the developed algorithm is named as MFG-based barrier-actor-critic-mass learning algorithm.

The key contributions of this paper are listed as follows:

- The challenge from unstructured environment has been overcome by using a novel barrier function based approach. Through barrier function, we are not only able to use mean field game theory for finding optimal solution for LS-MAS under structured environment, but also obtain the corresponding optimal design under unstructured environment.
- The novel barrier-actor-critic-mass algorithm is developed to solve the constrained HJB and FPK equation simultaneously and further obtain the optimal control for LS-MAS in real-time.

The structure of this paper is given as follows. Section II provides the problem formulation including barrier function as well as LS-MAS tracking optimal problem formulation. In Section III, the novel barrier-actor-critic-mass algorithm has been developed. Then, the numerical simulation is shown in Section IV to demonstrate the effectiveness of the proposed design.

II. Problem Formulation

Consider N represents the number of homogeneous agents that deployed in a 2D space, where N is a countably infinite number, i.e, $N \rightarrow \infty$. Now, an agent i is controlled by the stochastic differential equation with their states being constrained

$$dx_{s,i} = [f(x_{s,i}) + g(x_{s,i})u_{s,i}]dt + \sqrt{2\nu}d\sigma_i \qquad (1)$$

where $f(x_{s,i})$ and $g(x_{s,i}) \in \mathbb{R}^2$ are nonlinear functions, $x_{s,i} = \begin{bmatrix} x^1_{s,i} & x^2_{s,i} \end{bmatrix}^T \in \mathbb{R}^2$ is the agent state, which includes its position in x and y coordinates. Also, $u_{s,i} \in \mathbb{R}$ is the control input, σ_i is the standard Brownian motion which represents the process noise and ν is a non-negative parameter. Moreover, the objective of each agent is to reach a predefined destination while avoiding the obstacles in the unstructured environment. Let, there are multiple number of static obstacles in the unstructured space. Considering 2-D space, the boundary region of each obstacle can be defined by the following function

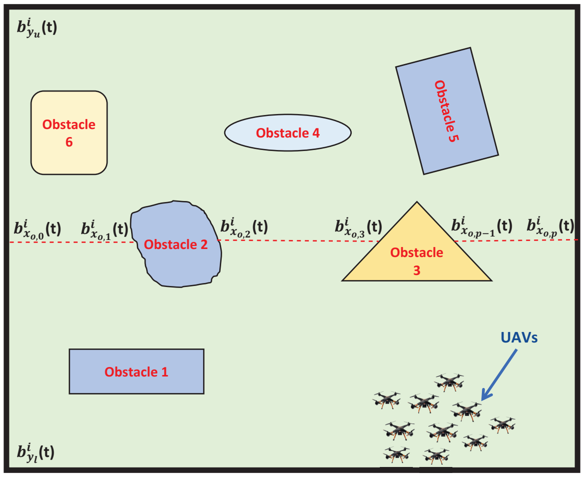

Fig. 1: An illustration of an unstructured environment with multiple obstacles

$$O_o(l) = \{[x_o(l), y_o(l)] \in \mathbb{R}^2 : y_o(l) = h(x_o(l))\} \qquad (2)$$

where, $l = 1, ..., N_o$ with N_o being the number of obstacles in unstructured environment.

A. Unstructured and Structured Space Transformation via Barrier Function

To avoid the collision with multiple distributed obstacles in the unstructured space, we introduce a time varying barrier function. The barrier function transform the system to a new free space, where the LS-MAS MFG control for the homogeneous agents are obtained. At first, the upper and lower bound of the barrier function has been evaluated for each agent with respect to their location at time t. In this study, we consider a two dimensional barrier function. The bounds of the barrier function in 2D space with x and y coordinates depends on the boundary of the obstacles and the configuration space. An environment has been illustrated with multiple obstacles and corresponding boundary positions in Figure 1. The bounds for any agent i can be evaluated by fixing any one coordinate, i.e., horizontal or vertical position. Let, the position of the bounds in y coordinate are the upper and lower boundary of the configuration space. Now, the bounds positions in the x coordinates for agent i at time t can be evaluated by using the boundary function (2). Let $\mathcal{B}^i_{x_o}(t)$ is a set of all boundary position in x coordinate as follows:

$$\mathcal{B}^i_{x_o}(t) = \{b^i_{x_{o,0}}(t), b^i_{x_{o,1}}(t), ..., b^i_{x_{o,p-1}}(t), b^i_{x_{o,p}}(t)\}$$

where $b_{x_{o,0}}(t)$ and $b_{x_{o,p}}(t)$ are the upper and lower bounds of the configuration space, and $p + 1$ is the total number of boundary points for agent i at time t. Now, there exist any subset $\mathcal{S}^i_{x_o}(t) \subseteq \mathcal{B}^i_{x_o}(t)$ and $\mathcal{S}^i_{x,o}(t) = \{(b^i_{x,u}, b^i_{x,l}) \in \mathbb{R} \times \mathbb{R} \mid b^i_{x,u} < x^1_i < b^i_{x,l}\}$, which contains the closest two points in x direction for agent i at time t. Now, the upper and lower bound for agent i at time t can be represented as $b_u^i = \begin{bmatrix} b_{x,u}^i & b_{y,u}^i \end{bmatrix}^T$ and $b_l^i = \begin{bmatrix} b_{x,l}^i & b_{y,l}^i \end{bmatrix}^T$, where $b_{y,u}^i$ and $b_{y,l}^i$ are the fixed upper and lower bound in y coordinate. Now, Let the Barrier function $B(.) : \mathbb{R} \rightarrow \mathbb{R}$ is defined on (b_l^i, b_u^i). Then the state of the agent i can be represented as

$$s_i(t) = B_i(x_{s,i}(t)) = \ln \frac{b_u^i(t)(b_l^i(t) - x_{s,i}(t))}{b_l^i(t)(b_u^i(t) - x_{s,i}(t))} \tag{3}$$

where, $s_i(t)$ is the transformed state of the agent i at time t. Besides that, the Barrier function is invertible on interval $(b_l^i(t), b_u^i(t))$, i.e.,

$$x_{s,i}(t) = B_i^{-1}(s_i(t)) = b_l^i(t) b_u^i(t) \frac{e^{\frac{s_i(t)}{2}} - e^{-\frac{s_i(t)}{2}}}{b_l^i(t) e^{\frac{s_i(t)}{2}} - b_u^i(t) e^{-\frac{s_i(t)}{2}}} \tag{4}$$

Remark 1: The barrier function $B(*)$ takes finite value when the arguments are within the above defined region and approaches to infinity as the state approaches the boundary of the defined region. It allows us to remove the obstacle from the unstructured space and further transform to structured space.

B. MFG-based Decentralized Optimal Tracking Control Problem Formulation under Structured Space

The dynamics of the transformed state s_i can be obtained by using following chain rule

$$\begin{aligned} ds_i &= [f(x_{s,i}) + g(x_{s,i}) u_{s,i}] \frac{{b_u^i}^2 e^{-s_i} - 2 b_l^i b_u^i + {b_l^i}^2 e^{s_i}}{b_u^i {b_l^i}^2 - b_l^i {b_u^i}^2} dt \\ &= [F(s_i) + G(s_i) u_i] dt + \sqrt{2 \upsilon} d\sigma_i \end{aligned} \tag{5}$$

where, $F(s_i) = f(x_{s,i}) \frac{{b_u^i}^2 e^{-s_i} - 2 b_l^i b_u^i + {b_l^i}^2 e^{s_i}}{b_u^i {b_l^i}^2 - b_l^i {b_u^i}^2}$ and
$G(s_i) = g(x_{s,i}) \frac{{b_u^i}^2 e^{-s_i} - 2 b_l^i b_u^i + {b_l^i}^2 e^{s_i}}{b_u^i {b_l^i}^2 - b_l^i {b_u^i}^2}$

Assumption 1. $F(s_i)$ is Lipschitz and there exists a constant a_f such that, for $s_i \in \Omega$, $\|F(s_i)\| \leq a_f \|s_i\|$, where Ω is a compact set containing the origin. Also, $G(s_i)$ is bounded on Ω, i.e., there exists a constant a_g such that $\|G(s_i)\| \leq a_g$. Moreover, the system in Eq. 1 is controllable over the compact set Ω.

Now, a predefined transformed reference trajectory s_r has been given in the unstructured space. Now the objective of each agent is to reach the destination by following the reference trajectory

$$e_i(t) = s_i(t) - s_r(t) \tag{6}$$

Next, the tracking error dynamics can be derived as:

$$\begin{aligned} de_i(t) &= ds_i(t) - ds_r(t) \\ &= [F_a(e_i) + G_a(e_i) u_i] dt + \sqrt{2\nu} d\sigma_i \end{aligned} \tag{7}$$

where, $F_a(e_i) = F(e_i + s_r) - (ds_r/dt)$ and $G_a(e_i) = G(e_i + s_r)$.

Now, the cost function in the transformed state can be represented as follows:

$$V_i(e_i, \rho) = E\Big\{ \int_0^\infty [\mathrm{L}(e_i, u_i) + \Phi(e_i, \rho)] dt \Big\} \tag{8}$$

where, $\rho(e_i, t)$ denotes the probability density function(PDF) of the population's tracking error at time t and $\mathbb{E}\{.\}$ is the expectation operator. Also, $\Phi(e_i, \rho)$ is the mean field coupling function which represents the interaction between agent i and the whole population of other agents. Since the dimension of the PDF and each agent state is same, the mean field coupling function can greatly reduce the computational complexity problem. Moreover, $\mathrm{L}(e_i, u_i) = \|e_i\|_Q^2 + \|u_i\|_R^2$, where Q and R are positive definite matrices with compatible dimensions.

Then, a Hamiltonian [12] can be defined as

$$H_i[e_i, DV_i(e_i, \rho, t)] = \mathbb{E}\Big\{ \begin{bmatrix} \mathrm{L}(e_i, u_i) + DV_i^T(e_i, \rho, t) \\ [F_a(e_i) + G_a(e_i) u_i] \end{bmatrix} \Big\} \tag{9}$$

Next, the optimal control for each agent can be derived as

$$u_i^*(e_i) = -\frac{1}{2} \mathbb{E}\Big\{ R^{-1} (G_a)^T(e_i) DV_i^*(e_i, \rho, t) \Big\} \tag{10}$$

Now, the corresponding HJB equation is obtained by substituting the optimal evaluation function into the Hamiltonian which is shown at the bottom of this page. To obtain the HJB equation, the probability density function (PDF), i.e. Mass function ρ is required. The mass function can be obtained by solving the FPK equation shown at the bottom of the page.

Remark 2: To obtain the optimal control policy, the coupled HJB-FPK equation need to be solved in real time. However, the HJB and FPK equations are multi-dimensional nonlinear PDEs whose solution is difficult to achieve with state constraints. Therefore, in this paper, a novel barrier-actor-critic-mass based NNs is developed to learn the solution of coupled HJB-FPK equations.

III. Barrier-Actor-Critic-Mass Based Optimal Control Design

In this section, the Barrier-Actor-Critic-Mass (BACM) algorithm is developed. In the BACM, each agent maintain three neural networks(NN). The actor NN approximate the optimal control policy, the critic NN approximate the optimal evaluation function and the mass NN estimate the density of the entire population. Meanwhile, the barrier function is applied into three NNs to ensure both state and density constraints being satisfied during the learning.

The optimal cost, control and mass function can be represented as:

$$\begin{aligned} \text{Critic: } V_i^*(e_i, \rho, t) &= \mathbb{E}\Big\{ W_{V,i}^T \phi_{V,i}(e_i, \rho) + \varepsilon_{\text{HJBi}} \Big\} \\ \text{Actor: } u_i^*(e_i, \rho, t) &= \mathbb{E}\Big\{ W_{u,i}^T \phi_{u,i}(e_i, \rho) + \varepsilon_{\text{u,i}} \Big\} \\ \text{Mass: } \rho(e_i, t) &= \mathbb{E}\Big\{ W_{\rho,i}^T \phi_{\rho,i}(e_i, V_i, t) + \varepsilon_{\text{FPKi}} \Big\} \end{aligned} \tag{13}$$

where, $W_{V,i}$, $W_{u,i}$ and $W_{u,i}$ are the critic, actor and mass weights, respectively. Also, $\phi_{V,i}$, $\phi_{u,i}$ and $\phi_{\rho,i}$ are the critic, actor and mass activation functions. Moreover, $\phi(.)$ is the

HJB:

$$\mathbb{E}\Big\{-\partial_t V_i^*(e_i,\rho,t)-\nu\Delta V_i^*(e_i,\rho,t)+H_i[e_i,DV_i^*(e_i,\rho,t)]\Big\}=\mathbb{E}\Big\{\Phi(e_i,\rho)\Big\} \tag{11}$$

FPK:

$$\mathbb{E}\Big\{\partial_t\rho(e_i,t)-\nu\Delta\rho(e_i,t)-div(\rho D_p H_i[e_i,DV_i^*(e_i,\rho,t)])\Big\}=0 \tag{12}$$

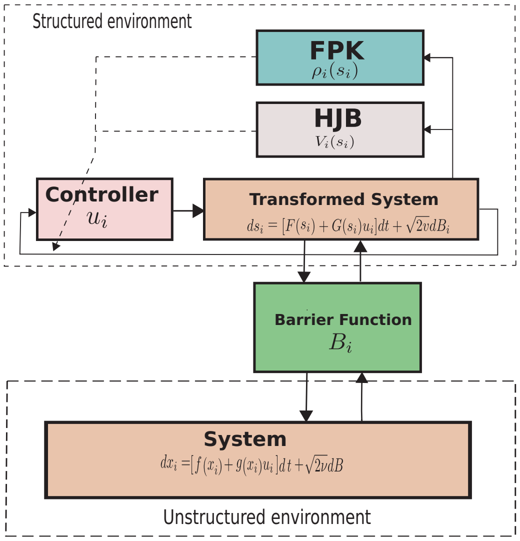

Fig. 2: Structure of Barrier-Actor-Critic-Mass system

bounded and continuous activation function of the respective neural networks. Furthermore, $\varepsilon_{\text{HJBi}}$, $\varepsilon_{\text{u,i}}$ and $\varepsilon_{\text{FPKi}}$ represents the reconstruction error of critic, actor and mass neural network, respectively.

Next, the optimal cost, control and mass distribution function can be approximated as:

$$\begin{aligned}\text{Critic: }\hat{V}_i(e_i,\hat{\rho}_i,t)&=\mathbb{E}\Big\{\hat{W}_{V,i}^T\hat{\phi}_{V,i}(e_i,\hat{\rho})\Big\}\\ \text{Actor: }\hat{u}_i(e_i,\hat{\rho}_i,t)&=\mathbb{E}\Big\{\hat{W}_{u,i}^T\hat{\phi}_{u,i}(e_i,\hat{\rho})\Big\}\\ \text{Mass: }\hat{\rho}(e_i,\bar{\rho},t)&=\mathbb{E}\Big\{\hat{W}_{\rho,i}^T\hat{\phi}_{\rho,i}(e_i,\hat{V}_i,t)\Big\}\end{aligned} \tag{14}$$

where, $\bar{\rho}$ is the averaged historical density defined as $\bar{\rho}=\frac{1}{T}\int_t^{t-T}\rho d\rho$ and T is a constant historical window.

By substituting equations (14) into the HJB, FPK and optimal control equations (11), (10) and (12), we encounter residuals errors which can be used to tune the critic, actor and mass neural networks:

$$\mathbb{E}\{e_{\text{HJBi}}\}=\mathbb{E}\Big\{\Phi(e_i,\hat{\rho}_i)+\hat{W}_{V,i}^T\begin{bmatrix}\partial_t\hat{\phi}_{V,i}+\nu\Delta\hat{\phi}_{V,i}\\-\hat{H}_{i,W}\end{bmatrix}\Big\} \tag{15}$$

$$\mathbb{E}\{e_{\text{FPKi}}\}=\mathbb{E}\Big\{\hat{W}_{\rho,i}^T\begin{bmatrix}\partial_t\hat{\phi}_{\rho,i}-\nu\Delta\hat{\phi}_{\rho,i}\\-div(\hat{\phi}_{\rho,i})D_p\hat{H}_i\end{bmatrix}\Big\} \tag{16}$$

$$\mathbb{E}\{e_{\text{u,i}}\}=\mathbb{E}\Big\{\hat{W}_{u,i}^T\hat{\phi}_{u,i}+\frac{1}{2}R^{-1}(G_a)^T(e_i)D\hat{V}_i(e_i,\hat{\rho}_i,t)\Big\} \tag{17}$$

with $\hat{H}_i=W_{V,i}^T\hat{H}_{i,W}$ and $\hat{H}_i=H_i[e_i,D\hat{\phi}_{V,i}]$.

Now, let

$$\begin{aligned}\mathbb{E}\Big\{\Psi_{V,i}(e_i,\hat{\rho}_i,t)\Big\}&=\mathbb{E}\Big\{\partial_t\hat{\phi}_{V,i}+\nu\Delta\hat{\phi}_{V,i}-\hat{H}_{i,W}\Big\}\\ \mathbb{E}\Big\{\Psi_{\rho,i}(e_i,\bar{\rho}_i,\hat{V}_i,t)\Big\}&=\mathbb{E}\Big\{\begin{bmatrix}\partial_t\hat{\phi}_{\rho,i}-\nu\Delta\hat{\phi}_{\rho,i}\\-div(\hat{\phi}_{\rho,i}D_p\hat{H}_i)\end{bmatrix}\Big\}\\ \mathbb{E}\Big\{\Phi(e_i,\tilde{\rho}_i)\Big\}&=\mathbb{E}\Big\{\Phi(e_i,\hat{\rho}_i)-\Phi(e_i,\rho_i)\Big\}\end{aligned}$$

with $H=W_{V,i}^TH_W$ and $\hat{H}=H[e_i,D_e\hat{\phi}_{V,i}]$.

Then, the HJB and FPK residual errors equation (15) and (16) can be simplified as follows:

$$\mathbb{E}\{e_{\text{HJBi}}\}=\mathbb{E}\Big\{\begin{bmatrix}\Phi(e_i,\rho_i)+\Phi(e_i,\tilde{\rho}_i)\\+\hat{W}_{V,i}^T\Psi_{V,i}(e_i,\hat{\rho}_i,t)\end{bmatrix}\Big\} \tag{18}$$

$$\mathbb{E}\{e_{\text{FPKi}}\}=\mathbb{E}\Big\{\hat{W}_{\rho,i}^T\Psi_{\rho,i}(e_i,\bar{\rho}_i,\hat{V}_i,t)\Big\} \tag{19}$$

Next, we consider the effect of reconstruction errors by substituting the optimal functions from (13) into (11) and (12)

$$\mathbb{E}\Big\{\Phi(e_i,\rho_i)+W_{V,i}^T\Psi_{V,i}(e_i,\rho_i,t)+\varepsilon_{\text{HJBi}}\Big\}=0 \tag{20}$$

$$\mathbb{E}\Big\{W_{\rho,i}^T\Psi_{\rho,i}(e_i,\bar{\rho}_i,V_i,t)+\varepsilon_{\text{FPKi}}\Big\}=0 \tag{21}$$

where, $\varepsilon_{\text{HJBi}}$ and $\varepsilon_{\text{FPKi}}$ are the reconstruction errors. Again, substitute (20) and (21) into (18) and (19)

$$\mathbb{E}\{e_{\text{HJBi}}\}=\mathbb{E}\Big\{\begin{bmatrix}\Phi(e_i,\tilde{\rho}_i)-\tilde{W}_{V,i}^T\Psi_{V,i}(e_i,\hat{\rho}_i,t)\\-W_{V,i}^T\tilde{\Psi}_{V,i}(e_i,\tilde{\rho}_i,t)-\varepsilon_{\text{HJBi}}\end{bmatrix}\Big\} \tag{22}$$

$$\mathbb{E}\{e_{\text{FPKi}}\}=\mathbb{E}\Big\{\begin{bmatrix}-\tilde{W}_{\rho,i}^T\Psi_{\rho,i}(e_i,\bar{\rho}_i,\hat{V}_i,t)\\-W_{\rho,i}^T\tilde{\Psi}_{\rho,i}(e_i,\bar{\rho}_i,\tilde{V}_i,t)\\-\varepsilon_{\text{FPKi}}\end{bmatrix}\Big\} \tag{23}$$

Similarly, we obtain

$$\mathbb{E}\{e_{\text{u,i}}\}=\mathbb{E}\Big\{\begin{bmatrix}-\tilde{W}_{u,i}^T\hat{\phi}_{u,i}(e_i,\hat{\rho}_i,t)\\-W_{u,i}^T\tilde{\phi}_{u,i}(e_i,\tilde{\rho}_i,t)\\-\frac{1}{2}R^{-1}(G_a)^T(e_i)D\tilde{V}_i(e_i,\tilde{\rho}_i,t)-\varepsilon_{\text{u,i}}\end{bmatrix}\Big\} \tag{24}$$

where,

$$\begin{aligned}\mathbb{E}\Big\{\tilde{W}_{V,i}\Big\}&=\mathbb{E}\Big\{W_{V,i}-\hat{W}_{V,i}\Big\}\\ \mathbb{E}\Big\{\tilde{W}_{u,i}\Big\}&=\mathbb{E}\Big\{W_{u,i}-\hat{W}_{u,i}\Big\}\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}\Big\{\tilde{W}_{\rho,i}\Big\} &= \mathbb{E}\Big\{W_{\rho,i} - \hat{W}_{\rho,i}\Big\} \\
\mathbb{E}\Big\{\tilde{\Psi}_{V,i}(e_i,\tilde{\rho}_i,t)\Big\} &= \mathbb{E}\Big\{\Psi_{V,i}(e_i,\rho_i,t) - \Psi_{V,i}(e_i,\hat{\rho}_i,t)\Big\} \\
\mathbb{E}\Big\{\tilde{\Psi}_{\rho,i}(e_i,\bar{\rho}_i,\tilde{V}_i,t)\mathbb{E}\Big\} &= \mathbb{E}\Big\{\begin{bmatrix}\Psi_{\rho,i}(e_i,\bar{\rho}_i,V_i,t) \\ -\Psi_{\rho,i}(e_i,\bar{\rho}_i,\hat{V}_i,t)\end{bmatrix}\Big\} \\
\mathbb{E}\Big\{\tilde{\phi}_{u,i}(e_i,\tilde{\rho}_i,t)\Big\} &= \mathbb{E}\Big\{\phi_{u,i}(e_i,\rho_i,t) - \phi_{u,i}(e_i,\hat{\rho}_i,t)\Big\}
\end{aligned} \tag{25}
$$

Next, applying the gradient descent algorithm, the critic, mass and actor update law can be derived as:

$$
\mathbb{E}\Big\{\dot{\hat{W}}_{V,i}\Big\} = \mathbb{E}\Big\{-\alpha_{V,i}\frac{\Psi_{V,i}(e_i,\hat{\rho}_i,t)e_{\text{HJBi}}^T}{1+\|\Psi_{V,i}(e_i,\hat{\rho}_i,t)\|^2}\Big\} \tag{26}
$$

$$
\mathbb{E}\Big\{\dot{\hat{W}}_{\rho,i}\Big\} = \mathbb{E}\Big\{-\alpha_{\rho,i}\frac{\Psi_{\rho,i}(e_i,\bar{\rho}_i,\hat{V}_i,t)e_{\text{FPKi}}^T}{1+\|\Psi_{\rho,i}(e_i,\bar{\rho}_i,\hat{V}_i,t)\|^2}\Big\} \tag{27}
$$

$$
\mathbb{E}\Big\{\dot{\hat{W}}_{u,i}\Big\} = \mathbb{E}\Big\{-\alpha_{u,i}\frac{\phi_{u,i}(e_i,\hat{\rho}_i,t)e_{\text{ui}}^T}{1+\|\phi_{u,i}(e_i,\hat{\rho}_i,t)\|^2}\Big\} \tag{28}
$$

where $\alpha_{V,i}$, $\alpha_{\rho,i}$ and $\alpha_{u,i}$ are the learning rates.

Algorithm 1 BACM Algorithm

```
Initialize agents $i$'s state $x_{s,i}$
Transform the state $x_{s,i}$ to $s_i$ using (3)
Calculate error $e_i$
Initialize NN weights $\hat{W}_{V,i}$, $\hat{W}_{\rho,i}$, $\hat{W}_{u,i}$ randomly
Initialize errors $e_{\text{FPKi}}$, $e_{\text{HJBi}}$, $e_{u,i} \leftarrow \infty$
Initialize thresholds $\delta_{\text{FPK}}$, $\delta_{\text{HJB}}$, $\delta_u$
while TRUE do
    while $e_{\text{FPKi}} \geq \delta_{\text{FPK}}, e_{\text{HJBi}} \geq \delta_{\text{HJB}}, e_{ui} \geq \delta_u$ do
        Update NN weights by solving (26), (27), and (28),
            $\mathbb{E}\Big\{\dot{\hat{W}}_{V,i}\Big\} = \mathbb{E}\Big\{-\alpha_{V,i}\frac{\Psi_{V,i}(e_i,\hat{\rho}_i,t)e_{\text{HJBi}}^T}{1+\|\Psi_{V,i}(e_i,\hat{\rho}_i,t)\|^2}\Big\}$
            $\mathbb{E}\Big\{\dot{\hat{W}}_{\rho,i}\Big\} = \mathbb{E}\Big\{-\alpha_{\rho,i}\frac{\Psi_{\rho,i}(e_i,\bar{\rho}_i,\hat{V}_i,t)e_{\text{FPKi}}^T}{1+\|\Psi_{\rho,i}(e_i,\bar{\rho}_i,\hat{V}_i,t)\|^2}\Big\}$
            $\mathbb{E}\Big\{\dot{\hat{W}}_{u,i}\Big\} = \mathbb{E}\Big\{-\alpha_{u,i}\frac{\phi_{u,i}(e_i,\hat{\rho}_i,t)e_{\text{ui}}^T}{1+\|\phi_{u,i}(e_i,\hat{\rho}_i,t)\|^2}\Big\}$
        Update NN errors by (22), (23), and (24)
    end while
    $\hat{u}_i(e_i,\hat{\rho}_i,t) \leftarrow \hat{W}_{u,i}^T\hat{\phi}_{u,i}(e_i,\hat{\rho})$
    Execute the control $\hat{u}_i$
    Observe new state $s_i$
end while
Transform the state $s_i$ to $x_{s,i}$ using (4)
```

The performance of all neural networks are given as follows: **Theorem 1:** Let $\mathbb{E}\{\hat{W}_{V,i}\}$ can be updated as (26), where the learning rate $\alpha_{V,i} > 0$. Now, we can say that the error between actual and approximated critic NN's weights $\mathbb{E}\{\tilde{W}_{V,i}\}$ and also the optimal evaluation function approximation errors, i.e., $\mathbb{E}\{\tilde{V}_i\} = \mathbb{E}\{V_i - \hat{V}_i\}$ are uniformly ultimately bounded(UUB). Moreover, $\mathbb{E}\{\tilde{W}_{V,i}\}$ and $\tilde{V}_i$ are asymptotically stable if the NN's structures are selected perfectly. Also, if the reconstruction errors are sufficiently small, then the corresponding bounds of the critic weight and optimal evaluation function i.e., $b_{W_{V,i}}$ and $b_{V,i}$ are insignificant.

Theorem 2: Let $\mathbb{E}\{\hat{W}_{\rho,i}\}$ can be updated as (27) where the learning rate $\alpha_{\rho,i} > 0$. Now, we can say that the error between actual and approximated mass NN's weights $\mathbb{E}\{\tilde{W}_{\rho,i}\}$ and also the mass function approximation errors, i.e., $\mathbb{E}\big\{\tilde{\rho}_i\big\} = \mathbb{E}\big\{\rho_i - \hat{\rho}_i\big\}$ are uniformly ultimately bounded(UUB). Moreover, $\tilde{W}_{\rho,i}$ and $\mathbb{E}\{\tilde{W}_{\rho,i}\}$ are asymptotically stable if the NN's structures are selected perfectly. Also, if the reconstruction errors are sufficiently small, then the corresponding bounds $b_{W_{\rho,i}}$ and $b_{\rho,i}$ are insignificant.

Theorem 3: Let $\mathbb{E}\{\hat{W}_{u,i}\}$ can be updated as (28), where the learning rate $\alpha_{u,i} > 0$. Now, we can say that the error between actual and approximated actor NN's weights $\mathbb{E}\{\tilde{W}_{u,i}\}$ and also the optimal control approximation errors, i.e., $\mathbb{E}\big\{\tilde{u}_i\big\} = \mathbb{E}\big\{u_i - \hat{u}_i\big\}$ are uniformly ultimately bounded(UUB). Moreover, $\mathbb{E}\{\tilde{W}_{u,i}\}$ and $\tilde{u}_i$ are asymptotically stable if the NN's structures are selected perfectly. Also, if the reconstruction errors are sufficiently small, then the corresponding bounds $b_{W_{u,i}}$ and $b_{u,i}$ are insignificant.

Lemma 1: There exist optimal control policies for agent u_i^* for the stochastic system dynamics equations given in (7)

$$
\mathbb{E}\Big\{e_i^T[F_a(e_i(t)) + G_a(e_i(t))u_i^*(t) + \frac{\sqrt{2\nu}d\sigma_i}{dt}]\Big\} \leq \gamma\mathbb{E}\{\|e_i\|^2\} \tag{29}
$$

Theorem 4 (Closed-Loop Stability): Let the crtic, mass and actor NNs' weights can be updated as (26)- (28). Also, we assume that learning rates $\alpha_{V,i}$, $\alpha_{\rho,i}$ and $\alpha_{u,i}$ are greater than zero. Then, $\mathbb{E}\{\tilde{W}_{V,i}\}$, $\mathbb{E}\{\tilde{W}_{\rho,i}\}$, $\mathbb{E}\{\tilde{W}_{u,i}\}$ and $\mathbb{E}\{e_i\}$ are all UUB. Moreover, $\mathbb{E}\{\tilde{W}_{V,i}\}$, $\mathbb{E}\{\tilde{W}_{\rho,i}\}$, $\mathbb{E}\{\tilde{W}_{u,i}\}$ and $\mathbb{E}\{e_i\}$ are asymptotically stable if the NN's structures are selected perfectly.

Proof: See Appendix A

IV. Simulation Results

In this section, the developed BACM algorithm has been implemented into the large scale multi-UAV system with unstructured environment to address the decentralized mean field based optimal control problem. A total of 500 Unmanned Aerial Vehicles(UAVs) are deployed with system dynamics under an unstructured environment with multiple obstacles. The goal of each UAV is to reach a destination while avoiding multiple obstacles during the mission. Therefore, the movements of all UAVs are limited to a fixed area with specific boundary constraints. The initial positions of all UAVs are generated randomly following a normal distribution with mean 0.5 and variance 0.16. Then, the system of the UAVs are transformed to an unconstrained system by using the barrier function. The bounds of the barrier function varies with time depending on the boundary of the obstacles and configuration space. After transforming the unstructured space to a structured space, a reference trajectory has been given ahead of the mission planning to all the UAVs. The objective of each UAV is to track this reference trajectory in order to reach the goal position.

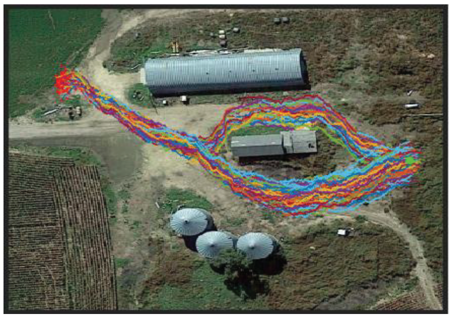

Fig. 3: Trajectory of Large number of UAVs with time while avoiding multiple obstacles

The agents intrinsic dynamics are given as:

$$f(x_{s,i}) = \begin{bmatrix} x_{s,2} - x_{s,1} \\ x_{s,2} - 2x_{s,1}^2 \end{bmatrix} \quad g(x_{s,i}) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Also, the dynamics of the transformed state are as follows:

$$F(s_i) = f(B_i^{-1}(s_i)) \frac{{b_u^i}^2 e^{-s_i} - 2b_l^i b_u^i + {b_l^i}^2 e^{s_i}}{b_u^i {b_l^i}^2 - b_l^i {b_u^i}^2}$$

$$G(s_i) = g(B_i^{-1}(s_i)) \frac{{b_u^i}^2 e^{-s_i} - 2b_l^i b_u^i + {b_l^i}^2 e^{s_i}}{b_u^i {b_l^i}^2 - b_l^i {b_u^i}^2}$$

The non-negative parameter ν is selected as 0.02. The mean field cost function is selected as $\Phi(e_i, \rho) = \|e_i - \mathbb{E}\{\rho(e_i)\}\|$, which represents the difference between agent i's current tracking error and the current average tracking error of the whole population. Moreover, $\rho(e) = 1$ denotes that the tracking error of all agents are same.

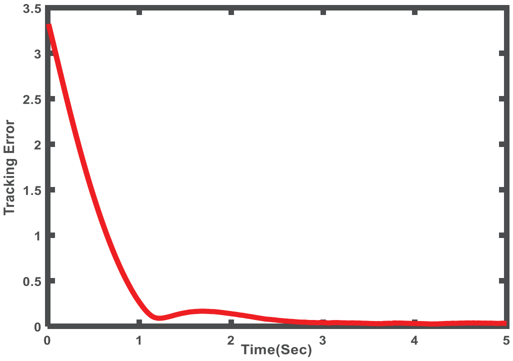

Fig. 4: Average tracking error of all UAVs

The coefficients to evaluate the cost of actions and tracking errors are selected as $R = 1$, $Q = 1$. The learning rate of the neural network are defined as $\alpha_{u,i} = 2 \times 10^{-4}$, $\alpha_{V,i} = 2 \times 10^{-6}$, and $\alpha_{\rho,i} = 1 \times 10^{-3}$. Also, the thresholds are defined as $\delta_u = 1 \times 10^{-3}$, $\delta_{FPK} = 1 \times 10^{-3}$, and $\delta_{HJB} = 1 \times 10^{-4}$.

Firstly, the overall performance of developed BACM based decentralized optimal tracking control is shown in Fig.3. It is clear that the developed algorithm can force all the UAVs to go to the destination while avoiding the obstacles in the environment.

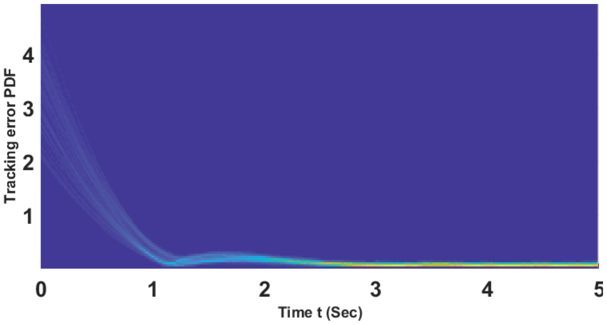

Fig. 5: Tracking error PDF of all UAVs

Secondly, the tracking errors of all UAVs has been analyzed. Fig.4 shows the average tracking error of all UAVs. This figure clearly shows that the tracking errors converge to near zero along with time. It indicates that the designed algorithm can track the reference trajectory to reach to a destination and avoid the obstacles in real time.

Also, Fig.5 shows the distribution of all UAVs tracking error, where the yellow color shows the tracking errors with higher probabilities. This figure clearly shows that the initial tracking error is high and randomly distributed. However, the variance of the tracking errors decrease to zero with time. This also proves that the system can track the reference trajectory successfully.

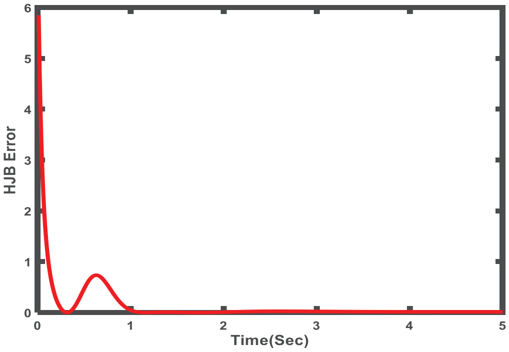

Fig. 6: HJB equation errors

Next, the neural networks performance is demonstrated by analyzing the HJB equation error along with the FPK equation error of UAVs. Without loss of generality, we consider the optimality for UAV 1. In Fig.6, the HJB equation error of the UAV 1 has been plotted. From this figure, it is clear that the error converge to near zero after 1s. Similarly, in Fig.7, the FPK equation error of the UAV 1 has been plotted to demonstrate the mean field error. From Fig.6 and Fig.7, it is clear that the HJB and mean field equation errors for agent 1 converge to near zero with time, which proves that the solution of the HJB-FPK coupled equation system is successfully approximated. Therefore the ε- Nash Equilibrium has been reached.

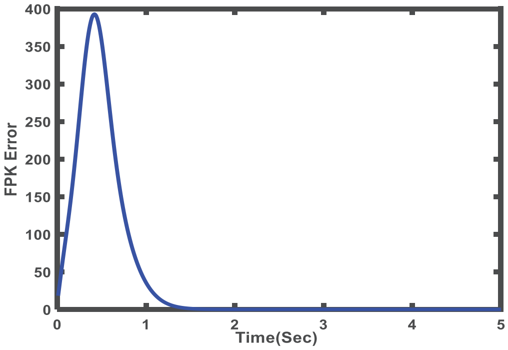

Fig. 7: FPK equation errors

V. Conclusions

In this paper, a novel mean field game based barrier-actor-critic-mass (BACM) learning algorithm has been developed to obtain the decentralized optimal tracking control for large scale multi-agent system (LS-MAS) under unstructured environment. In this BACM algorithm, a barrier function is used to tranform the unstructured LS-MAS state space to structured space. Then, decentralized optimal tracking control for LS-MAS under structured space can be obtained by solving the coupled HJB-FPK equations obtained from mean-field game theory. Specifically, to solve the barrier function based mean field game, three neural networks (NN) are employed, i.e., the actor NN for learning optimal control, the critic NN for estimating optimal cost function, and the mass NN for approximating the LS-MAS's probaiblity density function, i.e. Mass. Next, using the barrier function again, we can transform back and obtain the decentralized optimal tracking control for LS-MAS under unstructured environment. The effectiveness of the developed technique has been ensured through a closed-loop stability analysis and a series of numerical simulations.

Appendix A
Proof of Theorem 1

Consider the Lyapunov function as

$$\begin{aligned} L_{\text{sys}}(t) &= \frac{\beta_1}{2}\text{tr}\Big(\mathbb{E}\Big\{e_i^T(t)e_i(t)\Big\}\Big) + \frac{\beta_2}{2}\text{tr}\Big(\mathbb{E}\Big\{\tilde{W}_{V,i}^T(t) \\ &\times \tilde{W}_{V,i}(t)\Big\}\Big) + \frac{\beta_3}{2}\text{tr}\Big(\mathbb{E}\Big\{\{\tilde{W}_{\rho,i}^T(t)\tilde{W}_{\rho,i}(t)\Big\}\Big) + \frac{\beta_4}{2}\text{tr} \\ &\times \Big(\mathbb{E}\Big\{\{\tilde{W}_{u,i}^T(t)\tilde{W}_{u,i}(t)\Big\}\Big) \end{aligned} \tag{30}$$

Taking the first derivative and substituting Lemma 1

$$\begin{aligned} &\dot{L}_{\text{sys}}(t) = \beta_1\text{tr}\Big(\mathbb{E}\Big\{\{e_i^T(t)\dot{e}_i(t)\Big\}\Big) + \beta_2\text{tr}\Big(\mathbb{E}\Big\{\{\tilde{W}_{V,i}^T(t) \\ &\times \dot{\tilde{W}}_{V,i}(t)\Big\}\Big) + \beta_3\text{tr}\Big(\mathbb{E}\Big\{\{\tilde{W}_{\rho,i}^T(t)\dot{\tilde{W}}_{\rho,i}(t)\Big\}\Big) + \beta_4\text{tr} \\ &\times \Big(\mathbb{E}\Big\{\tilde{W}_{u,i}^T(t)\dot{\tilde{W}}_{u,i}(t)\Big\}\Big) \\ &\leq -\frac{\gamma\beta_1}{2}\mathbb{E}\{\|e_i\|^2\} - \kappa_{V,i}\mathbb{E}\{\|\tilde{W}_{V,i}\|^2\} - \kappa_{u,i}\mathbb{E}\{\|\tilde{W}_{u,i}\|^2\} \\ &- \kappa_{\rho,i}\mathbb{E}\Big\{\|\tilde{W}_{\rho,i}\|^2\Big\} + \varepsilon_{CS} \end{aligned} \tag{31}$$

The derivative of Lyapunov function $\dot{L}_{sys}(t)$ is less than zero outside a compact set, i.e.,

$$\mathbb{E}\Big\{\|e_i\|\Big\} > \sqrt{\frac{2}{\gamma\beta_1}\varepsilon_{CS}} \quad \text{or} \quad \mathbb{E}\Big\{\|\tilde{W}_{V,i}\|\Big\} > \sqrt{\frac{1}{\kappa_{V,i}}\varepsilon_{CS}}$$

or

$$\mathbb{E}\Big\{\|\tilde{W}_{u,i}\|\Big\} > \sqrt{\frac{1}{\kappa_{u,i}}\varepsilon_{CS}} \quad \text{or} \quad \mathbb{E}\Big\{\|\tilde{W}_{\rho,i}\|\Big\} > \sqrt{\frac{1}{\kappa_{\rho,i}}\varepsilon_{CS}}$$

with,

$$\begin{aligned} b_1 &= -\frac{\gamma\beta_1}{2}\mathbb{E}\Big\{\|e_i\|^2\Big\} + \frac{2\beta_1 g_l^2}{\gamma}\mathbb{E}\|\tilde{u}_i\|^2 - \frac{\beta_2\alpha_{V,i}}{4} \\ &\times \mathbb{E}\Big\{\frac{\|\hat{\Psi}_{V,i}\|^2\|\tilde{W}_{V,i}\|^2}{1+\|\hat{\Psi}_{V,i}\|^2}\Big\} \\ b_2 &= \alpha_{V,i}\beta_2\frac{[l_{\Phi,i} + l_{\Psi_{V,i}}\mathbb{E}\{\|W_{V,i}\|^2\}]}{1+\mathbb{E}\{\|\hat{\Psi}_{V,i}\|^2\}} \\ b_3 &= -\frac{\beta_3\alpha_{\rho,i}}{2}\mathbb{E}\Big\{\frac{\|\hat{\Psi}_{\rho,i}\|^2\|\tilde{W}_{\rho,i}\|^2}{1+\|\hat{\Psi}_{\rho,i}\|^2}\Big\} - \frac{\beta_4\alpha_{u,i}}{4} \\ &\times \mathbb{E}\Big\{\frac{\|\hat{\phi}_{u,i}\|^2\|\tilde{W}_{u,i}\|^2}{1+\|\hat{\phi}_{u,i}\|^2}\Big\} + \beta_2\alpha_{V,i}\mathbb{E}\Big\{\frac{\|\varepsilon_{\text{HJBi}}\|^2}{1+\|\hat{\Psi}_{V,i}\|^2}\Big\} \\ &+ \beta_3\alpha_{\rho,i}\mathbb{E}\Big\{\frac{\|\varepsilon_{\text{FPKi}}\|^2}{1+\|\hat{\Psi}_{\rho,i}\|^2}\Big\} + \beta_4\alpha_{u,i}\mathbb{E}\Big\{\frac{\|\varepsilon_{\text{u,i}}\|^2}{1+\|\hat{\phi}_{u,i}\|^2}\Big\} \\ b_4 &= \beta_3\alpha_{\rho,i}\frac{[l_{\Psi_{\rho,i}}\mathbb{E}\{\|W_{\rho,i}\|^2\}]}{1+\mathbb{E}\{\|\hat{\Psi}_{\rho,i}\|^2\}} + \beta_4\alpha_{u,i} \\ &\times \mathbb{E}\Big\{\frac{\|R^{-1}G_a^T\|^2}{1+\|\hat{\phi}_{u,i}\|^2}\Big\} \end{aligned} \tag{32}$$

$$\begin{aligned} &\kappa_{V,i} = -\frac{\beta_2\alpha_{V,i}}{4}\mathbb{E}\Big\{\frac{\|\hat{\Psi}_{V,i}\|^2\|\tilde{W}_{V,i}\|^2}{1+\|\hat{\Psi}_{V,i}\|^2}\Big\} - 3b_4\mathbb{E}\{\|\hat{\Psi}_{V,i}\|^2\} \\ &\kappa_{u,i} = \frac{\beta_4\alpha_{u,i}}{4}\mathbb{E}\Big\{\frac{\|\hat{\phi}_{u,i}\|^2\|\tilde{W}_{u,i}\|^2}{1+\|\hat{\phi}_{u,i}\|^2}\Big\} - \frac{6\beta_1 g_l^2}{\gamma}\mathbb{E}\{\|\hat{\phi}_{u,i}\|^2\} \\ &\kappa_{\rho,i} = \frac{\beta_3\alpha_{\rho,i}}{2}\mathbb{E}\Big\{\frac{\|\hat{\Psi}_{\rho,i}\|^2\|\tilde{W}_{\rho,i}\|^2}{1+\|\hat{\Psi}_{\rho,i}\|^2}\Big\} - \frac{6\beta_1 g_l^2}{\gamma}l_{\phi_{u,i}}^2 \\ &\mathbb{E}\{\|W_{u,i}\|^2\|\hat{\Psi}_{\rho,i}\|^2\} - 6b_4 l_{\Psi_{V,i}}^2\mathbb{E}\{\|W_{V,i}\|^2\|\hat{\Psi}_{\rho,i}\|^2\} \\ &- 2b_2\mathbb{E}\{\|\hat{\Psi}_{\rho,i}\|^2\} \\ &\varepsilon_{CS} = \frac{6\beta_1 g_l^2}{\gamma}l_{\phi_{u,i}}^2\mathbb{E}\{\|W_{u,i}\|^2\}\mathbb{E}\{\|\varepsilon_{\text{FPKi}}\|^2\} + \frac{6\beta_1 g_l^2}{\gamma} \end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}\big\{\|\varepsilon_{\text{u,i}}\|^2\big\} + \beta_2\varepsilon_{\text{NHJBi}} + \beta_3\varepsilon_{\text{NFPKi}} + \beta_4\varepsilon_{\text{Nu,i}} + 6b_4 l^2_{\Psi_{V,i}} \\
&\mathbb{E}\big\{\|W_{V,i}\|^2\big\}\mathbb{E}\big\{\|\varepsilon_{\text{FPKi}}\|^2\big\} + 2b_2\mathbb{E}\big\{\|\varepsilon_{\text{FPKi}}\|^2\big\} \\
&3b_4\mathbb{E}\big\{\|\varepsilon_{\text{HJBi}}\|^2\big\}
\end{aligned}
\tag{33}
$$

and,

$$
\begin{aligned}
\varepsilon_{\text{NHJBi}} &= \alpha_{V,i}\frac{\|\varepsilon_{\text{HJB}_i}\|^2}{1+\|\hat{\Psi}_{V,i}\|^2} \\
\varepsilon_{\text{Nu,i}} &= \alpha_{u,i}\frac{\|\varepsilon_{\text{u,i}}\|^2}{1+\|\hat{\phi}_{u,i}\|^2} \\
\varepsilon_{\text{NFPKi}} &= \alpha_{\rho,i}\frac{\|\varepsilon_{\text{FPKi}}\|^2}{1+\|\hat{\Psi}_{\rho,i}\|^2}
\end{aligned}
$$

where, $l_{\Psi_{\rho,i}}$, $l_{\Psi_{V,i}}$ and $l_{\phi_{u,i}}$ are the Lipschitz constants and g_l is the upper bound of the function $G_a(e_i)$.

References

[1] A. Dorri, S. S. Kanhere, and R. Jurdak, "Multi-agent systems: A survey," *IEEE Access*, vol. 6, pp. 28 573–28 593, 2018.

[2] J.-W. Zhu, Y.-P. Yang, W.-A. Zhang, L. Yu, and X. Wang, "Cooperative attack tolerant tracking control for multi-agent system with a resilient switching scheme," *Neurocomputing*, vol. 409, pp. 372–380, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220310754

[3] P. Gong and W. Lan, "Adaptive robust tracking control for uncertain nonlinear fractional-order multi-agent systems with directed topologies," *Automatica*, vol. 92, pp. 92–99, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005109818300700

[4] Z. Zhou and H. Xu, "Decentralized adaptive optimal tracking control for massive multi-agent systems: An actor-critic-mass algorithm," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 1231–1236.

[5] T. Feng, H. Zhang, Y. Luo, and H. Liang, "Globally optimal distributed cooperative control for general linear multi-agent systems," *Neurocomputing*, vol. 203, pp. 12–21, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231216302302

[6] W. Liu, J. Liu, J. Peng, and Z. Zhu, "Cooperative multi-agent traffic signal control system using fast gradient-descent function approximation for v2i networks," in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 2562–2567.

[7] D. Simões, N. Lau, and L. Paulo Reis, "Multi-agent actor centralized-critic with communication," *Neurocomputing*, vol. 390, pp. 40–56, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220301314

[8] P. Cannarsa, R. Capuani, and P. Cardaliaguet, "Mean Field Games with state constraints: from mild to pointwise solutions of the PDE system," *Calculus of Variations and Partial Differential Equations*, vol. 60(3), pp. 1–33, 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01964755

[9] A. R. Mészáros and F. J. Silva, "A variational approach to second order mean field games with density constraints: The stationary case," *Journal de Mathématiques Pures et Appliquées*, vol. 104, no. 6, pp. 1135–1159, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021782415000926

[10] Y. Yang, D.-W. Ding, H. Xiong, Y. Yin, and D. C. Wunsch, "Online barrier-actor-critic learning for h control with full-state constraints and input saturation," *Journal of the Franklin Institute*, vol. 357, no. 6, pp. 3316–3344, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016003219309044

[11] Z. Zhou and H. Xu, "Large-scale multiagent system tracking control using mean field games," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–9, 2021.

[12] K. G. Vamvoudakis and F. L. Lewis, "Online actor critic algorithm to solve the continuous-time infinite horizon optimal control problem," *2009 International Joint Conference on Neural Networks*, pp. 3180–3187, 2009.