

# Scalable Task-Driven Robotic Swarm Control via Collision Avoidance and Learning Mean-Field Control

Kai Cui, Mengguang Li, Christian Fabian and Heinz Koepl

**Abstract**—In recent years, reinforcement learning and its multi-agent analogue have achieved great success in solving various complex control problems. However, multi-agent reinforcement learning remains challenging both in its theoretical analysis and empirical design of algorithms, especially for large swarms of embodied robotic agents where a definitive toolchain remains part of active research. We use emerging state-of-the-art mean-field control techniques in order to convert many-agent swarm control into more classical single-agent control of distributions. This allows profiting from advances in single-agent reinforcement learning at the cost of assuming weak interaction between agents. However, the mean-field model is violated by the nature of real systems with embodied, physically colliding agents. Thus, we combine collision avoidance and learning of mean-field control into a unified framework for tractably designing intelligent robotic swarm behavior. On the theoretical side, we provide novel approximation guarantees for general mean-field control both in continuous spaces and with collision avoidance. On the practical side, we show that our approach outperforms multi-agent reinforcement learning and allows for decentralized open-loop application while avoiding collisions, both in simulation and real UAV swarms. Overall, we propose a framework for the design of swarm behavior that is both mathematically well-founded and practically useful, enabling the solution of otherwise intractable swarm problems.

## I. INTRODUCTION

Over the past decades, the field of swarm robotics [1]–[3] has received considerable attention [4]. Various areas of potential applications include for example industrial inspection tasks [5], such as for turbines, cooperative object transport [6]–[8], agriculture [9], aerial combat [10], and cooperative search [11]. A recent promising approach for engineering many-agent systems such as intelligent robot swarms is multi-agent reinforcement learning (MRL) [12], which has found success in diverse complex problems such as strategic video games [13], communication networks [14] or traffic control [15]. However, MRL algorithms suffer from issues such as credit assignment, non-stationarity and scalability to many agents [12]. Meanwhile, robotic swarms such as fleets of unmanned aerial vehicles (UAVs) usually consist of many

This work has been co-funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center, the State of Hesse and HOLM as part of the "Innovations in Logistics and Mobility" programme of the Hessian Ministry of Economics, Energy, Transport and Housing (HA project no.: 1010/21-12), and the Hessian Ministry of Science and the Arts (HMWK) within the projects "The Third Wave of Artificial Intelligence - 3AI" and hessian.AI. The authors acknowledge the Lichtenberg high performance computing cluster of the TU Darmstadt for providing computational facilities for the calculations of this research.

The authors are with the Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, 64287 Darmstadt, Germany. (e-mail: {kai.cui, mengguang.li, christian.fabian, heinz.koepl}@tu-darmstadt.de).

interacting UAVs and remain of considerable interest due to their inherent robustness, scalability to large-scale deployment and decentralization, which can be considered the ultimate goal of the study of swarm intelligence and robotics [1], [16]. Here, scalable control approaches and highly general toolchains for swarm robotics remain to be established [2].

A classical approach to formulate systems with large numbers of agents with low complexity is via mean-field models, describing swarms of drones by their distribution, see also [17] and [18] for reviews on mean-field swarm robotics and mean-field control (MFC). However, most prior literature is based on analytic derivations and continuous-time models, which are less conducive to advances in MARL. For example, stabilizing control of swarms to distributions are designed in [19]–[21]. Other works such as [22], [23] consider population density estimates via collisions for task allocation problems, while [24] study robots for stick-pulling. Lastly, a variety of approaches use PDE-based formulations, e.g. [25], [26] for density control, or [27], [28] for general analytic frameworks, though they are significantly more difficult to treat both rigorously and from a learning perspective. Especially mean-field-based learning algorithms often remain restricted to competitive settings such as mean-field games [29], [30] by learning e.g. Nash [31]–[35], regularized [36], [37] or correlated equilibria [38], [39]. For instance, works such as [40] or [41] investigate trajectory control of selfish UAV agents, while [42] considers formation flight in dense environments. Although selfish control problems are interesting for many applications, aligning selfish or local cost functions with a certain cooperative, global behavior can be difficult [43]. Solutions for cooperative joint objectives without necessity of manual cost function tuning are therefore of practical interest for artificially engineering swarm behaviors.

In this work, we propose a discrete-time MFC-based swarm robotics framework that is conducive to powerful deep reinforcement learning (RL) techniques. Only very recently were MFC [44]–[46] and related histogram observations for MARL [47] proposed as a potential solution to cooperative scalable MARL, which could enable both the solution of otherwise intractable tasks as well as model-free application to swarms, adapting to environments and tasks. However, an eminent issue of MFC for robotic systems is violation of the MFC model due to physical collisions between robots. To solve this issue, we combine MFC with deep RL and collision avoidance algorithms. Here, collision avoidance algorithms could range from classical rule-based [48] over planning-based [49] to learning-based approaches [50], [51], and similarly for RL, see e.g. [52]. Importantly, our approach

(i) is able to utilize advances in RL, circumventing MARL and solving otherwise difficult swarm problems without extensive manual and analytical design of algorithms, and (ii) closes the gap between mean-field models and reality, as collisions between agents violate the weak interaction principle of mean-field models and are usually to be avoided, e.g. in UAVs. As a result, our approach is highly practical, with the advantage of automatic design of swarm algorithms for swarm problems.

Our contribution can be summarized as follows: (i) We combine RL with MFC and collision avoidance algorithms for general task-driven control of robotic swarms; (ii) We give novel theoretical approximation guarantees of MFC in finite swarms as well as in the presence of additional collision avoidance maneuvers; (iii) We demonstrate in a variety of tasks that MFC outperforms state-of-the-art MARL, can be applied in a decentralized open-loop manner and avoids collisions, both in simulation and real UAV swarms. Overall, we provide a general framework for tractable swarm control that could be applied directly to swarms of UAVs.

## II. SWARM MODEL

In order to tractably describe a plethora of swarm tasks, we formulate a mean-field model where all agents are anonymous and it is sufficient to consider their distribution.

### A. Finite swarm model

Formally, we consider compact state and action spaces  $\mathcal{X}, \mathcal{U} \subseteq \mathbb{R}^2$  (though our results are easily extended to  $\mathbb{R}^3$ ) representing possible locations and movement choices of an agent. For any  $N \in \mathbb{N}$ , at each time  $t = 0, 1, \dots$ , the states and actions of agent  $i = 1, \dots, N$  are denoted by  $x_t^{i,N}$  and  $u_t^{i,N}$ . We denote by  $\mathcal{P}(\mathcal{X})$  the space of probability measures on  $\mathcal{X}$ , equipped with the topology of weak convergence. Define the empirical state distribution  $\mu_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{x_t^{i,N}} \in \mathcal{P}(\mathcal{X})$ , which represents all agents anonymously by their states. We consider policies  $\pi = \{\pi_t\}_{t \geq 0} \in \Pi$  from a space of policies  $\Pi$  with shared Lipschitz constant, such that agents act on their location and the distribution of all agents,  $\pi_t: \mathcal{X} \times \mathcal{P}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{U})$ . The assumption of Lipschitz continuity is standard in the literature, includes e.g. neural networks [46], [53], [54], and may allow approximation of less regular policies.

Under a policy  $\pi \in \Pi$ , the finite swarm system shall evolve by sampling an initial state  $x_0^{i,N} \sim \mu_0$  from an initial distribution  $\mu_0$  of agents, and subsequently taking movement actions  $u_t^{i,N} \sim \pi_t(x_t^{i,N}, \mu_t^N)$ , resulting in new states  $x_{t+1}^{i,N} = x_t^{i,N} + u_t^{i,N} + \epsilon_t^i$  for all agents  $i$  with optional i.i.d. Gaussian noise  $\epsilon_t^i \sim \mathcal{N}(0, \Sigma)$  and diagonal covariance matrix  $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ . In other words, each drone can move a distance limited to  $\mathcal{U}$ , up to some smoothing or inaccuracy  $\epsilon_t^i$ . In simulation, we further clip agent positions to stay inside  $\mathcal{X}$ . The objective is then given by an arbitrary function  $r: \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}$  of the spatial distribution of agents, giving rise to the infinite-horizon discounted objective

$$J^N(\pi) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(\mu_t^N) \right]. \quad (1)$$

Since MARL can be difficult in the presence of many agents (see e.g. combinatorial nature in [12]), we will formulate and verify a limiting infinite-agent system.

### B. Mean-field swarm model

In the limit as  $N \rightarrow \infty$ , single agents become indiscernible and we need only model their distribution (mean-field)  $\mu_t \in \mathcal{P}(\mathcal{X})$ . Starting at  $\mu_0$ , under policy  $\pi \in \Pi$ , deterministically

$$\begin{aligned} \mu_{t+1} &= T^{\pi_t}(\mu_t) \equiv T^{\pi_t(\cdot|\cdot, \mu_t)}(\mu_t) \\ &:= \int \mathcal{N}(x+u, \sigma^2) \pi_t(du | x, \mu_t) \mu_t(dx) \end{aligned} \quad (2)$$

with shorthand of the deterministic mean-field transition operator  $T^{\pi_t}(\mu_t) \equiv T^{\pi_t(\cdot|\cdot, \mu_t)}(\mu_t)$ ,  $\pi_t(\cdot | \cdot, \mu_t) \in \mathcal{P}(\mathcal{U})^{\mathcal{X}}$ , giving way to the MFC problem with objective function

$$J(\pi) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(\mu_t) \right]. \quad (3)$$

*Remark 1:* A dependence of  $r$  on joint state-action distributions in  $\mathcal{P}(\mathcal{X} \times \mathcal{U})$  can be modelled by splitting time steps into two and using the new state space  $\mathcal{X} \cup (\mathcal{X} \times \mathcal{U})$ .

For simplicity of analysis, we assume absence of common noise, leading to a deterministic mean-field limit, though in our experiments we also allow reactions to a random external environment. Under a mild continuity assumption, weaker than the common Lipschitz assumption in existing literature [46], [54], we obtain rigorous approximation guarantees.

*Assumption 1:* The reward function  $r$  is continuous.

By compactness of  $\mathcal{P}(\mathcal{X})$ ,  $r$  is bounded. As long as  $r$  is continuous, i.e. small changes in the agent distribution lead to small changes in reward, the MFC model is a good approximation for large swarms and its solution solves the finite agent system approximately optimally. As existing approximation properties still remain limited to finite  $\mathcal{X}$ ,  $\mathcal{U}$  [45], [46], we give a brief, novel proof for compact spaces.

*Theorem 1:* Under Assumption 1, at all times  $t \in \mathcal{T}$ , the empirical reward  $r(\mu_t^N)$  converges weakly and uniformly to the limiting reward  $r(\mu_t)$  as  $N \rightarrow \infty$ , i.e.

$$\sup_{\pi \in \Pi} \mathbb{E} [ |r(\mu_t^N) - r(\mu_t)| ] \rightarrow 0. \quad (4)$$

*Proof:* We can metrize  $\mathcal{P}(\mathcal{X})$  via the metric  $d(\mu, \nu) := \sum_{m=1}^{\infty} 2^{-m} |\mu(f_m) - \nu(f_m)|$  for a sequence of continuous and bounded  $f_m: \mathcal{X} \rightarrow \mathbb{R}$ ,  $|f_m| \leq 1$  (cf. [55, Theorem 6.6]).

Consider any (uniformly) equicontinuous set  $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{P}(\mathcal{X})}$  of functions, i.e. there exists an increasing (concave, cf. [56, p. 41])  $\omega_{\mathcal{F}}: [0, \infty) \rightarrow [0, \infty)$  (modulus of continuity) such that  $\omega_{\mathcal{F}}(x) \rightarrow 0$  when  $x \rightarrow 0$  and  $|f(\mu) - f(\nu)| \leq \omega_{\mathcal{F}}(d(\mu, \nu))$  for all  $f \in \mathcal{F}$ . We show inductively for  $t \geq 0$  that

$$\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [ |f(\mu_t^N) - f(\mu_t)| ] \rightarrow 0, \quad (5)$$

which implies the desired property, since  $r$  is uniformly continuous by compactness of  $\mathcal{P}(\mathcal{X})$  and Assumption 1.

At time  $t = 0$ , the proof follows from the weak law of large numbers (LLN) argument (see (7) and below). For the induction step,

$$\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [ |f(\mu_{t+1}^N) - f(\mu_{t+1})| ] \quad (6)$$

$$\leq \sup_{\pi \in \Pi} \mathbb{E} [\omega_{\mathcal{F}}(d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N)))] \quad (7)$$

$$+ \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [|f(T^{\pi_t}(\mu_t^N)) - f(\mu_{t+1})|] \quad (8)$$

where for the first term (7), by Jensen's inequality we obtain

$$\mathbb{E} [\omega_{\mathcal{F}}(d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N)))] \leq \omega_{\mathcal{F}}(\mathbb{E} [d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N))])$$

for concave  $\omega_{\mathcal{F}}$ . Abbreviating  $x_t^N \equiv \{x_t^{i,N}\}_{i \in [N]}$ , we have

$$\begin{aligned} & \mathbb{E} [d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N))] \\ &= \sum_{m=1}^{\infty} 2^{-m} \mathbb{E} [|\mu_{t+1}^N(f_m) - T^{\pi_t}(\mu_t^N)(f_m)|] \\ &\leq \sup_{m \geq 1} \mathbb{E} [\mathbb{E} [|\mu_{t+1}^N(f_m) - T^{\pi_t}(\mu_t^N)(f_m)| \mid x_t^N]], \end{aligned}$$

where by the weak LLN argument, the squared term

$$\begin{aligned} & \mathbb{E} [|\mu_{t+1}^N(f_m) - T^{\pi_t}(\mu_t^N)(f_m)| \mid x_t^N]^2 \\ &\leq \mathbb{E} \left[ \left| \frac{1}{N} \sum_{i=1}^N (f_m(x_{t+1}^{i,N}) - \mathbb{E} [f_m(x_{t+1}^{i,N}) \mid x_t^N]) \right|^2 \mid x_t^N \right] \\ &= \frac{1}{N^2} \sum_{i=1}^N \mathbb{E} \left[ (f_m(x_{t+1}^{i,N}) - \mathbb{E} [f_m(x_{t+1}^{i,N}) \mid x_t^N])^2 \mid x_t^N \right] \\ &\leq \frac{4}{N} \rightarrow 0 \end{aligned}$$

since for any  $f_m$ , the cross-terms are zero and  $|f_m| \leq 1$ .

For the second term (8), by induction assumption we have

$$\begin{aligned} & \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [|f(T^{\pi_t}(\mu_t^N)) - f(\mu_{t+1})|] \\ &\leq \sup_{\pi \in \Pi} \sup_{g \in \mathcal{G}} \mathbb{E} [|g(\mu_t^N) - g(\mu_t)|] \rightarrow 0 \end{aligned}$$

using  $g = f \circ T^{\pi_t}$  and the corresponding class  $\mathcal{G}$  of functions with modulus of continuity  $\omega_{\mathcal{G}} := \omega_{\mathcal{F}} \circ \omega_T$ , where  $\omega_T$  denotes the uniform modulus of continuity of  $T^{\pi_t}$  by uniform Lipschitz continuity of  $\pi \in \Pi$ . ■

As a result, the MFC approach is a theoretically rigorous approach to approximately optimally solving large-scale swarm problems with complexity independent of  $N$ .

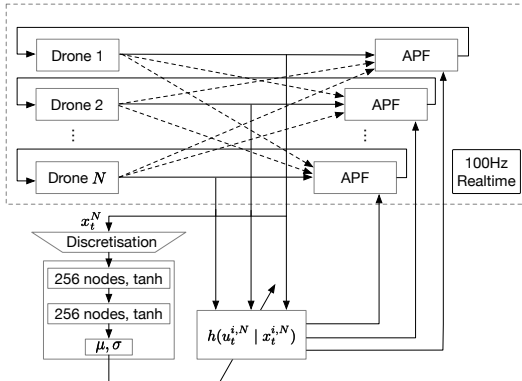


Fig. 1. A hierarchical overview of our approach. The learned high-level mean-field control policy sends movement instructions to the (UAV) swarm, while each agent uses a real-time collision avoidance algorithm – here artificial potential fields (APF) – to avoid collisions with others.

TABLE I

HYPERPARAMETER CONFIGURATIONS FOR PPO.

Symbol	Name	Value
$\gamma$	Discount factor	0.99
$\lambda$	GAE lambda	1
$\beta$	KL coefficient	0.03
$\epsilon$	Clip parameter	0.2
$l_r$	Learning rate	0.00005
$B_b$	Training batch size	4000
$B_m$	Minibatch size	1000
$T_b$	Updates per training batch	5

*Corollary 1:* Under Assumption 1, an optimal solution  $\pi^* \in \Pi$  to the MFC problem constitutes an  $\epsilon$ -optimal solution to the finite swarm problem, where  $\epsilon \rightarrow 0$  as  $N \rightarrow \infty$ .

*Proof:* For any  $\pi \in \Pi$  and  $\epsilon > 0$ , we can choose  $T$  such that  $\sum_{t=T+1}^{\infty} \gamma^t \mathbb{E} [|r(\mu_t^N) - r(\mu_t)|] \leq 2^{-T} \max_{\mu} 2|r(\mu)| < \frac{\epsilon}{4}$ , and for sufficiently large  $N$   $\sum_{t=0}^T \gamma^t \mathbb{E} [|r(\mu_t^N) - r(\mu_t)|] < \frac{\epsilon}{4}$  by Theorem 1. Therefore, we have  $J^N(\pi^*) - \max_{\pi \in \Pi} J^N(\pi) = \min_{\pi \in \Pi} (J^N(\pi^*) - J^N(\pi)) \geq \min_{\pi \in \Pi} (J^N(\pi^*) - J(\pi^*)) + \min_{\pi \in \Pi} (J(\pi^*) - J(\pi)) \geq -\frac{\epsilon}{2} + 0 - \frac{\epsilon}{2} = -\epsilon$  by the prequel and optimality of  $\pi^*$  in the MFC problem. ■

### III. METHODOLOGY

In order to remove the two remaining obstacles of (i) solving the MFC problem, and (ii) resolving the real-world gap of MFC for embodied agents, we combine MFC with arbitrary powerful RL and collision avoidance techniques. The overall hierarchical structure is found in Fig. 1. The MFC solution is learned via RL and gives high-level directions, which are realized by each agent while avoiding collisions.

#### A. Reinforcement learning

For the MFC problem, it is known that there exists an optimal stationary solution [44, Theorem 19], which may be found by solving the MFC Markov decision problem (MDP), a single-agent but infinite-dimensional RL problem with  $\mathcal{P}(\mathcal{X})$ -valued states  $\mu_t$  and  $\mathcal{P}(\mathcal{U})^{\mathcal{X}}$ -valued actions  $h_t$  evolving according to  $\mu_{t+1} = T^h(\mu_t)$ . To deal with the infinite dimensionality of  $\mathcal{P}(\mathcal{X})$  and  $\mathcal{P}(\mathcal{U})^{\mathcal{X}}$ , we discretize  $\mathcal{X}$  and use a binned histogram of  $\mathcal{P}(\mathcal{X})$  as in [44] with  $M = 6^2 = 36$  bins by trading off between tractability (good training, low  $M$ ) and performance (high  $M$ ), while  $\mathcal{P}(\mathcal{U})$  is parametrized by Gaussians with means  $\theta \in \mathcal{U}$  and diagonal covariances  $\sigma_1, \sigma_2 \in (0, 0.25]$ , of which the samples  $u_t^i \sim h(\cdot \mid x_t^i) = \mathcal{N}(\theta, \text{diag}(\sigma_1, \sigma_2))$  are clipped to  $\mathcal{U}$ . As exact computation of  $\mu_t$  is difficult, we use the finite system with  $N = 300$  agents (though less works fine) and their empirical distribution analogous to particle filtering, which can be understood as directly learning on a large finite swarm.

We use the RLlib 1.13.0 implementation [57] of proximal policy optimization (PPO) RL [58] and a diagonal Gaussian neural network policy with two hidden tanh-layers of 256 nodes, sampling clipped values in  $[-1, 1]$  affinely transformed to  $\theta, \sigma_1, \sigma_2$ . Hyperparameters are printed in Table I, of which sufficiently high minibatch sizes appeared most important.

### B. Collision avoidance subroutine

A solution of the mean-field system does not directly translate into applicable real-world behavior, since the mean-field solution ignores physical constraints. While e.g. UAVs could fly at different heights, a general swarm algorithm should explicitly avoid collisions in order to guarantee suitability of the weakly-interacting MFC model. This is done by separating concerns, decomposing the issue into MFC plus sequences of collision-avoiding navigation subproblems between decision epochs. For example, we could choose  $\mathcal{U}$  slightly smaller than the maximum speed range to allow for additional avoidance maneuvers. Then, assuming the time  $\Delta t$  between two MFC decisions  $t$  and  $t + 1$  is sufficiently long, and that agents have finer, direct control over their positions, a collision-avoiding navigation subroutine could approximately achieve the desired positions up to an error that becomes arbitrarily small with agent radius  $r$ .

For  $N$  drones and agent radius  $r$  we hence assume existence of such a subroutine  $F$  which mildly perturbs all positions and their distribution  $\mu_t^N$  at each time step and thereby achieves a collision-free mean field, which we write as  $F(\mu_t^N)$ , such that  $\|x_i^i - x_j^j\|_2 > 2r$  for all  $i, j$ . We further assume that  $F$  is near-optimal, i.e. each drone's position is perturbed at most by a distance of  $4Nr$ . Indeed, this is possible for sufficiently small  $r$ , e.g. if  $\mathcal{X} = [-m, m]^2$  for  $m > 0$ : At any  $x \in \mathcal{X}$ , on an arbitrary line of length greater  $2m$  passing through  $x$ , we can always choose a position that is at most  $4Nr$  away from  $x$ , as in the worst case all other  $N - 1$  drones are located on the line along which  $F$  moves the drone and have a distance of slightly less than  $4r$  between each other. Under  $F$ , we can show that a collision-avoiding finite swarm of sufficiently many small agents is solved well by our approach.

*Theorem 2:* Let  $\pi \in \Pi$  be an optimal solution to the MFC problem, and let  $F$  be the near-optimal collision avoidance subroutine as defined above. Then for each  $\varepsilon > 0$  there exists an  $N'$  such that for all  $N \geq N'$  and agent radii  $r_{N,\varepsilon}$ , the solution  $\pi$  gives an  $\varepsilon$ -optimal solution to the finite swarm problem with collision avoidance.

*Proof:* The definition of  $F$  allows us to define new model dynamics with new random mean field variables denoted by  $\mu_t^N$ , where we leave out the definition of each agent variable for brevity. For the new dynamics, at each time step  $t$  we apply function  $F$  to the current mean field  $\mu_t^N$  and underlying positions. Subsequently, the mean field  $\mu_{t+1}^N$  is obtained by applying the usual transition dynamics.

Now, we show via induction over  $t$  that for all  $t$ ,

$$\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [|f(\mu_t^N) - f(F(\mu_t^N))|] \rightarrow 0. \quad (9)$$

Analogous to the proof of Theorem 1, the induction start follows from a weak LLN argument. For the induction step,

$$\begin{aligned} & \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [|f(\mu_{t+1}^N) - f(F(\mu_{t+1}^N))|] \\ & \leq \sup_{\pi \in \Pi} \mathbb{E} [\omega_{\mathcal{F}}(d(\mu_{t+1}^N, T^{\pi t}(\mu_t^N)))] \\ & + \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [|f(T^{\pi t}(\mu_t^N)) - f(T^{\pi t}(F(\mu_t^N)))|] \end{aligned}$$

$$\begin{aligned} & + \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [|f(T^{\pi t}(F(\mu_t^N))) - f(\mu_{t+1}^N)|] \\ & + \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E} [|f(\mu_{t+1}^N) - f(F(\mu_{t+1}^N))|] \end{aligned} \quad (10)$$

where the first two summands converge to zero by arguments as in the proof of Theorem 1. The third term converges to zero by a weak LLN argument while the fourth summand is bounded by  $\omega_{\mathcal{F}}(4Nr_{N,\varepsilon})$ , see the explanation above. By choosing  $r_{N,\varepsilon} = o(1/N)$ , the last summand in (10) converges to 0. This concludes the induction.

For  $\varepsilon$ -optimality, we proceed as in Corollary 1 and obtain

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma^t |r(\mu_t) - r(F(\mu_t^N))| \right] < \frac{\varepsilon}{2}$$

for  $N$  large enough by applying statement (9). The terms beyond  $T - 1$  can be bounded by  $\varepsilon/2$  as in Corollary 1. ■

Hence, for a given allowed sub-optimality specification  $\varepsilon$ , we can find a number  $N$  and size  $r$  of drones such that solving the MFC problem is  $\varepsilon$ -optimal in the finite swarm system. In practice, this means that if we can use sufficiently many sufficiently small drones, MFC provides good solutions.

In this work, for simplicity we use artificial potential fields (APF) as in [59] with attractive velocity  $F_d = 1.5(\hat{x}_t^i - x_t^i)$  in simulation, where  $\hat{x}_t^i$  denotes the MFC-based target position, and similarly repulsive velocity from agent  $j$  on agent  $i$ ,  $F_{ji} = 1.5c_{\text{rep}} \cdot \left( \frac{1}{\|x_t^i - x_t^j\|_2} - 1 \right) \cdot \frac{x_t^i - x_t^j}{\|x_t^i - x_t^j\|_2^3}$  whenever  $\|x_t^i - x_t^j\|_2 \leq 1$  and zero otherwise, where  $c_{\text{rep}} > 0$  is a variable repulsion coefficient. However, we stress that other more advanced collision avoidance algorithms could be used.

## IV. EXPERIMENTS

In this section, we verify the usefulness of MFC-based robotic swarm control experimentally.

### A. Problems

We consider three problems of increasing complexity to demonstrate our approach. In the following, we consider uniform initial state distributions  $\mu_0 = \text{Unif}(\mathcal{X})$  and let  $\mathcal{X} = [-2, 2]^2$ , allowing circular-constrained, noise-free movement, i.e. circular  $\mathcal{U}$  such that  $\|u_t^i\|_2 \leq 0.2$  with  $\epsilon_t^i \equiv 0$ .

*a) Aggregation:* In the simple Aggregation or Rendezvous [47] problem, the goal of agents is to aggregate into a point while minimizing movement. Hence, we choose rewards  $r(\nu_t) = \iint -\|x - \int x \nu_t(dx, du)\|_2 - 0.3\|u\|_2 \nu_t(dx, du)$  for joint state-actions  $\nu_t = \mu_t \otimes h_t \in \mathcal{P}(\mathcal{X} \times \mathcal{U})$  (see Remark 1).

*b) Formation:* In the Formation problem, the goal is to achieve an anonymous formation flight of large swarms, i.e. matching the distribution of agent positions with a given distribution – e.g. for providing coverage for surveillance or communication. The rewards are given by the Wasserstein distance  $r(\mu_t) = \inf_{X,Y: \mathcal{L}(X)=\mu_t, \mathcal{L}(Y)=\mu^*} \mathbb{E} [\|X - Y\|_2]$  [60] between agent distribution  $\mu_t$  and e.g. a Gaussian mixture  $\mu^* = \frac{1}{2}\mathcal{N}(e_1, \text{diag}(0.05, 0.05)) + \frac{1}{2}\mathcal{N}(-e_1, \text{diag}(0.05, 0.05))$  with unit vector  $e_1$ , computed via the empirical Wasserstein distance between agents and 300 samples of  $\mu^*$ . In principle, it is also possible to add movement costs as in Aggregation.

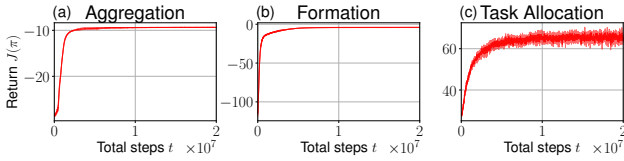


Fig. 2. Training curves of the MFC algorithm trained on  $N = 300$ , plotting the average achieved objective over time steps taken, together with its standard deviation over 3 seeds. The MFC approach leads to very stable learning results for all of our considered problems. (a): Aggregation; (b): Formation; (c): Task Allocation.

c) *Task allocation*: Lastly, we formulate a problem with stochasticity even in the limit. Consider randomly generated, spatially localized tasks such as providing a UAV-based communication uplink, or emergency operations for clearing rubble and firefighting. We add spatially localized tasks to the model which are observed via an additional histogram of task locations. Here, in each time step,  $N_t = \text{Pois}(0.4)$  tasks  $l$  arrive at uniformly random points  $x^l \in \mathcal{X}$ , up to a maximum of 5 total tasks. Each task  $l$  begins with length  $L_t = 10$  and at each time step is processed abstractly by proximity of nearby agents according to  $L_{t+1}^l = L_t^l - \Delta L^l(\mu_t)$ ,  $\Delta L^l(\mu_t) := \min(1, \int (1 - 2\|x - x^l\|_2) \mathbf{1}_{\|x - x^l\|_2 \leq 0.5} \mu_t(dx)$ , until it is fully processed and disappears. The reward is defined by the processed task lengths  $r(\mu_t) = \sum_l \Delta L^l(\mu_t)$ .

### B. Experimental results

In the following, we show results demonstrating the power of our MFC framework for task-driven swarm control, namely their theoretical and numerical advantage over standard MARL, the potential for decentralized open-loop control, and the influence of collision avoidance on optimality, both in simulation and in the real world.

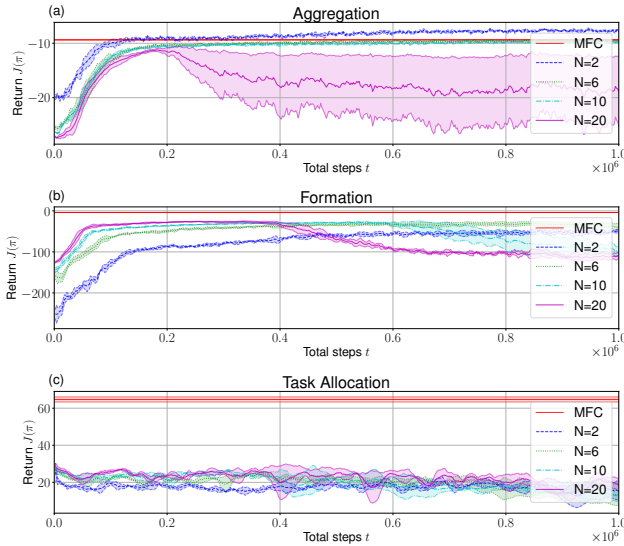


Fig. 3. Training curves of the  $N$ -agent MARL algorithm, plotting the average achieved objective over time steps taken, together with its standard deviation over 3 seeds and compared to final MFC performance (red,  $N = 300$ ). (a): In the simple Aggregation task, MARL and MFC are comparable for few agents, but MARL fails for many agents. (b-c): In more complex scenarios, MFC converges to a better solution than common MARL.

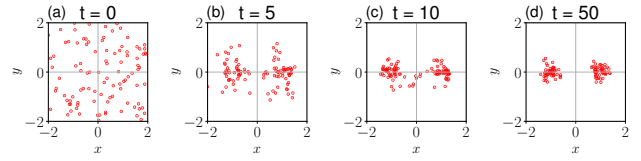


Fig. 4. One sample run of the MFC solution to the Formation problem, applied to a system with  $N = 100$  agents and plotted at times  $t \in \{0, 5, 10, 50\}$ . Agents successfully form a mixture of two Gaussians.

a) *Training results*: In our implementation, each training episode consists of 50, 100 and 200 time steps for the Aggregation, Formation and Task Allocation problems respectively, of which the average sum of rewards will constitute the return values shown in the following figures. As can be seen in Fig. 2, the learning curve of PPO in the MFC problem is smoothly increasing as expected, since the MFC MDP leads to a single-agent problem solvable via standard RL with better understood theory than MARL, e.g. [61].

In contrast, state-of-the-art MARL techniques miss theoretical guarantees. We compare to PPO with parameter-sharing [62] and independent learning [63], which has repeatedly achieved state-of-the-art performance in benchmarks [64]–[67] and remains applicable to arbitrary numbers of homogeneous agents. For comparability, we use the same architecture and implementation as in our MFC experiments, outputting parameters of a Gaussian over actions. Each agent simply observes the same information plus the agent’s own position.

As seen in Fig. 3, MARL works well in the very simple Aggregation task, but becomes increasingly unstable for many agents, especially in the more complex Formation scenario, finally failing entirely in Task Allocation due to non-stationarity of learning [12]. Although MARL could work for other hyperparameter configurations, it shows that standard MARL can suffer from worse stability than single-agent RL in even high-dimensional single-agent MFC MDPs, congruent with the outstanding issue of theoretical MARL convergence guarantees [12]. As seen exemplarily for the Formation problem in Fig. 4 and a variation of the problem with real drones (later) in Fig. 6, the MFC solution successfully achieves the desired mixture of Gaussian formation of agents. In Fig. 5, it can be seen that (i) the MFC solution outperforms MARL, and (ii) the MFC solution quickly converges to the limiting deterministic objective in Fig. 2 as  $N$  grows large, verifying the MFC approximation properties in Theorem 1.

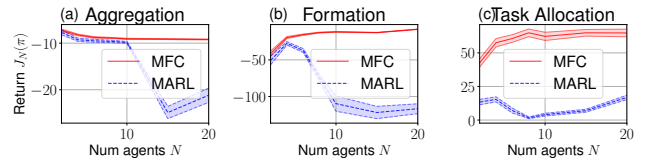


Fig. 5. Comparison of achieved objectives in the finite swarm system of MFC and MARL solutions over 100 sample episodes, with 95% confidence interval (shaded). The MFC algorithm quickly converges to the deterministic, limiting mean-field objective as  $N$  becomes large. In simple scenarios such as Aggregation (a), MARL outperforms MFC in the finite system, while in more complex scenarios (b-c), MFC outperforms MARL (at end of training).

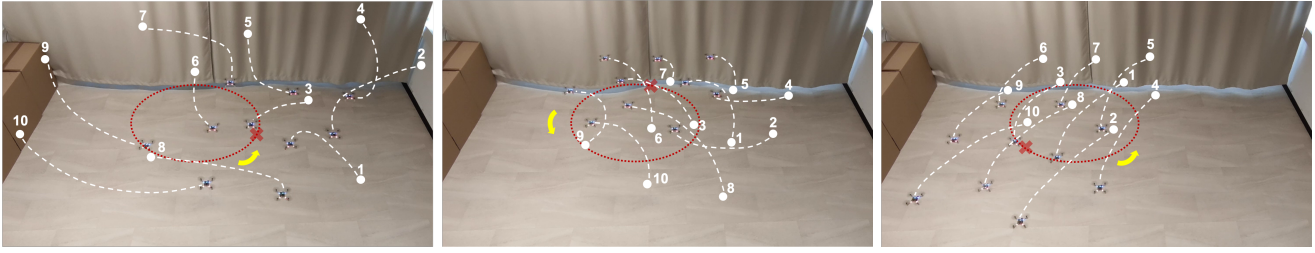


Fig. 6. Real world coverage experiment with a swarm of 10 Crazyflie nano-quadcopters and a variant of the Formation problem where agents track a single time-varying Gaussian distribution (current center of Gaussian shown as red cross), moving counter-clockwise on a circle (red dotted line). The drones successfully track the time-varying Gaussian distribution using the open-loop control policy without collision. Time progresses from left to right.

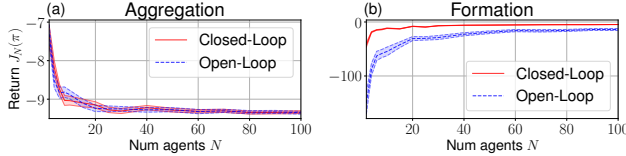


Fig. 7. Comparison of mean objectives in the finite swarm system of closed-loop and open-loop MFC over 100 sample episodes, with 95% confidence interval (shaded). In Aggregation (a), little difference can be seen between the closed-loop and open-loop performance. In Formation (b), the open-loop policy is unable to react to stochastic initialization effects of finite swarm size, only approaching optimality in the large swarm limit.

*b) Decentralized open-loop control:* In the absence of global information, it makes sense for large swarms to let agents act stochastically and independently, especially since agents are interchangeable and anonymous. For this purpose, as long as the limiting MFC is deterministic (e.g. Aggregation and Formation), we can compute an optimal open-loop control sequence  $h_0, h_1, \dots$  of MFC actions  $h_t \in \mathcal{P}(\mathcal{U})^X$  for a given starting  $\mu_0$ , and apply  $h_t$  to each agent. This results in both open-loop and decentralized control, as each agent moves depending on its own local position only. As expected by determinism of MFC, in Fig. 7 we observe that the open-loop performance becomes practically indistinguishable from the closed-loop performance in Aggregation, as well as approaches it in Formation for sufficiently large swarms. We note that at least for finite spaces, very recently a similar decentralization result was also rigorously shown [54].

*c) Influence of collision avoidance:* For MFC with collision avoidance, we simulate  $\Delta t = 2$  and 100 explicit Euler steps of length 0.02 between each decision epoch  $t$ .

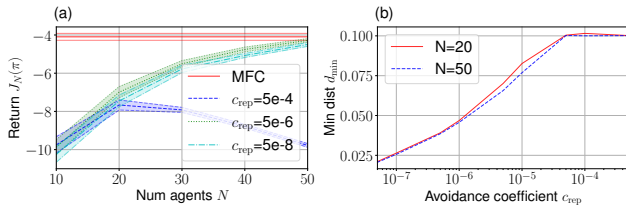


Fig. 8. Comparison of results in the finite swarm system of MFC solution with collision avoidance for various collision avoidance coefficients  $c_{\text{rep}}$  in the Formation problem. (a): Mean objectives averaged over 100 sample episodes, with 95% confidence interval (shaded). (b): Minimum occurring inter-agent distance over 100 sample episodes.

Furthermore, to avoid bad initializations, we resample initial states until the minimal inter-agent distance is above 0.1. As seen in Fig. 8, the minimal inter-agent distance is easily tuned via  $c_{\text{rep}}$ , rising up to the initialization distance 0.1. We find that for strong collision avoidance, the performance deteriorates in the presence of many agents, whereas for smaller collision avoidance coefficients the performance approaches the mean-field limit, verifying Theorem 2.

*d) Illustrative real-life experiment:* Lastly, we show the results of applying a variant of the Formation task – tracking a single time-variant Gaussian moving on a circle – to a real fleet of Crazyflie quadcopters [68], each peer-to-peer-broadcasting only their local Lighthouse-based state estimates [69]. Here, we use the aforementioned decentralized open-loop control and APF-based collision avoidance. Although our experiments remain small-scale due to space constraints and downwash effects, we nonetheless show that our approach works in practice and can be applied to even small swarm sizes. In the future, we imagine similar approaches to be scaled up to larger fleets. As can be seen in Fig. 6, the agents successfully track the formation without colliding.

## V. CONCLUSION

In this work, we have proposed a scalable task-driven approach to robotic swarm control that allows for model-free solution of swarm tasks while remaining applicable in practice by using deep RL, MFC and collision avoidance. Our approach is hierarchical, in principle allowing to profit from any state-of-the-art RL and collision avoidance techniques. Our work is a step towards general toolchains for robotic swarm control, which yet remain part of active research [2]. We have solved part of the limitations of mean-field theory for embodied agents by integrating collision avoidance into the toolchain, but more work on more sparsely interacting mean-field models may be necessary, e.g. for UAV-based communication with strongly neighbor-dependent interaction, by incorporating graph structure [70]–[72]. Extensions to non-linear dynamics and dynamical constraints may be fruitful. Lastly, while our Gaussian parametrization of  $\mathcal{P}(\mathcal{U})$  is efficient, the state discretization still suffers from a curse of dimensionality, as the number of bins rises quickly with fineness of discretization, which is state of the art [44], [45] and could be supplemented e.g. by visual techniques [73].

## REFERENCES

- [1] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, "Swarm robotics: a review from the swarm engineering perspective," *Swarm Intell.*, vol. 7, no. 1, pp. 1–41, 2013.
- [2] M. Schranz, M. Umlauf, M. Sende, and W. Elmenreich, "Swarm robotic behaviors and current applications," *Front. Robot. AI*, vol. 7, p. 36, 2020.
- [3] S.-J. Chung, A. A. Paranjape, P. Dames, S. Shen, and V. Kumar, "A survey on aerial swarm robotics," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 837–855, 2018.
- [4] M. Dorigo, G. Theraulaz, and V. Trianni, "Reflections on the future of swarm robotics," *Science Robotics*, vol. 5, no. 49, p. eabe4385, 2020.
- [5] N. Correll and A. Martinoli, "System identification of self-organizing robotic swarms," in *Distributed Autonomous Robotic Systems 7*. Springer, 2006, pp. 31–40.
- [6] E. Tuci, M. H. Alkilabi, and O. Akanyeti, "Cooperative object transport in multi-robot systems: A review of the state-of-the-art," *Front. Robot. AI*, vol. 5, p. 59, 2018.
- [7] R. Gross and M. Dorigo, "Evolution of solitary and group transport behaviors for autonomous robots capable of self-assembling," *Adapt. Behav.*, vol. 16, no. 5, pp. 285–305, 2008.
- [8] —, "Towards group transport by swarms of robots," *Int. J. Bio-Inspired Comput.*, vol. 1, no. 1/2, pp. 1–13, 2009.
- [9] D. Albani, T. Manoni, D. Nardi, and V. Trianni, "Dynamic UAV swarm deployment for non-uniform coverage," in *Proc. AAMAS*, 2018, pp. 523–531.
- [10] D. Xing, Z. Zhen, and H. Gong, "Offense–defense confrontation decision making for dynamic UAV swarm versus UAV swarm," *Proc. Inst. Mech. Eng. G*, vol. 233, no. 15, pp. 5689–5702, 2019.
- [11] P. Vincent and I. Rubin, "A framework and analysis for cooperative search using UAV swarms," in *Proc. ACM Symp. Appl. Comput.*, 2004, pp. 79–86.
- [12] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.
- [13] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse *et al.*, "Dota 2 with large scale deep reinforcement learning," *arXiv:1912.06680*, 2019. [Online]. Available: <https://arxiv.org/abs/1912.06680>
- [14] H. A. Al-Rawi, M. A. Ng, and K.-L. A. Yau, "Application of reinforcement learning to routing in distributed wireless networks: A review," *Artif. Intell. Rev.*, vol. 43, no. 3, pp. 381–416, 2015.
- [15] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Trans. Transp. Syst.*, vol. 21, no. 3, pp. 1086–1095, 2019.
- [16] H. Hamann, *Swarm Robotics: A Formal Approach*. Springer, 2018.
- [17] K. Elamvazhuthi and S. Berman, "Mean-field models in swarm robotics: A survey," *Bioinspir. Biomim.*, vol. 15, no. 1, p. 015001, 2019.
- [18] A. Bensoussan, J. Frehse, P. Yam *et al.*, *Mean field games and mean field type control theory*. Springer, 2013, vol. 101.
- [19] K. Elamvazhuthi, M. Kawski, S. Biswal, V. Deshmukh, and S. Berman, "Mean-field controllability and decentralized stabilization of markov chains," in *Proc. IEEE CDC*, 2017, pp. 3131–3137.
- [20] V. Deshmukh, K. Elamvazhuthi, S. Biswal, Z. Kakish, and S. Berman, "Mean-field stabilization of markov chain models for robotic swarms: Computational approaches and experimental results," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1985–1992, 2018.
- [21] K. Elamvazhuthi, S. Biswal, and S. Berman, "Mean-field stabilization of robotic swarms to probability distributions with disconnected supports," in *Proc. IEEE ACC*, 2018, pp. 885–892.
- [22] S. Mayya, P. Pierpaoli, G. Nair, and M. Egerstedt, "Localization in densely packed swarms using interrobot collisions as a sensing modality," *IEEE Trans. Robot.*, vol. 35, no. 1, pp. 21–34, 2018.
- [23] S. Mayya, S. Wilson, and M. Egerstedt, "Closed-loop task allocation in robot swarms using inter-robot encounters," *Swarm Intell.*, vol. 13, no. 2, pp. 115–143, 2019.
- [24] K. Lerman, A. Galstyan, A. Martinoli, and A. Ijspeert, "A macroscopic analytical model of collaboration in distributed robotic systems," *Artif. Life*, vol. 7, no. 4, pp. 375–393, 2001.
- [25] U. Eren and B. Açıkmüşe, "Velocity field generation for density control of swarms using heat equation and smoothing kernels," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 9405–9411, 2017.
- [26] T. Zheng, Q. Han, and H. Lin, "Transporting robotic swarms via mean-field feedback control," *IEEE Trans. Autom. Control*, 2021.
- [27] D. Milutinović and P. Lima, "Modeling and optimal centralized control of a large-size robotic population," *IEEE Trans. Robot.*, vol. 22, no. 6, pp. 1280–1285, 2006.
- [28] H. Hamann and H. Wörn, "A framework of space–time continuous models for algorithm design in swarm robotics," *Swarm Intell.*, vol. 2, no. 2, pp. 209–239, 2008.
- [29] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese J. Math.*, vol. 2, no. 1, pp. 229–260, 2007.
- [30] M. Huang, R. P. Malhamé, P. E. Caines *et al.*, "Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle," *Commun. Inf. Syst.*, vol. 6, no. 3, pp. 221–252, 2006.
- [31] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," *arXiv:1802.05438*, 2018. [Online]. Available: <https://arxiv.org/abs/1802.05438>
- [32] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games," in *Proc. NeurIPS*, 2019, pp. 4966–4976.
- [33] J. Subramanian and A. Mahajan, "Reinforcement learning in stationary mean-field games," in *Proc. AAMAS*, 2019, pp. 251–259.
- [34] J. Pérolat, S. Perrin, R. Elie, M. Laurière, G. Piliouras, M. Geist, K. Tuyls, and O. Pietquin, "Scaling mean field games by online mirror descent," in *Proc. AAMAS*, vol. 21, 2022, pp. 1028–1037.
- [35] X. Guo, A. Hu, and J. Zhang, "MF-OMO: An optimization formulation of mean-field games," *arXiv:2206.09608*, 2022. [Online]. Available: <https://arxiv.org/abs/2206.09608>
- [36] B. Anaharci, C. D. Kariksiz, and N. Saldi, "Q-learning in regularized mean-field games," *Dyn. Games and Appl.*, pp. 1–29, 2022.
- [37] K. Cui and H. Koepl, "Approximately solving mean field games via entropy-regularized deep reinforcement learning," in *Proc. AISTATS*, 2021, pp. 1909–1917.
- [38] L. Campi and M. Fischer, "Correlated equilibria and mean field games: a simple model," *Math. Oper. Res.*, 2022.
- [39] P. Muller, R. Elie, M. Rowland, M. Lauriere, J. Perolat, S. Perrin, M. Geist, G. Piliouras, O. Pietquin, and K. Tuyls, "Learning correlated equilibria in mean-field games," *arXiv:2208.10138*, 2022. [Online]. Available: <https://arxiv.org/abs/2208.10138>
- [40] H. Gao, W. Lee, Y. Kang, W. Li, Z. Han, S. Osher, and H. V. Poor, "Energy-efficient velocity control for massive numbers of UAVs: A mean field game approach," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6266–6278, 2022.
- [41] H. Shiri, J. Park, and M. Bennis, "Massive autonomous UAV path planning: A neural network based mean-field game theoretic approach," in *Proc. IEEE GLOBECOM*, 2019, pp. 1–6.
- [42] G. Wang, W. Yao, X. Zhang, and Z. Li, "A mean-field game control for large-scale swarm formation flight in dense environments," *Sensors*, vol. 22, no. 14, p. 5437, 2022.
- [43] A. Šošić, A. M. Zoubir, and H. Koepl, "Reinforcement learning in a continuum of agents," *Swarm Intell.*, vol. 12, no. 1, pp. 23–51, 2018.
- [44] R. Carmona, M. Laurière, and Z. Tan, "Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning," *arXiv:1910.12802*, 2019. [Online]. Available: <https://arxiv.org/abs/1910.12802>
- [45] H. Gu, X. Guo, X. Wei, and R. Xu, "Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis," *SIAM J. Math. Data Sci.*, vol. 3, no. 4, pp. 1168–1196, 2021.
- [46] W. U. Mondal, M. Agarwal, V. Aggarwal, and S. V. Ukkusuri, "On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC)," *J. Mach. Learn. Res.*, vol. 23, no. 129, pp. 1–46, 2022.
- [47] M. Hüttenrauch, S. Adrian, G. Neumann *et al.*, "Deep reinforcement learning for swarm systems," *J. Mach. Learn. Res.*, vol. 20, no. 54, pp. 1–31, 2019.
- [48] P. Fiorini and Z. Shiller, "Motion planning in dynamic environments using velocity obstacles," *Int. J. Robot. Res.*, vol. 17, no. 7, pp. 760–772, 1998.
- [49] M. Hamer, L. Widmer, and R. D'andrea, "Fast generation of collision-free trajectories for robot swarms using gpu acceleration," *IEEE Access*, vol. 7, pp. 6679–6690, 2018.
- [50] M. Everett, Y. F. Chen, and J. P. How, "Collision avoidance in pedestrian-rich environments with deep reinforcement learning," *IEEE Access*, vol. 9, pp. 10357–10377, 2021.
- [51] R. Ourari, K. Cui, A. Elshamhory, and H. Koepl, "Nearest-neighbor-based collision avoidance for quadrotors via reinforcement learning," in *Proc. IEEE ICRA*, 2022, pp. 293–300.

- [52] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, 2017.
- [53] B. Pasztor, I. Bogunovic, and A. Krause, "Efficient model-based multi-agent mean-field reinforcement learning," *arXiv:2107.04050*, 2021. [Online]. Available: <https://arxiv.org/abs/2107.04050>
- [54] W. U. Mondal, V. Aggarwal, and S. Ukkusuri, "On the near-optimality of local policies in large cooperative multi-agent reinforcement learning," *Trans. Mach. Learn. Res.*, 2022. [Online]. Available: <https://openreview.net/forum?id=t5HkgbxZp1>
- [55] K. R. Parthasarathy, *Probability measures on metric spaces*. American Mathematical Soc., 2005, vol. 352.
- [56] R. A. DeVore and G. G. Lorentz, *Constructive approximation*. Springer Science & Business Media, 1993, vol. 303.
- [57] E. Liang, R. Liaw, R. Nishihara, P. Moritz, R. Fox, K. Goldberg, J. Gonzalez, M. Jordan, and I. Stoica, "RLlib: Abstractions for distributed reinforcement learning," in *Proc. ICML*, 2018, pp. 3053–3062.
- [58] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017. [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [59] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," in *Proc. IEEE ICRA*, vol. 2, 1985, pp. 500–505.
- [60] C. Villani, *Optimal transport: old and new*. Springer, 2009, vol. 338.
- [61] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. ICML*. PMLR, 2015, pp. 1889–1897.
- [62] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in *Proc. AAMAS*, 2017, pp. 66–83.
- [63] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proc. ICML*, 1993, pp. 330–337.
- [64] C. S. de Witt, T. Gupta, D. Makoviichuk, V. Makoviychuk, P. H. Torr, M. Sun, and S. Whiteson, "Is independent learning all you need in the Starcraft multi-agent challenge?" *arXiv:2011.09533*, 2020. [Online]. Available: <https://arxiv.org/abs/2011.09533>
- [65] C. Yu, A. Velu, E. Vinitzky, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative, multi-agent games," *Proc. NeurIPS Datasets and Benchmarks*, 2022.
- [66] G. Papoudakis, F. Christianos, L. Schäfer, and S. V. Albrecht, "Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks," in *Proc. NeurIPS Datasets and Benchmarks*, 2021.
- [67] W. Fu, C. Yu, Z. Xu, J. Yang, and Y. Wu, "Revisiting some common practices in cooperative multi-agent reinforcement learning," in *Proc. ICML*, 2022, pp. 6863–6877.
- [68] W. Giernacki, M. Skwierczyński, W. Witwicki, P. Wroński, and P. Koziński, "Crazyflie 2.0 quadrotor as a platform for research and education in robotics and control engineering," in *Proc. IEEE MMAR Conf.*, 2017, pp. 37–42.
- [69] M. Greiff, A. Robertsson, and K. Berntorp, "Performance bounds in positioning with the vive lighthouse system," in *Proc. IEEE FUSION*, 2019, pp. 1–8.
- [70] P. E. Caines and M. Huang, "Graphon mean field games and the GMFG equations:  $\epsilon$ -Nash equilibria," in *Proc. IEEE CDC*, 2019, pp. 286–292.
- [71] K. Cui and H. Koepl, "Learning graphon mean field games and approximate Nash equilibria," in *Proc. ICLR*, 2022, pp. 1–31.
- [72] C. Duan, T. Nishikawa, and A. E. Motter, "Prevalence and scalable control of localized networks," *PNAS*, vol. 119, no. 32, p. e2122566119, 2022.
- [73] S. Perrin, M. Laurière, J. Pérolat, R. Élie, M. Geist, and O. Pietquin, "Generalization in mean field games by learning master policies," in *Proc. AAAI*, vol. 36, no. 9, 2022, pp. 9413–9421.