2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Modeling and Optimal Control for a Class of One Dimensional Counter-Swarm Problems with Distributed Point Actuation

Aditya A. Paranjape

***Abstract*— This paper presents a class of counter-swarm problems wherein the mean-field behavior of the swarm is modeled as an infinite dimensional system. This work considers two classes of problems, one in which the spatial distribution of the agents is fixed and one in which it is dynamic and driven by the standard continuity equation in mechanics. The counter-swarm objectives are formulated as optimal control problems and solved numerically using deep Q-networks.**

I. Introduction

Robotic swarms are evolving into a promising technology for carrying out complex surveillance and related tasks using a large number of simple, low-cost robots [1]. In the context of military applications, the development of robust swarming methods has been complemented by the development of methods to control or disrupt swarms using adversarial agents. Such engagements take the form of herding [1], [2], [3], [4] or attrition-driven combat [5]. While it is challenging enough to herd swarms using a small number of pursuers, additional difficulties arise from the need to estimate the dynamics of the swarms [6] and dealing with instances where the swarm might prefer to split in order to maximize its effectiveness [3].

Graph-based models for swarms have been used to study a wide range of control and estimation problems [7]. However, as the number of agents increases, these methods suffer from the curse of dimensionality and from difficulties associated with generating the inter-agent communication graphs. As an alternative to graph-based approaches, systems with a large number of almost homogeneous agents can be examined using mean-field techniques. These techniques convert the governing ordinary differential equations into a system of partial differential equations (PDEs) that governs the mean-field behavior of the swarm. The resulting equation is usually some form of the Fokker-Planck equation, and the system theoretic properties of several classes of such systems have been examined in the literature [8], [9]. This includes controllability in the sense of whether or not the swarm can achieve a certain target density distributions [10] and its amenability to stabilization via optimal control [11], [12].

A. Contribution

In this paper, we consider two classes of counter-swarm problems, recognizing that the states of a swarm consisting of physical or virtual agents can be split into two groups. *Longitudinal* states describe the spatial distribution of the agents, where the notion of space can be abstract. We refer to all other states as *transverse* states. For instance, the density of the swarm is a longitudinal state, while the velocity of the agents is a transverse state.

We develop PDE-based mean field models for two classes of swarms based on longitudinal and transverse dynamics, respectively. The longitudinal dynamics are described by the well-known continuity equation from fluid mechanics, which takes the form of a nonlinear, integro-differential equation. The transverse model is a linear integro-differential equation. Next, we pose optimal control problems representative of typical counter-swarm objectives for each of these systems. We solve these problems numerically using the well-known Q-learning method.

The work presented in this paper is preliminary in nature. To the best of the author's knowledge, the two systems presented here are novel despite apparent similarities with those described in the prior work (for instance, the aforementioned references). We do not attempt a formal analysis of the well-posedness of these systems and leave it is as an open problem. We focus on the formulation and the numerical solution of the optimal (counter-swarm) control problem.

The rest of the paper is organized as follows. We present the preliminaries in Sec. II. We present our models, together with a few properties, in Sec. III. After presenting the problem formulation in Sec. IV, we present numerical results in Sec. V and conclude the paper thereafter.

II. Preliminaries

A. Notation

Definition 1: Let $\mathcal{U} = \{u_1, \ldots, u_m\}$ where $u_i \in \mathbb{R}$ and m need not be finite. Let $P_\tau([0,T];\mathcal{U})$ denote the set of all $\mathcal{U}$-valued piecewise constant functions with a dwell time of $\tau > 0$ over an interval $[0,T] \subset \mathbb{R}$; i.e., if a function $f \in P_\tau$, then, $f(s) = f(t)$ for all s, t that satisfy $\lfloor (s/\tau) \rfloor = \lfloor (t/\tau) \rfloor$.

B. First order Reynolds' model with actuation

Reynolds' models [13] are commonly employed to study flocks and swarms. A second-order Reynolds' model, informally, describes the net acceleration of a robot or an agent in the swarm as a combination of accelerations due to nearest neighbor tracking (cohesion), safe distance-keeping from other agents (repulsion), collision-avoidance, and goal seeking. In this paper, we prescribe a first-order model which embodies the same behavior and include external actuation

AP is with TCS Research, Tata Consultancy Services Ltd, Pune, India. He is also Honorary Lecturer in the Department of Aeronautics at Imperial College London and Visiting Associate Professor at the Indian Institute of Technology (IIT) Bombay, Mumbai, India. Email: aditya.paranjape@tcs.com.

so that Reynolds' equation takes the following form:

$$\dot{z}_i(t) = \sum_{j \in N_i} \left(1 - \frac{1}{f(r_{ij})}\right)(z_j - z_i) + v_g + \sum_p U(r_{pi}) \quad (1)$$

where $z_i \in \mathbb{R}^n$ is the the motion coordinate of interest of the i^{th} agent, $v_g \in \mathbb{R}^n$ is goal-seeking term (velocity), $r_{ij} \in \mathbb{R}_{\geq 0}$ denotes the distance between agents i and j, and N_i is the set of agents in the sensing neighborhood of j. The function $f(\cdot) \in \mathcal{K}^\infty(\mathbb{R}_{\geq 0})$. The function $U(r_{pi})$ captures the influence of external control inputs, where r_{pi} denotes the distance between the p^{th} point of actuation and the i^{th} agent. We will specify $U(\cdot)$ on a case-by-case basis.

C. Approximate dynamic programming using DQN

Let $z \in \mathbb{R}^n$ denote the state of the system and let $u \in \mathcal{U} = \{u_1, \ldots, u_m\}$, $u_i \in \mathbb{R}$ and m finite, denote the control input. Consider the optimal control problem (OCP)

$$\max_{v[0:T-1]} \sum_{k=0}^{T-1} r(z[k], u[k], z[k+1]),$$
$$\text{s.t.} \qquad z[k+1] = F(z[k], u[k])$$

where $F(\cdot, \cdot)$ is assumed to be known to the control designer and the reward function $r : \mathbb{R}^n \times \mathcal{U} \times \mathbb{R}^n \to \mathbb{R}$ is a deterministic function of its arguments. When m is sufficiently small, approximate dynamic programming via the deep Q network (DQN) technique can be used to estimate the optimal control law [14]. The dynamic program to be solved is given by

$$V(z) = \max_{u \in \mathcal{U}} (r(z, u, z') + \gamma V(z')), \;\; z' = F(z, u)$$

where V is the cost-to-go (or the value function), $\gamma \in (0, 1]$ is the discount, and, for compactness, we denote $z \equiv z[k]$, $u \equiv u[k]$, $z' \equiv z[k+1]$. We define the Q function, $Q : (\mathbb{R}^n \times \mathcal{U}) \to \mathbb{R}$ as

$$Q(z, u) = r(z, u, z') + \gamma V(z') \quad (2)$$
$$\implies \qquad Q(z, u) = r(z, u, z') + \gamma \max_{v'} Q(z', v')$$

The optimal control law is given by

$$u^*(z) = \arg\max_u Q(z, u)$$

For computational purposes, we approximate Q using two identical neural networks parametrized by their weights θ and μ. We rewrite (2) as:

$$Q_\theta(z, u) = r(z, u, z') + \gamma \max_{v'} Q_\mu(z, u') \quad (3)$$

The weights θ and μ are initialized randomly and adjusted recursively out a large number of simulations. During each simulation, or at the end of a prescribed number of simulations, the weights θ are updated using stochastic gradient descent to minimize the empirical loss function

$$L_\theta = \sum_{i=1}^{N_s} (Q_\theta(z, u) - r(z, u, z') - \gamma \max_{v'} Q_\mu(z, u'))^2$$

where N_s denotes the number of samples, each of which is a combination (z, u, z'). The weights μ are set to θ after every N_u training episodes.

III. Continuum Swarm Models

Aerial swarms are typically 3-dimensional in nature. We present a simplified case wherein the swarm is assumed to be 1-dimensional: informally, this allows us to focus on our essential ideas without getting bogged down by the dimensionality of the problem.

Definition 2: A 1-d swarm is a 1-dimensional continuum over $X \subseteq \mathbb{R}$, with $\rho(x) \in \mathbb{R}_{\geq 0}$ denoting the density of the agents at $x \in X$.

A. Transverse dynamics

The simplest swarm model is one where the density $\rho(x)$ in Definition 2 is constant for all $x \in X = [0, 1]$. A canonical example is a string of (uncountably infinite) particles, each of which is free to vibrate perpendicular to the length of the string. Information broadcast and consensus can also be modeled in this framework. Let $y(t, x) \in \mathbb{R}$ denote the state of the swarm at coordinate x and time t. We will define the space in which y lies presently. Let $\mathcal{U} = \{u_1, \ldots, u_m, u_\emptyset\}$, $u_i \in [0, 1]$ for $i \in \{1, \ldots, m\}$ (m finite), denote candidate locations for applying the control input, where $u_\emptyset$ denotes the case where no control input is applied (realized, in practice, by setting the $u_\emptyset \gg 1$).

Remark 1: The choice of piecewise constant control inputs is motivated by the practical problem [1] where the pursuer engages the swarm by hopping between different locations in relation to the swarm while spending a finite, non-zero amount of time at each location.

We write the following dynamics for the transverse dynamics of the state $y(t, x)$:

$$\frac{\partial y}{\partial t}(t, x) = \int_0^1 \psi(\|z - x\|)(y(t, z) - y(t, x))\, dz$$
$$+ U(\|u(t) - x\|), \quad y(0, \cdot) \in H_1([0, 1]; \mathbb{R}) \quad (4)$$

where $u \in P_\tau([0, T]; \mathcal{U})$ and H_1 is the space of square integrable functions over $[0, 1]$ with square integrable first derivatives. The functions $\psi(z)$ and $U(z)$ are of the form

$$\psi(q),\, U(q) = \beta_{\{\psi, U\}} \exp(-\alpha_{\{\psi, U\}} q^2), \quad q \in [0, 1] \quad (5)$$

where $\beta_{\{\}}, \alpha_{\{\}} > 0$ are constants.

Consider the uncontrolled system found by setting $U(\cdot) \equiv 0$ in (4), and assume that it is well-posed. The first result shows that the average value of $y(t, x)$, $x \in [0, 1]$ remains constant in the absence of external control. This is a well-known property in systems whose dynamics are governed by undirected graph Laplacians.

Lemma 1: Consider the governing equation (4) with $U(\cdot) = 0$ and the initial conditions $y(0, \cdot) \in H_1([0, 1]; \mathbb{R})$. Let $w(t) = \int_0^1 y(t, z)\, dz$. Then, $w(t) \equiv w(0)$ for all t.

Proof: Let $\psi(z, x) = \psi(x, z) = \psi(\|z - x\|)$ for $x, z \in [0, 1]$. Differentiating $w(t)$, using (4), and applying Fubini's

theorem, we get

$$\frac{dw}{dt} = \int_0^1 \int_0^1 \psi(z,s)(y(t,z) - y(t,s))\, ds\, dz$$
$$= \int_0^1 \int_0^1 \psi(z,s) y(t,z)\, dz\, ds - \int_0^1 \int_0^1 \psi(s,z) y(t,s)\, ds\, dz$$
$$= 0$$

where we have made use of the fact that $\psi(z,s) = \psi(s,z) = \psi(\|z-s\|)$. This completes the proof. ■

Before proving the next result, we observe that

$$\int_0^1 y(t,x) \int_0^1 \psi(z,x)(y(t,z) - y(t,x))\, dz\, dx$$
$$= \underbrace{\int_0^1 \int_0^1 \psi(z,x) y(t,x)(y(t,z) - y(t,x))\, dx\, dz}_{\text{T1}}$$
$$= -\int_0^1 \int_0^1 \psi(z,x)(y(t,z) - y(t,x))^2\, dx\, dz$$
$$- \underbrace{\int_0^1 \int_0^1 \psi(z,x) y(t,z)(y(t,x) - y(t,z))\, dx\, dz}_{\text{T2}}$$

where the addition and subtraction of $y(t,z)$ in T1 is clearly evident. Notice that the terms T1 and T2 in the equation above are identical since $\psi(z,x) = \psi(x,z)$. It follows that

$$\int_0^1 y(t,x) \int_0^1 \psi(z,x) y(t,z)\, dz\, dx$$
$$= -\frac{1}{2} \int_0^1 \int_0^1 \psi(t,z)(y(t,z) - y(t,x))^2\, dz\, dx \quad (6)$$

We are now ready to prove that the system (4) is asymptotically stable. The following result is an analogue of the well-known property of consensus on connected graphs.

Theorem 1: Consider the governing equation (4) with $U(\cdot) = 0$ and the initial conditions $y(0,\cdot) \in H_1([0,1];\mathbb{R})$. The state $y(t,\cdot)$ converges asymptotically to $w = \int_0^1 y(0,z)\, dz$, a constant.

Proof: From Lemma 1, we know that $w(t) = \int_0^1 y(t,z)\, dz$ is constant for all t. Let us denote the constant value by w. We start by defining the Lyapunov function

$$V(t) = \frac{1}{2} \int_0^1 (y(t,x) - w)^2\, dt, \quad w = \int_0^1 y(0,z)\, dz$$

Differentiating $V(t)$, we get

$$\dot{V} = \int_0^1 (y(t,x) - w) \frac{\partial y}{\partial t}(t,x)\, dx$$

Notice that $\int_0^1 (\partial y/\partial t)\, dx = \dot{w}(t) = 0$. It follows that

$$\dot{V} = \int_0^1 y(t,x) \int_0^1 \psi(z,x)(y(t,z) - y(t,x))\, dz\, dx$$
$$= -\frac{1}{2} \int_0^1 \int_0^1 \psi(z,x)(y(t,z) - y(t,x))^2\, dz\, dx \quad (7)$$

where the last equality follows from (6). From La Salle's invariance principle, and using the fact that $y(t,\cdot) \in H_1([0,1];\mathbb{R})$, it follows that $\lim_{t\to\infty} y(t,x) = w$ for all x. This completes the proof. ■

Next, we add the control actuators; i.e., we reintroduce the term $U(\cdot)$ from (4). We start by considering the effect of the control signal on the average value of the state. Let

$$w(t) = \int_0^1 y(t,x)\, dx$$

Then, it follows from (4) that

$$\dot{w}(t) = \int_0^1 \int_0^1 \psi(z,x)(y(t,z) - y(t,x))\, dz\, dx$$
$$+ \int_0^1 \sum_{i=1}^{m+1} U(\|u_i - x\|)\sigma_i(t) \quad (8)$$

where $\sigma_i(t) = 1$ if $u(t) = u_i$ and 0 otherwise. Recall that $u_{m+1} = u_\emptyset$. The first term on the right hand side of (8) is 0 because of $\psi(z,x) = \psi(x,z)$. Further, let

$$g_i = \int_0^1 U(\|u_i - x\|)\, dx$$

which depends on the actuator placement. It follows that

$$\dot{w}(t) = \sum_{i=1}^{m+1} g_i \sigma_i(t) \quad (9)$$
$$\text{s.t. } \sum_i \sigma_i(t) = 1,\ \sigma_i \in P_\tau([0,T]|\{0,1\})\ \forall\, i$$

This is the well-studied herding problem wherein the objective is to shift the centre of mass of the swarm. The maximizing solution for w is clearly to set $\sigma_k(t) = 1$ for $k = \arg\max(g_i)$. We consider other, non-trivial objective functions in Sec. V.

B. Longitudinal motion

Consider the governing equation for a finite swarm described by (1). It is clear that it would not generalize trivially to a continuum due to the presence of the repulsion term $1/f(r_{ij})$, where $f(0) = 0$. Instead, we observe that its role in a linear swarm is to ensure that the agents are equispaced when the swarm is in equilibrium. The continuum analog of equispaced agents is uniform linear density.

Furthermore, although the swarm may start in a finite domain (i.e., without loss of generality, $\rho(0,x) > 0$ only if $x \in [0,1]$), a continuum analog of (1) need not restrict the agents to $[0,1]$. Instead, we assume that the swarm is contained in $[0,1]$ through an external influence (see, for e.g., [15]) which we do not explicitly model here. Thus, we prescribe that the speed of the agents at $x = 0$ and $x = 1$ should be zero. With this understanding, we assume that the longitudinal motion of a swarm is driven by the difference between the local, point-wise density $\rho(\cdot,x)$, with $x \in [0,1]$, and a constant c. The governing equations for the speed an

agent and the density of the swarm are thus given by

$$\begin{aligned}
&\frac{\partial \rho}{\partial t}(t,x) + v(t,x)\frac{\partial \rho}{\partial x}(t,x) + \rho(t,x)\frac{\partial v}{\partial x}(t,x) = 0 \\
&\frac{\partial \rho}{\partial t}(t,\{0,1\}) + \rho(t,\{0,1\})\frac{\partial v}{\partial x}(t,\{0,1\}) = 0 \\
&v(t,x) = h(x)\Bigg(\int_0^1 \phi(z-x)\left(1-\frac{\rho(t,z)}{c}\right)dz \\
&+U(u(t)-x)\Bigg)
\end{aligned} \tag{10}$$

where the equation for $\rho(t,x)$ is the well-known continuity equation from mechanics. The function $h(x) \in C^1([0,1];\mathbb{R})$ is chosen to ensures that $v(t,0) = v(t,1) = 0 \ \forall t$ and the swarm remains bounded in $[0,1]$. The control input $u \in P([0,T];\mathcal{U})$ enters via $U : \mathbb{R} \to \mathbb{R}$. We assume that $\phi(-p) = -\phi(p)$ and $U(p) = -U(-p)$ for all $p \in \mathbb{R}$.

Remark 2: One candidate for $h \in C^1([0,1];\mathbb{R})$ which satisfies the boundary conditions $h(0) = h(1) = 0$ is

$$h(x) = \begin{cases} \frac{x}{\epsilon}\left(2-\frac{x}{\epsilon}\right) & 0 \le x \le \epsilon \\ 1 & \epsilon < x \le 0.5 \\ h(1-x) & 0.5 < x \le 1 \end{cases} \tag{11}$$

where $0 < \epsilon \ll 1$ is arbitrary.

Remark 3: Equation (10) is the degenerate Fokker-Planck equation found by setting Brownian motion to zero. We note that it is nonlinear in ρ due to the fact that $v(t,x)$ is a linear function of ρ. As a result, well-established results for the well-posedness and control of systems described by the Fokker-Planck equation, such as those in [11], [12], cannot be applied directly to our problem. In this paper, we do not address the well-posedness problem and leave it as a subject of future work. We turn to the optimal control problems under the assumption of well-posedness.

IV. Optimal Control Problem Formulation

The counter-swarm optimal control problem (OCP) deals with the maximization of a given metric, subject to the constraints imposed by the dynamics of the swarm. Let $s(t) \in S$ denote the state of the swarm, where S is a suitable Hilbert space. The metrics, in terms of norms (if they are well-posed), are typically:

1) The mean $\langle s(t) \rangle = \int_0^1 s(t,x)\,dx$, which equivalent to the classic problem of herding the swarm [1].
2) The dissonance metric $\|s(t) - \langle s(t) \rangle\|_{\{\cdot\}}$ in terms of a suitable norm. In herding problems, we seek to *minimize* the dissonance metric (i.e., achieve consensus) while moving the swarm in a predictable manner. On the other hand, in problems where the connectivity of the swarm is to be minimized, we seek to maximize the dissonance metric.

Let $r : S \to \mathbb{R}$ denote the metric of interest, and suppose that we wish to maximize it. Note that $r(t)$ need not be positive for all t: this allows us to accommodate soft constraints through non-positive penalties. Thus, we wish to solve the following OCP:

$$\max_{u \in P_\tau([0,T];\mathcal{U})} \int_0^T r(s(t))\,dt \tag{12}$$

subject to (4) or (10), as per the system of interest, and we assume that U is finite.

Since $u \in P_\tau([0,T];\mathcal{U})$, we replace (12) with its discrete-time version

$$\max_{u \in P_\tau([0,T];\mathcal{U})} \sum_{k=1}^{(T/\tau)} r(q(k\tau))\,dt \tag{13}$$

where we have assumed that T/τ is an integer. If T is not a multiple of τ, we can replace T/τ with its floored value which is an integer. The state transition $s(k\tau) \mapsto s((k+1)\tau)$ is modeled as an integration of the system ((4) or (10)) over the time interval $[k\tau, (k+1)\tau]$.

It is clear that the resulting problem is a combinatorial optimization problem, albeit with $(\mathrm{card}(\mathcal{U}))^{(T/\tau)}$ solutions. We use the approach described in Sec. II-C to solve this problem numerically since a purely analytical solution is cumbersome to derive in all but the simplest cases.

V. Numerical Results

A. Implementation of DQN

In our implementation of DQN, we approximate the Q function via a neural network with three hidden layers of width 24, 12, and 6, respectively. The inputs to the neural network include the remaining time horizon $T - k\tau$ and the values of the state at 10 equispaced points in the spatial dimension. The outputs are passed through a softmax to extract the optimal decision at a given time. We use the Keras API for TensorFlow [1] to implement the neural network. The weights are updated via stochastic gradient descent.

In order to implement the system dynamics, we discretize the PDEs using 100 grid points along the spatial dimension. The spatial derivatives are computed using finite differentiation. Integration with respect to time is carried out using a 2^{nd} order Runge-Kutta scheme.

B. Transverse dynamics

For transverse dynamics (4), we consider two objective functions. We set $\tau = 2$ units. Thus, the control inputs are piecewise constant with "dwell times" of 2 s. Recall that this dwell time has a physical interpretation of the time spent by a pursuer in engaging with agents at a given actuation point. We consider five actuation points located at $\{0, 0.25, 0.5, 0.75, 1.0\}$, so that the number of points available to the pursuer is 6 including $u_\emptyset = 100$.

The two objective metrics of interest (see (13)) are $r_1(y(t)) = \max_x |y(t,x)|$ and $r_2(y(t)) = -\max_x \big(y(t,x) - \langle y(t) \rangle\big) + \langle y(t) \rangle$. Notice that r_2 corresponds to the problem of herding while ensuring the minimum possible dissonance. We denote the net reward by

$$J_i(T) = \sum_{k=1}^{T/\tau} r_i(k\tau),\ i \in \{1, 2\} \tag{14}$$

[1]URL: `https://keras.io/about/`

Figure 1 shows the growth of $J_1(T)$ over 1000 training episodes of $T = 20$ units each. The corresponding trajectory of $y(t)$ is depicted via a series of snapshots in Fig. 2, together with the control inputs. Notice that, in this particular case, the optimal policy involves the pursuer placing itself $x = 0.25$ for the entire duration of the pursuit.

Figure 3 shows the growth of J_2 during the course of training over 1000 episodes. Notice the negative values early in the training phase: these occur because of the high level of local push compared to the average movement of the swarm, leading to higher dissonance. The training works to rectify the dissonance, and the optimal trajectory in Fig. 4 reflects this learning. We notice that the pursuer moves around on the x axis as it seeks to push all sections of the swarm more or less evenly.

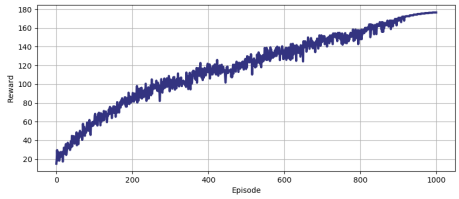


Fig. 1: Episodic reward J_1 from (14) during the training.

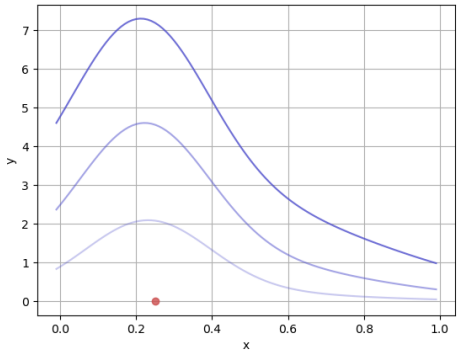


Fig. 2: Snapshots of y at three instants of time: near the initial time (light blue), midway to terminal time, and at the terminal time (dark blue). The red dot shows the position of the actuation - in this case, the position does not change with time.

Fig. 3: Episodic reward during the training.

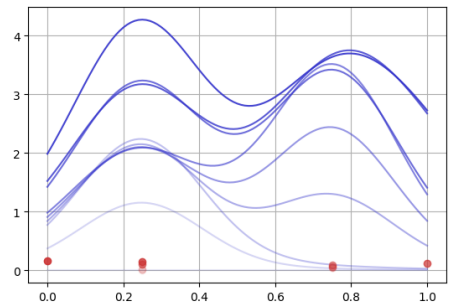


Fig. 4: Snapshots of ρ at 10 instants of time, with darker shades indicating later times. The red dots show the x coordinate of the actuation. The red dots move up with time only for the sake of visualization.

C. Longitudinal dynamics

Next, we consider the longitudinal dynamics with the objective function

$$J = \sum_{k=1}^{(T/\tau)} r(\rho(k\tau)), \quad r(\rho(k\tau)) = \max_x \rho(k\tau, x) \tag{15}$$

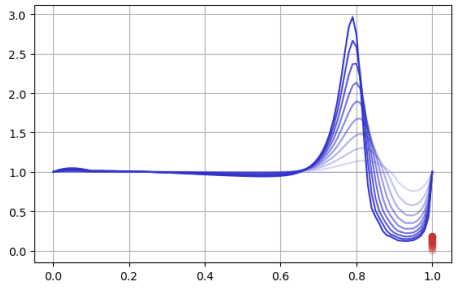
The swarm has an equilibrium at $\rho(x) = 1 \ \forall\, x$. Moreover, (10) ensures that $\int_0^1 \rho(t, x)\, dx$ is constant for all t. Thus, the objective function in (15) can be interpreted as that of maximizing the dissonance in the swarm, since a higher value of ρ in any region of $[0, 1]$ is accompanied by a reduction elsewhere. This objective is of interest because the higher density of agents can actually result in a destabilization of the swarm in practice. This is evident when one considers Reynolds' rules in (1) and their general second-order form [13]: when an agent approaches another agent with excessive closure, it experiences a strong rebound, which can have ripple effect all over the swarm.

The growth of the net reward (15) with training is depicted in Fig. 5, and two sets of snapshots are shown in Fig. 6. The first plot shows the optimal outcome after 500 training episodes, while the second shows the optimal trajectories after 1000 training episodes. In both cases, the optimal approach for the pursuer involves positioning itself at a single location. The location changes between the two plots, and we notice an improvement of approximately 30% when 1000 episodes are employed for training. In both plots, it is worth noting the minimum value of the density in addition to the maximum value. In addition to the aforementioned discussion about the relevance of the maximum value, the permissible range of minimum values can be bounded below as a direct measure of the connectivity of the agent graph. Thus, the second plot of Fig. 6 can be interpreted not only as showing a high degree of dissonance, but also a greater chance of loss of connectivity in the swarm.

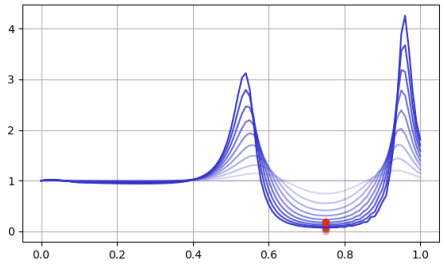
VI. Conclusions

In this paper, we presented two classes of counter-swarm problems. In both cases, the mean field behavior of the

Fig. 5: Episodic reward during the training for the longitudinal dynamics with the reward function (15).



(a) After 500 episodes



(b) After 1000 episodes

Fig. 6: Snapshots of $\rho(t, x)$ at 10 instants of time, with darker shades indicating later times, for two different training durations. The red dots show the x coordinate of the actuation point. The red dots move up with time only for the sake of visualization.

swarm was modeled using infinite dimensional integro-differential equation. The two classes represented dynamics with and without the movement and mixing of agents. The counter-swarm problems were posed as optimal control problems, which we solved numerically using a variant of approximate dynamic programming.

The preliminary work presented in the paper presents several lines for further development. The well-posedness of the nonlinear longitudinal dynamics is of immediate interest to place the results presented here on a theoretically sound footing. The second objective is that of extending the class of objective functions to cover the connectivity of the swarm. Although we mentioned this point briefly in Sec. V, to the best of our knowledge, there does not exist a formal equivalent for the connectivity of the graph in the context of the mean field representation of the swarm.

REFERENCES

[1] A. A. Paranjape, S.-J. Chung, K. Kim, and D. H. Shim, "Robotic herding of a flock of birds using an unmanned aerial vehicle," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 901–915, 2018.

[2] V. S. Chipade and D. Panagou, "Multiagent planning and control for swarm herding in 2-d obstacle environments under bounded inputs," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1956–1972, 2021.

[3] V. S. Chipade, V. S. A. Marella, and D. Panagou, "Aerial swarm defense by stringnet herding: Theory and experiments," *Frontiers in Robotics and AI*, vol. 8, p. 640446, 2021.

[4] C. Walton, I. Kaminer, Q. Gong, A. H. Clark, and T. Tsatsanifos, "Defense against adversarial swarms with parameter uncertainty," *Sensors*, vol. 22, no. 13, p. 4773, 2022.

[5] T. Tsatsanifos, A. H. Clark, C. Walton, I. Kaminer, and Q. Gong, "Modeling and control of large-scale adversarial swarm engagements," *arXiv preprint arXiv:2108.02311*, 2021.

[6] Q. Gong, W. Kang, C. Walton, I. Kaminer, and H. Park, "Partial observability analysis of an adversarial swarm model," *Journal of Guidance, Control, and Dynamics*, vol. 43, no. 2, pp. 250–261, 2020.

[7] S.-J. Chung, A. A. Paranjape, P. Dames, S. Shen, and V. Kumar, "A survey of aerial swarm robotics," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 837–855, 2018.

[8] J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil, "Particle, kinetic, and hydrodynamic models of swarming," *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*, pp. 297–336, 2010.

[9] K. Elamvazhuthi and S. Berman, "Mean-field models in swarm robotics: A survey," *Bioinspiration & Biomimetics*, vol. 15, no. 1, p. 015001, 2019.

[10] K. Elamvazhuthi, H. Kuiper, and S. Berman, "Controllability to equilibria of the 1-d Fokker-Planck equation with zero-flux boundary condition," in *2017 IEEE Conference on Decision and Control (CDC)*, 2017, pp. 2485–2491.

[11] T. Breiten, K. Kunisch, and L. Pfeiffer, "Control strategies for the Fokker- Planck equation," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 24, no. 2, pp. 741–763, 2018.

[12] T. Breiten and K. Kunisch, "Improving the convergence rates for the kinetic Fokker-Planck equation by optimal control," *arXiv preprint arXiv:2205.01369*, 2022.

[13] C. W. Reynolds, "Flocks, herds, and schools: A distributed behavioral model," in *SIGGRAPH'87, Computer Graphics*, M. C. Stone, Ed., vol. 21, no. 4, 1987, pp. 25–34.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[15] S. Gade, A. A. Paranjape, and S.-J. Chung, "Herding a flock of birds approaching an airport using an unmanned aerial vehicle," in *AIAA Guidance, Navigation, and Control Conference*, 2015, p. 1540.