Learning to flock through reinforcement

Mihir Durve
Department of Physics, Università degli studi di Trieste, Trieste, Italy 34127 and
Quantitative Life Sciences Unit, The Abdus Salam International Centre for Theoretical Physics - ICTP, Trieste, Italy 34151

Fernando Peruani
Laboratoire J. A. Dieudonné, Université Côte d'Azur,
UMR 7351 CNRS, Parc Valrose, F-06108 Nice Cedex 02, France

Antonio Celani
Quantitative Life Sciences Unit, The Abdus Salam International Centre for Theoretical Physics - ICTP, Trieste, Italy 34151
(Dated: November 6, 2019)

arXiv:1911.01697v1 [physics.soc-ph] 5 Nov 2019

Flocks of birds, schools of fish, insects swarms are examples of coordinated motion of a group that arises spontaneously from the action of many individuals. Here, we study flocking behavior from the viewpoint of multi-agent reinforcement learning. In this setting, a learning agent tries to keep contact with the group using as sensory input the velocity of its neighbors. This goal is pursued by each learning individual by exerting a limited control on its own direction of motion. By means of standard reinforcement learning algorithms we show that: i) a learning agent exposed to a group of teachers, i.e. hard-wired flocking agents, learns to follow them, and ii) that in the absence of teachers, a group of independently learning agents evolves towards a state where each agent knows how to flock. In both scenarios, i) and ii), the emergent policy (or navigation strategy) corresponds to the polar velocity alignment mechanism of the well-known Vicsek model. These results show that a) such a velocity alignment may have naturally evolved as an adaptive behavior that aims at minimizing the rate of neighbor loss, and b) prove that this alignment does not only favor (local) polar order, but it corresponds to best policy/strategy to keep group cohesion when the sensory input is limited to the velocity of neighboring agents. In short, to stay together, steer together.

The spectacular collective behavior observed in insect swarms, birds flocks, and ungulate herds have long fascinated and inspired researchers [1–6]. There are many long-standing and challenging questions about collective animal behavior: How do so many animals achieve such a remarkably synchronized motion? What do they perceive from their environment and how do they use and share this information within the group in order to coordinate their motion? Are there any general rules of motion that individuals obey while exhibiting collective behavior? Since the last few decades, these questions have been addressed with systematic field observations coupled with mathematical models of animal behavior. Data from experimental observations have been analyzed in order to infer the rules that individuals follow in a group [7–11] and numerous models have been proposed to explain the observed flocking behavior [12–18]. The great majority of flocking studies [1–17], except for few exceptions (see [18] references therein), are based on a velocity alignment mechanism that ensures that neighboring individuals move in the same direction. However, the origin of such a velocity alignment, from a cognitive point of view, is not known, and neither its biological function.

The natural mathematical language that we will use here to discuss collective motion is the framework of Multi Agent Reinforcement Learning (MARL) [19, 20]. In this scheme, the agents can perform actions in response to external cues that they can sense from the environment as well as from other agents. The goal of each agent is to achieve a given objective. In the case at hand, the agents are individuals who can observe the behavior of their close neighbors and react by steering according to some rule. Since it has been hypothesized that there exist many benefits associated to group-living, such as predator avoidance [21] and collective foraging [22], we assume that the objective of the agents is to increase or maintain the cohesion of the group. The essence of Reinforcement Learning (RL) is that, by repeated trial and error, the agents can learn how to behave in an approximately optimal way so as to achieve their goals [23]. Here, we show that velocity alignment emerges spontaneously in a RL process from the minimization of the rate of neighbor loss, and represents a optimal strategy to keep group cohesion.

In the following we will consider individual agents that move at constant speed in a two-dimensional box with periodic boundary conditions. The density of agents is kept fixed to $\rho = 2$ agents/(unit length)^2. Updates are performed at discrete time steps as follows. For the i-th agent, the position update is:

$$\mathbf{r}_i^{t+1} = \mathbf{r}_i^t + v_0 \mathbf{v}_i^t \Delta t \ , \tag{1}$$

where $\mathbf{r}_i^t$ and $\mathbf{v}_i^t$, with $||\mathbf{v}_i^t|| = 1$, are the position and moving direction, respectively, of the agent at time t; the term v_0 corresponds to the speed, which we fix to $v_0 = 0.5$, and $\Delta t = 1$. At each time step, each agents makes a decision on whether keeping the current heading direction or performing a turn. The decision-making process is based on the sensorial input of the

agent, which corresponds to the angular difference between the (normalized) average velocity defined by $\mathbf{P}_i = (\sum_{|\mathbf{r}_j - \mathbf{r}_i| < R} \mathbf{v}_j^t)/n_i$ (with n_i the number of neighbors of agent i within its perception range R) and the moving direction of the agent $\mathbf{v}_i^t$. Below, we take $R = 1$. We can express the state as

$$s_i^t = \arg(\mathbf{P}_i, \mathbf{v}_i^t)\,, \tag{2}$$

where the function $\arg(\mathbf{P}_i, \mathbf{v}_i^t)$ is defined as $\arccos(\mathbf{P}_i \cdot \mathbf{v}_i^t / ||\mathbf{P}_i||)$ for $\mathbf{P}_i \cdot (\mathbf{v}_i^t)_\perp > 0$ with $(\mathbf{v}_i^t)_\perp$ obtained by rotating $\pi/2$ counter clockwise the unit vector $\mathbf{v}_i^t$, and minus this quantity otherwise. This means that $s_i^t \in [-\pi, \pi)$. For computational simplicity, we discretize s_i^t by dividing 2π into K_s equally spaced elements. In the RL language, the relative angle s_i^t is the contextual information that defines the current state of the i-th agent. Knowing s_i^t, the agent updates $\mathbf{v}_i^t$ by turning this vector an angle a_i^t

$$\mathbf{v}_{t+1} = \mathcal{R}(a_i^t)\mathbf{v}_t\ , \tag{3}$$

where $\mathcal{R}(a_i^t)$ is a rotation matrix. Note that there are K_a possible turning angles, equally spaced in $[-\theta_{max}, \theta_{max}]$. In the RL jargon, choosing the turning angle a_i^t represents an "action" performed by the agent. The association of a given state s_t^i with an action a_t^i is called a policy.

Policy evaluation takes place at each time step as the agent receives a (negative) reinforcement signal in the form of a cost c_i^{t+1} for losing neighbors within its perception range R

$$c_i^{t+1} = \begin{cases} 1, & \text{if } n_i^{t+1} < n_i^t \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where n_i^t is the current number of neighbors. The goal of the learning agent is to find a policy that minimizes the cost. To achieve this goal the agent makes use of a simple learning rule [23, 24]. The i-th learning agent keeps in memory a table of values $Q_i(s, a)$ for each state-action pair (here a matrix $K_a \times K_s$) which is updated at each step of the dynamics – only for the entry that corresponds to the state just visited and the action just taken – according to

$$Q_i(s_i^t, a_i^t) \leftarrow Q_i(s_i^t, a_i^t) + \alpha[c_i^{t+1} - Q_i(s_i^t, a_i^t)]\,. \tag{5}$$

This update rule effectively constructs an estimator for the expected cost that will be incurred by starting in a given state and performing a given action. The policy at each time-step is based upon the current Q_i according to so-called $\epsilon-$greedy exploration scheme:

$$a_i^t = \begin{cases} \underset{a'}{\text{argmin}}\ Q_i(s_i^t, a') & \text{with prob. } 1-\epsilon \\ \text{an action at random} & \text{with prob. } \epsilon \end{cases}\ . \tag{6}$$

In the simulations we have used $\alpha = 0.005$ and various different schedules for the exploration probability [23].

We start by considering the case when there is a single learning agent in a crowd of N teachers (see Figure 1)

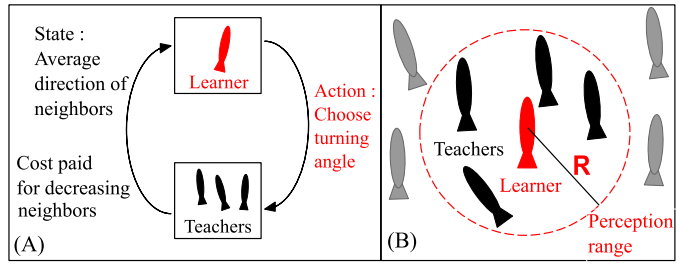

FIG. 1. Learning to flock within a group of teachers. (A) Scheme of reinforcement learning. (B) Neighbors (black) within the perception range of the learner (red).

who have a hard-wired policy:

$$a_i^t(s_i^t) = \begin{cases} s_i^t & \text{if } |s_i^t| \le \theta_{max} \\ \theta_{max} & \text{if } s_i^t > \theta_{max} \\ -\theta_{max} & \text{if } s_i^t < -\theta_{max} \end{cases}\ . \tag{7}$$

This decision rule is nothing else but a version of the Vicsek model of flocking with a discrete number of possible moving direction and limited angular speed. Thus, teachers display robust collective motion. The learning agent, on the contrary, does not have a fixed policy, but one that evolves over time as it acquires experience, i.e. by visiting a large number states, and evaluating the outcome of executing its actions. In this case we find that for suitably chosen learning rates α and exploration probability ϵ the algorithm approaches an approximately optimal solution to the decision-making problem after some period of training.

In simulations, we break the training session into a number of training episodes of equal prescribed duration of 10^4 time steps. In each episode the teachers start with random initial positions and velocities. After a transient, they form an ordered flock and at this time we introduce the learner and implement the learning algorithm. At the very beginning, the learner starts with a Q-matrix with all zero entries, which in a case of optimistic initialization (the naive learner expects to incur no costs), a choice that is known to favor exploration [23]. From one episode to the following, the learner keeps in memory the Q-matrix that it has learned so far. During the training session, we measure the success of the learning process with the average cost that a learner pays per time step, that is the rate at which it is losing contact with the teachers (see Fig. 2A). As the training progresses, the rate at which neighbors are lost starts from an initial value of 0.5, meaning that on average the learner loses contact with some neighbor every other step, to decrease and eventually saturate down to a value around 0.1 meaning that the contact is kept for 90% of the time. In the insets of Fig 2A we show samples of short trajectories of the learner and some teachers at the early and later phase of the training process. We observe that in the early phase of the training, the learner essentially moves at random (see movie1.mp4) and eventually it learns to stay within the flock (see movie2.mp4). In Fig 2C we

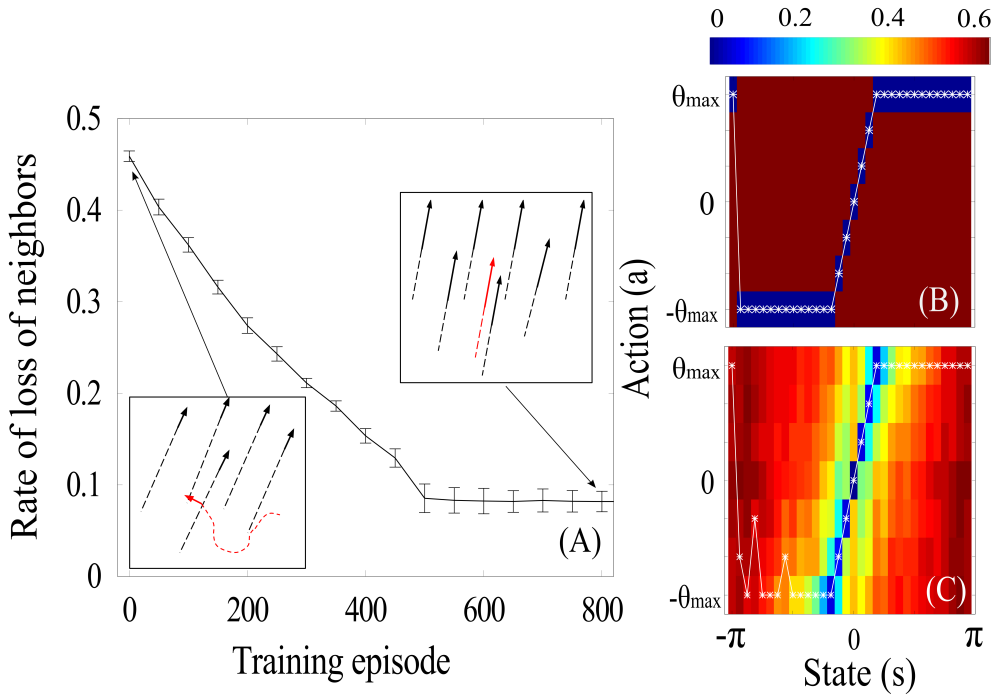

FIG. 2. Single learner results. (A) Performance of the learner as training progresses. The error bars indicate standard deviation in the values in 20 training sessions. In the inset, some short trajectories of the learning agent (red) and teachers (black) at the stages indicated by the arrows. The number of teachers is $N = 200$. The maximal turning angle is $\theta_{max} = 3\pi/16$. (B) The Q-matrix, i.e the average cost incurred for a given state-action pair by the teachers. White stars shows the action a taken by the teacher when in state s. (C) Q-matrix of the learner at the end the training session. White stars denote the best estimated action of the learner for each state.

show that the policy discovered by the learner is identical with the pre-defined policy of the teachers, see Eq. (7) and Fig. 2B. It is important to remark that the one and only goal of the learner is to keep contact with its neighboring teachers, not to imitate their behavior. It simply turns out that the best strategy for the learner is in fact the underlying rule that was assigned to teachers.

Now, let us move our focus to the situation where there are no teachers, but only N independently learning agents (see Figure 3). A distinctive difficulty of applying reinforcement learning to the multi-agent setting is that all individuals have to concurrently discover the best way to behave and cannot rely upon knowledge previously acquired by their peers. However, we find that N learning agents are able to overcome this hurdle and are actually capable of learning to flock even in the case when all of them start as absolute beginners (all Q-matrices initialized to zero).

To characterize the performance of the learners, we measure the average rate of loss of neighbors. In Fig 4A we show the average cost for various groups sizes and state-action space discretizations $\{K_s, K_a\}$. The cost reaches a small and steady value after few hundreds of episodes. As the group size grows, the performance remains essentially the same. Conversely, refining the discretization allows to further reduce the costs: for 128 relative alignment angles and 28 turning angles the agents do not lose neighbors for about 97% of the time.

The resulting Q-matrix at the end of the training, averaged over all learners, is shown in Fig. 4B. The colors represent the numerical values in the Q-matrix and the discovered policy is shown with white points. We observe that the discovered policy is the same that the one learned by the single agent with N teachers.

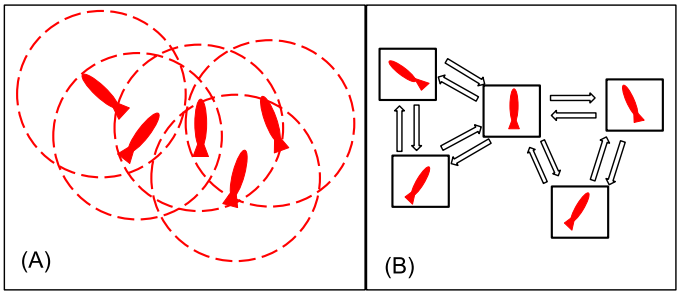

FIG. 3. Multi-agent concurrent learning. (A) Multiple agents with their perception range. (B) Agents interact with short-range, reciprocal interactions.

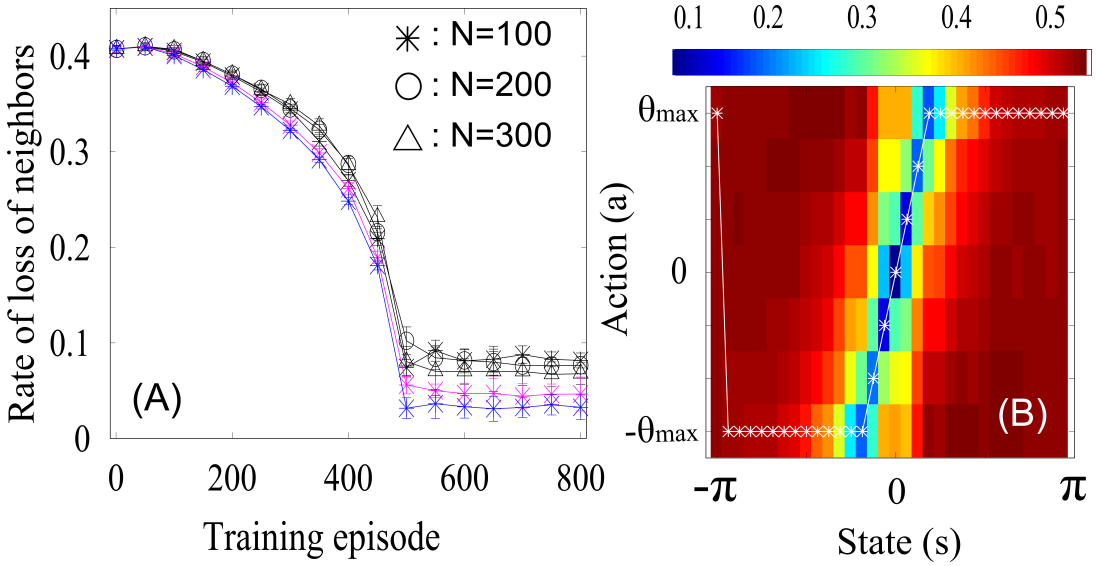

FIG. 4. Results for the multi-agent concurrent learning. (A) Average performance of learners in groups of different sizes. Black, magenta and blue colors corresponds to state-action spaces of size $\{Ks, Ka\}$=$\{32, 7\}$, $\{64, 14\}$, $\{128, 28\}$ respectively. Error bars indicate standard deviation in the average values for each agent. (B) Average Q-matrix at the end of the training for $N = 200$ agents with combination of $\{Ks, Ka\}$=$\{32, 7\}$. White points indicate actions with estimated minimum cost for given state. The colors represent values in the Q-matrix.

It is worth stressing that all agents independently learn the same strategy. We have collected the values of the Q-matrix for a given state ($s = 0$) and different actions, for all agents, at the end of training. The histogram for the frequency of Q values is shown in Fig. 5 where one can observe that there is a clear gap that separates the estimated costs for the optimal turning angle, which lie around 0.1, from the suboptimal actions that have significantly larger costs.

A customary measure of the degree of alignment is the polar order parameter:

$$\psi(t) = \frac{1}{N} \left\| \sum_{i=1}^{N} \mathbf{v}_i^t \right\|. \tag{8}$$

If all the agents are oriented randomly then, as $N \to \infty$, $\psi \to 0$ whereas if all the agents are oriented in the same direction then $\psi = 1$. In Fig 6 we show the evolution of order parameter versus the average cost as the multi-agent learning is advancing. We observe that in the early phases of training the rate of loss of neighbors is comparatively high and the direction consensus among the agents is low, in agreement with the notion that the agents are behaving randomly (see movie3.mp4). As the learning progresses, the agents discover how to keep cohesion, and in doing so they achieve a highly ordered state (see movie4.mp4).

Therefore we conclude that the obtained results proves that the velocity alignment mechanism of the Vicsek model (see Eq. (7)) – based on energy minimization of spin-spin interaction of the XY model – can spontaneously emerge, counterintuitively, from the minimization of neighbor-loss rate, and furthermore represents an optimal strategy to keep group cohesion when the perception is limited to velocity of neighboring agents. In

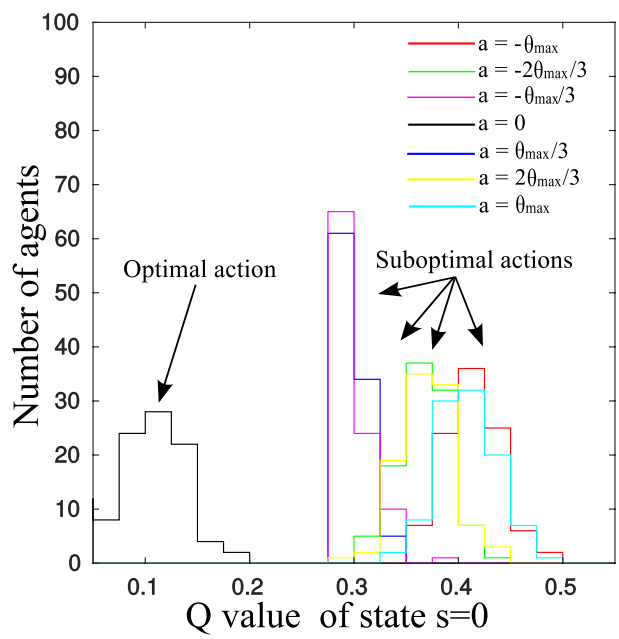

FIG. 5. All agents independently learn the same optimal strategy. The histogram shows the frequency of a give numerical value of $Q(0, a)$ across all learners, at the end of training. The best action $a = 0$ always performs better than any other action. The same holds for other states (not shown). Data obtained with $N = 100$ agents, $K_s = 32$, $K_a = 7$.

summary, if the agents want to stay together, they must learn that they have to steer together.

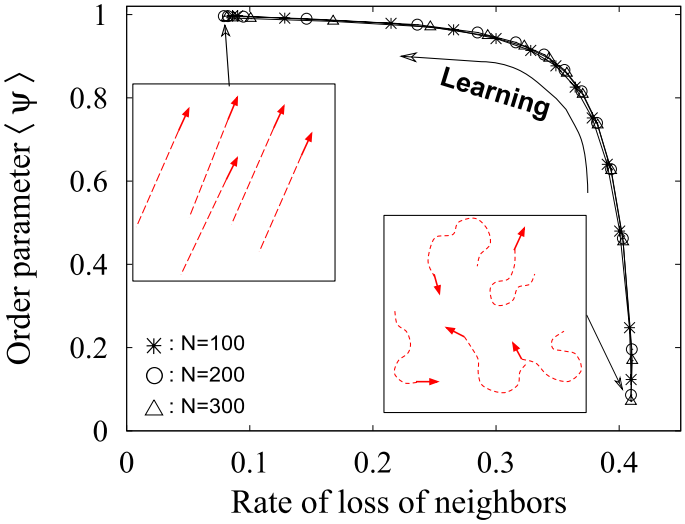

FIG. 6. Average polar order parameter $\langle\psi\rangle$ versus the rate of loss of neighbors. In the insets we show a few short trajectories of naive and trained agents.

In more general terms, we have shown that Multi-Agent Reinforcement Learning can provide an effective way to deal with questions about the emergence of collective behaviors and the driving forces behind them. Our present contribution is just an initial step in this direction and we feel that prospective applications of this approach remain largely unexplored.

For instance, in the present work we have decided at the outset the structure of the perceptual space of the agents, namely the choice of the radius of perception as the relevant parameter and of the relative angle as the relevant state variable. In doing so, we bypassed fundamental questions like: Is the metric distance the most appropriate choice for ranking neighbors ? How should the information given by other individuals be discounted depending on their ranking ? A more ambitious approach would tackle these issues directly through MARL and try to learn from experience what are better choices of the state variable that allow to achieve optimal cohesion.

As another example, here we have tasked our agents with the goal of keeping contact with neighbors, which in itself is understood to be a secondary goal motivated by the primary need of avoiding predators (safety by the numbers) or of increasing the efficiency of foraging. Can one recapitulate the congregation behavior by tasking agents with the primary goal itself ? More explicitly, would agents learn to align themselves by rewarding behaviors that reduce their risk of being predated or increase their chance of getting some food ?

Also, in this paper we have considered a group of identical agents. When agents differ for their perceptual abilities or their dexterity in taking the appropriate actions, then competitive behaviors may arise within the group and the problem acquires a new challenging dimension. How much heterogeneity and competition can be tolerated before it starts impacting the benefit of staying in a group ?

These and many other questions lend themselves to be attacked by the techniques of MARL and we believe that the approach that we have delineated here will show its full potential in the near future.

M. D. acknowledges fruitful discussion with A. Pezzotta, M. Adorisio and A. Mazzolini. M. D. is grateful for the graduate fellowship by the ICTP and University of Trieste. M. D. and A. C. acknowledge hospitality and support from Laboratoire J. A. Dieudonné, Université Côte d'Azur, France.

[1] T. Vicsek and A. Zafeiris, Phys. Rep. **517**, 71 (2012).
[2] J. Buhl, D. Sumpter, I. Couzin, , J. Hale, E. Despland, E. Miller, and S. Simpson., Science **312**, 1402 (2006).
[3] J. Parrish, S. Viscido, and D. Grünbaum, Biol. Bull **202**, 296 (2002).
[4] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, PNAS **105**, 1232 (2008).
[5] J. Puckett, D. Kelley, and N. Ouellette, Scientific Reports **4** (2014).
[6] F. Ginelli, F. Peruani, M.-H. Pillot, H. Chaté, G. Theraulaz, and R. Bon, Proceedings of the National Academy of Sciences **112**, 12729 (2015).
[7] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet, Proceedings of the National Academy of Sciences **107**, 12576 (2010).
[8] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and R. Tavarone, Mathematical Models and Methods in Applied Sciences **20**, 1491 (2010).
[9] J. Herbert-Read, A. Perna, R. Mann, T. Schaerf, D. Sumpter, and A. Ward, Proceedings of the National Academy of Sciences **108**, 18726 (2011).
[10] Y. Katz, K. Tunstrøm, C. Ioannou, C. Huepe, and I. Couzin, Proceedings of the National Academy of Sciences **108**, 18720 (2011).
[11] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. Walczak, Proceedings of the National Academy of Sciences **109**, 4786 (2012).
[12] I. Aoki, Bulletin of the Japanese Society of Scientific Fisheries **48(8)**, 1081 (1982).
[13] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, Phys. Rev. Lett **75**, 1226 (1995).
[14] I. D. Couzin and J. Krause, Adv Study Behav **32**, 1 (2003).
[15] H. Hildenbrandt, C. Carere, and C. Hemelrijk, Behavioral Ecology **21**, 1349 (2010).
[16] A. Cavagna, L. D. Castello, I. Giardina, T. Grigera, A. Jelic, S. Melillo, T. Mora, L. Parisi, E. Silvestri, and M. V. A. Walczak, J. Stat. Phys. **158**, 601 (2005).

[17] D. J. Pearce, A. M. Miller, G. Rowlands, and M. S. Turner, Proceedings of the National Academy of Sciences **111**, 10422 (2014).
[18] L. Barberis and F. Peruani, Phys Rev Lett **117**, 248001 (2016).
[19] L. Panait and S. Luke, Autonomous Agents and Multi-Agent Systems **11**, 387 (2005).
[20] L. Buşoniu, R. Babŭska, and B. Schutter, IEEE transactions on systems, man, and cybernetics-part C: application and reviews **38** (2008).
[21] M. Milinski and R. Heller, Nature **275**, 642 (1978).
[22] T. Pitcher, A. Magurran, and I. Winfield, Behav Ecol Sociobiol **10**, 149 (1982).
[23] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
[24] C. Watkins, Machine Learning **8**, 279 (1992).