PHYSICAL REVIEW LETTERS

༄

Flow Navigation by Smart Microswimmers via Reinforcement Learning

Simona Colabrese,[1,*] Kristian Gustavsson,[1,2] Antonio Celani,[3] and Luca Biferale[1]

[1]*Department of Physics and INFN, University of Rome Tor Vergata, Via della Ricerca Scientifica 1, 00133 Rome, Italy*
[2]*Department of Physics, University of Gothenburg, Origovägen 6 B, 41296 Göteborg, Sweden*
[3]*Quantitative Life Sciences, The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy*

Smart active particles can acquire some limited knowledge of the fluid environment from simple mechanical cues and exert a control on their preferred steering direction. Their goal is to learn the best way to navigate by exploiting the underlying flow whenever possible. As an example, we focus our attention on smart gravitactic swimmers. These are active particles whose task is to reach the highest altitude within some time horizon, given the constraints enforced by fluid mechanics. By means of numerical experiments, we show that swimmers indeed learn nearly optimal strategies just by experience. A reinforcement learning algorithm allows particles to learn effective strategies even in difficult situations when, in the absence of control, they would end up being trapped by flow structures. These strategies are highly nontrivial and cannot be easily guessed in advance. This Letter illustrates the potential of reinforcement learning algorithms to model adaptive behavior in complex flows and paves the way towards the engineering of smart microswimmers that solve difficult navigation problems.

Swimming microorganisms can take advantage of environmental stimuli to bias their motility patterns in order to achieve some biologically relevant goal, some examples being chemotaxis, phototaxis, and gravitaxis [1–3]. Taking inspiration from nature, artificial micro- and nanoswimmers with active internal or external controls could be engineered to execute specialized tasks in complex environments [4–11]. These tasks could be, for example, exploiting advection by the flow to reach specific regions, enhancing transport and mixing, or escaping from potentially dangerous hydrodynamical fluctuations [12–16]. Here, the general questions that we want to address are: can these smart particles learn how to escape their hydrodynamical fate just by sensing simple environmental cues and by reacting to these with the modification of a few control parameters of their dynamics? Is learning also feasible in complex flows, which unavoidably lead to poor performances in the absence of control? What do good strategies look like? Could they be easily intuited *a priori*? To what extent can strategies that perform well in a given environment provide an advantage also under other conditions?

In this Letter, we advocate for the use of reinforcement learning as a general framework to construct efficient strategies for microscopic motility and to train smart particles to accomplish long-term tasks. Reinforcement learning is based on the prolonged and continued interaction between agent and environment, during which the agent—here, a particle or microorganism—learns how to behave optimally by trial and error [17]. The great potential of this approach has been recently demonstrated in the very complex navigation task of thermal soaring in large-scale

turbulent environments [18]. Here, we show by means of numerical experiments that it can be successfully applied to the microscopic problem of gravitaxis in a flowing fluid. We consider active particles that swim with constant speed while being carried away by the underlying flow. The direction of the swimming velocity is determined by the competition between a stabilizing torque that tries to align the particle with a preferred swimming direction—one might think about it as the orientation of a rudder—and the rotation induced by the flow vorticity which could favor or oppose this alignment. If the particle has some control on the preferred direction, how should it operate to achieve its goal, that is, to obtain, in the long run, the largest possible progression in the upward direction? In a quiescent fluid, the optimal choice for the preferred direction is to steadily point upwards. This is realized in *gyrotactic particles* by means of an uneven distribution of mass, see for example [19]. In the presence of an underlying flow, this strategy may reveal to be highly ineffective [20–22]. Indeed, naive gyrotactic particles in a steady flow with horizontal vortex rolls can aggregate in tight clusters and remain trapped at a given height (see Ref. [23] and Fig. 1 below).

Smart gravitactic particles, on the contrary, are endowed with the ability of obtaining some partial information about the regions of the flow that they are visiting. They can use this knowledge to choose directions that maximize the total ascent in the long run, which allows them to escape trapping regions and seek "elevator" regions of the flow. This might be seen as a basic implementation of more realistic behavior as the one given by some species of phytoplankton that are able to actively reorganize its internal organelles in response to fluid mechanical cues

158004-1

[16]. Reinforcement learning provides a way to construct these efficient strategies just by accumulating experience.

Gyrotactic swimmers.—We consider pointlike neutrally buoyant particles that are small enough for inertial effects to be ignored. The flow is not affected by the particles. The trajectories $\boldsymbol{x}(t)$ are determined by a superposition of the fluid velocity $\boldsymbol{u}$ and the swimming velocity $v_s\boldsymbol{p}$,

$$\dot{\boldsymbol{x}} = \boldsymbol{u} + v_s\boldsymbol{p} + \sqrt{2D_0}\boldsymbol{\eta}. \qquad (1)$$

Here, v_s is a constant speed, $\boldsymbol{\eta}(t)$ is Gaussian white noise, $\langle \eta_i(t)\eta_j(t')\rangle = \delta_{ij}\delta(t-t')$, and D_0 is the translational diffusivity. The direction of the swimming velocity $\boldsymbol{p}$ obeys [1]

$$\dot{\boldsymbol{p}} = \frac{1}{2B}[\boldsymbol{k}_a - (\boldsymbol{k}_a \cdot \boldsymbol{p})\boldsymbol{p}] + \frac{1}{2}\boldsymbol{\omega} \times \boldsymbol{p} + \sqrt{2D_R}\boldsymbol{\xi}, \qquad (2)$$

where $\boldsymbol{k}_a$ is a unit vector that defines the preferred direction, B is the time scale of alignment, $\boldsymbol{\xi}(t)$ is a white-in-time Gaussian noise, and D_R is the rotational diffusivity. The index a runs over a discrete set (the actions), each corresponding to a possible choice for the preferred direction. For a flow with a characteristic velocity u_0 and vorticity ω_0, the particle motion is characterized by two dimensionless parameters: the swimming number, $\Phi = v_s/u_0$, quantifying the swimming speed relative to the ambient flow and the stability number, $\Psi = B\omega_0$, measuring the strength of the viscous torque exerted by vorticity relative to the stabilizing torque. The dimensionless translational and rotational diffusivities are chosen to be small (see section 3 in Supplemental Material [24]).

Naive gyrotactic particles have only one single $\boldsymbol{k}_a$ that always points upwards. A systematic exploration of the parameter space for a gyrotactic particle in a Taylor-Green flow (TGF) made of a periodic array of counterrotating vortices has been performed in Ref. [23], and it has revealed the existence of different regimes. The flow configuration can be considered a model for convection in two dimensions [25,26] and can be realized in a laboratory with rotating cylinders or in ion solutions in an array of magnets [27,28]. In the top panel of Fig. 1, we show the phase diagram for naive gyrotactic particles in the TGF, $\boldsymbol{u} = (u_0/2)[-\cos x \sin z, \sin x \cos z]$. The results coincide qualitatively with those presented in Ref. [23] with the only difference that the noise terms in Eqs. (1) and (2) remove the occurrence of strictly periodic orbits and other fragile behaviors. Notably, the bottom part of phase space ($\Phi \ll 1$), where the flow strongly affects the dynamics, is characterized by a strong reduction in vertical motion either because of confinement (for fast realignment $\Psi \ll 1$) or due to random undirected motion (for slow realignment $\Psi \gg 1$).

Smart gravitactic particles should be able to significantly improve the ascent by appropriately choosing their
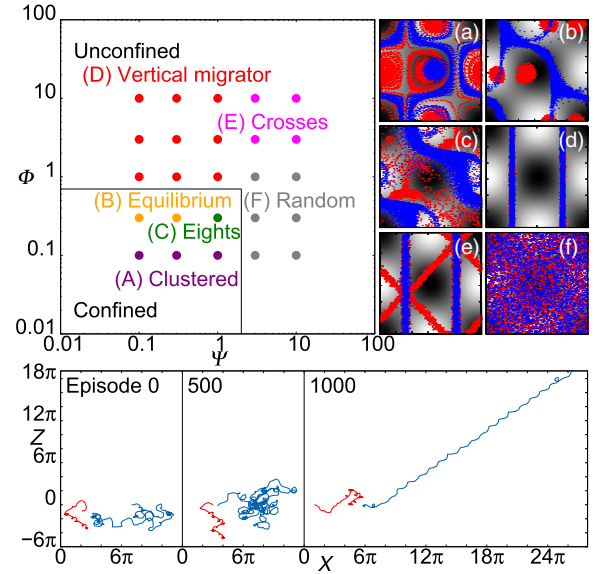

FIG. 1. Top left: phase diagram of gyrotactic particles in a Taylor-Green vortex flow. Each circle represents one of the 25 investigated parameter pairs. Six distinct patterns emerge: confined (a)–(c) and unconfined (d)–(f). Top right: gyrotactic trajectories (red) for each of the 6 patterns plotted on a periodic domain. For the sake of representation, unconfined trajectories are reinjected in the basic $2\pi \times 2\pi$ domain. Trajectories for smart gravitactic particles, after learning the optimal policy to choose the preferred direction $\boldsymbol{k}_a$, are shown in blue. The vorticity field is shown in grey scale. Bottom: a set of representative trajectories at different learning episodes for smart gravitactic particles (blue) compared with typical trajectories for naive gyrotactic particles (red) confined in a trapping dynamics (case C above).

preferred direction $\boldsymbol{k}_a$, according to the environmental cues that they are receiving. Anticipating our results—that will be presented below in full detail—we show in the bottom panel of Fig. 1 the trajectories of smart particles at different stages of the learning process for a given point (labeled C in the left top panel) in the parameter space. Evidently, as experience is accumulated, the smart particle performs better and better and eventually achieves a large upward drift. For different values of the parameters, different gains in vertical motion are obtained, but in all cases, we observe at least some improvement. In the top right panel of Fig. 1, we show a comparison between the spatial distribution patterns of naive gyrotactic and smart gravitactic particles in various regimes. For parameters leading to confinement of naive particles (a)–(c), one can appreciate how smart particles have learned how to concentrate preferentially in regions where the underlying flow facilitates ascent.

Learning gravitaxis using smart particles.—Key ingredients in the reinforcement learning framework are to identify what environmental cues the agent can sense (the states s), what it can do (the actions a), and what reinforcement signals it receives in response to its behavior (the rewards r) [17]. In our setup, particles can perceive only a crude representation of their current swimming

direction $\boldsymbol{p}$ and of the flow. We choose the set of possible states to be the product of two subsets, one indexing the vorticity level of the underlying flow (the vorticity is $\omega = \nabla \times \boldsymbol{u}$) and the other one labeling the instantaneous swimming direction. In short, the discrete state space is $\mathcal{S} = \mathcal{S}_\omega \times \mathcal{S}_p$, where $\mathcal{S}_\omega = \{\omega_-, \omega_0, \omega_+\}$ are three coarse-grained vorticity states (negative, close to zero, and positive), and $\mathcal{S}_p = \{\uparrow, \downarrow, \rightarrow, \leftarrow\}$ are four coarse-grained directions of the instantaneous swimming velocity, pointing mainly upwards, downwards, rightwards, or leftwards (see section 1 in the Supplemental Material [24] for a detailed description). The set of actions comprises four preferred swimming directions, $\boldsymbol{k}_a \in \mathcal{A} = \{\uparrow, \downarrow, \rightarrow, \leftarrow\}$. Particles evolve according to Eqs. (1) and (2) with a given action a that is chosen according to some strategy. When a particle changes state, $s_n \rightarrow s_{n+1}$, because it has moved into a region with a different vorticity level or its swimming direction points to a different angular sector, a reward r_{n+1} is issued. The reward is given by the net increase in altitude experienced by the particle while being in the old state,

$$r_{n+1} = z(s_{n+1}) - z(s_n),$$

where $z(s_n)$ is the initial z coordinate of the particle in state s_n. In the new state, a new action is chosen and the cycle is repeated. The final goal is to maximize the expected return, in this case, the average global long-term vertical displacement,

$$R_{\text{tot}} = \left\langle \sum_{n=1}^{N_s} r_n \right\rangle,$$

where the sum extends up to a horizon $N_s \gg 1$, and the average is over the realizations of the noise in Eqs. (1) and (2) and over initial conditions.

Among the many reinforcement learning algorithms that can produce approximately optimal actions, we adopted Q learning. In a nutshell, it constructs an approximation $Q(s, a)$ of the quality matrix, which is the maximal return that can be achieved starting in state s and taking action a. Given a state s_n and a current approximation Q_n, the action a_n is selected with a bias to favor actions with higher values of Q_n. When the state changes, the current estimate is updated on the basis of the reward that has just been received: if it is larger than expected on the basis of Q_n, the quality function is increased accordingly, otherwise it is decreased. Iterating this procedure, nearly optimal strategies can be obtained (see section 2 in the Supplemental Material [24]).

Operationally, we broke the training sessions into subsequences, called episodes E, with $E = 1, \ldots, N_E$, where N_E is the total number of episodes of each session. The first episode is initialized with an optimistic Q; i.e., all entries are equal and very large. This has the effect to encourage exploration and avoid local maxima. Each episode ends

after a fixed number of total state changes N_s and is followed by a new episode with a random restart of the initial position and orientation of the particle. The initial Q of each restarted episode is given by that obtained at the end of the previous episode (see Supplemental Material [24]).

In order to quantify the success of the learning process, we introduce the *learning gain* of an episode $\Sigma(E)$. It measures the relative increase in the return for smart gravitaxis compared to the return for naive gyrotaxis $R_{\text{tot},g}$,

$$\Sigma(E) = \frac{R_{\text{tot}}}{R_{\text{tot},g}} - 1. \tag{3}$$

In Fig. 2, we show the evolution of $\Sigma(E)$ for ten different training sessions and for a given choice of parameters (Φ, Ψ) in regime C. We see that the smart particle learns how to improve its performance in a robust way. The differences in the asymptotic values across trials are due to the greedy choice of actions based on Q that we have adopted in this particular numerical experiment. Results with better exploratory choices of actions, such as ϵ greedy, support the same conclusions (see section 3 in the Supplemental Material [24]).

In the left panel of Fig. 3, we present a global overview of the gain for all points in phase space that appear in Fig. 1. When the naive gyrotactic particles are confined or move randomly, the gain is very high, while it is just moderate or low—but always positive—when the naive strategy is already performing well.

It is interesting to notice the nontrivial character of the best strategies, which makes them hard to guess *a priori*. This can be appreciated by visualization of the optimal
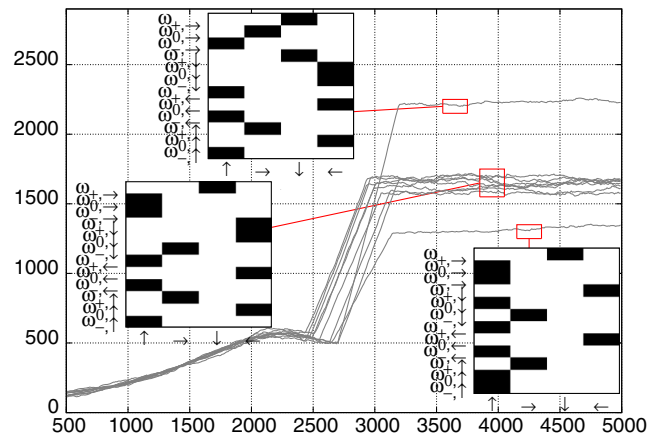

FIG. 2. Dependence of the learning gain $\Sigma(E)$ in percentage vs E for 10 different learning processes (grey curves). Point ($\Psi = 1$, $\Phi = 0.3$) region C in Fig. 1. The value of $\Sigma(E)$ that is visualized is averaged locally on a window of 500 episodes. The insets highlight which preferred directions the smart particle takes for each of the 12 states, according to three final approximately optimal strategies [these are the highest values of the $Q(s, a)$ matrices].
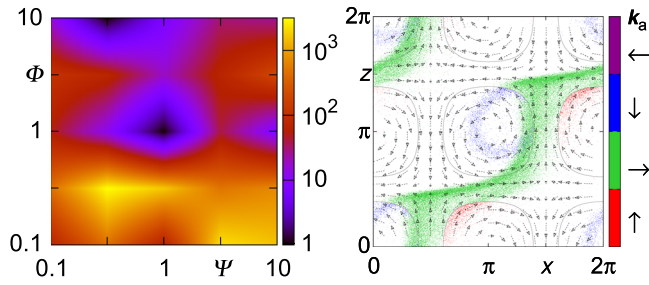
FIG. 3. Left: final averaged learning gain Σ_{avg} (in percentage) for different regions in the parameter space, where the average is made out of 10 different learning experiments. Right: example of the optimal actions for the smart gravitactic particle that succeeded to escape the confinement, parameters ($\Psi = 1$, $\Phi = 0.3$) in region C in Fig. 1. Data obtained from an optimal training using ϵ-greedy exploration. Notice that strategies obtained by permutations that respect the symmetries of the underlying flow would lead to the same learning gain.
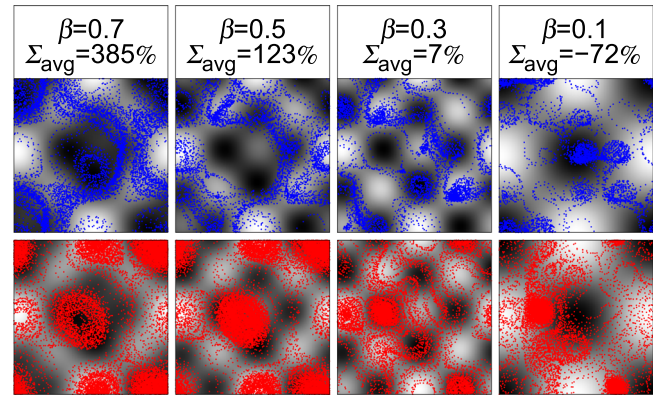

FIG. 4. Comparison between trajectories of naive gyrotactic (red) and smart gravitactic particles (blue) in a perturbed flow. Smart particles move with a policy obtained for the original flow. Above each column, we put β values and the corresponding Σ_{avg}. Case ($\Psi = 0.3$, $\Phi = 0.1$) region A in Fig. 1.

actions taken in different regions of space (Fig. 3, right panel). We observe that the trajectories of smart gravitactic particles have high density in the up-welling regions, showing that they exploit the "elevators" of the flow to reach high altitude. In particular, when the dynamics bring the particle in a recirculation region, the optimal strategy attempts at steering away in the shortest possible time. Sometimes, this might require a seemingly ineffective choice in the short run, such as pointing downwards; the usefulness of this action can be appreciated only over a long horizon and is therefore difficult to guess in advance.

Specialized strategies and flow perturbations.—Given that particles are trained in a specific environment, it is natural to ask how the optimal policy will perform under perturbations of the underlying flow. In general, over-specialized strategies may fail when they have to deal with environments that are wildly different from the ones where the agent has been trained. However, for reasonable classes of environments, they may also display a significant degree of robustness. We addressed this point by evaluating the performance of swimmers trained in the basic Taylor-Green flow when confronted with a more general flow with vorticity, $\omega(x,z) = \beta\omega_1 + (1-\beta)\omega_2$, where $\omega_1(x,z) = -u_0 \cos x \cos z$ is the original flow, and $\omega_2(x,z) = -u_0 \cos 2(x - \Delta x)\cos 2(y - \Delta y)$ is a rescaled and dephased version of the original flow, with $(\Delta x, \Delta y) = (3.35, 1.83)$. The parameter $1 - \beta$ controls the intensity of the perturbation, with $\beta = 1$ corresponding to the original training environment. In Fig. 4, we show the spatial distribution of particles and see that the strategy outperforms naive gyrotaxis even down to values $\beta \sim 0.3$, for which the learning gain is of 7%. Moreover, in section 4 of the Supplemental Material [24], we also explore the robustness and efficiency of the optimal policy in the presence of time-dependent perturbations of the TGF, i.e., even in the case

when tracers might have chaotic evolution. We used either the same strategy optimized without the time-dependent perturbation, or we allowed the smart particle to learn a new policy. In both cases, the smart microswimmers are able to outperform the unskilled ones, at least for the explored phase-space points.

Conclusions.—In this Letter, we have shown how smart particles can learn to accomplish difficult navigation tasks in complex fluid flows. We made no attempt at a fully realistic description of the particle dynamics nor at the actual complexity of real flows, let alone the actual technological implementation of our approach. Rather, our goal was to provide a proof of concept for the possibility to engineer smart microswimmers and to make a case for the use of reinforcement learning algorithms for this purpose. There is enormous room for improvement in many directions: better algorithms, more realistic sensory inputs, and more refined control mechanisms. For instance, Q learning algorithms can be implemented to teach particles that can control their relative density with respect to the underlying fluid to target specific flow configurations. Work in this direction will be reported elsewhere. We hope that our Letter will spur further research on this field at the interface between fluid mechanics, engineering, and computer science.

*simona.colabrese@roma2.infn.it

[1] T. J. Pedley and J. O. Kessler, Hydrodynamic phenomena in suspensions of swimming microorganisms, Annu. Rev. Fluid Mech. **24**, 313 (1992).

[2] T. Fenchel, Microbial behavior in a heterogeneous world, Science **296**, 1068 (2002).

[3] T. Kiørboe and G. A. Jackson, Marine snow, organic solute plumes, and optimal chemosensory behavior of bacteria, Limnol. Oceanogr. **46**, 1309 (2001).

[4] E. Lauga and T. R. Powers, The hydrodynamics of swimming microorganisms, Rep. Prog. Phys. **72**, 096601 (2009).

[5] S. J. Ebbens and J. R. Howse, In pursuit of propulsion at the nanoscale, Soft Matter **6**, 726 (2010).

[6] A. Ghosh and P. Fischer, Controlled propulsion of artificial magnetic nanostructured propellers, Nano Lett. **9**, 2243 (2009).

[7] L. O Mair, B. Evans, A. R. Hall, J. Carpenter, A. Shields, K. Ford, M. Millard, and R. Superfine, Highly controllable near-surface swimming of magnetic janus nanorods: Application to payload capture and manipulation, J. Phys. D **44**, 125001 (2011).

[8] P. Fischer and A. Ghosh, Magnetically actuated propulsion at low reynolds numbers: Towards nanoscale control, Nanoscale **3**, 557 (2011).

[9] J. Wang and W. Gao, Nano/microscale motors: Biomedical opportunities and challenges, ACS Nano **6**, 5745 (2012).

[10] M. Gazzola, B. Hejazialhossein, and P. Koumoutsakos, Reinforcement learning and wavelet adapted vortex methods for simulations of self-propelled swimmers, SIAM J. Sci. Comput. **36**, B622 (2014).

[11] M. Gazzola, A. A. Tchieu, D. Alexeev, A. de Brauer, and P. Koumoutsakos, Learning to school in the presence of hydrodynamic interactions, J. Fluid Mech. **789**, 726 (2016).

[12] F.-G. Michalec, S. Souissi, and M. Holzner, Turbulence triggers vigorous swimming but hinders motion strategy in planktonic copepods, J. R. Soc. Interface **12**, 20150158 (2015).

[13] M. Tanyeri, E. M. Johnson-Chavarria, and C. M. Schroeder, Hydrodynamic trap for single particles and cells, Appl. Phys. Lett. **96**, 224101 (2010).

[14] A. Genin, J. S. Jaffe, R. Reef, C. Richter, and P. J. S. Franks, Swimming against the flow: A mechanism of zooplankton aggregation, Science **308**, 860 (2005).

[15] M. J. Zirbel, F. Veron, and M. I. Latz, The reversible effect of flow on the morphology of ceratocorys horrida (peridiniales, dinophyta), J. Phycol. **36**, 46 (2000).

[16] A. Sengupta, F. Carrara, and R. Stocker, Phytoplankton can actively diversify their migration strategy in response to turbulent cues, Nature **543**, 555 (2017).

[17] R. S. Sutton and A. G. Barto, Reinforcement learning: An Introduction (MIT press, Cambridge, 1998).

[18] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, Learning to soar in turbulent environments, Proc. Natl. Acad. Sci. U.S.A. **113**, E4877 (2016).

[19] J. O. Kessler, Hydrodynamic focusing of motile algal cells, Nature (London) **313**, 218 (1985).

[20] F. Santamaria, F. De Lillo, M. Cencini, and G. Boffetta, Gyrotactic trapping in laminar and turbulent kolmogorov flow, Phys. Fluids **26**, 111901 (2014).

[21] W. M. Durham, J. O. Kessler, and R. Stocker, Disruption of vertical motility by shear triggers formation of thin phytoplankton layers, Science **323**, 1067 (2009).

[22] W. M. Durham, E. Climent, M. Barry, F. De Lillo, G. Boffetta, M. Cencini, and R. Stocker, Turbulence drives microscale patches of motile phytoplankton, Nat. Commun. **4**, 2148, (2013).

[23] W. M. Durham, E. Climent, and R. Stocker, Gyrotaxis in a Steady Vortical Flow, Phys. Rev. Lett. **106**, 238102 (2011).

[24] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.118.158004 for a full description of the state space, a more in-depth discussion of Q learning algorithm and numerical details, some results obtained with a non-greedy policy and the evidence of robustness of the optimal policy in the case of time-dependent perturbations of the TGF.

[25] W. Young, A. Pumir, and Y. Pomeau, Anomalous diffusion of tracer in convection rolls, Phys. Fluids A **1**, 462 (1989).

[26] A. Sarracino, F. Cecconi, A. Puglisi, and A. Vulpiani, Nonlinear Response of Inertial Tracers in Steady Laminar Flows: Differential and Absolute Negative Mobility, Phys. Rev. Lett. **117**, 174501 (2016).

[27] T. H. Solomon and J. P. Gollub, Chaotic particle transport in time-dependent rayleigh-bénard convection, Phys. Rev. A **38**, 6280 (1988).

[28] P. Tabeling, Two-dimensional turbulence: A physicist approach, Phys. Rep. **362**, 1 (2002).