Optimal control of point-to-point navigation in turbulent flows using Reinforcement Learning

M. Buzzicotti[a], L. Biferale[a], F. Bonaccorso[a,b], P. Clark di Leoni[c] and K. Gustavsson[d]

[a]*Dept. Physiscs and INFN, University of Rome Tor Vergata, Via della Ricerca Scientifica 1, 00133 Rome -Italy*
[b]*Center for Life Nano Science@La Sapienza, Istituto Italiano di Tecnologia, 00161 Roma, Italy*
[c]*Department of Mechanical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA.*
[d]*Dept. of Physics, University of Gothenburg, Gothenburg, 41296, Sweden.*

Abstract

We present theoretical and numerical results concerning the problem to find the path that minimizes the time to navigate between two given points in a complex fluid and under realistic navigation constraints. We contrast deterministic Optimal Navigation (ON) control with stochastic policies obtained by Reinforcement Learning (RL) algorithms. We show that Actor-Critic RL algorithms are able to find quasi-optimal solutions in the presence of either time-independent or chaotically evolving flow configurations. For our application, ON solutions develop unstable behaviour within the typical duration of the navigation process, and are therefore not useful in practice. The explored setup consists of using a constant propulsion speed to navigate a turbulent flow. Based on a discretized phase-space the propulsion direction is adjusted with the aim to minimize the time spent to reach the target. Our approach can be generalized to other set-ups, for example unmanned navigation with minimal energy consumption under imperfect environmental forecast or with different models for the moving vessel.

Keywords

Optimal Control, Reinforcement Learning, Turbulence, Unmanned Navigation

1. Introduction

Controlling and planning paths of small autonomous marine vehicles [1] such as wave and current gliders [2], active drifters [3], buoyant underwater explorers, and small swimming drones is important for many geo-physical [4] and engineering [5] applications. In realistic open environments, these vessels are affected by disturbances like wind, waves and ocean currents, characterized by unpredictable (chaotic) trajectories. Furthermore, active control is also limited by engineering and budget aspects as for the important case of unmanned drifters for oceanic exploration [6, 7]. The problem of (time) optimal point-to-point navigation in a flow, known as Zermelo's problem [8], is interesting *per se* in the framework of Optimal Control Theory [9]. In this paper, we report the results from a recent theoretical and numerical study [10], tackling the Zermelo's problem for the navigation in a two-dimensional fully turbulent flow in the presence of an inverse energy cascade, i.e. with chaotic, multi-scale and rough

The 7th Italian Workshop on Artificial Intelligence and Robotics (AIRO 2020)
EMAIL: michele.buzzicotti@roma2.infn.it (M. Buzzicotti); biferale@roma2.infn.it (L. Biferale); fabio.bonaccorso@roma2.infn.it (F. Bonaccorso); pato@jhu.edu (P. C. d. Leoni); kristian.gustafsson@physics.gu.se (K. Gustavsson)

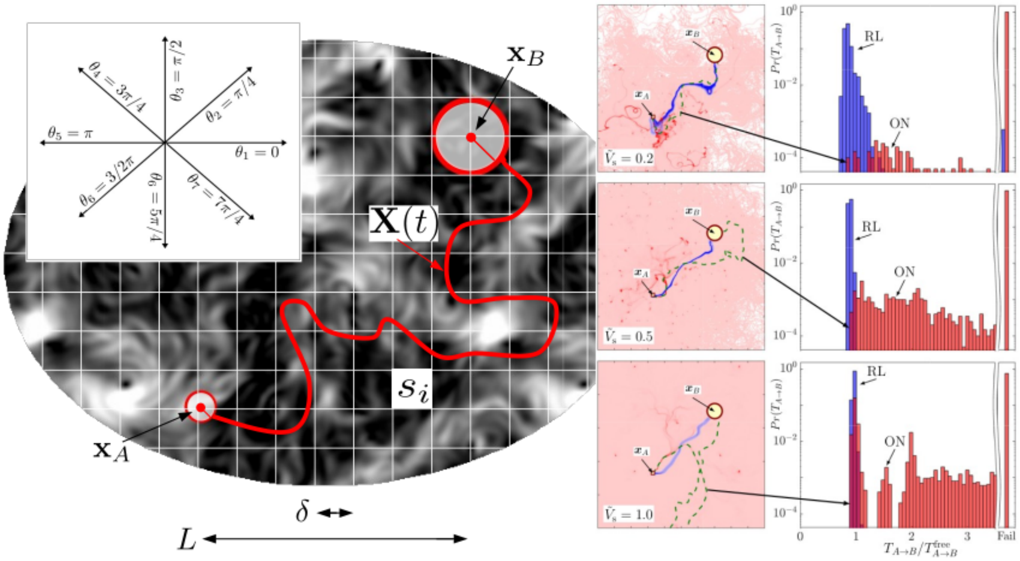

Figure 1: Left: Image of one turbulent snapshot used as the advecting flow, with the starting, $\boldsymbol{x}_A$, and ending point, $\boldsymbol{x}_B$, of our problem. We also show an illustrative navigation trajectory $\boldsymbol{X}_t$. The flow is obtained from a spatially periodic snapshot of a 2D turbulent configuration in the inverse energy cascade regime with a multi-scale power-law Fourier spectrum, $E(k) = \sum_{k<\boldsymbol{k}<k+1} |\boldsymbol{u}(\boldsymbol{k})|^2 \sim k^{-5/3}$. For RL optimization, the initial conditions are taken randomly inside a circle of radius d_A centered around $\boldsymbol{x}_A$. Similarly, the final target is the circle of radius d_B centered around $\boldsymbol{x}_B$. The flow area is covered by a grid-world with tiles s_i with $i = 1, \dots, N_s$ and $N_s = 900$ of size $\delta \times \delta$ which identify the state-space for the RL protocol. The large-scale periodicity of the underlying flow is L, and we fixed $\delta = L/10$. Every time interval Δt, the unmanned vessel selects one of the 8 possible actions a_j with $j = 1 \dots 8$ (the steering directions θ_j depicted in left top inset according to a policy $\pi(a|s)$, where π is the probability distribution of the action a given the current state s of the agent at that time. The policy is optimized during the learning to maximizes the total reward, r_{tot}, proportional to minus the navigation time, $r_{tot} \sim -T_{\boldsymbol{x}_A \to \boldsymbol{x}_B}$, such as the maximum reward corresponds to the time-optimal trajectory. To reach the policy convergence the actor-critic method requires to accumulate experience over a number of the order of 1000 different trajectories, with small variations depending on the values of $\tilde{V}_s$ and the specific flow properties. Right: spatial concentrations of trajectories for three values of $\tilde{V}_s$. The flow region is color coded proportionally to the time the trajectories spend in each pixel area for both ON (red) and RL (blue). Light colors refer to low occupation and bright to high occupation. The green-dashed line shows the best ON out the 20000 trajectories. Right histograms: arrival time distribution for ON (red) and RL (blue). Probability of not reaching the target within the upper time limit is plotted in the *Fail* bar.

velocity distributions [11], see Fig. 1 for a summary of the problem. In such a flow, even for time-independent configurations, trivial or naive navigation policies can be extremely inefficient and ineffective if the set of actions by the vessel are limited. We show that an approach based on semi-supervised AI algorithms using actor-critic Reinforcement Learning (RL) [12] is able to find robust quasi-optimal -stochastic- policies that accomplish the task. Furthermore, we compare RL with solutions from Optimal Navigation (ON) theory [13] and show that the latter is of almost no practical use for the case of navigation in turbulent waters due to strong sensitivity to the initial (and final) conditions, in contrast to what happens for simpler advecting flows [14]. RL has shown to have promising potential to similar problems, such as the training of smart inertial particles or swimming objects navigating intense vortex regions [15, 16, 17].

We present here results from navigating one static snapshot of 2D turbulence (for time-dependent flows see [10]). In Fig. 1 we show a sketch of the set-up. Our goal is to find (if they exist) trajectories that join the region close to $\boldsymbol{x}_A$ with a target close to $\boldsymbol{x}_B$ in the shortest time supposing that the vessels obey the following equations of motion:

$$\begin{cases} \dot{\boldsymbol{X}}_t = \boldsymbol{u}(\boldsymbol{X}_t, t) + \boldsymbol{U}^{ctrl}(\boldsymbol{X}_t) \\ \boldsymbol{U}^{ctrl}(\boldsymbol{X}_t) = V_{\rm s}\boldsymbol{n}(\boldsymbol{X}_t) \end{cases} \tag{1}$$

where $\boldsymbol{u}(\boldsymbol{X}_t, t)$ is the velocity of the underlying 2D advecting flow, and $\boldsymbol{U}^{ctrl}(\boldsymbol{X}_t) = V_{\rm s}\boldsymbol{n}(\boldsymbol{X}_t)$ is the control slip velocity of the vessel with fixed intensity $V_{\rm s}$ and varying steering direction: $\boldsymbol{n}(\boldsymbol{X}_t) = (\cos[\theta_t], \sin[\theta_t])$, where the angle is evaluated along the trajectory, $\theta_t = \theta(\boldsymbol{X}_t)$. We introduce a dimensionless slip velocity by normalizing with the maximum velocity $u_{\rm max}$ of the underlying flow: $\tilde{V}_{\rm s} = V_{\rm s}/u_{\rm max}$. Zermelo's problem reduces to optimize the steering direction θ in order to reach the target [8]. For time independent flows, optimal navigation (ON) control theory gives a general solution[18, 19]. Assuming that the angle θ is controlled continuously in time, the optimal steering angle must satisfy the following time-evolution:

$$\dot{\theta}_t = A_{21}\sin^2\theta_t - A_{12}\cos^2\theta_t + (A_{11} - A_{22})\cos\theta_t \sin\theta_t\,, \tag{2}$$

where $A_{ij} = \partial_j u_i(\boldsymbol{X}_t)$ is evaluated along the agent trajectory $\boldsymbol{X}_t$ obtained from Eq. (1). The set of equations (1-2) may lead to chaotic dynamics even for time-independent flows in two spatial dimensions. Due to the sensitivity to small perturbations in chaotic systems the ON approach becomes useless for many practical applications.

2. Methods

RL applications [12] are based on the idea that an optimal solution can be obtained by learning from continuous interactions of an agent with its environment. The agent interacts with the environment by sampling its states s, performing actions a and collecting rewards r. In our case the vessel acts as the agent and the two-dimensional flow as the environment. In the approach used here, actions are chosen randomly with a probability that is given by the policy $\pi(a|s)$, given the current flow-state s. The goal is to find the optimal policy $\pi^*(a|s)$ that maximizes the total reward, $r_{\rm tot} = \sum_t r_t$, accumulated along one episode, see Fig. 1 for precise definition of flow-states and agent-actions. To identify a time-optimal trajectory we use a potential based reward shaping [20] at each time t during the learning process, see [10] for details. An episode is finalized when the trajectory reaches the circle of radius d_B around the target. In order to converge to robust policies each episode is started with a uniformly random position within a given radius, d_A, from the starting point. To estimate the expected total future reward we follow the one-step actor-critic method [12] based on a gradient ascent in the policy parametrization.

3. Results

In the right part of Fig. 1 we show the main results comparing RL and ON approaches [10]. The minimum time taken by the best trajectory to reach the target is of the same order for the two methods. The most important difference between RL and ON lies in their robustness as seen by plotting the spatial density of trajectories in the right part of Fig. 1 for the optimal policies of ON and RL with three values of $\tilde{V}_s$. We observe that the RL trajectories (blue coloured area) form a much more coherent cloud in space, while the ON trajectories (red coloured area) fill space almost uniformly. Moreover, for small navigation velocities, many trajectories in the ON system approach regular attractors, as visible by the high-concentration regions. The rightmost histograms in Fig. 1 show a comparison between the probability of arrival times for the trajectories illustrated in the two-dimensional domain, providing a quantitative estimation of the better robustness of RL compared to ON. Other RL algorithms, such as Q-learning[12], could also be implemented and compared with other path search algorithms such as A^* which is often used in many fields of computer science [21, 22].
To conclude, we have discussed a systematic investigation of Zermelo's time-optimal navigation problem in a realistic 2D turbulent flow, comparing both RL and ON approaches [10]. We showed that RL stochastic algorithms are key to bypass unavoidable instability given by the chaoticity of the environment and/or by the strong sensitivity of ON approaches in the presence of non-linear flow configurations. RL methods offer also a wider flexibility, being applicable also to energy-minimization problems and in situation where the flow evolution is known only in statistical sense as in partially observable Markov processes. Let us stress that it is possible to implement RL strategies aimed to improve a-priori policy designed to particular problems instead of staring from a completely random policy. For example one can imagine to use an RL approach to optimize an initial "trivial" policy, where the navigation angle is selected as the action that points most directly toward the target.

Acknowledgments

L.B. and M.B. acknowledge funding from the European Union Programme (FP7/2007-2013) AdG ERC grant No.339032. K.G. acknowledges funding from the Knut and Alice Wallenberg Foundation, Grant No. KAW 2014.0048, and Vetenskapsrådet, Grant No. 2018-03974. F.B acknowledges funding from the European Research Council under the European Union's Horizon 2020 Framework Programme (No. FP/2014-2020) ERC Grant Agreement No.739964 (COPMAT).

References

[1] C. Petres, Y. Pailhas, P. Patron, Y. Petillot, J. Evans, D. Lane, Path planning for autonomous underwater vehicles, IEEE Transactions on Robotics 23 (2007) 331–341.

[2] N. D. Kraus, Wave glider dynamic modeling, parameter identification and simulation, Ph.D. thesis, [Honolulu]:[University of Hawaii at Manoa],[May 2012], 2012.

[3] R. Lumpkin, M. Pazos, Measuring surface currents with surface velocity program drifters:

the instrument, its data, and some recent results, Lagrangian analysis and prediction of coastal and ocean dynamics (2007) 39–67.

[4] P. F. Lermusiaux, D. Subramani, J. Lin, C. Kulkarni, A. Gupta, A. Dutt, T. Lolla, P. Haley, W. Ali, C. Mirabito, et al., A future for intelligent autonomous ocean observing systems, Journal of Marine Research 75 (2017) 765–813.

[5] C. Bechinger, R. Di Leonardo, H. Löwen, C. Reichhardt, G. Volpe, G. Volpe, Active particles in complex and crowded environments, Reviews of Modern Physics 88 (2016) 045006.

[6] L. R. Centurioni, Drifter technology and impacts for sea surface temperature, sea-level pressure, and ocean circulation studies, in: Observing the Oceans in Real Time, Springer, 2018, pp. 37–57.

[7] D. Roemmich, G. C. Johnson, S. Riser, R. Davis, J. Gilson, W. B. Owens, S. L. Garzoli, C. Schmid, M. Ignaszewski, The argo program: Observing the global ocean with profiling floats, Oceanography 22 (2009) 34–43.

[8] E. Zermelo, Über das navigationsproblem bei ruhender oder veränderlicher windverteilung, ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik 11 (1931) 114–124.

[9] A. E. Bryson, Y. Ho, Applied optimal control: optimization, estimation and control, New York: Routledge, 1975.

[10] L. Biferale, F. Bonaccorso, M. Buzzicotti, P. Clark Di Leoni, K. Gustavsson, Zermelo's problem: Optimal point-to-point navigation in 2d turbulent flows using reinforcement learning, Chaos: An Interdisciplinary Journal of Nonlinear Science 29 (2019) 103138.

[11] A. Alexakis, L. Biferale, Cascades and transitions in turbulent flows, Physics Reports 767-769 (2018) 1 – 101.

[12] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[13] L. S. Pontryagin, Mathematical theory of optimal processes, Routledge, 2018.

[14] E. Schneider, H. Stark, Optimal steering of a smart active particle, arXiv preprint arXiv:1909.03243 (2019).

[15] S. Colabrese, K. Gustavsson, A. Celani, L. Biferale, Smart inertial particles, Physical Review Fluids 3 (2018) 084301.

[16] S. Colabrese, K. Gustavsson, A. Celani, L. Biferale, Flow navigation by smart microswimmers via reinforcement learning, Physical review letters 118 (2017) 158004.

[17] K. Gustavsson, L. Biferale, A. Celani, S. Colabrese, Finding efficient swimming strategies in a three-dimensional chaotic flow by reinforcement learning, The European Physical Journal E 40 (2017) 110.

[18] L. Techy, Optimal navigation in planar time-varying flow: Zermelo's problem revisited, Intelligent Service Robotics 4 (2011) 271–283.

[19] G. Mannarini, N. Pinardi, G. Coppini, P. Oddo, A. Iafrati, Visir-i: small vessels–least-time nautical routes using wave forecasts, Geoscientific Model Development 9 (2016) 1597–1625.

[20] Y. N. Andrew, D. Harada, S. Russelt, Policy invariance under reward transformations: Theory and application to reward shaping, ICML 99 (1999) 278.

[21] S. Russell, P. Norvig, Artificial intelligence: a modern approach (2002).

[22] J. Lerner, D. Wagner, K. Zweig, Algorithmics of large and complex networks: design, analysis, and simulation, volume 5515, Springer, 2009.