Letter

Reinforcement learning with thermal fluctuations at the nanoscale

Francesco Boccardo[1,2] and Olivier Pierre-Louis[1,*]

[1]*Institut Lumière Matière, UMR5306, Université Lyon 1 - CNRS, Villeurbanne, France*

[2]*MaLGa, Department of Civil, Chemical and Environmental Engineering, University of Genoa, Genoa, Italy*

(Received 18 December 2023; revised 8 February 2024; accepted 6 August 2024; published 30 August 2024)

Reinforcement Learning offers a framework to learn to choose actions in order to control a system. However, at small scales Brownian fluctuations limit the control of nanomachine actuation or nanonavigation and of the molecular machinery of life. We analyze this regime using the general framework of Markov decision processes. We show that at the nanoscale, while optimal control actions should bring an improvement proportional to the small ratio of the applied force times a length scale over the temperature, the learned improvement is smaller and proportional to the square of this small ratio. Consequently, the efficiency of learning, which compares the learning improvement to the theoretical optimal improvement, drops to zero. Nevertheless, these limitations can be circumvented by using actions learned at a lower temperature. These results are illustrated with simulations of the control of the shape of small particle clusters.

DOI: 10.1103/PhysRevE.110.L023301

Reinforcement learning (RL), the process by which an agent learns to choose actions on a dynamical system to maximize rewards from this system, is one of the main machine learning paradigms [1]. Following its long established use for games [2] and robotics [3,4], RL has recently paved the way for many breakthroughs in the control of systems that are constrained by various physical and biological environments. A few examples include the control of fluid flows [5,6], biological and artificial navigation [7–16], organization of active assemblies [17,18], and order and shape of colloidal clusters [19].

Many of these systems involve objects with very small sizes. However, the development of RL for micro and nanometer scale objects faces challenging regimes where thermal fluctuations lead to erratic Brownian motion that dominates the dynamics, and actions then only produce a small bias in stochastic rewards. For instance, nanomotors must be actuated in a robust way within these fluctuating environments [20–23]. Such noise could be detrimental for learning, and indeed a large body of work has been devoted to the robustness of learning in the presence of adverse noise [24–28]. However, it is known that noise can also help learning by enhancing exploration of new states [29,30] or by regularizing the learning process [31,32]. In addition, external fields to manipulate colloids and nanomachines must be weak to achieve nanorobot navigation for drug delivery or remote-controlled surgery [33–36] without damaging the surrounding soft living environment. These systems and the associated emerging technologies call for a better fundamental understanding of the performance of RL when the actuation forces are small as compared to thermal fluctuations.

Previous works on navigation in gridworlds (where space is represented by a finite lattice) suggest that learning is more difficult when increasing temperature [9,37]. Temperature indeed has a strong effect on the speed of the dynamics, which is usually faster at high temperatures and slower at low temperatures. This leads to difficulties in observation, as transitions become too fast or too slow to be observed in experiments. In the following, we show that beyond these difficulties in observability, there are intrinsic physical constraints that make the efficiency of RL vanish in the presence of strong thermal fluctuations.

We investigate this question using simulations of a specific problem that can be considered as a prototype of a nanorobot, where the shape of a fluctuating few-particle cluster is controlled with a macroscopic field [38]. The macroscopic field biases the stochastic configurational changes of the cluster, as expected for atomic or colloidal clusters in the presence, e.g., of an electric field [19,39–42] or light [43]. Our aim is to set the macroscopic field as a function of the observed shape to reach an arbitrary target shape in minimum time. The experimental realization of this system when downscaling particle sizes toward the atomic size is a challenge due to the weakness of available driving forces such as electromigration [38]. This problem can be formulated within the general framework of Markov decision processes, i.e., a Markov chain where the choice of an action—called the policy—is made as a function of the observed state. Hence, our results can be transposed to other RL problems and pertain to all systems that can be modeled with Markov decision processes [1,9], including for example actuation of molecular machines, navigation, and controlled assembly.

In order to grasp the effect of temperature on learning, we do not employ the most sophisticated and powerful RL methods. Instead, we use two of the most elementary ones [1]: Monte Carlo learning (MCL) and Q-learning (QL), which are based on ε-greedy policies. These are standard methods to deal with discrete sets of states and actions, which is our focus in the following. We find that the amount of control achieved

*Contact author: olivier.pierre-louis@univ-lyon1.fr

by RL depends crucially on the dimensionless ratio between the work of the applied force during an elementary move and the thermal energy. When this ratio is small, corresponding to small forces, small scales, or high temperatures, the efficiency of RL is proportional to it. However, this inefficiency of learning at high temperatures can be circumvented by using actions learned at a lower temperature.

Our analysis is based on the comparison of RL to the optimal solution of the control problem. This solution is obtained with dynamic programming (DP) [1,38,44–46], a model-based method that relies on the full knowledge of the laws that govern the system, while RL is a model-free approach that only relies on the observation of the response of the system to some actions without prior knowledge on the governing laws.

a. Cluster model. The dynamics of a fluctuating few-particle cluster are modeled using on-lattice edge-diffusion dynamics [38,41,47,48]. Edge diffusion was observed in metal atomic monolayer islands [49,50] and colloids [51]. Our approach can be readily applied to other types of dynamics for clusters or molecules that preserve the number of particles, such as dislocation-induced events in metal clusters, colloidal clusters, and Wigner crystals, detachment-diffusion-reattachment dynamics inside vacancies in particle monolayers [41,52], or dynamics of polymer and proteins [53].

In this model, particles hop to nearest neighbor sites along the cluster edge and cannot detach from the cluster or break the cluster [38]. The particle hopping rate follows an Arrhenius law [38,47,48]

$$\gamma = \nu \exp[-(nJ - \mathbf{F}\cdot\boldsymbol{u}d)/k_BT], \tag{1}$$

where J is the bond energy, n is the number of bonds of the particle before hopping, k_BT is the thermal energy, $\mathbf{F}$ is the macroscopic force field, d is the lattice constant, and $\boldsymbol{u}d$ is the vector from the initial to the saddle position of the moving particle. For the sake of simplicity, we assume that the saddle point of the diffusion energy landscape is halfway between the initial and final positions [38].

b. RL algorithms. In the language of Markov decision processes [1], the configuration or shape of the cluster is the state s of the system. Moreover, we consider a discrete set of actions labeled by the index a. For simplicity, we consider only three possible actions $a = -1, 0, +1$, that respectively correspond to setting the force to the left, to zero, or to the right, i.e., $\mathbf{F} = aF\mathbf{e}_x$, where $F > 0$ and $\mathbf{e}_x$ is the unit vector along the (10) lattice direction.

The policy $\pi(a|s)$ is the probability to choose action a in state s, and the reward is minus the residence time $\hat{r} = -\hat{t}$ in this state. Here and in the following, a hat indicates a stochastic variable. An episode is a sample of the dynamics consisting of a list of states, actions, and rewards $\{\hat{s}_{k-1}, \hat{a}_{k-1}, \hat{r}_k; k = 1, \ldots, K\}$. The sum of all future rewards is $\hat{g}_k = \sum_{p=k+1}^{K} \hat{r}_p$. Episodes terminate when the state s reaches the target state $\bar{s}$ or when k reaches the maximum allowed number of steps M. Assuming that M is large enough for the target state $\bar{s}$ to be reached with high probability before the end of the episode, minus the average of $\hat{g}$ is a good approximation of the expected first passage time $\tau_\pi(s;\bar{s})$, i.e.,

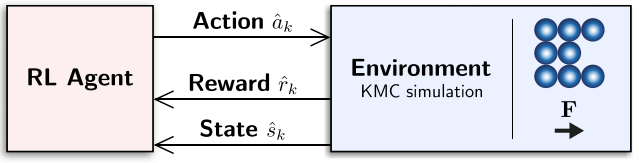

FIG. 1. Schematic of the agent-environment interface in RL.

$$\tau_\pi(s;\bar{s}) \approx -\mathbb{E}_\pi[\hat{g}_k|\hat{s}_k = s], \tag{2}$$

where $\mathbb{E}_\pi[\cdot]$ is the expected value under the policy π.

The optimal policy π_* is a deterministic policy which minimizes the first passage time to the target state: $\pi_* \in \mathrm{argmin}_\pi \tau_\pi(s;\bar{s})$. The optimal first passage times $\tau_*(s;\bar{s})$ obey the Bellman optimality equation [1,54]

$$\tau_*(s;\bar{s}) = \min_a \left[t(s,a) + \sum_{s'\in\mathcal{B}_s} p(s'|s,a)\tau_*(s;\bar{s}) \right], \tag{3}$$

where $\mathcal{B}_s$ is the set of states that can be reached from s in one move, $t(s,a)$ and $p(s'|s,a)$ are respectively the average residence time and the transition probability to state s' when the system is in state s with the action a. This equation can be solved numerically by DP [1,38], which here consists in iterating Eq. (3). Optimal policies for five and seven-particle clusters are shown in Figs. S1 and S2 of the Supplemental Material (SM) [55]. The number of states increases exponentially with the number of particles in the cluster [38], and we have investigated systems up to ten particles ($\sim 3 \times 10^4$ states).

In contrast, RL algorithms aim to find $\tau_*(s;\bar{s})$ and π_* from a set of episodes where they choose the actions without prior knowledge of the model, i.e., of $t(s,a)$ and $p(s'|s,a)$. Here, RL only observes trajectories produced by Kinetic Monte Carlo (KMC), as schematized in Fig. 1. We use two well-known RL algorithms [1] based on the evaluation of the action-value function $q_\pi(s,a;\bar{s})$, which is the expected value of $\hat{g}$ starting from state s, taking the action a first, and then following the policy π for subsequent actions. The first one, MCL, evaluates $q_\pi(s,a;\bar{s})$ from a direct estimate of Eq. (2) using averages over the trajectories that lead to the target in the episodes. In contrast, QL updates an approximation of $q_\pi(s,a;\bar{s})$ so as to reduce the so-called time-difference error [1].

In both MCL and QL, the actions are chosen in such a way to explore the states using an ε-greedy policy: as the exploration parameter ε decreases during learning, actions are chosen very randomly initially, but gradually more greedily, i.e., in a way that minimizes the evaluation of the first passage time to target based on the current approximate estimate of the action-value function. This gradual decrease of the randomness of the policy reflects the standard machine learning tradeoff between exploration and exploitation. At the end of the simulation, we obtain a deterministic policy that approximates an optimal one. Details of these well-known algorithms are provided in Sec. II of the SM [55].

Following Refs. [38,52,56], we use the return time to target $\tau^{\mathrm{r}}(\bar{s})$, defined as the time of first return to the target after leaving it, as a simple representative value of the first passage

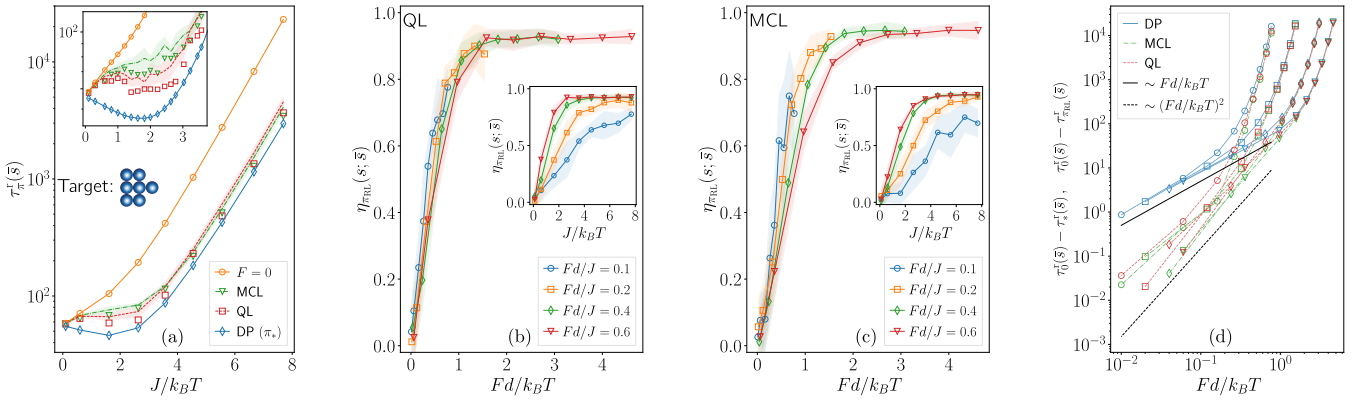

FIG. 2. Learning efficiency. (a) Return time to target for no force, DP, MCL, and QL policies (evaluated with DP) as a function of J/k_BT, with fixed $Fd/k_BT = 0.4$. Inset: zoom in the intermediate/high temperature regime. The markers represent the shortest return time out of ten independent learning runs and the dashed line corresponds to the average. (b) Average learning efficiency $\eta_{\pi_{\rm RL}}(\bar{s})$ [defined in Eq. (4)] as a function of Fd/k_BT for QL. Inset: as a function of J/k_BT. (c) Learning efficiency $\eta_{\pi_{\rm RL}}(\bar{s})$ as a function of Fd/k_BT for MCL. Inset: as a function of J/k_BT. In (a)–(c) the shaded area represents the standard deviation over the ten learning runs. (d) Reduction of the return time obtained with DP: $\tau_0^{\rm r}(\bar{s}) - \tau_{\pi_*}^{\rm r}(\bar{s})$ (blue markers), and with RL: $\tau_0^{\rm r}(\bar{s}) - \tau_{\pi_{\rm RL}}^{\rm r}(\bar{s})$. Markers indicate the values of Fd/J, as in (b) and (c).

times $\tau_\pi(s,\bar{s})$ from all the states s. We use the improvement of $\tau^{\rm r}(\bar{s})$ to define our learning convergence criterion. To avoid limitations due to observability, this criterion is unaware of elapsed physical time, and is only based on the number of episodes, as discussed in Sec. II of the SM [55]. Furthermore, to keep simulation costs down, we focus on clusters with at most ten particles [38].

c. Vanishing learning efficiency at high temperatures and small forces. In Fig. 2(a), the return time obtained with DP, MCL, and QL is shown for a given target state as a function of the inverse temperature J/k_BT for a fixed value of $Fd/J = 0.4$. The case of zero force is also shown, and will serve as a reference for the value of the return time without any optimization (a random policy could also be used, as discussed in Sec. IV of the SM [55]). Strikingly, while the return time obtained with MCL and QL at low temperatures is close to the optimal result provided by DP, no learning is obtained at high temperatures.

The effectiveness of learning can be quantified by the efficiency of the learned policy $\pi_{\rm RL}$, defined as the ratio of the reduction of the return time obtained with $\pi_{\rm RL}$ over the optimal value of this reduction

$$\eta_{\pi_{\rm RL}}(\bar{s}) = \frac{\tau_0^{\rm r}(\bar{s}) - \tau_{\pi_{\rm RL}}^{\rm r}(\bar{s})}{\tau_0^{\rm r}(\bar{s}) - \tau_*^{\rm r}(\bar{s})}, \tag{4}$$

where $\tau_0^{\rm r}(\bar{s})$, $\tau_*^{\rm r}(\bar{s})$, and $\tau_{\pi_{\rm RL}}^{\rm r}(\bar{s})$ are the return times with zero force, optimal policy obtained by DP, and RL policy, respectively. In Figs. 2(b) and 2(c), all efficiencies for QL and MCL for different values of Fd/J are seen to drop to zero when Fd/k_BT is small. This drop of efficiency is also seen in the return time for other targets with various sizes as shown in Fig. S3 of the SM [55], showing that it is not affected by the complexity of the learning task. In addition, it is also found for first passage times from other states, as shown in Fig. S4 of the SM [55], showing that the return time is a good probe for the properties of first passage times.

A heuristic reasoning can rationalize this drop. In our model Eq. (1), Fda corresponds to the change of the diffusion barrier due to the work of the force Fa on the distance $\sim d$ to the energy saddle point. Since energy barriers are divided by k_BT in Eq. (1), an expansion of the rates for small Fd/k_BT leads to a first-order correction proportional to Fda/k_BT. Such correction $\sim Fda/k_BT$ corresponds to a broad class of systems and is also recovered, e.g., with a simple continuum model with no barriers in Sec. VI of the SM [55]. Here, since $a \in \{-1, 0, 1\}$, the corrections of the rates are $\sim Fda/k_BT \sim Fd/k_BT$. Hence, the optimal reduction of the first passage times—which depends on all corrections of the transition rates—is proportional to Fd/k_BT, i.e., $\tau_0^{\rm r}(\bar{s}) - \tau_*^{\rm r}(\bar{s}) = O(Fd/k_BT)$. In contrast, during learning, the actions a are extracted from the policy $\pi_{\rm RL}$. For small Fd/k_BT, $\pi_{\rm RL}$ is dominated by noise, i.e., all actions are equiprobable up to a small correction proportional to Fd/k_BT. The expectation of the action in a given state $\mathbb{E}_{\pi_{\rm RL}}[a]$ is therefore small and proportional to Fd/k_BT. Hence, the expectation of the change in the rates is $\sim Fd\,\mathbb{E}_{\pi_{\rm RL}}[a]/k_BT \sim (Fd/k_BT)^2$. We thus have $\tau_0^{\rm r}(\bar{s}) - \tau_{\pi_{\rm RL}}^{\rm r}(\bar{s}) = O(Fd/k_BT)^2$. As a consequence, the efficiency $\eta_{\pi_{\rm RL}}(\bar{s})$, which is the ratio of the learned over the optimal reduction of the return time, is proportional to Fd/k_BT and drop to zero. Such a behavior of the optimal and learned reductions is confirmed by simulations in Fig. 2(d), and a more rigorous derivation of the drop of efficiency based on an expansion to first order in Fd/k_BT is reported in Sec. IV of the SM [55].

Our results show that the efficiency drops at small Fd/k_BT. Since in practice this means that learning does not converge, such a statement seems to be in contradiction with well known proofs of convergence for the RL methods that we use [1]. However, there is no contradiction because it is always possible to use a stronger convergence criterion to improve learning. In other words, the resources needed to converge at small Fd/k_BT increase, and in practical applications the difficulty of learning in this regime therefore has to be faced.

d. Transferability of learning from low temperatures. Remarkably, one can circumvent the drop of efficiency at high temperature by learning at a lower temperature. Indeed, policies learned at $k_BT \ll Fd$ can exhibit efficiencies up to 0.4

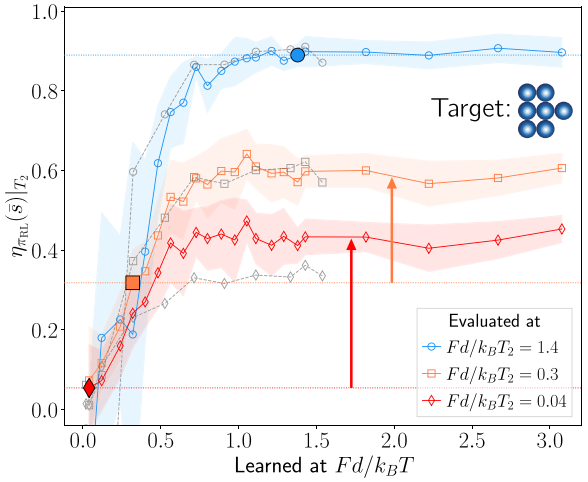

FIG. 3. Transfer learning for a seven-particle target. For each curve, the value of Fd/J is fixed: $Fd/J = 0.4$ (continuous lines) and $Fd/J = 0.2$ (dashed gray lines). The efficiency $\eta_{\pi_{\rm RL}}(\bar{s})|_{T_2}$ is evaluated with DP at a fixed temperature T_2 with policies $\pi_{\rm RL}$ learned at various temperatures T with the same value of Fd/J. For each point, we perform ten independent learning runs. The average efficiency $\eta_{\pi_{\rm RL}}(\bar{s})|_{T_2}$ as a function of Fd/k_BT is shown with empty symbols. The width of the shaded area represents the standard deviation. The efficiency increases and reaches a plateau for low learning temperatures, roughly corresponding to $Fd/k_BT > 1$. The larger full symbols correspond to the average efficiency when learning and evaluation are done at the same temperature $T = T_2$. The arrows represent the gain of efficiency when learning at low temperatures $Fd/k_BT > 1$.

when evaluated at a higher temperature T_2 such that $k_BT_2 \gg Fd$, while the efficiency would be vanishingly small if the policy was learned at T_2. In Fig. 3, the efficiency of a policy $\pi_{\rm RL}$ learned at a temperature T and evaluated at a temperature T_2 for different forces F is seen to depend only weakly on J/k_BT (similar figures for first passage times from various targets are shown in Fig. S5 of the SM [55]). For $Fd/k_BT > 1$, the efficiency saturates to a maximum value. Note that learning at $Fd/k_BT > 1$ means that the efficiency of learning is good, i.e., that the learned policy is close to the optimal policy at the learning temperature. Such a transferability of the learned policy therefore suggests transferability of the optimal policy when varying temperature.

However, the policy changes in a finite fraction of the states when the temperature is varied, as shown in details in Sec. V of the SM [55]. To elucidate this apparent contradiction, we propose that this transferability is caused by the tendency of the optimal policy close to the target to exhibit both a larger policy robustness and a larger relevance for the efficiency. The larger relevance is due to the fact that one necessarily has to go through these states close to the target to reach the target. The larger robustness is caused by the fact that the optimal choice of the force tends to be always directed toward the target in states that are close to the target, with a small sensitivity to the values of the transition rates. In Sec. V of the SM [55], we provide some numerical results following hints from an expansion at small Fd/k_BT that support these assumptions. The optimal policies at different temperatures are similar in states that are most relevant for the efficiency, and these states are on average closer to the target (in terms of distance measured by the ring index [57]). Since the optimal actions in states that are close to the target are similar at high and low temperatures, one can learn them at low temperatures where RL performs better, and transfer them to higher temperatures. Such transferability was unexpected because cluster fluctuation dynamics exhibit different regimes as temperature is varied [58–60].

e. Discussion. Our main result is that reinforcement learning becomes inefficient at high temperatures, small length scales, and small forces. Moreover, since the states that are close to the target exhibit both higher robustness and larger relevance, transfer learning is possible from conditions where learning can be achieved, such as lower temperatures.

The feasibility of some experimental RL control of colloidal clusters was demonstrated in Ref. [19], focusing on relaxation of large clusters (of a few hundreds of particles) toward a circular shape. Despite major differences with our work—we consider arbitrary target shapes and few-particle clusters—we speculate that better learning at low temperatures and transferability of low-temperature policies to high temperatures could be observed in this type of experimental system.

Furthermore, navigation policies for directed or active colloids were investigated experimentally in Ref. [9]. Two-dimensional continuum positions were discretized with a coarse-grained square lattice. In such settings, the coarse-grained lattice parameter should play the role of d, and we again expect the efficiency to drop as Fd/k_BT. In Sec. VI of the SM [55], a similar drop is recovered from the analysis of a simplified one-dimensional continuum model. Learned policies were also found to be transferable from one temperature to another in Ref. [9]. We propose that this is caused by the persistence of the optimal policy in the most relevant states. However, observations become more difficult at high temperatures due to faster particle motion. Disentangling the observation-induced failure of the learning process and the intrinsic difficulty of learning at high temperatures is an open fundamental challenge that would require further theoretical developments.

Moreover, when noise is not too strong and when the systems or the learning algorithms are complex enough, noise is not always detrimental and can help exploration or can regularize the learning process [30,31,61]. However, our results, which have been checked here for MCL and QL, only rely on the use of an ε-greedy policy based on the q function in the limit of strong thermal noise. We therefore expect them to be robust and relevant for a wider range of RL algorithms based on the q function, including advanced RL algorithms that are coupled to artificial neural networks, such as deep Q-learning [2,62]. Hence, our study on elementary RL methods provides directions for better performances and transfer learning strategies [63,64] for advanced RL algorithms applied to nanoscale systems.

The authors wish to thank Y. Benamara and L. Matignon for useful discussions, and D. Rodney for comments on the manuscript.

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (The MIT Press, Cambridge, MA, 2018).

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, arXiv:1312.5602.

[3] J. Kober, J. A. Bagnell, and J. Peters, Int. J. Robot. Res. **32**, 1238 (2013).

[4] W. Zhao, J. P. Queralta, and T. Westerlund, in *Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (IEEE, Canberra, ACT, Australia, 2020), pp. 737–744.

[5] P. Garnier, J. Viquerat, J. Rabault, A. Larcher, A. Kuhnle, and E. Hachem, Computers & Fluids **225**, 104973 (2021).

[6] G. Beintema, A. Corbetta, L. Biferale, and F. Toschi, J. Turbul. **21**, 585 (2020).

[7] S. H. Singh, F. van Breugel, R. P. Rao, and B. W. Brunton, Nat. Mach. Intell. **5**, 58 (2023).

[8] M. Nasiri and B. Liebchen, New J. Phys. **24**, 073042 (2022).

[9] S. Muiños-Landin, A. Fischer, V. Holubec, and F. Cichos, Sci. Rob. **6**, eabd9285 (2021).

[10] F. Cichos, K. Gustavsson, B. Mehlig, and G. Volpe, Nat. Mach. Intell. **2**, 94 (2020).

[11] Y. Yang, M. A. Bevan, and B. Li, Adv. Inte. Systems **2**, 1900106 (2020).

[12] G. Novati, L. Mahadevan, and P. Koumoutsakos, Phys. Rev. Fluids **4**, 093902 (2019).

[13] E. Schneider and H. Stark, Europhys. Lett. **127**, 64003 (2019).

[14] G. Reddy, J. Wong-Ng, A. Celani, T. J. Sejnowski, and M. Vergassola, Nature (London) **562**, 236 (2018).

[15] G. Reddy, A. Celani, T. J. Sejnowski, and M. Vergassola, Proc. Natl. Acad. Sci. USA **113**, E4877 (2016).

[16] S. Colabrese, K. Gustavsson, A. Celani, and L. Biferale, Phys. Rev. Lett. **118**, 158004 (2017).

[17] M. Durve, F. Peruani, and A. Celani, Phys. Rev. E **102**, 012601 (2020).

[18] Z. Young and H. M. La, Int. J. Adv. Rob. Syst. **17**, 1729881420960342 (2020).

[19] J. Zhang, J. Yang, Y. Zhang, and M. A. Bevan, Sci. Adv. **6**, eabd6716 (2020).

[20] S. P. Fletcher, F. Dumur, M. M. Pollard, and B. L. Feringa, Science **310**, 80 (2005).

[21] T. E. Mallouk and A. Sen, Sci. Am. **300**, 72 (2009).

[22] L. Shao and M. Käll, Adv. Funct. Mater. **28**, 1706272 (2018).

[23] B. Robertson, M.-J. Huang, J.-X. Chen, and R. Kapral, Acc. Chem. Res. **51**, 2355 (2018).

[24] A. Nilim and L. El Ghaoui, Oper. Res. **53**, 780 (2005).

[25] G. N. Iyengar, Math. Oper. Res. **30**, 257 (2005).

[26] J. Wang, Y. Liu, and B. Li, Proc. AAAI Conf. Artif. Intell. **34**, 6202 (2020).

[27] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, Annu. Rev. Control Robot. Auton. Syst. **5**, 411 (2022).

[28] S. Dridi and L. Lehmann, Animal Behaviour **104**, 87 (2015).

[29] J. Hao, T. Yang, H. Tang, C. Bai, J. Liu, Z. Meng, P. Liu, and Z. Wang, IEEE Trans. Neural Netw. Learn. Syst. **35**, 8762 (2024).

[30] W. Barfuss and J. M. Meylahn, Sci. Rep. **13**, 1309 (2023).

[31] M. Arjovsky and L. Bottou, arXiv:1701.04862.

[32] D. Lanzoni, O. Pierre-Louis, and F. Montalenti, J. Chem. Phys. **159**, 144109 (2023).

[33] F. Zhang, Z. Li, Y. Duan, A. Abbas, R. Mundaca-Uribe, L. Yin, H. Luan, W. Gao, R. H. Fang, L. Zhang, and J. Wang, Sci. Rob. **7**, eabo4160 (2022).

[34] J. Li, I. Rozen, and J. Wang, ACS Nano **10**, 5619 (2016).

[35] S. Li, Q. Jiang, S. Liu, Y. Zhang, Y. Tian, C. Song, J. Wang, Y. Zou, G. J. Anderson, J.-Y. Han *et al.*, Nat. Biotechnol. **36**, 258 (2018).

[36] M. Zarei and M. Zarei, Small **14**, 1800912 (2018).

[37] M. A. Larchenko, P. Osinenko, G. Yaremenko, and V. V. Palyulin, IEEE Access **9**, 159349 (2021).

[38] F. Boccardo and O. Pierre-Louis, Phys. Rev. Lett. **128**, 256102 (2022).

[39] P. Kuhn, J. Krug, F. Hausser, and A. Voigt, Phys. Rev. Lett. **94**, 166105 (2005).

[40] M. Mahadevan and R. M. Bradley, Phys. Rev. B **59**, 11037 (1999).

[41] O. Pierre-Louis and T. L. Einstein, Phys. Rev. B **62**, 13697 (2000).

[42] S. Curiotto, F. Leroy, P. Müller, F. Cheynis, M. Michailov, A. El-Barraj, and B. Ranguelov, J. Cryst. Growth **520**, 42 (2019).

[43] P. McCormack, F. Han, and Z. Yan, J. Phys. Chem. Lett. **9**, 545 (2018).

[44] Y. Xue, D. J. Beltran-Villegas, X. Tang, M. A. Bevan, and M. A. Grover, IEEE Trans. Control Syst. Technol. **22**, 1956 (2014).

[45] Y. Xue, D. J. Beltran-Villegas, M. A. Bevan, and M. A. Grover, in *Proceedings of the 2013 American Control Conference* (IEEE, Washington, DC, 2013), pp. 3397–3402.

[46] A. G. Banerjee, A. Pomerance, W. Losert, and S. K. Gupta, IEEE Trans. Auto. Sci. Eng. **7**, 218 (2009).

[47] S. Glasstone, K. J. Laidler, and H. Eyring, *The Theory of Rate Processes: The Kinetics of Chemical Reactions, Viscosity, Diffusion and Electrochemical Phenomena*, International chemical series (McGraw-Hill Book Company, Inc., New York, 1941).

[48] D.-J. Liu and J. D. Weeks, Phys. Rev. B **57**, 14891 (1998).

[49] M. Giesen, Prog. Surf. Sci. **68**, 1 (2001).

[50] C. Tao, W. G. Cullen, and E. D. Williams, Science **328**, 736 (2010).

[51] B. C. Hubartt and J. G. Amar, J. Chem. Phys. **142**, 024709 (2015).

[52] F. Boccardo, Y. Benamara, and O. Pierre-Louis, Phys. Rev. E **106**, 024120 (2022).

[53] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).

[54] R. E. Bellman, *Dynamic Programming* (Dover Publications, Inc., USA, 2003).

[55] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.110.L023301 for Figs. S1 and S2 (Markov Decision Processes graphs for five and seven particles); Fig. S3 (return time with MCL and QL for different targets); Fig. S4 (first passage time efficiency); Fig. S5 (transfer learning for first passage times); RL algorithms; definitions for Markov Decision Processes; drop of efficiency at small F/k_BT (expansion and drop of efficiency for an ϵ-greedy policy); transfer learning (expansion around the optimal policy and linear policy interpolation); and expansion of effective transition rates for small Fd/k_BT in a coarse-grained continuum model and which also includes Refs. [65–67].

[56] F. Boccardo and O. Pierre-Louis, J. Stat. Mech.: Theory Exp. (2022) 103205.

[57] A. Baronchelli and V. Loreto, Phys. Rev. E **73**, 026103 (2006).
[58] S. V. Khare, N. C. Bartelt, and T. L. Einstein, Phys. Rev. Lett. **75**, 2148 (1995).
[59] O. Pierre-Louis, Phys. Rev. Lett. **87**, 106104 (2001).
[60] M. Giesen-Seibert, F. Schmitz, R. Jentjens, and H. Ibach, Surf. Sci. **329**, 47 (1995).
[61] M. Igl, K. Ciosek, Y. Li, S. Tschiatschek, C. Zhang, S. Devlin, and K. Hofmann, Adv. Neural Info. Proc. Syst. **32**, 13956 (2019).
[62] H. van Hasselt, A. Guez, and D. Silver, Proc. AAAI Conf. Art. Intell. **30**, 2094 (2016).
[63] M. E. Taylor and P. Stone, J. Machine Learn. Res. **10**, 1633 (2009).
[64] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, Proc. IEEE **109**, 43 (2020).
[65] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, The Netherlands, 1992).
[66] M. Kac, Bull. Am. Math. Soc. **53**, 1002 (1947).
[67] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2017), Vol. 30, pp. 5947–5956.