Run-and-tumble chemotaxis using reinforcement learning

Ramesh Pramanik,[1] Shradha Mishra,[2] and Sakuntala Chatterjee[1]
[1]*Physics of Complex Systems, S.N. Bose National Centre for Basic Sciences, Salt Lake Sector 3, Kolkata 700106, India*
[2]*Department of Physics, Indian Institute of Technology (BHU), Varanasi 221005, India*

(Received 30 August 2024; accepted 10 December 2024; published 3 January 2025)

Bacterial cells use run-and-tumble motion to climb up attractant concentration gradient in their environment. By extending the uphill runs and shortening the downhill runs the cells migrate towards the higher attractant zones. Motivated by this, we formulate a reinforcement learning (RL) algorithm where an agent moves in one dimension in the presence of an attractant gradient. The agent can perform two actions: either persistent motion in the same direction or reversal of direction. We assign costs for these actions based on the recent history of the agent's trajectory. We ask the question: which RL strategy works best in different types of attractant environment. We quantify efficiency of the RL strategy by the ability of the agent (a) to localize in the favorable zones after large times, and (b) to learn about its complete environment. Depending on the attractant profile and the initial condition, we find an optimum balance is needed between exploration and exploitation to ensure the most efficient performance.

DOI: 10.1103/PhysRevE.111.014106

I. INTRODUCTION

Reinforcement learning (RL) is a branch of machine learning [1–6], where an agent makes a choice to execute suitable actions to maximize its reward, while learning from its past [7]. In previous studies RL has been used in many areas, such as game theory [8], operations research, information theory [9], statistical physics [7], etc. In recent years, RL has been successfully used to understand and control different aspects of many equilibrium and nonequilibrium systems. Recently in [10,11] RL is used to predict the nucleation and phase transition in two-dimensional Ising model. The results obtained from RL were in good match with the predictions from classical nucleation theory and from exact results. The learning strategy were also used in nonequilibrium systems like directed percolation in one and two dimensions [12] and RL proved to be very successful in predicting the critical points. In [13] an efficient RL strategy is used in active flow control and recently RL is also used to navigate microswimmers to swim in response to hydrodynamic flow, shear flow, and shear gradient [14]. It was found that swimmers use their instantaneous orientation direction to obtain the optimal policy to perform efficient swimming in different environments.

In this work, we use reinforcement learning to study a model that has been motivated by the phenomenon of bacterial chemotaxis [15–17]. Certain bacteria like *E.coli, S.typhimurium, B.subtilis* are known to show chemotaxis where they can move along a chemical gradient in their environment [18–20]. When these microorganisms experience concentration gradient of an attractant chemical in their surroundings, they show a tendency to migrate towards regions of higher attractant concentration [21–24]. This migration happens via run-and-tumble motion, which is characterized by persistent movement along a particular direction (run), punctuated by abrupt change of direction (tumble) [25,26]. In a homogeneous attractant environment, after a large number of runs and tumbles the net displacement of the cell is zero. But in presence of an attractant concentration gradient, runs in the favorable direction are extended and those in the opposite direction are shortened, giving rise to a chemotactic drift [27–31].

Among all types of microorganisms which show chemotaxis, *E.coli* is the most well-studied one [32–37]. The signaling network inside an *E.coli* cell senses the attractant concentration gradient in the extra-cellular environment and regulates the run-and-tumble motion of the cell [38–41]. The detailed structure of the network has been studied both experimentally and by using theoretical models [42–47]. The spatial variation of attractant chemical across the cell body is negligible due to small size of the cell [48]. Instead, *E.coli* senses the spatial gradient by comparing the attractant level experienced in the recent past and distant past along its run-and-tumble trajectory [49–52]. If in the recent past attractant level is higher than that in the distant past, it generally means the cell is running up the concentration gradient. In that case, the cell tends to run for longer. If on the other hand, recent attractant level is lower than what the cell has experienced earlier, it tends to tumble quickly and chooses a new random direction to run. This mechanism allows the cell to accumulate in the favorable region [53–56]. After a long enough time has passed, the cell density becomes large in the region with high concentration of attractant.

Motivated by this, in the present paper, we formulate a reinforcement learning (RL) strategy where an agent performs run-and-tumble motion in an environment with inhomogeneous concentration of attractant. For simplicity, we consider one spatial dimension here. The agent can either persist moving in the same direction, or can reverse its direction. We define a cost matrix which assigns a cost to each of these two actions, depending on the recent history of the agent's trajectory. With a small probability ϵ the agent "explores" its surroundings by performing a random action (persist or

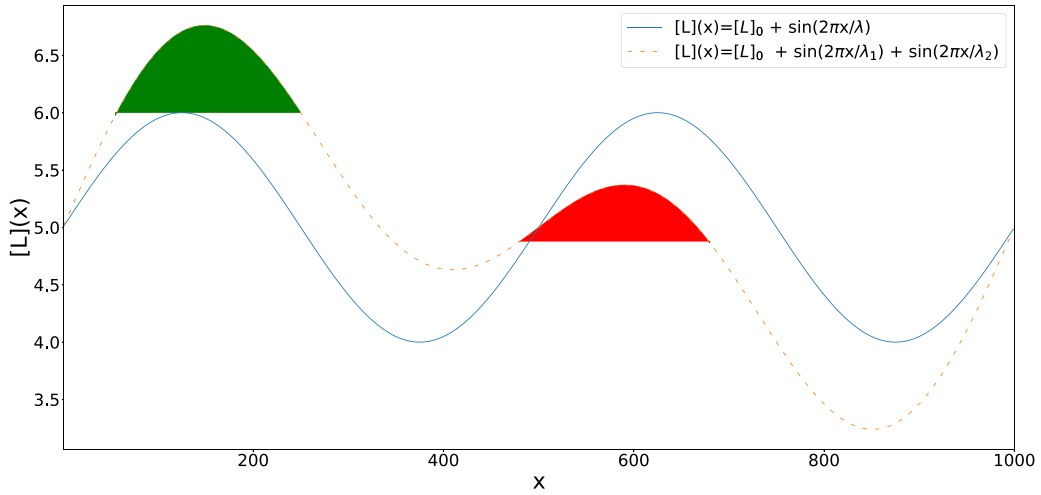

FIG. 1. Attractant concentration profile $[L](x)$. Solid line corresponds to $[L](x) = [L]_0 + \sin(2\pi x/\lambda)$ and dashed line corresponds to $[L](x) = [L]_0 + \sin(2\pi x/\lambda_1) + \sin(2\pi x/\lambda_2)$. Here, $[L]_0$ represents the background concentration, which also makes sure $[L](x)$ never becomes negative. Unless mentioned otherwise, we throughout use $[L]_0 = 5$, $\lambda = \lambda_1 = 500$, $\lambda_2 = 1000$, system size $L = 1000$ with periodic boundary conditions.

reverse) irrespective of its cost, and with the remaining probability $(1 - \epsilon)$, the agent "exploits" its previous learning experience to decide its next action. The agent uses "Q-learning" method to learn and optimize its action based on its experience. Q-learning employs a nonsupervised learning method and it is a simple version of RL which is based on optimizing a value function with respect to a given environment [6].

We are interested in the long time behavior of the agent. In particular, we measure the effectiveness of the RL algorithm by evaluating the performance of the agent at large times. We use different performance criteria like how strongly the agent is able to localize in the high attractant zones, or how quickly it is able to find the region where the attractant concentration is highest, etc. The efficiency of the RL strategy can also be measured from how well the agent has learnt about its environment and whether this learning is used in its long time behavior. If the agent is found to behave based on only partial knowledge of its environment even after a long time has passed, then the RL strategy is deemed rather inefficient.

We consider two different types of attractant concentration profiles. In one case all the concentration peaks are of the same height, and in the other case they have unequal heights. We use a sine function, and superposition of two sine functions to represent these two cases (also see Fig. 1) but our conclusions do not depend on the specific functional forms used. We consider a uniform initial condition, where the initial position of the agent is equally likely to be anywhere in the system, and a nonuniform initial condition where the agent starts from the vicinity of one particular peak of the attractant profile. For an attractant profile whose peak heights are equal, a uniform initial condition gives rise to a long time performance that gets better with less exploration (smaller ϵ) and higher learning rate. However, for a nonuniform initial condition, a competition between exploration and exploitation ensues, which yields optimum ranges for exploration parameter and learning rate at which the RL strategy is most successful. When the attractant profile consists of peaks of different heights, even a uniform initial condition leads to optimal values for exploration and learning parameters. For a nonuniform initial condition when the agent starts near the lower attractant peak, we find that even after a long time, the agent is not able to preferentially localize itself near the higher attractant peak, unless exploration and learning parameters are chosen from an optimal range. For optimal choices of these two parameters, the long time behavior of the agent takes into account its full environment. Outside this optimal range, RL algorithm works inefficiently and the performance of the agent dips. We also measure the mean first passage time of the agent starting from the lower attractant peak to the higher peak. The mean first passage time shows a minimum for specific values of exploration and learning parameters. For these values the agent acts as the most efficient searcher and finds the higher peak in the shortest possible time.

In Sec. II, we describe the model in details and explain the RL algorithm. In Sec. III, we present our results for the sine wave attractant profile and in Sec. IV, we take up superposition of two sine waves. We include some concluding remarks in Sec. V.

II. FORMULATION OF RL ALGORITHM

We consider a one dimensional system of size L with periodic boundary condition across which a spatially varying attractant concentration profile $[L](x)$ is set up. A successful RL strategy should allow the agent to efficiently localize near the maxima of $[L](x)$. Here, we mainly consider two scenarios: one in which $[L](x)$ has multiple maxima of the same height, and the other in which different maxima of $[L](x)$ have different heights. We choose simple functional forms, a sine wave and superposition of two sine waves to represent the above two scenarios, although our results and arguments for these two cases remain valid for any profile with equal and unequal peaks, respectively. In Fig. 1, we plot these two functions. For the sake of comparison we also consider a flat attractant profile, where $[L](x)$ is constant for all x.

TABLE I. Table for model parameters.

Symbol	Value
L	1000
Δ_1	1
Δ_2	2
p_0	0.9
α	Range [0, 0.5]
ϵ	Range [0, 1]
λ	500
λ_1	500
λ_2	1000

A particle (or agent) can be in two different states, high or low, depending on the local attractant level. If the attractant concentration at the location of the agent is equal to or higher than the background level $[L]_0$, the agent is assigned a state "high," and if the local concentration is lower than $[L]_0$, the state is "low." The agent moves on the one dimensional system, either leftward or rightward, with fixed velocity v. In a time step dt its displacement is vdt. This sets the lattice spacing in our system and we use $|v|dt$ as the unit of length. Time is measured in units of dt. During a time step, the agent can perform two actions: persist or reverse, i.e., it can either retain its current run direction, or it can start running in the opposite direction. Below we describe how to calculate the cost for each of these actions.

In line with the strategy employed by the bacteria during chemotaxis, in our model, the agent calculates the cost for each action based on its recent history. By comparing the attractant concentration experienced in the recent past and distant past, the agent assigns a cost for persisting in the current direction and reversing its direction. Define $x_1 \equiv x(t-\Delta_1)$ as the position of the agent at a time Δ_1 back in the past, and similarly, $x_2 \equiv x(t-\Delta_2)$ with $\Delta_2 > \Delta_1$. If $[L](x_1) \geqslant [L](x_2)$ then we assign a cost 0 for persisting and cost 1 for reversing. On the other hand, if $[L](x_1) < [L](x_2)$, the cost for persisting is 1 and cost for reversing is 0. This method of assigning cost is consistent with the fact that in bacterial chemotaxis the runs in favorable directions are extended and those in unfavorable directions are shortened. Unless mentioned otherwise, throughout this work we use $\Delta_1 = 1$ and $\Delta_2 = 2$, i.e., the agent compares the attractant encountered between last two time steps (see Sec. V also for a detailed discussion on this choice). Our method of cost calculation also shows that the background concentration $[L]_0$ does not affect anything, since we are only interested in the difference between $[L](x_1)$ and $[L](x_2)$. In Table I, we show the relevant parameters.

In any run-tumble motion, the probability to tumble in a given time step remains much less than the probability to keep running. When these two probabilities become equal, it ceases to be a run-tumble motion and becomes ordinary diffusion instead. Consistent with this, we assume that at every time step, the agent persists in the same direction with a probability $p_0 \gg 0.5$ and with the remaining probability $(1-p_0)$ it employs the RL algorithm to decide its course of action.

In RL algorithm with Q-learning, the information about previous learning experience is encoded in the Q-matrix whose row and column indices correspond to different states and actions, respectively. For our model with two possible states (high/low) and two possible actions (persist/reverse), the Q-matrix is 2×2 dimensional, with rows (columns) representing the states (actions). $S = 1(2)$ corresponds to high (low) state and $A = 1(2)$ represent persist (reverse) action. At each time-step the agent at state S estimates the cost of both actions $c[S, A]$ and estimates the Q matrix elements corresponding to the row S according to the formula

$$Q_{t+1}[S, A] \leftarrow Q_t[S, A] + \alpha(c[S, A] - Q_t[S, A]) \qquad (1)$$

with $A = 1, 2$, and the parameter α is called the learning rate. The agent follows the ϵ-greedy algorithm, where with probability $(1-\epsilon)$ the agent exploits the previous learning experience and selects the action for which the Q-matrix element is minimum. With probability ϵ the agent explores its environment by choosing an action at random, irrespective of its cost. If at time t the agent persists in the same direction, then $Q_{t+1}(S, 1)$ is updated according to Eq. (1) and $Q_{t+1}(S, 2)$ value reverts back to $Q_t(S, 2)$, since that action was not executed. Similarly, if the agent reverses its direction at time t, then $Q_{t+1}(S, 2)$ is updated and $Q_{t+1}(S, 1)$ remains same as $Q_t(S, 1)$. We start with the initial condition that at $t = 0$ all elements of the Q-matrix are zero. Below we illustrate the algorithm with a specific example.

(1) Suppose at time t the agent is found at position $x(t)$ with velocity v which can be ± 1. With probability p_0 the agent persists in the same direction, and $x(t+1) = x(t) + v$. No cost calculation is done in this case.

(2) With probability $(1-p_0)$ the agent performs exploitation or exploration. Let us consider the specific example, where the agent is in the high state, $S = 1$ and that it had been going uphill for some time in the past, such that $[L](x_1) > [L](x_2)$. Then the cost $c[1, 1]$ for persisting is 0 and the cost for reversal $c[1, 2] = 1$. The agent computes $Q_{t+1}(1, 1)$ and $Q_{t+1}(1, 2)$ according to Eq. (1).

(3) With probability $(1-\epsilon)$ the agent performs exploitation. If $Q_{t+1}(1, 1) \leqslant Q_{t+1}(1, 2)$, the agent persists in the same direction with probability $(1-p_0)$. If $Q_{t+1}(1, 1) > Q_{t+1}(1, 2)$, then the agent reverses its direction with probability $(1-p_0)$.

(4) With probability ϵ the agent performs exploration. In this case, it chooses one of the two actions randomly with probability $\frac{1-p_0}{2}$ for each.

(5) If persistence is chosen (either in exploitation or in exploration), the position is updated as $x(t+1) = x(t) + v$, and $Q_{t+1}(1, 1)$ is updated, while $Q_{t+1}(1, 2)$ is reassigned its old value $Q_t(1, 2)$. If reversal is chosen, then $x(t+1) = x(t) - v$, and $Q_{t+1}[1, 2]$ is updated but $Q_{t+1}[1, 1]$ goes back to $Q_t[1, 1]$.

In a similar manner, the steps for other cases, like $[L](x_1) \leqslant [L](x_2)$ or $S = 2$ can be outlined.

To test the above model, we consider the simple case of a spatially homogeneous attractant environment. According to our definition, the system can only be in $S = 1$ in this case, and can never visit $S = 2$. Since $[L](x_1) = [L](x_2)$ at all times, persistence (reversal) cost remains 0(1) always. If we start with the initial condition that all Q-matrix elements are zero, then they remain zero at all later times. For any other choice of initial Q, it is easy to see that $Q_t[1, 1]$ and $Q_t[1, 2]$ vanish for

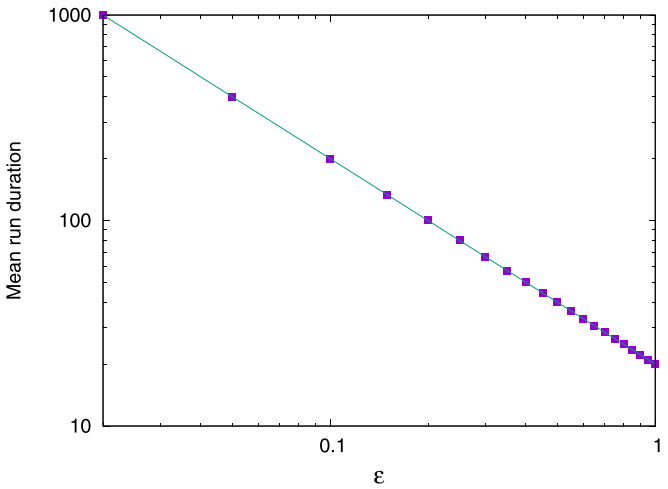

FIG. 2. The average run duration in a homogeneous medium is inversely proportional to ϵ. The discrete points show simulation data, and the continuous line shows analytical expression. Here we have used $p_0 = 0.9$.

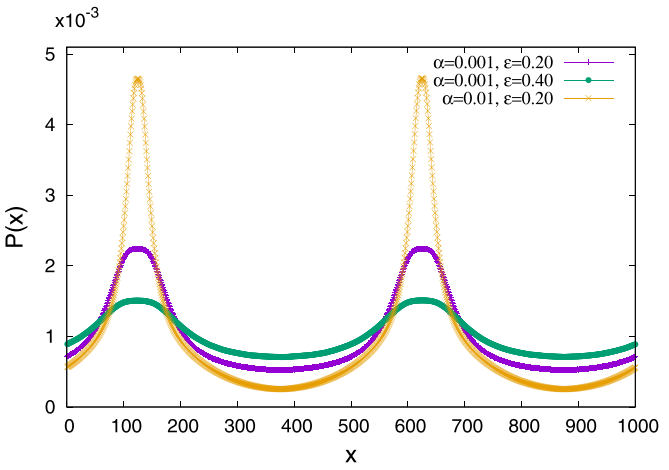

FIG. 3. The position distribution of the agent after a long time in a sine wave attractant profile. The agent localizes more strongly near the two attractant peaks as ϵ decreases and/or α increases. Here we have used uniform initial condition and $\Delta_1 = 1$, $\Delta_2 = 2$, and $p_0 = 0.90$.

large t, while $Q_t[2, 1]$ and $Q_t[2, 2]$ never change with time, since that state is never visited.

From our algorithm described above, it follows that during exploitation the agent can never tumble in this case. Velocity reversal is possible if and only if the agent chooses reversal during exploration. The tumbling probability then becomes $\frac{\epsilon(1-p_0)}{2}$. The average number of steps between two successive tumbling events, or the mean run duration is then $\frac{2}{\epsilon(1-p_0)}$. Therefore, in a homogeneous environment, we retrieve the simple run-tumble motion with constant tumbling rate, as expected. We explicitly compare our calculation with simulation results, in Fig. 2.

III. PERFORMANCE OF RL AGENT IN SINE WAVE ATTRACTANT ENVIRONMENT

In this section, we present our results for the case when the attractant concentration shows a sinusoidal variation in space, as shown in Fig. 1 (solid line). In Fig. 3, we plot the position distribution $P(x)$ of the agent in the long time limit. We find that RL algorithm can successfully generate chemotactic behavior in this system, i.e., the particle is more likely to be found near the maxima of the sine curve. However, depending on the exploration parameter and learning rate, interesting variations are observed.

A. Shortest run for a specific exploration parameter

In Fig. 4, we plot the mean run duration τ as a function of ϵ and find a minimum. Our data also show that τ for $\epsilon = 0$ and 1 have similar values, particularly when Δ_1 and Δ_2 are small (purple squares). For $\epsilon = 1$, the agent only explores and never exploits. At every time step, the agent persists in the same direction with probability $\frac{1+p_0}{2}$ and reverses its direction with probability $\frac{1-p_0}{2}$, irrespective of the local attractant environment. In this case, its behavior is same as in the homogeneous case. τ therefore has the value $\frac{2}{1-p_0}$. For our choice of $p_0 = 0.9$ we find very good agreement with the simulation data.

The $\epsilon = 0$ limit on the other hand, means that the agent never explores and only exploits at all times. When both Δ_1 and Δ_2 are smaller than τ, then during an uphill run, reversal has cost 1 and persistence has cost 0. The agent in this case persists with probability 1 until it reaches the point where attractant concentration is maximum. Beyond this point, the run continues as downhill run. Although the cost for persisting now becomes 1 and reversal has no cost, the agent can still persist with probability $(1 - p_0)$. The run terminates when the agent has crossed an average number of $\frac{1}{1-p_0}$ steps after crossing the peak. A new run starts in the opposite direction, this time going uphill with zero tumbling probability, and after crossing the attractant peak, the run again terminates after an average duration of $\frac{1}{1-p_0}$. This means in the steady state

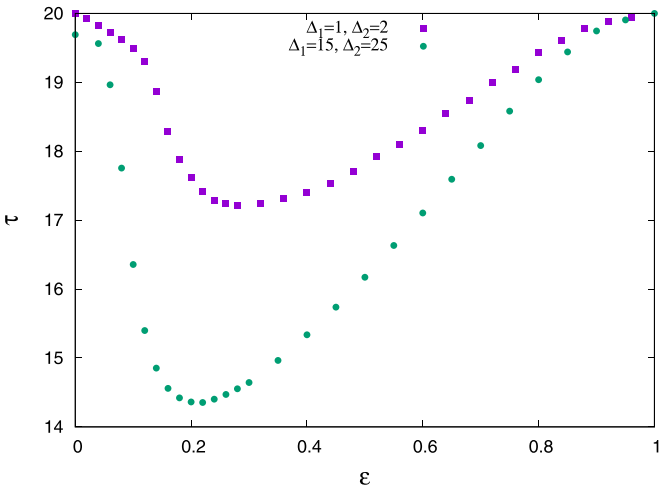

FIG. 4. Mean run duration τ shows a minimum with ϵ. These data are for sine wave attractant profile with $p_0 = 0.90$ and $\alpha = 0.001$.

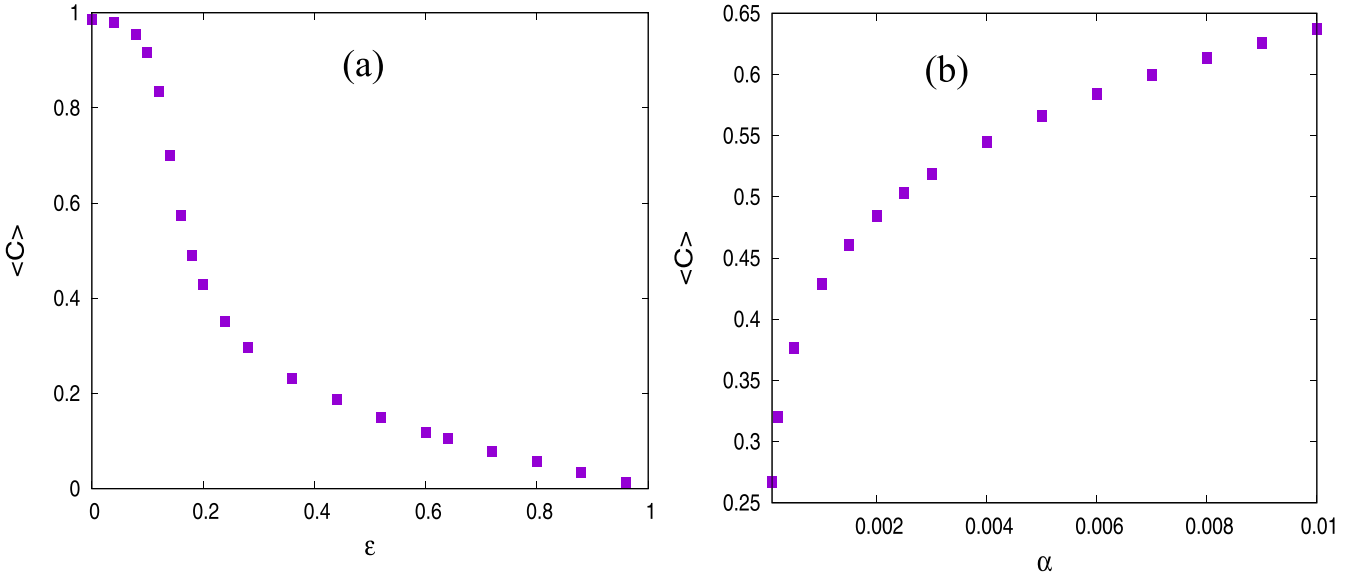

FIG. 5. Uptake $\langle C\rangle$ decreases with ϵ and increases with α. In panel (a) we have used $\alpha = 0.001$ and in (b) we have $\epsilon = 0.20$. All other simulation parameters are as in Fig. 3.

for $\epsilon = 0$ the agent hovers around the peak of the attractant concentration profile and the mean separation between two successive tumbles, i.e., $\tau = \frac{2}{1-p_0}$, same as we found for $\epsilon = 1$. This is consistent with our data in Fig. 4 (purple squares).

Note that equal values for $\epsilon = 0$ and 1 means there has to be some maximum or minimum in between. It is possible to argue that τ must decrease with ϵ when ϵ is small but nonzero. In this case, because of exploration, the agent has a finite tumbling probability $\frac{(1-p_0)\epsilon}{2}$ during an uphill run. During a downhill run the tumbling probability is $(1-p_0)(1-\frac{\epsilon}{2})$. For a simple run-and-tumble motion with these tumbling rates in two directions, one can easily calculate the mean run duration and show that it decreases with ϵ. Although in our present system the dynamics involves additional steps like Q matrix calculation, etc., the qualitative trend for small ϵ is still captured by the above simplified description. This explains the minimum of τ as a function of ϵ. In this argument, we assumed small values of Δ_1 and Δ_2, such that most of the time the comparison between recent and distant past is performed within the same run. When at least one of Δ_1 and Δ_2 becomes comparable to τ, above argument does not work very well. In that case, we find (Fig. 4 green circles) some deviation of the observed run duration from the analytical prediction $\frac{2}{1-p_0}$.

B. Exploration-exploitation competition affects performance

To quantitatively evaluate the performance of the agent, we define the following quantity, known as "uptake" [30,51]:

$$\langle C\rangle = \int_0^L dx P(x)\{[L](x) - [L]_0\}. \tag{2}$$

Large uptake means in the long time limit the agent is more likely to be found in regions with high concentration of attractant, which indicates good performance. The maximum possible value of uptake is reached when the agent is localized at the peak of the attractant profile with all probability. It follows from Eq. (2)) that the uptake in this limit is given by $[L](x)|_{\max} - [L]_0$. In the case of a sine wave attractant profile which we consider here, largest possible uptake is 1. Similarly, the lowest uptake is obtained when the agent localizes at the minimum of $[L](x)$ with all probability. In this case, the lowest uptake is -1. A negative uptake generally indicates a chemo-repellent environment (e.g., some toxic chemical) from which the agent prefers to stay away. When the uptake is zero, it means the agent is completely insensitive to the concentration gradient present in its environment and is equally likely to be found anywhere in the system.

In Figs. 5(a) and 5(b), we plot the variation of uptake $\langle C\rangle$ as a function of the exploration parameter ϵ and learning parameter α, respectively. Here, we have considered a uniform initial condition, *i.e.* the initial position of the agent is chosen from a uniform distribution. We find that for a fixed α (ϵ) uptake decreases (increases) with ϵ (α). These data are consistent with our plots for $P(x)$ shown in Fig. 3. Both the Figs. 3 and 5 show that increasing the learning rate α and/or decreasing the exploration parameter ϵ helps the agent to strongly localize near the peaks of $[L](x)$. However, we show below that for a different choice of initial condition, conclusions can be different.

Note that larger α or smaller ϵ inhibit the agent from sampling the whole environment. The agent remains confined near the peak of $[L](x)$. This can significantly affect its performance if the initial position of the agent is in the vicinity of one particular peak of $[L](x)$ (instead of being uniformly distributed over the whole system). In that case, the agent is not able to visit the other peaks of $[L](x)$ when α is too large or ϵ is too small. This can have a potentially adverse effect on the agent's performance when $[L](x)$ has peaks of different heights. The agent may get stuck near a local maximum and may not be able to reach the highest peak of $[L](x)$ even after a long time has elapsed. We explicitly consider such a case in Sec. IV. For our present choice of a sinusoidal form of

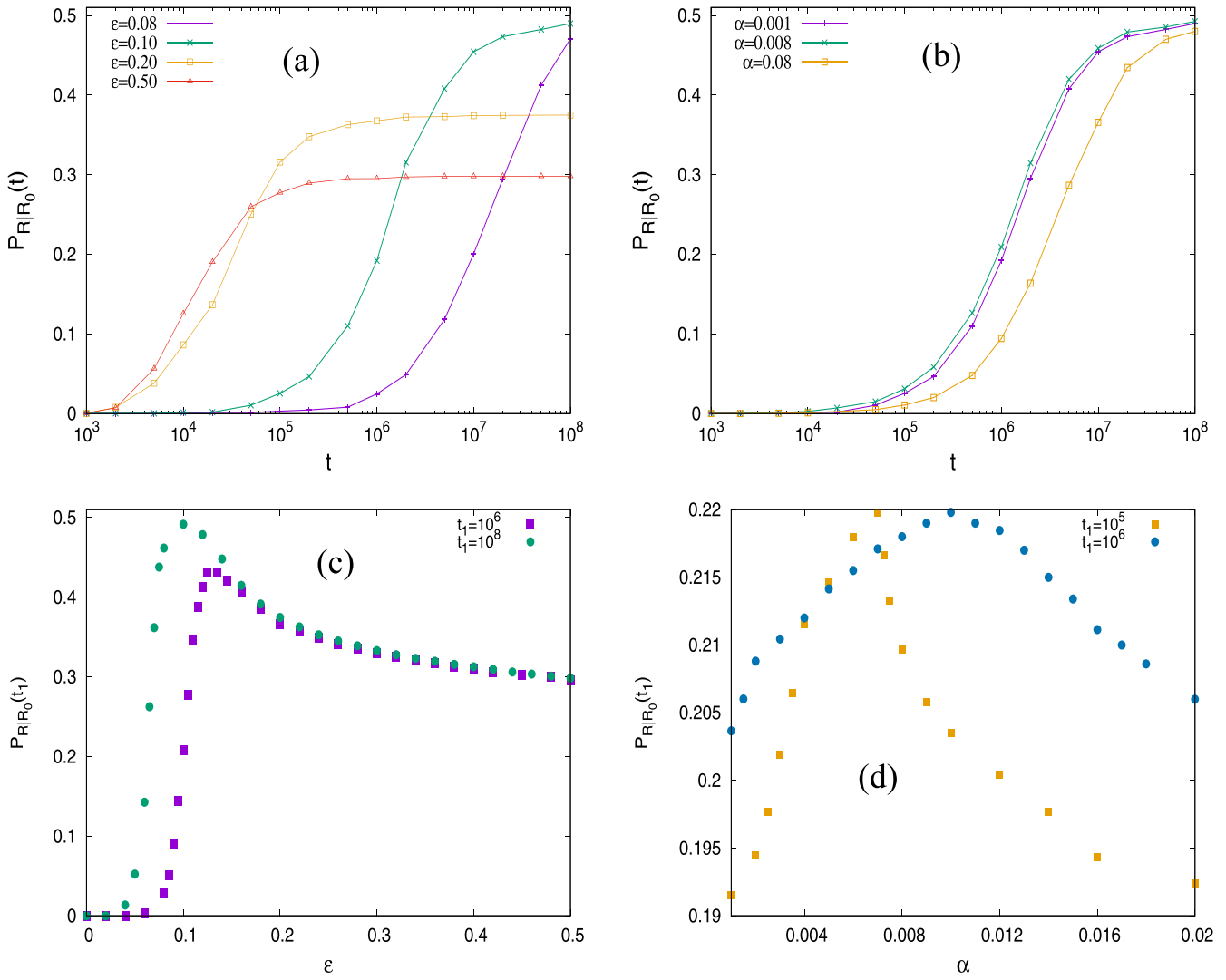

FIG. 6. Starting from the neighborhood of one attractant peak, the probability to find the agent near the other peak at large times. (a) $\mathcal{P}_{R|R_0}(t)$ increases with t and saturates for large ϵ but as ϵ becomes small, due to trapping effect $\mathcal{P}_{R|R_0}(t)$ increases very slowly. Here, α is held fixed at 0.001 (b) Similar effect observed as ϵ is held fixed at 0.1 and α is varied. But α-dependence is relatively weaker. (c) $\mathcal{P}_{R|R_0}(t = t_1)$ after a large time t_1 shows a peak as a function of ϵ. Peak shifts to smaller ϵ for larger t_1. (d) $\mathcal{P}_{R|R_0}(t = t_1)$ also shows a peak with α. For larger t_1 peak position shifts towards smaller α values. We have rescaled the data for $t_1 = 10^5$ by a constant factor 6.77 in order to show both the plots in the same scale. In (a) and (c), we have used $\alpha = 0.001$ and in (b) and (d), $\epsilon = 0.1$. Other simulation parameters are as in Fig. 3.

$[L](x)$, all the peaks are of same height. Therefore, even if the agent gets stuck near one peak, it will not affect a quantity like uptake.

To capture the trapping effect, therefore, we define another performance criterion, which measures if all the favorable regions in the environment has been sampled by the agent. Starting from the vicinity of one peak of $[L](x)$, we measure the probability to find the agent in the vicinity of another peak, as a function of time. More specifically, we choose the initial condition where the agent is equally likely to be located in a region R_0 with initial position lying in the range $0 < x < \frac{\lambda}{2}$. As a function of time we measure the probability to find the agent in the region R which is in the neighborhood of the next peak with $\lambda < x < \frac{3\lambda}{2}$. We denote this probability as $\mathcal{P}_{R|R_0}(t)$, which is expected to grow with time t and then saturate as steady state is reached. However, as seen in our data in Figs. 6(a) and 6(b) the saturation can be logarithmically slow with time, specially when ϵ (α) is small (large). It is expected, since going from one peak to another, the agent needs to cross a region where $[L](x)$ is quite small. This means the agent has to execute long downhill runs which get increasingly difficult for small (large) ϵ (α). For a logarithmically slow relaxation, it is often not feasible to evaluate the performance of the agent based on its steady state properties. Rather, the behavior of $\mathcal{P}_{R|R_0}(t)$ for large times (but not large enough to reach steady state) can be studied. If it is found that for a certain large value of $t = t_1$, the probability $\mathcal{P}_{R|R_0}(t_1)$ has low value, then that means the agent has not yet learned about its full environment, although a long time has passed. This indicates inefficient learning. On the other hand, a large value of $\mathcal{P}_{R|R_0}(t_1)$ means that the agent managed to learn about its environment and determined its preference for a particular

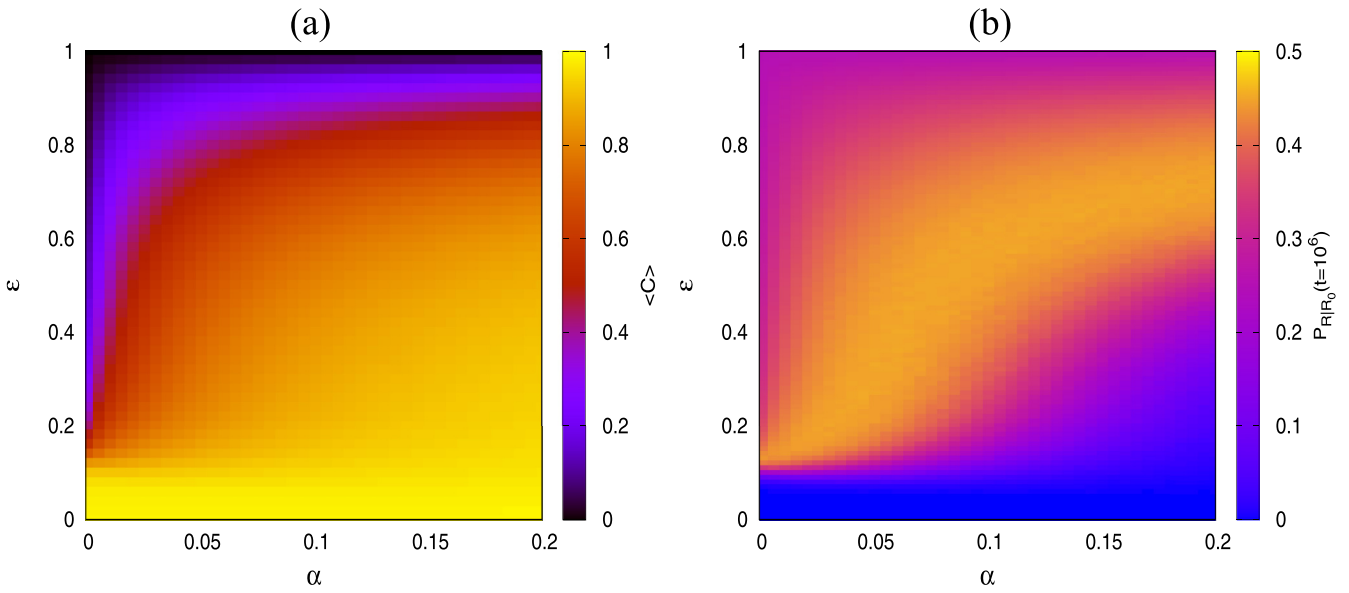

FIG. 7. Heat map of uptake $\langle C\rangle$ and $\mathcal{P}_{R|R_0}(t=t_1)$ in the ϵ-α plane. Here we have used $t_1 = 10^6$. While uptake is highest in the small-ϵ and large-α corner, $\mathcal{P}_{R|R_0}(t=t_1)$ reaches its maximum value for an optimal ϵ-α zone. Other simulation parameters are as in Fig. 3.

location based on this knowledge. The RL strategy in this case is working well.

In Fig. 6(a), we plot $\mathcal{P}_{R|R_0}(t)$ versus t for different ϵ values and a fixed α. As discussed in the previous paragraph, we find for small ϵ the probability increases almost logarithmically slowly with time but reaches high value eventually, whereas for larger ϵ the probability quickly saturates but the saturation value is lower. In Fig. 6(c), we plot $\mathcal{P}_{R|R_0}(t=t_1)$ as a function of ϵ for fixed α and show that the probability has a maximum value at a specific ϵ. In other words, the performance measured after a certain large time t_1 is best at a specific ϵ value. As t_1 increases, the best performance is obtained at lower ϵ. This trend is consistent with our expectation that eventually, for $t_1 \to \infty$ $\mathcal{P}_{R|R_0}(t=t_1)$ will not depend on t_1 anymore and will saturate with time, with the highest saturation value obtained for $\epsilon = 0$.

We find a similar qualitative behavior when we keep ϵ fixed and measure $\mathcal{P}_{R|R_0}(t)$ for different values of α. For small α, the probability saturates faster but the saturation value is low. When α is large, $\mathcal{P}_{R|R_0}(t)$ increases logarithmically slowly and ultimately saturates to a large value for very large times. When we measure $\mathcal{P}_{R|R_0}(t=t_1)$ for a large t_1, we find a peak at a specific α, as shown in Fig. 6(d). However, the α dependence is rather weak, unlike the pronounced variation against ϵ shown by $\mathcal{P}_{R|R_0}(t)$. In Fig. 7, we present the heat map of two performance criteria, uptake $\langle C\rangle$ (for uniform initial condition) and $\mathcal{P}_{R|R_0}(t=t_1)$ (for initial position in the vicinity of one maximum of $[L](x)$) in the ϵ-α plane. While uptake is maximum for smallest ϵ and largest α considered, $\mathcal{P}_{R|R_0}(t=t_1)$ shows a peak for intermediate values of ϵ and α.

IV. ATTRACTANT PROFILE WITH DIFFERENT PEAK HEIGHTS

In this section, we consider an attractant concentration profile with two peaks of different heights, as shown by the dashed line in Fig. 1. The functional form we consider here is given by $[L](x) = [L]_0 + \sin\frac{2\pi x}{\lambda} + \sin\frac{\pi x}{\lambda}$. We consider two different initial conditions, where the initial position of the agent is chosen from (a) a uniform distribution over the entire system, and (b) a uniform distribution in the vicinity of the lower peak (see Fig. 1, region shaded by red). Like the previous section, we are interested in the long time performance of the agent for different choices of ϵ and α. Due to the presence of high and low attractant peaks, the competition between exploration and exploitation is more clearly visible here. Even when we start from a uniform initial condition, for low ϵ and/or high α, the agent may get trapped in the lower attractant peak. Uniform initial condition makes sure that the agent is equally likely to start from the vicinity of the lower and higher attractant peaks. For those trajectories which start

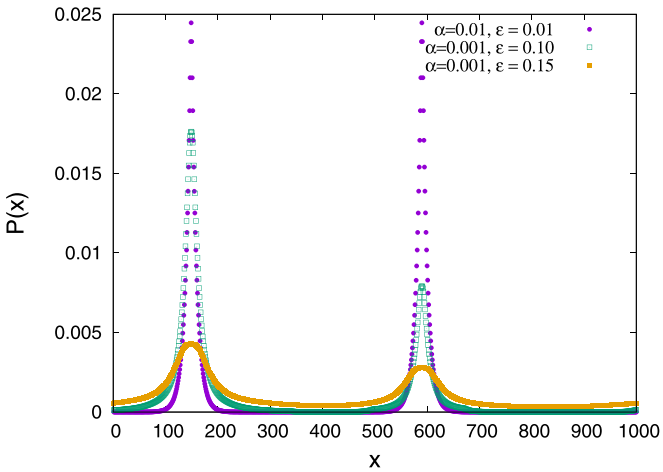

FIG. 8. Position distribution measured after time $t = 10^6$ starting from a uniform initial condition. The attractant profile has peaks of unequal heights here. But for small ϵ and large α the agent shows almost equal affinity for both peaks, even after a long time. Here we have used $p_0 = 0.9$, $\Delta_1 = 1$, and $\Delta_2 = 2$.

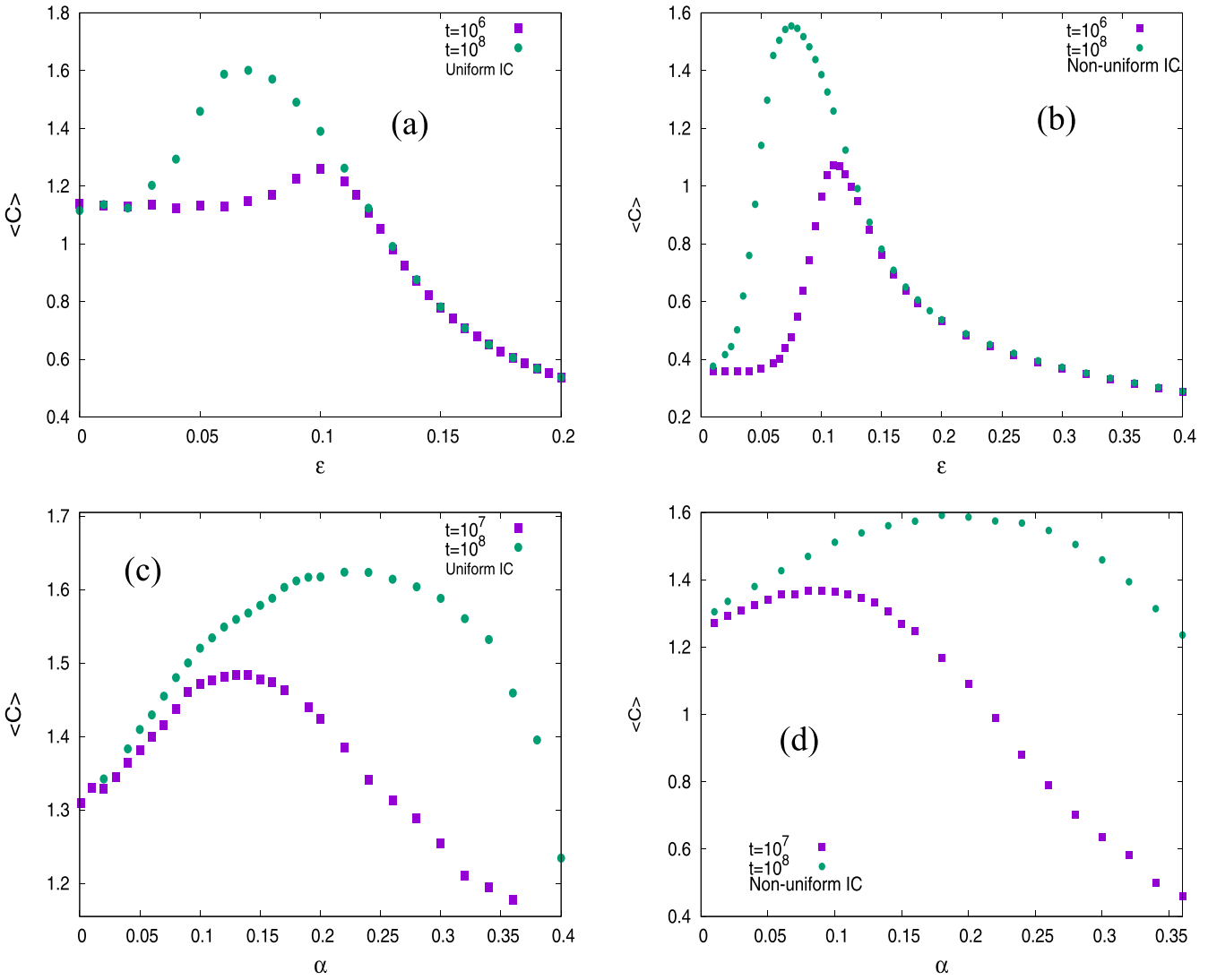

FIG. 9. Uptake $\langle C \rangle$ measured after a long time t starting from uniform and nonuniform initial conditions. As a function of ϵ and α the uptake shows a peak. For larger t, the peak shifts to smaller ϵ and larger α. In (a) and (b), we have used $\alpha = 0.001$ and in (c) and (d), we have $\epsilon = 0.1$. Other simulation parameters are as in Fig. 8.

near the lower (higher) peak, the agent quickly localizes itself at the lower (higher) peak, and remains trapped there for long time when ϵ (α) is small (large). Thus even after a long time the agent is not able to learn about its complete environment and shows roughly the same preference for both peaks. Our data in Fig. 8 show this trend clearly.

A. Performance peak for optimal ϵ and α

In Fig. 9, we plot the value of uptake, measured after long times, for different choices of ϵ, α and starting with both types of initial conditions. In Fig. 9(a) uptake at two different times are plotted as a function of ϵ, and for a fixed α. These data are for uniform initial condition. As explained above, even after a long time, almost half of the population remains trapped near the lower peak of the attractant, when ϵ is small. This yields low value of uptake. On the other hand, when ϵ is too large, the exploitation loses out to exploration and the agent becomes less sensitive to the attractant gradient which decreases uptake. For intermediate ϵ uptake shows a peak. When a longer time has passed, a fraction of the population which was trapped near the lower attractant peak, gets a chance to find the higher peak. The uptake peak thus shifts to lower ϵ values. Figure 9(b) shows the data for a nonuniform initial condition (where the agent starts near the lower attractant peak, shown by the region shaded in red in Fig. 1). While the qualitative behavior remains similar, the uptake peaks at different times are much more pronounced here. This is not surprising since the trapping effect is now felt by the entire population. Figures 9(c) and 9(d) show the plots for uptake measured at different times, as a function of α and for a fixed ϵ, for uniform and nonuniform initial conditions, respectively. In this case, for large α effect of confinement is felt more strongly and uptake decreases. For very small α, learning happens at a negligible rate, and the agent's preference for high attractant regions is low. Uptake is small in this limit too. For an intermediate α uptake has a peak and with time that peak shifts rightward towards larger α values.

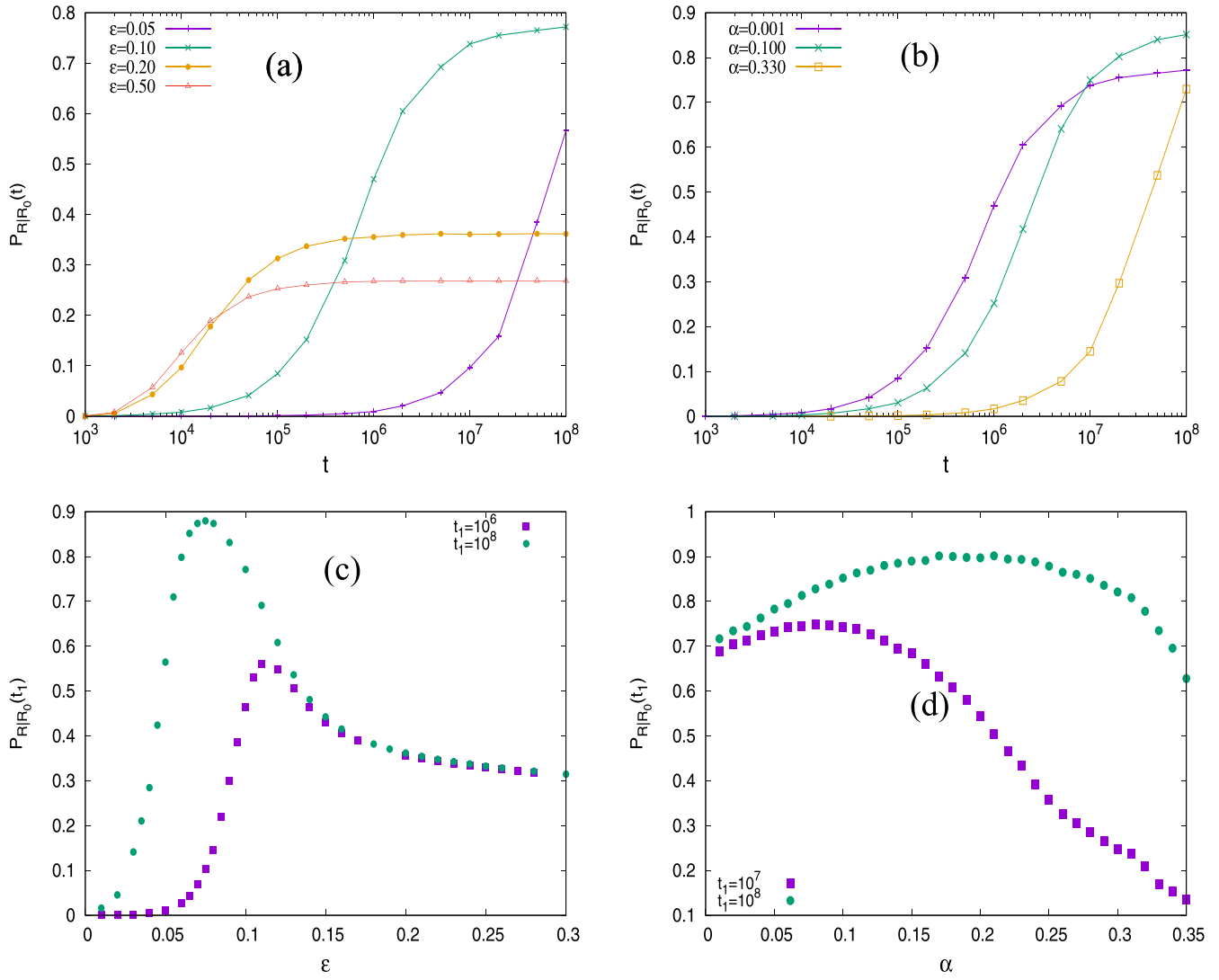

FIG. 10. Starting from a nonuniform initial condition, $\mathcal{P}_{R|R_0}(t)$ is measured for different ϵ and α. (a) $\mathcal{P}_{R|R_0}(t)$ increases with t and saturates quickly when ϵ is large. However, as ϵ decreases the saturation comes at a later time and saturation value is also higher. Finally, for very small ϵ, $\mathcal{P}_{R|R_0}(t)$ increases logarithmically slowly with time. (b) Similar trend is observed as ϵ is held fixed and α is increased. (c) $\mathcal{P}_{R|R_0}(t = t_1)$ shows a peak as a function of ϵ and the peak position shifts leftward for larger t_1. (d) The peak of $\mathcal{P}_{R|R_0}(t = t_1)$ in α similarly shifts rightward towards larger α values when t_1 is increased. In (a) an (c), we have $\alpha = 0.001$ and in (b) and (d), we have $\epsilon = 0.1$.

We also measure the conditional probability $\mathcal{P}_{R|R_0}(t)$, as defined in Sec. III, for different ϵ-α value. Here, R_0 represents the region shaded by red in Fig. 1 and the region R is marked by green shade. The agent starts in the neighborhood of the lower peak of the attractant concentration and we measure the probability to find it in the vicinity of the higher peak as a function of time. The effect of trapping is much more pronounced in this case. While for large ϵ or small α, Figs. 10(a) and 10(c) show that the probability $\mathcal{P}_{R|R_0}(t)$ increases with time and saturates quickly at a lower value, as ϵ decreases and/or α increases, $\mathcal{P}_{R|R_0}(t)$ shows a much slower growth with time, due to long trapping time of the agent at the lower peak. Figure 10(b) plots the value of $\mathcal{P}_{R|R_0}(t = t_1)$ at a fixed time t_1 as a function of ϵ and we find a peak. As t_1 increases, the peak shifts to lower ϵ values. Figure 10(d) similarly shows $\mathcal{P}_{R|R_0}(t = t_1)$ as a function of α and in this case peak shifts to larger α values. These trends are consistent with our data in Fig. 6.

In Fig. 11, we show, in the ϵ-α plane, the variation of uptake and $\mathcal{P}_{R|R_0}(t)$, measured after a long time. We find both these quantities reach a maximum value for an optimal range of ϵ and α. This shows that the RL strategy employed in this system works best in this range.

B. First passage time from lower peak to higher peak of attractant profile

To further probe the role of exploitation-exploration competition on the performance of the RL agent, we measure its mean first passage time from the lower attractant peak to the higher one. Starting from a local maximum, how long the agent needs to find the global maximum, is an important quantity to characterize its efficiency. For small ϵ the agent is not able to explore its surroundings and remains trapped near the local maximum. This increases its mean first passage time. On the other hand, for $\epsilon \to 1$, the agent shows little

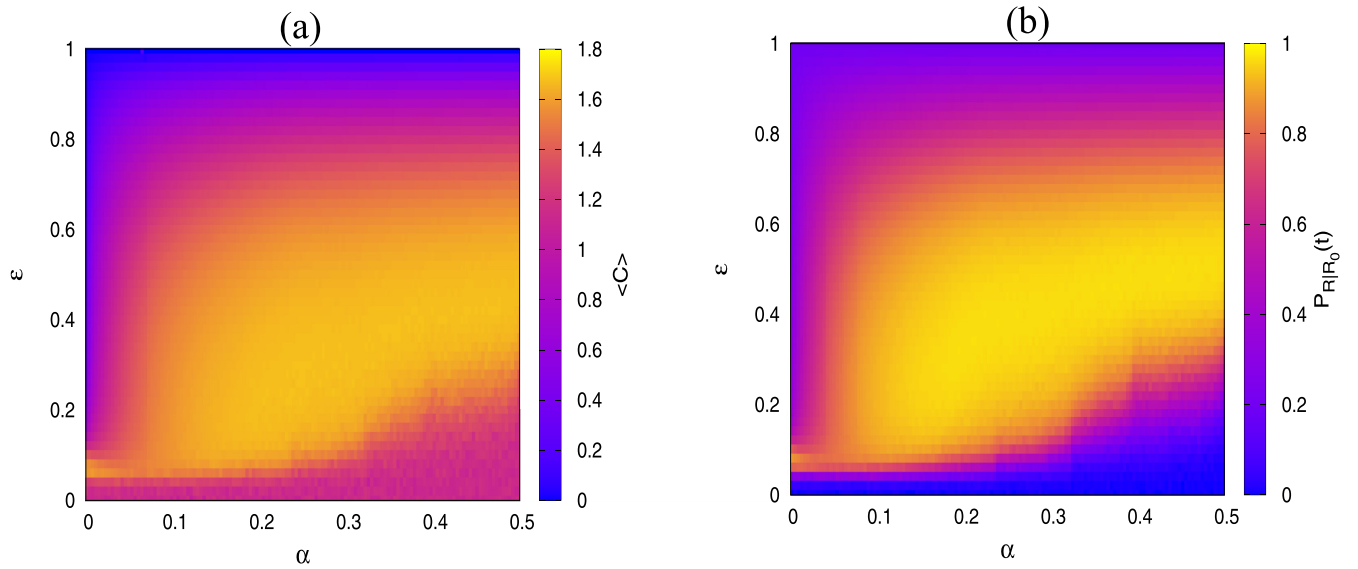

FIG. 11. (a) In the ϵ-α plane, heat map of $\langle C \rangle$ after time $t = t_1$ starting from uniform initial condition. (b) Heat map of $\mathcal{P}_{R|R_0}(t = t_1)$ in the same plane starting with nonuniform initial condition. We have used $t_1 = 10^8$ here. Both plots show an optimum region on ϵ-α plane where RL algorithm is working best.

affinity towards attractant peaks. In this limit, its motion is mainly diffusive which also corresponds to a large value of mean first passage time. For an intermediate ϵ, mean first passage time shows a minimum [Fig. 12(a)]. For a given α, there is an optimal value of ϵ for which the agent performs the quickest search. Figure 12(b) shows the plot of mean first passage time as a function of α, for a fixed ϵ. Even in this case, we find an optimum α corresponding to the quickest search.

V. DISCUSSIONS

In this work, we have considered an RL agent which is exploring its environment via run-and-tumble motion. We are interested in the question: under what condition the RL strategy is most efficient. We quantify efficiency by the probability to find the agent in the attractant-rich region in the long time limit, and also by how much the agent has learnt about its environment. A successful RL strategy allows the agent to quickly learn about its attractant environment and localize in the favorable region. We find depending on the nature of the attractant concentration profile, different RL strategies work best.

In the case when all the concentration peaks are of the same size, then starting from a uniform initial position, the agent is able to localize most strongly near the peak regions when exploration rate is low and learning rate is high. However, when the agent starts from the vicinity of one particular attractant peak, then it remains trapped there and even after a large time has passed, the agent is not able to learn about its complete

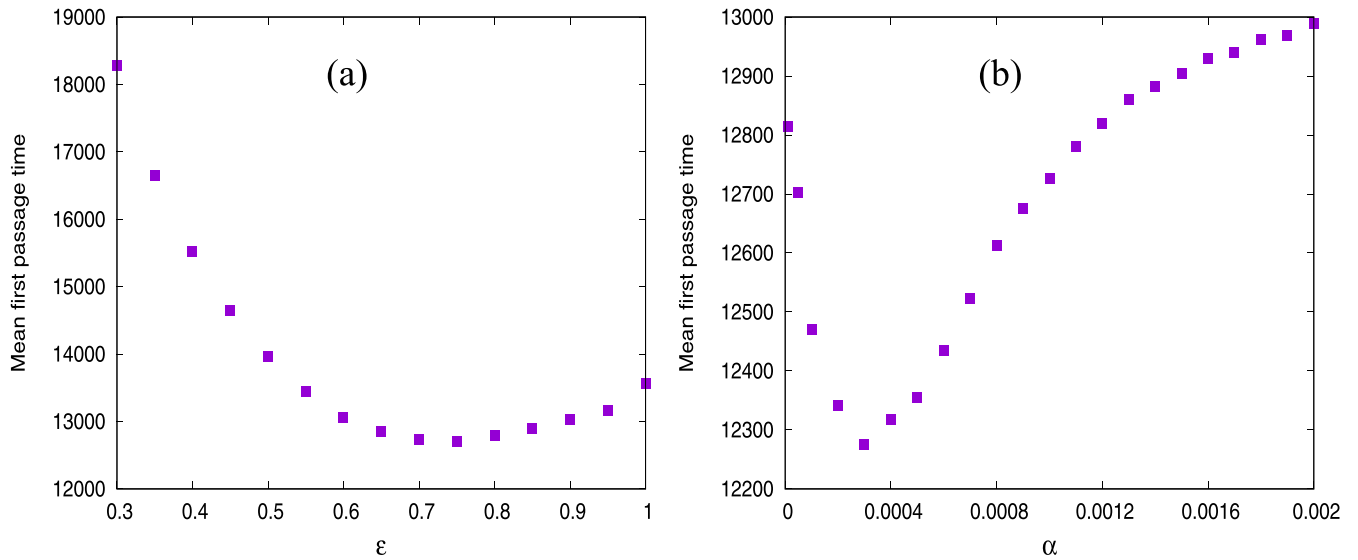

FIG. 12. Mean first passage time starting from the lower peak to the higher peak of the attractant profile. The quantity shows a minimum as a function of both ϵ and α. For (a) α is fixed at 0.001 and for (b) ϵ is fixed at 0.7. All other simulation parameters are as in Fig. 8.

environment if the exploration rate is low or learning rate is high. In this case, an optimum range for these two rates works best for the agent.

When the attractant profile has peaks of different sizes, the exploration-exploitation trade off is even more pronounced. Even when the initial position of the agent is uniformly distributed throughout the system, the agent is not able to learn about its complete environment and gets trapped in nearest attractant peak, for low exploration rate and high learning rate. As a result, even after a long time has passed, the agent may show same affinity for the lower and higher attractant peaks, which is detrimental for its performance. An optimum balance between exploitation and exploration helps the agent overcome the trapping effect and show its best performance. For nonuniform initial condition, when the agent starts near the lower attractant peak, its probability to localize near the higher peak after a long time, shows a maximum for intermediate rates of exploration and learning.

Our simple model has few limitations. For example, there is an important difference between our cost calculation method and the decision making process that an *E.coli* cell follows during chemotaxis. As explained in Sec. II, our RL algorithm just takes into account the sign of the concentration difference experienced in the recent and distant past and does not care about the magnitude of the difference. However, an *E.coli* cell modulates its tumbling rate by taking into account both the sign and magnitude of the difference in concentration experienced in recent and distant past [32]. Most of our results presented here are for the case when the agent uses attractant levels at last two time-steps to read off the local gradient. This is the best choice to read off the sign of the local attractant gradient accurately. However, most of our qualitative conclusions remain valid (data not shown) even for larger values of Δ_1 and Δ_2, as long as they do not exceed typical run duration. For very large Δ_1 and Δ_2, the memory goes too far back in the past. The gradient calculated from $[L](x_1) - [L](x_2)$ may not accurately represent the current environment since the agent may have reversed its direction once or multiple times meanwhile.

Throughout this work, we have considered one spatial dimension. Many experiments measure run-tumble motion of *E.coli* in narrow microfluidic channel whose width is comparable to or smaller than typical run length [46,57–59], and therefore, the motion effectively takes place in one dimension. However, it is an interesting question to ask whether in a more natural setting, where the cell moves in two or three dimensions, our RL description might work. In two dimensions, for example, after each tumble the direction of the cell is randomized. Therefore the effect of a tumble can not be simply described as velocity reversal, as we have done in the present model. In this case, one can classify the actions of the RL agent as "persist" and "randomize," where the former means persisting in the same direction, and the latter means choosing a random direction. During exploration, the agent can either persist or randomize with equal probability. Note that the possibility of randomizing direction after each tumble means that the trapping effect seen in Figs. 6 and 10 is expected to be weaker in two dimensions. Indeed, our preliminary data (not shown here) indicate that even when the exploration parameter ϵ is significantly small, the agent is able to escape a local attractant peak and explore its full environment. This affects the exploration-exploitation competition and exploration starts winning over exploitation at a much lower ϵ value, compared to what we have seen in one dimension. However, we verify (data not shown) that apart from some quantitative differences, all our qualitative conclusions remain valid in higher dimension as well.

Possible future directions include generalizing our simple model where state description can have a finer resolution (instead of only two possible values, high and low), or cost function can be assigned real values instead of 0 and 1, or using a learning parameter which evolves with time, etc. Also, the current study is limited for a single particle in a time-independent environment, the study can be extended for many interacting particles moving in an environment that changes with time. Expanding the model for the case where particles also learn from each other might be interesting to explore in future studies.

ACKNOWLEDGMENTS

R.P. acknowledges the research fellowship [Grant No. 09/0575(13013)/2022-EMR-I] from the Council of Scientific and Industrial Research (CSIR), India. S.M. thanks DST-SERB India, ECR/2017/000659, CRG/2021/006945 and MTR/2021/000438 for financial support. S.M. also thanks S.N. Bose National Centre for Basic Sciences, Kolkata for providing kind hospitality. S.C. acknowledges support from Anusandhan National Research Foundation (ANRF), India (Grant No. CRG/2023/000159).

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT press, Cambridge, Massachusetts, 2018).

[2] L. Panait and S. Luke, Cooperative multi-agent learning: The state of the art, Auton Agent Multi-Agent Syst **11**, 387 (2005).

[3] L. Busoniu, R. Babuska, and B. De Schutter, A comprehensive survey of multiagent reinforcement learning, IEEE Trans. Syst. Man, Cybern. C **38**, 156 (2008).

[4] M. Kubat, *An Introduction to Machine Learning* (Springer, Cham, Switzerland, 2017).

[5] D. E. Goldberg and J. H. Holland, Mach. Learn. **3**, 95 (1988).

[6] C. J. C. H. Watkins and P. Dayan, Q-learning, Mach. Learn. **8**, 279 (1992).

[7] M. Durve, F. Peruani, and A. Celani, Learning to flock through reinforcement, Phys. Rev. E **102**, 012601 (2020).

[8] M. Zinkevich, A. Greenwald, and M. Littman, Cyclic equilibria in markov games, in *Advances in Neural Information Processing Systems*, edited by In Y. Weiss, B. Schölkopf, and J. Platt (MIT Press, 2005), Vol. 18.

[9] I. M. de Abril and R. Kanai, Curiosity-driven reinforcement learning with homeostatic regulation, in *2018 International Joint Conference on Neural Networks, IJCNN 2018*, (IEEE, Brazil, 2018), pp. 1–6.

[10] S. Huang, W. Klein, and H. Gould, Predicting nucleation using machine learning in the ising model, Phys. Rev. E **103**, 033305 (2021).
[11] P. B. Sampat, A. Verma, R. Gupta, and S. Mishra, Ordering through learning in two-dimensional ising spins, Phys. Rev. E **106**, 054149 (2022).
[12] J. Shen, W. Li, S. Deng, and T. Zhang, Supervised and unsupervised learning of directed percolation, Phys. Rev. E **103**, 052140 (2021).
[13] A. Kubo and M. Shimizu, Efficient reinforcement learning with partial observables for fluid flow control, Phys. Rev. E **105**, 065101 (2022).
[14] K. Sankaewtong, J. J. Molina, M. S. Turner, and R. Yamamoto, Learning to swim efficiently in a nonuniform flow field, Phys. Rev. E **107**, 065102 (2023).
[15] M. Eisenbach, *Chemotaxis* (Imperial College Press, London, 2004).
[16] J. Adler, G. L. Hazelbauer, and M. M. Dahl, Chemotaxis toward sugars in *Escherichia coli*, J. Bacteriol. **115**, 824 (1973).
[17] J. Adler, A method for measuring chemotaxis and use of the method to determine optimum conditions for chemotaxis by *Escherichia coli*, Microbiology **74**, 77 (1973).
[18] H. C. Berg, *E. coli in Motion* (Springer-Verlag, New York, 2004).
[19] R. J. Galloway and B. L. Taylor, Histidine starvation and adenosine 5'-triphosphate depletion in chemotaxis of salmonella typhimurium, J. Bacteriol. **144**, 1068 (1980).
[20] M. Sidortsov, Y. Morgenstern, and A. Be'Er, Role of tumbling in bacterial swarming, Phys. Rev. E **96**, 022407 (2017).
[21] R. Colin and V. Sourjik, Emergent properties of bacterial chemotaxis pathway, Curr. Opin. Microbiol. **39**, 24 (2017).
[22] Y. Tu, Quantitative modeling of bacterial chemotaxis: signal amplification and accurate adaptation, Annu. Rev. Biophys. **42**, 337 (2013).
[23] G. Lan and Y. Tu, Information processing in bacteria: memory, computation, and statistical physics: a key issues review, Rep. Prog. Phys. **79**, 052601 (2016).
[24] J. S. Parkinson, G. L. Hazelbauer, and J. J. Falke, Signaling and sensory adaptation in escherichia coli chemoreceptors: 2015 update, Trends Microbiol. **23**, 257 (2015).
[25] H. C. Berg and D. A. Brown, Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking, Nature (London) **239**, 500 (1972).
[26] S. Dev and S. Chatterjee, Run-and-tumble motion with step-like responses to a stochastic input, Phys. Rev. E **99**, 012402 (2019).
[27] P.-G. de Gennes, Chemotaxis: the role of internal delays, Eur. Biophys. J. **33**, 691 (2004).
[28] S. Chatterjee, R. A. da Silveira, and Y. Kafri, Chemotaxis when bacteria remember: drift versus diffusion, PLoS Comput. Biol. **7**, e1002283 (2011).
[29] N. Vladimirov, D. Lebiedz, and V. Sourjik, Predicted auxiliary navigation mechanism of peritrichously flagellated chemotactic bacteria, PLoS Comput. Biol. **6**, e1000717 (2010).
[30] S. Dev and S. Chatterjee, Optimal methylation noise for best chemotactic performance of *E. coli*, Phys. Rev. E **97**, 032420 (2018).
[31] S. D. Mandal and S. Chatterjee, Effect of receptor clustering on chemotactic performance of *E. coli*: Sensing versus adaptation, Phys. Rev. E **103**, L030401 (2021).
[32] S. M. Block, J. E. Segall, and H. C. Berg, Impulse responses in bacterial chemotaxis, Cell **31**, 215 (1982).
[33] J. E. Keymer, R. G. Endres, M. Skoge, Y. Meir, and N. S. Wingreen, Chemosensing in *Escherichia coli*: two regimes of two-state receptors, Proc. Natl. Acad. Sci. USA **103**, 1786 (2006).
[34] T. S. Shimizu, Y. Tu, and H. C. Berg, A modular gradient-sensing network for chemotaxis in *Escherichia coli* revealed by responses to time-varying stimuli, Mol. Syst. Biol. **6**, 382 (2010).
[35] Y. Tu, T. S. Shimizu, and H. C. Berg, Modeling the chemotactic response of *Escherichia coli* to time-varying stimuli, Proc. Natl. Acad. Sci. USA **105**, 14855 (2008).
[36] S. Bi and Victor Sourjik, Stimulus sensing and signal processing in bacterial chemotaxis, Curr. Opin. Microbiol. **45**, 22 (2018).
[37] V. Sourjik and H. C. Berg, Receptor sensitivity in bacterial chemotaxis, Proc. Natl. Acad. Sci. USA **99**, 123 (2002).
[38] U. Alon, M. G. Surette, N. Barkai, and S. Leibler, Robustness in bacterial chemotaxis, Nature (London) **397**, 168 (1999).
[39] N. Barkai and S. Leibler, Robustness in simple biochemical networks, Nature (London) **387**, 913 (1997).
[40] V. Frank, G. E. Piñas, H. Cohen, J. S. Parkinson, and A. Vaknin, Networked chemoreceptors benefit bacterial chemotaxis performance, mBio **7**, e01824-16 (2016).
[41] Y. Meir, V. Jakovljevic, O. Oleksiuk, V. Sourjik, and N. S. Wingreen, Precision and kinetics of adaptation in bacterial chemotaxis, Biophys. J. **99**, 2766 (2010).
[42] D. Bray, M. D. Levin, and C. J. Morton-Firth, Receptor clustering as a cellular mechanism to control sensitivity, Nature (London) **393**, 85 (1998).
[43] J. Liu, B. Hu, D. R. Morado, S. Jani, M. D. Manson, and W. Margolin, Molecular architecture of chemoreceptor arrays revealed by cryoelectron tomography of *Escherichia coli* minicells, Proc. Natl. Acad. Sci. USA **109**, E1481 (2012).
[44] C. H. Hansen, R. G. Endres, and N. S. Wingreen, Chemotaxis in *Escherichia coli*: a molecular model for robust precise adaptation, PLoS Comput. Biol. **4**, e1 (2008).
[45] V. Sourjik and H. C. Berg, Functional interactions between receptors in bacterial chemotaxis, Nature (London) **428**, 437 (2004).
[46] L. Jiang, Q. Ouyang, and Y. Tu, Quantitative modeling of *Escherichia coli* chemotactic motion in environments varying in space and time, PLoS Comput. Biol. **6**, e1000735 (2010).
[47] S. D. Mandal and S. Chatterjee, Effect of receptor cooperativity on methylation dynamics in bacterial chemotaxis with weak and strong gradient, Phys. Rev. E **105**, 014411 (2022).
[48] M. Li and G. L. Hazelbauer, Cellular stoichiometry of the components of the chemotaxis signaling complex, J. Bacteriol. **186**, 3687 (2004).
[49] H. C. Berg and P. M. Tedesco, Transient response to chemotactic stimuli in *Escherichia coli*, Proc. Natl. Acad. Sci. USA **72**, 3235 (1975).
[50] Y. Kafri and R. A. da Silveira, Steady-state chemotaxis in *Escherichia coli*, Phys. Rev. Lett. **100**, 238101 (2008).
[51] A. Celani and M. Vergassola, Bacterial strategies for chemotaxis response, Proc. Natl. Acad. Sci. USA **107**, 1391 (2010).

[52] S. Dev and S. Chatterjee, Optimal search in *E. coli* chemotaxis, Phys. Rev. E **91**, 042714 (2015).

[53] N. Vladimirov, L. Løvdok, D. Lebiedz, and V. Sourjik, Dependence of bacterial chemotaxis on gradient shape and adaptation rate, PLoS Comput. Biol. **4**, e1000242 (2008).

[54] F. Matthäus, M. Jagodič, and J. Dobnikar, *E. coli* superdiffusion and chemotaxis–search strategy, precision, and motility, Biophys. J. **97**, 946 (2009).

[55] J. Long, S. W. Zucker, and T. Emonet, Feedback between motion and sensation provides nonlinear boost in run-and-tumble navigation, PLoS Comput. Biol. **13**, e1005429 (2017).

[56] W. Sun and M. Tang, Macroscopic limits of pathway-based kinetic models for *E. coli* chemotaxis in large gradient environments, Multiscale Model. Simul. **15**, 797 (2017).

[57] S. Samanta, R. Layek, S. Kar, M. K. Raj, S. Mukhopadhyay, and S. Chakraborty, Predicting *Escherichia coli's* chemotactic drift under exponential gradient, Phys. Rev. E **96**, 032409 (2017).

[58] S. Samanta, R. Layek, S. Kar, S. Mukhopadhyay, and S. Chakraborty, Transient drift of escherichia coli under a diffusing step nutrient profile, Phys. Rev. E **98**, 052413 (2018).

[59] Y. V. Kalinin, L. Jiang, Y. Tu, and M. Wu, Logarithmic sensing in *Escherichia coli* bacterial chemotaxis, Biophys. J. **96**, 2439 (2009).