

Optimality Theory of Stigmergic Collective Information Processing by Chemotactic Cells

Masaki Kato^{1,*} and Tetsuya J. Kobayashi^{2,1,3,†}

¹*Department of Mathematical Informatics, The Graduate School of Information Science and Technology,
The University of Tokyo, Tokyo, Japan*

²*Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan*

³*Universal Biology Institute, The University of Tokyo, Tokyo 113-8654, Japan*



(Received 19 February 2025; accepted 25 August 2025; published 15 October 2025)

Collective information processing is fundamental in various biological systems, where the cooperation of multiple cells results in complex functions beyond individual capabilities. A distinctive example is collective exploration in which chemotactic cells locate sparsely distributed targets by sharing information of them through active generation and modulation of endogenous chemical cues. While the optimality of sensing target-derived gradients has been studied extensively in the context of single-cell information processing, the optimality of collective information processing that self-organizes endogenous cue gradient towards the targets remains underexplored. In this study, we employ the reinforcement learning (RL) framework to formalize the collective exploration, in which a population adaptively optimize its navigation toward targets through active cue modulation. Based on RL theory, we derive the optimal exploration dynamics of agents and identify their structural correspondence with the Keller-Segel model, the established phenomenological model of collective cellular dynamics. Our theory identifies an optimal coupling relationship between self-organized gradient generation and sensing and demonstrates that the optimal cue-generation strategy changes qualitatively depending on whether cells sense gradients linearly or logarithmically. The underlying RL structure is leveraged to compare the derived collective dynamics with single-agent searching dynamics, showing that distributed information processing by the population enables a fraction of agents to reach the target robustly. Our formulation provides a foundation for understanding collective exploration mediated by active modulation and sensing of cues.

DOI: [10.1103/tvfy-lbbl](https://doi.org/10.1103/tvfy-lbbl)

I. INTRODUCTION

Collective information processing by a population of cells is a widespread phenomenon throughout biology—from bacterial populations [1–3] to developing tissues [4–6], immune systems [7,8], and metastatic tumors [4,9,10]—where the cooperation of multiple cells leads to complex biological functions beyond individual capabilities. Among others, collective exploration for sparsely distributed targets mediated by active cue modulation is a representative class of information processing. Targets (e.g., invaded pathogens or damaged cells in tissue) to be explored usually emit exogenous chemical signals; however, these signals are often severely localized only around the targets due to their instability and limited diffusivity. To effectively share and amplify the target information, cells actively produce and degrade endogenous chemical cues, thereby forming a dynamic cue field in the environment. Alongside, they navigate towards the targets by sensing the resulting cue gradient [Fig. 1(a)]. Active cue degradation by cells in regions lacking targets enables other cells to recog-

nize the absence of targets, efficiently navigating them toward unexplored areas [11]. In contrast, cue production around a target can amplify the presence and magnitude of the target, recruiting other cells and enhancing their accumulation at the target, even when the target leaves only a faint and localized signal [12]. As a combination of modulation and sensing of cues, a population of *Dictyostelium discoideum* can cooperatively locate the goal of complex mazes [13]. Similar strategies may be employed by eukaryotic cell populations in tissues [11], including immune [14,15] and cancer cells [16,17].

Despite such accumulating experimental insights into collective information processing driven by cue modulation, we still lack the theoretical basis to investigate its functionality and efficiency. Recent studies on cellular information processing of external chemicals have mainly focused on single-cell chemotaxis, particularly bacterial responses to ligand gradients. Their optimality has been verified [18–21] by comparing optimal responses and behaviors derived theoretically with those measured and estimated by experiments [22–25]. Nonetheless, extending those studies on single-cell responses to collective behaviors remains limited. Existing works on collective cellular behaviors investigated only their phenomenological aspects: cell population motility [26–30], impacts of individual traits on population behavior [31,32], and the interplay of population behaviors with growth [33–35]. Only one pioneering work [36] addressed the functionality and efficiency of collective behaviors.

With regard to the efficiency of collective chemotaxis, there are two problems that require insights beyond single-cell

*Contact author: m_kato@sat.t.u-tokyo.ac.jp

†Contact author: tetsuya@mail.crmind.net

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

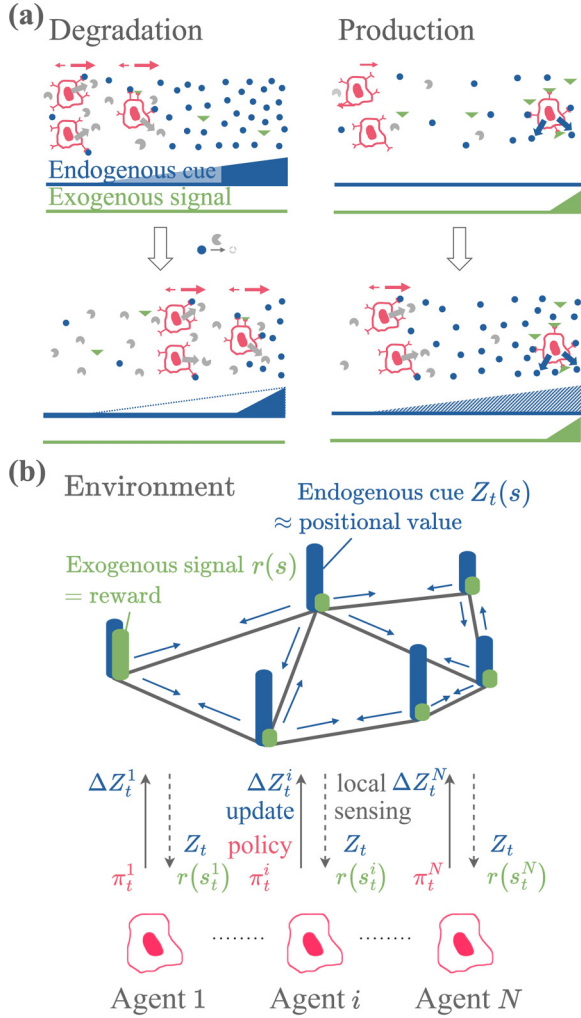


FIG. 1. (a) Schematics of an endogenous cue (attractant) modulation by a migrating cell population. Degradation by enzymes in the absence of an exogenous signal forms a gradient of the endogenous cue, guiding cells toward unexplored areas. Production based on the detected exogenous signal establishes a gradient of the endogenous cue toward the source of the exogenous signal, recruiting cells to the source. (b) A schematic diagram of an agent population which cooperatively learns optimal navigation toward the target locations with higher exogenous signal concentration (reward) r . The optimal state (positional) value V^* for navigation represents the maximum long-term desirability of each location for gaining given the fixed exogenous cue concentration r . The agents learn V^* by associating endogenous cue concentration Z_t with their provisional estimation V_t of V^* ($V_t = \mathcal{F}[Z_t] \approx V^*$). At time t , each agent $i \in \{1, \dots, N\}$ senses surrounding Z_t and local $r(s_t^i)$. Based on the sensed r and Z_t , each agent transits between vertices according to its policy π_t^i and individually modulates local $Z_t(s_t^i)$ by $\Delta Z_t^i = \Delta Z_t(s_t^i)$, where $\Delta Z_t^i < 0$ corresponds to degradation; $\Delta Z_t^i > 0$, to production.

studies. One is the interdependence between the generation of cue concentration field and the responses to the field for efficient navigation. Focusing only on the responses, as in the case of single-cell studies, is insufficient to capture collective chemotaxis; optimal migration and exploration require consistency between shaping the cue field and responding to it, as already suggested in the context of traveling wave

formation [37]. A recent study addressed the optimality of cue-induced collective search using optimal control theory [36]. They derived an optimal gradient sensing to a collectively generated cue field, yet the generation process of the field was incorporated heuristically. Thus, the consistency problem remains unresolved.

The other problem is the benefits and drawbacks of being a collective over being an individual for solving information processing tasks. Some organisms, such as human beings, possess high computational capabilities as individual agents [38], whereas others have low capabilities yet work collectively. Both of them have been maintained evolutionarily. Understanding the difference between individual and collective information processing is fundamental to comprehend the evolvability of biological computation [39,40].

To address these problems in this study, we formulate a cell population that cooperatively improves navigation toward sparsely distributed targets through cue modulation as an agent population that performs reinforcement learning (RL) of optimal navigation in a distributed manner. This formulation allows us to clarify the optimal coupling between modulation and sensing of the cue and the functionality of distributed computation. Specifically, we associate the cue concentration Z_t with the state value V_t in RL (the long-term likelihood of accumulating at the targets) and show that the cooperative cue modulation by cells can be interpreted as a collective learning of the optimal state value [Fig. 1(b)]. The derived dynamics are structurally equivalent to the Keller-Segel model, a widely used mechanical model for explaining experiments of collective chemotaxis, suggesting that cell populations exhibiting these dynamics can engage in distributed learning. Furthermore, we derive two qualitatively different equations for optimal cue modulation, which optimally couples with linear and logarithmic sensing, respectively. Finally, we compare a single smart agent, which possesses the internal memory of entire environment, with a population of nonsmart agents, which has no internal memory yet can learn collectively via an external cue in the environment. We demonstrate that despite individual limitations, the population can achieve accumulation at the target more robustly. Our results provide a quantitative foundation for understanding collective information processing mediated by cue modulation and sensing.

II. MODEL

A. Chemical-mediated collective exploration by chemotactic agents

To formulate an environment exploration by a population of biological agents such as cells, we represent the environment as a connected undirected graph $G = (\mathcal{S}, \mathcal{E})$ with a finite set $\mathcal{S}$ of vertices and a finite set $\mathcal{E} \subset \mathcal{S} \times \mathcal{S}$ of edges, which allows to incorporate general spatial topologies explored by the agent population. In the environment, $N \in \mathbb{Z}_+$ homogeneous and independent agents are located at the vertices of the graph with overlap. The location of agent $i \in \{1, 2, \dots, N\}$ at time $t \in \{0, 1, \dots\}$ is indicated by state $s_t^i \in \mathcal{S}$. Within a small unit time, agent i stochastically chooses either to stay at $s' = s_t^i$ or to move to a neighboring vertex $s' \in \mathcal{N}(s_t^i)$ with probability $\pi_t^i(s'|s_t^i) \in [0, 1]$, where $\mathcal{N}(s)$ is the set of neigh-

boring vertices of s . The transition probability $\pi_t^i : \mathcal{S} \rightarrow \Delta(\mathcal{S})$ is referred to as the agent's policy.

Each agent is assumed to learn a better policy π_t^i to reach and accumulate at targets in the environment through exploration. If the agent is a higher organism with sufficient internal memory, then it can memorize the past experience, update π_t^i , and guide itself individually. However, if the agent is a simple organism such as a cell, then a versatile memory is not necessarily available.

Instead, the policy (stochastic behaviors) of a simple agent is typically controlled and guided by various external chemicals [41]. One of these chemicals is a primary exogenous guidance signal, which conveys information about the targets' positions and magnitudes. We specifically refer to the signal generated by the targets as exogenous to discriminate it from the chemical cues produced by the agents themselves. Bacterial chemotaxis to sugars is an example in which an agent is guided by exogenous signals [42]. We consider a single exogenous signal for simplicity, whose spatial distribution is represented by $r : \mathcal{S} \rightarrow \mathbb{R}_{\geq 0}$. For a biological instance, $r(s) = \exp(-l(s; s^0))$ represents fMLP concentration originating from damaged tissue at s^0 , at distance $l(s; s^0)$ from the tissue [43]. The concentration can also vary depending on the size of the damage [44]. While exogenous signals may be affected by agents, we assume for simplicity that they are not altered by the agents and dependent only on the targets' positions and magnitudes.

Although the exogenous signal is the primary information about the targets, they may decay rapidly away from the targets due to harsh environments and their instabilities, being less likely to spread over long distances. As a result, a cell can detect such a signal only when it reaches the target's vicinity. As such, cell populations often utilize endogenous cues—generated and modulated by individual cells—to amplify the localized target signal and share the strengthened information among the population. For example, neutrophils detect exogenous signals, such as fMLP, from damaged tissue, and they also actively secrete more stable endogenous cues, such as LTB4, packaged in vesicles or exosomes to convey the information to distant immune cells [45–48]. For simplicity, we here consider only one endogenous cue, whose spatial distribution is represented by $Z_t : \mathcal{S} \rightarrow \mathbb{R}_{\geq 0}$. In addition, we exclusively consider the nontrivial case in which the exogenous signal is localized at the targets' position. Without the endogenous cue, each cell can only walk randomly until it reaches the target by chance. However, with the endogenous cue, cells may efficiently navigate towards the target as a population.

Finally, we assume that the policy π_t^i of agent i at vertex s_t^i and time t is solely determined by the concentrations of the exogenous signal $r(s_t^i)$ at s_t^i and the endogenous cue around its vicinity $(Z_t(s))_{s \in \{s_t^i\} \cup \mathcal{N}(s_t^i)}$. This corresponds to the situation where each agent has no internal memory and no direct interaction with the others. Yet, it can guide itself using locally sensed exogenous and endogenous chemicals, the latter effectively working as a shared memory deployed in the environment.

B. Characterization of optimal exploration

Although each agent lacks an internal memory, the agents may still effectively navigate toward targets through explo-

ration if they collectively and optimally modulate and sense the endogenous cue based on the locally detected exogenous signal. To characterize navigation efficiency, we define an optimal policy for accumulating at targets as one that maximizes the gain of the exogenous signal r , since reaching targets coincides with acquiring high concentration r . Focusing only on immediate gain would fail in navigation to sparsely distributed targets and compromises persistent target accumulation [18]. Rather, the agent may integrate the localized signal r over multiple periods, facilitating long-range navigation and amplifying accumulation. As such, we associate the exogenous signal concentration r with the reward signal in RL and formulate the policy to be learned through exploration as the optimal policy $\pi^* = \operatorname{argmax}_{\pi} \mathcal{J}[\pi]$, where $\mathcal{J}[\pi]$ is the expected cumulative reward with discounting and control penalty:

$$\mathcal{J}[\pi] := E \left[\sum_{t=0}^{\infty} \gamma^t R^{\pi}(s_t, s_{t+1}) \right], \quad (1)$$

$$R^{\pi}(s_t, s_{t+1}) = r(s_t) - \frac{1}{\beta} \log \frac{\pi(s_{t+1}|s_t)}{p(s_{t+1}|s_t)}, \quad (2)$$

where the expected value $E[f(s_{0:t})]$ is the path average of $f(s_{0:t})$ on the state sequences $s_0, s_1, \dots, s_t$. The discount factor $\gamma \in [0, 1)$ is an intrinsic parameter common to all agents. This weights more on the current reward than on the future one, thereby compromising transient navigation and steady-state accumulation. It would be natural for simple organisms with limited lifespan to prioritize the present over the distant future. The additional term $\frac{1}{\beta} \log \frac{\pi}{p}$ is a motility control penalty for the cue-dependent policy π to deviate from the intrinsic (stationary) transition probability $p : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$. The intrinsic transition probability would correspond to the random motility spontaneously exhibited by cells without cues. This form of cost was employed as biological control costs in bacterial chemotaxis [21] and immunological learning [49]. Note that the inverse control weight $\beta \in (0, \infty)$ represents the ease of motility control. The formulation in Eq. (1) is equivalent to the entropy-regularized Markov decision process $(\mathcal{S}, r, p, \gamma, \beta)$ [50] (Appendix A b in the Supplemental Material (SM) [51]). Defining the state value function $V^{\pi} : \mathcal{S} \rightarrow \mathbb{R}$ as

$$V^{\pi}(s) := E \left[\sum_{t=0}^{\infty} \gamma^t R^{\pi}(s_t, s_{t+1}) \mid s_0 = s \right], \quad s \in \mathcal{S}, \quad (3)$$

and the optimal state value function $V^* : \mathcal{S} \rightarrow \mathbb{R}$ as $V^*(s) := \max_{\pi} V^{\pi}(s)$ for $s \in \mathcal{S}$, the optimal policy π^* can be explicitly expressed using V^* as follows:

$$\pi^*(s'|s) = \frac{p(s'|s) \exp(\beta \gamma V^*(s'))}{\sum_{s' \in \mathcal{S}} p(s'|s) \exp(\beta \gamma V^*(s'))}, \quad s, s' \in \mathcal{S}, \quad (4)$$

where $V^*(s)$ represents the maximum long-term desirability of vertex s for gaining reward r . The functional form of optimal policy π^* means that the agent should bias its intrinsic transition $p(s'|s)$ to the surrounding vertices with higher $V^*(s')$. This behavior can be interpreted as a biased random walk along the local gradient of $V^*(s')$.

Optimal value V^* is determined by the distribution of exogenous signals and environment geometry; however, both are initially unknown to the agents. The agents have to learn this

optimal V^* by updating a provisional, nonoptimal V_t . If agents encode V_t through the endogenous cue concentration Z_t such that $V_t(s) = \mathcal{F}[Z_t](s)$, then they can update $V_t(s)$ by modulating the endogenous cue via its degradation and production while moving by following the suboptimal policy with V_t in Eq. (4). When $\mathcal{F}[Z_t](s)$ closely approximates $V^*(s)$, agents move to vertices with higher state values, thus approaching the target.

C. Optimal coupling of cue sensing and modulation

The optimal cue modulation is dependent on the actual form of the mapping $\mathcal{F}$, which boils down to the optimal coupling of cue modulation and chemotactic sensing behavior as we demonstrate in this subsection.

A biologically relevant mapping is the one associating the logarithm of the endogenous cue concentration with the estimate of the optimal state value: $V_t(s) = \mathcal{F}[Z_t](s) := \log(Z_t(s))/\beta$ for $s \in \mathcal{S}$. This mapping aligns with Weber Fechner response observed across a wide range of concentrations in various chemotactic cells [23,24,52,53]. Under this logarithmic mapping, the corresponding agents' policy is reduced to

$$s_{t+1}^i \sim \pi_t(\cdot | s_t^i) := \frac{p(\cdot | s_t^i) Z_t^\gamma(\cdot)}{\sum_{s' \in \mathcal{S}} p(s' | s_t^i) Z_t^\gamma(s')}, \quad (5)$$

Meanwhile, the endogenous cue Z_t should be modulated by its degradation and production for learning Z_t which better approximates V^* . We consider the simplest and biologically plausible dynamics of Z_t that approximately minimize the deviation between $V_t = \mathcal{F}[Z_t]$ and V^* under the physical constraint of diffusivity of Z_t (see Appendix B a 1 in the SM [51] for derivation):

$$Z_{t+1}(s) - Z_t(s) = - \sum_{i=1}^N \mathbb{1}_{s_t^i=s} \alpha \Delta Z_t(s_t^i) - D \sum_{s' \in \mathcal{S}} L_{ss'} Z_t(s'), \quad s \in \mathcal{S}, \quad (6a)$$

$$\Delta Z_t(s) := Z_t(s) - \exp(\beta r(s)) \sum_{s' \in \mathcal{S}} p(s' | s) Z_t^\gamma(s'), \quad (6b)$$

where $\mathbb{1}_{s_t^i=s}$ is an indicator function that takes 1 if agent i is at vertex s at time t and 0 otherwise, $\alpha \in (0, 1/N)$ is the speed of the endogenous cue degradation and production by agents, and L is the graph Laplacian of the environment $\mathcal{G}$. In Eq. (6a), each agent at s_t^i individually modulates $Z_t(s_t^i)$ by $\alpha \Delta Z_t(s_t^i)$, which is dependent exponentially on the locally detected r and nonlinearly on surrounding Z_t . In addition, Z_t diffuses to neighbors independently of the agents at a constant rate D [Fig. 1(b)].

This learning dynamics is a distributed [54] and sequentially regularized version of greedy Z-learning [50], which have two advantages from both biological and learning theory perspectives: First, the optimal behavior learned by the dynamics accommodates the cost of agent's motility control as KL divergence from the intrinsic motility [Eq. (2)]. Second, it allows learning of the optimal policy (4) through a sequential process while maintaining an exploratory policy (5) like the

Boltzmann policy. We refer to this coupling of cue sensing (5) and modulation (6) as the log-exp coupling model from the biological interpretation of its continuous limit (7) obtained in the next subsection.

D. Continuous limit of optimal cue sensing and modulation

To demonstrate the biological relevance of the policy (5) and cue dynamics (6), we derive its continuous limit, which is structurally equivalent to the well-established Keller-Segel models for collective chemotaxis [26–28].

For simplicity, we consider a one-dimensional lattice $\mathcal{S} = S^h := \{0, \pm h, \pm 2h, \dots\}$ with a uniform spacing h and define the agents' intrinsic transition probability $p^h : S^h \times S^h \rightarrow [0, 1]$ as a lazy random walk, such that $p^h(x|x) = \varepsilon^h(x) = O(h)$, $p^h(x \pm h|x) = (1 - \varepsilon^h(x))/2$. Under an appropriate timescale and in the limit of infinite agents, the continuous space-time limits $h \rightarrow 0$ of the log-exp coupling models becomes the following pair of one-dimensional partial differential equations (derivation is provided in Appendix B b in the SM [51]):

$$\partial_t \mu = -\nabla \cdot (\mu \gamma \sigma^2 \nabla \log Z) + \frac{\sigma^2}{2} \nabla^2 \mu, \quad (7a)$$

$$\partial_t Z = -[\alpha \mu Z] + [\alpha \mu \exp(\beta r) Z^\gamma] + D \sigma^2 \nabla^2 Z, \quad (7b)$$

where $\mu(t, x) \in \mathbb{R}_{\geq 0}$ is the agents density and $Z(t, x) \in \mathbb{R}_{\geq 0}$ and $r(x) \in \mathbb{R}_{\geq 0}$ are the concentration of endogenous cue and exogenous signal, respectively. The parameter $\sigma \in \mathbb{R}_{\geq 0}$ is the randomness of the agents' motility, determined by the scaling of time and space.

Equation (7) is structurally homologous to the Keller-Segel models [26–28], which was originally introduced to model microbial populations that migrate by sensing and degrading an attractant cue:

$$\partial_t n = -\nabla \cdot (n \chi \nabla \log s) + D_n \nabla^2 n, \quad (8a)$$

$$\partial_t s = -[\eta n s] + [-\kappa s] + D_s \nabla^2 s, \quad (8b)$$

where $n(t, x)$ and $s(t, x)$ denote the microbial density and the attractant concentration, respectively. The parameter $D_n \in \mathbb{R}_{\geq 0}$ represents the randomness of the microbes, and $D_s \in \mathbb{R}_{\geq 0}$ is the diffusion coefficient of the attractant. Apart from linear diffusion terms appearing in both equations, Eqs. (7) and (8) share the logarithmic sensing terms, $\nabla \log Z$ and $\nabla \log s$ and the density-dependent linear degradation term, $-[\alpha \mu Z]$ and $-[\eta n s]$. The original Keller-Segel model does not include the production term, and its modified models introduced the term heuristically without considering the correspondence with the sensing term [36,55]. In contrast, the production term we derived, $[\alpha \mu \exp(\beta r) Z^\gamma]$, has an explicit meaning as optimally coupled cue production with logarithmic sensing. Notably, our production term depends on the endogenous cue concentration itself, which aligns with the signal relay mechanism in neutrophil swarming [56].

The optimality of Keller-Segel dynamics was highlighted in Pezzotta *et al.* [36], where, by using optimal control and mean-field games theories, they demonstrated that a cell population ascending an optimized endogenous cue gradient can be interpreted as an optimally controlled agent population following the optimal state value gradient. Yet the production

of the cue in their model was not theoretically principled. Our result extends their approach to reinforcement learning theory by considering an agent population where each agent ascends the endogenous cue gradient while cooperatively updating the endogenous cue (the estimated state value) based on the detected exogenous signal. This extension encompasses not only the optimal behaviors but also the learning process toward optimality, enabling the derivation of the optimal relationship between sensing and modulation of the endogenous cue gradient.

E. Another optimal coupling

The log-exp coupling is obtained by assuming the logarithmic mapping of Z_t to V_t . We can obtain another coupling by considering a different mapping, which could expand the applicability of our theory to diverse biological situations.

Another biologically relevant mapping is the one that linearly associates the endogenous cue concentration with the estimate of the state value $\mathcal{F}[Z_t](s) := Z_t(s)/\beta$. This mapping aligns with the absolute concentration sensing observed in various chemotactic behaviors [23,24,57,58]. The corresponding policy becomes

$$s_{t+1}^i \sim \pi_t(\cdot | s_t^i) := \frac{p(\cdot | s_t^i) \exp(\gamma Z_t(\cdot))}{\sum_{s' \in \mathcal{S}} p(s' | s_t^i) \exp(\gamma Z_t(s'))}. \quad (9)$$

Similarly to the log-exp coupling, the optimal cue modulation dynamics is obtained (see Appendix B a 1 in the SM [51] for derivation):

$$Z_{t+1}(s) - Z_t(s) = - \sum_{i=1}^N \mathbb{1}_{s_t^i=s} \alpha \Delta Z_t(s_t^i) - D \sum_{s' \in \mathcal{S}} L_{ss'} Z_t(s'), \quad s \in \mathcal{S}, \quad (10a)$$

$$\Delta Z_t(s) := Z_t(s) - \beta r(s) - \ln \sum_{s' \in \mathcal{S}} p(s' | s) \exp(\gamma Z_t(s')). \quad (10b)$$

This learning dynamics is equivalent to a distributed version of greedy G-learning [59] with a sequential regularization.

The continuous limit of Eqs. (9) and (10) is obtained (see Appendix B b in the SM [51] for derivation) as

$$\partial_t \mu = -\nabla \cdot (\mu \gamma \sigma^2 \nabla Z) + \frac{\sigma^2}{2} \nabla^2 \mu, \quad (11a)$$

$$\partial_t Z = -[\alpha(1 - \gamma)\mu Z] + [\alpha\beta\mu r] + D\sigma^2 \nabla^2 Z. \quad (11b)$$

Since the cue sensing and production are linear with respect to Z and r , respectively, we refer to this coupling of sensing (9) and modulation (10) as the linear-linear (lin-lin) coupling. Although production term is not included in the classical Keller-Segel model (8b), the linear production term $\alpha\beta\mu r$ of the lin-lin coupling model (11) corresponds to the production term μr of heuristically extended models [36,55].

Note that our formulation yields not only the two optimal couplings—log-exp and lin-lin—discussed in the main text but also additional optimal ones arising from alternative mappings $\mathcal{F}$. For instance, a mapping with the form of the Michaelis-Menten response leads to a learning dynam-

ics involving the response in both sensing and production, reflecting the underlying receptor kinetics [27,35,60] (see Appendix B a 2 in the SM [51]).

By introducing chemotactic response function $\psi(\nabla Z, \nabla \log Z)$, cue-degradation function $H(\mu, Z)$, and cue-production function $G(\mu, Z, r)$, the log-exp coupling, lin-lin coupling, and Keller-Segel models are integratively described as

$$\partial_t \mu = -\nabla \cdot (\mu \psi(\nabla Z, \nabla \log Z)) + \frac{\sigma^2}{2} \nabla^2 \mu, \quad (12a)$$

$$\partial_t Z = -H(\mu, Z) + G(\mu, Z, r) + D\sigma^2 \nabla^2 Z, \quad (12b)$$

where the functional forms of $\psi(\nabla Z, \nabla \log Z)$, $H(\mu, Z)$, and $G(\mu, Z, r)$ are summarized in Table I. From this table, we can identify the correspondences of the two couplings to phenomenological Keller-Segel models.

Then, the next question would be how the coupling between cue sensing and modulation affects collective exploration behaviors by agents and their efficiency. Through numerical experiments in the following Secs. III A and III B, we highlight both the behavioral and functional importance of optimal couplings to achieve optimal navigation.

III. NUMERICAL EXPERIMENTS

In this section, we conduct numerical experiments using the derived learning dynamics to evaluate their behaviors and optimality in biologically relevant problems.

A. Maze-solving problem and impact of optimal coupling

Biological systems often encounter exploration problems in complex environments. For instance, neutrophils detect and accumulate at infected tissues [41], and germ cells migrate to the gonad in embryos [61]. Navigation in a complex environment can be framed as maze-solving, and various cells were experimentally verified to navigate themselves in artificial mazes [13,16,62]. In this section, we employ a complex maze for investigating the properties and optimality of the derived learning dynamics.

We adopted a maze with $|\mathcal{S}| = 274$ vertices shown in Fig. 2, which models the maze used in an experimental work for maze solving by *Dictyostelium* (see Fig. 6 in Ref. [13]). In the following numerical experiments, we assume a lazy random walk for the intrinsic dynamics p of agents: $p(s|s) = \varepsilon$ and $p(s'|s) = (1 - \varepsilon)/|\mathcal{N}(s)|$, where $s' \in \mathcal{N}(s)$. The parameter $\varepsilon \in [0, 1)$ represents the sluggishness of migration, reflecting that the migration speed varies across species, even in the same geometry [13]. The exogenous signal concentration r is assumed to be high only at a goal (target) vertex $s^{\text{goal}} \in \mathcal{S}$ and uniformly low at the other vertices: $r(s^{\text{goal}}) = r_{\text{target}}$ and $r(s) = r_{\text{default}}$ for $s \neq s^{\text{goal}}$. This represents that the exogenous signal is produced only at s^{goal} and decays rapidly away from there. Initially, the endogenous cue is uniformly distributed ($Z_0(s) = Z_0$ for $s \in \mathcal{S}$), and all N agents start exploring from the vertex $s_0^i = s^{\text{start}} \in \mathcal{S}$. Then, agents collectively modify the endogenous cue concentration to reach the goal s^{goal} [Fig. 2].

Figures 2(a)–2(c) show temporal snapshots of the concentration distribution Z_t (a), agent distribution μ_t (b), and time-

TABLE I. Component-wise comparison via (12) of the continuous limits of the two learning dynamics with the classical Keller-Segel equation (8). The mapping $\mathcal{F}[Z]$ associates the endogenous cue concentration with the estimated state value. The term $\psi(\nabla Z, \nabla \log Z)$ represents the chemotactic response to endogenous cue concentration, and $H(\mu, Z)$ and $G(\mu, Z, r)$ do the degradation and production of the endogenous cue, respectively. Phenomenological meanings of parameters are chemotactic coefficient χ , cellular degradation rate η , and decay rate κ . Computational meanings of parameters are inverse control weight β , discount factor γ , agent's motility σ , and learning rate α .

	Keller-Segel model	log-exp coupling	lin-lin coupling
Mapping from concentration to value	$\mathcal{F}[Z]$	—	$\log(Z)/\beta$
Chemotactic response	$\psi(\nabla Z, \nabla \log Z)$	$\chi \nabla \log Z$	$\gamma \sigma^2 \nabla \log Z$
Degradation	$H(\mu, Z)$	$(\eta\mu + \kappa)Z$	Z/β
Production	$G(\mu, Z, r)$	(in extensions, μr)	$\alpha(1 - \gamma)\mu Z$
			$\alpha\beta\mu r$

averaged regularized reward distribution $(R^i[s_{0:t}^i, \pi_{0:t-1}])_{i=1}^N$ (c) for one trial of the log-exp coupling dynamics (5) and (6). The time-averaged regularized reward for agent i at time t is defined as

$$R^i[s_{0:T}^i, \pi_{0:T-1}] := \frac{1}{T} \sum_{t'=0}^{T-1} \left(r(s_{t'}^i) - \frac{1}{\beta} \log \frac{\pi_{t'}(s_{t'+1}^i | s_{t'}^i)}{p(s_{t'+1}^i | s_{t'}^i)} \right), \quad (13)$$

which measures the difference between the cumulative exogenous signal concentration detected by the agent and the motility control cost incurred by the agent from the start to time T along the individual state history. While the time-averaged reward $R^i[s_{0:T}^i, \pi_{0:T-1}]$ assigns less weight on early rewards than the original discounted cumulative reward $\mathcal{J}[\pi]$ (1), the maximizer of $R^i[s_{0:T}^i, \pi_{0:T-1}]$ can still correspond to

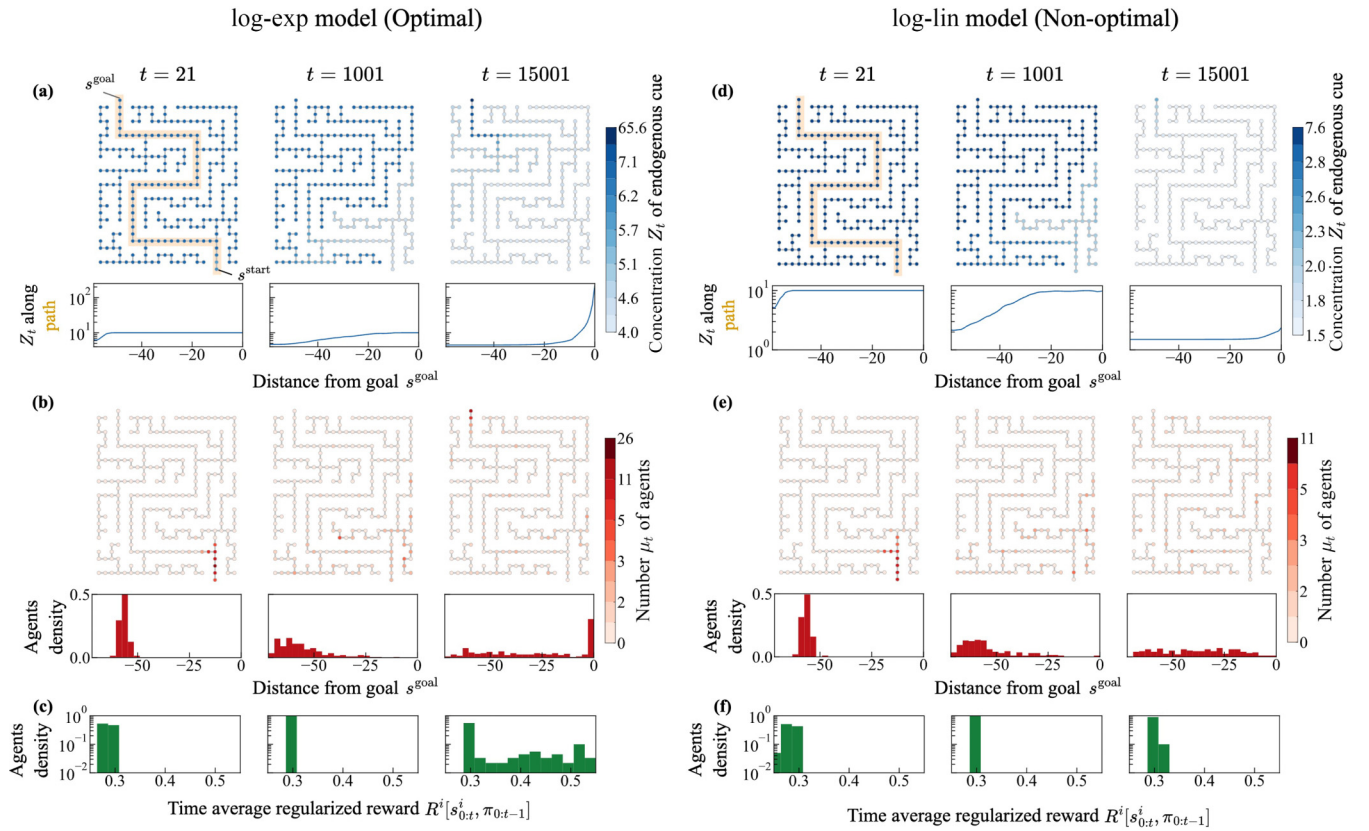


FIG. 2. Snapshots at $t = 21, 1001, 15001$ for one trial of the maze exploration by the optimal log-exp coupling dynamics [(5) and (6)] and the nonoptimal log-lin coupling model [(5) and (10)]. The distributions of the endogenous cue Z_t are shown on the maze (upper panel) and on the path from s^{start} to s^{goal} (lower panel) for the log-exp (a) and the log-lin (d) couplings. The path is highlighted with orange background in the maze and the distribution along the path is shown as a function from the goal s^{goal} . The agent distribution $\mu_t: \mathcal{S} \rightarrow \{0, 1, \dots, N\}$ defined as $\mu_t(s) := \sum_{i=1}^N \mathbb{1}_{s_i=s}$ is similarly shown for the log-exp (b) and the log-lin (e) couplings. Time-averaged regularized reward distribution $(R^i[s_{0:t}^i, \pi_{0:t-1}])_{i=1}^N$ for the log-exp (c) and the log-lin (f) couplings. The vertices are color coded by the value of cue concentration in (a) and (c) and that of agent density in (b) and (e). The parameter values used are $\alpha = 0.0098$, $N = 100$, $D = 0.01$, $\gamma = 0.8$, $Z_0 = 10.0$, $\varepsilon = 0.5$, $\beta = 1$, $r_{\text{target}} = 1.5$, and $r_{\text{default}} = 0.3$.

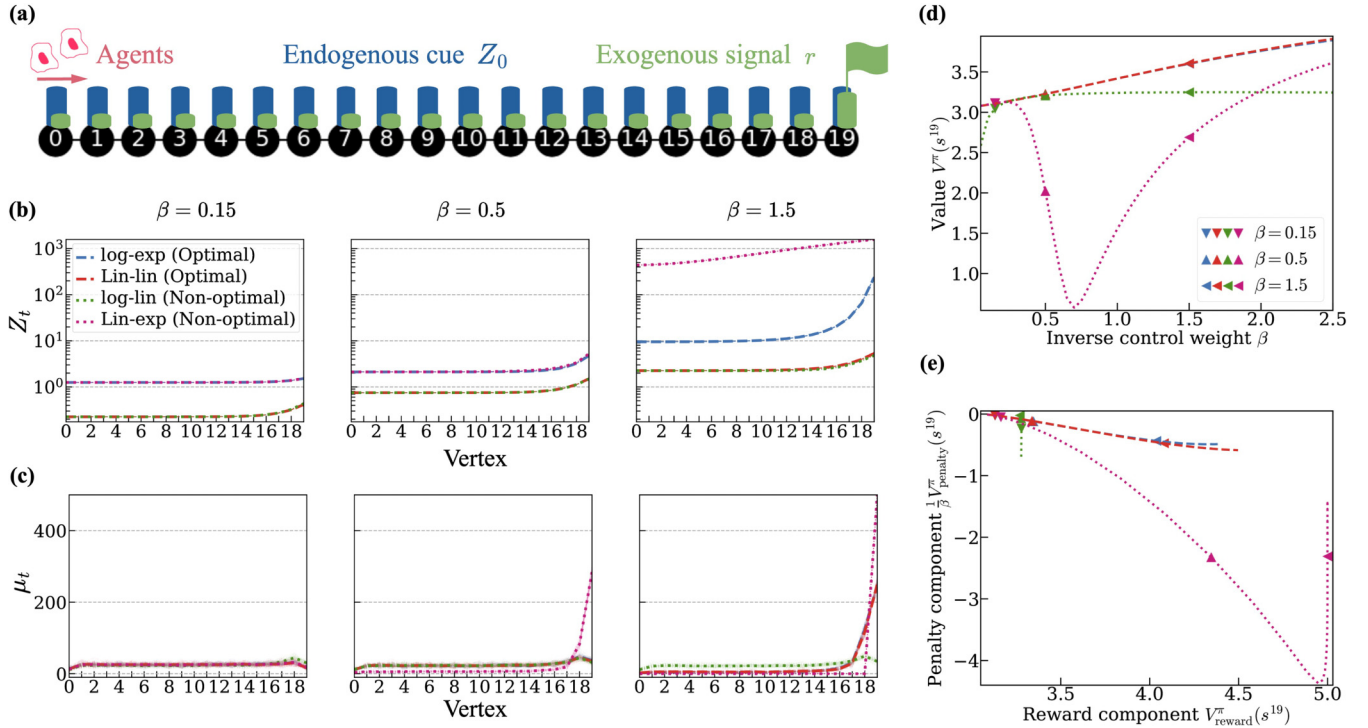


FIG. 3. (a) An environment represented by a graph with 20 vertices $\mathcal{S} = \{s^0, s^1, \dots, s^{19}\}$ connected in a chain. All agents start from vertex s^0 , and the target is located only at vertex s^{19} . The exogenous signal concentration (reward) is set $r(s^{19}) = r_{\text{target}} = 1.0$ and $r(s) = r_{\text{default}} = 0.3$ for $s \neq s^{19}$. (b) and (c) Endogenous cue concentration Z_T (b) and agent distribution μ_T (c) at $T = 8000$ for 50 trials with $\beta \in \{0.15, 0.5, 1.5\}$ for the optimal and nonoptimal couplings. The optimal couplings are log-exp coupling [(5) and (6)] and lin-lin coupling [(9) and (10)]; the nonoptimal couplings are log-lin coupling [(5) and (6)] and lin-exp coupling [(9) and (10)]. Each trial's Z_T and μ_T are shown with thin solid curves, and the trial averages $\bar{Z}_T$ and $\bar{\mu}_T$ are shown with thick solid curves. (d) State value $V^\pi(s^{19})$ at vertex s^{19} for each policy $\pi[\bar{Z}_T]$ based on the trial-averaged steady concentration $\bar{Z}_T$ for the four couplings with $\beta \in [0.25, 2.5]$. (e) The relationship between the reward component $V^\pi_{\text{reward}}(s^{19})$ and penalty component $\frac{1}{\beta} V^\pi_{\text{penalty}}(s^{19})$ of $V^\pi(s^{19})$ for the four couplings with $\beta \in [0.25, 2.5]$. The parameters values are $N = 500$, $\varepsilon = 0.5$, $\alpha = 0.00196$, $D = 0.01$, $\gamma = 0.8$, $\beta \in [0.25, 2.65]$.

the maximizer of $\mathcal{J}[\pi]$ under certain conditions (cf. Sec. 8.4 in Ref. [63]).

We observe that the agents expand their exploration area by dispersing at branching vertices and shaping a gradient toward unexplored directions through cue degradation [Figs. 2(a) and 2(b) at $t = 21$ and $t = 1001$]. Upon reaching s^{goal} , they detect the exogenous signal and produce the endogenous cue, strengthening the gradient toward s^{goal} . Eventually, the gradient aligns from s^{start} to s^{goal} and major fraction of agents accumulate near s^{goal} [Figs. 2(a) and 2(b) at $t = 15001$], resulting in high R^i for many agents [Fig. 2(c) at $t = 15001$]. These behaviors are qualitatively consistent with experimentally observed phenomena, such as exploration via endogenous cue degradation [11] and accumulation at target through endogenous cue production [12]. Similar behaviors are also observed in the optimal lin-lin coupling dynamics (9) and (10) (see Fig. S1 in the SM [51]).

Optimal couplings, in principle, outperform any nonoptimal coupling under the objective (1) across all parameters sets, provided $D/(\alpha N)$ is small, thanks to our optimality-based construction (see Appendix Cc in the SM). To highlight the behavioral importance of the optimal couplings, we adopt a simple nonoptimal benchmark comprising the logarithmic sensing policy (5) and the linear cue production (9), hereafter termed the log-lin coupling. Figures 2(d) and 2(e) show that

the agents with the log-lin coupling disperse, degrade the endogenous cue, and generate a cue gradient toward s^{goal} up to around $t = 1001$, similarly to the optimal log-exp coupling. However, the generated endogenous cue gradient is shallow and most agents fail to accumulate at the target even after the cue concentration has almost reached the steady state, resulting in a low R^i for most agents [Figs. 2(d)–2(f) at $t = 15001$]. For another nonoptimal coupling of linear sensing policy (9) and exponential production (6), we observed an overaccumulation (see Fig. S1 in the SM [51]). These results highlight the importance of optimal coupling between cue sensing and modulation for achieving efficient exploration and navigation in a complex environment.

B. Coupling optimality and its consequence

To scrutinize how the optimality of coupling affects the agent's behaviors, we compare the steady states in a simpler one-dimensional environment for the four couplings. Each coupling is obtained by combinatorially pairing sensing (5) and (9), with modulations (6) and (10). The environment is a graph with 20 vertices $\mathcal{S} = \{s^0, s^1, \dots, s^{19}\}$ connected in a chain, where a target locates at the vertex s^{19} [Fig. 3(a)]. We also modulate the relative importance of the reward and the control cost by changing the parameter β .

For the optimal couplings, i.e., log-exp and lin-lin couplings, the steady agent distributions μ_T are almost identical while the steady cue distribution Z_T are different, for each β [Fig. 3(c)]. The agents accumulate more at the target as β increases, indicating that accumulation is encouraged if the control cost becomes lower and the agents can deviate more from the intrinsic lazy random walk.

However, for the nonoptimal couplings, i.e., log-lin and lin-exp couplings, dispersal and overaccumulation in μ_T were observed, respectively [Fig. 3(b)], which are consistent with the results in the maze environment [Fig. 2(e)]. For the log-lin coupling, the agents remain dispersed while the cue distribution is almost identical to that of the optimal lin-lin coupling. Because the logarithmic sensing requires a steeper gradient than linear one to achieve the same chemotactic behavior, the agents with log-lin coupling were less sensitive to the gradient than lin-lin coupling and thus failed to accumulate at the target. For the lin-exp coupling, agents overaccumulated at the target compared with the optimal log-exp coupling when β is high. As linear sensing is more sensitive than logarithmic sensing especially when the basal cue concentration is high, this oversensitivity results in the overaccumulation. These results demonstrate that the disruption of the optimal coupling substantially alters the optimal exploration and exploitation behavior of agents.

To quantitatively evaluate the optimality of the four couplings in the steady states within our optimization framework, we compare, for each policy π_T , its state value function V^π (3), along with its reward component V_{reward}^π and the penalty (control cost) component V_{penalty}^π . Here we approximate the fluctuating steady concentration Z_T , which determines π_T , with the trial average $\tilde{Z}_T$, and define $V_{\text{reward}}^\pi(s) := E[\sum_{t=0}^{\infty} \gamma^t r(s_t) \mid s_0 = s]$, and $V_{\text{penalty}}^\pi(s) := -E[\sum_{t=0}^{\infty} \gamma^t \log \pi(s_{t+1}|s_t)/p(s_{t+1}|s_t) \mid s_0 = s]$. These values are obtained by value iteration method (Appendix A b in the SM [51]), and additional state values are provided in Fig. S2 in the SM [51].

The state value V^π , the optimization metric, is larger for the optimal log-exp and lin-lin couplings than for the nonoptimal log-lin and lin-exp couplings at any β [Fig. 3(d)]. The reward component V_{reward}^π and the penalty component $V_{\text{penalty}}^\pi/\beta$ of V^π are balanced for the optimal couplings [Fig. 3(e)] such that V_{reward}^π and $V_{\text{penalty}}^\pi/\beta$ trade off along the change in β . Despite the different shapes of Z_T [Fig. 3(b)], the two optimal couplings exhibit almost identical values of V^π and achieve the same balancing curve of V_{reward}^π and $V_{\text{penalty}}^\pi/\beta$ for all β . This is a natural consequence of the fact that the two optimal couplings solve the same optimization problem, while their representations of V via Z are different.

In contrast, we observe low V^π and imbalances between V_{reward}^π and $V_{\text{penalty}}^\pi/\beta$ for the nonoptimal couplings. V_{reward}^π of the nonoptimal log-lin coupling remains constant regardless of β , while the absolute value of $V_{\text{penalty}}^\pi/\beta$ increases as β decreases [Fig. 3(e)]. This indicates that agent dynamics are not navigated appropriately to exploit V_{reward}^π due to the nonoptimal coupling. For the nonoptimal lin-exp coupling, V_{reward}^π is far below that of the optimal coupling and changes nonmonotonically as a function of β [Fig. 3(d)]. Figure 3(e) indicates that the lin-exp coupling incurs a greater penalty than the optimal ones to gain the same value of V_{reward}^π [Fig. 3(e)]. This

imbalance between V_{reward}^π and $V_{\text{penalty}}^\pi/\beta$ results in a deficiency in overall performance. This analysis verifies that nonoptimal couplings are indeed inferior in terms of optimization for accumulating at targets, highlighting the importance to optimally couple cue sensing and modulation.

These results clarify that the optimal endogenous cue modulation qualitatively differs depending on the gradient sensing characteristics: The logarithmic sensing policy optimally couples with production that is exponentially dependent on exogenous signal concentration, whereas the linear sensing policy optimally couples with production that is linearly dependent on the concentration. These two are almost identical in terms of the optimization criterion, and there may exist other optimal couplings, corresponding to different choices of the mapping $\mathcal{F}$. This flexibility in the mapping allows optimization-based modeling to incorporate biological details, thereby expanding its applicability.

C. Comparison with a single-agent system

We have analyzed distributed learning in which agents with limited memory share and learn environmental information cooperatively through the endogenous cue to search for targets. However, there exist higher organisms that have sufficient memory and computational capacity as a single agent to store and learn environmental information within each individual's internal model [38]. An unsolved issue in the evolution of biological information processing is the fitness advantages and disadvantages of distributed information processing by agents with limited computational capacity compared with the centralized one by a single smart agent with sufficient capacity [39,40,64]. Even if both systems achieve the same objective, metrics such as robustness can depend on the system's architecture, and their comparative advantages can vary based on these metrics [65] (cf. no free lunch theorem for search [66]). We address this problem by leveraging the fact that both of our distributed model and the usual single-agent learning model are based on the same RL framework.

Specifically, we consider a single agent with an internal memory to store the estimate of state value function $Z_t : \mathcal{S} \rightarrow \mathbb{R}_{\geq 0}$ at each time step $t = 0, 1, \dots$. The agent has a state $s_t \in \mathcal{S}$ and transits between the states based on Z_t following logarithmic sensing policy (5), and update Z_t as

$$Z_{t+1}(s) - Z_t(s) = -\mathbb{1}_{s_t=s} \alpha' \Delta Z_t(s), \quad (14)$$

which is nearly identical to the exponential cue production (6) with $N = 1$, except the diffusion term (6b). This update rule is also identical to the standard greedy Z-learning [50]. For comparison, the learning rate $\alpha' \in (0, 1)$ for the single agent is set to $\alpha' = \alpha N$, indicating that it has the same learning capability as the entire agent population.

Figure 4(a) shows the exploration behavior of the single agent, which should be compared with its distributed counterpart [Fig. 2(a)]. In this one-trial comparison, the overall update speed of Z_t by the single agent is slower than the agent population due to its limited exploration range [Fig. 4(a) $t = 1001$]. At time $t = 15001$, while the agent population has mostly completed learning and some agents have accumulated at the target [Figs. 2(a) and 2(b)], the single agent is still exploring

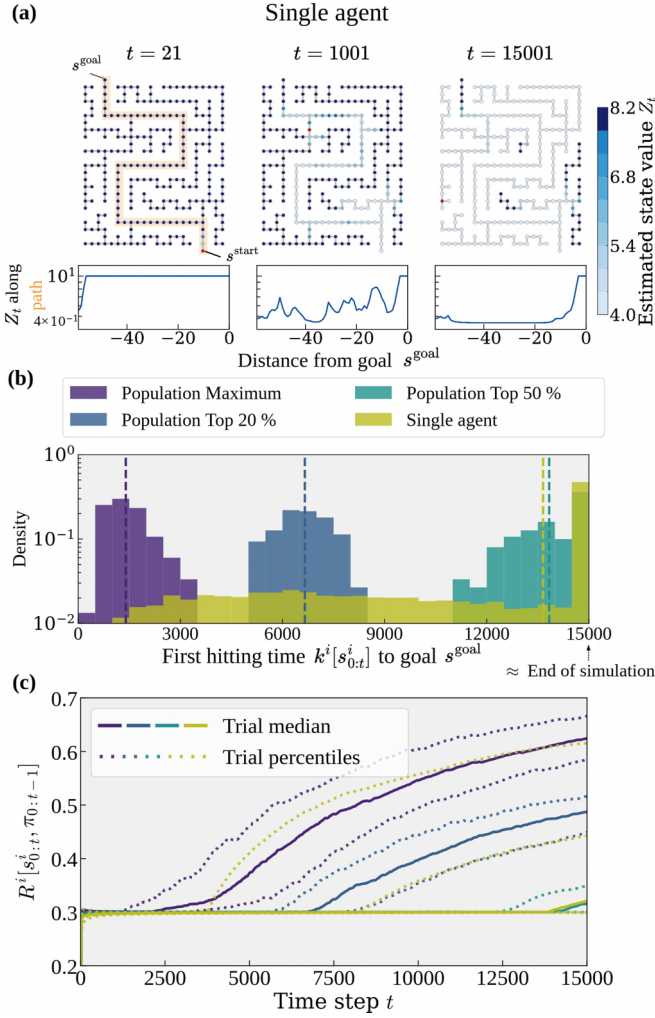


FIG. 4. (a) Snapshots of the estimated state value Z_t at $t = 21, 1001, 15001$ for one trial by a single smart agent with policy (5) and update rule (14). The upper panels show Z_t and the agent's location s_t on the maze. The vertex s_t , where the agent is located, is marked in red, while the other vertices $s \neq s_t$ are color coded based on $Z_t(s)$. The lower panels show the values of Z_t along the path highlighted by orange as a function of the distance from the goal s_{goal} . (b) Distribution of the first hitting time to s_{goal} computed from multiple trials by the single agent (yellow), and by the agent at the maximum (purple), 20th percentile (blue), and 50th percentile (green) of the population. The first hitting time $k^i[s_0:t]$ is calculated as $k^i[s_0:t] = \min\{t \in \{1, 2, \dots, T\} \mid s_t^i = s_{\text{goal}}\}$ if $s_{\text{goal}} \in \{s_t^i\}_{t=0:T}$; otherwise, $k^i[s_0:t] = T$. Each dashed line indicates the median value in the distributions. (c) Temporal variation of the time-averaged regularized reward $R^i[s_0:t, \pi_{0:t-1}]$ of the single agent (yellow) and of the agent at the maximum (purple), the 20th percentile (blue), and 50th percentile (green) of the population. Each solid line represents the trial median at each time step. Dashed lines represent the trial 10th, 30th, 70th, and 90th percentiles for the single agent (yellow), and the trial 10th and 90th percentiles for each population percentiles (purple, blue, and green). The parameters used are $T = 15001$, $(\alpha, N, D, \# \text{trials}) = (0.0098, 100, 0.01, 150)$ for the agent population, and $(0.98, 1, 0, 15000)$ for the single agent. The other parameters are same as Fig. 2.

the target. Furthermore, the value distribution becomes irregular [Fig. 4(a) $t = 1001$] due to random exploration histories

and the lack of a diffusion term. These observations suggest that the single smart agent may not explore and learn as effectively as the less-smart agent population.

For a more quantitative evaluation, Fig. 4(b) summarizes the hitting time statistics for multiple trials. The first hitting time of the smart single agent has a broad distribution ranging over $[1500, 15000]$, indicating that the success of the exploration by a single agent is extremely stochastic. For the agent population, the first hitting time fluctuates not only among trials, but also among agents in the population [Fig. 4(b)]. However, if we focus on the agent who hit the target first among the population, then its hitting time distribution for the multiple trials consolidates around $t = 1500$ and its median is close to the shortest hitting time of the single agent among the trials. Although the median of this distribution increases as we focus on hitting times ranked at 20% and 50% among the population, the value of 50% is still comparable to that of the single agent. Thus, population-wise exploration can work more robustly than its single-agent counterpart. We further evaluate the performance of the learning process using the time-averaged regularized reward R^i (13) of each agent, which is related to the optimization metric $\mathcal{J}$ (1) [Fig. 4(c)]. The bottom 50% of R^i in the agent population is smaller than its median of the single agent's trials, which corresponds to the fraction of agents trapped in dead ends [Fig. 2(b) $t = 15001$]. However, the top 20% of R^i in the agent population are larger than the top 30% of the single agent's trials for almost 90% of trials.

These results indicate that population-wise exploration can work more robustly in each trial than the single-agent-wise counterpart at the cost of the fraction of agents who could not reach the target. Such a fraction can be compensated effectively by the self-replication of the other fraction that reached the target given the cost for replication is cheap. This suggests that distributed exploration becomes more advantageous for simple and fertile organisms. In real biological problems such as neutrophil accumulation at a wound site, a prompt arrival to the target even by a fraction of immune cells is effective against exponentially growing pathogens entering from the wound.

IV. SUMMARY AND DISCUSSION

To formulate a cell population that explores the environment and learns optimal navigation toward targets through active cue modulation, we leveraged reinforcement learning theory and derived two qualitatively different optimal dynamics of the population with linear and logarithmic gradient sensing. These dynamics have continuous limits that are structurally equivalent to the phenomenological Keller-Segel model, indicating that optimal learning could be achieved by biologically relevant cell populations. We also verified that the optimal coupling of the modulation and sensing of cues is crucial for achieving efficient exploration by the population. Additionally, we have demonstrated that an agent population with limited computational capability can achieve more robust learning via distributed modulation of external cues than a smart single agent with complete internal memory.

A. Extensions of the formulation

Although we derived only two optimal couplings, our formulation can be extended beyond them by incorporating various biological factors, providing a foundation for the quantitative understanding of collective information processing mediated by endogenous chemicals. For instance, we can derive a learning dynamics involving a Hill function in chemotactic response and production, reflecting the underlying receptor kinetics [27,35,60] (see Appendix B a 2 in the SM [51]). Furthermore, we may explore various collective dynamics from the optimality principle by modifying (i) the cue-value mapping and control penalty, (ii) regularization term, and (iii) agent's interactions and heterogeneity.

Nonlinear mappings that align with control penalties enable the derivation of nonlinear dynamics. In this work, we used a log-form control penalty (KL divergence), which aligns with the logarithmic mapping, to obtain the log-exp model. Alternatively, other control penalties, such as Tsallis relative entropy [67], could be employed to explore different nonlinear dynamics through their associated nonlinear mappings. The diffusivity of the endogenous cue, which was regarded as a physical constraint in Appendix B a 1 in the SM [51], can also be interpreted as Laplacian regularization, which ensures continuity and smoothness of the value between adjacent vertices in the graph [68,69]. We may exploit other types of regularization. For instance, L^2 regularization ($\lambda\|Z_t\|_2^2$, $\lambda \in \mathbb{R}_{\geq 0}$) can suppress extreme values and results in another sequential regularization term $-\lambda Z_t(s)$. This aligns with the cue decay [$\lambda = \kappa$, Eq. (8b)] in Keller-Segel models, or the pheromones evaporation in ant population dynamics [70]. Moreover, agent-dependent regularization might be employed to account for the alternation of cue diffusivity via the enzyme produced by cells [71].

Incorporation of agents' interactions and heterogeneity is more challenging theoretically. We obtained homogeneous agents that interact only through one endogenous cue, which stems from our assumption that the reward function depends only on the targets' location and magnitude. However, agents' behaviors can be influenced by integrated information from other agents, which is often biologically mediated via a variety of cues, including repellents [72]. For instances, immune responses are modulated by chemokines released by other immune cells to control inflammation [73]. Microorganisms recognize and aggregate in response to the states of others to adapt to starvation [74]. In development, cells repel each other to form tissues [75]. Moreover, biological systems often

involve heterogeneous cell populations, where different subgroups play distinct roles, such as leaders navigating gradients at the front and followers trailing behind [76]. Extending our formulation to include interactions [77] and heterogeneity [78] is essential to comprehend the computational aspects of complex and collective behaviors observed in heterogeneous biological populations.

B. Evolutionary advantages of distributed computation

Understanding the evolution of biological information processing lies at the intersection of biophysics, evolutionary biology, and computational neuroscience. Our result demonstrates that a population of simple agents can excel at learning environmental information robustly compared to a single smart agent. This aligns with the redundancy principle [79], which states that having many particles with the same objective can increase the probability of quickly achieving the objective as a population. This principle has been applied to various biological processes from synaptic signaling to embryo fertilization, where the arrival time of the fastest individuals is more critical than the population average. According to this biophysical principle, distributed information processing may have an evolutionary benefit for agents with limited capabilities yet low replication costs. For example, neutrophils may achieve a robust population response by dispersing numerous genetically identical cells. Also, integrating learning and evolutionary theories [80–83] would be a promising direction for advancing our understanding of the principles behind the emergence and persistence of diverse forms of biological information processing.

ACKNOWLEDGMENTS

The authors thank Shuhei A. Horiguchi for discussions. M.K. received the JSPS Research Fellowship Grant No. 24KJ0846 and was partially supported by the World-leading Innovative Graduate Study Program in Proactive Environmental Studies (WINGS-PES), The University of Tokyo. This work was supported by JST CREST JPMJCR2011 and JSPS KAKENHI Grants No. 24H01465 and No. 25H01365.

DATA AVAILABILITY

No data were created or analyzed in this study.

-
- [1] J. A. Shapiro, Thinking about bacterial populations as multicellular organisms, *Annu. Rev. Microbiol.* **52**, 81 (1998).
 - [2] R. Popat, D. M. Cornforth, L. McNally, and S. P. Brown, Collective sensing, and collective responses in quorum-sensing bacteria, *J. R. Soc. Interface* **12**, 20140882 (2015).
 - [3] H.-C. Flemming, J. Wingender, U. Szewzyk, P. Steinberg, S. A. Rice, and S. Kjelleberg, Biofilms: An emergent form of bacterial life, *Nat. Rev. Microbiol.* **14**, 563 (2016).
 - [4] P. Friedl and D. Gilmour, Collective cell migration in morphogenesis, regeneration, and cancer, *Nat. Rev. Mol. Cell Biol.* **10**, 445 (2009).
 - [5] E. Theveneau and R. Mayor, Neural crest delamination, and migration: From epithelium-to-mesenchyme transition to collective cell migration, *Dev. Biol.* **366**, 34 (2012).
 - [6] D. Ellison, A. Mugler, M. D. Brennan, S. H. Lee, R. J. Huebner, E. R. Shamir, L. A. Woo, J. Kim, P. Amar, I. Nemenman, A. J. Ewald, and A. Levchenko, Cell-cell

- communication enhances the capacity of cell ensembles to sense shallow gradients during morphogenesis, *Proc. Natl. Acad. Sci. USA* **113**, E679 (2016).
- [7] A. D. Luster, R. Alon, and U. H. von Andrian, Immune cell migration in inflammation: Present, and future therapeutic targets, *Nat. Immunol.* **6**, 1182 (2005).
- [8] A. Mantovani, M. A. Cassatella, C. Costantini, and S. Jaillon, Neutrophils in the activation, and regulation of innate, and adaptive immunity, *Nat. Rev. Immunol.* **11**, 519 (2011).
- [9] R. Axelrod, D. E. Axelrod, and K. J. Pienta, Evolution of cooperation among tumor cells, *Proc. Natl. Acad. Sci. USA* **103**, 13474 (2006).
- [10] T. S. Deisboeck and I. D. Couzin, Collective behavior in cancer cell populations, *BioEssays* **31**, 190 (2009).
- [11] R. H. Insall, P. Paschke, and L. Tweedy, Steering yourself by the bootstraps: How cells create their own gradients for chemotaxis, *Trends Cell Biol.* **32**, 585 (2022).
- [12] S. de Oliveira, E. E. Rosowski, and A. Huttenlocher, Neutrophil migration in infection, and wound repair: Going forward in reverse, *Nat. Rev. Immunol.* **16**, 378 (2016).
- [13] L. Tweedy, P. A. Thomason, P. I. Paschke, K. Martin, L. M. Machesky, M. Zagnoni, and R. H. Insall, Seeing around corners: Cells solve mazes, and respond at a distance using attractant breakdown, *Science* **369**, eaay9792 (2020).
- [14] B. Bréart, W. D. Ramos-Perez, A. Mendoza, A. K. Salous, M. Gobert, Y. Huang, R. H. Adams, J. J. Lafaille, D. Escalante-Alcalde, A. J. Morris, and S. R. Schwab, Lipid phosphate phosphatase 3 enables efficient thymic egress, *J. Exp. Med.* **208**, 1267 (2011).
- [15] T. Lämmermann, P. V. Afonso, B. R. Angermann, J. M. Wang, W. Kastenmüller, C. A. Parent, and R. N. Germain, Neutrophil swarms require LTB₄, and integrins at sites of cell death *in vivo*, *Nature (London)* **498**, 371 (2013).
- [16] C. Scherber, A. J. Aranyosi, B. Kulemann, S. P. Thayer, M. Toner, O. Iliopoulos, and D. Irimia, Epithelial cell guidance by self-generated egf gradients, *Integr. Biol.* **4**, 259 (2012).
- [17] O. Susanto, Y. W. H. Koh, N. Morrice, S. Tumanov, P. A. Thomason, M. Nielson, L. Tweedy, A. J. Muinonen-Martin, J. J. Kamphorst, G. M. Mackay, and R. H. Insall, LPP3 mediates self-generation of chemotactic LPA gradients by melanoma cells, *J. Cell Sci.* **130**, 3455 (2017).
- [18] D. A. Clark and L. C. Grant, The bacterial chemotactic response reflects a compromise between transient, and steady-state behavior, *Proc. Natl. Acad. Sci. USA* **102**, 9150 (2005).
- [19] A. Celani and M. Vergassola, Bacterial strategies for chemotaxis response, *Proc. Natl. Acad. Sci. USA* **107**, 1391 (2010).
- [20] H. H. Mattingly, K. Kamino, B. B. Machta, and T. Emonet, *Escherichia coli* chemotaxis is information limited, *Nat. Phys.* **17**, 1426 (2021).
- [21] K. Nakamura and T. J. Kobayashi, Optimal sensing, and control of run-and-tumble chemotaxis, *Phys. Rev. Res.* **4**, 013120 (2022).
- [22] J. E. Segall, S. M. Block, and H. C. Berg, Temporal comparisons in bacterial chemotaxis, *Proc. Natl. Acad. Sci. USA* **83**, 8987 (1986).
- [23] Y. V. Kalinin, L. Jiang, Y. Tu, and M. Wu, Logarithmic sensing in *Escherichia coli* bacterial chemotaxis, *Biophys. J.* **96**, 2439 (2009).
- [24] M. D. Lazova, T. Ahmed, D. Bellomo, R. Stocker, and T. S. Shimizu, Response rescaling in bacterial chemotaxis, *Proc. Natl. Acad. Sci. USA* **108**, 13870 (2011).
- [25] Y. Tu, Quantitative modeling of bacterial chemotaxis: Signal amplification, and accurate adaptation, *Annu. Rev. Biol.* **42**, 337 (2013).
- [26] E. F. Keller and L. A. Segel, Model for chemotaxis, *J. Theor. Biol.* **30**, 225 (1971).
- [27] M. J. Tindall, P. K. Maini, S. L. Porter, and J. P. Armitage, Overview of mathematical approaches used to model bacterial chemotaxis II: Bacterial populations, *Bull. Math. Biol.* **70**, 1570 (2008).
- [28] K. J. Painter, Mathematical models for chemotaxis, and their applications in self-organisation phenomena, *J. Theor. Biol.* **481**, 162 (2019).
- [29] R. Alert, A. Martínez-Calvo, and S. S. Datta, Cellular sensing governs the stability of chemotactic fronts, *Phys. Rev. Lett.* **128**, 148101 (2022).
- [30] T. V. Phan, H. H. Mattingly, L. Vo, J. S. Marvin, L. L. Looger, and T. Emonet, Direct measurement of dynamic attractant gradients reveals breakdown of the Patlak–Keller–Segel chemotaxis model, *Proc. Natl. Acad. Sci. USA* **121**, e2309251121 (2024).
- [31] X. Fu, S. Kato, J. Long, H. H. Mattingly, C. He, D. C. Vural, S. W. Zucker, and T. Emonet, Spatial self-organization resolves conflicts between individuality, and collective migration, *Nat. Commun.* **9**, 2177 (2018).
- [32] H. H. Mattingly and T. Emonet, Collective behavior, and nongenetic inheritance allow bacterial populations to adapt to changing environments, *Proc. Natl. Acad. Sci. USA* **119**, e2117377119 (2022).
- [33] D. T. Fraebel, H. Mickalide, D. Schnitkey, J. Merritt, T. E. Kuhlman, and S. Kuehn, Environment determines evolutionary trajectory in a constrained phenotypic space, *Elife* **6**, e24669 (2017).
- [34] W. Liu, J. Cremer, D. Li, T. Hwa, and C. Liu, An evolutionarily stable strategy to colonize spatially extended habitats, *Nature (London)* **575**, 664 (2019).
- [35] J. Cremer, T. Honda, Y. Tang, J. Wong-Ng, M. Vergassola, and T. Hwa, Chemotaxis as a navigation strategy to boost range expansion, *Nature (London)* **575**, 658 (2019).
- [36] A. Pezzotta, M. Adorisio, and A. Celani, Chemotaxis emerges as the optimal solution to cooperative search games, *Phys. Rev. E* **98**, 042401 (2018).
- [37] D. Horstmann, and A. Stevens, A constructive approach to traveling waves in chemotaxis, *J. Nonlinear Sci.* **14**, 1 (2004).
- [38] T. E. J. Behrens, T. H. Muller, J. C. R. Whittington, S. Mark, A. B. Baram, K. L. Stachenfeld, and Z. Kurth-Nelson, What is a cognitive map? Organizing knowledge for flexible behavior, *Neuron* **100**, 490 (2018).
- [39] R. Solé, M. Moses, and S. Forrest, Liquid brains, solid brains, *Philos. Trans. R. Soc. B* **374**, 20190040 (2019).
- [40] M. G. Navas-Zuloaga, T. P. Pavlic, and B. H. Smith, Alternative model systems for cognitive variation: eusocial-insect colonies, *Trends Cogn. Sci.* **26**, 836 (2022).
- [41] E. Kolaczowska and P. Kubes, Neutrophil recruitment, and function in health, and inflammation, *Nat. Rev. Immunol.* **13**, 159 (2013).
- [42] H. C. Berg, *E. coli in Motion* (Springer, Berlin, 2004).

- [43] A. C. Szatmary, R. Nossal, C. A. Parent, and R. Majumdar, Modeling neutrophil migration in dynamic chemoattractant gradients: Assessing the role of exosomes during signal relay, *Mol. Biol. Cell* **28**, 3457 (2017).
- [44] S. Uderhardt, A. J. Martins, J. S. Tsang, T. Lämmermann, and R. N. Germain, Resident macrophages cloak tissue microlesions to prevent neutrophil-driven inflammatory damage, *Cell* **177**, 541 (2019).
- [45] L. G. Ng, J. S. Qin, B. Roediger, Y. Wang, R. Jain, L. L. Cavanagh, A. L. Smith, C. A. Jones, M. De Veer, M. A. Grimbaldston, E. N. Meeusen, and W. Weninger, Visualizing the neutrophil response to sterile tissue injury in mouse dermis reveals a three-phase cascade of events, *J. Invest. Dermatol.* **131**, 2058 (2011).
- [46] P. V. Afonso, M. Janka-Junttila, Y. J. Lee, C. P. McCann, C. M. Oliver, K. A. Aamer, W. Losert, M. T. Cicerone, and C. A. Parent, LTB4 is a signal-relay molecule during neutrophil chemotaxis, *Dev. Cell* **22**, 1079 (2012).
- [47] K. M. Glaser, M. Mhlan, and T. Lämmermann, Positive feedback amplification in swarming immune cell populations, *Curr. Opin. Cell Biol.* **72**, 156 (2021).
- [48] S. X. Tamás, B. T. Roux, B. Vámosi, F. G. Dehne, A. Török, L. Fazekas, and B. Enyedi, A genetically encoded sensor for visualizing leukotriene b4 gradients *in vivo*, *Nat. Commun.* **14**, 4610 (2023).
- [49] T. Kato and T. J. Kobayashi, Understanding adaptive immune system as reinforcement learning, *Phys. Rev. Res.* **3**, 013222 (2021).
- [50] E. Todorov, Efficient computation of optimal actions, *Proc. Natl. Acad. Sci. USA* **106**, 11478 (2009).
- [51] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/tvfy-lbbl> for additional details on RL theory, derivation, and experiments, which also contains Refs. [84–100].
- [52] C. Janetopoulos, L. Ma, P. N. Devreotes, and P. A. Iglesias, Chemoattractant-induced phosphatidylinositol 3, 4, 5-trisphosphate accumulation is spatially amplified, and adapts, independent of the actin cytoskeleton, *Proc. Natl. Acad. Sci. USA* **101**, 8951 (2004).
- [53] Y. Tu and W.-J. Rappel, Adaptation in living systems, *Annu. Rev. Condens. Matter Phys.* **9**, 183 (2018).
- [54] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in *Proceedings of The 33rd International Conference on Machine Learning* (PMLR, Cambridge, MA, 2016), Vol. 48, pp. 1928–1937.
- [55] T. J. Newman and R. Grima, Many-body theory of chemotactic cell-cell interactions, *Phys. Rev. E* **70**, 051916 (2004).
- [56] E. Strickland, D. Pan, C. Godfrey, J. S. Kim, A. Hopke, W. Ji, M. Degrange, B. Villavicencio, M. K. Mansour, C. S. Zerbe, D. Irimia, A. Amir, and O. D. Weiner, Self-extinguishing relay waves enable homeostatic control of human neutrophil swarming, *Dev. Cell* **59**, 2659 (2024).
- [57] D. Fuller, W. Chen, M. Adler, A. Groisman, H. Levine, W.-J. Rappel, and W. F. Loomis, External, and internal constraints on eukaryotic chemotaxis, *Proc. Natl. Acad. Sci. USA* **107**, 9656 (2010).
- [58] A. S. Perelson, *Theoretical Immunology, Part Two* (Addison-Wesley, Boston, 1988).
- [59] R. Fox, A. Pakman, and N. Tishby, Taming the noise in reinforcement learning via soft updates, in *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence* (AUAI Press, Arlington, VA, 2016).
- [60] J. Monod, On the nature of allosteric transitions: A plausible model, *J. Mol. Biol.* **12**, 88 (1965).
- [61] A. D. Renault, P. S. Kunwar, and R. Lehmann, Lipid phosphate phosphatase activity regulates dispersal, and bilateral sorting of embryonic germ cells in *Drosophila*, *Development* **137**, 1815 (2010).
- [62] T. V. Phan, R. Morris, M. E. Black, T. K. Do, K.-C. Lin, K. Nagy, J. C. Sturm, J. Bos, and R. H. Austin, Bacterial route finding, and collective escape in mazes, and fractals, *Phys. Rev. X* **10**, 031017 (2020).
- [63] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. (John Wiley & Sons, New York, 1994).
- [64] W. F. Vining, F. Esponda, M. E. Moses, and S. Forrest, How does mobility help distributed systems compute? *Philos. Trans. R. Soc. B* **374**, 20180375 (2019).
- [65] M. E. Moses, J. L. Cannon, D. M. Gordon, and S. Forrest, Distributed adaptive search in T cells: Lessons from ants, *Front. Immunol.* **10**, 1357 (2019).
- [66] D. H. Wolpert and W. G. Macready, No Free Lunch Theorems for Search, Technical Report 12, Santa Fe Institute (1995).
- [67] Y. Chow, O. Nachum, and M. Ghavamzadeh, Path consistency learning in Tsallis entropy regularized MDPs, in *Proceedings of the 35th International Conference on Machine Learning* (PMLR, Cambridge, MA, 2018), Vol. 80, pp. 979–988.
- [68] K. Lang and G. E. Hinton, Dimensionality reduction, and prior knowledge in E-Set Recognition, in *Advances in Neural Information Processing Systems* (Morgan Kaufmann, San Francisco, CA, 1989), Vol. 2, pp. 178–185.
- [69] X. Zhu, Z. Ghahramani, and J. D. Lafferty, Semi-supervised learning using Gaussian fields, and harmonic functions, in *Proceedings of the 20th International Conference on Machine Learning* (AAAI Press, Washington, DC, 2003), pp. 912–919.
- [70] D. Chowdhury, A. Schadschneider, and K. Nishinari, Physics of transport, and traffic phenomena in biology: From molecular motors, and cells to organisms, *Phys. Life Rev.* **2**, 318 (2005).
- [71] J. Weickert, A review of nonlinear diffusion filtering, in *Proceedings of International Conference on Scale-space Theories in Computer Vision* (Springer, Berlin, 1997), pp. 1–28.
- [72] C. E. Hughes and R. J. B. Nibbs, A guide to chemokines, and their receptors, *FEBS J.* **285**, 2944 (2018).
- [73] A. Hopke, A. Scherer, S. Kreuzburg, M. S. Abers, C. S. Zerbe, M. C. Dinauer, M. K. Mansour, and D. Irimia, Neutrophil swarming delays the growth of clusters of pathogenic fungi, *Nat. Commun.* **11**, 2031 (2020).
- [74] B. J. Crespi, The evolution of social behavior in microorganisms, *Trends Ecol. Evol.* **16**, 178 (2001).
- [75] B. Stramer and R. Mayor, Mechanisms and in vivo functions of contact inhibition of locomotion, *Nat. Rev. Mol. Cell Biol.* **18**, 43 (2017).
- [76] E. Theveneau, B. Steventon, E. Scarpa, S. Garcia, X. Treppe, A. Streit, and R. Mayor, Chase-and-run between adjacent cell populations promotes directional collective migration, *Nat. Cell Biol.* **15**, 763 (2013).
- [77] M. Laurière, S. Perrin, M. Geist, and O. Pietquin, Learning mean fieldgames: A survey, [arXiv:2205.12944](https://arxiv.org/abs/2205.12944).

- [78] S. G. Subramanian, P. Poupart, M. E. Taylor, and N. Hegde, Multi type mean field reinforcement learning, in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20* (IFAAMAS, Richland, SC, Auckland, New Zealand, 2020), pp. 411–419.
- [79] Z. Schuss, K. Basnayake, and D. Holcman, Redundancy principle, and the role of extreme statistics in molecular, and cellular biology, *Phys. Life Rev.* **28**, 52 (2019).
- [80] P. Turney, D. Whitley, and R. W. Anderson, Evolution, learning, and instinct: 100 years of the Baldwin effect, *Evol. Comput.* **4**, 271 (1996).
- [81] B. Xue and S. Leibler, Evolutionary learning of adaptation to varying environments through a transgenerational feedback, *Proc. Natl. Acad. Sci. USA* **113**, 11266 (2016).
- [82] S. Khadka and K. Tumer, Evolution-guided policy gradient in reinforcement learning, in *Advances in Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, 2018).
- [83] S. Nakashima, and T. J. Kobayashi, Acceleration of evolutionary processes by learning, and extended Fisher’s fundamental theorem, *Phys. Rev. Res.* **4**, 013069 (2022).
- [84] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 2018).
- [85] S. C. Tanaka, K. Doya, G. Okada, K. Ueda, Y. Okamoto, and S. Yamawaki, Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops, *Nat. Neurosci.* **7**, 887 (2004).
- [86] J. E. M. Bennett, A. Philippides, and T. Nowotny, Learning with reinforcement prediction errors in a model of the *Drosophila* mushroom body, *Nat. Commun.* **12**, 2569 (2021).
- [87] W. Schultz, Dopamine reward prediction-error signalling: a two-component response, *Nat. Rev. Neurosci.* **17**, 183 (2016).
- [88] B. D. Ziebart, Modeling purposeful adaptive behavior with the principle of maximum causal entropy, Ph.D. thesis, Carnegie Mellon University, 2010.
- [89] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, Reinforcement learning with deep energy-based policies, in *Proceedings of the 34th International Conference on Machine Learning* (PMLR, Cambridge, MA, 2017), Vol. 70, pp. 1352–1361.
- [90] M. G. Azar, V. Gómez, and H. J. Kappen, Dynamic policy programming, *J. Mach. Learn. Res.* **13**, 3207 (2012).
- [91] W.-W. Tso and J. Adler, Negative chemotaxis in *Escherichia coli*, *J. Bacteriol.* **118**, 560 (1974).
- [92] D. Precup, R. S. Sutton, and S. P. Singh, Eligibility traces for off-policy policy evaluation, in *Proceedings of the Seventeenth International Conference on Machine Learning* (Morgan Kaufmann Publishers, San Francisco, CA, 2000), pp. 759–766.
- [93] T. Sun, Y. Sun, and W. Yin, On Markov chain gradient descent, in *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY, 2018), pp. 9918–9927.
- [94] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint* (Springer, 2008).
- [95] A. Stevens, The derivation of chemotaxis equations as limit dynamics of moderately interacting stochastic many-particle systems, *SIAM J. Appl. Math.* **61**, 183 (2000).
- [96] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed. (Springer, 2001), Vol. 24.
- [97] B. Oksendal, *Stochastic Differential Equations: An Introduction with Applications*, 3rd ed. (Springer-Verlag, 1992).
- [98] R. Grima and T. J. Newman, Accurate discretization of advection-diffusion equations, *Phys. Rev. E* **70**, 036703 (2004).
- [99] H. Meyer and H. Rieger, Alignment interaction and band formation in assemblies of autochemorepulsive walkers, *Phys. Rev. E* **108**, 034604 (2023).
- [100] M. Kato and T. J. Kobayashi, Optimality theory of stigmergic collective information processing by chemotactic cells, [arXiv:2407.15298](https://arxiv.org/abs/2407.15298).