On Optimal Foraging and Multi-armed Bandits

Vaibhav Srivastava Paul Reverdy Naomi E. Leonard

Abstract— **We consider two variants of the standard multi-armed bandit problem, namely, the multi-armed bandit problem with transition costs and the multi-armed bandit problem on graphs. We develop block allocation algorithms for these problems that achieve an expected cumulative regret that is uniformly dominated by a logarithmic function of time, and an expected cumulative number of transitions from one arm to another arm uniformly dominated by a double-logarithmic function of time. We observe that the multi-armed bandit problem with transition costs and the associated block allocation algorithm capture the key features of popular animal foraging models in literature.**

I. Introduction

Foraging is a fundamental animal behavior that pertains to searching out food resources and exploiting them. Foraging behavior is studied in *behavioral ecology* using economic principles, i.e., the foraging decisions are evaluated based on their effects on certain pay-off functions. At the heart of a foraging decision is the tradeoff between exploration (to search for a better food resource) and exploitation (to stick with the best known food resource).

In the engineering literature, a benchmark setup to study the exploration-exploitation tradeoff is the multi-armed bandit problem. The multi-armed bandit problem models a class of resource allocation problems in which a decision-maker allocates a single resource by sequentially choosing one among a set of competing alternative options called arms. In the so-called stationary multi-armed bandit problem, a decision-maker at each discrete time instant chooses an arm and collects a reward drawn from an unknown stationary probability distribution associated with the selected arm. The objective of the decision-maker is to maximize the total reward aggregated over the sequential allocation process.

The fundamental exploration-exploitation tradeoff in foraging can be modeled as a multi-armed bandit problem, and the effectiveness of the foraging decisions can be measured by comparing them to the optimal decisions for the multi-armed bandit problem. In this paper, we explore this connection and argue that the solution to a Bayesian multi-armed bandit problem captures the qualitative features of the foraging behavior in some animals.

Literature review: The multi-armed bandit problem has been extensively studied; a survey is presented in [1]. In their seminal work, Lai and Robbins [2] established a logarithmic lower bound on the expected number of times a sub-optimal arm needs to be selected by an optimal policy. Since [2], a considerable emphasis has been on the design of simple heuristic policies that achieve the logarithmic lower bound on the expected number of selection instances of any suboptimal arm. To this end, Auer *et al.* [3] developed upper confidence bound (UCB) algorithms for multi-armed bandits with bounded reward that achieve logarithmic expected cumulative regret uniformly in time. Recently, Srinivas *et al.* [4] developed asymptotically optimal UCB algorithms for Gaussian process optimization. Kauffman *et al.* [5] developed a generic Bayesian UCB algorithm and established its optimality for binary bandits with uniform prior. Reverdy *et al.* [6] established the optimality of a Bayesian UCB algorithm for Gaussian rewards and drew several connections between these algorithms and human decision-making. They also elucidated the role of priors in decision-making performance.

Some variations of the multi-armed bandit problem have been studied as well. Agarwal *et al.* [7] studied the multi-armed bandit problem with transition costs, i.e., the multi-armed bandit problem in which a certain penalty is imposed each time the decision-maker switches from the currently selected arm, and developed an asymptotically optimal block allocation algorithm. In this paper, we consider the Gaussian multi-armed bandit problem with transition costs and develop a block allocation algorithm that achieves an expected cumulative regret that is uniformly dominated by a logarithmic term. Moreover, the block allocation scheme designed in this paper incurs smaller expected transition costs than the block allocation scheme in [7].

Kleinberg *et al.* [8] considered the multi-armed bandit problem in which every arm is not available for selection at each time (sleeping experts), and they analyzed the performance of the UCB algorithms. In contrast to the temporal unavailability of arms in [8], we consider a spatial unavailability of arms. We propose a novel multi-armed bandit problem, namely, the *graphical multi-armed bandit* problem, in which only a subset of the arms can be selected at the next allocation instance given the currently selected arm. We develop a block allocation algorithm for such a problem that achieves expected cumulative regret that is uniformly dominated by a logarithmic term.

Foraging has been extensively studied in the behavioral ecology literature [9], [10], [11], [12], [13]. A particular emphasis has been on *optimal foraging theory* [9], [10] that studies foraging behavior based on economic principles. Traditional works [9], [10] in optimal foraging theory have studied the optimal behavior by (i) picking an appropriate currency; (ii) establishing appropriate cost-benefit functions; and (iii) determining the optimal policies. Typically the currency is chosen as the net rate of energy intake and the fundamental hypothesis is that this intake rate is maximized.

This research has been supported in part by ONR grant N00014-09-1-1074 and ARO grant W911NG-11-1-0385. P. Reverdy is supported through an NDSEG Fellowship.

V. Srivastava, P. Reverdy, and N. E. Leonard are with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA 08544 {vaibhavs, preverdy, naomi} @princeton.edu

The fundamental questions studied in optimal foraging theory include (i) which environment patch should the animal visit next? (ii) how long should the animal stay in that patch? and (iii) which foraging path should the animal choose in each patch?

In recent years, a significant focus has been on the macroscopic properties of foraging. It has been observed that *Lévy flights* are efficient search mechanisms, and it has been hypothesized that animal foraging has evolved into a Lévy flight [14], [11]. An alternative macroscopic model to the Lévy flight model is the *intermittent search model* [15]. The intermittent search model views foraging in two alternating phases. In the first phase the animal performs a local Brownian search, and in the second phase the animal performs a ballistic relocation. In both the Lévy flight and intermittent search models, the key macroscopic observation is that the animal performs a local exploration for some time and then moves to a far-off location.

While these macroscopic models capture the general characteristics of foraging well, they do not provide insights into the decision mechanisms used by the animal. There have been significant efforts to understand the decision mechanisms in foraging; see, e.g., [16], [17], [18]. Of particular interest here are the foraging studies in the multi-armed bandit problem setting. Krebs *et al.* [16] studied foraging in great-tits in a two-armed bandit setting and found that the foraging policy of great-tits is close to the optimal policy for the two-armed bandit problem. Keasar [17] explored the foraging behavior of bumblebees in a two-armed bandit setting and discussed plausible decision-making mechanisms.

Contributions: In this paper, we study the multi-armed bandit problem with Gaussian rewards. In animal foraging, the energy aggregated from a patch can be thought of as the reward from the patch, and the animal's objective is to maximize intake energy rate, while minimizing expenditure in time and energy. In robotic foraging, the robot searches an area, and the reward is the aggregated evidence. Analogous to the animal, the robot's objective is typically to maximize evidence collected, while minimizing expenditure of time and energy.

To address this common problem, we consider two particular extensions of the standard multi-armed bandit problem, namely, the multi-armed bandit problem with transition costs, and the graphical multi-armed bandit problem. We justify the need for the extensions as follows. In the standard multi-armed bandit problem, the decision-maker can switch between two arms any number of times; while in the robotic as well as the animal foraging task, a higher number of switches between arms is undesirable because it results in a higher travel time that leads to a smaller energy/evidence aggregation rate and a larger fuel cost. Thus the foraging objective is equivalent to maximizing the aggregated reward while minimizing the switches between the arms; this is addressed by our first extension. Another shortcoming of the standard multi-armed bandit problem is that it assumes that each arm can be directly visited from another arm; while this is true for any convex environment, non-convex environments require extra care. A well known technique to handle non-convex environments is the *occupancy grid* [19] that constructs a graph associated with the non-convex environment. Accordingly, our second extension to the multi-armed bandit problem on graphs enables study of the foraging problem in non-convex environments.

The major contributions of this work are threefold. First, we study the Gaussian multi-armed bandit problem with transition costs and extend the Bayesian-UCB algorithm in [6] to a block allocation strategy that uniformly achieves an expected cumulative regret that is dominated by a logarithmic term and an expected number of transitions between arms that is dominated by a double-logarithmic term. Second, we study the graphical Gaussian multi-armed bandit problem and extend the block allocation strategy to this problem. We show that even for the graphical multi-armed bandit problem, the block allocation strategy uniformly achieves an expected cumulative regret that is dominated by a logarithmic term. Third, we draw connections between animal foraging behavior and the behavior of the proposed policies for the multi-armed bandits. We argue that the multi-armed bandits and the associated block allocation algorithms qualitatively capture the foraging behavior of some animals. In particular, we observe that the multi-armed bandit problem setup has the potential to provide an overarching framework that brings together the classical optimal foraging theory, the Lévy flight based macroscopic search models, and the decision-mechanism based search models.

Paper structure: The remainder of the paper is organized as follows. We review standard Gaussian multi-armed bandits in Section II. The Gaussian multi-armed bandits with transition costs and the graphical Gaussian multi-armed bandits are studied in Section III and IV, respectively. We draw comparisons between the behavior of the block allocation algorithm and animal foraging in Section V and conclude in Section VI.

II. Review of Bandits with Gaussian Rewards

Consider an N-armed bandit problem, i.e., a multi-armed bandit problem with N arms. The reward associated with arm $i \in \{1, \dots, N\}$ is a Gaussian random variable with an unknown mean m_i, and a known variance σ_s^2. The mean of the Gaussian reward at arm i can be interpreted as the signal strength at the arm, while the variance can be interpreted as the sampling noise that is the same at each arm. Let the agent choose arm i_t at time $t \in \{1, \dots, T\}$ and receive a reward $r_t \sim \mathcal{N}(m_{i_t}, \sigma_s^2)$. The decision-maker's objective is to choose a sequence of arms $\{i_t\}_{t \in \{1,\dots,T\}}$ that maximizes the expected cumulative reward $\sum_{t=1}^{T} m_{i_t}$, where T is the horizon length of the sequential allocation process.

For a multi-armed bandit, the expected *regret* at time t is defined by $R_t = m_{i^*} - m_{i_t}$, where $m_{i^*} = \max\{m_i \mid i \in \{1, \dots, N\}\}$. The objective of the decision-maker can be equivalently defined as minimizing the expected cumulative regret defined by $\sum_{t=1}^{T} R_t = \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_i^T]$, where n_i^T is the cumulative number of times option i has been chosen until time T and $\Delta_i = m_{i^*} - m_i$ is the expected regret due to picking arm i instead of arm i^*.

A. Bound on Optimal Performance

Lai and Robbins [2] showed that any asymptotically efficient algorithm for the multi-armed bandit problem must choose suboptimal arms for an expected number of times that is at least logarithmic in time. That is,

$$\mathbb{E}[n_i^T] \geq \Big(\frac{1}{D(p_i||p_{i^*})} + o(1)\Big)\log T,$$

where $o(1) \to 0$ as $T \to +\infty$ and $D(\cdot||\cdot) \mapsto \mathbb{R}_{\geq 0} \cup \{+\infty\}$, is defined by

$$D(p_i||p_{i^*}) = \int p_i(r) \log \frac{p_i(r)}{p_{i^*}(r)} \mathrm{d}r,$$

is the Kullback-Leibler divergence between the reward density p_i of any suboptimal option and the reward density p_{i^*} of the optimal arm. For the Gaussian reward structure considered in this paper, the Kullback-Leibler divergence is equal to $D(p_i||p_{i^*}) = \Delta_i^2/2\sigma_s^2$, and consequently, $\mathbb{E}[n_i^T] \geq (2\sigma_s^2/\Delta_i^2 + o(1))\log T$. This leads to a lower bound on the cumulative regret given by

$$\sum_{t=1}^{T} R_t \geq \sum_{i=1}^{N} \Big(\frac{2\sigma_s^2}{\Delta_i} + o(1)\Big)\log T.$$

B. Upper Credible Limit Algorithm for Gaussian Bandits

Let the prior on the mean reward at arm i be a Gaussian random variable with mean μ_i^0 and variance σ_0^2. We are particularly interested in the case of an uninformative prior, i.e., $\sigma_0^2 \to +\infty$. Let the number of times arm i has been chosen until time t be denoted by n_i^t. Let the empirical mean of the rewards from arm i until time t be $\bar{m}_i^t$. Conditioned on the number of visits n_i^t to arm i and the empirical mean $\bar{m}_i^t$, the posterior distribution of the mean reward (M_i) at arm i at time t is a Gaussian random variable with mean and variance

$$\mu_i^t := \mathbb{E}[M_i|n_i^t, \bar{m}_i^t] = \frac{\delta^2\mu_i^0 + n_i^t\bar{m}_i^t}{\delta^2 + n_i^t}, \text{ and}$$

$$\left(\sigma_i^t\right)^2 := \mathrm{Var}[M_i|n_i^t, \bar{m}_i^t] = \frac{\sigma_s^2}{\delta^2 + n_i^t},$$

respectively, where $\delta^2 = \sigma_s^2/\sigma_0^2$.

The UCL algorithm, proposed in [6], at each (discrete) time t first computes the $(1-1/Kt)$-upper credible limit Q_i^t associated with each arm $i \in \{1,\dots,N\}$ defined by

$$Q_i^t := \mu_i^t + \frac{\sigma_s}{\sqrt{\delta^2 + n_i^t}}\Phi^{-1}\Big(1 - \frac{1}{Kt}\Big),$$

where $K > 0$ is a constant and $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function for the standard normal random variable. The UCL algorithm then selects an arm $i_t :=$ arg $\max\{Q_i^t \mid i \in \{1,\dots,N\}\}$. For the uninformative prior, i.e., $\delta^2 \to 0^+$, the UCL algorithm achieves a logarithmic expected cumulative regret for a multi-armed bandit problem with Gaussian rewards. In particular, the regret satisfies the following uniform upper bound:

$$\sum_{t=1}^{T} R_t^{\mathrm{UCL}} \leq \sum_{i=1}^{N} \Delta_i\Bigg(\Big(\frac{8\beta^2\sigma_s^2}{\Delta_i^2} + \frac{2}{\sqrt{2\pi e}}\Big)\log T + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2 - \log\log T) + 1 + \frac{2}{\sqrt{2\pi e}}\Bigg),$$

where R_t^{UCL} is the regret of the UCL algorithm at time t, and $\beta = 1.02$.

III. Gaussian multi-armed bandits with transition costs

Consider the N-armed bandit problem described in Section II. Suppose the decision-maker incurs a random transition cost $c_{ij} \in \mathbb{R}_{\geq 0}$ for a transition from arm i to arm j. No cost is incurred if the same arm as at the previous time instant is chosen, i.e., $c_{ii} = 0$. Such a cost structure corresponds to a search problem in which the N arms correspond to N spatially distributed regions and the transition cost c_{ij} correspond to the travel cost from region i to region j.

A. The Block UCL Algorithm

For such Gaussian bandits with transition costs, we develop a block allocation strategy that extends the UCL algorithm of Section II-B. To develop this strategy, we divide the set of natural numbers (allocation instances) into frames $\{f_k \mid k \in \mathbb{N}\}$ such that frame f_k starts at time 2^{k-1} and ends at time $2^k - 1$. Thus, the length of frame f_k is 2^{k-1}. We subdivide frame f_k into blocks that we call rounds of allocation. Let the first $\lfloor 2^{k-1}/k \rfloor$ blocks in frame f_k have length k and the remaining allocation instances in frame f_k constitute a single block of length $2^{k-1} - \lfloor 2^{k-1}/k \rfloor k$. The total number of allocation rounds (blocks) in frame f_k is $b_k = \lceil 2^{k-1}/k \rceil$. Let $\ell \in \mathbb{N}$ be the smallest index such that $T < 2^\ell$. Note that each round of allocation is characterized by the tuple (k, r), for some $k \in \{1,\dots,\ell\}$, and $r \in \{1,\dots,b_k\}$. The block UCL algorithm at each round of allocation selects the arm with the maximum upper limit to the smallest $(1 - 1/K\tau_{kr})$-credible set Q_i^{kr} (defined below), where τ_{kr} is the time at allocation round (k, r), and chooses it for the length of that round (block).

B. Regret Analysis of the Block UCL Algorithm

In this section, we analyze the regret of the block UCL algorithm. We first introduce some notation. Let Q_i^{kr} be the maximum upper limit to the smallest $(1 - 1/K\tau_{kr})$-credible set for the mean of arm i at allocation round (k, r), where $K = \sqrt{2\pi e}$ is the credible limit parameter. Let n_i^{kr} be the number of times arm i has been chosen until allocation round (k, r). Let s_i^t be the number of times the decision-maker transitions to arm i from another arm $j \in \{1,\dots,N\} \setminus \{i\}$ until time t. Let the empirical mean of the rewards from arm i until allocation round (k, r) be $\bar{m}_i^{kr}$. Conditioned on the number of visits n_i^{kr} to arm i and the empirical mean $\bar{m}_i^{kr}$, the posterior distribution of the mean reward (M_i) at arm i at allocation round (k, r) is a Gaussian random variable with mean and variance

$$\mu_i^{kr} := \mathbb{E}[M_i|n_i^{kr}, \bar{m}_i^{kr}] = \frac{\delta^2\mu_i^0 + n_i^{kr}\bar{m}_i^{kr}}{\delta^2 + n_i^{kr}}, \text{ and}$$

$${\sigma_i^{kr}}^2 := \mathrm{Var}[M_i|n_i^{kr}, \bar{m}_i^{kr}] = \frac{\sigma_s^2}{\delta^2 + n_i^{kr}},$$

respectively. Moreover,

$$\mathbb{E}[\mu_i^{kr}|n_i^{kr}] = \frac{\delta^2\mu_i^0 + n_i^{kr}m_i}{\delta^2 + n_i^{kr}} \text{ and } \mathrm{Var}[\mu_i^{kr}|n_i^{kr}] = \frac{n_i^{kr}\sigma_s^2}{(\delta^2 + n_i^{kr})^2}.$$

Accordingly, the $(1-\frac{1}{K\tau_{kr}})$-upper credible limit Q_i^{kr} is

$$Q_i^{kr} = \mu_i^{kr} + \frac{\sigma_s}{\sqrt{\delta^2 + n_i^{kr}}}\Phi^{-1}\Big(1 - \frac{1}{K\tau_{kr}}\Big).$$

Also, for each $i \in \{1, \dots, N\}$, we define constants

$$\begin{aligned}
\gamma_1^i &= \frac{8\beta^2\sigma_s^2}{\Delta_i^2} + \frac{1}{\log 2} + \frac{2}{K},\\
\gamma_2^i &= \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2) + 2 + \frac{8}{K} + \frac{\log 4}{K},\\
\gamma_3^i &= \gamma_1^i \log 2(2 - \log\log 2)\\
&\qquad - \Big(\frac{4\beta^2\sigma_s^2\gamma_1^i}{\Delta_i^2}\log\log 2 - \gamma_2^i\Big)\Big(1 + \frac{\pi^2}{6}\Big), \text{ and}\\
\bar{c}_i^{\max} &= \max\{\mathbb{E}[c_{ij}] \mid j \in \{1, \dots, N\}\}.
\end{aligned}$$

Let $\{R_t^{\text{BUCL}}\}_{t\in\{1,\dots,T\}}$ be the sequence of the expected regret of the block UCL algorithm, and $\{S_t^{\text{BUCL}}\}_{t\in\{1,\dots,T\}}$ be the sequence of expected transition costs. The Block UCL algorithm achieves a logarithmic expected cumulative regret as formalized in the following theorem.

*Theorem 1 (**Regret of Block UCL Algorithm**):* The following statements hold for the Gaussian multi-armed bandit problem with transition costs and the block UCL algorithm with an uninformative prior:

(i) the expected number of times a suboptimal arm i is chosen until time T satisfies

$$\mathbb{E}[n_i^T] \leq \gamma_1^i \log T - \frac{4\beta^2\sigma_s^2}{\Delta_i^2}\log\log T + \gamma_2^i;$$

(ii) the expected number of transitions to a suboptimal arm i from another arm until time T satisfies

$$\mathbb{E}[s_i^T] \leq (\gamma_1^i \log 2)\log\log T + \gamma_3^i;$$

(iii) the cumulative regret and the cumulative transition cost until time T satisfy

$$\begin{aligned}
\sum_{t=1}^{T} R_t^{\text{BUCL}} &\leq \sum_{i=1}^{N} \Delta_i\Big(\gamma_1^i \log T - \frac{4\beta^2\sigma_s^2}{\Delta_i^2}\log\log T + \gamma_2^i\Big),\\
\sum_{t=1}^{T} S_t^{\text{BUCL}} &\leq \sum_{i=1, i\neq i^*}^{N} (\bar{c}_i^{\max} + \bar{c}_{i^*}^{\max})\times\\
&\qquad\qquad ((\gamma_1^i \log 2)\log\log T + \gamma_3^i) + \bar{c}_{i^*}^{\max}.
\end{aligned}$$

Proof: See Appendix. ■

IV. Graphical Gaussian Bandits

We now consider multi-armed bandits with Gaussian rewards in which the decision-maker cannot move to every other arm from the current arm. Let the arms that can be visited from arm i be $\text{ne}(i) \subseteq \{1, \dots, N\}$. Such a multi-armed bandit can be represented by a graph $\mathcal{G}$ with node set $\{1, \dots, N\}$ and edge set $\mathcal{E} = \{(i, j) \mid j \in \text{ne}(i), i \in \{1, \dots, N\}\}$. We assume that the graph is connected in the sense that there exists at least one path from each node $i \in \{1, \dots, N\}$ to every other node $j \in \{1, \dots, N\}$.

A. The Graphical Block UCL Algorithm

For the graphical Gaussian bandits, we develop an algorithm similar to the block allocation algorithm, namely, the graphical block UCL algorithm. Similar to the block allocation algorithm, at each comparison block, the arm with the maximum upper credible limit is determined. Since the arm with the maximum upper credible limit may not be immediately reached from the current arm, the graphical block UCL algorithm traverses a shortest path from the current arm to the arm with the maximum upper credible limit. The key intuition behind the algorithm is that the block allocation strategy results in an expected number of transitions that is sub-logarithmic in the horizon length. In the context of graphical bandits, sub-logarithmic transitions result in sub-logarithmic *undesired* visits to the arms on the chosen shortest path to the *desired* arm. Consequently, the regret of the algorithm is dominated by the logarithmic term.

B. Regret Analysis of the Graphical Block UCL Algorithm

We now analyze the performance of the graphical block UCL algorithm. Let $\{R_t^{\text{GUCL}}\}_{t\in\{1,\dots,T\}}$ be the sequence of expected regret of the graphical block UCL algorithm. The graphical block UCL algorithm achieves a logarithmic expected cumulative regret as formalized in the following theorem.

*Theorem 2 (**Regret of Graphical Block UCL Algorithm**):* The following statements hold for the graphical Gaussian multi-armed bandit problem with the graphical block UCL algorithm and an uninformative prior:

(i) the expected number of times a suboptimal arm i is chosen until time T satisfies

$$\begin{aligned}
\mathbb{E}[n_i^T] &\leq \gamma_1^i \log T - \frac{4\beta^2\sigma_s^2}{\Delta_i^2}\log\log T + \gamma_2^i\\
&\quad + \sum_{i=1, i\neq i^*}^{N} \big((2\gamma_1^i \log 2)\log\log T + 2\gamma_3^i\big) + 1;
\end{aligned}$$

(ii) the cumulative regret until time T satisfies

$$\begin{aligned}
\sum_{t=1}^{T} R_t^{\text{GUCL}} &\leq \sum_{i=1}^{N} \Big(\gamma_1^i \log T - \frac{4\beta^2\sigma_s^2}{\Delta_i^2}\log\log T\\
&\quad + \gamma_2^i + \sum_{i=1, i\neq i^*}^{N} \big((2\gamma_1^i \log 2)\log\log T + 2\gamma_3^i\big) + 1\Big)\Delta_i;
\end{aligned}$$

Proof: See Appendix. ■

V. Comparison with Animal Foraging

In this section, we compare the behavior of the block allocation algorithm for the multi-armed bandits with the animal foraging behavior reported in the literature. Consider the foraging environment as composed of patches and each patch has sources of energy that are modeled by Gaussian random variables with an unknown mean and a known variance. The exploration-exploitation tradeoff in the foraging problem can be modeled by the multi-armed bandit problem. In particular, the foraging objective of animals is to maximize the net energy accumulation rate which in the multi-armed bandit setting maps to maximizing the expected cumulative reward while minimizing the travel time, i.e., minimizing the number of transitions among arms.

The solution to the multi-armed bandit problem naturally answers the first two fundamental questions studied in optimal foraging theory: (i) which environment patch should the animal visit next? (ii) how long should the animal stay in that patch? Although the solution to the multi-armed bandit problem does not answer the third fundamental question: which foraging path should the animal choose in each patch? To understand the third question, it is natural to envision that points within a patch are highly correlated in terms of the energy accumulation, i.e., each point within a patch provides energy at somewhat the same rate, and accordingly the energy can be accumulated, e.g., via an ergodic random walk.

For simplicity of analysis, in this paper, we assume that the arms are uncorrelated and the prior is uninformative. In general, the prior may be informative and arms may be correlated. The algorithm proposed in this paper extends to this case by simply replacing the N univariate inference procedures with an N-variate inference procedure. The correlation structure captures the structure of the environment: higher correlation describes a smoother environment, while lower correlation describes a rougher environment.

In a sufficiently correlated environment, the block allocation algorithm at allocation round (k, r) picks an arm with highest value of Q_i^{kr} and samples it k times. At the subsequent allocation instance, due to the correlation structure the uncertainty in the estimates for the nearby locations will go down while the uncertainty in the far-off locations would remain high. Consequently, the component of Q_i^{kr} associated with the width of the credible set will be higher for the far-off locations than the nearby locations. If the prior means are assumed to be uniform, the block allocation strategy at the next allocation instance will select a location far-off from the current location. This is a central feature of the macroscopic foraging models, including the Lévy flight model and the intermittent search model. Thus, the Bayesian multi-armed bandit problem and the associated block allocation strategy qualitatively captures the behavior of Lévy flights and related macroscopic models for search.

Overall, the multi-armed bandit problem with transition costs models the fundamental foraging objective as defined in the optimal foraging literature, and its solution yields search trajectories akin to those described by macroscopic search models. Moreover, the solution to the multi-armed bandit problem with transition costs naturally provides the decision mechanisms involved with the search process. Therefore, the multi-armed bandit problem setup has the potential to provide an overarching framework that brings together the classical optimal foraging theory, the Lévy flight based macroscopic search models, and the decision-mechanism based search models.

VI. Conclusions

We studied two variations of the Gaussian multi-armed bandit problem, namely, the Gaussian multi-armed bandit problem with transition cost, and the graphical Gaussian multi-armed bandit problem and developed block allocation algorithms that uniformly achieve an expected cumulative regret dominated by a logarithmic function of time, and a number of expected cumulative transitions among the arms dominated by a double-logarithmic function of time. We drew some qualitative connections between foraging behavior of some animals and the behavior of the block allocation algorithm. In particular, we argued that the multi-armed bandit problem models the foraging objective in optimal foraging theory well and the associated block allocation strategy captures the key features of popular macroscopic search models.

At this stage, we observe and point out the potential of the multi-armed bandit problem and the associated block allocation algorithm to bridge the gap between classical optimal foraging theory and recent macroscopic search models. This suggests an exciting new avenue of inquiry in which the bandit model may prove valuable for future study of animal foraging. In the future, we plan to investigate the bandit model more extensively in the context of empirical work on both animal and robotic foraging.

Appendix

A. Proof of regret of the block UCL algorithm

Proof of Theorem 1: We start by establishing the first statement. For a given t, let (k_t, r_t) be the lexicographically maximum tuple such that $\tau_{k_t r_t} \leq t$. We note that

$$\begin{aligned} n_i^T &= \sum_{t=1}^{T} \mathbf{1}(i_t = i) \\ &= \sum_{t=1}^{T} \big(\mathbf{1}(i_t = i \;\&\; n_i^{k_t r_t} < \eta) + \mathbf{1}(i_t = i \;\&\; n_i^{k_t r_t} \geq \eta)\big) \\ &\leq \eta + \ell + \sum_{t=1}^{T} \mathbf{1}(i_t = i \;\&\; n_i^{k_t r_t} \geq \eta) \\ &\leq \eta + \ell + \sum_{k=1}^{\ell} \sum_{r=1}^{b_k} k\mathbf{1}(i_{\tau_{kr}} = i \;\&\; n_i^{kr} \geq \eta). \qquad (1) \end{aligned}$$

It can be shown (see [20] for details) that if we choose $\eta = \lceil \frac{8\beta^2\sigma_s^2}{\Delta_i^2}(\log T - \frac{1}{2}\log\log T) + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2)\rceil$, then

$$\mathbb{E}[n_i^T] \leq \eta + \ell + \frac{2}{K}\sum_{k=1}^{\ell}\sum_{r=1}^{b_k}\frac{k}{\tau_{kr}}. \qquad (2)$$

We now focus on the term $\sum_{k=1}^{\ell}\sum_{r=1}^{b_k}\frac{k}{\tau_{kr}}$. We note that $\tau_{kr} = 2^{k-1} + (r-1)k$, and hence

$$\begin{aligned} \sum_{r=1}^{b_k}\frac{k}{\tau_{kr}} &= \sum_{r=1}^{b_k}\frac{k}{2^{k-1} + (r-1)k} \\ &\leq \frac{k}{2^{k-1}} + \int_1^{b_k}\frac{k}{k(x-1) + 2^{k-1}}\mathrm{d}x \\ &\leq \frac{k}{2^{k-1}} + \log 2. \qquad (3) \end{aligned}$$

Since $T \geq 2^{\ell-1}$, it follows that $\ell \leq 1 + \log_2 T =: \bar{\ell}$. Therefore, inequalities (2) and (3) yield

$$
\begin{aligned}
\mathbb{E}[n_i^T] &\leq \eta + \bar{\ell} + \frac{2}{K}\sum_{k=1}^{\bar{\ell}}\Big(\frac{k}{2^{k-1}} + \log 2\Big) \\
&\leq \eta + \bar{\ell} + \frac{8}{K} + \frac{2\log 2}{K}\bar{\ell} \\
&\leq \gamma_1^i \log T - \frac{4\beta^2\sigma_s^2}{\Delta_i^2}\log\log T + \gamma_2^i.
\end{aligned}
$$

We now establish the second statement. In the spirit of [7], we note that the number of times the decision-maker transitions to arm i from another arm in frame f_k is equal to the number of times arm i is selected in frame k divided by the length of each block is frame f_k. Consequently,

$$
\begin{aligned}
s_i^T &\leq \sum_{k=1}^{\ell}\frac{n_i^{2^k} - n_i^{2^{k-1}}}{k} = \sum_{k=1}^{\ell}\frac{n_i^{2^k}}{k} - \sum_{k=1}^{\ell-1}\frac{n_i^{2^k}}{k+1} \\
&= \frac{n_i^{2^\ell}}{\ell} + \sum_{k=1}^{\ell-1} n_i^{2^k}\Big(\frac{1}{k} - \frac{1}{k+1}\Big) \leq \frac{n_i^{2^\ell}}{\ell} + \sum_{k=1}^{\ell-1}\frac{n_i^{2^k}}{k^2}.
\end{aligned}
$$

Therefore, it follows that

$$
\mathbb{E}[s_i^T] \leq \frac{\mathbb{E}[n_i^{2^\ell}]}{\ell} + \sum_{k=1}^{\ell-1}\frac{\mathbb{E}[n_i^{2^k}]}{k^2}. \tag{4}
$$

Substituting $\mathbb{E}[n_i^{2^k}]$ in inequality (4) with the derived upper bounds and performing some algebraic manipulations, we obtain

$$
\mathbb{E}[s_i^T] \leq (\gamma_1^i \log 2)\log\log T + \gamma_3^i.
$$

We now establish the last statement. The bound of the cumulative regret follows from the definition and the first statement. To establish the bound on the cumulative switching cost, we note that

$$
\begin{aligned}
\sum_{t=1}^{T} S_t^{\text{BUCL}} &\leq \sum_{i=1, i\neq i^*}^{N} \bar{c}_i^{\max}\mathbb{E}[s_i^T] + \bar{c}_{i^*}^{\max}\mathbb{E}[s_{i^*}^T] \\
&\leq \sum_{i=1, i\neq i^*}^{N} (\bar{c}_i^{\max} + \bar{c}_{i^*}^{\max})\mathbb{E}[s_i^T] + \bar{c}_{i^*}^{\max}, \qquad (5)
\end{aligned}
$$

where the second inequality follows from the observation that $s_{i^*}^T \leq \sum_{i=1, i\neq i^*}^{T} s_i^T + 1$. The final expression follows from inequality (5) and the second statement. □

B. Proof of regret of the graphical block UCL algorithm

Proof of Theorem 2: We start by establishing the first statement. We classify the selection of arms in two categories, namely, the *goal* selection and the *transient* selection. The goal selection of an arm corresponds to the situation in which the arm has the maximum upper credible limit, while the transient selection corresponds to the situation in which the arm is selected because it belongs to the chosen shortest path to the arm with the maximum credible limit. We note that due to transient selections, the number of frames until time T are at most equal to the number of frames if there are no transient selections. Consequently, the expected number of goal selections of a suboptimal arm i are upper bounded by the expected number of selections of arm i in the block allocation algorithm, i.e.,

$$
\mathbb{E}[n_{\text{goal},i}^T] \leq \gamma_1^i \log T - \frac{4\beta^2\sigma_s^2}{\Delta_i^2}\log\log T + \gamma_2^i.
$$

Moreover, the number of transient selections of arm i are upper bounded by the total number of transitions from an arm to another arm in the block allocation algorithm, i.e.,

$$
\mathbb{E}[n_{\text{transient},i}^T] \leq \sum_{i=1, i\neq i^*}^{N}\big((2\gamma_1^i \log 2)\log\log T + 2\gamma_3^i\big) + 1.
$$

The expected number of selections of arm i is the sum of the expected number of transient selections and the expected number of goal selections, and thus the first statement follows.

The second statement follows immediately from the definition of the cumulative regret. □

References

[1] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

[2] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[4] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

[5] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Int. Conf. on Artificial Intelligence and Statistics*, pages 592–600, La Palma, Canary Islands, Spain, April 2012.

[6] P. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision-making in multi-armed bandits. In *Multidisciplinary Conf. on Reinforcement Learning and Decision Making*, Princeton, NJ, USA, Oct 2013.

[7] R. Agrawal, M. V. Hedge, and D. Teneketzis. Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10):899–906, 1988.

[8] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2-3):245–272, 2010.

[9] G. H. Pyke, H. R. Pulliam, and E. L. Charnov. Optimal foraging: a selective review of theory and tests. *Quarterly Review of Biology*, pages 137–154, 1977.

[10] D. W. Stephens, , and J. R. Krebs. *Foraging theory*. Princeton University Press, 1986.

[11] G. M. Viswanathan, M. G. E. da Luz, E. P. Raposo, and H. E. Stanley. *The Physics of Foraging: An Introduction to Random Searches and Biological Encounters*. Cambridge University Press, 2011.

[12] E. Gelenbe, N. Schmajuk, J. Staddon, and J. Reif. Autonomous search by robots and animals: A survey. *Robotics and Autonomous Systems*, 22(1):23–34, 1997.

[13] S. R. X. Dall, L. Giraldeau, O. Olsson, J. M. McNamara, and D. W. Stephens. Information and its use by animals in evolutionary ecology. *Trends in Ecology & Evolution*, 20(4):187–193, 2005.

[14] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. G. E. da Luz, E. P. Raposo, and H. E. Stanley. Optimizing the success of random searches. *Nature*, 401(6756):911–914, 1999.

[15] O. Bénichou, C. Loverdo, M. Moreau, and R. Voituriez. Intermittent search strategies. *Reviews of Modern Physics*, 83(1):81, 2011.

[16] J. R. Krebs, A. Kacelnik, and P. Taylor. Test of optimal sampling by foraging great tits. *Nature*, 275(5675):27–31, 1978.

[17] T. Keasar, E. Rashkovich, D. Cohen, and A. Shmida. Bees in two-armed bandit situations: Foraging choices and possible decision mechanisms. *Behavioral Ecology*, 13(6):757–765, 2002.

[18] A. M. Hein and S. A. McKinley. Sensing and decision-making in random search. *Proceedings of the National Academy of Sciences*, 109(30):12070–12074, 2012.

[19] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT Press, 2005.

[20] P. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision-making in generalized Gaussian multi-armed bandits. *arXiv preprint arXiv:1307.6134*, July 2013.