Modeling Human Decision-making in Generalized Gaussian Multi-armed Bandits

Paul Reverdy Vaibhav Srivastava Naomi Ehrich Leonard

***Abstract*—We present a formal model of human decision-making in explore-exploit tasks using the context of multi-armed bandit problems, where the decision-maker must choose among multiple options with uncertain rewards. We address the standard multi-armed bandit problem, the multi-armed bandit problem with transition costs, and the multi-armed bandit problem on graphs. We focus on the case of Gaussian rewards in a setting where the decision-maker uses Bayesian inference to estimate the reward values. We model the decision-maker's prior knowledge with the Bayesian prior on the mean reward. We develop the upper credible limit (UCL) algorithm for the standard multi-armed bandit problem and show that this deterministic algorithm achieves logarithmic cumulative expected regret, which is optimal performance for uninformative priors. We show how good priors and good assumptions on the correlation structure among arms can greatly enhance decision-making performance, even over short time horizons. We extend to the stochastic UCL algorithm and draw several connections to human decision-making behavior. We present empirical data from human experiments and show that human performance is efficiently captured by the stochastic UCL algorithm with appropriate parameters. For the multi-armed bandit problem with transition costs and the multi-armed bandit problem on graphs, we generalize the UCL algorithm to the block UCL algorithm and the graphical block UCL algorithm, respectively. We show that these algorithms also achieve logarithmic cumulative expected regret and require a sub-logarithmic expected number of transitions among arms. We further illustrate the performance of these algorithms with numerical examples.**

***Index Terms*—multi-armed bandit, human decision-making, machine learning, adaptive control**

I. Introduction

Imagine the following scenario: you are reading the menu in a new restaurant, deciding which dish to order. Some of the dishes are familiar to you, while others are completely new. Which dish do you ultimately order: a familiar one that you are fairly certain to enjoy, or an unfamiliar one that looks interesting but you may dislike?

Your answer will depend on a multitude of factors, including your mood that day (Do you feel adventurous or conservative?), your knowledge of the restaurant and its cuisine (Do you know little about African cuisine, and everything looks new to you?), and the number of future decisions the outcome is likely to influence (Is this a restaurant in a foreign city you are unlikely to visit again, or is it one that has newly opened close to home, where you may return many times?). This scenario encapsulates many of the difficulties faced by a decision-making agent interacting with his/her environment, e.g. the role of prior knowledge and the number of future choices (time horizon).

The problem of learning the optimal way to interact with an uncertain environment is common to a variety of areas of study in engineering such as adaptive control and reinforcement learning [3]. Fundamental to these problems is the tradeoff between exploration (collecting more information to reduce uncertainty) and exploitation (using the current information to maximize the immediate reward). Formally, such problems are often formulated as Markov Decision Processes (MDPs). MDPs are decision problems in which the decision-making agent is required to make a sequence of choices along a process evolving in time [4]. The theory of dynamic programming [5], [6] provides methods to find optimal solutions to generic MDPs, but is subject to the so-called *curse of dimensionality* [4], where the size of the problem often grows exponentially in the number of states.

The curse of dimensionality makes finding the optimal solution difficult, and in general intractable for finite-horizon problems of any significant size. Many engineering solutions of MDPs consider the infinite-horizon case, i.e., the limit where the agent will be required to make an infinite sequence of decisions. In this case, the problem simplifies significantly and a variety of reinforcement learning methods can be used to converge to the optimal solution, for example [7], [6], [4], [3]. However, these methods only converge to the optimal solution asymptotically at a rate that is difficult to analyze. The UCRL algorithm [8] addressed this issue by deriving a heuristic-based reinforcement learning algorithm with a provable learning rate.

However, the infinite-horizon limit may be inappropriate for finite-horizon tasks. In particular, optimal solutions to the finite-horizon problem may be strongly dependent on the task horizon. Consider again our restaurant scenario. If the decision is a one-off, we are likely to be conservative, since selecting an unfamiliar option is risky and even if we choose an unfamiliar dish and like it, we will have no further opportunity to use the information in the same context. However, if we are likely to return to the restaurant many times in the future, discovering new dishes we enjoy is valuable.

Although the finite-horizon problem may be intractable to computational analysis, humans are confronted with it all the time, as evidenced by our restaurant example. The fact that they are able to find efficient solutions quickly with inherently limited computational power suggests that

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. Preliminary versions of parts of this work were presented at IEEE CDC 2012 [1] and Allerton 2013 [2]. In addition to improving on the ideas in [1], [2], this paper improves the analysis of algorithms and compares the performance of these algorithms against empirical data. The human behavioral experiments were approved under Princeton University Institutional Review Board protocol number 4779.

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

humans employ relatively sophisticated heuristics for solving these problems. Elucidating these heuristics is of interest both from a psychological point of view where they may help us understand human cognitive control and from an engineering point of view where they may lead to development of improved algorithms to solve MDPs [9]. In this paper, we seek to elucidate the behavioral heuristics at play with a model that is both mathematically rigorous and computationally tractable.

Multi-armed bandit problems [10] constitute a class of MDPs that is well suited to our goal of connecting biologically plausible heuristics with mathematically rigorous algorithms. In the mathematical context, multi-armed bandit problems have been studied in both the infinite-horizon and finite-horizon cases. There is a well-known optimal solution to the infinite-horizon problem [11]. For the finite-horizon problem, the policies are designed to match the best possible performance established in [12]. In the biological context, the decision-making behavior and performance of both animals and humans have been studied using the multi-armed bandit framework.

In a multi-armed bandit problem, 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. We will refer to this as the *standard* multi-armed bandit problem, and we will consider variations that add transition costs or spatial unavailability of arms. A classical example of a standard multi-armed bandit problem is the evaluation of clinical trials with medical patients described in [13]. The decision-maker is a doctor and the options are different treatments with unknown effectiveness for a given disease. Given patients that arrive and get treated sequentially, the objective for the doctor is to maximize the number of cured patients, using information gained from successive outcomes.

Multi-armed bandit problems capture the fundamental exploration-exploitation tradeoff. Indeed, they model a wide variety of real-world decision-making scenarios including those associated with foraging and search in an uncertain environment. The rigorous examination in the present paper of the heuristics that humans use in multi-armed bandit tasks can help in understanding and enhancing both natural and engineered strategies and performance in these kinds of tasks. For example, a trained human operator can quickly learn the relevant features of a new environment, and an efficient model for human decision-making in a multi-armed bandit task may facilitate a means to learn a trained operator's task-specific knowledge for use in an autonomous decision-making algorithm. Likewise, such a model may help in detecting weaknesses in a human operator's strategy and deriving computational means to augment human performance.

Multi-armed bandit problems became popular following the seminal paper by Robbins [14] and found application in diverse areas including controls, robotics, machine learning, economics, ecology, and operational research [15], [16], [17], [18], [19]. For example, in ecology the multi-armed bandit problem was used to study the foraging behavior of birds in an unknown environment [20]. The authors showed that the optimal policy for the two-armed bandit problem captures well the observed foraging behavior of birds. Given the limited computational capacity of birds, it is likely they use simple heuristics to achieve near-optimal performance. The development of simple heuristics in this and other contexts has spawned a wide literature.

Gittins [11] studied the infinite-horizon multi-armed bandit problem and developed a dynamic allocation index (Gittins' index) for each arm. He showed that selecting an arm with the highest index at the given time results in the optimal policy. The dynamic allocation index, while a powerful idea, suffers from two drawbacks: (i) it is hard to compute, and (ii) it does not provide insight into the nature of the optimal policies.

Much recent work on multi-armed bandit problems focuses on a quantity termed *cumulative expected regret*. The cumulative expected regret of a sequence of decisions is simply the cumulative difference between the expected reward of the options chosen and the maximum reward possible. In this sense, expected regret plays the same role as expected value in standard reinforcement learning schemes: maximizing expected value is equivalent to minimizing cumulative expected regret. Note that this definition of regret is in the sense of an omniscient being who is aware of the expected values of all options, rather than in the sense of an agent playing the game. As such, it is not a quantity of direct psychological relevance but rather an analytical tool that allows one to characterize performance.

In a ground-breaking work, Lai and Robbins [12] established a logarithmic lower bound on the expected number of times a sub-optimal arm needs to be sampled by an optimal policy, thereby showing that cumulative expected regret is bounded below by a logarithmic function of time. Their work established the best possible performance of any solution to the standard multi-armed bandit problem. They also developed an algorithm based on an upper confidence bound on estimated reward and showed that this algorithm achieves the performance bound asymptotically. In the following, we use the phrase *logarithmic regret* to refer to cumulative expected regret being bounded above by a logarithmic function of time, i.e., having the same order of growth rate as the optimal solution. The calculation of the upper confidence bounds in [12] involves tedious computations. Agarwal [21] simplified these computations to develop sample mean-based upper confidence bounds, and showed that the policies in [12] with these upper confidence bounds achieve logarithmic regret asymptotically.

In the context of bounded multi-armed bandits, i.e., multi-armed bandits in which the reward is sampled from a distribution with a bounded support, Auer *et al.* [22] developed upper confidence bound-based algorithms that achieve logarithmic regret uniformly in time; see [23] for an extensive survey of upper confidence bound-based algorithms. Audibert *et al.* [24] considered upper confidence bound-based algorithms that take into account the empirical variance of the various arms. In a related work, Cesa-Bianchi *et al.* [25] analyzed

a Boltzman allocation rule for bounded multi-armed bandit problems. Garivier *et al.* [26] studied the KL-UCB algorithm, which uses upper confidence bounds based on the Kullback-Leibler divergence, and advocated its use in multi-armed bandit problems where the rewards are distributed according to a known exponential family.

The works cited above adopt a frequentist perspective, but a number of researchers have also considered MDPs and multi-armed bandit problems from a Bayesian perspective. Dearden *et al.* [27] studied general MDPs and showed that a Bayesian approach can substantially improve performance in some cases. Recently, Srinivas *et al.* [28] developed asymptotically optimal upper confidence bound-based algorithms for Gaussian process optimization. Agrawal *et al.* [29] proved that a Bayesian algorithm known as Thompson Sampling is near-optimal for binary bandits with a uniform prior. Kauffman *et al.* [30] developed a generic Bayesian upper confidence bound-based algorithm and established its optimality for binary bandits with a uniform prior. In the present paper we develop a similar Bayesian upper confidence bound-based algorithm for Gaussian multi-armed bandit problems and show that it achieves logarithmic regret for uninformative priors uniformly in time.

Some variations of these multi-armed bandit problems have been studied as well. Agarwal *et al.* [31] studied multi-armed bandit problems with transition costs, i.e., the multi-armed bandit problems in which a certain penalty is imposed each time the decision-maker switches from the currently selected arm. To address this problem, they developed an asymptotically optimal block allocation algorithm. Banks and Sundaram [32] show that, in general, it is not possible to define dynamic allocation indices (Gittins' indices) which lead to an optimal solution of the multi-armed bandit problem with switching costs. However, if the cost to switch to an arm from any other arm is a stationary random variable, then such indices exist. Asawa and Teneketzis [33] characterize qualitative properties of the optimal solution to the multi-armed bandit problem with switching costs, and establish sufficient conditions for the optimality of limited lookahead based techniques. A survey of multi-armed bandit problems with switching costs is presented in [34]. In the present paper, we consider Gaussian multi-armed bandit problems with transition costs and develop a block allocation algorithm that achieves logarithmic regret for uninformative priors uniformly in time. Our block allocation scheme is similar to the scheme in [31]; however, our scheme incurs a smaller expected cumulative transition cost than the scheme in [31]. Moreover, an asymptotic analysis is considered in [31], while our results hold uniformly in time.

Kleinberg *et al.* [35] considered multi-armed bandit problems in which arms are not all available for selection at each time (sleeping experts) and analyzed the performance of upper confidence bound-based algorithms. In contrast to the temporal unavailability of arms in [35], we consider a spatial unavailability of arms. In particular, 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 problems that achieves logarithmic regret for uninformative priors uniformly in time.

Human decision-making in multi-armed bandit problems has also been studied in the cognitive psychology literature. Cohen *et al.* [9] surveyed the exploration-exploitation tradeoff in humans and animals and discussed the mechanisms in the brain that mediate this tradeoff. Acuña *et al.* [36] studied human decision-making in multi-armed bandits from a Bayesian perspective. They modeled the human subject's prior knowledge about the reward structure using conjugate priors to the reward distribution. They concluded that a policy using Gittins' index, computed from approximate Bayesian inference based on limited memory and finite step look-ahead, captures the empirical behavior in certain multi-armed bandit tasks. In a subsequent work [37], they showed that a critical feature of human decision-making in multi-armed bandit problems is structural learning, i.e., humans learn the correlation structure among different arms.

Steyvers *et al.* [38] considered Bayesian models for multi-armed bandits parametrized by human subjects' assumptions about reward distributions and observed that there are individual differences that determine the extent to which people use optimal models rather than simple heuristics. In a subsequent work, Lee *et al.* [39] considered latent models in which there is a latent mental state that determines if the human subject should explore or exploit. Zhang *et al.* [40] considered multi-armed bandits with Bernoulli rewards and concluded that, among the models considered, the knowledge gradient algorithm best captures the trial-by-trial performance of human subjects.

Wilson *et al.* [41] studied human performance in two-armed bandit problems and showed that at each arm selection instance the decision is based on a linear combination of the estimate of the mean reward of each arm and an ambiguity bonus that depends on the value of the information from that arm. Tomlin *et al.* [42] studied human performance on multi-armed bandits that are located on a spatial grid; at each arm selection instance, the decision-maker can only select the current arm or one of the neighboring arms.

In this paper, we study multi-armed bandits with Gaussian rewards in a Bayesian setting, and we develop upper credible limit (UCL)-based algorithms that achieve efficient performance. We propose a deterministic UCL algorithm and a stochastic UCL algorithm for the standard multi-armed bandit problem. We propose a block UCL algorithm and a graphical block UCL algorithm for the multi-armed bandit problem with transitions costs and the multi-armed problem on graphs, respectively. We analyze the proposed algorithms in terms of the cumulative expected regret, i.e., the cumulative difference between the expected received reward and the maximum expected reward that could have been received. We compare human performance in multi-armed bandit tasks with the performance of the proposed stochastic UCL algorithm and show that the algorithm with the right choice of parameters efficiently models human decision-making performance. The major contributions of this work are fourfold.

First, we develop and analyze the deterministic UCL al-

gorithm for multi-armed bandits with Gaussian rewards. We derive a novel upper bound on the inverse cumulative distribution function for the standard Gaussian distribution, and we use it to show that for an uninformative prior on the rewards, the proposed algorithm achieves logarithmic regret. To the best of our knowledge, this is the first confidence bound-based algorithm that provably achieves logarithmic cumulative expected regret uniformly in time for multi-armed bandits with Gaussian rewards.

We further define a *quality* of priors on rewards and show that for small values of this quality, i.e., good priors, the proposed algorithm achieves logarithmic regret uniformly in time. Furthermore, for good priors with small variance, a slight modification of the algorithm yields sub-logarithmic regret uniformly in time. Sub-logarithmic refers to a rate of expected regret that is even slower than logarithmic, and thus performance is better than with uninformative priors. For large values of the quality, i.e., bad priors, the proposed algorithm can yield performance significantly worse than with uninformative priors. Our analysis also highlights the impact of the correlation structure among the rewards from different arms on the performance of the algorithm as well as the performance advantage when the prior includes a good model of the correlation structure.

Second, to capture the inherent noise in human decision-making, we develop the stochastic UCL algorithm, a stochastic arm selection version of the deterministic UCL algorithm. We model the stochastic arm selection using softmax arm selection [4], and show that there exists a feedback law for the cooling rate in the softmax function such that for an uninformative prior the stochastic arm selection policy achieves logarithmic regret uniformly in time.

Third, we compare the stochastic UCL algorithm with the data obtained from our human behavioral experiments. We show that the observed empirical behaviors can be reconstructed by varying only a few parameters in the algorithm.

Fourth, we study the multi-armed bandit problem with transition costs in which a stationary random cost is incurred each time an arm other than the current arm is selected. We also study the graphical multi-armed bandit problem in which the arms are located at the vertices of a graph and only the current arm and its neighbors can be selected at each time. For these multi-armed bandit problems, we extend the deterministic UCL algorithm to block allocation algorithms that for uninformative priors achieve logarithmic regret uniformly in time.

In summary, the main contribution of this work is to provide a formal algorithmic model (the UCL algorithms) of choice behavior in the exploration-exploitation tradeoff using the context of the multi-arm bandit problem. In relation to cognitive dynamics, we expect that this model could be used to explain observed choice behavior and thereby quantify the underlying computational anatomy in terms of key model parameters. The fitting of such models of choice behavior to empirical performance is now standard in cognitive neuroscience. We illustrate the potential of our model to categorize individuals in terms of a small number of model parameters by showing that the stochastic UCL algorithm can reproduce canonical classes of performance observed in large numbers of subjects.

The remainder of the paper is organized as follows. The standard multi-armed bandit problem is described in Section II. The salient features of human decision-making in bandit tasks are discussed in Section III. In Section IV we propose and analyze the regret of the deterministic UCL and stochastic UCL algorithms. In Section V we describe an experiment with human participants and a spatially-embedded multi-armed bandit task. We show that human performance in that task tends to fall into one of several categories, and we demonstrate that the stochastic UCL algorithm can capture these categories with a small number of parameters. We consider an extension of the multi-armed bandit problem to include transition costs and describe and analyze the block UCL algorithm in Section VI. In Section VII we consider an extension to the graphical multi-armed bandit problem, and we propose and analyze the graphical block UCL algorithm. Finally, in Section VIII we conclude and present avenues for future work.

II. A REVIEW OF MULTI-ARMED BANDIT PROBLEMS

Consider a set of N options, termed *arms* in analogy with the lever of a slot machine. A single-levered slot machine is termed a *one-armed bandit*, so the case of N options is often called an N-armed bandit. The N-armed bandit problem refers to the choice among the N options that a decision-making agent should make to maximize the cumulative reward.

The agent collects reward $r_t \in \mathbb{R}$ by choosing arm i_t at each time $t \in \{1, \dots, T\}$, where $T \in \mathbb{N}$ is the horizon length for the sequential decision process. The reward from option $i \in \{1, \dots, N\}$ is sampled from a stationary distribution p_i and has an unknown mean $m_i \in \mathbb{R}$. The decision-maker's objective is to maximize the cumulative expected reward $\sum_{t=1}^{T} m_{i_t}$ by selecting a sequence of arms $\{i_t\}_{t\in\{1,\dots,T\}}$. Equivalently, defining $m_{i^*} = \max\{m_i \mid i \in \{1, \dots, N\}\}$ and $R_t = m_{i^*} - m_{i_t}$ as the expected *regret* at time t, the objective can be formulated as minimizing the cumulative expected regret defined by

$$\sum_{t=1}^{T} R_t = T m_{i^*} - \sum_{i=1}^{N} m_i \mathbb{E}\left[n_i^T\right] = \sum_{i=1}^{N} \Delta_i \mathbb{E}\left[n_i^T\right],$$

where n_i^T is the total 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^*. Note that in order to minimize the cumulative expected regret, it suffices to minimize the expected number of times any suboptimal option $i \in \{1, \dots, N\} \setminus \{i^*\}$ is selected.

The multi-armed bandit problem is a canonical example of the exploration-exploitation tradeoff common to many problems in controls and machine learning. In this context, at time t, exploitation refers to picking arm i_t that is estimated to have the highest mean at time t, and exploration refers to picking any other arm. A successful policy balances the exploration-exploitation tradeoff by exploring enough to learn which arm is most rewarding and exploiting that information by picking the best arm often.

A. Bound on optimal performance

Lai and Robbins [12] showed that, for any algorithm solving the multi-armed bandit problem, the expected number of times a suboptimal arm is selected is at least logarithmic in time, i.e.,

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

for each $i \in \{1, \ldots, N\} \setminus \{i^*\}$, where $o(1) \to 0$ as $T \to +\infty$. $D(p_i||p_{i^*}) := \int p_i(r) \log \frac{p_i(r)}{p_{i^*}(r)} \mathrm{d}r$ is the Kullback-Leibler divergence between the reward density p_i of any suboptimal arm and the reward density p_{i^*} of the optimal arm. The bound on $\mathbb{E}\left[n_i^T\right]$ implies that the cumulative expected regret must grow at least logarithmically in time.

B. The Gaussian multi-armed bandit task

For the Gaussian multi-armed bandit problem considered in this paper, the reward density p_i is Gaussian with mean m_i and variance σ_s^2. The variance σ_s^2 is assumed known, e.g., from previous observations or known characteristics of the reward generation process. Therefore

$$D(p_i||p_{i^*}) = \frac{\Delta_i^2}{2\sigma_s^2}, \tag{2}$$

and accordingly, the bound (1) is

$$\mathbb{E}\left[n_i^T\right] \geq \left(\frac{2\sigma_s^2}{\Delta_i^2} + o(1)\right) \log T. \tag{3}$$

The insight from (3) is that for a fixed value of σ_s, a suboptimal arm i with higher Δ_i is easier to identify, and thus chosen less often, since it yields a lower average reward. Conversely, for a fixed value of Δ_i, higher values of σ_s make the observed rewards more variable, and thus it is more difficult to distinguish the optimal arm i^* from the suboptimal ones.

C. The Upper Confidence Bound algorithms

For multi-armed bandit problems with bounded rewards, Auer *et al.* [22] developed upper confidence bound-based algorithms, known as the UCB1 algorithm and its variants, that achieve logarithmic regret uniformly in time. UCB1 is a heuristic-based algorithm that at each time t computes a heuristic value Q_i^t for each option i. This value provides an upper bound for the expected reward to be gained by selecting that option:

$$Q_i^t = \bar{m}_i^t + C_i^t, \tag{4}$$

where $\bar{m}_i^t$ is the empirical mean reward and C_i^t is a measure of uncertainty in the reward of arm i at time t. The UCB1 algorithm picks the option i_t that maximizes Q_i^t. Figure 1 depicts this logic: the confidence intervals represent uncertainty in the algorithm's estimate of the true value of m_i for each option, and the algorithm optimistically chooses the option with the highest upper confidence bound. This is an example of a general heuristic known in the bandit literature as *optimism in the face of uncertainty* [23]. The idea is that one should formulate the set of possible environments that are consistent with the observed data, then act as if the true environment were the most favorable one in that set.

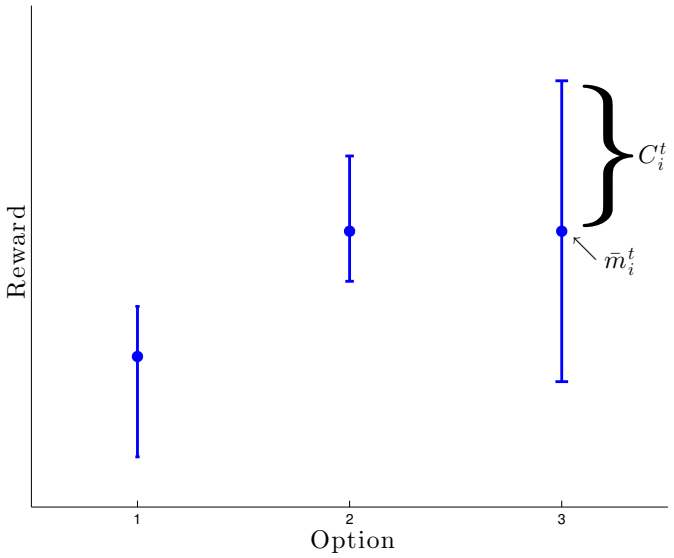


Fig. 1. Components of the UCB1 algorithm in an $N = 3$ option (arm) case. The algorithm forms a confidence interval for the mean reward m_i for each option i at each time t. The heuristic value $Q_i^t = \bar{m}_i^t + C_i^t$ is the upper limit of this confidence interval, representing an optimistic estimate of the true mean reward. In this example, options 2 and 3 have the same mean $\bar{m}$ but option 3 has a larger uncertainty C, so the algorithm chooses option 3.

Auer *et al.* [22] showed that for an appropriate choice of the uncertainty term C_i^t, the UCB1 algorithm achieves logarithmic regret uniformly in time, albeit with a larger leading constant than the optimal one (1). They also provided a slightly more complicated policy, termed UCB2, that brings the factor multiplying the logarithmic term arbitrarily close to that of (1). Their analysis relies on Chernoff-Hoeffding bounds which apply to probability distributions with bounded support.

They also considered the case of multi-armed bandits with Gaussian rewards, where both the mean (m_i in our notation) and sample variance (σ_s^2) are unknown. In this case they constructed an algorithm, termed UCB1-Normal, that achieves logarithmic regret. Their analysis of the regret in this case cannot appeal to Chernoff-Hoeffding bounds because the reward distribution has unbounded support. Instead their analysis relies on certain bounds on the tails of the χ^2 and the Student t-distribution that they could only verify numerically. Our work improves on their result in the case σ_s^2 is known by constructing a UCB-like algorithm that provably achieves logarithmic regret. The proof relies on new tight bounds on the tails of the Gaussian distribution that will be stated in Theorem 1.

D. The Bayes-UCB algorithm

UCB algorithms rely on a frequentist estimator $\bar{m}_i^t$ of m_i and therefore must sample each arm at least once in an initialization step, which requires a sufficiently long horizon, i.e., $N < T$. Bayesian estimators allow the integration of prior beliefs into the decision process. This enables a Bayesian UCB algorithm to treat the case $N > T$ as well as to capture the initial beliefs of an agent, informed perhaps through prior

experience. Kauffman *et al.* [30] considered the N-armed bandit problem from a Bayesian perspective and proposed the quantile function of the posterior reward distribution as the heuristic function (4).

For every random variable $X \in \mathbb{R} \cup \{\pm\infty\}$ with probability distribution function (pdf) $f(x)$, the associated cumulative distribution function (cdf) $F(x)$ gives the probability that the random variable takes a value of at most x, i.e., $F(x) = \mathbb{P}(X \leq x)$. See Figure 2. Conversely, the *quantile* function $F^{-1}(p)$ is defined by

$$F^{-1} : [0, 1] \to \mathbb{R} \cup \{\pm\infty\},$$

i.e., $F^{-1}(p)$ inverts the cdf to provide an upper bound for the value of the random variable $X \sim f(x)$:

$$\mathbb{P}\left(X \leq F^{-1}(p)\right) = p. \tag{5}$$

In this sense, $F^{-1}(p)$ is an *upper confidence bound*, i.e., an upper bound that holds with probability, or *confidence level*, p. Now suppose that $F(r)$ is the cdf for the reward distribution $p_i(r)$ of option i. Then, $Q_i = F^{-1}(p)$ gives a bound such that $\mathbb{P}(m_i > Q_i) = 1 - p$. If $p \in (0, 1)$ is chosen large, then $1 - p$ is small, and it is unlikely that the true mean reward for option i is higher than the bound. See Figure 3.

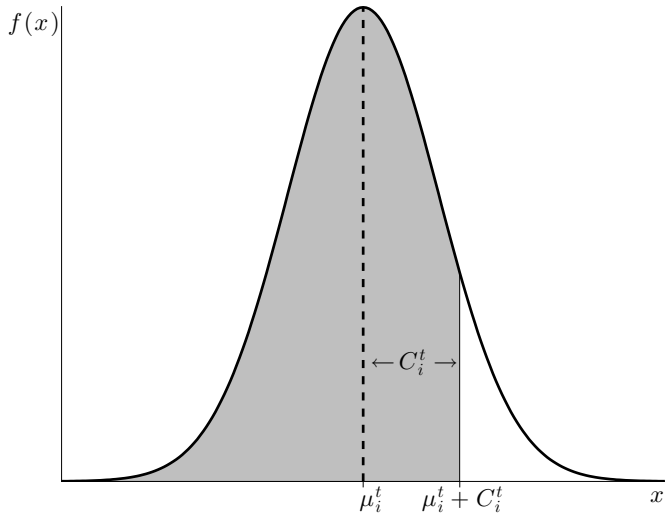


Fig. 2. The pdf $f(x)$ of a Gaussian random variable X with mean μ_i^t. The probability that $X \leq x$ is $\int_{-\infty}^{x} f(X)\,\mathrm{d}X = F(x)$. The area of the shaded region is $F(\mu_i^t + C_i^t) = p$, so the probability that $X \leq \mu_i^t + C_i^t$ is p. Conversely, $X \geq \mu_i^t + C_i^t$ with probability $1 - p$, so if p is close to 1, X is almost surely less than $\mu_i^t + C_i^t$.

In order to be increasingly sure of choosing the optimal arm as time goes on, [30] sets $p = 1 - \alpha_t$ as a function of time with $\alpha_t = 1/(t(\log T)^c)$, so that $1 - p$ is of order $1/t$. The authors termed the resulting algorithm Bayes-UCB. In the case that the rewards are Bernoulli distributed, they proved that with $c \geq 5$ Bayes-UCB achieves the bound (1) for uniform (uninformative) priors.

The choice of $1/t$ as the functional form for α_t can be motivated as follows. Roughly speaking, α_t is the probability of making an error (i.e., choosing a suboptimal arm) at time t. If a suboptimal arm is chosen with probability $1/t$, then the expected number of times it is chosen until time T will follow the integral of this rate, which is $\sum_1^T 1/t \approx \log T$, yielding a logarithmic functional form.

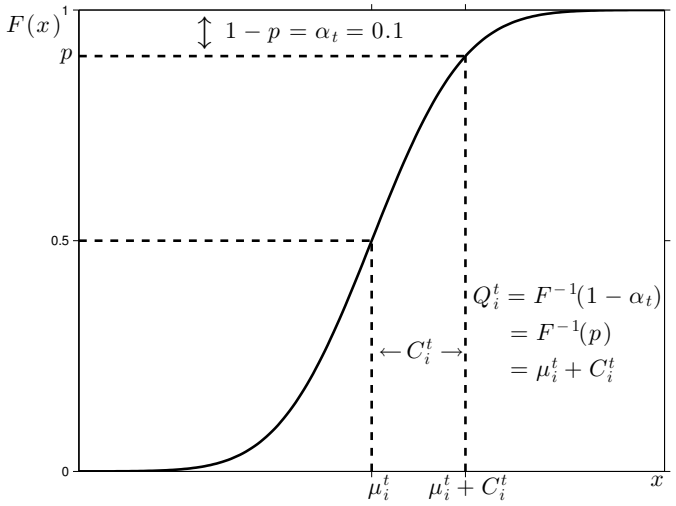


Fig. 3. Decomposition of the Gaussian cdf $F(x)$ and relation to the UCB/Bayes-UCB heuristic value. For a given value of α_t (here equal to 0.1), $F^{-1}(1 - \alpha_t)$ gives a value $Q_i^t = \mu_i^t + C_i^t$ such that the Gaussian random variable $X \leq Q_i^t$ with probability $1 - \alpha_t$. As $\alpha_t \to 0$, $Q_i^t \to +\infty$ and X is almost surely less than Q_i^t.

III. Features of human decision-making in multi-armed bandit tasks

As discussed in the introduction, human decision-making in the multi-armed bandit task has been the subject of numerous studies in the cognitive psychology literature. We list the five salient features of human decision-making in this literature that we wish to capture with our model.

(i) **Familiarity with the environment:** Familiarity with the environment and its structure plays a critical role in human decision-making [9], [38]. In the context of multi-armed bandit tasks, familiarity with the environment translates to prior knowledge about the mean rewards from each arm.

(ii) **Ambiguity bonus:** Wilson *et al.* [41] showed that the decision at time t is based on a linear combination of the estimate of the mean reward of each arm and an *ambiguity bonus* that captures the value of information from that arm. In the context of UCB and related algorithms, the ambiguity bonus can be interpreted similarly to the C_i^t term of (4) that defines the size of the upper bound on the estimated reward.

(iii) **Stochasticity:** Human decision-making is inherently noisy [9], [36], [38], [40], [41]. This is possibly due to inherent limitations in human computational capacity, or it could be the signature of noise being used as a cheap, general-purpose problem-solving algorithm. In the context of algorithms for solving the multi-armed bandit problem, this can be interpreted as picking arm i_t at time t using a stochastic arm selection strategy rather than a deterministic one.

(iv) **Finite-horizon effects:** Both the level of decision noise and the exploration-exploitation tradeoff are sensitive to the

time horizon T of the bandit task [9], [41]. This is a sensible feature to have, as shorter time horizons mean less time to take advantage of information gained by exploration, therefore biasing the optimal policy towards exploitation. The fact that both decision noise and the exploration-exploitation tradeoff (as represented by the ambiguity bonus) are affected by the time horizon suggests that they are both working as mechanisms for exploration, as investigated in [1]. In the context of algorithms, this means that the uncertainty term C_i^t and the stochastic arm selection scheme should be functions of the horizon T.

(v) **Environmental structure effects:** Acuña *et al.* [37] showed that an important aspect of human learning in multi-armed bandit tasks is structural learning, i.e., humans learn the correlation structure among different arms, and utilize it to improve their decision.

In the following, we develop a plausible model for human decision-making that captures these features. Feature (i) of human decision-making is captured through priors on the mean rewards from the arms. The introduction of priors in the decision-making process suggests that non-Bayesian upper confidence bound algorithms [22] cannot be used, and therefore, we focus on Bayesian upper confidence bound (upper credible limit) algorithms [30]. Feature (ii) of human decision-making is captured by making decisions based on a metric that comprises two components, namely, the estimate of the mean reward from each arm, and the width of a credible set. It is well known that the width of a credible set is a good measure of the uncertainty in the estimate of the reward. Feature (iii) of human decision-making is captured by introducing a stochastic arm selection strategy in place of the standard deterministic arm selection strategy [22], [30]. In the spirit of Kauffman *et al.* [30], we choose the credibility parameter α_t as a function of the horizon length to capture feature (iv) of human decision-making. Feature (v) is captured through the correlation structure of the prior used for the Bayesian estimation. For example, if the arms of the bandit are spatially embedded, it is natural to think of a covariance structure defined by $\Sigma_{ij} = \sigma_0^2 \exp(-|x_i - x_j|/\lambda)$, where x_i is the location of arm i and $\lambda \geq 0$ is the correlation length scale parameter that encodes the spatial smoothness of the rewards.

IV. The Upper Credible Limit (UCL) Algorithms for Gaussian Multi-armed Bandits

In this section, we construct a Bayesian UCB algorithm that captures the features of human decision-making described above. We begin with the case of deterministic decision-making and show that for an uninformative prior the resulting algorithm achieves logarithmic regret. We then extend the algorithm to the case of stochastic decision-making using a Boltzmann (or softmax) decision rule, and show that there exists a feedback rule for the temperature of the Boltzmann distribution such that the stochastic algorithm achieves logarithmic regret. In both cases we first consider uncorrelated priors and then extend to correlated priors.

A. The deterministic UCL algorithm with uncorrelated priors

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 selected 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 mean reward at arm i at time t is a Gaussian random variable (M_i) with mean and variance

$$\mu_i^t := \mathbb{E}[M_i | n_i^t, \bar{m}_i^t] = \frac{\delta^2 \mu_i^0 + n_i^t \bar{m}_i^t}{\delta^2 + n_i^t}, \text{ and}$$
$$\left(\sigma_i^t\right)^2 := \text{Var}[M_i | n_i^t, \bar{m}_i^t] = \frac{\sigma_s^2}{\delta^2 + n_i^t},$$

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

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

We now propose the UCL algorithm for the Gaussian multi-armed bandit problem. At each decision instance $t \in \{1, \ldots, T\}$, the UCL algorithm selects an arm with the maximum value of the upper limit of the smallest $(1 - 1/Kt)$-credible interval, i.e., it selects an arm $i_t = \text{argmax}\{Q_i^t \mid i \in \{1, \ldots, N\}\}$, where

$$Q_i^t = \mu_i^t + \sigma_i^t \Phi^{-1}(1 - 1/Kt),$$

$\Phi^{-1} : (0, 1) \to \mathbb{R}$ is the inverse cumulative distribution function for the standard Gaussian random variable, and $K \in \mathbb{R}_{>0}$ is a tunable parameter. For an explicit pseudocode implementation, see Algorithm 1 in Appendix F. In the following, we will refer to Q_i^t as the $(1-1/Kt)$-*upper credible limit* (UCL).

It is known [43], [28] that an efficient policy to maximize the total information gained over sequential sampling of options is to pick the option with highest variance at each time. Thus, Q_i^t is the weighted sum of the expected gain in the total reward (exploitation), and the gain in the total information about arms (exploration), if arm i is picked at time t.

B. Regret analysis of the deterministic UCL Algorithm

In this section, we analyze the performance of the UCL algorithm. We first derive bounds on the inverse cumulative distribution function for the standard Gaussian random variable and then utilize it to derive upper bounds on the cumulative expected regret for the UCL algorithm. We state the following theorem about bounds on the inverse Gaussian cdf.

Theorem 1 (***Bounds on the inverse Gaussian cdf***)**.** *The following bounds hold for the inverse cumulative distribution function of the standard Gaussian random variable for each $\alpha \in (0, 1/\sqrt{2\pi})$, and any $\beta \geq 1.02$:*

$$\Phi^{-1}(1 - \alpha) < \beta \sqrt{-\log(-(2\pi\alpha^2)\log(2\pi\alpha^2))}, \text{ and} \quad (6)$$
$$\Phi^{-1}(1 - \alpha) > \sqrt{-\log(2\pi\alpha^2(1 - \log(2\pi\alpha^2)))}. \quad (7)$$

Proof: See Appendix A. ∎

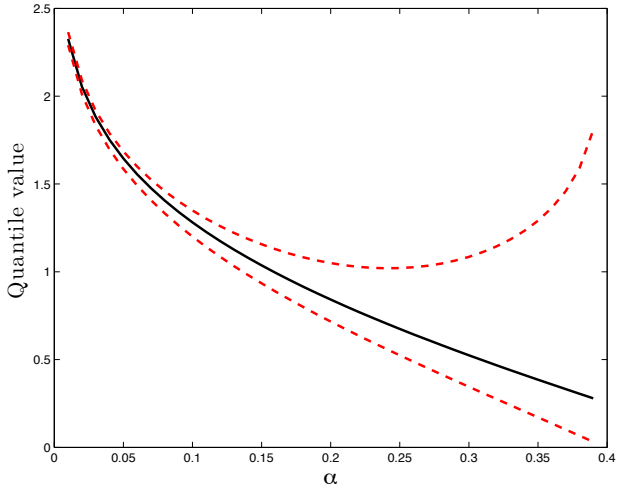


Fig. 4. Depiction of the normal quantile function $\Phi^{-1}(1-\alpha)$ (solid line) and the bounds (6) and (7) (dashed lines), with $\beta = 1.02$.

The bounds in equations (6) and (7) were conjectured by Fan [44] without the factor β. In fact, it can be numerically verified that without the factor β, the conjectured upper bound is incorrect. We present a visual depiction of the tightness of the derived bounds in Figure 4.

We now analyze the performance of the UCL algorithm. We define $\{R_t^{\text{UCL}}\}_{t\in\{1,\ldots,T\}}$ as the sequence of expected regret for the UCL algorithm. The UCL algorithm achieves logarithmic regret uniformly in time as formalized in the following theorem.

Theorem 2 (***Regret of the deterministic UCL algorithm***)**.** *The following statements hold for the Gaussian multi-armed bandit problem and the deterministic UCL algorithm with uncorrelated uninformative prior and $K = \sqrt{2\pi e}$:*

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

$$\mathbb{E}\left[n_i^T\right] \leq \left(\frac{8\beta^2\sigma_s^2}{\Delta_i^2} + \frac{2}{\sqrt{2\pi e}}\right)\log T + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1-\log 2-\log\log T) + 1 + \frac{2}{\sqrt{2\pi e}};$$

(ii) *the cumulative expected regret until time T satisfies*

$$\sum_{t=1}^{T} R_t^{\text{UCL}} \leq \sum_{i=1}^{N}\Delta_i\left(\left(\frac{8\beta^2\sigma_s^2}{\Delta_i^2} + \frac{2}{\sqrt{2\pi e}}\right)\log T + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1-\log 2-\log\log T) + 1 + \frac{2}{\sqrt{2\pi e}}\right).$$

Proof: See Appendix B. ■

Remark 3 (***Uninformative priors with short time horizon***)**.** When the deterministic UCL algorithm is used with an uncorrelated uninformative prior, Theorem 2 guarantees that the algorithm incurs logarithmic regret uniformly in horizon length T. However, for small horizon lengths, the upper bound on the regret can be lower bounded by a super-logarithmic curve. Accordingly, in practice, the cumulative expected regret curve may appear super-logarithmic for short time horizons. For example, for horizon T less than the number of arms N, the cumulative expected regret of the deterministic UCL algorithm grows at most linearly with the horizon length. □

Remark 4 (***Comparison with UCB1***)**.** In view of the bounds in Theorem 1, for an uninformative prior, the $(1-1/Kt)$-upper credible limit obeys

$$Q_i^t < \bar{m}_i^t + \beta\sigma_s\sqrt{\frac{1+2\log t-\log\log et^2}{n_i^t}}.$$

This upper bound is similar to the one in UCB1, which sets

$$Q_i^t = \bar{m}_i^t + \sqrt{\frac{2\log t}{n_i^t}}.$$ □

Remark 5 (***Informative priors***)**.** For an uninformative prior, i.e., very large variance σ_0^2, we established in Theorem 2 that the deterministic UCL algorithm achieves logarithmic regret uniformly in time. For informative priors, the cumulative expected regret depends on the quality of the prior. The quality of a prior on the rewards can be captured by the metric $\zeta := \max\{|m_i - \mu_i^0|/\sigma_0 \mid i \in \{1,\ldots,N\}\}$. A *good prior* corresponds to small values of ζ, while a *bad prior* corresponds to large values of ζ. In other words, a good prior is one that has (i) mean close to the true mean reward, or (ii) a large variance. Intuitively, a good prior either has a fairly accurate estimate of the mean reward, or has low confidence about its estimate of the mean reward. For a good prior, the parameter K can be tuned such that

$$\Phi^{-1}\left(1-\frac{1}{Kt}\right) - \max_{i\in\{1,\ldots,N\}}\frac{\sigma_s(|m_i-\mu_i^0|)}{\sigma_0^2} > \Phi^{-1}\left(1-\frac{1}{\bar{K}t}\right),$$

where $\bar{K} \in \mathbb{R}_{>0}$ is some constant, and it can be shown, using the arguments of Theorem 2, that the deterministic UCL algorithm achieves logarithmic regret uniformly in time. A bad prior corresponds to a fairly inaccurate estimate of the mean reward and high confidence. For a bad prior, the cumulative expected regret may be a super-logarithmic function of the horizon length. □

Remark 6 (***Sub-logarithmic regret for good priors***)**.** For a good prior with a small variance, even uniform sub-logarithmic regret can be achieved. Specifically, if the variable Q_i^t in Algorithm 1 is set to $Q_i^t = m_i^t + \sigma_i^t\Phi^{-1}(1-1/Kt^2)$, then an analysis similar to Theorem 2 yields an upper bound on the cumulative expected regret that is dominated by (i) a sub-logarithmic term for good priors with small variance, and (ii) a logarithmic term for uninformative priors with a higher constant in front than the constant in Theorem 2. Notice that such good priors may correspond to human operators who have previous training in the task. □

C. The stochastic UCL algorithm with uncorrelated priors

To capture the inherent stochastic nature of human decision-making, we consider the UCL algorithm with stochastic arm selection. Stochasticity has been used as a generic optimization mechanism that does not require information about the objective function. For example, simulated annealing [45], [46],

[47] is a global optimization method that attempts to break out of local optima by sampling locations near the currently selected optimum and accepting locations with worse objective values with a probability that decreases in time. By analogy with physical annealing processes, the probabilities are chosen from a Boltzmann distribution with a dynamic temperature parameter that decreases in time, gradually making the optimization more deterministic. An important problem in the design of simulated annealing algorithms is the choice of the temperature parameter, also known as a *cooling schedule*.

Choosing a good cooling schedule is equivalent to solving the explore-exploit problem in the context of simulated annealing, since the temperature parameter balances exploration and exploitation by tuning the amount of stochasticity (exploration) in the algorithm. In their classic work, Mitra *et al.* [46] found cooling schedules that maximize the rate of convergence of simulated annealing to the global optimum. In a similar way, the stochastic UCL algorithm (see Algorithm 2 in Appendix F for an explicit pseudocode implementation) extends the deterministic UCL algorithm (Algorithm 1) to the stochastic case. The stochastic UCL algorithm chooses an arm at time t using a Boltzmann distribution with temperature υ_t, so the probability P_{it} of picking arm i at time t is given by

$$P_{it} = \frac{\exp(Q_i^t/\upsilon_t)}{\sum_{j=1}^N \exp(Q_j^t/\upsilon_t)}.$$

In the case $\upsilon_t \to 0^+$ this scheme chooses $i_t = \operatorname{argmax}\{Q_i^t \mid i \in \{1,\dots,N\}\}$ and as υ_t increases the probability of selecting any other arm increases. Thus Boltzmann selection generalizes the maximum operation and is sometimes known as the soft maximum (or softmax) rule.

The temperature parameter might be chosen constant, i.e., $\upsilon_t = \upsilon$. In this case the performance of the stochastic UCL algorithm can be made arbitrarily close to that of the deterministic UCL algorithm by taking the limit $\upsilon \to 0^+$. However, [46] showed that good cooling schedules for simulated annealing take the form

$$\upsilon_t = \frac{\nu}{\log t},$$

so we investigate cooling schedules of this form. We choose ν using a feedback rule on the values of the heuristic function $Q_i^t, i \in \{1,\dots,N\}$ and define the cooling schedule as

$$\upsilon_t = \frac{\Delta Q_{\min}^t}{2\log t},$$

where $\Delta Q_{\min}^t = \min\{|Q_i^t - Q_j^t| \mid i,j \in \{1,\dots,N\}, i \neq j\}$ is the minimum gap between the heuristic function value for any two pairs of arms. We define $\infty - \infty = 0$, so that $\Delta Q_{\min}^t = 0$ if two arms have infinite heuristic values, and define $0/0 = 1$.

D. Regret analysis of the stochastic UCL algorithm

In this section we show that for an uninformative prior, the stochastic UCL algorithm achieves efficient performance. We define $\{R_t^{\text{SUCL}}\}_{t\in\{1,\dots,T\}}$ as the sequence of expected regret for the stochastic UCL algorithm. The stochastic UCL algorithm achieves logarithmic regret uniformly in time as formalized in the following theorem.

Theorem 7 (***Regret of the stochastic UCL algorithm***)**.** *The following statements hold for the Gaussian multi-armed bandit problem and the stochastic UCL algorithm with uncorrelated uninformative prior and $K = \sqrt{2\pi e}$:*

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

$$\mathbb{E}\left[n_i^T\right] \le \Big(\frac{8\beta^2\sigma_s^2}{\Delta_i^2} + \frac{2}{\sqrt{2\pi e}}\Big)\log T + \frac{\pi^2}{6} + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2 - \log\log T) + 1 + \frac{2}{\sqrt{2\pi e}};$$

(ii) *the cumulative expected regret until time T satisfies*

$$\sum_{t=1}^T R_t^{\text{SUCL}} \le \sum_{i=1}^N \Delta_i \left(\Big(\frac{8\beta^2\sigma_s^2}{\Delta_i^2} + \frac{2}{\sqrt{2\pi e}}\Big)\log T + \frac{\pi^2}{6} + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2 - \log\log T) + 1 + \frac{2}{\sqrt{2\pi e}} \right).$$

Proof: See Appendix C. ∎

E. The UCL algorithms with correlated priors

In the preceding sections, we consider the case of uncorrelated priors, i.e., the case with diagonal covariance matrix of the prior distribution for mean rewards $\Sigma_0 = \sigma_0^2 I_N$. However, in many cases there may be dependence among the arms that we wish to encode in the form of a non-diagonal covariance matrix. In fact, one of the main advantages a human may have in performing a bandit task is their prior experience with the dependency structure across the arms resulting in a good prior correlation structure. We show that including covariance information can improve performance and may, in some cases, lead to sub-logarithmic regret.

Let $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ and $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_{0d})$ be correlated and uncorrelated priors on the mean rewards from the arms, respectively, where $\boldsymbol{\mu}_0 \in \mathbb{R}^N$ is the vector of prior estimates of the mean rewards from each arm, $\Sigma_0 \in \mathbb{R}^{N\times N}$ is a positive definite matrix, and Σ_{0d} is the same matrix with all its non-diagonal elements set equal to 0. The inference procedure described in Section IV-A generalizes to a correlated prior as follows: Define $\{\boldsymbol{\phi}_t \in \mathbb{R}^N\}_{t\in\{1,\dots,T\}}$ to be the indicator vector corresponding to the currently chosen arm i_t, where $(\boldsymbol{\phi}_t)_k = 1$ if $k = i_t$, and zero otherwise. Then the belief state $(\boldsymbol{\mu}_t, \Sigma_t)$ updates as follows [48]:

$$\begin{aligned} \mathbf{q} &= \frac{r_t\boldsymbol{\phi}_t}{\sigma_s^2} + \Lambda_{t-1}\boldsymbol{\mu}_{t-1} \\ \Lambda_t &= \frac{\boldsymbol{\phi}_t\boldsymbol{\phi}_t^T}{\sigma_s^2} + \Lambda_{t-1},\ \Sigma_t = \Lambda_t^{-1} \\ \boldsymbol{\mu}_t &= \Sigma_t\mathbf{q}, \end{aligned} \tag{8}$$

where $\Lambda_t = \Sigma_t^{-1}$ is the *precision* matrix.

The upper credible limit for each arm i can be computed based on the univariate Gaussian marginal distribution of the posterior with mean μ_i^t and variance $(\sigma_i^t)^2 = (\Sigma_t)_{ii}$. Consider the evolution of the belief state with the diagonal (uncorrelated) prior Σ_{0d} and compare it with the belief state based on the non-diagonal Σ_0 which encodes information

about the correlation structure of the rewards in the off-diagonal terms. The additional information means that the inference procedure will converge more quickly than in the uncorrelated case, as seen in Theorem 8. If the assumed correlation structure correctly models the environment, then the inference will converge towards the correct values, and the performance of the UCL and stochastic UCL algorithms will be at least as good as that guaranteed by the preceding analyses in Theorems 2 and 7.

Denoting ${\sigma_i^t}^2 = (\Sigma_t)_{ii}$ as the posterior at time t based on Σ_0 and ${\sigma_{id}^t}^2 = (\Sigma_{td})_{ii}$ as the posterior based on Σ_{0d}, for a given sequence of chosen arms $\{i_\tau\}_{\tau\in\{1,\dots,T\}}$, we have that the variance of the non-diagonal estimator will be no larger than that of the diagonal one, as summarized in the following theorem:

Theorem 8 (***Correlated versus uncorrelated priors***)**.** *For the inference procedure in* (8)*, and any given sequence of selected arms* $\{i_\tau\}_{\tau\in\{1,\dots,T\}}$*,* ${\sigma_i^t}^2 \leq {\sigma_{id}^t}^2$*, for any* $t \in \{0,\dots,T\}$*, and for each* $i \in \{1,\dots,N\}$*.*

Proof: We use induction. By construction, ${\sigma_i^0}^2 = {\sigma_{id}^0}^2$, so the statement is true for $t = 0$. Suppose the statement holds for some $t \geq 0$ and consider the update rule for Σ_t. From the Sherman-Morrison formula for a rank-1 update [49], we have

$$(\Sigma_{t+1})_{jk} = (\Sigma_t)_{jk} - \left(\frac{\Sigma_t\boldsymbol{\phi}_t\boldsymbol{\phi}_t'\Sigma_t}{\sigma_s^2 + \boldsymbol{\phi}_t'\Sigma_t\boldsymbol{\phi}_t}\right)_{jk}.$$

We now examine the update term in detail, starting with its denominator:

$$\boldsymbol{\phi}_t'\Sigma_t\boldsymbol{\phi}_t = (\Sigma_t)_{i_ti_t},$$

so $\sigma_s^2 + \boldsymbol{\phi}_t'\Sigma_t\boldsymbol{\phi}_t = \sigma_s^2 + (\Sigma_t)_{i_ti_t} > 0$. The numerator is the outer product of the i_t-th column of Σ_t with itself, and can be expressed in index form as

$$(\Sigma_t\boldsymbol{\phi}_t\boldsymbol{\phi}_t'\Sigma_t)_{jk} = (\Sigma_t)_{ji_t}(\Sigma_t)_{i_tk}.$$

Note that if Σ_t is diagonal, then so is Σ_{t+1} since the only non-zero update element will be $(\Sigma_t)_{i_ti_t}^2$. Therefore, Σ_{td} is diagonal for all $t \geq 0$.

The update of the diagonal terms of Σ only uses the diagonal elements of the update term, so

$${\sigma_i^{(t+1)}}^2 = (\Sigma_{t+1})_{ii} = (\Sigma_t)_{ii} - \frac{1}{\sigma_s^2 + \boldsymbol{\phi}_t'\Sigma_t\boldsymbol{\phi}_t}\sum_j (\Sigma_t)_{ji_t}(\Sigma_t)_{i_tj}.$$

In the case of Σ_{td}, the sum over j only includes the $j = i_t$ element whereas with the non-diagonal prior Σ_t the sum may include many additional terms. So we have

$$\begin{aligned}{\sigma_i^{(t+1)}}^2 = (\Sigma_{t+1})_{ii} &= (\Sigma_t)_{ii} - \frac{1}{\sigma_s^2 + \boldsymbol{\phi}_t'\Sigma_t\boldsymbol{\phi}_t}\sum_j (\Sigma_t)_{ji_t}(\Sigma_t)_{i_tj} \\ &\leq (\Sigma_{td})_{ii} - \frac{1}{\sigma_s^2 + \boldsymbol{\phi}_t'\Sigma_{td}\boldsymbol{\phi}_t}(\Sigma_{td})_{i_ti_t}^2 \\ &= {\sigma_{id}^{(t+1)}}^2,\end{aligned}$$

and the statement holds for $t + 1$. ∎

Note that the above result merely shows that the belief state converges more quickly in the case of a correlated prior, without making any claim about the correctness of this convergence. For example, consider a case where the prior belief is that two arms are perfectly correlated, i.e., the relevant block of the prior is a multiple of $\left(\begin{smallmatrix}1 & 1\\ 1 & 1\end{smallmatrix}\right)$, but in actuality the two arms have very different mean rewards. If the algorithm first samples the arm with lower reward, it will tend to underestimate the reward to the second arm. However, in the case of a well-chosen prior the faster convergence will allow the algorithm to more quickly disregard related sets of arms with low rewards.

V. Classification of human performance in multi-armed bandit tasks

In this section, we study human data from a multi-armed bandit task and show how human performance can be classified as falling into one of several categories, which we term *phenotypes*. We then show that the stochastic UCL algorithm can produce performance that is analogous to the observed human performance.

A. Human behavioral experiment in a multi-armed bandit task

In order to study human performance in multi-armed bandit tasks, we ran a spatially-embedded multi-armed bandit task through web servers at Princeton University. Human participants were recruited using Amazon's Mechanical Turk (AMT) web-based task platform [50]. Upon selecting the task on the AMT website, participants were directed to follow a link to a Princeton University website, where informed consent was obtained according to protocols approved by the Princeton University Institutional Review Board.

After informed consent was obtained, participants were shown instructions that told them they would be playing a simple game during which they could collect points, and that their goal was to collect the maximum number of total points in each part of the game.

Each participant was presented with a set of $N = 100$ options in a 10×10 grid. At each decision time $t \in \{1,\dots,T\}$, the participant made a choice by moving the cursor to one element of the grid and clicking. After each choice was made a numerical reward associated to that choice was reported on the screen. The time allowed for each choice was manipulated and allowed to take one of two values, denoted fast and slow. If the participant did not make a choice within 1.5 (fast) or 6 (slow) seconds after the prompt, then the last choice was automatically selected again. The reward was visible until the next decision was made and the new reward reported. The time allotted for the next decision began immediately upon the reporting of the new reward. Figure 5 shows the screen used in the experiment.

The dynamics of the game were also experimentally manipulated, although we focus exclusively here on the first dynamic condition. The first dynamic condition was a standard bandit task, where the participant could choose any option at each decision time, and the game would immediately sample that option. In the second and third dynamic conditions, the participant was restricted in choices and the game responded

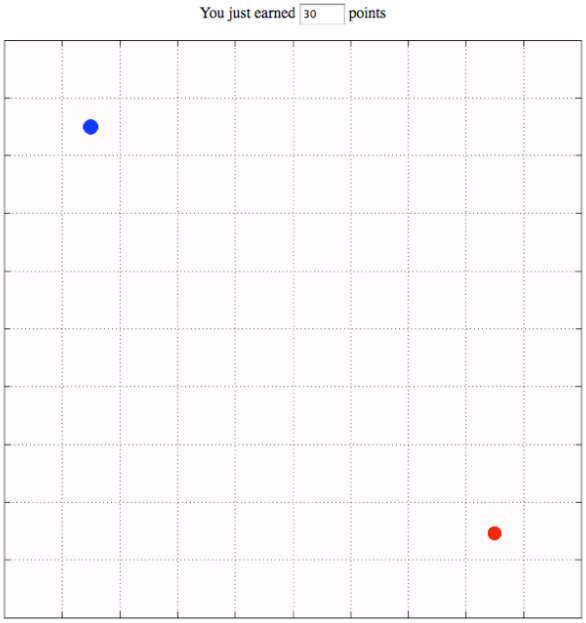


Fig. 5. The screen used in the experimental interface. Each square in the grid corresponded to an available option. The text box above the grid displayed the most recently received reward, the blue dot indicated the participant's most recently recorded choice, and the smaller red dot indicated the participant's next choice. In the experiment, the red dot was colored yellow, but here we have changed the color for legibility. When both dots were located in the same square, the red dot was superimposed over the blue dot such that both were visible. Initially, the text box was blank and the two dots were together in a randomly chosen square. Participants indicated a choice by clicking in a square, at which point the red dot would move to the chosen option. Until the time allotted for a given decision had elapsed, participants could change their decision without penalty by clicking on another square, and the red dot would move accordingly. When the decision time had elapsed, the blue dot would move to the new square, the text box above the grid would be updated with the most recent reward amount, and the choice would be recorded.

in different ways. These two conditions are beyond the scope of this paper.

Participants were first trained with three training blocks of $T = 10$ choices each, one for each form of the game dynamics. Subsequently, the participants performed two task blocks of $T = 90$ choices each in a balanced experimental design. For each participant, the first task had parameters randomly chosen from one of the 12 possible combinations (2 timing, 3 dynamics, 2 landscapes), and the second task was conditioned on the first so that the alternative timing was used with the alternative landscape and the dynamics chosen randomly from the two remaining alternatives. In particular, only approximately 2/3 of the participants were assigned a standard bandit task, while other subjects were assigned other dynamic conditions. The horizon $T < N$ was chosen so that prior beliefs would be important to performing the task. Each training block took 15 seconds and each task block took 135 (fast) or 540 (slow) seconds. The time between blocks was negligible, due only to network latency.

Mean rewards in the task blocks corresponded to one of two landscapes: Landscape A (Figure 6(a)) and Landscape B (Figure 6(b)). Each landscape was flat along one dimension and followed a profile along the other dimension. In the two task blocks, each participant saw each landscape once, presented in random order. Both landscapes had a mean value of 30 points and a maximum of approximately 60 points, and the rewards r_t for choosing an option i_t were computed as the sum of the mean reward m_{i_t} and an integer chosen uniformly from the range $[-5, 5]$. In the training blocks, the landscape had a mean value of zero everywhere except for a single peak of 100 points in the center. The participants were given no specific information about the value or the structure of the reward landscapes.

To incentivize the participants to make choices to maximize their cumulative reward, the participants were told that they were being paid based on the total reward they collected during the tasks. As noted above, due to the multiple manipulations, not every participant performed a standard bandit task block. Data were collected from a total of 417 participants: 326 of these participants performed one standard bandit task block each, and the remaining 91 participants performed no standard bandit task blocks.

B. Phenotypes of observed performance

For each 90 choice standard bandit task block, we computed observed regret by subtracting the maximum mean cumulative reward from the participant's cumulative reward, i.e.,

$$\mathcal{R}(t) = m_{i^*} t - \sum_{\tau=1}^{t} r_\tau.$$

The definition of $\mathcal{R}(t)$ uses received rather than expected reward, so it is not identical to cumulative expected regret. However, due to the large number of individual rewards received and the small variance in rewards, the difference between the two quantities is small.

We study human performance by considering the functional form of $\mathcal{R}(t)$. Optimal performance in terms of regret corresponds to $\mathcal{R}(t) = \mathcal{C} \log t$, where $\mathcal{C}$ is the sum over i of the factors in (1). The worst-case performance, corresponding to repeatedly choosing the lowest-value option, corresponds to the form $\mathcal{R}(t) = \mathcal{K}t$, where $\mathcal{K} > 0$ is a constant. Other bounds in the bandit literature (e.g. [28]) are known to have the form $\mathcal{R}(t) = \mathcal{K}\sqrt{t}$.

To classify types of observed human performance in bandit tasks, we fit models representing these three forms to the observed regret from each task. Specifically, we fit the three models

$$\mathcal{R}(t) = a + bt \tag{9}$$

$$\mathcal{R}(t) = at^b \tag{10}$$

$$\mathcal{R}(t) = a + b\log(t) \tag{11}$$

to the data from each task and classified the behavior according to which of the models (9)–(11) best fit the data in terms of squared residuals. Model selection using this procedure is tenable given that the complexity or number of degrees of freedom of the three models is the same.

Of the 326 participants who performed a standard bandit task block, 59.2% were classified as exhibiting linear regret (model (9)), 19.3% power regret (10), and 21.5% logarithmic regret (11). This suggests that 40.8% of the participants

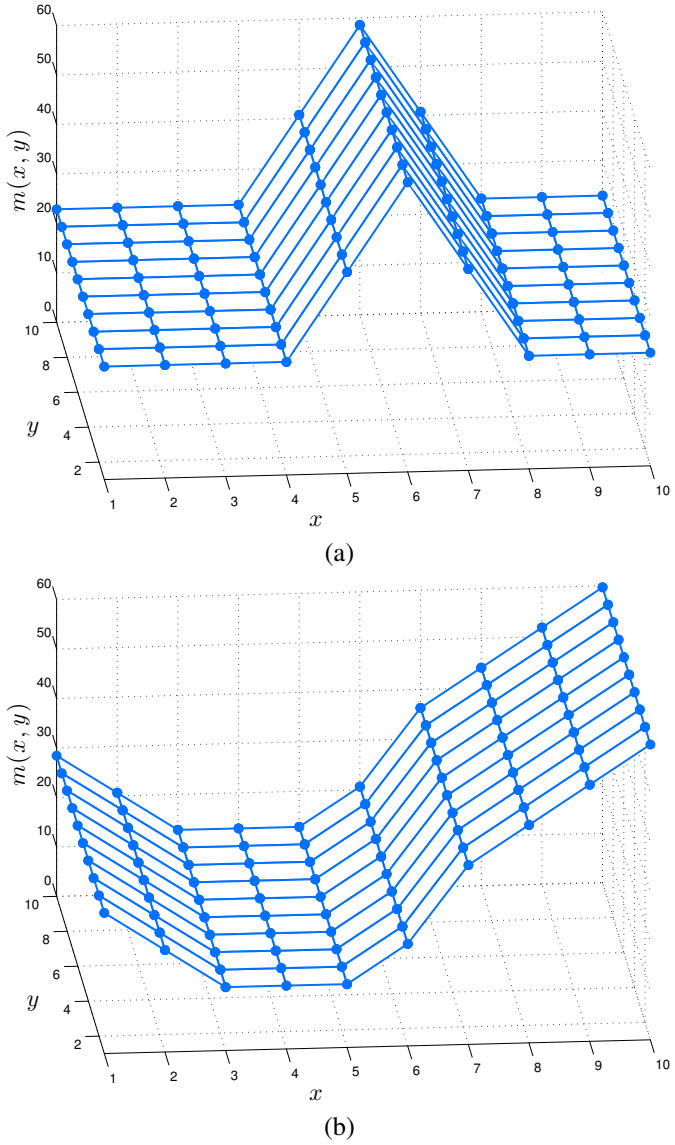


Fig. 6. The two task reward landscapes: (a) Landscape A, (b) Landscape B. The two-dimensional reward surfaces followed the profile along one dimension (here the x direction) and were flat along the other (here the y direction). The Landscape A profile is designed to be simple in the sense that the surface is concave and there is only one global maximum ($x = 6$), while the Landscape B profile is more complicated since it features two local maxima ($x = 1$ and 10), only one of which ($x = 10$) is the global maximum.

performed well overall and 21.5% performed very well. We observed no significant correlation between performance and timing, landscape, or order (first or second) of playing the standard bandit task block.

Averaging across all tasks, mean performance was best fit by a power model with exponent $b \approx 0.9$, so participants on average achieved sub-linear regret, i.e., better than linear regret. The nontrivial number of positive performances are noteworthy given that $T < N$, i.e., a relatively short time horizon which makes the task challenging.

Averaging, conditional on the best-fit model, separates the performance of the participants into the three categories of regret performance as can be observed in Figure 7. The difference between linear and power-law performance is not statistically significant until near the task horizon at $t = 90$, but log-law performance is statistically different from the other two, as seen using the confidence intervals in the figure. We therefore interpret the linear and power-law performance phenotypes as representing participants with low performance and the log-law phenotype as representing participants with high performance. Interestingly, the three models are indistinguishable for time less than sufficiently small $t \lesssim 30$. This may represent a fundamental limit to performance that depends on the complexity of the reward surface: if the surface is smooth, skilled participants can quickly find good options, corresponding to a small value of the constant $\mathcal{K}$, and thus their performance will quickly be distinguished from less skilled participants. However, if the surface is rough, identifying good options is harder and will therefore require more samples, i.e., a large value of $\mathcal{K}$, even for skilled participants.

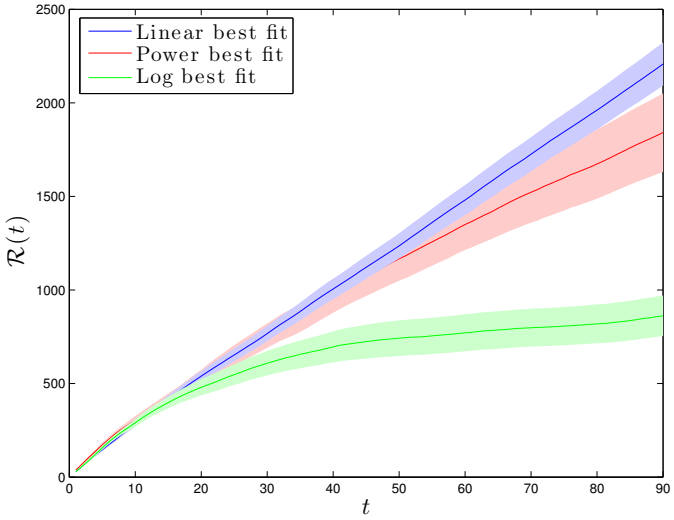


Fig. 7. Mean observed regret $\mathcal{R}(t)$ conditional on the best-fit model (9)–(11), along with bands representing 95% confidence intervals. Note how the difference between linear and power-law regret is not statistically significant until near the task horizon $T = 90$, while logarithmic regret is significantly less than that of the linear and power-law cases.

C. Comparison with UCL

Having identified the three phenotypes of observed human performance in the above section, we show that the stochastic UCL algorithm (Algorithm 2) can produce behavior corresponding to the linear-law and log-law phenotypes by varying a minimal number of parameters. Parameters are used to encode the prior beliefs and the decision noise of the participant. A minimal set of parameters is given by the four scalars μ_0, σ_0, λ and υ, defined as follows.

(i) **Prior mean** The model assumes prior beliefs about the mean rewards to be a Gaussian distribution with mean $\boldsymbol{\mu}_0$ and covariance Σ_0. It is reasonable to assume that participants set $\boldsymbol{\mu}_0$ to the uniform prior $\boldsymbol{\mu}_0 = \mu_0 \mathbf{1}_N$, where $\mathbf{1}_N \in \mathbb{R}^N$ is the vector with every entry equal to 1. Thus, $\mu_0 \in \mathbb{R}$ is a single parameter that encodes the participants' beliefs about the mean value of rewards.

(ii,iii) **Prior covariance** For a spatially-embedded task, it is reasonable to assume that arms that are spatially close will have similar mean rewards. Following [51] we choose the elements of Σ_0 to have the form

$$\Sigma_{ij} = \sigma_0^2 \exp(-|x_i - x_j|/\lambda), \tag{12}$$

where x_i is the location of arm i and $\lambda \geq 0$ is the correlation length scale parameter that encodes the spatial smoothness of the reward surface. The case $\lambda = 0$ represents complete independence of rewards, i.e., a very rough surface, while as λ increases the agent believes the surface to be smoother. The parameter $\sigma_0 \geq 0$ can be interpreted as a confidence parameter, with $\sigma_0 = 0$ representing absolute confidence in the beliefs about the mean $\boldsymbol{\mu}_0$, and $\sigma_0 = +\infty$ representing complete lack of confidence.

(iv) **Decision noise** In Theorem 7 we show that for an appropriately chosen cooling schedule, the stochastic UCL algorithm with softmax action selection achieves logarithmic regret. However, the assumption that human participants employ this particular cooling schedule is unreasonably strong. It is of great interest in future experimental work to investigate what kind of cooling schedule best models human behavior. The Bayes-optimal cooling schedule can be computed using variational Bayes methods [52]; however, for simplicity, we model the participants' decision noise by using softmax action selection with a constant temperature $\upsilon \geq 0$. This yields a single parameter representing the stochasticity of the decision-making: in the limit $\upsilon \to 0^+$, the model reduces to the deterministic UCL algorithm, while with increasing υ the decision-making is increasingly stochastic.

With this set of parameters, the prior quality ζ from Remark 5 reduces to $\zeta = (\max_i |m_i - \mu_0|)/\sigma_0$. Uninformative priors correspond to very large values of σ_0. Good priors, corresponding to small values of ζ, have μ_0 close to $m_{i^*} = \max_i m_i$ or little confidence in the value of μ_0, represented by large values of σ_0.

By adjusting these parameters, we can replicate both linear and logarithmic observed regret behaviors as seen in the human data. Figure 8 shows examples of simulated observed regret $\mathcal{R}(t)$ that capture linear and logarithmic regret, respectively. In both examples, Landscape B was used for the mean rewards. The example with linear regret shows a case where the agent has fairly uninformative and fully uncorrelated prior beliefs (i.e., $\lambda = 0$). The prior mean $\mu_0 = 30$ is set equal to the true surface mean, but with $\sigma_0^2 = 1000$, so that the agent is not very certain of this value. Moderate decision noise is incorporated by setting $\upsilon = 4$. The values of the prior encourage the agent to explore most of the $N = 100$ options in the $T = 90$ choices, yielding regret that is linear in time. As emphasized in Remark 3, the deterministic UCL algorithm (and any agent employing the algorithm) with an uninformative prior cannot in general achieve sub-linear cumulative expected regret in a task with such a short horizon. The addition of decision noise to this algorithm will tend to increase regret, making it harder for the agent to achieve sub-linear regret.

In contrast, the example with logarithmic regret shows how an informative prior with an appropriate correlation structure can significantly improve the agent's performance. The prior mean $\mu_0 = 200$ encourages more exploration than the previous value of 30, but the smaller value of $\sigma_0^2 = 10$ means the agent is more confident in its belief and will explore less. The correlation structure induced by setting the length scale $\lambda = 4$ is a good model for the reward surface, allowing the agent to more quickly reject areas of low rewards. A lower softmax temperature $\upsilon = 1$ means that the agent's decisions are made more deterministically. Together, these differences lead to the agent's logarithmic regret curve; this agent suffers less than a third of the total regret during the task as compared to the agent with the poorer prior and linear regret.

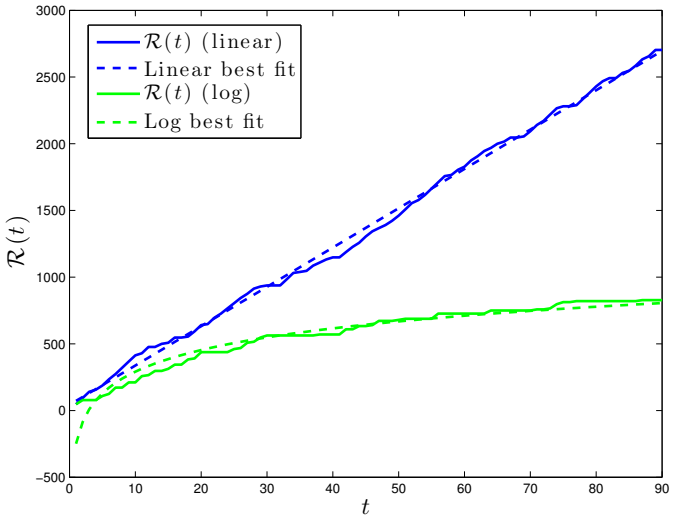


Fig. 8. Observed regret $\mathcal{R}(t)$ from simulations (solid lines) that demonstrate linear (9), blue curves, and log (11), green curves, regret. The best fits to the simulations are shown (dashed lines). The simulated task parameters were identical to those of the human participant task with Landscape B from Figure 6(b). In the example with linear regret, the agent's prior on rewards was the uncorrelated prior $\mu_0 = 30$, $\sigma_0^2 = 1000, \lambda = 0$. Decision noise was incorporated using softmax selection with a constant temperature $\upsilon = 4$. In the example with log regret, the agent's prior on rewards was the correlated prior with uniform $\mu_0 = 200$ and Σ_0 an exponential prior (12) with parameters $\sigma_0^2 = 10, \lambda = 4$. The decision noise parameter was set to $\upsilon = 1$.

VI. Gaussian multi-armed bandit problems with transition costs

Consider an N-armed bandit problem as described in Section II. Suppose that 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 decision-maker chooses the same arm as the previous time instant, and accordingly, $c_{ii} = 0$. Such a cost structure corresponds to a search problem in which the N arms may correspond to N spatially distributed regions and the transition cost c_{ij} may correspond to the travel cost from region i to region j.

To address this variation of the multi-armed bandit problem, we extend the UCL algorithm to a strategy that makes use of block allocations. Block allocations refer to sequences in which the same choice is made repeatedly; thus, during a block no transition cost is incurred. The UCL algorithm is used to

make the choice of arm at the beginning of each block. The design of the (increasing) length of the blocks makes the block algorithm provably efficient. This model can be used in future experimental work to investigate human behavior in multi-armed bandit tasks with transition costs.

A. The Block UCL Algorithm

For Gaussian multi-armed bandits with transition costs, we develop a block allocation strategy described graphically in Figure 9 and in pseudocode in Algorithm 3 in Appendix F. The intuition behind the strategy is as follows. The decision-maker's objective is to maximize the total expected reward while minimizing the number of transitions. As we have shown, maximizing total expected reward is equivalent to minimizing expected regret, which we know grows at least logarithmically with time. If we can bound the number of expected cumulative transitions to grow less than logarithmically in time, then the regret term will dominate and the overall objective will be close to its optimum value. Our block allocation strategy is designed to make transitions less than logarithmically in time, thereby ensuring that the expected cumulative regret term dominates.

We know from the Lai-Robbins bound (1) that the expected number of selections of suboptimal arms i is at least $\mathcal{O}(\log T)$. Intuitively, the number of transitions can be minimized by selecting the option with the maximum upper credible limit $\lceil \log T \rceil$ times in a row. However, such a strategy will have a strong dependence on T and will not have a good performance uniformly in time. To remove this dependence on T, we divide the set of natural numbers (choice 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 each of which will correspond to a sequence of choices of the same option. Let the first $\lfloor 2^{k-1}/k \rfloor$ blocks in frame f_k have length k and the remaining choices in frame f_k constitute a single block of length $2^{k-1} - \lfloor 2^{k-1}/k \rfloor k$. The time associated with the choices made within frame f_k is $\mathcal{O}(2^k)$. Thus, following the intuition in the last paragraph, the length of each block in frame f_k is chosen equal to k, which is $\mathcal{O}(\log(2^k))$.

The total number of 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 $T < 2^\ell$. Each block is characterized by the tuple (k, r), for some $k \in \{1, \ldots, \ell\}$, and $r \in \{1, \ldots, b_k\}$, where k identifies the frame and r identifies the block within the frame. We denote the time at the start of block r in frame f_k by $\tau_{kr} \in \mathbb{N}$. The block UCL algorithm at time τ_{kr} selects the arm with the maximum $(1 - 1/K\tau_{kr})$-upper credible limit and chooses it k times in a row ($\leq k$ times if the block r is the last block in frame f_k). The choice at time τ_{kr} is analogous to the choice at each time instant in the UCL algorithm.

Next, we analyze the regret of the block UCL algorithm. We first introduce some notation. Let Q_i^{kr} be the $(1 - 1/K\tau_{kr})$-upper credible limit for the mean reward 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 time τ_{kr} (the start of block (k, r)). Let s_i^t be the number of times the decision-maker transitions to arm i from another arm $j \in \{1, \ldots, N\} \setminus \{i\}$ until time t. Let the empirical mean of the rewards from arm i until time τ_{kr} 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 mean reward at arm i at time τ_{kr} is a Gaussian random variable (M_i) 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 $(1 - 1/K\tau_{k,r})$-upper credible upper 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 $i \in \{1, \ldots, N\}$, we define constants

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

Let $\{R_t^{\mathrm{BUCL}}\}_{t \in \{1,\ldots,T\}}$ be the sequence of the expected regret of the block UCL algorithm, and $\{S_t^{\mathrm{BUCL}}\}_{t \in \{1,\ldots,T\}}$ be the sequence of expected transition costs. The block UCL algorithm achieves logarithmic regret uniformly in time as formalized in the following theorem.

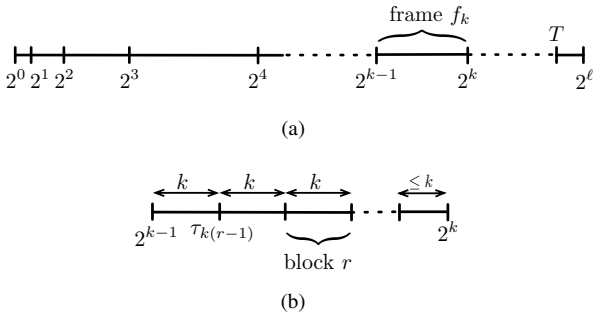


Fig. 9. The block allocation scheme used in the block UCL algorithm. Decision time t runs from left to right in both panels. Panel (a) shows the division of the decision times $t \in \{1, \ldots, T\}$ into frames $k \in \{1, \ldots, \ell\}$. Panel (b) shows how an arbitrary frame k is divided into blocks. Within the frame, an arm is selected at time τ_{kr}, the start of each block r in frame k, and that arm is selected for each of the k decisions in the block.

Theorem 9 (*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 uncorrelated 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 expected regret and the cumulative transition cost until time T satisfy*

$$\sum_{t=1}^{T} R_t^{\text{BUCL}} \leq \sum_{i=1}^{N} \Delta_i \Big(\gamma_1^i \log T - \frac{4\beta^2\sigma_s^2}{\Delta_i^2} \log\log T + \gamma_2^i \Big),$$

$$\sum_{t=1}^{T} S_t^{\text{BUCL}} \leq \sum_{i=1, i\neq i^*}^{N} (\bar{c}_i^{\max} + \bar{c}_{i^*}^{\max}) \times ((\gamma_1^i \log 2) \log\log T + \gamma_3^i) + \bar{c}_{i^*}^{\max}.$$

Proof: See Appendix D. ■

Figures 10 and 11 show, respectively, the cumulative expected regret and the cumulative transition cost of the block UCL algorithm on a bandit task with transition costs. For comparison, the figures also show the associated bounds from statement (iii) of Theorem 9. Cumulative expected regret was computed using 250 runs of the block UCL algorithm. Variance of the regret was minimal. The task used the reward surface of Landscape B from Figure 6(b) with sampling noise variance $\sigma_s^2 = 1$. The algorithm used an uncorrelated prior with $\mu_0 = 200$ and $\sigma_0^2 = 10^6$. Transition costs between options were equal to the distance between them on the surface.

The variance of the cumulative regret is relatively small, i.e., the cumulative regret experienced in a given task is close to the expected value. Also the bound on transition costs is quite loose. This is due to the loose bound on the expected number of transitions to the optimal arm. More detailed analysis of the total number of transitions would allow the bound to be tightened.

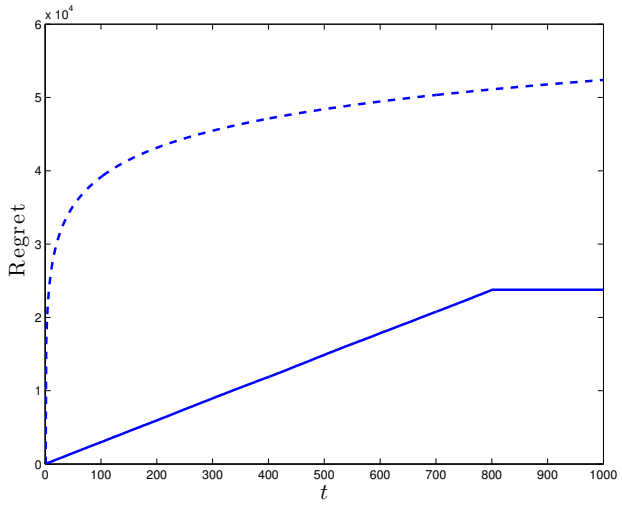


Fig. 10. Cumulative expected regret (solid line) and the associated bound (dashed line) from Theorem 9. Expected regret was computed using 250 runs of the block UCL algorithm; variance of the regret was minimal. The task used the reward surface from Figure 6(b) with sampling noise variance $\sigma_s^2 = 1$. The algorithm used an uncorrelated prior with $\mu_0 = 200$ and $\sigma_0^2 = 10^6$.

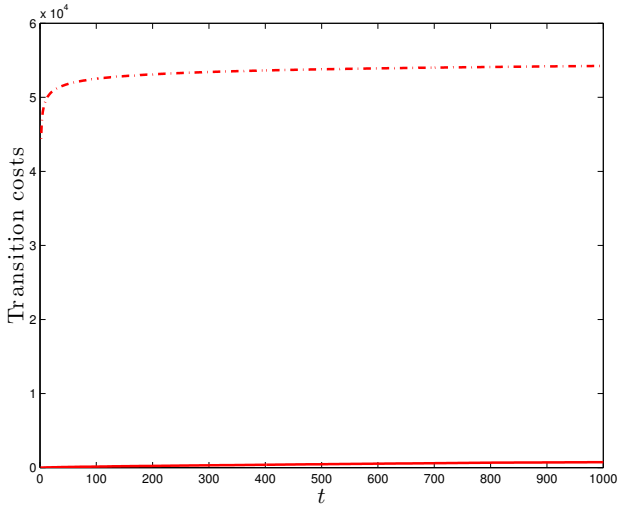


Fig. 11. Cumulative transition cost (solid line) and the associated bound (dashed line) from Theorem 9. Transition costs were computed using 250 runs of the block UCL algorithm with the same parameters as in Figure 10. Transition costs between any two arms i and j were deterministic and set equal to $|x_i - x_j|$, where x_i is the location of arm i in the grid.

VII. Graphical Gaussian multi-armed bandit problems

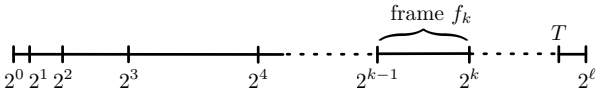
We now consider multi-armed bandits with Gaussian rewards in which the decision-maker cannot move to every other arm from the current arm. Let the set of arms that can be visited from arm i be $\text{ne}(i) \subseteq \{1, \ldots, N\}$. Such a multi-armed bandit can be represented by a graph $\mathcal{G}$ with node set $\{1, \ldots, N\}$ and edge set $\mathcal{E} = \{(i, j) \mid j \in \text{ne}(i), i \in \{1, \ldots, N\}\}$. We assume that the graph is connected in the sense that there exists at least one path from each node $i \in \{1, \ldots, N\}$ to every other node $j \in \{1, \ldots, N\}$. Let $\mathcal{P}^{ij}$ be the set of intermediary nodes in a shortest path from node i to node j. Note that the set $\mathcal{P}^{ij}$ does not contain node i nor node j. We denote the cardinality of the set $\mathcal{P}^{ij}$ by p_{ij} and accordingly, the elements of the set $\mathcal{P}^{ij}$ by $\{P_1^{ij}, \ldots, P_{p_{ij}}^{ij}\}$.

A. The graphical block UCL algorithm

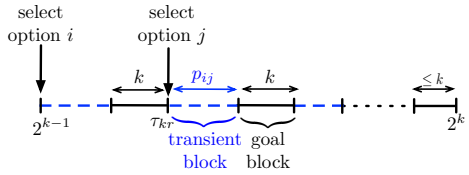
For graphical Gaussian multi-armed bandits, we develop an algorithm similar to the block allocation Algorithm 3, namely, the graphical block UCL algorithm, described in pseudocode in Algorithm 4 in Appendix F. Similar to the block allocation algorithm, at each block, the arm with maximum upper credible 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 maximum upper credible limit. Traversing a shortest path will mean making as many as $N - 2$ visits to undesirable arms ($N - 2$ is the worst case in a line graph where the current location is

at one end of the line and the desired arm is at the other end of the line). Thus, we apply a block allocation algorithm to limit the number of transitions as in the case of Algorithm 3 for the bandit problem with transition costs.

We classify the selection of arms in two categories, namely, *goal* selection and *transient* selection. The goal selection of an arm corresponds to the situation in which the arm is selected because it 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 shortest path to the arm with the maximum credible limit. Accordingly, we define the block associated with the goal selection of an arm as the *goal block*, and the block associated with arms on the shortest path between two arms associated with consecutive goal blocks as the *transient block*. The design of the blocks is pictorially depicted in Figure 12.



(a) Frames associated with goal selection of arms



(b) Goal and transient blocks within each frame

Fig. 12. The block allocation scheme used in the graphical block UCL algorithm. Decision time t runs from left to right in both panels. Panel (a) shows the division of goal selection instances of the arms into frames $k \in \{1, \ldots, \ell\}$. The frame f_k corresponds to 2^{k-1} goal selections of the arms. Panel (b) shows how an arbitrary frame f_k is divided into goal and transient blocks. The goal blocks are selected as they are in the block allocation strategy, while the transient blocks correspond to the shortest path between two arms associated with consecutive goal blocks. Only the goal selections are counted to compute the length of the frames.

The goal selection instances of arms are subdivided into frames $f_k, k \in \{1, \ldots, \ell\}$, where the length of the frame f_k is 2^{k-1}. Note that only goal selections are counted to compute the length of each frame. The length of the goal blocks within each frame is chosen as it is in the block allocation strategy. We denote the time at the start of the transient block before goal block r in frame f_k by $\tau_{kr} \in \mathbb{N}$. The graphical block allocation algorithm at time τ_{kr} (i) determines the arm with the maximum $(1-1/K\tau_{kr})$-upper credible limit, (ii) traverses the shortest path to the arm, (iii) picks the arm k times ($\leq k$ times if the goal block r is the last goal block in frame f_k). In Figure 12 the goal block only shows the choices of the goal selection. The transient block, shown prior to the corresponding goal block, accounts for the selections along a shortest path.

The key idea 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 the graphical bandit, sub-logarithmic transitions result in sub-logarithmic *undesired* visits to the arms on the chosen shortest path to the *desired* arm with maximum upper credible limit. Consequently, the cumulative expected regret of the algorithm is dominated by a logarithmic term.

B. Regret analysis of the graphical block UCL algorithm

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

Theorem 10 (***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 uncorrelated uninformative prior:*

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

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

(ii) *the cumulative expected regret until time T satisfies*

$$\sum_{t=1}^{T} R_t^{\text{GUCL}} \leq \sum_{i=1}^{N} \Big(\gamma_1^i \log T - \frac{4\beta^2\sigma_s^2}{\Delta_i^2}\log\log T + \gamma_2^i + \sum_{i=1, i\neq i^*}^{N} \left((2\gamma_1^i \log 2)\log\log T + 2\gamma_3^i\right) + 1\Big)\Delta_i.$$

Proof: See Appendix E. ∎

Figure 13 shows cumulative expected regret and the associated bound from Theorem 10 for the graphical block UCB algorithm. The underlying graph topology was chosen to be a line graph, so the algorithm could only choose to move one step forwards or backwards at each time. Expected regret was computed using 250 runs of the graphical block UCL algorithm. Each task consisted of $N = 10$ bandits with mean rewards set equal to the reward profile along the x-axis of Figure 6(b). Reward variance was $\sigma_s^2 = 6.25$, while the agent used the uncorrelated prior with $\mu_0 = 40$ and $\sigma_0^2 = 10^6$. Note that the regret bound is quite loose, as in the case of transition costs for the block UCL algorithm. This is because the regret bound uses the same bound on switching costs as in Theorem 9 to bound the regret incurred by traversing the graph.

VIII. Conclusions

In this paper, we considered multi-armed bandit problems with Gaussian rewards and studied them from a Bayesian perspective. We considered three particular multi-armed bandit problems: the standard multi-armed bandit problem, the multi-armed bandit problem with transition costs, and the graphical multi-armed bandit problem. We developed two UCL algorithms, namely, the deterministic UCL algorithm and the stochastic UCL algorithm, for the standard multi-armed bandit problem. We extended the deterministic UCL algorithm to the

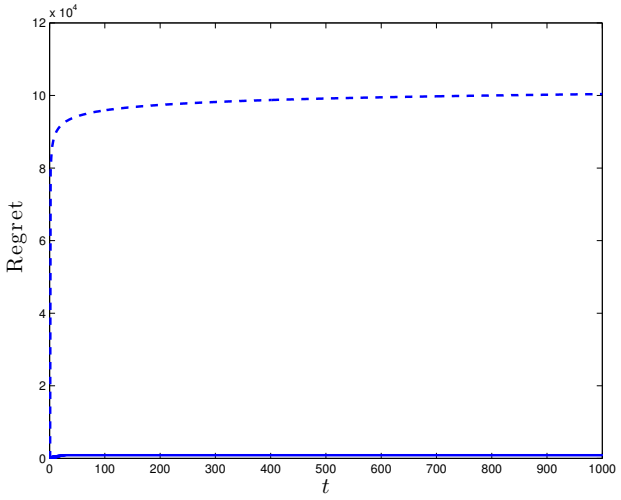


Fig. 13. Cumulative expected regret (solid line) and the associated bound (dashed line) from Theorem 10. Expected regret was computed from 250 simulated tasks played using the graphical block UCL algorithm. Each task consisted of $N = 10$ bandits with mean rewards set equal to the reward profile from Figure 6(b). The graph topology was a line graph, so the agent could only move one step forwards or backwards at each time. Reward variance was $\sigma_s^2 = 6.25$, while the agent used the uncorrelated prior with $\mu_0 = 40$ and $\sigma_0^2 = 10^6$.

block UCL algorithm and the graphical block UCL algorithm for the multi-armed bandit problem with transition costs, and the graphical multi-armed bandit problem, respectively. We established that for uninformative priors, each of the proposed algorithms achieves logarithmic regret uniformly in time, and moreover, the block UCL algorithm achieves a sub-logarithmic expected number of transitions among arms. We elucidated the role of general priors and the correlation structure among arms, showing how good priors and good assumptions on the correlation structure among arms can greatly enhance decision-making performance of the proposed deterministic UCL algorithm, even over short time horizons.

We drew connections between the features of the stochastic UCL and human decision-making in multi-armed bandit tasks. In particular, we showed how the stochastic UCL algorithm captures five key features of human decision-making in multi-armed bandit tasks, namely, (i) familiarity with the environment, (ii) ambiguity bonus, (iii) stochasticity, (iv) finite-horizon effects, and (v) environmental structure effects. We then presented empirical data from human decision-making experiments on a spatially-embedded multi-armed bandit task and demonstrated that the observed performance is efficiently captured by the proposed stochastic UCL algorithm with appropriate parameters.

This work presents several interesting avenues for future work in the design of human-automata systems. The model phenotypes discussed in Section V provide a method for assessing human performance in real time, and the experimental results presented in that section suggest that some humans use informative priors for spatial search tasks which allow them to achieve better performance than a similar algorithm using uninformative priors. Therefore, a useful goal for human-automata systems would be to develop a means to learn the humans' informative priors and use them to improve the performance of the overall system.

This work also presents several interesting avenues for future psychological research. First, in this work, we relied on certain functional forms for the parameters in the algorithms, e.g., we considered credibility parameter $\alpha_t = 1/Kt$ and cooling schedule $\upsilon_t = \nu/\log t$. It is of interest to perform thorough experiments with human subjects to ascertain the correctness of these functional forms. Second, efficient methods for estimation of parameters in the proposed algorithms need to be developed.

Overall, the proposed algorithms provide ample insights into plausible decision mechanisms involved with human decision-making in tasks with an explore-exploit tension. We envision a rich interplay between these algorithms and psychological research.

ACKNOWLEDGEMENTS

The authors wish to thank John Myles White, Robert C. Wilson, Philip Holmes and Jonathan D. Cohen for their input, which helped make possible the strong connection of this work to the psychology literature. We also thank the editor and two anonymous referees, whose comments greatly strengthened and clarified the paper. The first author is grateful to John Myles White and Daniel T. Swain for their help with implementing the online experiment.

REFERENCES

[1] P. Reverdy, R. C. Wilson, P. Holmes, and N. E. Leonard. Towards optimization of a human-inspired heuristic for solving explore-exploit problems. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2820–2825, Maui, HI, USA, December 2012.

[2] V. Srivastava, P. Reverdy, and N. E. Leonard. On optimal foraging and multi-armed bandits. In *Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing*, pages 494–499, Monticello, IL, USA, 2013.

[3] F. L. Lewis, D. Vrabie, and K.G. Vamvoudakis. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Magazine*, 32(6):76–105, 2012.

[4] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[5] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[6] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[7] C. J. C. H Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[8] P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 49–56, Cambridge, MA, 2007. MIT Press.

[9] J. D. Cohen, S. M. McClure, and A. J. Yu. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007.

[10] J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed Bandit Allocation Indices*. Wiley, second edition, 2011.

[11] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

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

[13] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[14] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

[15] M. Babaioff, Y. Sharma, and A. Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, pages 79–88, Stanford, CA, USA, July 2009.

[16] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th International Conference on Machine Learning*, pages 784–791, Helsinki, Finland, July 2008.

[17] J. L. Ny, M. Dahleh, and E. Feron. Multi-UAV dynamic routing with partial observations using restless bandit allocation indices. In *Proceedings of the American Control Conference*, pages 4220–4225, Seattle, Washington, USA, June 2008.

[18] B. P. McCall and J. J. McCall. A sequential study of migration and job search. *Journal of Labor Economics*, 5(4):452–476, 1987.

[19] M. Y. Cheung, J. Leighton, and F. S. Hover. Autonomous mobile acoustic relay positioning as a multi-armed bandit with switching costs. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3368–3373, Tokyo, Japan, November 2013.

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

[21] R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

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

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

[24] J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[25] N. Cesa-Bianchi and P. Fischer. Finite-time regret bounds for the multiarmed bandit problem. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 100–108, Madison, Wisconsin, USA, July 1998.

[26] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *JMLR: Workshop and Conference Proceedings*, volume 19: COLT 2011, pages 359–376, 2011.

[27] R. Dearden, N. Friedman, and S. Russell. Bayesian Q-learning. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence, AAAI-98*, pages 761–768, 1998.

[28] 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.

[29] S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. In S. Mannor, N. Srebro, and R. C. Williamson, editors, *JMLR: Workshop and Conference Proceedings*, volume 23: COLT 2012, pages 39.1–39.26, 2012.

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

[31] 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.

[32] J. S. Banks and R. K. Sundaram. Switching costs and the gittins index. *Econometrica: Journal of the Econometric Society*, 62(3):687–694, 1994.

[33] M. Asawa and D. Teneketzis. Multi-armed bandits with switching penalties. *IEEE Transactions on Automatic Control*, 41(3):328–348, 1996.

[34] T. Jun. A survey on the bandit problem with switching costs. *De Economist*, 152(4):513–541, 2004.

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

[36] D. Acuña and P. Schrater. Bayesian modeling of human sequential decision-making on the multi-armed bandit problem. In B. C. Love, K. McRae, and V. M. Sloutsky, editors, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 2065–2070, Washington, DC, USA, July 2008.

[37] D. E. Acuña and P. Schrater. Structure learning in human sequential decision-making. *PLoS Computational Biology*, 6(12):e1001003, 2010.

[38] M. Steyvers, M. D. Lee, and E. Wagenmakers. A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168–179, 2009.

[39] M. D. Lee, S. Zhang, M. Munro, and M. Steyvers. Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, 12(2):164–174, 2011.

[40] S. Zhang and A. J. Yu. Cheap but clever: Human active learning in a bandit setting. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1647–1652, Berlin, Germany, Aug 2013.

[41] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen. Why the grass is greener on the other side: Behavioral evidence for an ambiguity bonus in human exploratory decision-making. In *Neuroscience 2011 Abstracts*, Washington, DC, November 2011.

[42] D. Tomlin, A. Nedic, R. C. Wilson, P. Holmes, and J. D. Cohen. Group foraging task reveals separable influences of individual experience and social information. In *Neuroscience 2012 Abstracts*, New Orleans, LA, October 2012.

[43] A. Krause and C. E. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 324–331, Edinburgh, Scotland, July 2005.

[44] P. Fan. New inequalities of Mill's ratio and its application to the inverse Q-function approximation. *arXiv preprint arXiv:1212.4899*, Dec 2012.

[45] D. Bertsimas and J. N. Tsitsiklis. Simulated annealing. *Statistical Science*, 8(1):10–15, 1993.

[46] D. Mitra, F. Romeo, and A. Sangiovanni-Vincentelli. Convergence and finite-time behavior of simulated annealing. *Advances in Applied Probability*, 18(3):747–771, 1986.

[47] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[48] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume I : Estimation Theory*. Prentice Hall, 1993.

[49] J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 21(1):124–127, 1950.

[50] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[51] N. E. Leonard, D. A. Paley, F. Lekien, R. Sepulchre, D. M Fratantoni, and R. E Davis. Collective motion, sensor networks, and ocean sampling. *Proceedings of the IEEE*, 95(1):48–74, Jan. 2007.

[52] K. Friston, P. Schwartenbeck, T. Fitzgerald, M. Moutoussis, T. Behrens, and R. J. Dolan. The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience*, 7:598, 2013.

APPENDIX

A. Proof of inverse Gaussian tail bound

Proof of Theorem 1: We start by establishing inequality (6). It suffices to establish this inequality for $\beta = 1.02$. Since the cumulative distribution function for the standard normal random variable is a continuous and monotonically increasing function, it suffices to show that

$$\Phi(\beta\sqrt{-\log(-2\pi\alpha^2\log(2\pi\alpha^2))}) + \alpha - 1 \geq 0, \qquad (13)$$

for each $\alpha \in (0,1)$. Equation (13) can be equivalently written as $h(x) \geq 0$, where $x = 2\pi\alpha^2$ and $h : (0,1) \to (0, 1/\sqrt{2\pi})$ is defined by

$$h(x) = \Phi(\beta\sqrt{-\log(-x\log x)})) + \frac{\sqrt{x}}{\sqrt{2\pi}} - 1.$$

Note that $\lim_{x\to 0^+} h(x) = 0$ and $\lim_{x\to 1^-} h(x) = 1/\sqrt{2\pi}$. Therefore, to establish the theorem, it suffices to establish that h is a monotonically increasing function. It follows that

$$g(x) := 2\sqrt{2\pi}h'(x) = \frac{1}{\sqrt{x}} + \frac{\beta(-x\log x)^{\beta^2/2-1}(1+\log x)}{\sqrt{-\log(-x\log x)}}.$$

Note that $\lim_{x\to 0^+} g(x) = +\infty$ and $\lim_{x\to 1^-} g(x) = 1$. Therefore, to establish that h is monotonically increasing, it suffices to show that g is non-negative for $x \in (0,1)$. This is the case if the following inequality holds:

$$g(x) = \frac{1}{\sqrt{x}} + \frac{\beta(-x\log x)^{\beta^2/2-1}(1+\log x)}{\sqrt{-\log(-x\log x)}} \geq 0,$$

which holds if

$$\frac{1}{\sqrt{x}} \geq -\frac{\beta(-x\log x)^{\beta^2/2-1}(1+\log x)}{\sqrt{-\log(-x\log x)}}.$$

The inequality holds if the right hand side is negative. If it is positive, one can take the square of both sides and the inequality holds if

$$\begin{aligned}-\log(-x\log x) &\geq \beta^2 x(1+\log x)^2(-x\log x)^{\beta^2-2} \\ &= \beta^2 x(1+2\log x+(\log x)^2)(-x\log x)^{\beta^2-2}.\end{aligned}$$

Letting $t = -\log x$, the above inequality transforms to

$$-\log(te^{-t}) \geq \beta^2 e^{-t}(1-2t+t^2)(te^{-t})^{\beta^2-2},$$

which holds if

$$-\log t \geq \beta^2 t^{\beta^2-2}(1-2t+t^2)e^{-(\beta^2-1)t} - t.$$

Dividing by t, this is equivalent to

$$-\frac{\log t}{t} \geq \beta^2 t^{\beta^2-3}(1-2t+t^2)e^{-(\beta^2-1)t} - 1,$$

which is true if

$$\inf_{t\in[1,+\infty)} -\frac{\log t}{t} \geq \max_{t\in[1,+\infty)} \beta^2 t^{\beta^2-3}(1-2t+t^2)e^{-(\beta^2-1)t} - 1. \tag{14}$$

These extrema can be calculated analytically, so we have

$$\inf_{t\in[1,+\infty)} -\frac{\log t}{t} = -\frac{1}{e} \approx -0.3679$$

for the left hand side and

$$\begin{aligned}t^* &= \operatorname*{argmax}_{t\in[1,+\infty)} \beta^2 t^{\beta^2-3}(1-2t+t^2)e^{-(\beta^2-1)t} - 1 \\ &= 1+\sqrt{2/(\beta^2-1)} \\ \implies &\max_{t\in[1,+\infty)} \beta^2 t^{\beta^2-3}(1-2t+t^2)e^{-(\beta^2-1)t} - 1 \approx -0.3729,\end{aligned}$$

for the right hand side of (14). Therefore, (14) holds. In consequence, $g(x)$ is non-negative for $x \in (0,1)$, $h(x)$ is a monotonically increasing function. This establishes inequality (6). Inequality (7) follows analogously. ■

B. Proof of regret of the deterministic UCL algorithm

Proof of Theorem 2: We start by establishing the first statement. In the spirit of [22], we bound n_i^T as follows:

$$\begin{aligned}n_i^T &= \sum_{t=1}^{T} \mathbf{1}(i_t = i) \\ &\leq \sum_{t=1}^{T} \mathbf{1}\big(Q_i^t > Q_{i^*}^t\big) \\ &\leq \eta + \sum_{t=1}^{T} \mathbf{1}\Big(Q_i^t > Q_{i^*}^t, n_i^{(t-1)} \geq \eta\Big),\end{aligned}$$

where η is some positive integer and $\mathbf{1}(x)$ is the indicator function, with $\mathbf{1}(x) = 1$ if x is a true statement and 0 otherwise.

At time t, the agent picks option i over i^* only if

$$Q_{i^*}^t \leq Q_i^t.$$

This is true when at least one of the following equations holds:

$$\mu_{i^*}^t \leq m_{i^*} - C_{i^*}^t \tag{15}$$

$$\mu_i^t \geq m_i + C_i^t \tag{16}$$

$$m_{i^*} < m_i + 2C_i^t \tag{17}$$

where $C_i^t = \frac{\sigma_s}{\sqrt{\delta^2+n_{it}}}\Phi^{-1}(1-\alpha_t)$ and $\alpha_t = 1/Kt$. Otherwise, if none of the equations (15)-(17) holds,

$$Q_{i^*}^t = \mu_{i^*}^t + C_{i^*}^t > m_{i^*} \geq m_i + 2C_i^t > \mu_i^t + C_i^t = Q_i^t,$$

and option i^* is picked over option i at time t.

We proceed by analyzing the probability that Equations (15) and (16) hold. Note that the empirical mean $\bar{m}_i^t$ is a normal random variable with mean m_i and variance σ_s^2/n_i^t, so, conditional on n_i^t, μ_i^t is a normal random variable distributed as

$$\mu_i^t \sim \mathcal{N}\left(\frac{\delta^2\mu_i^0 + n_i^t m_i}{\delta^2+n_i^t}, \frac{n_i^t\sigma_s^2}{(\delta^2+n_i^t)^2}\right).$$

Equation (15) holds if

$$\begin{aligned}&m_{i^*} \geq \mu_{i^*}^t + \frac{\sigma_s}{\sqrt{\delta^2+n_i^t}}\Phi^{-1}(1-\alpha_t) \\ \iff &m_{i^*} - \mu_{i^*}^t \geq \frac{\sigma_s}{\sqrt{\delta^2+n_i^t}}\Phi^{-1}(1-\alpha_t) \\ \iff &z \leq -\sqrt{\frac{n_{i^*}^t+\delta^2}{n_{i^*}^t}}\Phi^{-1}(1-\alpha_t) + \frac{\delta^2}{\sigma_s}\frac{\Delta m_{i^*}}{\sqrt{n_{i^*}^t}},\end{aligned}$$

where $z \sim \mathcal{N}(0,1)$ is a standard normal random variable and $\Delta m_{i^*} = m_{i^*} - \mu_{i^*}^0$. For an uninformative prior $\delta^2 \to 0^+$, and consequently, equation (15) holds if and only if $z \leq -\Phi(1-\alpha_t)$. Therefore, for a uninformative prior,

$$\mathbb{P}(\text{Equation (15) holds}) = \alpha_t = \frac{1}{Kt} = \frac{1}{\sqrt{2\pi et}}.$$

Similarly, Equation (16) holds if

$$\begin{aligned}&m_i \leq \mu_i^t - \frac{\sigma_s}{\sqrt{\delta^2+n_{it}}}\Phi^{-1}(1-\alpha_t) \\ \iff &\mu_i^t - m_i \geq \frac{\sigma_s}{\sqrt{\delta^2+n_i^t}}\Phi^{-1}(1-\alpha_t) \\ \iff &z \geq \sqrt{\frac{n_i^t+\delta^2}{n_i^t}}\Phi^{-1}(1-\alpha_t) + \frac{\delta^2}{\sigma_s}\frac{\Delta m_i}{\sqrt{n_i^t}},\end{aligned}$$

where $z \sim \mathcal{N}(0,1)$ is a standard normal random variable and $\Delta m_i = m_i - \mu_i^0$. The analogous argument to that for the above case shows that, for an uninformative prior,

$$\mathbb{P}(\text{Equation (16) holds}) = \alpha_t = \frac{1}{Kt} = \frac{1}{\sqrt{2\pi et}}.$$

Equation (17) holds if

$$\begin{aligned}
& m_{i^*} < m_i + \frac{2\sigma_s}{\sqrt{\delta^2 + n_i^t}}\Phi^{-1}(1-\alpha_t) \\
\iff & \Delta_i < \frac{2\sigma_s}{\sqrt{\delta^2 + n_i^t}}\Phi^{-1}(1-\alpha_t) \\
\iff & \frac{\Delta_i^2}{4\beta^2\sigma_s^2}(\delta^2 + n_i^t) < -\log(-2\pi\alpha_t^2\log(2\pi\alpha_t^2)) \qquad (18) \\
\implies & \frac{\Delta_i^2}{4\beta^2\sigma_s^2}(\delta^2 + n_i^t) < \log(et^2) - \log\log(et^2) \\
\implies & \frac{\Delta_i^2}{4\beta^2\sigma_s^2}(\delta^2 + n_i^t) < \log(eT^2) - \log\log(eT^2) \qquad (19) \\
\implies & \frac{\Delta_i^2}{4\beta^2\sigma_s^2}(\delta^2 + n_i^t) < 1 + 2\log T - \log 2 - \log\log T
\end{aligned}$$

where $\Delta_i = m_{i^*} - m_i$, the inequality (18) follows from the bound (6), and the inequality (19) follows from the monotonicity of the function $\log x - \log\log x$ in the interval $[e, +\infty)$. Therefore, for an uninformative prior, inequality (17) never holds if

$$n_i^t \geq \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 + 2\log T - \log 2 - \log\log T).$$

Setting $\eta = \lceil \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 + 2\log T - \log 2 - \log\log T) \rceil$, we get

$$\begin{aligned}
\mathbb{E}\left[n_i^T\right] &\leq \eta + \sum_{t=1}^{T} \mathbb{P}(Q_i^t > Q_{i^*}^t, n_i^{(t-1)} \geq \eta) \\
&= \eta + \sum_{t=1}^{T} \mathbb{P}(\text{Equation (15) holds}, n_i^{(t-1)} \geq \eta) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}(\text{Equation (16) holds}, n_i^{(t-1)} \geq \eta) \\
&< \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 + 2\log T - \log 2 - \log\log T) \\
&\quad + 1 + \frac{2}{\sqrt{2\pi e}} \sum_{t=1}^{T} \frac{1}{t}.
\end{aligned}$$

The sum can be bounded by the integral

$$\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \int_1^T \frac{1}{t}\mathrm{d}t = 1 + \log T,$$

yielding the bound in the first statement

$$\begin{aligned}
\mathbb{E}\left[n_i^T\right] \leq & \left(\frac{8\beta^2\sigma_s^2}{\Delta_i^2} + \frac{2}{\sqrt{2\pi e}}\right)\log T \\
& + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2 - \log\log T) + 1 + \frac{2}{\sqrt{2\pi e}}.
\end{aligned}$$

The second statement follows from the definition of the cumulative expected regret. ■

C. Proof of regret of the stochastic UCL algorithm

Proof of Theorem 7: We start by establishing the first statement. We begin by bounding $\mathbb{E}[n_i^T]$ as follows

$$\mathbb{E}\left[n_i^T\right] = \sum_{t=1}^{T} \mathbb{E}\left[P_{it}\right] \leq \eta + \sum_{t=1}^{T} \mathbb{E}\left[P_{it}\mathbf{1}\big(n_i^t \geq \eta\big)\right], \qquad (20)$$

where η is a positive integer.

Now, decompose $\mathbb{E}\left[P_{it}\right]$ as

$$\begin{aligned}
\mathbb{E}[P_{it}] &= \mathbb{E}\left[P_{it}|Q_i^t \leq Q_{i^*}^t\right]\mathbb{P}(Q_i^t \leq Q_{i^*}^t) \\
&\quad + \mathbb{E}\left[P_{it}|Q_i^t > Q_{i^*}^t\right]\mathbb{P}(Q_i^t > Q_{i^*}^t) \\
&\leq \mathbb{E}\left[P_{it}|Q_i^t \leq Q_{i^*}^t\right] + \mathbb{P}(Q_i^t > Q_{i^*}^t). \qquad (21)
\end{aligned}$$

The probability P_{it} can itself be bounded as

$$P_{it} = \frac{\exp(Q_i^t/\upsilon_t)}{\sum_{j=1}^{N}\exp(Q_j^t/\upsilon_t)} \leq \frac{\exp(Q_i^t/\upsilon_t)}{\exp(Q_{i^*}^t/\upsilon_t)}. \qquad (22)$$

Substituting the expression for the cooling schedule in inequality (22), we obtain

$$P_{it} \leq \exp\left(-\frac{2(Q_{i^*}^t - Q_i^t)}{\Delta Q_{\min}^t}\log t\right) = t^{-\frac{2(Q_{i^*}^t - Q_i^t)}{\Delta Q_{\min}^t}}. \qquad (23)$$

For the purposes of the following analysis, define $\frac{0}{0} = 1$.

Since $\Delta Q_{\min}^t \geq 0$, with equality only if two arms have identical heuristic values, conditioned on $Q_{i^*}^t \geq Q_i^t$ the exponent on t can take the following magnitudes:

$$\frac{|Q_{i^*}^t - Q_i^t|}{\Delta Q_{\min}^t} = \begin{cases} \frac{0}{0} = 1, & \text{if } Q_{i^*}^t = Q_i^t, \\ +\infty, & \text{if } Q_{i^*}^t \neq Q_i^t \text{ and } \Delta Q_{\min}^t = 0, \\ x, & \text{if } \Delta Q_{\min}^t \neq 0, \end{cases}$$

where $x \in [1, +\infty)$. The sign of the exponent is determined by the sign of $Q_{i^*}^t - Q_i^t$.

Consequently, it follows from inequality (23) that

$$\sum_{t=1}^{T} \mathbb{E}[P_{it}|Q_{i^*}^t \geq Q_i^t] \leq \sum_{t=1}^{T} \frac{1}{t^2} \leq \frac{\pi^2}{6}.$$

It follows from inequality (21) that

$$\begin{aligned}
\sum_{i=1}^{T} \mathbb{E}[P_{it}] &\leq \frac{\pi^2}{6} + \sum_{i=1}^{T} \mathbb{P}(Q_i^t > Q_{i^*}^t) \\
&\leq \frac{\pi^2}{6} + \left(\frac{8\beta^2\sigma_s^2}{\Delta_i^2} + \frac{2}{\sqrt{2\pi e}}\right)\log T \\
&\quad + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2 - \log\log T) + 1 + \frac{2}{\sqrt{2\pi e}},
\end{aligned}$$

where the last inequality follows from Theorem 2. This establishes the first statement.

The second statement follows from the definition of the cumulative expected regret. ■

D. Proof of regret of the block UCL algorithm

Proof of Theorem 9: 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

$$\begin{aligned} n_i^T &= \sum_{t=1}^{T} \mathbf{1}(i_t = i) \\ &= \sum_{t=1}^{T} \big(\mathbf{1}(i_t = i \,\&\, n_i^{k_t r_t} < \eta) + \mathbf{1}(i_t = i \,\&\, n_i^{k_t r_t} \geq \eta)\big) \\ &\leq \eta + \ell + \sum_{t=1}^{T} \mathbf{1}(i_t = i \,\&\, n_i^{k_t r_t} \geq \eta) \\ &\leq \eta + \ell + \sum_{k=1}^{\ell} \sum_{r=1}^{b_k} k \mathbf{1}(i_{\tau_{kr}} = i \,\&\, n_i^{kr} \geq \eta). \end{aligned} \tag{24}$$

We note that $\mathbf{1}(i_{\tau_{kr}} = i) \leq \mathbf{1}(Q_i^{kr} > Q_{i^*}^{kr})$, where i^* is the optimal arm. We now analyze the event $\mathbf{1}(Q_i^{kr} > Q_{i^*}^{kr})$. It follows that $\mathbf{1}(Q_i^{kr} > Q_{i^*}^{kr}) = 1$ if the following inequalities hold:

$$\mu_{i^*}^{kr} \leq m_{i^*} - C_{i^*}^{kr} \tag{25}$$
$$\mu_i^{kr} \geq m_i + C_i^{kr} \tag{26}$$
$$m_{i^*} < m_i + 2C_i^{kr}, \tag{27}$$

where $C_i^{kr} = \frac{\sigma_s}{\sqrt{\delta^2 + n_i^{kr}}} \Phi^{-1}\big(1 - \frac{1}{K\tau_{kr}}\big)$. Otherwise if none of the inequalities (25)-(27) hold, then

$$Q_i^{kr} = \mu_i^{kr} + C_i^{kr} < \mu_{i^*}^{kr} + C_{i^*}^{kr} = Q_{i^*}^{kr}.$$

We now evaluate the probabilities of events (25)-(27). We note that

$$\begin{aligned} &\mathbb{P}(\mu_{i^*}^{kr} \leq m_{i^*} - C_{i^*}^{kr}) \\ &\leq \mathbb{P}\left(z \leq \frac{\delta^2 (m_{i^*} - \mu_{i^* 0})}{\sigma_s \sqrt{n_{i^*}^{kr}}} - \sqrt{\frac{\delta^2 + n_{i^*}^{kr}}{n_{i^*}^{kr}}} \Phi^{-1}\Big(1 - \frac{1}{K\tau_{kr}}\Big)\right), \end{aligned}$$

where $z \sim \mathcal{N}(0,1)$ is a standard normal random variable. Since $\delta^2 \to 0^+$ as $\sigma_0^2 \to +\infty$, it follows that

$$\mathbb{P}(\mu_{i^*}^{kr} \leq m_{i^*} - C_{i^*}^{kr}) \leq \mathbb{P}\Big(z \leq -\Phi^{-1}\Big(1 - \frac{1}{K\tau_{kr}}\Big)\Big) = \frac{1}{K\tau_{kr}}.$$

A similar argument shows that $\mathbb{P}(\mu_i^{kr} \geq m_i + C_i^{kr}) \leq 1/K\tau_{kr}$. We note that inequality (27) holds if

$$\begin{aligned} & m_{i^*} < m_i + 2\frac{\sigma_s}{\sqrt{\delta^2 + n_i^{kr}}} \Phi^{-1}\big(1 - \frac{1}{K\tau_{kr}}\big) \\ \implies & \Delta_i < 2\frac{\sigma_s}{\sqrt{\delta^2 + n_i^{kr}}} \Phi^{-1}\big(1 - \frac{1}{K\tau_{kr}}\big) \\ \implies & \Delta_i^2 < -4\frac{\sigma_s^2}{\delta^2 + n_i^{kr}} \beta^2 \log\Big(-\frac{2\pi}{K\tau_{kr}} \log\Big(\frac{2\pi}{K^2 \tau_{kr}^2}\Big)\Big) \\ & < \frac{4\beta^2 \sigma_s^2}{\delta^2 + n_i^{kr}} \Big(\log(e\tau_{kr}^2) - \log\log(e\tau_{kr}^2)\Big). \end{aligned}$$

Since $\log x - \log\log x$ achieves its minimum at $x = e$, it follows that $\log(e\tau_{kr}^2) - \log\log(e\tau_{kr}^2) \leq \log(eT^2) - \log\log(eT^2)$. Consequently, inequality (27) holds if

$$\begin{aligned} \Delta_i^2 &< \frac{4\beta^2\sigma_s^2}{\delta^2 + n_i^{kr}} \Big(1 + 2\log T - \log(1 + 2\log T)\Big) \\ &< \frac{4\beta^2\sigma_s^2}{\delta^2 + n_i^{kr}} \Big(1 + 2\log T - \log\log T - \log 2\Big). \end{aligned}$$

Since $\delta^2 \to 0^+$, it follows that inequality (27) does not hold if

$$n_i^{kr} \geq \frac{8\beta^2\sigma_s^2}{\Delta_i^2}\Big(\log T - \frac{1}{2}\log\log T\Big) + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2).$$

Therefore, if we choose $\eta = \lceil \frac{8\beta^2\sigma_s^2}{\Delta_i^2}(\log T - \frac{1}{2}\log\log T) + \frac{4\beta^2\sigma_s^2}{\Delta_i^2}(1 - \log 2) \rceil$, it follows from equation (24) that

$$\mathbb{E}[n_i^T] \leq \eta + \ell + \frac{2}{K} \sum_{k=1}^{\ell} \sum_{r=1}^{b_k} \frac{k}{\tau_{kr}}. \tag{28}$$

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

$$\begin{aligned} \sum_{r=1}^{b_k} \frac{k}{\tau_{kr}} &= \sum_{r=1}^{b_k} \frac{k}{2^{k-1} + (r-1)k} \\ &\leq \frac{k}{2^{k-1}} + \int_1^{b_k} \frac{k}{k(x-1) + 2^{k-1}} \mathrm{d}x \\ &= \frac{k}{2^{k-1}} + \log \frac{2^{k-1} + k(b_k - 1)}{2^{k-1}} \\ &\leq \frac{k}{2^{k-1}} + \log 2. \end{aligned} \tag{29}$$

Since $T \geq 2^{\ell - 1}$, it follows that $\ell \leq 1 + \log_2 T =: \bar{\ell}$. Therefore, inequalities (28) and (29) 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 [31], 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 in 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

$$\mathbb{E}[s_i^T] \leq \frac{\mathbb{E}[n_i^{2^\ell}]}{\ell} + \sum_{k=1}^{\ell-1} \frac{\mathbb{E}[n_i^{2^k}]}{k^2}. \tag{30}$$

We now analyze inequality (30) separately for the three terms in the upper bound on $\mathbb{E}[n_i^T]$. For the first logarithmic term, the right hand side of inequality (30) yields

$$\frac{\gamma_1^i \log 2^\ell}{\ell} + \sum_{k=1}^{\ell-1} \frac{\gamma_1^i \log 2^k}{k^2} = \gamma_1^i \log 2 \Big(1 + \sum_{k=1}^{\ell-1} \frac{1}{k}\Big) \leq \gamma_1^i \log 2 (\log\log T + 2 - \log\log 2). \quad (31)$$

For the second sub-logarithmic term, the right hand side of inequality (30) is equal to

$$\begin{aligned} &- \frac{4\beta^2\sigma_s^2}{\Delta_i^2}\Big(\frac{(\log \ell + \log\log 2)}{\ell} + \sum_{k=1}^{\ell-1} \frac{(\log k + \log\log 2)}{k^2}\Big) \\ &\leq -\frac{4\beta^2\sigma_s^2}{\Delta_i^2}\Big(\frac{(\log\log 2)}{\ell} + \sum_{k=1}^{\ell-1} \frac{\log\log 2}{k^2}\Big) \\ &\leq -\frac{4\beta^2\sigma_s^2}{\Delta_i^2}\Big(1 + \frac{\pi^2}{6}\Big)\log\log 2. \end{aligned} \quad (32)$$

Similarly, for the constant term γ_2, the right hand side of inequality (30) is equal to

$$\frac{\gamma_2^i}{\ell} + \sum_{k=1}^{\ell-1} \frac{\gamma_2^i}{k^2} \leq \gamma_2^i \Big(1 + \frac{\pi^2}{6}\Big). \quad (33)$$

Collecting the terms from inequalities (31)-(33), it follows from inequality (30) that

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

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

$$\begin{aligned} \sum_{t=1}^T S_t^{\mathrm{B}} &\leq \sum_{i=1, i\neq i^*}^N \bar{c}_i^{\max} \mathbb{E}[s_i^T] + \bar{c}_{i^*}^{\max} \mathbb{E}[s_{i^*}^T] \\ &\leq \sum_{i=1, i\neq i^*}^N (\bar{c}_i^{\max} + \bar{c}_{i^*}^{\max}) \mathbb{E}[s_i^T] + \bar{c}_{i^*}^{\max}, \end{aligned} \quad (34)$$

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 (34) and the second statement. ∎

E. Proof of regret of the graphical block UCL algorithm

Proof of Theorem 10: We start by establishing the first statement. 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 UCL Algorithm 3, 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 UCL Algorithm 3, 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. ∎

F. Pseudocode implementations of the UCL algorithms

Algorithm 1: *Deterministic UCL Algorithm*

Input : prior $\mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 I_N)$, variance σ_s^2;
Output : allocation sequence $\{i_t\}_{t\in\{1,\ldots,T\}}$;

set $n_i \leftarrow 0, \bar{m}_i \leftarrow 0$, for each $i \in \{1,\ldots,N\}$;
set $\delta^2 = \frac{\sigma_s^2}{\sigma_0^2}$; $K \leftarrow \sqrt{2\pi e}$; $T_0^{\text{end}} \leftarrow 0$;
% at each time pick the arm with maximum upper credible limit
for $\tau \in \{1,\ldots,T\}$ **do**
  for $i \in \{1,\ldots,N\}$ **do**
    $Q_i \leftarrow \frac{\delta^2\mu_i^0 + n_i\bar{m}_i}{\delta^2 + n_i} + \frac{\sigma_s}{\sqrt{\delta^2 + n_i}}\Phi^{-1}\Big(1 - \frac{1}{K\tau}\Big)$;
  $i_\tau \leftarrow \text{argmax}\{Q_i \mid i \in \{1,\ldots,N\}\}$;
  collect reward m^{real};
  $\bar{m}_{i_\tau} \leftarrow \frac{n_{i_\tau}\bar{m}_{i_\tau} + m}{n_{i_\tau} + 1}$;
  $n_{i_\tau} \leftarrow n_{i_\tau} + 1$;

Algorithm 2: *Stochastic UCL Algorithm*

Input : prior $\mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 I_N)$, variance σ_s^2;
Output : allocation sequence $\{i_t\}_{t\in\{1,\ldots,T\}}$;

set $n_i \leftarrow 0, \bar{m}_i \leftarrow 0$, for each $i \in \{1,\ldots,N\}$;
set $\delta^2 = \frac{\sigma_s^2}{\sigma_0^2}$; $K \leftarrow \sqrt{2\pi e}$; $T_0^{\text{end}} \leftarrow 0$;
% at each time pick an arm using Boltzmann probability distribution
for $\tau \in \{1,\ldots,T\}$ **do**
  for $i \in \{1,\ldots,N\}$ **do**
    $Q_i \leftarrow \frac{\delta^2\mu_i^0 + n_i\bar{m}_i}{\delta^2 + n_i} + \frac{\sigma_s}{\sqrt{\delta^2 + n_i}}\Phi^{-1}\Big(1 - \frac{1}{K\tau}\Big)$;
  $\Delta Q_{\min} = \min_{i,t} |Q_i - Q_j|$;
  $\upsilon_\tau \leftarrow \frac{\Delta Q_{\min}}{2\log\tau}$;
  select i_τ with probability $p_i \propto \exp(Q_i/\upsilon_\tau)$;
  collect reward m^{real};
  $\bar{m}_{i_\tau} \leftarrow \frac{n_{i_\tau}\bar{m}_{i_\tau} + m}{n_{i_\tau} + 1}$;
  $n_{i_\tau} \leftarrow n_{i_\tau} + 1$;

Algorithm 3: *Block UCL Algorithm*

Input : prior $\mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 I_N)$, variance σ_s^2 ;
Output : allocation sequence $\{i_t\}_{t\in\{1,\ldots,T\}}$;

set $n_i \leftarrow 0, \bar{m}_i \leftarrow 0, \forall\ i \in \{1,\ldots,N\}$; $\delta^2 \leftarrow \dfrac{\sigma_s^2}{\sigma_0^2}$; $K \leftarrow \sqrt{2\pi e}$;

% at each allocation round pick the arm with maximum UCL

for $k \in \{1,\ldots,\ell\}$ **do**
  for $r \in \{1,\ldots,b_k\}$ **do**
    $\tau \leftarrow 2^{k-1} + (r-1)k$
    $Q_i \leftarrow \dfrac{\delta^2\mu_i^0 + n_i\bar{m}_i}{\delta^2 + n_i} + \dfrac{\sigma_s}{\sqrt{\delta^2+n_i}}\Phi^{-1}\Big(1 - \dfrac{1}{K\tau}\Big)$;
    $\hat{i} \leftarrow \text{argmax}\{Q_i \mid i \in \{1,\ldots,N\}\}$;
    if $2^k - \tau \geq k$ **then**
      set $i_t \leftarrow \hat{i}$, for each $t \in \{\tau,\ldots,\tau+k\}$;
      collect reward $m_t^{\text{real}},\ t \in \{1,\ldots,k\}$;
      $\bar{m}_{\hat{i}} \leftarrow \dfrac{n_{\hat{i}}\bar{m}_{\hat{i}} + \sum_{t=1}^{k} m_t^{\text{real}}}{n_{\hat{i}} + k}$;
      $n_{\hat{i}} \leftarrow n_{\hat{i}} + k$;
    else
      set $i_t \leftarrow \hat{i}$, for each $t \in \{\tau,\ldots,2^k-1\}$;
      collect reward $m_t^{\text{real}},\ t \in \{1,\ldots,2^k-\tau\}$;
      $\bar{m}_{\hat{i}} \leftarrow \dfrac{n_{\hat{i}}\bar{m}_{\hat{i}} + \sum_{t=1}^{2^k-\tau} m_t^{\text{real}}}{n_{\hat{i}} + 2^k - \tau}$;
      $n_{\hat{i}} \leftarrow n_{\hat{i}} + 2^k - \tau$;

Algorithm 4: *Graphical Block UCL Algorithm*

Input : prior $\mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 I_N)$, variance σ_s^2 ;
Output : allocation sequence $\{i_t\}_{t\in\{1,\ldots,T\}}$;

set $n_i \leftarrow 0, \bar{m}_i \leftarrow 0, \forall\ i \in \{1,\ldots,N\}$; $\delta^2 \leftarrow \dfrac{\sigma_s^2}{\sigma_0^2}$; $K \leftarrow \sqrt{2\pi e}$;

set $\tau \leftarrow 1$; $i_0 \leftarrow 1$;

% at each allocation round pick the arm with maximum UCL

for $k \in \{1,\ldots,\ell\}$ **do**
  for $r \in \{1,\ldots,b_k\}$ **do**
    $Q_i \leftarrow \dfrac{\delta^2\mu_i^0 + n_i\bar{m}_i}{\delta^2 + n_i} + \dfrac{\sigma_s}{\sqrt{\delta^2+n_i}}\Phi^{-1}\Big(1 - \dfrac{1}{K\tau}\Big)$;
    $\hat{i} \leftarrow \text{argmax}\{Q_i \mid i \in \{1,\ldots,N\}\}$;
    % reach node $\hat{i}$ using the shortest path
    for $t \in \{\tau,\ldots,\tau + p_{i_\tau\hat{i}} - 1\}$ **do**
      set $i_t \leftarrow P_{t-\tau+1}^{i_\tau\hat{i}}$;
      collect rewards m^{real};
      $\bar{m}_{i_t} \leftarrow \dfrac{n_{i_t}\bar{m}_{i_t} + m^{\text{real}}}{n_{i_t}+1}$;
      $n_{i_t} \leftarrow n_{i_t} + 1$;
    set $\tau \leftarrow \tau + p_{i_\tau\hat{i}}$;
    if $2^{k-1} - (r-1)k \geq k$ **then**
      set $i_t \leftarrow \hat{i}$, for each $t \in \{\tau,\ldots,\tau+k-1\}$;
      collect reward $m_t^{\text{real}},\ t \in \{1,\ldots,k\}$;
      $\bar{m}_{\hat{i}} \leftarrow \dfrac{n_{\hat{i}}\bar{m}_{\hat{i}} + \sum_{t=1}^{k} m_t^{\text{real}}}{n_{\hat{i}} + k}$;
      $n_{\hat{i}} \leftarrow n_{\hat{i}} + k$;
      $\tau \leftarrow \tau + k$;
    else
      set $i_t \leftarrow \hat{i}$, for each $t \in \{2^{k-1} + (r-1)k,\ldots,2^k-1\}$;
      collect reward $m_t^{\text{real}},\ t \in \{1,\ldots,2^{k-1} - (r-1)k\}$;
      $\bar{m}_{\hat{i}} \leftarrow \dfrac{n_{\hat{i}}\bar{m}_{\hat{i}} + \sum_{t=1}^{2^{k-1}-(r-1)k} m_t^{\text{real}}}{n_{\hat{i}} + 2^{k-1} - (r-1)k}$;
      $n_{\hat{i}} \leftarrow n_{\hat{i}} + 2^{k-1} - (r-1)k$;
      $\tau \leftarrow \tau + 2^{k-1} - (r-1)k$;