Algorithmic models of human decision making in Gaussian multi-armed bandit problems

Paul Reverdy, Vaibhav Srivastava and Naomi E. Leonard

***Abstract*— We consider a heuristic Bayesian algorithm as a model of human decision making in multi-armed bandit problems with Gaussian rewards. We derive a novel upper bound on the Gaussian inverse cumulative distribution function and use it to show that the algorithm achieves logarithmic regret. We extend the algorithm to allow for stochastic decision making using Boltzmann action selection with a dynamic temperature parameter and provide a feedback rule for tuning the temperature parameter such that the stochastic algorithm achieves logarithmic regret. The stochastic algorithm encodes many of the observed features of human decision making.**

I. INTRODUCTION

The multi-armed bandit problem has been extensively studied in the machine learning and controls community [3], [4], [6], [9]. It is a canonical model of decision making under uncertainty where the explore-exploit tradeoff is central. At each of a sequence of times, the decision maker chooses one among finite options (arms), with uncertain associated rewards, aiming to maximize accumulated reward over the whole sequence. When the decision maker chooses the most rewarding among known options, the strategy is called *exploitation*, and when the decision maker chooses a poorly known but potentially revealing option, the strategy is called *exploration*. Good strategies balance exploration to reduce uncertainty and exploitation to accumulate high reward.

In the controls literature, the multi-armed bandit problem is a model problem for adaptive control [2], [9] and has been applied to a variety of problems, including multi-agent task assignment [11] and channel allocation for networks [1].

The performance of algorithms solving the multi-armed bandit problem can be characterized in terms of *regret*, which is the accumulated difference between the highest available reward and the expected reward of the algorithm. Lai and Robbins [10] proved that any algorithm solving the multi-armed bandit problem must incur regret that grows logarithmically with time, and they provided an algorithm that asymptotically achieves that bound. Since then, a significant line of research has focused on providing algorithms that uniformly achieve the Lai-Robbins bound.

One such class of algorithms is the so-called Upper Confidence Bound (UCB) algorithms, first introduced by Auer *et al.* [3]. For each decision time t, these algorithms compute a heuristic value for each option i which provides an upper bound for the expected reward to be gained by selecting that option:

$$Q_i^t = \mu_i^t + C_i^t, \tag{1}$$

where μ_i^t is the expected reward and C_i^t is a measure of uncertainty in the reward of option i at time t. The algorithm's decision at time t is to pick the option i that maximizes Q_i^t.

UCB1, the main algorithm introduced in [3], is designed for the case where rewards are drawn from a distribution with bounded support. In this case, Auer *et al.* proved that UCB1 achieves logarithmic regret. They also considered the case where rewards are drawn from a Gaussian distribution with unknown variance and introduced an algorithm they call UCB1-Normal to solve it. They analyzed the performance of UCB1-Normal and showed that it achieves logarithmic regret, but their proof relies on several conjectures about Student and χ^2 random variables that they only verify numerically. Liu and Zhao [12] studied multi-armed bandit problems where the rewards are drawn from a light-tailed distribution, which includes Gaussian distributions with known variance as a special case. For such light-tailed rewards, they extended UCB1 to achieve logarithmic regret. UCB1 and its variants rely on frequentist estimators, and therefore cannot incorporate prior knowledge about the rewards.

Recent work in neuroscience [15] showed that human decision making in multi-armed bandit problems is consistent with a stochastic heuristic similar to (1). In the present paper we construct an algorithmic model of human decision making that formalizes the connection between the heuristics used by humans and the UCB algorithms. Our model uses stochastic decision making and Bayesian estimators to incorporate prior knowledge about the rewards.

In the present work, we provide a Bayesian algorithm for the Gaussian multi-armed bandit problem and prove that it achieves logarithmic regret in certain cases. The algorithm is derived by applying the ideas in [7] to the case of bandits with Gaussian rewards, and adding a noise model to the decision process. Rather than following the analysis in [7], our analysis follows Auer *et al.* [3] and facilitates extensions to the case of stochastic decision making as well as to the case of the multi-armed bandit with transition costs and the graphical multi-armed bandit, considered in [14]. We make use of a novel upper bound on the inverse cumulative distribution function for the standard Gaussian distribution, which we present in Theorem 1. The bound is tighter than the one used by Liu and Zhao [12], and allows us to achieve a smaller leading factor in the case of Gaussian rewards.

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 a NDSEG Fellowship.

Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA {preverdy,vaibhavs,naomi}@princeton.edu

The contributions of this paper are threefold. First, we provide a deterministic Bayesian algorithm that provably achieves uniform logarithmic regret in the case of Gaussian bandits. Second, we show how to extend this algorithm to employ stochastic policies while still achieving logarithmic regret. Third, we show how the stochastic algorithm can be used as a model of human decision making in multi-armed bandit problems.

The remainder of the paper is organized as follows. In Section II we describe the multi-armed bandit problem. In Section III we review the algorithm of [7] and apply it to the case of Gaussian bandits. In Section IV we analyze the finite-time properties of the model and prove that it achieves logarithmic regret in the case of deterministic decision making. We extend the model to the case of stochastic decision making using Boltzmann action selection with a dynamic temperature parameter and provide a feedback rule that tunes the temperature parameter such that the model again achieves logarithmic regret. Finally, we conclude in Section V.

II. GAUSSIAN BANDITS PROBLEM

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 $N > 1$ options is often called a multi-armed bandit. In the multi-armed bandit problem, the decision-making agent must choose, at each of a sequence of times, one among N arms. Each arm i has an associated mean reward m_i, which is unknown to the agent and remains fixed for the duration of the problem.

The agent collects rewards by choosing arm i_t at each time $t = 1, 2, \ldots, T$ and receiving reward r_t, which is the mean reward associated with the arm plus Gaussian noise: $r_t \sim \mathcal{N}(m_{i_t}, \sigma_r)$. The noise variance σ_r is assumed known, e.g. from previous observations or known characteristics of the reward generation process.

The agent's objective is to maximize cumulative expected reward by choosing a sequence of arms $\{i_t\}$:

$$\max_{\{i_t\}} J, \quad J = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \sum_{t=1}^{T} m_{i_t}. \tag{2}$$

In this context exploitation refers to picking arm i_t which appears to have the highest mean at time t, and exploration refers to picking any other arm.

Equivalently, defining $m_{i^*} = \max_i m_i$ 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

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

where n_i^T is the cumulative number of times arm i has been chosen up to time T and $\Delta_i = m_{i^*} - m_i$ is the expected regret due to picking arm i instead of arm i^*.

A. Bound on optimal performance

Lai and Robbins [10] showed that any algorithm solving the multi-armed bandit problem must choose suboptimal arms at a rate that is at least logarithmic in time:

$$\mathbb{E}\left[n_i^T\right] \geq \left(\frac{1}{D(p_i||p_{i^*})} + o(1)\right) \log T, \tag{3}$$

where $o(1) \to 0$ as $T \to +\infty$ and $D(p_i||p_{i^*}) := \int p_i(r) \log \frac{p_i(r)}{p_{i^*}(r)} \mathrm{d}r$ is the Kullback-Liebler 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 a bound on cumulative regret J_R, showing that it must grow at least logarithmically with time.

In the present case where $r_i \sim \mathcal{N}(m_i, \sigma_r)$, the Kullback-Liebler divergence is equal to

$$D(p_i||p_{i^*}) = \frac{\Delta_i^2}{2\sigma_r^2}, \tag{4}$$

so the bound is

$$\mathbb{E}\left[n_i^T\right] \geq \left(\frac{2\sigma_r^2}{\Delta_i^2} + o(1)\right) \log T. \tag{5}$$

The intuition is that for a fixed value of σ_r, a suboptimal arm i with higher Δ_i is easier to identify since it yields a lower average reward. Conversely, for a fixed value of Δ_i, higher values of σ_r mean that the observed rewards are more variable, making it more difficult to distinguish the optimal arm i^* from the suboptimal ones.

B. Bayes-UCB

For every probability distribution $f(x)$ with associated cumulative distribution function (cdf) $F(x)$, the *quantile* function $F^{-1}(p)$ inverts the cdf to provide an upper bound for the value of the random variable $X \sim f(x)$:

$$\Pr\left[X \leq F^{-1}(p)\right] = p. \tag{6}$$

In this sense, $F^{-1}(p)$ is an *upper confidence bound*, an upper bound that holds with probability, or *confidence level*, p. The authors of [7] considered the multi-armed bandit problem from a Bayesian perspective and suggested using $F^{-1}(p)$ of the posterior reward distribution as the heuristic function (1). The intuition is that $Q_i = F^{-1}(p)$ gives a bound such that $\Pr\left[m_i > Q_i\right] = 1 - p$, so that if $p < 1$ is large, then $1 - p$ is small and it is unlikely that the true mean reward for arm i is higher than the bound.

In order to be increasingly sure of choosing the optimal arm as time goes on, the algorithm in [7] 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 term the resulting algorithm Bayes-UCB, and in the case that the rewards are Bernoulli distributed they proved that with $c \geq 5$ Bayes-UCB achieves the bound (3).

III. THE UPPER CREDIBLE LIMIT ALGORITHM

We apply Bayes-UCB to the case of bandits with Gaussian rewards of known variance σ_r^2 and term the resulting algorithm the *Upper Credible Limit algorithm*, or UCL. We then consider an extension of UCL to a stochastic policy by using Boltzmann action selection and term the resulting algorithm *stochastic UCL*.

A. Inference algorithm

We begin by assuming that the agent's prior distribution of $\mathbf{m}$ (i.e. the agent's initial beliefs about the mean reward values $\mathbf{m}$ and their covariance Σ) is multivariate Gaussian with mean $\boldsymbol{\mu_0}$ and covariance Σ_0:

$$\mathbf{m} \sim \mathcal{N}(\boldsymbol{\mu_0}, \Sigma_0),$$

where $\boldsymbol{\mu_0} \in \mathbb{R}^N$ and $\Sigma_0 \in \mathbb{R}^{N\times N}$ is a positive-definite matrix. Note that this does not assume that the rewards are truly described by these statistics, simply that these are the agent's initial beliefs, informed perhaps by previous measurements of the mean value and covariance.

With this prior, the posterior distribution is also Gaussian, so the Bayesian optimal inference algorithm is linear and can be written down as follows. At each time t, the agent selects arm i_t and receives a reward r_t. Recall that n_i^t is defined as the number of times the agent has selected arm i up to time t, and let $\bar{m}_i^t$ be the empirical mean reward observed for arm i. Let $\mathbf{n}_t$ and $\bar{\mathbf{m}}_t$ be the corresponding vectors with components $n_i^t, \bar{m}_i^t$, respectively.

Then the belief state $(\boldsymbol{\mu}_t, \Sigma_t)$ updates as follows:

$$\Lambda_t = \frac{\mathrm{diag}(\mathbf{n}_t)}{\sigma_r^2} + \Lambda_0,\ \Sigma_t = \Lambda_t^{-1} \tag{7}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \Sigma_t \frac{\mathrm{diag}(\mathbf{n}_t)}{\sigma_r^2}(\bar{\mathbf{m}}_t - \boldsymbol{\mu}_0), \tag{8}$$

where $\Lambda_t = \Sigma_t^{-1}$ is the *precision* matrix. As noted above, this assumes that the sampling noise σ_r is known, e.g. from previous observations or known sensor characteristics.

The above holds for general $\Sigma_0 > 0$, but for simplicity of exposition we will specialize in the following to the case where $\Sigma_0 = \sigma_0^2 I$, so the agent believes the mean rewards to be independent. In this case the belief state update equations simplify to

$$\mathrm{Var}(m_i^t|\bar{m}_i^t) = \left(\sigma_i^t\right)^2 = \frac{\sigma_r^2}{\delta^2 + n_i^t}$$

$$\mathbb{E}\left[m_i^t|\bar{m}_i^t\right] = \mu_i^t = \frac{\delta^2\mu_i^0 + n_i^t\bar{m}_i^t}{\delta^2 + n_i^t},$$

where $\delta^2 = \sigma_r^2/\sigma_0^2$.

B. Quantile function

With the assumption of independence made above, the posterior distribution of the mean m_i at time t is

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

so the $(1-\alpha)^{th}$ quantile of the distribution is given by

$$F^{-1}(1-\alpha) = \mu_i^t + \frac{\sigma_r}{\sqrt{\delta^2 + n_i^t}}\Phi^{-1}(1-\alpha), \tag{9}$$

where $\Phi^{-1}(p)$ is the inverse of the cdf of the normal distribution, also known as the probit function.

C. Decision heuristic

In the case of deterministic decision making, the decision at time t is given by maximizing the heuristic:

$$i_t = \arg\max_i Q_i^t, \tag{10}$$

where the heuristic function is defined by the quantile (9),

$$Q_i^t = \mu_i^t + \frac{\sigma_r}{\sqrt{\delta^2 + n_i^t}}\Phi^{-1}(1-\alpha_t), \tag{11}$$

with $\alpha_t = 1/Kt$ and $K > 1$ is a constant.

We extend the algorithm to the case of stochastic decision making using Boltzmann action selection, as is used in simulated annealing [13], [8]. The choice of arm is made stochastically 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 reduces to the deterministic scheme (10), and as υ_t increases the probability of selecting any other arm increases. In this way, Boltzmann selection generalizes the maximum operation and is sometimes known as the soft maximum (or softmax) rule.

In the context of simulated annealing, the choice of υ_t is known as a cooling schedule. In their classic work, Mitra *et al.* [13] showed that good cooling schedules for simulated annealing take the form

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

so we study cooling schedules of this form. We choose ν using a feedback rule on the values of the heuristic function Q_i^t and define the cooling schedule as

$$\upsilon_t = \frac{\Delta Q_{\min}^t D}{2\log t}, \tag{12}$$

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

Much of the bandits literature considers only deterministic maximization rules; for example, Bayes-UCB, as presented in [7], is a deterministic decision rule. However, several authors have considered stochastic decision rules in adversarial contexts, where it is advantageous to avoid making predictable decisions. See Chapter 3 of the recent review [4] and references therein.

D. Application to human decision making

Human decision making in multi-armed bandit problems is well modeled by a heuristic similar to that of UCL (11) and humans are sensitive to the parameters of the problem [15]. In particular, both the uncertainty measure and the level of decision noise increase with problem horizon T.

Stochastic UCL can be used to model human decision making. By choosing the parameters K and D as increasing

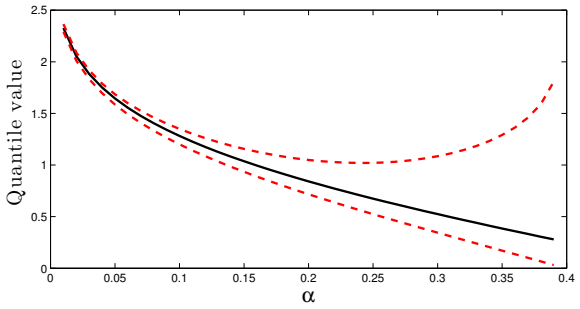

Fig. 1. Depiction of the normal quantile function $\Phi^{-1}(1-\alpha)$ (solid line) and the bounds (13) and (14) (dashed lines).

functions of the horizon T, the stochastic UCL algorithm presented here captures the effect of the horizon and other important features of human decision making in multi-armed bandit problems, as studied in more detail in [14].

IV. REGRET ANALYSIS

In this section we first consider UCL and bound its cumulative expected regret. We show the bound is logarithmic in horizon length T with proportionality constant within a constant factor of the best possible bound $2\sigma_r^2/\Delta_i^2$ (cf. (5)). We then consider the case of stochastic UCL where the cooling schedule follows (12) and show that the regret is again bounded by a logarithmic function of the horizon length T.

A. Deterministic decision making

Before analyzing the regret of our model in the case of deterministic decision making, we state the following bounds on the values of the normal quantile function $\Phi^{-1}(1-\alpha)$.

Theorem 1 (Bounds on the Gaussian inverse cdf): The following bounds hold when $\alpha < 1/\sqrt{2\pi}$ and $\beta \geq 1.02$:

$$\Phi^{-1}(1-\alpha) < \beta\sqrt{-\log(-(2\pi\alpha^2)\log(2\pi\alpha^2))} \tag{13}$$

$$\Phi^{-1}(1-\alpha) > \sqrt{-\log(2\pi\alpha^2(1-\log(2\pi\alpha^2)))}. \tag{14}$$

Fan [5] posed these bounds (without the factor β in (13)) as conjectures without proof. In fact, the factor β is necessary to get a correct upper bound, as we prove in the Appendix. See Figure 1 for a visual depiction of the bounds.

Turning to the regret analysis of the UCL algorithm, we consider the case of an uninformative prior, i.e., $\sigma_0^2 \to +\infty$. In the case of an uninformative prior and setting $K = \sqrt{2\pi e}$, the following performance bound holds with $\beta = 1.02$:

Theorem 2 (Regret for deterministic decision making): Let $\beta = 1.02$. The expected number of draws of any sub-optimal arm i is bounded by

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

Proof: In the spirit of [3], we bound n_i^T as follows:

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

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 arm i over i^* only if

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

This is true when at least one of the following 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 $C_i^t = \frac{\sigma_r}{\sqrt{\delta^2+n_i^t}}\Phi^{-1}(1-\alpha_t)$. 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 arm i^* is picked over arm 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_r^2/n_i^t, so, conditional on n_i^t, μ_i^t is a normal random variable distributed as

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

Equation (15) holds if

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

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 an uninformative prior,

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

Similarly, Equation (16) holds if

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

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,

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

Equation (17) holds if

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

where $\Delta_i = m_{i^*} - m_i$ and the inequality (18) follows from the bound (13). Therefore, for an uninformative prior, inequality (17) never holds if

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

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

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

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 desired bound

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

■

Thus, we have shown that in the case of deterministic decision making, the model achieves logarithmic regret uniformly in T with a constant which agrees with the best possible one (5) up to a constant factor. As the following section shows, the analysis extends to the case of stochastic decision making in a straightforward way.

B. Stochastic decision making

In the case where υ_t is defined by (12), a similar analysis holds. Again considering the case of an uninformative prior and setting the parameters $K = \sqrt{2\pi e}$ and $D = 1$, the following performance bound holds.

Theorem 3 (Regret for stochastic decision making): The expected number of draws of a suboptimal arm i satisfies

$$\begin{aligned} \mathbb{E}\left[n_i^T\right] \leq & \left(\frac{8\beta^2\sigma_r^2}{\Delta_i^2} + \frac{2}{\sqrt{2\pi e}}\right)\log T + \frac{\pi^2}{6} \\ & + \frac{4\beta^2\sigma_r^2}{\Delta_i^2}(1 - \log 2 - \log\log T) + 1 + \frac{2}{\sqrt{2\pi e}}. \end{aligned}$$

Proof: See Appendix. ■

Note that the bound on regret of the stochastic decision-making algorithm only differs from that of the deterministic decision-making algorithm by a constant equal to $\pi^2/6$. Therefore, by using the dynamic feedback rule (12) in the cooling schedule, the algorithm only pays a small performance penalty for the use of a stochastic maximization in the decision step. Human decision making is inherently stochastic. While it is unlikely humans are using this specific form of feedback rule, Theorem 3 shows that a stochastic decision rule can achieve near-optimal performance.

V. CONCLUSION

In conclusion, we propose the UCL algorithm for multi-armed Gaussian bandit problems, and we analyze its performance in terms of expected regret. We show that, using an uninformative prior, it achieves logarithmic regret. We extend the algorithm to incorporate stochastic policies using Boltzmann action selection and develop a feedback law to dynamically tune the temperature parameter of the selection rule such that the stochastic algorithm achieves logarithmic regret. As shown further in [14], with appropriate choices of parameter values, stochastic UCL is a model for human decision making in multi-armed bandit problems.

APPENDIX

Proof: [Proof of Theorem 1] Since the cdf for the standard normal random variable is a continuous and monotonically increasing function, it suffices to show that

$$\Phi(\beta\sqrt{-\log(-2\pi\alpha^2\log(2\pi\alpha^2))}) + \alpha - 1 \geq 0, \qquad (19)$$

for each $\alpha \in (0,1)$. Equation (19) can be equivalently written as $f(x) \geq 0$, where $x = 2\pi\alpha^2$ and f is defined by

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

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

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

Note that $\lim_{x\to 0^+} g(x) = +\infty$ and $\lim_{x\to 1^-} g(x) = 1$. Therefore, to establish that f 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

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

Letting $t = -\log(x)$, the above inequality is transformed to

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

which holds if

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

which is true if

$$\inf_{t\in[1,\infty)} -\frac{\log t}{t} \geq \max_{t\in[1,\infty)} \beta^2 t^{\beta^2-3}(1-2t+t^2)e^{-(\beta^2-1)t} - 1. \tag{20}$$

The 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

$$\max_{t\in[1,\infty)} \beta^2 t^{\beta^2-3}(1-2t+t^2)e^{-(\beta^2-1)t} - 1 \approx -0.3729$$

for the right hand side of (20), so (20) holds.

Therefore, $g(x)$ is non-negative for $x \in (0,1)$, $f(x)$ is a monotonically increasing function, and the theorem holds. ■

Proof: [Proof of Theorem 3] We begin by bounding $\mathbb{E}[n_i^T]$ as follows

$$\mathbb{E}\left[n_i^T\right] = \sum_{t=1}^{T} \mathbb{E}\left[P_{it}\right] \leq \eta + \sum_{t=1}^{T} \mathbb{E}\left[P_{it}\mathbf{1}\left(n_i^t \geq \eta\right)\right], \tag{21}$$

where η is a positive integer. Now, decompose $\mathbb{E}\left[P_{it}\right]$ as

$$\begin{aligned} \mathbb{E}[P_{it}] &= \mathbb{E}\left[P_{it}|Q_i^t \leq Q_{i^*}^t\right]\mathbb{P}(Q_i^t \leq Q_{i^*}^t) \\ &\quad + \mathbb{E}\left[P_{it}|Q_i^t > Q_{i^*}^t\right]\mathbb{P}(Q_i^t > Q_{i^*}^t) \\ &\leq \mathbb{E}\left[P_{it}|Q_i^t \leq Q_{i^*}^t\right] + \mathbb{P}(Q_i^t > Q_{i^*}^t). \end{aligned} \tag{22}$$

The probability P_{it} can itself be bounded as

$$P_{it} = \frac{\exp(Q_i^t/\upsilon_t)}{\sum_{j=1}^{N}\exp(Q_j^t/\upsilon_t)} \leq \frac{\exp(Q_i^t/\upsilon_t)}{\exp(Q_{i^*}^t/\upsilon_t)}. \tag{23}$$

Substituting the expression for the cooling schedule in inequality (23), we obtain

$$P_{it} \leq \exp\left(-\frac{2(Q_{i^*}^t - Q_i^t)}{\Delta Q_{\min}^t}\log t\right) = t^{-\frac{2(Q_{i^*}^t - Q_i^t)}{\Delta Q_{\min}^t}}. \tag{24}$$

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 $Q_{i^*}^t - Q_i^t$.

Once each arm has been picked once, the probability of ties between any pair of the Q_is is zero, i.e., $\Delta Q_{\min}^t = 0$ is zero. Consequently, it follows from inequality (22) 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 (24) that

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

where the last inequality follows from Theorem 2. ■

REFERENCES

[1] J. Ai and A.A. Abouzeid. Opportunistic spectrum access based on a constrained multi-armed bandit formulation. *J. of Communications and Networks*, 11(2):134–147, 2009.

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

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

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

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

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

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

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

[9] P. Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23(3):329–380, 1985.

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

[11] J. Le Ny, M. Dahleh, and E. Feron. Multi-agent task assignment in the bandit framework. In *IEEE Conf. on Decision and Control*, pages 5281–5286, 2006.

[12] K. Liu and Q. Zhao. Extended UCB policy for multi-armed bandit with light-tailed reward distributions. *arXiv:1112.1768*, 2011.

[13] 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.

[14] P. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision-making in generalized Gaussian multi-armed bandits. *Proc. IEEE*, 102(4):544–571, 2014.

[15] 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, 2011. Society for Neuroscience.