# Satisficing in Gaussian bandit problems

Paul Reverdy and Naomi E. Leonard

*Abstract*— **We propose a satisficing objective for the multi-armed bandit problem, i.e., where the objective is to achieve performance above a given threshold. We show that this new problem is equivalent to a standard multi-armed bandit problem with a maximizing objective and use this equivalence to find bounds on performance in terms of the satisficing objective. For the special case of Gaussian rewards we show that the satisficing problem is equivalent to a related standard multi-armed bandit problem again with Gaussian rewards. We apply the Upper Credible Limit (UCL) algorithm to this standard problem and show how it achieves optimal performance in terms of the satisficing objective.**

## I. Introduction

Engineering solutions to decision-making problems are often designed to maximize an objective function. However, in many contexts maximization of an objective function is an unreasonable goal, either because the objective itself is poorly defined or because solving the resulting optimization problem is intractable or costly. In these contexts, it is valuable to consider alternative decision-making frameworks.

Herbert Simon considered [16] alternative models of rational decision making with the goal of making them "compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist." A major feature of the models he considered is what he called "satisficing". In [16], Simon discussed in very broad terms a variety of simplifications to the classical economic concept of rationality, most importantly the idea that payoffs should be simple, defined by doing well relative to some threshold value. In [17], he introduced the word "satisficing" to refer to this thresholding concept and considered an ecological example of food foraging behavior in detail using mathematical terms. He also briefly discussed how satisficing relates to problems in inventory control and more complicated decision processes like playing chess.

Since Simon's pioneering work, satisficing has been studied in many fields such as psychology [15], economics [3], management science [10], [21], and ecology [20], [4]. In engineering, satisficing is of interest for the same reasons that motivated its introduction in the social science literature, specifically that it can simplify decision-making problems. Furthermore, many engineering problems are naturally posed using a satisficing objective, for example design problems that have to meet given specifications. A design that meets all the required specifications is acceptable, and the designers may be indifferent between any such design. In this context, optimization may be poorly defined, for example if there are several competing performance measures that trade off in complicated ways. Satisficing can be a simpler decision paradigm than maximizing, which requires additional information about preferences among possible tradeoffs.

Satisficing has been studied in the engineering literature in several contexts. In [11], Nakayama studied design optimization using a satisficing objective and found that it is effective in many practical fields. In [6], the authors studied control theory using a satisficing objective function, and in [22], the authors used satisficing to study optimal software design.

Satisficing can be implemented in a variety of ways. In this paper, we consider the stochastic multi-armed bandit problem [14], where a decision maker sequentially chooses one of a set of alternative options, or arms, and earns a reward drawn from a stationary probability distribution associated with that arm. The standard multi-armed bandit problem uses a maximizing objective, for which there is a known performance bound. We propose a satisficing objective for the multi-armed bandit problem based on the number of times the decision maker receives a reward that is above a threshold value and show that the multi-armed bandit problem with this objective is equivalent to a related standard multi-armed bandit problem. We use the equivalent problem to derive a performance bound for the new satisficing problem.

For Gaussian bandit problems, i.e., where the reward distributions are Gaussian with unknown mean and known variance, we show that solving the problem with the satisficing objective is equivalent to solving a standard Gaussian multi-armed bandit problem. We then apply the UCL algorithm we developed in previous work [13] to the standard problem, and show how this algorithm achieves optimal performance in terms of the original satisficing objective.

The remainder of the paper is structured as follows. In Section II we review the standard stochastic multi-armed bandit problem and the associated performance bounds. In Section III we propose the satisficing objective and bound performance in terms of this objective by defining a notion of satisficing regret. In Section IV we specialize to the case of Gaussian rewards and show that solving the satisficing problem is equivalent to solving a standard problem with Gaussian rewards. In Section V we review the UCL algorithm and show how applying it to the problem with Gaussian rewards achieves optimal performance in terms of the satisficing objective. Section VI shows the results of numerical simulations and Section VII concludes.

This research has been supported in part by ONR grants N00014-09-1-1074 and N00014-14-10635 and ARO grant W911NG-11-1-0385. P. Reverdy is supported through a NDSEG Fellowship. The authors are with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA {preverdy,naomi}@princeton.edu

## II. The stochastic multi-armed bandit problem

The stochastic multi-armed bandit problem is a decision-making problem in which the decision maker sequentially chooses one among a set of $N$ options, called *arms* in analogy with the lever of a slot machine. A single-levered slot machine is sometimes called a *one-armed bandit*, so the case of $N$ options is often called an $N$-armed bandit.

The decision-making agent collects reward $r_t \in \mathbb{R}$ by choosing arm $i_t$ at each time $t \in \{1, \cdots, T\}$, where $T \in \mathbb{N}$ is the horizon length for the sequential decision process. The reward from option $i \in \{1, \cdots, N\}$ is sampled from a stationary probability distribution $p_i$ and has an unknown mean $m_i \in \mathbb{R}$. The decision-maker's objective is to maximize some function of the sequence of rewards $\{r_t\}$.

### A. Maximization objective

In the standard multi-armed bandit problem, the agent's objective is to maximize the expected cumulative reward

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

Equivalently, by defining $m_{i^*} = \max_i m_i$ and $R_t = m_{i^*} - m_{i_t}$ as the *expected regret* at time $t$, the objective (1) can be formulated as minimizing the cumulative expected regret defined by

$$\sum_{t=1}^{T} R_t = T m_{i^*} - \sum_{i=1}^{N} m_i \mathbb{E}\left[n_i^T\right] = \sum_{i=1}^{N} \Delta_i \mathbb{E}\left[n_i^T\right], \tag{2}$$

where $n_i^T$ is the number of times arm $i$ has been chosen up to time $T$, $\Delta_i = m_{i^*} - m_i$ is the expected regret due to picking arm $i$ instead of arm $i^*$, and the expectation is over the possible rewards and decisions made by the agent.

The interpretation of (2) is that suboptimal arms $i \neq i^*$ should be chosen as rarely as possible. This is a non-trivial task since the mean rewards $m_i$ are initially unknown to the decision maker, who must try all arms to learn about their rewards while preferentially picking arms that appear more rewarding. The tension between these requirements is known as the *explore-exploit* tradeoff and is common to many problems in machine learning and adaptive control.

### B. Bound on optimal performance

Optimal performance in a bandit problem with the maximization objective (1) corresponds to picking suboptimal arms as rarely as possible, as shown by the last equality in (2). Lai and Robbins [9] studied the standard stochastic multi-armed bandit problem and showed that any policy solving the problem must pick each suboptimal arm $i \neq i^*$ a number of times that is at least logarithmic in the time horizon $T$, i.e.,

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

where $o(1) \to 0$ as $T \to +\infty$. The quantity $D(p_i||p_{i^*}) := \int p_i(r) \log \frac{p_i(r)}{p_{i^*}(r)} \mathrm{d}r$ is the Kullback-Leibler divergence between the reward density $p_i$ of a suboptimal arm $i$ 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.

The bound (3) is asymptotic in time, but a number of researchers (e.g., [2], [5], [13]) have constructed algorithms that achieve cumulative expected regret that is bounded by a logarithmic term uniformly in time, sometimes with the same constant as in (3). Cumulative expected regret that is uniformly bounded in time by a logarithmic term is often called *logarithmic regret* for short. In the literature, algorithms that achieve logarithmic regret with a leading term that is within a constant factor of that in (3) are considered to have optimal performance.

### C. Gaussian rewards

In this paper we focus on the case of Gaussian reward distributions, that is, the distribution $p_i$ of rewards associated with arm $i$ is Gaussian with unknown mean $m_i$ and known variance $\sigma_{s,i}^2$. In this case, the Kullback-Leibler divergence in (3) takes the value

$$D(p_i||p_{i^*}) = \frac{1}{2}\left(\frac{\Delta_i^2}{\sigma_{s,i^*}^2} + \frac{\sigma_{s,i}^2}{\sigma_{s,i^*}^2} - 1 - \log\frac{\sigma_{s,i}^2}{\sigma_{s,i^*}^2}\right). \tag{4}$$

This equation is more easily interpreted when the reward variances are uniform, i.e., $\sigma_{s,i}^2 = \sigma_s^2$ for each $i$. In this case, the divergence becomes

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

so the bound (3) is

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

This result can be interpreted as follows. For a given value of $\Delta_i$, a larger variance $\sigma_s^2$ makes the rewards more variable and therefore it is more difficult to distinguish between the arms. For a given value of $\sigma_s^2$, a larger value of $\Delta_i$ makes it easier to distinguish the optimal arm.

## III. The multi-armed bandit problem with satisficing objective

The standard multi-armed bandit problem is defined with the maximizing objective (1). We now propose a new satisficing objective for the multi-armed bandit problem and find bounds on optimal performance in terms of this new objective.

Consider an $N$-armed bandit problem. As before, the reward associated with each arm $i$ is drawn from a stationary probability distribution $p_i$, whose mean $m_i$ is unknown to the decision maker. At time $t \in \{1, \ldots, T\}$, the decision maker selects arm $i_t$ and receives a stochastic reward $r_t \in \mathbb{R}$.

The decision maker has a certain satisfaction level $M \in \mathbb{R}$, and is satisfied at time $t$ only if the reward $r_t$ is at least $M$. Let $s_t$ be the random variable denoting the decision maker's satisfaction at time $t$:

$$s_t = \begin{cases} 0, & r_t < M \\ 1, & r_t \geq M. \end{cases}$$

Then $s_t$ is a Bernoulli random variable with success probability $\pi_{i_t}$, where

$$\pi_i = \Pr[s_t = 1 | i_t = i] = \Pr[r_t \geq M | i_t = i] \quad (6)$$

is the probability of satisfaction upon picking arm $i$. We propose a satisficing objective in terms the number of times the satisfaction level is met.

**Definition 1** (Satisficing objective)**.** *The* satisficing objective *is to maximize the function*

$$\mathbb{E}\left[\sum_{t=1}^{T} s_t\right] = \sum_{t=1}^{T} \pi_{i_t}. \quad (7)$$

The satisficing objective differs from the maximization objective (1) in several important ways. First, it exhibits thresholding, that is, it is indifferent among rewards $r_t$ above the threshold value $M$. Second, it exhibits risk aversion, that is, it prefers smaller, consistent rewards (that will often be above the threshold) to larger, more variable ones (that may often be below it). Risk aversion is a characteristic often studied in economics and psychology [12], and is often incorporated in models of human decision making.

Since the satisficing objective consists of maximizing the number of times the agent is satisfied, it can be rewritten as follows. Let $\pi_{i^*} = \max_i \pi_i$ and define $\bar{\Delta}_i = \pi_{i^*} - \pi_i$ as the *expected satisficing regret* of selecting an arm $i$. We can rewrite (7) in terms of expected satisficing regret as

$$J_S = \mathbb{E}\left[\sum_{t=1}^{T} \bar{\Delta}_{i_t}\right] = \sum_{i=1}^{N} \bar{\Delta}_i \mathbb{E}\left[n_i^T\right], \quad (8)$$

where $n_i^T$ is the number of times arm $i$ has been chosen up to time $T$. This is a standard multi-armed bandit problem with Bernoulli rewards. Therefore the Lai-Robbins bound (3) holds, yielding a logarithmic lower bound on $\mathbb{E}\left[n_i^T\right]$ and cumulative expected satisficing regret:

**Corollary 1** (Satisficing regret bound)**.** *Any policy solving the multi-armed bandit problem with the satisficing objective* (8) *obeys*

$$\mathbb{E}\left[n_i^T\right] \geq \left(\frac{1}{D(\pi_i||\pi_{i^*})} + o(1)\right) \log T, \quad (9)$$

*for suboptimal arms $i \neq i^*$ where $D(\pi_i||\pi_{i^*}) = \pi_i \log\left(\frac{\pi_i}{\pi_{i^*}}\right) + (1-\pi_i)\log\left(\frac{1-\pi_i}{1-\pi_{i^*}}\right)$ is the Kullback-Leibler divergence between the two Bernoulli distributions with success probabilities $\pi_i$ and $\pi_{i^*}$.*

*Proof:* Apply the Lai-Robbins bound (3) to the standard multi-armed bandit problem with Bernoulli rewards.

We refer to cumulative expected satsificing regret that is uniformly bounded above in time by a logarithmic term as *logarithmic satisficing regret*. An algorithm that achieves logarithmic satisficing regret achieves *optimal satisficing performance*, i.e., optimal performance in terms of the satisfiicing objective (7).

The implication of writing the satisficing objective as the minimizing of cumulative regret is that if one can use the rewards $r_t$ to estimate the satisfaction probability $\pi_{i_t}$, one can use algorithms designed to solve the multi-armed bandit problem with a maximizing objective to solve the satisficing problem. In the next sections we study the Gaussian multi-armed bandit problem with a satisficing objective and show how to link rewards and probabilities in this case.

## IV. Satisficing with Gaussian rewards

In this section we study a Gaussian multi-armed bandit problem with the satisficing objective (8). By Gaussian multi-armed bandit problem, we mean that the reward $r_t$ due to selecting arm $i_t$ is $r_t \sim \mathcal{N}(m_{i_t}, \sigma^2_{s,i_t})$, where $\sigma^2_{s,i_t}$ is the known variance of arm $i_t$.

Define the quantity

$$x_i = \frac{m_i - M}{\sigma_{s,i}} \quad (10)$$

for each arm $i$. The following lemma states that the Gaussian multi-armed bandit problem with a satisficing objective is equivalent to a standard Gaussian multi-armed bandit problem with transformed reward distributions.

**Lemma 2** (Equivalence for Gaussian rewards)**.** *The Gaussian multi-armed bandit problem with satisficing objective is equivalent to a standard Gaussian multi-armed bandit problem with rewards $\tilde{r}_t \sim \mathcal{N}(x_{i_t}, 1)$ in the sense that the ordering of the arms in terms of $x_i$ is identical to the ordering in terms of $\pi_i$. In particular, the arm with maximal $x_i$ is the arm with maximal $\pi_i$*

*Proof:* With Gaussian rewards, the probability (6) of satisfaction from choosing arm $i$ is

$$\begin{aligned}\pi_i &= \Pr[m_i + \sigma_{s,i} z \geq M] \\ &= \Phi\left(\frac{m_i - M}{\sigma_{s,i}}\right) = \Phi(x_i),\end{aligned}$$

where $z \sim \mathcal{N}(0,1)$ is a standard normal random variable and $\Phi(z)$ is its cumulative distribution function. Let $i^* = \arg\max_i \pi_i$. The key insight is that $\Phi(\cdot)$ is a monotonically increasing function, which implies that the ordering of arms in terms of $\pi_i$ is identical to the ordering in terms of $x_i$. In particular, arm $i^*$ is the arm with maximal $x_i$. Therefore, the goal of an agent playing the satisficing bandit problem is to find the arm $i^*$ that maximizes $x_i$.

This is again a Gaussian bandit problem: consider the transformed reward

$$\tilde{r}_t = \frac{r_t - M}{\sigma_{s,i}},$$

which is a Gaussian random variable $\tilde{r}_t \sim \mathcal{N}(x_{i_t}, 1)$. The quantity $x_i$ plays the role of the mean reward $m_i$ from the original maximizing problem and the transformed rewards have uniform variance $\tilde{\sigma}^2_s = 1$. Solving this problem with a maximizing objective is equivalent to solving the original problem with the satisificing objective.

**Remark 3** (Location-scale families)**.** *The above analysis is easily generalized to reward distributions belonging to location-scale families. A location-scale family is a set of*

*probability distributions closed under affine transformations, i.e., if the random variable $X$ is in the family, so is the variable $Y = a + bX$, where $a, b \in \mathbb{R}$. Any random variable $X$ in such a family with mean $\mu$ and standard deviation $\sigma$ can be written as $X = \mu + \sigma Z$, where $Z$ is a zero-mean, unit-variance member of the family. Examples include the uniform or Student's $t$-distribution.*

## V. The UCL algorithm for Gaussian bandit problems

In this section we review the UCL algorithm, a Bayesian algorithm that we developed and analyzed in [13] to solve the standard Gaussian bandit problem. We then show that the UCL algorithm can be applied to the Gaussian satisficing problem of Section IV, achieving optimal performance. The algorithm maintains a belief about the mean rewards $\mathbf{m}$ by starting with a prior and updating it using Bayesian inference as new rewards are received. At each time $t$ the algorithm chooses arm $i_t$ using a heuristic which is a simple function of the current belief state. For uninformative priors, the UCL algorithm achieves logarithmic regret, i.e., optimal performance.

Uninformative priors correspond to having no information about the mean rewards. A major aspect of the UCL algorithm is its ability to incorporate information about the mean rewards through the use of a so-called "informative prior". In [13], we show that an appropriately chosen prior can significantly increase the performance of the UCL algorithm. Several different UCL algorithms are developed in [13]; here we cover only the deterministic UCL algorithm, which we refer to as the UCL algorithm for brevity.

### A. Prior

The prior distribution captures the agent's knowledge about the vector of mean rewards $\mathbf{m}$ before beginning the task. We assume that the prior distribution is multivariate Gaussian with mean $\boldsymbol{\mu}_0 \in \mathbb{R}^N$ and covariance $\Sigma_0 \in \mathbb{R}^{N \times N}$:

$$\mathbf{m} \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0). \tag{11}$$

The $i^{th}$ element of $\boldsymbol{\mu}_0$, denoted by $\mu_i^0$, represents the agent's mean belief of the reward $m_i$ associated with arm $i$. The $(i,i)$ element of $\Sigma_0$, denoted by $\left(\sigma_{ii}^0\right)^2$, represents the agent's uncertainty associated with that belief. Off-diagonal elements of $\Sigma_0$, e.g., $\sigma_{ij}^0$, represent the agent's perceived relationship between $m_i$ and $m_j$: if $\sigma_{ij}^0$ is positive, high values of $m_i$ are correlated with high values of $m_j$, while if it is negative, high values of $m_i$ correlate with low values of $m_j$. Any positive-definite matrix can be used as $\Sigma_0$, but several specific ones are of interest. An uninformative prior corresponds to a complete lack of certainty, i.e., $\left(\sigma_{ii}^0\right)^2 \to +\infty$, so one sets each element $\sigma_{ij}^0$ equal to $+\infty$.

### B. Inference update

At each time $t$ the agent picks an arm $i_t$ and receives a reward $r_t$ that is Gaussian distributed: $r_t \sim \mathcal{N}(m_{i_t}, \sigma_{s,i_t}^2)$. Bayesian inference provides an optimal solution to the problem of updating the belief state $(\boldsymbol{\mu}_t, \Sigma_t)$ to incorporate this new information. Given the Gaussian prior (11), the Bayesian update equations are linear [8]:

$$\begin{aligned} \mathbf{q} &= \frac{r_t \boldsymbol{\phi}_t}{\sigma_{s,i_t}^2} + \Lambda_{t-1}\boldsymbol{\mu}_{t-1} \\ \Lambda_t &= \frac{\boldsymbol{\phi}_t \boldsymbol{\phi}_t^T}{\sigma_{s,i_t}^2} + \Lambda_{t-1},\ \Sigma_t = \Lambda_t^{-1} \\ \boldsymbol{\mu}_t &= \Sigma_t \mathbf{q}. \end{aligned} \tag{12}$$

### C. Decision heuristic

At each time $t$ the UCL algorithm computes a value $Q_i^t$ for each arm $i$. The UCL algorithm picks the arm $i_t$ that maximizes $Q_i^t$. That is, it picks

$$i_t = \arg\max_i Q_i^t. \tag{13}$$

The heuristic value $Q_i^t$ is

$$Q_i^t = \mu_i^t + \sigma_i^t \Phi^{-1}(1 - \alpha_t), \tag{14}$$

where $\mu_i^t = (\boldsymbol{\mu}_t)_i$, $\left(\sigma_i^t\right)^2 = (\Sigma_t)_{ii}$, $\alpha_t = 1/Kt$, and $K > 0$ is a tunable parameter. The heuristic $Q_i^t$ is a Bayesian upper limit for the value of $m_i$ based on the information available at time $t$. It represents an optimistic assessment of the value of $m_i$. The decision made can be thought of as the most optimistic one consistent with the current information.

### D. Performance

In [13], we study the case of homogeneous sampling noise (i.e., $\sigma_{s,i}^2 = \sigma_s^2$ for each $i$) and show that the UCL algorithm achieves logarithmic cumulative expected regret uniformly in time. In particular, we prove that the following theorem holds for any $\beta \geq 1.02$.

**Theorem 4** (Regret of the deterministic UCL algorithm [13])**.** *The following statements hold for the Gaussian multi-armed bandit problem and the deterministic UCL algorithm with uncorrelated uninformative prior and $K = \sqrt{2\pi e}$:*

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

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

2) *the cumulative expected regret until time $T$ satisfies*

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

The implication of this theorem can be seen by comparing statement 1) with the Lai-Robbins bound (5): it shows that the UCL algorithm achieves logarithmic regret uniformly in time with a constant that differs from the optimal asymptotic one by a constant factor of $4\beta^2$, and therefore is considered to have optimal performance.

### E. Application to satisficing objective

In Section IV, we showed that solving the Gaussian multi-armed bandit problem with a satisficing objective is equivalent to a transformed standard Gaussian multi-armed bandit problem with maximizing objective. Therefore, we can apply the UCL algorithm to the satisficing problem. A prior belief $\mathbf{m} \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$ is transformed into prior beliefs on $\mathbf{x}$ by

$$\mathbf{x} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_0, \tilde{\Sigma}_0),$$

where $(\tilde{\boldsymbol{\mu}}_0)_i = ((\boldsymbol{\mu}_0)_i - M)/\sigma_{s,i}$ and $(\tilde{\Sigma}_0)_{ij} = (\Sigma_0)_{ij}/(\sigma_{s,i}\sigma_{s,j})$. Define $x_{i^*} = \max_i x_i$ and $\tilde{\Delta}_i = x_{i^*} - x_i$.

We refer to the UCL algorithm using the transformed reward $\tilde{r}_t$ and prior as the *satisficing UCL algorithm*. The satisficing UCL algorithm achieves logarithmic satisficing regret, as formalized in the following theorem.

**Theorem 5** (Regret of the satisficing UCL algorithm)**.** *The following statements hold for the Gaussian multi-armed bandit problem with a satisficing objective and the satisficing UCL algorithm with uncorrelated uninformative prior and* $K = \sqrt{2\pi e}$*:*

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

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

2) *the cumulative expected satisficing regret until time* $T$ *satisfies*

$$\sum_{t=1}^{T} R_t \leq \sum_{i=1}^{N} \tilde{\Delta}_i \left( \left(\frac{8\beta^2}{\tilde{\Delta}_i^2} + \frac{2}{\sqrt{2\pi e}}\right)\log T + \frac{4\beta^2}{\tilde{\Delta}_i^2}(1 - \log 2 - \log\log T) + 1 + \frac{2}{\sqrt{2\pi e}} \right). \tag{15}$$

*Proof:* Apply Theorem 4 to the Gaussian multi-armed bandit problem with mean rewards $\mathbf{x}$ and reward distributions $\tilde{r}_t \sim \mathcal{N}(x_{i_t}, 1)$ defined in Lemma 2.

The satisficing regret is upper bounded by a logarithmic function of $T$. Therefore, the satisficing UCL algorithm achieves optimal satisficing regret up to a constant factor.

## VI. Numerical example

In this section, we present the results of two numerical simulations of the satisficing UCL algorithm solving a multi-armed bandit problem with Gaussian rewards and the satisficing objective. The first simulation demonstrates the performance guarantees and allows us to compare the optimal regret bound (9) and the bound (15) obeyed by the satisficing UCL algorithm. The second simulation demonstrates the risk-averse nature of the satisficing objective.

For the simulations presented in Figure 1, we set $N = 4$. The satisfaction level $M$ was set equal to 2, the mean rewards $\mathbf{m}$ were equal to $[1\ 2\ 3\ 4]$ and the standard deviations equal to $[1\ 1\ 1\ 3]$, so $\mathbf{x} = [-1\ 0\ 1\ \frac{2}{3}]$ and $i^* = 3$ was the optimal arm. The algorithm used an uninformative prior. These values were chosen such that the arm with maximal mean reward was not the optimal arm, so satisficing induces different behavior than maximizing.

Figure 1 plots the mean cumulative satisficing regret incurred by the satisficing UCL algorithm over 100 simulations along with the two regret bounds (9) and (15). The mean regret obeys the performance bound (15) from Theorem 5 and is actually below the asymptotic lower bound (9) at initial times. This apparent violation of the bound is due to the fact that at initial times the system is not yet in the asymptotic regime where the bound applies.

For the simulations presented in Figure 2, we set $N = 2$. The mean rewards $\mathbf{m}$ were equal to $[12.2\ 12.1]$ and the standard deviations equal to $[10\ 1]$, so $\mathbf{x} = [0.02\ 0.1]$. This meant $i = 1$ was the optimal arm for the maximizing objective while $i = 2$ was the optimal arm for the satisficing objective. The algorithm used an uninformative prior. The problem was simulated 100 times with each objective.

Figure 2 demonstrates the risk aversion inherent in the satisficing objective by comparing the results of the same problem solved with the satisficing and the maximizing objectives. The satisfaction level $M$ was set equal to 12. We considered cumulative surplus (rewards in excess of the satisfaction level) for both objectives. Negative values of the surplus represent deficits, which are to be avoided. Results from the maximizing objective are presented in black. The solid line shows mean cumulative surplus and the shaded region shows the 95% confidence interval around that mean. Results from the satisficing objective are presented in blue. The solid line shows the mean cumulative surplus, and the dashed lines show the 95% confidence interval. The lower limit of the confidence intervals measures worst-case performance. The measure for the satisficing objective is consistently above the one for the maximizing objective, so satisficing results in better worst-case performance.

## VII. Conclusion

Satisficing, the concept of doing well relative to a reference value, is a useful alternative to maximizing that can be applied to a variety of decision-making scenarios. Considering satisficing objectives instead of maximizing ones can simplify decision-making problems and can result in policies that are more robust in the sense that they are risk-averse.

In this paper, we considered the multi-armed bandit problem using a satisficing objective by proposing a new notion of satisficing regret. We showed that there is an equivalence between minimizing satisficing regret and minimizing the standard notion of regret. Using this equivalence, we derived a logarithmic lower bound on satisficing regret and, in the case of Gaussian rewards, adapted the UCL algorithm [13] to achieve optimal satisficing performance.

This work opens the door to many future extensions. The satisficing objective with Gaussian rewards bears a strong resemblance to the CreditMetrics two-state credit risk model used in quantitative finance [7]. This could allow the credit

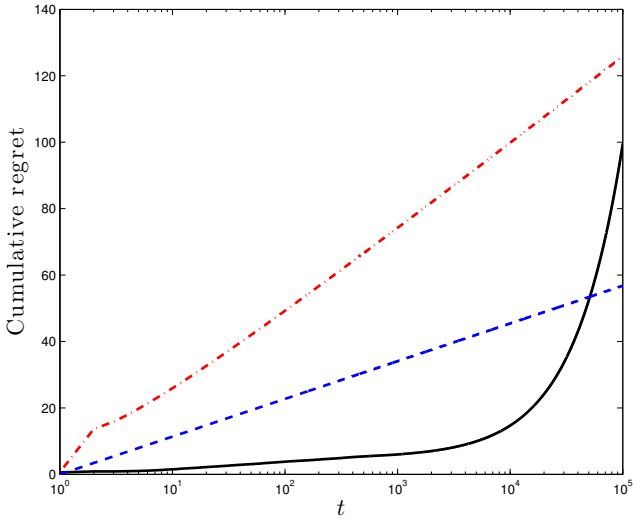


Fig. 1. Regret incurred by the satisficing UCL algorithm while solving a satisficing Gaussian multi-armed bandit problem, along with two theoretical bounds, plotted against time $t$ on a logarithmic scale. The solid black line shows mean cumulative expected regret from 100 simulations. The dashed line shows the asymptotic bound on regret (9), which appears as a straight line due to the scaling of the axes. The dash-dotted line shows the regret bound (15), which provides guarantees on the algorithm's performance.

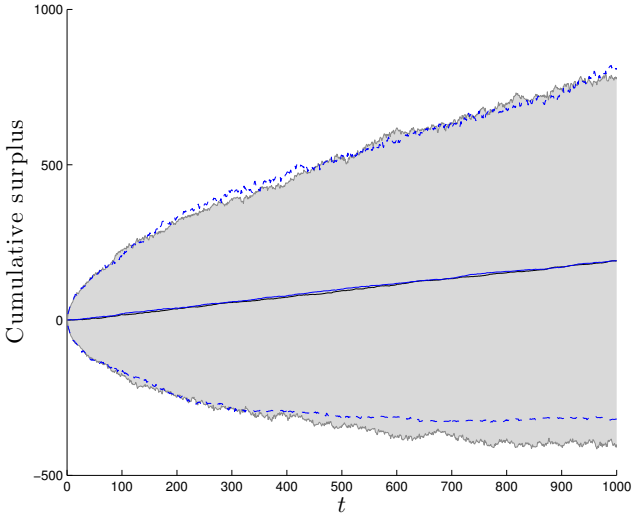


Fig. 2. Cumulative surplus earned by the UCL algorithm while solving a Gaussian multi-armed bandit problem, once with a satisficing (blue curves) and again with a maximizing objective (black curve and shaded region). Both objectives achieve similar mean performance (solid curves) but using the satisficing objective results in better worst-case performance. The shaded region (satisficing) and the blue dashed lines (maximizing) show the 95% confidence interval around the mean cumulative surplus. The lower limit of the confidence intervals measures worst-case performance. The lower limit for the satisficing objective is consistently above the one for the maximizing objective, so satisficing results in better worst-case performance.

investment portfolio problem studied in finance to be posed as a multi-armed bandit problem with satisficing objective.

The risk averse nature of satisficing objectives such as the one proposed in this paper will result in more robust policies for solving the multi-armed bandit problem in cases with reward variance $\sigma_s^2$ is heterogeneous across arms. Risk aversion and robustness are important for engineering applications (where standard bandit algorithms are known to have poor risk-aversion characteristics [1]) but also in the field of optimal foraging theory [4]. The multi-armed bandit framework has been used to study foraging [18] using a maximizing objective, but a satisficing objective is more ecologically plausible.

We developed a policy for the satisficing problem with Gaussian rewards, but development of optimal policies for the satisficing problem with other reward distributions remains an open problem. For all satisficing problems, picking the appropriate satisfaction level is a non-trivial problem in its own right, analogous to picking the error rates in the Sequential Probability Ratio Test [19].

## Acknowledgement

We thank Vaibhav Srivastava and Simon A. Levin for helpful discussions.

## References

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

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

[3] R. Bordley and M. LiCalzi. Decision analysis using targets instead of utility functions. *Decisions in Economics and Finance*, 23(1):53–74, 2000.

[4] Y. Carmel and Y. Ben-Haim. Info-gap robust-satisficing model of foraging behavior: Do foragers optimize or satisfice? *The Amer. Naturalist*, 166(5):633–641, 2005.

[5] 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.

[6] M. Goodrich, W. Stirling, and R. Frost. A theory of satisficing decisions and control. *IEEE Trans. Syst., Man and Cybern. A: Syst. Humans*, 28(6):763–779, Nov 1998.

[7] M. B. Gordy. A comparative anatomy of credit risk models. *J. of Banking & Finance*, 24(1):119–149, 2000.

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

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

[10] T. M. Moe. The new economics of organization. *Amer. J. of Political Sci.*, 28(4):739–777, 1984.

[11] H. Nakayama and Y. Sawaragi. Satisficing trade-off method for multiobjective programming. In *Interactive Decision Analysis*, pages 113–122. Springer, 1984.

[12] J. W. Pratt. Risk aversion in the small and in the large. *Econometrica: J. of the Econometric Soc.*, pages 122–136, 1964.

[13] 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.

[14] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the Amer. Math. Soc.*, 58:527–535, 1952.

[15] B. Schwartz, A. Ward, J. Monterosso, S. Lyubomirsky, K. White, and D. R. Lehman. Maximizing versus satisficing: happiness is a matter of choice. *J. Personality and Social Psychology*, 83(5):1178, 2002.

[16] H. A. Simon. A behavioral model of rational choice. *The Quarterly J. of Econ.*, 69(1):99–118, 1955.

[17] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129, 1956.

[18] V. Srivastava, P. Reverdy, and N. E. Leonard. On optimal foraging and multi-armed bandits. In *Proc. of the 51st Annu. Allerton Conf. on Commun., Control, and Computing*, pages 494–499, 2013.

[19] A. Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[20] D. Ward. The role of satisficing in foraging theory. *Oikos*, pages 312–317, 1992.

[21] S. G. Winter. The satisficing principle in capability learning. *Strategic Management Journal*, 21(10-11):981–996, 2000.

[22] B. Yin et al. Finding optimal solution for satisficing non-functional requirements via 0-1 programming. In *Proc. 2013 IEEE 37th Annu. Computer Software and Applications Conf.*, pages 415–424, 2013.