Gaussian multi-armed bandit problems with multiple objectives

Paul Reverdy

***Abstract*— Motivated by the goal of formally integrating human designers into computational systems for engineering design optimization, I study decision making under uncertainty with multiple objectives in the context of the multi-armed bandit problem. A key aspect of multi-objective optimization is the need for *scalarization*, i.e., a way to combine the various objectives into a single well-defined scalar objective function. I study the case where the multi-objective rewards are Gaussian distributed and the scalarization is linear and develop an algorithm that achieves optimal performance, i.e., converges to selecting the best arm at the highest possible rate.**

I. Introduction

Decision making is a central activity in engineering, ranging from the deliberative process of engineering design to the real-time decisions taken by control systems. Two key aspects of decision making that often arise in practice are the presence of uncertainty and the need to balance multiple objectives. In this paper, motivated by the Multidisciplinary Design Optimization (MDO) literature (e.g., [1]) on engineering design, I develop a formal model of multi-objective decision making under uncertainty based on the multi-armed bandit problem.

The multi-armed bandit problem [2] is a model of sequential decision making under uncertainty that has been the subject of a large and growing literature. The problem is simple enough to admit a bound on optimal performance [3] that allows one to have an objective benchmark on the performance of algorithms designed to solve it, and there has been associated development of algorithms, e.g., [3], [4], [5] that achieve optimal performance.

Furthermore, the multi-armed bandit problem has been used to develop models of human decision-making behavior under uncertainty [5]. Previous work has been done to develop methods for integrating human decision makers into the MDO decision-making process [6], and the present work provides a way to link this work with the rigorous multi-armed bandit approach. Our modeling work is also of interest to the broader community because it is relevant to the development of adaptive control systems that seek to maximize multiple objectives under uncertainty.

The remainder of the paper is structured as follows. In Section II I review the formulation of the multi-armed bandit problem for a single objective and define its extension to the multi-objective case. I then show how the multi- and single-objective cases are closely linked when the rewards follow a Gaussian distribution and the decision maker combines the vector-valued rewards using a linear function; I focus on this case in the rest of the paper. In Section III I review existing results on the performance of multi-armed bandit algorithms, particularly the Upper Credible Limit (UCL) algorithm from [5], and show that the results hold for the Gaussian–linear value function case. In Section IV I show how to extend the UCL algorithm to the MO-UCL algorithm for the multi-objective case, and in Section V show that the MO-UCL algorithm achieves optimal performance. In Section VI I show the results of a numerical simulation that demonstrates this optimal performance; I conclude in Section VII.

Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA preverdy@seas.upenn.edu

II. Problem formulation

The stochastic *multi-armed bandit problem* is a sequential decision-making task in which the decision-making agent sequentially chooses one among a set of N options, termed *arms* in analogy with the lever of a slot machine. As a slot machine is sometimes called a *one-armed bandit*, the problem with $N \geq 2$ arms is called a *multi-armed bandit*.

The decision-making agent performs the task by sequentially selecting an arm $i_t \in \{1, \dots, N\}$ for each of the decision times $t \in \{1, \dots, T\}$, where $T \in \mathbb{N}$ is the horizon of the task. The horizon T may in general tend towards infinity, but I consider only the finite-horizon case here. Upon selecting the arm i_t, the agent accrues a reward $\mathbf{r}_t \in \mathbb{R}^{n_o}$, which is sampled from the probability distribution ν_{i_t} associated with the arm i_t. Each distribution ν_i is stationary and has mean $\mathbf{m}_i \in \mathbb{R}^{n_o}$, which is constant but unknown to the agent. There are two distinct cases of the multi-armed bandit problem depending on whether the reward distribution is scalar- ($n_o = 1$) or vector-valued ($n_o > 1$).

A. Single objective

The standard multi-armed bandit problem introduced by Robbins [2] is the case of scalar ($n_o = 1$) rewards, i.e., a single objective. In the *single-objective multi-armed bandit problem*, the decision-making agent's objective for the task is to maximize the expected value of its cumulative rewards:

$$\max_{\{i_t\}_{t=1}^T} J, \; J = \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t\right] = \sum_{i=1}^{N} \mathbf{m}_i \mathbb{E}\left[n_i^T\right], \qquad (1)$$

where n_i^T is the number of times option i has been selected up to time T. The objective involves expectations since the observed rewards, and consequently the agent's choices made in response to them, are stochastic; the expectations are taken over the distribution of arm choices and reward outcomes. The task objective (1) is a well-defined optimization problem with a scalar objective function. The agent has to solve it by sequentially selecting arms i_t based on the information

available to it at time t. The core difficulty in solving problem (1) is that the agent does not know the mean rewards $\mathbf{m}_i$ and must pick arms often enough to learn their reward values while preferentially picking arms with high rewards. This tension between seeking information and seeking immediate reward is known as the *explore-exploit* tradeoff, and arises throughout the fields of machine learning and adaptive control.

B. Multiple objectives

I now consider the more general case of $n_o \geq 2$ objectives, i.e., vector rewards. Other authors, e.g., [7], [8], have also considered similar generalizations. The *multi-objective multi-armed bandit problem* is a multi-armed bandit problem where the rewards are vector-valued, so there is a vector of unknown means $\mathbf{m}_i \in \mathbb{R}^{n_o}$ for each arm i. The task objective (1) for $n_o = 1$ naturally generalizes to the multi-objective case as

$$\max_{\{i_t\}_{t=1}^T} \mathbf{J},\ \mathbf{J} = \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{r}_t\right] = \sum_{i=1}^{N} \mathbf{m}_i \mathbb{E}\left[n_i^T\right]. \tag{2}$$

In this case, the objective function $\mathbf{J}$ is vector-valued, so the optimization problem (2) is not well defined as stated.

There exist a number of methods, called *scalarizations*, to transform the ill-defined vector-valued optimization problem (2) into a well-defined scalar optimization problem [9], [10]. In this paper I consider a method called the *linear scalarization* [7], in which the scalar optimization objective function is defined as a linear combination of the individual components of the vector objective function.

Under the linear scalarization, each objective j is given a weight $w_j \geq 0$. The weights are normalized such that $\sum_{j=1}^{n_o} w_j = 1$, and I denote the vector of weights by $\mathbf{w}$. Selecting weights for a specific application is an important problem in its own right, analogous to selecting weighting matrices for the LQR problem. The scalar reward associated with a vector reward $\mathbf{r}_t \in \mathbb{R}^{n_o}$ is $\bar{r}_t = \mathbf{w}^\top \mathbf{r}_t \in \mathbb{R}$, which is a weighted sum of the individual objective rewards. For a given arm i, the unknown mean of the scalar reward associated with arm i is $\bar{m}_i = \mathbf{w}^\top \mathbf{m}_i$ due to the linearity of the expectations operator.

I define the vector $\mathbf{M} \in \mathbb{R}^{N \cdot n_o}$ of unknown vector means as

$$\mathbf{M} = \begin{bmatrix} \mathbf{m}_1^\top & \mathbf{m}_2^\top & \dots & \mathbf{m}_N^\top \end{bmatrix}^\top.$$

I define $\bar{\mathbf{m}} \in \mathbb{R}^N$ as the vector of unknown scalar means, whose i^{th} component is equal to $\bar{m}_i$. The two vectors can be related succinctly by

$$\bar{\mathbf{m}} = (I_N \otimes \mathbf{w}^\top)\mathbf{M},$$

where I_N is the identity matrix of size N and $\otimes$ is the Kronecker matrix product.

C. Linear scalarization with Gaussian rewards

In this paper I focus on the *Gaussian multi-armed bandit problem*, i.e., the case where the reward distributions are Gaussian. In this case, selecting an arm i_t at time t results in a reward with mean $\mathbf{m}_{i_t}$ and variance Σ_{s,i_t}:

$$\mathbf{r}_t \sim \mathcal{N}(\mathbf{m}_{i_t}, \Sigma_{s,i_t}).$$

I assume that the sampling variances $\Sigma_{s,i}$ are known to the agent, e.g., from prior experience or known sensor characteristics.

When the vector rewards are Gaussian distributed, the linear scalarization produces scalar rewards that are also Gaussian distributed. Specifically, selecting arm i_t at time t results in a reward

$$\bar{r}_t \sim \mathcal{N}(\bar{m}_{i_t}, \bar{\sigma}^2_{s,i_t}), \tag{3}$$

where $\bar{m}_i = \mathbf{w}^\top \mathbf{m}_i$ and $\bar{\sigma}^2_{s,i} = \mathbf{w}^\top \Sigma_{s,i} \mathbf{w}$ is the variance of the scalarized reward. Therefore, the linear scalarization reduces the multi-objective Gaussian multi-armed bandit problem to a single-objective Gaussian multi-armed bandit problem. This reduction allows us to apply algorithms developed for the single-objective Gaussian multi-armed bandit problem to the multi-objective problem.

The linear scalarization maintains the Gaussian nature of the rewards because a linear combination of Gaussian random variables is itself a Gaussian random variable. This property of closure under linear combination characterizes the general class of *stable* probability distributions. Precisely, let X_1 and X_2 be independent samples drawn from a probability distribution f and $a, b \in \mathbb{R}$. Let X be an independent sample from f: if there exist $c > 0, d \in \mathbb{R}$ such that $aX_1 + bX_2$ follows the same distribution as $cX + d$, then the distribution f is said to be *stable*. Other examples of stable distributions include the Cauchy and Lévy distributions [11].

III. Performance

In this section, I study the performance of algorithms designed to solve the multi-armed bandit problem. Having reduced the multi-objective problem (2) to a single-objective problem (1) by scalarization, I focus on the scalar, single objective, case. I review a result from the literature that shows the fundamental limits to performance of any algorithm solving the single-objective problem (1) and use it to derive a limit to the performance of any algorithm solving the Gaussian multi-objective problem with linear scalarization, where rewards follow (3). Finally, I review an algorithm from the literature that achieves optimal performance in the single-objective Gaussian multi-armed bandit problem.

A. Regret

An agent solving the single-objective multi-armed bandit problem (1) faces difficulty due to its lack of information about the mean rewards $\mathbf{m}_i$ associated with each arm i. I denote the optimal arm by $i^* = \arg\max_i \mathbf{m}_i$ and assume for convenience that it is unique. If the agent had full information, it would repeatedly select the optimal arm, i.e., set $i_t = i^*$ for each decision time $t \in \{1, \dots, T\}$. Since the agent does not have full information, it has to sample the arms to gain information about them while preferentially

selecting the arm that appears to have the highest rewards based on currently-available information.

The most common method used in the literature to characterize the performance of algorithms that solve the multi-armed bandit problem is to compare their rewards against those of the full-information optimal policy. To do so, one defines the *cumulative expected regret* as

$$J_R = T\mathbf{m}_{i^*} - J = \sum_{i=1}^{N}(\mathbf{m}_{i^*} - \mathbf{m}_i)\mathbb{E}\big[n_i^T\big] = \sum_{i=1}^{N}\Delta_i\mathbb{E}\big[n_i^T\big]\,, \tag{4}$$

where $\Delta_i = \mathbf{m}_{i^*} - \mathbf{m}_i \geq 0$ is the expected *regret* due to selecting arm i. Minimizing the cumulative expected regret J_R is equivalent to maximizing the cumulative expected reward J. By definition, the cumulative expected regret is non-negative, and depends on the number of times a suboptimal arm $i \neq i^*$ is selected.

B. Bound on optimal performance

The decision-making agent's regret is monotonically increasing in the number of times the agent selects a suboptimal arm. This number must be positive due to the agent's incomplete information, so a question of interest is how often suboptimal arms need to be selected. Lai and Robbins [3] studied this question and showed that suboptimal arms must be selected at a rate that is at least logarithmic with the horizon T. In particular, they showed that

$$\mathbb{E}\big[n_i^T\big] \geq \left(\frac{1}{D(\nu_i||\nu_{i^*})} + o(1)\right)\log T \tag{5}$$

holds for each suboptimal arm i, where $D(\nu_i||\nu_{i^*}) = \int \nu_i(r)\log\frac{\nu_i(r)}{\nu_{i^*}(r)}\mathrm{d}r$ is the Kullback-Leibler divergence between the reward distribution of arm i and that of the optimal arm i^* and $o(1)$ represents terms that obey $o(1) \to 0$ as $T \to +\infty$.

The bound (5) only holds asymptotically in time due to the $o(1)$ term. In the literature, an algorithm is considered to exhibit optimal performance if its cumulative expected regret is upper bounded by a logarithmic function of T with a leading constant that is within a constant factor of the optimal one defined in (5). The literature refers to such an algorithm as achieving *logarithmic regret*. Equation (5) can be thought of as a bound on the rate of convergence of an algorithm that solves the multi-armed bandit problem. Convergence for such an algorithm means consistently selecting the optimal arm i^* and not selecting suboptimal arms $i \neq i^*$. The bound shows that this convergence can only occur at a logarithmic rate.

C. Bound for Gaussian rewards

In the case of scalar Gaussian rewards, i.e., when the reward distributions ν_i are Gaussian with mean m_i and variance $\sigma_{s,i}^2$, the Kullback-Leibler divergence $D(\nu_i||\nu_{i^*})$ is

$$D(\nu_i||\nu_{i^*}) = \frac{1}{2}\left(\frac{\sigma_{s,i}^2}{\sigma_{s,i^*}^2} + \frac{\Delta_i^2}{\sigma_{s,i^*}^2} - 1 + \log\left(\frac{\sigma_{s,i^*}^2}{\sigma_{s,i}^2}\right)\right). \tag{6}$$

If the sampling variances are uniform, so $\sigma_{s,i}^2 = \sigma_s^2$ for each arm i, the expression simplifies to $D(\nu_i||\nu_{i^*}) = \Delta_i^2/(2\sigma_s^2)$.

D. Optimal performance with multiple objectives

In the multi-objective case, I define cumulative expected scalar regret (4) in terms of the scalarized rewards $\bar{m}_i$. In this context, let i^* be the arm with the maximal scalarized reward $\bar{m}_i$ and define the expected scalar regret associated with arm i as $\bar{\Delta}_i = \bar{m}_{i^*} - \bar{m}_i$. Because the linear scalarization maintains the Gaussian reward structure, the bound (5) holds, with the Kullback-Leibler divergence (6) being computed in terms of $\bar{\Delta}_i$ and $\sigma_{s,i}^2$.

E. An algorithm with optimal performance

Reverdy *et al.* [12], [5] studied the single-objective Gaussian multi-armed bandit problem from a Bayesian perspective and developed a family of Upper Credible Limit, or UCL, algorithms to solve it. The UCL algorithms define a decision heuristic function Q_i^t for each arm i at each time t and the deterministic UCL algorithm, which I focus on in the following, picks arm $i_t = \arg\max_i Q_i^t$. The Bayesian perspective allows the algorithms to incorporate prior knowledge that may be available about the rewards; if the prior knowledge is accurate, it can result in improved performance. Alternatively, the algorithms can use an uninformative prior that provides no prior knowledge. In this case, the UCL algorithms achieve logarithmic regret, as summarized by the theorem below for the deterministic UCL algorithm.

I define $\{R_t^{\mathrm{UCL}}\}_{t\in\{1,\ldots,T\}}$ as the sequence of expected regret for the deterministic UCL algorithm. The UCL algorithm achieves logarithmic regret uniformly in time as formalized by the following theorem, previously published as Theorem 1 of [12], which is itself a slightly modified version of Theorem 2 of [5].

Theorem 1 (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:

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

$$\mathbb{E}\big[n_i^T\big] \leq \Big(\frac{8\sigma_s^2}{\Delta_i^2} + 2\Big)\log T + 3\,;$$

2) the cumulative expected regret until time T satisfies

$$\sum_{t=1}^{T} R_t^{\mathrm{UCL}} \leq \sum_{i=1}^{N}\Delta_i\left(\Big(\frac{8\sigma_s^2}{\Delta_i^2} + 2\Big)\log T + 3\right).$$

IV. UCL FOR MULTIPLE OBJECTIVES

Because Gaussian distributions are stable distributions, applying the linear weighting scalarization transforms a Gaussian multiple-objective bandit problem to a single-objective bandit problem again with Gaussian rewards. While the Gaussian assumption is a strong one, the analysis in this section readily generalizes to other stable distributions. Furthermore, the location-scale structure of the Gaussian family facilitates encoding information about the correlation among rewards in the prior covariance matrix.

A. Priors

As shown in [5], prior information can greatly enhance performance by allowing an algorithm to quickly identify correlated groups of arms with low rewards. Therefore, a key element of generalizing the UCL algorithm to the case of multiple objectives is the design of priors. The generalization is non-trivial, because introducing multiple objectives to the multi-armed bandit problem means there are two sources of correlation: across-arm and within-arm correlations. I construct the prior by first considering within-arm correlations, then modeling the across-arm correlations.

For an individual arm i with unknown mean rewards $\mathbf{m}_i \in \mathbb{R}^{n_o}$, let the prior on $\mathbf{m}_i$ be multivariate Gaussian with mean $\boldsymbol{\mu}_i^0$ and covariance Σ_i^0:

$$\mathbf{m}_i \sim \mathcal{N}(\boldsymbol{\mu}_i^0, \Sigma_i^0). \tag{7}$$

The vector $\boldsymbol{\mu}_i^0 \in \mathbb{R}^{n_o}$ represents the mean belief about $\mathbf{m}_i$ and the positive-definite matrix $\Sigma_i^0 \in \mathbb{R}^{n_o \times n_o}$ represents correlations between the rewards associated with different objectives for the arm i. In a design application, these correlations might result, e.g., from the nature of the analysis tool that generates the rewards. For example, an arm might represent a particular design whose performance is analyzed through simulation. Repeated simulations will tend to report varying values of the performance metrics, and the variation will generally have some correlation structure.

Across-arm correlations model how rewards associated with one arm are related to those associated with another arm. For example, in previous work Reverdy *et al.* [5], [13] and others [14], [15] considered a spatially-embedded variant of the multi-armed bandit problem where each arm is embedded in an underlying metric space. In such a variant it is natural to assume that neighboring arms have similar reward values, and across-arm correlations provide a natural way to model these dependencies. Accordingly, I adopt the across-arm model of [5], which is structured as follows. Let i and j be two arms embedded in an underlying metric space at locations $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Then the correlation coefficient between the reward values associated with the k^{th} objective for these two arms is

$$\rho_{ij}^k = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\lambda_k}\right) \in (0, 1], \tag{8}$$

where $\|\cdot\|$ is the metric associated with the metric space and $\lambda_k \geq 0$ is the length scale associated with objective k. The limit $\lambda_k \to 0$ corresponds to the case where reward values associated with different arms are independent, while positive values of λ_k are associated with correlated values.

Using the across-arms correlation model (8), I can now construct the full prior covariance matrix. Recall that Σ_i^0 is the prior covariance matrix for arm i, and define, for each arm i, $\boldsymbol{\sigma}_i^0 = \operatorname{diag}(\Sigma_i^0) \in \mathbb{R}^{n_o}$ as the vector of its diagonal components. For each pair of distinct arms i and j, $i \neq j$, define $\boldsymbol{\rho}_{ij} \in \mathbb{R}^{n_o}$ as the vector of the n_o correlation coefficients ρ_{ij}^k defined by (8). This vector encodes all the correlation information between the rewards associated with arms i and j.

For a vector $\mathbf{y} \in \mathbb{R}^n$, define $\operatorname{diag}(\mathbf{y}) \in \mathbb{R}^{n \times n}$ as the diagonal matrix with the elements of $\mathbf{y}$ on its diagonal. The across-arm covariance matrix Σ_{ij}^0 can then be defined as:

$$\Sigma_{ij}^0 = \operatorname{diag}(\boldsymbol{\rho}_{ij}) \left(\operatorname{diag}(\boldsymbol{\sigma}_i^0) \operatorname{diag}(\boldsymbol{\sigma}_j^0)\right)^{1/2}.$$

The matrix square root is well defined because the prior covariances Σ_i^0 are positive-definite matrices. Also note that, by construction, Σ_{ij}^0 is symmetric under interchange of i and j, so $\Sigma_{ij}^0 = \Sigma_{ji}^0$. I then define the complete prior covariance matrix $\Sigma^0 \in \mathbb{R}^{n_o \cdot N \times n_o \cdot N}$ as the block matrix with Σ_i^0 as the i^{th} component of its diagonal and Σ_{ij}^0 as the i, j off-diagonal component:

$$\Sigma^0 = \sigma_0^2 \begin{bmatrix} \Sigma_1^0 & \Sigma_{12}^0 & \cdots & \Sigma_{1N}^0 \\ \Sigma_{21}^0 & \Sigma_2^0 & \cdots & \Sigma_{2N}^0 \\ \vdots & & \ddots & \vdots \\ \Sigma_{N1}^0 & \Sigma_{N2}^0 & \cdots & \Sigma_N^0 \end{bmatrix},$$

where $\sigma_0^2 > 0$ represents the strength of the agent's beliefs: larger values of σ_0 correspond to weaker beliefs. The limit $\sigma_0^2 \to +\infty$ corresponds to vanishingly weak initial beliefs, which I refer to as an *uninformative prior*. Finally, define $\boldsymbol{\mu}^0 \in \mathbb{R}^{n_o \cdot N}$ as the stacking of the individual mean belief vectors $\boldsymbol{\mu}_i^0$. Then the prior on $\mathbf{M}$, the vector of all the means, is the multivariate Gaussian distribution

$$\mathbf{M} \sim \mathcal{N}(\boldsymbol{\mu}^0, \Sigma^0). \tag{9}$$

Since the agent's beliefs are Gaussian, it suffices to keep track of the mean and variance. In the following, I refer to the mean-variance pair $(\boldsymbol{\mu}^0, \Sigma^0)$ as the *belief state* at the initial time $t = 0$.

B. Update rule

At each decision time t, the agent selects an arm i_t and receives a reward $\mathbf{r}_t$ drawn from the probability distribution ν_{i_t}. It must then update its belief state $(\boldsymbol{\mu}^t, \Sigma^t)$ to incorporate the new information gained from observing the reward. Bayesian inference provides an optimal solution to the update problem which I adopt in the following. The Gaussian prior (9) is conjugate to the Gaussian reward distributions ν, so the posterior distribution is also Gaussian. Furthermore, the update rule is linear and can be written down in closed form. The update equations for the single objective case, where $\mathbf{r}_t$ is scalar, are well known (see, e.g., [16, Theorem 10.3]). The generalization to the multiple objective case is straightforward and can be written down as follows.

Define $\boldsymbol{\phi}_i \in \mathbb{R}^N$ as the indicator vector whose i^{th} element is equal to one with all others equal to zero. Furthermore, define $\Lambda^t = (\Sigma^t)^{-1}$ and, for each arm i, $\Lambda_{s,i} = (\Sigma_{s,i})^{-1}$ as the *precision* matrices of the belief state and the reward distributions, respectively. Then the belief state $(\boldsymbol{\mu}^t, \Sigma^t)$ updates according to

$$\Lambda^{t+1} = \Lambda^t + \left(\boldsymbol{\phi}_{i_t} \boldsymbol{\phi}_{i_t}^\top\right) \otimes \Lambda_{s,i} \tag{10}$$

$$\Sigma^{t+1} = \left(\Lambda^{t+1}\right)^{-1} \tag{11}$$

$$\boldsymbol{\mu}^{t+1} = \Sigma^{t+1} \left(\Lambda^t \boldsymbol{\mu}^t + \boldsymbol{\phi}_{i_t} \otimes (\Lambda_{s,i_t} \mathbf{r}_t)\right), \tag{12}$$

where $\otimes$ is the Kronecker matrix product. Equations (10)–(12) can be viewed as a Kalman filter estimating the state $\mathbf{M}$, which has the trivial dynamics $\dot{\mathbf{M}} = 0$. The state Λ^t can be viewed as the information matrix of the Kalman filter, and $\boldsymbol{\phi}_{i_t}\boldsymbol{\phi}_{i_t}^\top$ as the observation matrix corresponding to observing the i_t^{th} arm.

The belief state tracks the agent's Gaussian belief about the mean rewards $\mathbf{M}$. The utilities $\bar{\mathbf{m}}$ are a linear combination of the mean rewards $\mathbf{M}$, so the agent's beliefs about $\bar{\mathbf{m}}$ are also Gaussian and can be represented by an analogous belief state $(\bar{\boldsymbol{\mu}}^t, \bar{\Sigma}^t)$:

$$\bar{\mathbf{m}} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}^t, \bar{\Sigma}^t).$$

This new belief state can be computed by taking the appropriate linear combinations, as follows:

$$\bar{\boldsymbol{\mu}}^t = \left(I_N \otimes \mathbf{w}^\top\right) \boldsymbol{\mu}^t \tag{13}$$

$$\bar{\Sigma}^t = \left(I_N \otimes \mathbf{w}^\top\right) \Sigma^t \left(I_N \otimes \mathbf{w}\right), \tag{14}$$

where $I_N \in \mathbb{R}^{N\times N}$ is the $N \times N$ identity matrix.

C. Decision heuristic

Once I have the Gaussian belief state about $\bar{\mathbf{m}}$, I can readily adapt the UCL decision heuristic to the multi-objective problem. Denote the i^{th} component of $\bar{\boldsymbol{\mu}}^t$ by $\bar{\mu}_i^t$ and the square root of the i^{th} component on the diagonal of $\bar{\Sigma}^t$ by $\bar{\sigma}_i^t$. These represent, respectively, the mean and standard deviation of the agent's belief about the weighted mean rewards $\bar{m}_i$ associated with arm i.

I then adapt the UCL decision heuristic as follows. At each decision time t, define the heuristic function Q_i^t for each arm i by

$$Q_i^t = \bar{\mu}_i^t + \bar{\sigma}_i^t \Phi^{-1}(1-\alpha_t), \tag{15}$$

where $\Phi^{-1}(\cdot)$ is the *inverse cumulative distribution* or *quantile* function of the Gaussian distribution and $\alpha_t = 1/t$ is a decreasing function of time.

At each decision time t, the MO-UCL algorithm picks the arm i_t with maximal heuristic value

$$i_t = \arg\max_i Q_i^t. \tag{16}$$

V. MO-UCL PERFORMANCE

In this section, I analyze the MO-UCL algorithm and show that it achieves efficient performance in terms of scalar regret. I use the following bound from [17] in our analysis.

For the standard normal (i.e., Gaussian) random variable z and a constant $w \in \mathbb{R}_{\geq 0}$,

$$\Pr\left[z \geq w\right] \leq \frac{2e^{-w^2/2}}{\sqrt{2\pi}(w + \sqrt{w^2 + 8/\pi})} \leq \frac{1}{2}e^{-w^2/2}. \tag{17}$$

It follows from (17) that for any $\alpha \in [0.5, 1]$,

$$\Phi^{-1}(1-\alpha) \leq \sqrt{-2\log\alpha}. \tag{18}$$

I define $\{R_t^{\text{MO-UCL}}\}_{t\in\{1,\ldots,T\}}$ as the sequence of expected regret for the MO-UCL algorithm. The MO-UCL algorithm achieves logarithmic expected scalar regret as formalized by the following theorem.

Theorem 2 (MO-UCL performance): Fix a weight vector $\mathbf{w} \in \mathbb{R}_+^{n_o}$. When the the MO-UCL algorithm is run with an uninformative prior, it achieves logarithmic cumulative expected scalar regret. In particular, for each suboptimal arm i, the following holds:

$$\mathbb{E}\left[n_i^T\right] \leq \left(\frac{8\bar{\sigma}_{s,i}^2}{\bar{\Delta}_i^2} + 2\right)\log T + 3.$$

Furthermore, the total cumulative expected scalar regret obeys

$$\mathbb{E}\left[\sum_{t=1}^T R_t^{\text{MO-UCL}}\right] \leq \sum_{i=1}^N \bar{\Delta}_i \left(\left(\frac{8\bar{\sigma}_{s,i}^2}{\bar{\Delta}_i^2} + 2\right)\log T + 3\right).$$

Proof: I begin by establishing the bound on $\mathbb{E}\left[n_i^T\right]$. In the spirit of [4], I 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 option i over i^* only if

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

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

$$\bar{\mu}_{i^*} \leq \bar{m}_{i^*} - C_{i^*}^t \tag{19}$$

$$\bar{\mu}_i^t \geq \bar{m}_i + C_i^t \tag{20}$$

$$\bar{m}_{i^*} < \bar{m}_i + 2C_i^t, \tag{21}$$

where $C_i^t = \bar{\sigma}_i^t \Phi^{-1}(1-\alpha_t)$ and $\alpha_t = 1/t$. Otherwise, if none of the equations (19)–(21) holds,

$$Q_{i^*}^t = \bar{\mu}_{i^*}^t + C_{i^*}^t > \bar{m}_{i^*} \geq \bar{m}_i + 2C_i^t > \bar{\mu}_i^t + C_i^t = Q_i^t,$$

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

I proceed by analyzing the probability that Equations (19) and (20) hold. Note that the empirical mean reward $\hat{\mathbf{m}}_i$ from arm i is a Gaussian random variable with mean $\mathbf{m}_i$ and variance $\Sigma_{s,i}/n_i^t$. Therefore, for an uninformative prior and conditional on n_i^t, $\boldsymbol{\mu}_i^t$ is distributed as

$$\boldsymbol{\mu}_i^t \sim \mathcal{N}(\mathbf{m}_i, \Sigma_{s,i}/n_i^t),$$

and the belief about the scalarized reward is distributed as

$$\bar{\mu}_i^t \sim \mathcal{N}(\bar{m}_i, \bar{\sigma}_{s,i}^2/n_i^t).$$

Equation (19) is equivalent to

$$\begin{aligned} \bar{m}_{i^*} + \frac{\bar{\sigma}_{s,i^*}}{\sqrt{n_{i^*}^t}} z &\leq \bar{m}_{i^*} - \frac{\bar{\sigma}_{s,i^*}}{\sqrt{n_{i^*}^t}} \Phi^{-1}(1-\alpha_t) \\ \Leftrightarrow z &\leq -\Phi^{-1}(1-\alpha_t), \end{aligned}$$

where z is a standard normal random variable. Consequently, for an uninformative prior,

$$\Pr\left[\text{Equation (19) holds}\right] = \alpha_t.$$

Similarly, Equation (20) is equivalent to

$$\bar{m}_i + \frac{\bar{\sigma}_{s,i}}{\sqrt{n_i^t}} z \geq \bar{m}_i + \frac{\bar{\sigma}_{s,i}}{\sqrt{n_i^t}} \Phi^{-1}(1-\alpha_t)$$
$$\Leftrightarrow z \geq \Phi^{-1}(1-\alpha_t),$$

where z is a standard normal random variable. Consequently, for an uninformative prior,

$$\Pr\left[\text{Equation (20) holds}\right] = \alpha_t.$$

Finally, I consider Equation (21), which holds if

$$\bar{m}_{i^*} < \bar{m}_i + 2\frac{\bar{\sigma}_{s,i}}{\sqrt{n_i^t}} \Phi^{-1}(1-\alpha_t)$$
$$\Leftrightarrow \bar{\Delta}_i < 2\frac{\bar{\sigma}_{s,i}}{\sqrt{n_i^t}} \Phi^{-1}(1-\alpha_t)$$
$$\Leftarrow \bar{\Delta}_i < 2\frac{\bar{\sigma}_{s,i}}{\sqrt{n_i^t}} \sqrt{-2\log\alpha_t}$$
$$\Leftrightarrow n_i^t < 4\frac{\bar{\sigma}_{s,i}}{\bar{\Delta}_i^2}(2\log(1/\alpha_t))$$
$$\Leftrightarrow n_i^t < 8\frac{\bar{\sigma}_{s,i}}{\bar{\Delta}_i^2}\log T,$$

where the third inequality holds because of the bound (18) and the final inequality holds because of the monotonicity of the log function. Therefore, Equation (21) never holds if

$$n_i^t \geq \lceil 8\frac{\bar{\sigma}_{s,i}}{\bar{\Delta}_i^2}\log T \rceil,$$

so set $\eta = 8\frac{\bar{\sigma}_{s,i}}{\bar{\Delta}_i^2}\log T + 1$ and the bound becomes

$$\mathbb{E}\left[n_i^T\right] \leq 8\frac{\bar{\sigma}_{s,i}}{\bar{\Delta}_i^2}\log T + 1 + \sum_{t=1}^{T} 2\alpha_t.$$

The sum can be bounded by the integral

$$\sum_{t=1}^{T}\frac{2}{t} \leq 2\left(1 + \int_1^T \frac{1}{t}\mathrm{d}t\right) = 2(1+\log T),$$

which yields the final result. The bound on the total cumulative expected scalar regret follows from its definition. ■

VI. Numerical Results

In this section, I present the results of a numerical simulation demonstrating the result of Theorem 2. The multi-armed bandit has $N = 4$ arms, each of which has an $n_o = 3$-dimensional Gaussian reward distribution. Each arm $i \in \{1, 2, 3, 4\}$ has a mean of $\mathbf{m}_i = i\,[1,\ 2,\ 3]$ and a variance of $\Sigma_{s,i} = \mathrm{diag}([1,\ 1.5,\ 2])$. The agent uses scalarization weights $\mathbf{w} = [1/2,\ 1/3,\ 1/6]$, which results in a vector of scalar means $\bar{\mathbf{m}} = [5/3,\ 10/3,\ 5,\ 20/3]$ and scalarized reward variance $\bar{\sigma}_{s,i}^2 = 17/36$ for each arm i. The optimal arm is $i^* = 4$.

Figure 1 shows mean cumulative scalar regret from 100 simulations where a simulated agent used the MO-UCL algorithm with an uninformative prior for $T = 100$ decisions. The dashed blue line represents the Lai-Robbins lower bound on regret (5), while the dash-dotted red line represents the upper bound on regret from Theorem 2. Due to the logarithmic scaling of the horizontal axis, both bounds appear as straight lines. The algorithm's cumulative scalar regret remains within a constant factor of the Lai-Robbins bound, thereby demonstrating optimal performance.

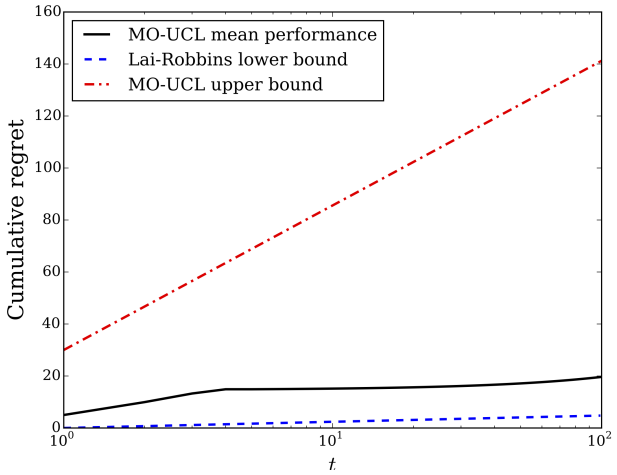

Fig. 1. Simulation results of the MO-UCL algorithm being applied on an $N = 4$-armed multi-objective multi-armed bandit problem with $n_o = 3$ objectives. The algorithm achieves a scalar cumulative regret that is within a constant factor of the Lai-Robbins lower bound (5), thereby demonstrating optimal performance.

VII. Conclusions

In this paper I studied multi-objective decision making under uncertainty using the multi-armed bandit problem. I defined the multi-objective multi-armed bandit problem as an extension of the standard multi-armed bandit problem with a single scalar reward objective. In the special case of Gaussian rewards and a linear scalarization, I showed that the multi-objective problem reduces to a single-objective bandit problem again with Gaussian rewards. I used this result to bound optimal performance in terms of the scalar rewards and developed an algorithm that achieves that optimum.

The case of Gaussian rewards and linear scalarization is somewhat limited, but two things should be noted. First, the location-scale property of the Gaussian rewards greatly facilitates modeling and the development of priors, which the MO-UCL algorithm is designed to take advantage of. Second, it is relatively straightforward to extend the ideas presented here to the case of a more general scalarization function, as long as it maps to a finite range of scalar values. Analysis using Hoeffding's inequality along the lines of [4] can deal with this case.

Going forward, I hope to use the model presented here to combine the rigorous study of human decision making under uncertainty of [5] with the heuristic approach to integrating human designers in the engineering MDO decision-making process studied in [6].

References

[1] J. R. Martins and A. B. Lambe, "Multidisciplinary design optimization: a survey of architectures," *AIAA J.*, vol. 51, no. 9, pp. 2049–2075, 2013.

[2] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin Amer. Math. Soc.*, vol. 58, pp. 527–535, 1952.

[3] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.
[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learning*, vol. 47, no. 2, pp. 235–256, 2002.
[5] P. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision making in generalized multi-armed bandits," *Proc. IEEE*, vol. 102, no. 4, pp. 544–571, 2014.
[6] P. Reverdy, A. Reddy, L. Martinelli, and N. E. Leonard, "Integrating a human designer's preferences in multidisciplinary design optimization," in *Proc. 15th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conf.*, no. AIAA 2014-2167, 2014.
[7] M. Drugan and A. Nowe, "Designing multi-objective multi-armed bandits algorithms: A study," in *2013 Int. Joint Conf. Neural Networks*, Aug 2013, pp. 1–8.
[8] K. V. Moffaert, K. Van Vaerenbergh, P. Vrancx, and A. Nowé, "Multi-objective $\mathcal{X}$-armed bandits," in *Proc. 2014 IEEE World Congress on Computational Intelligence*, 2014.
[9] R. L. Keeney and H. Raiffa, "Decision analysis with multiple conflicting objectives," *Wiley & Sons, New York*, 1976.
[10] ——, *Decisions with multiple objectives: preferences and value tradeoffs*. Cambridge University Press, 1993.
[11] J. P. Nolan, *Stable Distributions - Models for Heavy Tailed Data*. Boston: Birkhauser, 2015, in progress, Chapter 1 online at academic2.american.edu/∼jpnolan.
[12] P. Reverdy, V. Srivastava, R. C. Wilson, and N. E. Leonard, "Human decision making and the explore-exploit tradeoff: Algorithmic models for multi-armed bandit problems," in *Cognitive Dynamic Systems*, ser. Wiley Series on Adaptive and Cognitive Dynamic Systems. IEEE Press/Wiley, 2015.
[13] P. Reverdy, "Human-inspired algorithms for search: A framework for human-machine multi-armed bandit problems," Ph.D. dissertation, Princeton Univ., 2014.
[14] R. Kleinberg and A. Slivkins, "Sharp dichotomies for regret minimization in metric spaces," in *Proc. ACM-SIAM Symp. Discrete Algorithms, SODA*, 2010, pp. 827–846.
[15] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari, "$\mathcal{X}$-armed bandits," *J. Mach. Learning Research*, vol. 12, pp. 1655–1695, 2011.
[16] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I : Estimation Theory*. Prentice Hall, 1993.
[17] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1964.