Estimating Quality in User-Guided
Multi-Objective Bandits Optimization

Audrey Durand and Christian Gagné
Computer Vision and Systems Laboratory
Université Laval, Québec (QC), Canada
{ audrey.durand.2@ulaval.ca, christian.gagne@gel.ulaval.ca }

January 5, 2017

Abstract

Many real-world applications are characterized by a number of conflicting performance measures. As optimizing in a multi-objective setting leads to a set of non-dominated solutions, a preference function is required for selecting the solution with the appropriate trade-off between the objectives. This preference function is often unknown, especially when it comes from an expert human user. However, if we could provide the expert user with a proper estimation for each action, she would be able to pick her best choice. The question is: how good do these estimations have to be in order for her choice to remain the same as if she had access to the exact values? In this paper, we introduce the concept of preference radius to characterize the robustness of the preference function and provide guidelines for controlling the quality of estimations in the multi-objective setting. More specifically, we provide a general formulation of multi-objective optimization under the bandits setting and the pure exploration setting with user feedback for articulating the preferences. We show how the preference radius relates to the optimal gap and how it can be used to analyze algorithms in the bandits and pure exploration settings. We finally present experiments in the bandits setting, where we evaluate the impact of noise and delayed expert user feedback, and in the pure exploration setting, where we compare multi-objective Thompson sampling with uniform sampling.

1 Introduction

Multi-objective optimization (MOO) [5] is a topic of great importance for real-world applications. Indeed, optimization problems are characterized by a number of conflicting, even contradictory, performance measures relevant to the task at hand. For example, when deciding on the healthcare treatment to follow for a given sick patient, a trade-off must be made between the efficiency of the treatment to heal the sickness, the side effects of the treatment, and the treatment cost. In this work, we consider problems where outcomes are stochastic and costly to evaluate (e.g., involving a human in the loop). The challenge is therefore to identify the best solutions given random observations sampled from different (unknown) density distributions. We formulate this problem as multi-objective bandits, where the performance of the solutions evaluated during the optimization is important, or pure exploration, where only identifying the best final solution within a fixed number of attempts is under consideration.
MOO is often tackled by combining the objective into a single measure (a.k.a. scalarization). Such approaches are said to be *a priori*, as the preferences over the objectives is defined before doing the optimization itself. The challenge lies in the determination of the appropriate scalarization function to use and its parameterization. Another way to conduct MOO consists in learning the optimal trade-offs (the so-called Pareto-optimal set). Once the optimization is completed, techniques from the field of multi-criteria decision-making are applied to help the user to select the final solution from the Pareto-optimal set. These *a posteriori* techniques may require a huge number of evaluations to have a reliable estimation of the objective values over all potential solutions. Indeed, the Pareto-optimal set can be quite large, encompassing a majority, if not all, of the potential solutions. One must recall that in the end, a single solution is of interest. Therefore, the search could rather focus on a specific zone of the objective space. In this work, we tackle the MOO problem where the scalarization function is unknown and a user acts as a black box for articulating preferences. Integrating the user to the learning loop, she can provide feedback by selecting her preferred choice given a set of options – the scalarization function lying in her head.

Let the *right choice* denote the option that the user would select given that she had knowledge of the Pareto-optimal set. A learning algorithm for the multi-objective bandits setting aims at learning good-enough estimations of the available options to allow the user to make the right choices and its performance depends on the robustness of the user preference function to the quality of estimations. We therefore need a measure for characterizing the quality of estimations required to allow the user to make good decisions. For that purpose, we propose the concept of preference radius providing the tolerance range over objective value estimations, such that the user preference would remain the same as if the Pareto-optimal set was known.

The original contributions of the paper consist in:

- providing a general formulation of the MOO under the bandits and pure exploration settings;
- proposing the preference radius to characterize the robustness of the preference function to the estimations quality;
- relating the preference radius to the optimal gap and showing how it can be used for analyzing the regret;
- providing an empirical evaluation of the impact of noise in the user preference function and of delayed expert user feedback in the bandits setting, and linking the results to the analysis.

## 2 Multi-Objective Bandits

A multi-objective bandits problem is described by a (finite) set of actions $\mathcal{A}$, also referred to as the *design space*, and a sequence

$$\{(z_a(t))_{a \in \mathcal{A}}\}_{t \geq 1},$$

where $z_a(t) = (z_{a,1}(t), \ldots, z_{a,d}(t))^1$ denotes the $d$-dimensional multi-objective random output yielded by action $a$ at time $t$, for some $d \in \mathbb{N}$. We consider the stochastic bandit where, for each action $a$,
there is a $d$-dimensional multivariate distribution $\nu_a$ with unknown mean $\mu_a = (\mu_{a,1}, \ldots, \mu_{a,d}) \in \mathcal{X} \in \mathbb{R}^d$ and $z_a(t) \sim \nu_a$ are i.i.d. random vectors. For simplicity, we assume that the objective space $\mathcal{X} = [0, 1]^d$.

**Pareto-optimality:** Given two $d$-dimensional options $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$. $x$ is said to dominate, or Pareto-dominate, $y$ (denoted $x \succeq y$) if and only if $x_i > y_i$ for at least one $i$ and $x_i \geq y_i$ otherwise. The dominance is strict (denoted $x > y$) if and only if $x_i > y_i$ for all $i = 1, \ldots, d$. Finally, the two vectors are incomparable (denoted $x \parallel y$) if $x \not> y$ and $y \not> x$. Pareto-optimal options represent the best compromises amongst the objectives and are the only options that need to be considered in an application. We say that these options constitute the Pareto front $\mathcal{P} = \{a : \nexists b \in \mathcal{A} \text{ s.t. } g(b) \preceq g(a)\}$. Fig. 1 shows an example of dominated and non-dominated expected outcomes in a $d = 2$ objectives space. A user facing a multi-criteria decision making problem must select its preferred non-dominated option. Dominated options are obviously discarded by default.

**Setting:** In this episodic game, an agent interacts with a user characterized by an unknown (to the agent and to the user) preference function $f$. The agent iteratively chooses to perform an action and observes its outcome. The agent faces a trade-off between choosing to sample a distribution that the user might prefer, and choosing to sample a relatively unexplored distribution to improve its estimate. If users were able to express clearly their preference function, the situation would easily be solved as a standard bandit problem. Unfortunately, this is not the case as users are generally unable to scalarize their choices and preferences. Therefore, they cannot provide a function of reward (or regret) that can be used directly. However, given several options, users can tell which one they prefer and therefore can be used as a black box to provide feedback in the learning loop. Algorithm 1 describes this multi-objective bandits problem.

An algorithm for a multi-objective bandits problem is a (possibly randomized) method for choosing which distribution to sample from next, given a history of previous choices and obtained outcomes, $\mathcal{H}_{t-1} = \{a(s), z_{a(s)}(s)\}_{s=1}^{t-1}$, and preference indications. Let $\mathcal{O} = \arg\max_{a \in \mathcal{A}} f(\mu_a) \subseteq \mathcal{P}$ and let $\star \in \mathcal{O}$ denote the optimal action. Recall that we refer as the optimal action the option that an
Algorithm 1 Multi-objective bandits setting

On each episode $t \geq 1$:

1. The agent selects an action $a(t)$ to play.
2. The agent observes the outcome $z_{a(t)}(t)$.
3. The agent updates its estimates.
4. The agent shows options $\{\theta_a(t)\}_{a \in A}$ to the expert user.
5. The expert user indicates its preference $O(t) = \arg\max_{a \in A} f(\theta_a(t))$.

The expert user would select assuming that she had knowledge of the true Pareto-optimal set. The optimal gap $\Delta_a = f(\mu_\star) - f(\mu_a)$ measures the expected loss of playing playing action $a$ instead of the optimal action.

2.1 Performance Metrics

The performance depends on the capability of the expert user to select $\star$ among the available options.

**Bandit** In the bandit setting, the agent’s goal is to design an algorithm with low expected (cumulative) regret:\(^2\):

$$E[R(T)] = E \left[ \sum_{t=1}^{T} (f(\mu_\star) - f(z_{a(t)})) \right] = \sum_{a \in A} \sum_{t=1}^{T} P[a(t) = a] \Delta_a. \quad (1)$$

This quantity measures the expected performance of the algorithm compared to the expected performance of an optimal algorithm given knowledge of the outcome distributions and the unknown function $f$, i.e., always sampling from the distribution with the expectation maximizing $f$.

**Pure exploration** The pure exploration setting assumes that the algorithm has $T$ (possibly unknown) free episodes to learn before recommending the action $a_T$ that is going to be played forever. The agent’s goal is therefore to design an algorithm with low expected simple regret:

$$E[R_s(T)] = E[f(\mu_\star) - f(\mu_{a_T})] = \sum_{a \in A} P[a_T = a] \Delta_a. \quad (2)$$

2.2 Related Works

In a first formulation of multi-objective bandits [6], the goal was to minimize a Pareto-regret metric. Also known as the $\epsilon$-distance [8], the Pareto-regret associated with playing action $a$ is the minimum value $\epsilon_a$ such that $\mu_a + \epsilon_a$ is not dominated by any other actions. In other words, any actions standing on the front is considered equally good by the expert user. This is a specific instance of our more general setting where $O = \mathcal{P}$. We can represent this by considering the preference function

$$f(\mu_\star) = 1 \quad \text{and} \quad f(\mu_a) = 1 - \epsilon_a,$$

\(^2\) Also known as the scalarized regret [6].
such that $\Delta_a = \epsilon_a$. With this formulation, any standard bandit algorithm that could optimize a single objective could minimize the regret regardless of the other objectives.

In MOO settings [10], the goal is to identify the Pareto-optimal set $P$ without evaluating all actions. The quality of a solution $S$ is typically given by the hypervolume error $V(P) - V(S)$, where the $V(P)$ is the volume enclosed between the origin and $\{\mu_a\}_{a \in P}$ (and similarly for $S$). However, the hypervolume error does not give information about the quality of the estimation of actions. Identifying the Pareto front alone does not guarantee that the actions are well estimated and, therefore, that an expert user choice based on these estimations would lead to the right choice.

3 Preference Radius

Let $\theta_a(t)$ denote the estimation associated with action $a$ on episode $t$. The set of options $\{\theta_a(t)\}_{a \in A}$ is presented to the expert user on step 4 of Algorithm 1. Let $P(t) = \{a : \exists \theta_b(t) > \theta_a(t) \forall b \in A\}_{a \in A}$ denote the estimated Pareto front given these options. From step 5, we then obtain the optimal options $O(t) = \arg\max_{a \in A} f(\theta_a(t)) \subseteq P(t)$. Let

$$B(c, r) \subseteq \{x \in \mathcal{X} : |x_i - c_i| < r, i = 1, \ldots, d\}$$

denote a ball of center $c$ and radius $r$. In order to characterize the difficulty of a multi-objective bandits setting, we define the following problem-dependent quantity.

**Definition 1** For each action $a \in A$, we define the preference radius $\rho_a$ as any radius such that if $\theta_a(t) \in B(\mu_a, \rho_a)$ for all actions, then

$$\exists \star \in O : \star \in O(t) \quad \text{and} \quad a \notin O(t) \quad \forall a \in A, a \notin O.$$

The radii correspond to the robustness of the preference function, that is how much simultaneous variation of each action can happen before the expert user preference changes. The radius $\rho_a$ is directly linked to the gap $\Delta_a = f(\mu_\star) - f(\mu_a)$. For a suboptimal action $a \notin O$, a large $\rho_a$ indicates that action $a$ is far from being optimal. Therefore, $f(\mu_a)$ must be low and its gap $\Delta_a$ must be large. Also, the preference radii of suboptimal actions depend on the preference radius of the optimal action(s). Larger optimal action radii imply smaller radii for suboptimal actions.

Let $\alpha_1, \ldots, \alpha_d \in [0, 1]$ denote weights such that $\sum_{i=1}^d \alpha_i = 1$. The weighted $L_p$ metric $f(x) = (\sum_{i=1}^d \alpha_i x_i^p)^{1/p}$ with $p \geq 1$ is often used to represent decision functions. This function is known as the linear scalarization when $p = 1$ and as the Chebyshev scalarization when $p = \infty$. The following examples show the link between the preference radii and the gap for these two common functions.

**Example 1 (Linear scalarization)** The linear scalarization function is given by

$$f(x) = \sum_{i=1}^d \alpha_i x_i.$$

Consider the optimal action $\star$ and the suboptimal action $a$. By definition of the preference radii,
Figure 2: Examples of preference radii around the optimal (white) and suboptimal (black) actions given the linear preference function $f(x) = 0.4x_1 + 0.6x_2$.

we have that

$$\min_{\theta_* \in B(\mu_*, \rho_*)} f(\theta_*) > \max_{\theta_a \in B(\mu_a, \rho_a)} f(\theta_a)$$

$$\sum_{i=1}^{d} (\alpha_i \mu_{*,i} - \alpha_i \rho_*) > \sum_{i=1}^{d} (\alpha_i \mu_{a,i} + \alpha_i \rho_a)$$

$$f(\mu_*) - \rho_* > f(\mu_a) + \rho_a$$

$$\Delta_a > \rho_* + \rho_a.$$  

Figure 2 shows examples of preference radii with a linear preference function.

**Example 2 (Chebyshev scalarization)** The Chebyshev scalarization \([2]\) function is given by

$$f(x) = \max_{1 \leq i \leq d} \alpha_i x_i.$$  

Consider the optimal and suboptimal actions $\star$ and $a$, and let

$$i_* = \arg\max_{1 \leq i \leq d} \alpha_i (\mu_{*,i} - \rho_*) \quad i_a = \arg\max_{1 \leq i \leq d} \alpha_i (\mu_{a,i} - \rho_a).$$

By definition of the preference radii, we have that

$$\min_{\theta_* \in B(\mu_*, \rho_*)} f(\theta_*) > \max_{\theta_a \in B(\mu_a, \rho_a)} f(\theta_a)$$

$$\max_{1 \leq i \leq d} \alpha_i (\mu_{*,i} + \rho_*) > \max_{1 \leq i \leq d} \alpha_i (\mu_{a,i} - \rho_a)$$

$$\alpha_{i_*} \mu_{*,i} - \alpha_{i_*} \rho_* > \alpha_{i_a} \mu_{a,i} + \alpha_{i_a} \rho_a$$

$$f(\mu_*) - \alpha_{i_*} \rho_* > f(\mu_a) + \alpha_{i_a} \rho_a$$

$$\Delta_a > \alpha_{i_*} \rho_* + \alpha_{i_a} \rho_a.$$  

The difficulty here is that $i_*$ and $i_a$ respectively depend on $\rho_*$ and $\rho_a$. Consider a 2-objectives setting, we can define

$$\tau_* = \frac{\alpha_2 \mu_{*,2} - \alpha_1 \mu_{*,1}}{\alpha_2 - \alpha_1}, \quad \tau_a = \frac{\alpha_1 \mu_{a,1} - \alpha_2 \mu_{a,2}}{\alpha_2 - \alpha_1}$$
Figure 3: Examples of preference radii around the optimal action (white) and suboptimal actions (black) given a Chebyshev function with $\alpha_1 = 0.4$ and $\alpha_2 = 0.6$.

as thresholds such that

$$i_* = \begin{cases} 1 & \text{if } \rho_* > \tau_* \\ 0 & \text{otherwise} \end{cases}, \quad i_a = \begin{cases} 1 & \text{if } \rho_a < \tau_a \\ 0 & \text{otherwise} \end{cases}.$$  

Figure 3 shows examples of preference radii with a Chebyshev preference function.

Outside $L_p$ metrics, other scalarization functions are often based on constraints. For example, using the $\epsilon$-constraint scalarization technique, a user assigns a constraint to every objectives except a target objective $\ell$. All options that fail to respect one of the constraints get a value of 0, while the options that respect all constraints get a value of $x_\ell$. The following example shows the relation between the preference radius and the gap given a preference function that is articulated as an $\epsilon$-constraint scalarization technique.

**Example 3 (Epsilon-constraint)** The $\epsilon$-constraint function is given by

$$f(x) = \begin{cases} x_\ell & \text{if } x_i \geq \epsilon_i \quad \forall i \in \{1, \ldots, d\}, i \neq \ell \\ 0 & \text{otherwise.} \end{cases}$$

Consider the optimal and suboptimal actions $*$ and $a$. By definition of the preference radii, we have that

$$\rho_* \leq \min_{1 \leq i \leq d, i \neq \ell} \mu_{*,i} - \epsilon_i.$$  

We decompose $\rho_a = \rho_a + \tilde{\rho}_a$ such that

$$\rho_a = \min\{0, \max_{1 \leq i \leq d, i \neq \ell} \epsilon_i - \mu_a,i\}$$

denotes the radius required in order for action $a$ to respect the constraints, that is to get $f(\mu_a) > 0$, and $\tilde{\rho}_a$ denotes the leftover leading to a gap reduction. Finally, we have that

$$\mu_{*,\ell} - \rho_* > \mu_{a,\ell} + \rho_a + \tilde{\rho}_a \quad \text{and} \quad \Delta_a > \rho_* + \rho_a.$$  

Figure 4 shows examples of preference radii with an $\epsilon$-constraint preference functions.
4 Controlling the Regret

In this section, we use the preference radius to analyze the algorithms in the bandits and pure exploration settings. Essentially, we show that the preference radius allows to analyze the multi-objective setting very similarly as standard, single-objective settings. Let

\[ S_a(t) = \sum_{s=1}^{t-1} z_a(s) I\{a(s) = a\} \quad \text{and} \quad N_a(t) = \sum_{s=1}^{t-1} I\{a(s) = a\} \]

respectively denote the sum of observed outcomes for action \( a \) up to episode \( t \) and the number of times action \( a \) has been played up to episode \( t \). The empirical average and covariance are given by

\[ \hat{\mu}_a(t) = S_a(t)/N_a(t) \quad \text{and} \quad \hat{\Sigma}_a(t) = \frac{1}{N_a(t) - 1} \sum_{s=1}^{t-1} (z_a(s) - \hat{\mu}_a(t))(z_a(s) - \hat{\mu}_a(t))^T I\{a(s) = a\}. \]

4.1 Bandit Setting

We want to bound the expected regret (Equation 1). Let us consider the typical Thompson sampling [9] algorithm that selects the next action according to its probability of being optimal. The idea is to maintain a posterior distribution \( \pi_a(t) \) given a prior \( \pi_a(0) = \pi_0 \) and the history of observations \( \mathcal{H}_{t-1} \). On episode \( t \), one option \( \theta_a(t) \) is sampled from each posterior distribution \( \pi_a(t) \).

Example 4 (Multiple Bernoulli) Consider the setting where observations are random variables such that \( z_{a,i}(t) \sim B(\mu_{a,i}) \). The posterior distribution over the outcome expectation for each dimension corresponds to a Beta distribution conditioned using the observations in this dimension only, independently of other dimensions. The options are therefore sampled from one independent Beta distribution.
distribution per dimension:
\[
\theta_{a,i}(t) \sim \text{Beta} \left( \alpha_0 + S_{a,i}(t), \beta_0 + N_a(t) - S_{a,i}(t) \right),
\]
where \( \alpha_0 \) and \( \beta_0 \) are priors.

**Example 5 (Multivariate normal)** With multivariate normal outcome distributions, options are sampled from a multivariate normal distribution:
\[
\theta_a(t) \sim N\left( \tilde{\mu}_a(t), \tilde{\Sigma}_a(t) \right), \quad \text{where}
\]
\[
\tilde{\Sigma}_a(t) = \left( \Sigma_0^{-1} + N_a(t)\Sigma_a(t)^{-1} \right)^{-1},
\]
\[
\tilde{\mu}_a(t) = \Sigma_a(t) \left( \Sigma_0^{-1} \mu_0 + N_a(t)\Sigma_a(t)^{-1} \tilde{\mu}_a(t) \right).
\]

Options \( \{ \theta_a(t) \}_{a \in A} \) are presented to the user (step 4 of Algorithm 1). The options can be seen as estimates of the means given the prior and observations. The algorithm chooses to sample \( a(t) \in O(t) \). Therefore \( P[a(t) = a] \) is proportional to the posterior probability that \( a \) is preferred given the history \( H_{t-1} \).

**Definition 2 (Quantity \( x_a \))** For each action \( a \), we choose a radius threshold \( x_a \) such that \( \mu_a \in B(\mu_a, x_a) \in B(\mu_a, \rho_a) \).

**Definition 3 (Events \( E_{\mu}^a(t), E_{\theta}^a(t) \))** Define \( E_{\mu}^a(t) \) as the event that \( \tilde{\mu}_a(t) \in B(\mu_a, x_a) \). Define \( E_{\theta}^a(t) \) as the event that \( \theta_a(t) \in B(\mu_a, \rho_a) \). Intuitively, \( E_{\mu}^a(t) \) indicates that the empirical mean is well concentrated around the true mean and \( E_{\theta}^a(t) \) indicates that \( \theta_a(t) \) is close enough to \( \mu_a \).

For all suboptimal actions \( a \not\in O \), we can decompose
\[
\sum_{t=1}^{T} P[a(t) = a] \leq \sum_{t=1}^{T} P\left[ E_{\mu}^a(t), E_{\theta}^a(t) \right] + \sum_{t=1}^{T} P\left[ E_{\mu}^a(t), \bar{E}_{\theta}^a(t) \right] + \sum_{t=1}^{T} P\left[ \bar{E}_{\mu}^a(t) \right].
\]

Three situations arise:

A the probability of playing \( a \) when \( \mu_a \) is well estimated and \( \theta_a(t) \) is sampled close to the true mean;

B the probability of playing \( a \) when \( \mu_a \) is well estimated and \( \theta_a(t) \) is sampled far from the true mean;

C the probability of playing \( a \) when \( \mu_a \) is not well estimated.

It boils down to the same expression as previous work [1, 7] for single-objective bandits settings, where concentration and posterior concentration inequalities of multivariate distributions can be used to bound each situation independently.
Noise Tolerance  Now suppose that the expert function is noisy and let
\[ \tilde{f}(\cdot) \sim \mathcal{N}(f(\cdot), \sigma^2) \]
denote an observation from this function, where \( \sigma^2 \) is the noise. In this noisy setting, the user expert indicates its preference \( \tilde{O}(t) = \arg \max_{a \in A} \tilde{f}(\theta_a(t)) \) on step 5 of Algorithm 1. We have that
\[
P\left[ \tilde{f}(\theta_a(t)) > \tilde{f}(\theta_*(t)) \right] = P\left[ \tilde{f}(\theta_a(t)) - \tilde{f}(\theta_*(t)) > 0 \right] = 1 - P\left[ \tilde{f}(\theta_a(t)) - \tilde{f}(\theta_*(t)) \leq 0 \right].
\]
By independence, \( \tilde{f}(\theta_a(t)) - \tilde{f}(\theta_*(t)) \) is normally distributed with mean \( f(\theta_a(t)) - f(\theta_*(t)) \) and variance \( 2\sigma^2 \). Therefore,
\[
\frac{\tilde{f}(\theta_a(t)) - \tilde{f}(\theta_*(t)) - f(\theta_a(t)) + f(\theta_*(t))}{\sqrt{2\sigma^2}} \sim \mathcal{N}(0, 1)
\]
and
\[
P\left[ \tilde{f}(\theta_a(t)) - \tilde{f}(\theta_*(t)) \leq 0 \right] = P\left[ \frac{\tilde{f}(\theta_a(t)) - \tilde{f}(\theta_*(t)) - f(\theta_a(t)) + f(\theta_*(t))}{\sqrt{2\sigma^2}} \leq \frac{f(\theta_*(t)) - f(\theta_a(t))}{\sqrt{2\sigma^2}} \right] = \Phi\left( \frac{f(\theta_*(t)) - f(\theta_a(t))}{\sqrt{2\sigma^2}} \right),
\]
where \( \Phi \) is the cumulative distribution function (CDF) of the standard normal distribution \( \mathcal{N}(0, 1) \). Hence we have that
\[
P\left[ \tilde{f}(\theta_a(t)) > \tilde{f}(\theta_*(t)) \right] = 1 - P\left[ \tilde{f}(\theta_a(t)) - \tilde{f}(\theta_*(t)) \leq 0 \right] = 1 - \Phi\left( \frac{f(\theta_*(t)) - f(\theta_a(t))}{\sqrt{2\sigma^2}} \right). \tag{3}
\]
Recall that
\[
\Phi(x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x}{\sqrt{2}} \right) \right], \quad \text{where} \quad \text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} \, dt
\]
is the error function and \( \text{erfc}(x) = 1 - \text{erf}(x) \). We know [4] that \( \text{erfc}(x) \leq e^{-x^2} \) for \( x > 0 \) such that
\[
\Phi(x) \geq \frac{1}{2} \left[ 1 + 1 - e^{-x^2/2} \right] = 1 - \frac{1}{2} e^{-x^2/2}. \tag{4}
\]
Using Equation 4 into Equation 3, we obtain
\[
P\left[ \tilde{f}(\theta_a(t)) > \tilde{f}(\theta_*(t)) \right] \leq \frac{1}{2} \exp \left( - \frac{(f(\theta_*(t)) - f(\theta_a(t)))^2}{4\sigma^2} \right) \tag{5}
\]
for \( f(\theta_*(t)) - f(\theta_a(t)) > 0 \). In the ideal case, we would have
\[
P\left[ \tilde{f}(\mu_a) > \tilde{f}(\mu_*) \right] \leq \frac{1}{2} \exp \left( - \frac{\Delta_a^2}{4\sigma^2} \right) \quad \text{and} \quad \sigma^2 \leq - \frac{\Delta_a^2}{4 \ln 2\delta}, \tag{6} \tag{7}
\]
where \( \delta = P\left[ \tilde{f}(\mu_a) > \tilde{f}(\mu_*) \right] \leq 0.5. \)
4.2 Pure Exploration

We want to bound the expected simple regret (Equation 2). Given that the expert user selects its preference among the mean estimators after \( T \) episodes, we can decompose

\[
P[a_T = a] = P[a_T = a | \mu_a(T) \in B(\mu_a, \rho_a)] + P[a_T = a | \mu_a(T) \notin B(\mu_a, \rho_a)]
\]

\[
\leq P[a_T = a | \mu_a(T) \notin B(\mu_a, \rho_a)] + P[a_T = a | \mu_a(T) \notin B(\mu_a, \rho_a)]
\]

where the first inequality comes from the definition of the preference radius.

**Lemma 1 (Hoeffding’s Inequality)** Let \( X_1, \ldots, X_N \) be independent i.i.d. random variables with values in \([0, 1]\) and let \( \delta > 0 \). Then we have

\[
P\left[ \left| \mathbb{E}[X] - \frac{1}{N} \sum_{n=1}^{N} X_n \right| \geq \delta \right] \leq e^{-2N \delta^2}.
\]

Now consider the multivariate setting where \( x_1, \ldots, x_N \) are independent i.i.d. random variables with values in \([0, 1]^d\). We deduce that

\[
P\left[ \left| \mathbb{E}[X_i] - \frac{1}{N} \sum_{n=1}^{N} x_{n,i} \right| < \delta \quad \forall i = 1, \ldots, d \right] \geq \left( 1 - e^{-2N \delta^2} \right)^d,
\]

\[
P\left[ \left| \mathbb{E}[X_i] - \frac{1}{N} \sum_{n=1}^{N} x_{n,i} \right| \geq \delta \quad \forall i = 1, \ldots, d \right] \leq de^{-2N \delta^2}.
\]

From the Hoeffding’s inequality, we can bound the probability of estimating each action inside and outside its preference radius as

\[
P[\hat{\mu}_a(T) \in B(\mu_a, \rho_a)] \geq \left( 1 - e^{-2N_a(T)\rho_a^2} \right)^d,
\]

\[
P[\hat{\mu}_a(T) \notin B(\mu_a, \rho_a)] \leq de^{-2N_a(T)\rho_a^2}.
\]

Using Equation 10 in Equation 8, then in Equation 2, we get

\[
P[a_T = a] \leq de^{-2N_a(T)\rho_a^2} + de^{-2N_a(T)\rho_a^2} \Delta_a.
\]

Another way of analyzing the simple regret is through the probability of obtaining proper estimations after \( T \) episodes. Let us consider that a proper estimation after \( T \) episodes is such that the option \( \mu_a(T) \) of each action \( a \) is contained in its ball \( B(\mu_a, \rho_a) \). By definition of the preference radius, we know that the simple regret is null with proper estimations. Using Equation 9, we lower bound the probability of converging to the optimal solution:

\[
P[\mathbb{E}[R_s(T)] = 0] \geq P[\hat{\mu}_a(T) \in B(\mu_a, \rho_a) \forall a \in A] \geq \prod_{a \in A} \left( 1 - e^{-2N_a(T)\rho_a^2} \right)^d.
\]
There exists $\psi > 0$ and $\delta \in [0, 1]$ such that $N_a(T)\rho_a^2 \geq \psi$ for all $a \in A$ and

$$\mathbb{P} [\hat{\mu}_a(T) \in B(\mu_a, \rho_a) \forall a \in A] \geq \left(1 - e^{-2\psi}\right)^{dA} > \delta.$$  

We find out that we have a proper estimation after $T$ episodes with probability higher than $\delta$ if

$$\psi > \frac{\log (1 - \delta \pi)}{-2} \quad \text{and} \quad T \geq \psi \sum_{a \in A} \frac{1}{\rho_a^2},$$

When $T$ is large enough such that $T/A \geq \psi / \min_{a \in A} \rho_a^2$, a uniform allocation across all actions converges to the optimal solution with (high) probability $\delta$. It is therefore a good baseline in the pure exploration setting, as indicated previous results in the single-objective setting [3].

5 Experiments

We randomly generate a 10-arm setting with $d = 2$ objectives, such that the objective space is $\mathcal{X} = [0, 1]^2$. We consider settings where outcomes are sampled from multiple Bernoulli distributions and from multivariate normal distributions with covariance

$$\Sigma_a = \begin{bmatrix} 0.10 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}, \quad \forall a \in A.$$  

Experiments are conducted using the linear preference function

$$f(x) = 0.4x_1 + 0.6x_2, \quad x \in \mathcal{X},$$

and the $\epsilon$-constraint preference function

$$f(x) = \begin{cases} 
  x_2 & \text{if } x_1 \geq 0.5 \\
  0 & \text{otherwise}
\end{cases}, \quad x \in \mathcal{X}.$$  

Figure 5 shows the expected outcomes $\mu_a$ for all actions $a \in A$ and illustrates the preference function. We observe that the optimal action is different for the two preference functions. Table 1 shows the expected outcomes $\mu_a$ for all actions $a \in A$ along with the associated value and gap given the preference function. Notice that the given linear preference function leads to smaller gap on that setting.

We consider Thompson sampling (described in Section 4.1) for recommending actions in the bandit and pure exploration settings given the user interactive scheme depicted by Algorithm 1. Options $\{\theta_a(t)\}_{a \in A}$ are sampled from multiple Beta distributions in the multiple Bernoulli setting (Example 4) and from a multivariate normal distribution in the multivariate normal setting (Example 5), respectively with non-informative priors $\alpha_0 = \beta_0 = 1$ and

$$\mu_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$  

12
Figure 5: Expected outcomes for optimal (white) and suboptimal (black) actions. The dotted line shows the preference function on the left and the $\epsilon_1$ constraint on the right.

Table 1: Expected outcomes with values and gap for both preference functions. The expected outcome for the optimal action is shown in bold.

| $\mu$           | $f(\mu)$ | $\Delta$ |
|-----------------|----------|----------|
| Linear $\epsilon$-constraint | Linear $\epsilon$-constraint |
| (0.56,0.46)     | 0.50     | 0.17     | 0.26 |
| (0.75,0.26)     | 0.46     | 0.26     | 0.21 | 0.46 |
| (0.34,0.79)     | 0.61     | 0.00     | 0.06 | 0.72 |
| (0.67,0.50)     | 0.56     | 0.50     | 0.11 | 0.22 |
| (0.70,0.42)     | 0.54     | 0.42     | 0.13 | 0.29 |
| (0.54,0.72)     | 0.65     | **0.72** | 0.02 | 0.00 |
| (0.49,0.62)     | 0.57     | 0.00     | 0.10 | 0.72 |
| (0.13,0.84)     | 0.56     | 0.00     | 0.11 | 0.72 |
| (0.78,0.60)     | **0.67** | 0.60     | 0.00 | 0.12 |
| (0.63,0.44)     | 0.51     | 0.44     | 0.16 | 0.28 |

5.1 Bandits

We evaluate the performance of Thompson sampling in the bandits setting, where the goal is to minimize the cumulative regret (Equation 1). We consider the case where the expert user feedback may be noisy such that the preference $\hat{O}(t) = \arg\max_{a \in A} \hat{f}(\theta_a(t))$, where $\hat{f}(\cdot) \sim N(f(\cdot),\sigma^2)$, for $\sigma^2 \in \{0,10^{-4},10^{-3}\}$. For all the experiments presented in this subsection, we perform 100 repetitions over 10,000 episodes. Repetitions have been made such that the outcomes $z_a(t)$ are the same for all tested approaches on the same repetition. Therefore, we can compare the performance of different approaches on the same repetition.

Figure 6 shows the cumulative regret of Thompson sampling averaged over the repetitions for both outcome distributions and preference functions in the bandit setting. We observe that the cumulative regret grows logarithmically with the number of episodes, as expected from previous results in the single-objective setting [1, 7], unless the noise is too high. Recall Equation 6 that
bounds the probability that the expert user prefers the suboptimal action given that the expectations $\mu_a$ are well estimated and that the sampled options $\theta_a = \mu_a$. Consider the suboptimal action $a = \arg\min_{a \in \mathcal{A} \setminus \{\star\}} \Delta_a$, that is the closest suboptimal action from the optimal action $\star$. For the linear setting, from Table 1, for the suboptimal action $a = (0.54, 0.72)$ with the smallest gap $\Delta_a = 0.02$, we have $P[\tilde{f}(\mu_a) > \tilde{f}(\mu_\star)] \leq 0.18$, $P[\tilde{f}(\mu_a) > \tilde{f}(\mu_\star)] \leq 0.45$ using $\sigma^2 = 10^{-4}$, $\sigma^2 = 10^{-3}$ respectively in Equation 6. For the $\epsilon$-constraint setting, also from Table 1, for the suboptimal action $a = (0.78, 0.60)$ with the smallest gap $\Delta_a = 0.12$, we have $P[\tilde{f}(\mu_a) > \tilde{f}(\mu_\star)] \simeq 0$, $P[\tilde{f}(\mu_a) > \tilde{f}(\mu_\star)] \leq 0.01$ using $\sigma^2 = 10^{-4}$, $\sigma^2 = 10^{-3}$ respectively in Equation 6. This highlights the impact of the closer gap between $a$ and $\star$ in the noise tolerance and explains why the linear preference function is less tolerant to noise than the $\epsilon$-constraint on these specific settings.

A common issue in user-guided systems is the need for user assistance. Asking an expert user to provide feedback on each episode may not be realistic. More specifically, we consider the case where the expert user can only provide feedback every $E$ episodes. We evaluate the following behaviors to deal with expert user delayed feedback. Let $\tau$ denote the last time preference feedback was provided by the expert user.
Table 2: Ratio of repetitions for which each approach is the best in the noiseless setting, after 10,000 episodes. Boldface and upper line respectively indicate that \( p < 5\% \) and \( p < 10\% \) given a Welch’s t-test against the case without delay.

| Multi-Bernoulli          | No delay | Replicate | Linear | Squared Euclidean | Chebyshev |
|--------------------------|----------|-----------|--------|-------------------|-----------|
| Linear                   | 0.12     | 0.06      | 0.40   | 0.21              | \textbf{0.21} |
| \( \epsilon \)-constraint | 0.23     | 0.09      | \textbf{0.00} | 0.34              | 0.34      |
| Multivariate normal      |          |           |        |                   |           |
| Linear                   | 0.14     | 0.04      | \textbf{0.37} | 0.26              | \textbf{0.19} |
| \( \epsilon \)-constraint | 0.32     | 0.16      | \textbf{0.00} | 0.28              | 0.24      |

**Replicate** This approach considers \( O(t) = O(\tau) \).

**Linear** This approach assumes that the expert user preference function is a linear function. It computes a linear function \( \hat{f} \) such that \( O(\tau) = \arg\max_{a \in A} \hat{f}(\theta_a(\tau)) \) and considers that \( O(t) = \arg\max_{a \in A} \hat{f}(\theta_a(t)) \). Here we consider the linear function \( \hat{f} \) that goes through \( \theta_a(\tau) \).

**Squared Euclidean** This approach considers \( O(t) = \arg\min_{a \in A} d(\theta_a(t), \theta_a(\tau)) \), where \( d \) is the squared Euclidean distance.

**Chebyshev** This approach considers \( O(t) = \arg\min_{a \in A} d(\theta_a(t), \theta_a(\tau)) \), where \( d \) is the Chebyshev distance.

We perform experiments where expert user feedback is provided every 10 episodes. This corresponds to \( \tau = t - 10 \) for \( t \geq 10 \), otherwise \( \tau = 0 \).

Figures 7 to 10 show the cumulative regret of Thompson sampling averaged over the repetitions for feedback delay addressed using the presented approaches. Table 2 shows the ratio of repetitions for which each approach gathers the less cumulative regret in the noiseless setting. They compare the cumulative regret after 10,000 episodes from Figures 6 to 10 in the noiseless setting. We notice that delayed feedback addressed using the linear assumption technique performs the best when the true preference function is also linear (as observed in Figures 7(b) and 9(b)). In fact, it performs even better that the setting without delay (Table 2, lines 2 and 4). This is suprising given that the delayed setting uses only 10\% of expert user feedback. This can be explained by the fact that the linear assumption technique considered here goes through the last selected option. Consider that \( a(\tau) = * \) and \( \theta_a(\tau)(\tau) = \mu_a(\tau) \), we would have

\[
\hat{f}(x) = 0.57x_1 + 0.43x_2
\]

for the setting with linear user preference function shown by Figure 5(a). Figure 11(a) shows the resulting setting corresponding to \( \hat{f} \). Given this function, the smallest suboptimal gap would be of 0.08, which is higher than the smallest suboptimal gap of 0.02 (Table 1) with regards to the true preference function. Therefore, the closest sub optimal action is easier to distinguish in the resulting setting, allowing to obtain a lower cumulative regret in the delayed settings compared with the associated non-delayed settings. Increasing the suboptimal gap also makes the delayed settings more robust to noise, as discussed previously. On the other hand, assuming a linear preference function can lead to desastrous performance when the underlying preference function is not linear.
Figure 7: Cumulative regret over episodes with delayed feedback for multiple Bernoulli outcome distributions and linear preference function.
Figure 8: Cumulative regret over episodes with delayed feedback for multiple Bernoulli outcome distributions and $\epsilon$-constraint preference functions.
Figure 9: Cumulative regret over episodes with delayed feedback for multivariate normal outcome distributions and linear preference function.
Figure 10: Cumulative regret over episodes with delayed feedback for multivariate normal outcome distributions and $\epsilon$-constraint preference functions.
Figure 11: Optimal (white) and suboptimal (black) actions given the preference function $\hat{f}$.

(as observed in Figures 8(b) and 10(b)). Consider that $a(\tau) = \star$ and $\theta_{a(\tau)}(\tau) = \mu_{a(\tau)}$, we would have

$$\hat{f}(\mathbf{x}) = 0.43x_1 + 0.57x_2$$

for the setting with $\epsilon$-constraint user preference function shown by Figure 5(b). Figure 11(b) shows the resulting setting corresponding to $\hat{f}$. Notice that the optimal action in the resulting setting is not the same as the optimal action given the true, $\epsilon$-constraint, preference function. Therefore, Thompson sampling converges to the wrong action. This explains the linear regret shown in Figures 8(b) and 10(b), and the fact that the linear assumption technique never beats the other approaches, as stated in Table 2.

Finally, an interesting result is that the technique based on the closest option given a Chebyshev distance in the delayed feedback setting performs as good and sometimes better than the setting without delay. This is very interesting given that it requires, in the current case, only 10% of expert user feedback. This would definitely need further investigation.

5.2 Pure Exploration

We evaluate the performance of Thompson sampling in the pure exploration setting, where the goal is to minimize the simple regret (Equation 2). The uniform sampling strategy is considered as the baseline. We perform 1000 repetitions over 10,000 episodes.

Figure 12 shows the expected simple regret that would be obtained given that learning was conducted over different number of episodes, using Thompson sampling and uniform sampling, averaged over the repetitions for both outcome distributions and preference functions in the pure exploration setting. Recall from Table 1 that the smallest gap are 0.02 and 0.12 on linear and $\epsilon$-constraint settings, respectively, which corresponds to the simple regret incurred when recomending the best suboptimal action. Therefore, we observe that both approaches have reached convergence, on average, after 10,000 episodes.

Figure 13 shows explicitly the ratio of repetitions that converged to the optimal action after each number of episodes in the different settings. We observe that uniform sampling beats Thompson sampling.
Figure 12: Simple regret after episodes for tested outcome distributions and preference functions.
Figure 13: Ratio of repetitions that converged to the optimal action after episodes for tested outcome distributions and preference functions.
sampling in the setting with the $\epsilon$-constraint preference function, given enough episodes. These results corroborate previous findings [3] that optimal algorithms for the bandit setting are not optimal in the pure exploration setting. However, although Thompson sampling is not optimal in pure exploration, it can provide a lower simple regret than the uniform allocation given that the number of trials is limited for the difficulty of the setting.

6 Conclusion

In this work, we have addressed the online MOO problem with user-guided feedback. We have formulated the setting as a multi-objective bandit and we have introduced the concept of preference radius. These radii characterize the difficulty of a multi-objective setting through the robustness of the user decision function to the quality of estimations available. We have shown how this measure relates to the gap between the optimal action and the recommended action by a learning algorithm. We have shown that regret analysis in the multi-objective bandits setting can be brought back to a standard single-objective analysis using the preference radius. We have also provided a regret analysis in the pure exploration multi-objective setting using the preference radius and shown that the uniform allocation strategy is a good baseline in this setting. We have provided experimental results where we have considered two important issues that could be raised in any user-guided feedback: the impact of noise in the preference function and the impact of delayed feedback. We have linked the empirical results on the effects of noise to an analysis, which allowed us to explain the observed results. Results have suggested that it might be beneficial to delay the expert user feedback given that a proper interpolation technique is considered between feedbacks. This is an interesting finding as delayed feedback addresses a major problem of user-guided settings, that is the risk of annoying the user. Finally, experiments on the pure exploration setting have shown that Thompson sampling constitutes a fairly good baseline in this setting.

Future work includes a deeper investigation of approaches to deal with delayed feedback. It would also be good to validate the current results on a real world application.

Being able to evaluate multi-objective algorithms based on the unknown user decision function is very important since this function is often unavailable, except through the user. It is a novel way of approaching the MOO problem, which is otherwise usually tackled under a priori or a posteriori decision making methods. This paper provides tools for designing and analyzing algorithms under this new user-guided optimization scheme.

7 Acknowledgements

This work was supported through funding from NSERC (Canada). We also thank Julien-Charles Lévesque for insightful comments

References

[1] S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson Sampling. In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), pages 99–107, 2013.
[2] V. J. Bowman Jr. On the relationship of the Tchebycheff norm and the efficient frontier of multiple-criteria objectives. In *Multiple criteria decision making*, pages 76–86. 1976.

[3] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, pages 23–37, 2009.

[4] Marco Chiani, Davide Dardari, and Marvin K Simon. New exponential bounds and approximations for the computation of error probability in fading channels. *IEEE Transactions on Wireless Communications*, 2(4):840–845, 2003.

[5] C. A. C. Coello, G. B. Lamont, D. A. Van Veldhuizen, D. E. Goldberg, and J. R. Koza. *Evolutionary Algorithms for Solving Multi-Objective Problems*. 2nd edition, 2007. ISBN 9780387310299.

[6] M. M. Drugan and A. Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2013.

[7] N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

[8] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary computation*, 10(3):263–82, 2002.

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

[10] M. Zuluaga, G. Sergent, A. Krause, and M. Püschel. Active learning for multi-objective optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 462–470, 2013.