2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

The Scalarized Multi-Objective Multi-Armed Bandit Problem: An Empirical Study of its Exploration vs. Exploitation Tradeoff

Saba Q. Yahyaa, Madalina M. Drugan and Bernard Manderick

***Abstract*— The multi-armed bandit (MAB) problem is the simplest sequential decision process with stochastic rewards where an agent chooses repeatedly from different arms to identify as soon as possible the optimal arm, i.e. the one of the highest mean reward. Both the knowledge gradient (KG) policy and the upper confidence bound (UCB) policy work well in practice for the MAB-problem because of a good balance between exploitation and exploration while choosing arms.**

In case of the multi-objective MAB (or MOMAB)-problem, arms generate a vector of rewards, one per arm, instead of a single scalar reward. In this paper, we extend the KG-policy to address multi-objective problems using scalarization functions that transform reward vectors into single scalar reward. We consider different scalarization functions and we call the corresponding class of algorithms *scalarized KG*. We compare the resulting algorithms with the corresponding variants of the multi-objective UCB1-policy (MO-UCB1) on a number of MOMAB-problems where the reward vectors are drawn from a multivariate normal distribution. We compare experimentally the exploration versus exploitation trade-off and we conclude that scalarized-KG outperforms MO-UCB1 on these test problems.

I. INTRODUCTION

THE MULTI-ARMED BANDIT (MAB) is a sequential decision problem where an agent tries to optimize its decisions while improving its knowledge concerning the arms among which it has to choose. At each time step t, the agent pulls one arm from the set A of available arms and receives a reward as feedback signal. That reward is independent from the past rewards of the selected arm and all the other arms. The rewards from each arm i are drawn from a stationary probability distribution, e.g. the normal distribution $N(\mu_i, \sigma_i^2)$ with mean μ_i and variance σ_i^2 and we assume that these parameters are unknown to the agent. By pulling an arm i, the agent improves its estimates $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ of the true mean μ_i and variance σ_i^2, respectively.

The goal of the agent is to minimize the *total expected regret* $R_L = L\mu^* - \sum_{t=1}^{L} \mu(t)$ of not pulling the best arm i^* at all time steps L. In this expression, i^* is the arm with the maximum mean $\mu^* = \max_{i=1,\cdots,|A|} \mu_i$ and $\mu_i(t)$ is the mean of the selected arm i at time step t.

In the MAB-problem, at each time step t, the agent either selects the arm with the highest estimated mean $\hat{\mu}^*$ (exploitation of the greedy arm) or selects one of the other arms in order to improve its corresponding estimate (exploration of the other arms). And, the agent has to find a proper *trade-off between exploitation and exploration* [1] to minimize the total expected regret R_L. To find a good trade-off, [2] compares several action selection policies on the MAB-problem and shows that Knowledge Gradient (or KG)-policy [3] outperforms other MAB-policies including the Upper Confidence Bound (UCB1) policy [4]. UCB1 and KG are similar in the way they trade-off between exploitation and exploration, both add an exploration bonus to the estimated mean of each arm i and select the arm that has the highest combined value of estimated mean and exploration bonus. But in case of UCB1, the exploration bonus of arm i requires only knowledge about that arm itself, while in case of KG it also requires knowledge about the other arms.

The Multi-Objective Multi-Armed Bandit (MOMAB)-problem has a set of Pareto optimal arms (Pareto front) that all can be considered best since they are all non-dominated [6], [9]. The agent trades-off conflicting objectives of the mean reward vectors as follows: *exploration* equals finding the Pareto optimal arms while *exploitation* equals selecting fairly among these optimal arms.

A popular way to identify the Pareto optimal arm set is by using scalarized functions [5]. Scalarized functions transform the multi-objective space into a single-objective space, i.e. the mean reward vectors are transformed into scalar rewards. We consider linear and non-linear scalarization. Linear scalarization is simple and intuitive but cannot find all the Pareto optimal arms on a non-convex Pareto front. The Chebyshev scalarization function, which is a non-linear, has an extra parameter that can be tuned to find all arms in the Pareto front set, even if it is non-convex. Recently, [6] introduced a multi-objective version of UCB1, the multi-objective UCB1 (or MO-UCB1) and [9] adapted the KG policy for the MOMAB-problem.

In this paper, we introduce the *multi-objective knowledge gradient framework* and the *scalarized multi-objective knowledge gradient function* (scalarized-KG). And, we study the *exploration vs exploitation trade-off for the MOMAB* by comparing empirically scalarized-KG with multi-objective UCB1. We consider three variants of scalarized-KG: 1) linear scalarized-KG across arms and 2) across objectives, and 3) non-linear scalarized-KG. Scalarized-KG converts the multi-objective space into a single objective space and then adds an exploration bound in order to trade-off between exploration and exploitation.

The rest of the paper is organized as follows. Section II presents background information on the algorithms and the notation used. Section III introduces the MOMAB-problem. Section IV introduces the three variants of the scalarized KG functions mentioned above. Section V describes the

Saba Q. Yahyaa, Madalina M. Drugan and Bernard Manderick are with the Artificial Intelligent Lab., The Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium (email: {syahyaa, mdrugan,bmanderi}@vub.ac.be).

This work was supported by Vrije Universiteit Brussel

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

experimental set up followed by the experimental results. Finally, Section VI concludes the paper and discusses future work.

II. Background

We consider MOMAB-problems with $|A| \geq 2$ arms and with D objectives per arm. The mean reward vector of arm $i, i = 1 \leq i \leq |A|$ is represented as $\boldsymbol{\mu}_i = (\mu_i^1, \cdots, \mu_i^D)^T$, where T is the transpose. When the objectives are conflicting with one another then the mean reward component μ_i^d of arm i corresponding with objective $d, d \in D$, can be better than the component μ_j^d of another arm j but worse if we compare the components for another objective d': $\mu_i^d > \mu_j^d$ but $\mu_i^{d'} < \mu_j^{d'}$ for objectives d and d', respectively. If there is an arm k for which at least one component μ_k^d corresponding with one of the objectives d is strictly greater than the corresponding components μ_i^d of all other arms i, then that arm k is *Pareto optimal* and the set of all Pareto optimal arms is the *Pareto front* A^*.

A. Scalarized Functions

We consider scalarization functions that take the weighted sum of the components of the mean reward vector $\boldsymbol{\mu}$ and return a scalar value [5]. We discuss linear and Chebyshev scalarizations.

A *linear scalarization* $f_{\boldsymbol{w}}$ with predefined weights $\boldsymbol{w} = (w^1, \cdots, w^D)$ such that $\sum_{d=1}^{D} w^d = 1$ assigns to each component μ_i^d of the mean vector $\boldsymbol{\mu}_i$ of an arm i a weight w^d and returns the weighted sum of the means:

$$f_{\boldsymbol{w}}(\boldsymbol{\mu}_i) = w^1\mu_i^1 + \cdots + w^D\mu_i^D \tag{1}$$

Apart from the weights $\boldsymbol{w}$, the *Chebyshev scalarization* $f_{\boldsymbol{w}}$ also takes into account a D-dimensional reference point $\boldsymbol{z} = (z^1, \cdots, z^D)^T$ which has to be dominated by all elements in the Pareto front. Chebyshev scalarization for maximization problem is as follows [6]:

$$\begin{aligned} f_{\boldsymbol{w}}(\boldsymbol{\mu}_i) &= \min_{1 \leq d \leq D} w^d(\mu_i^d - z^d) \quad && \forall i \\ z^d &= \min_{1 \leq i \leq A} \mu_i^d - \epsilon && \forall d \end{aligned} \tag{2}$$

where ϵ, $\epsilon > 0$ is a small value. As a consequence, the reference point $\boldsymbol{z}$ is dominated by all the elements in the Pareto front. The parameter ϵ can be varied in order to find all the Pareto optimal arms in A^* [7]. Once the multi-objective MAB-problem is converted into a single-objective problem, the scalarization function $f_{\boldsymbol{w}}$ (linear or Chebyshev scalarization functions) select its arm $i^*_{f_{\boldsymbol{w}}}$ that maximizes the function $f_{\boldsymbol{w}}$:

$$i^*_{f_{\boldsymbol{w}}} = \underset{1 \leq i \leq A}{\operatorname{argmax}}\, f_{\boldsymbol{w}}(\boldsymbol{\mu}_i) \tag{3}$$

Scalarized functions convert MOMAB-problems into corresponding single objective MAB-problems that have in general a unique optimal arm as solution. In order to find all Pareto optimal arms in the Pareto front set A^*, we need a set of scalarization functions $f^s_{\boldsymbol{w}}, s = 1, \cdots, S$ that generates variety of elements belonging to the Pareto front set, like in multi-objective optimization the scalarization functions are uniformly random spread in the weighted space. Each $f^s_{\boldsymbol{w}}$ has the corresponding predefined weights $\boldsymbol{w}^s$.

B. Regret Metrics

To measure the performance of scalarized functions $f_{\boldsymbol{w}}$, the authors of [6] have proposed two regret metrics.

The *scalarized regret metric* measures the distance between the maximum of a scalarized function and the scalarized mean vector of the arm chosen at time step t. The scalarized regret $R_s(f_{\boldsymbol{w}})(t)$ for a scalarized function $f_{\boldsymbol{w}}$ at time step t is the difference between the maximum for that function $f_{\boldsymbol{w}}$ and the scalarized mean vector for the arm k chosen at time step t by the scalarized function $f_{\boldsymbol{w}}$

$$R_s(f_{\boldsymbol{w}})(t) = \max_{1 \leq i \leq A} f_{\boldsymbol{w}}(\boldsymbol{\mu}_i) - f_{\boldsymbol{w}}(\boldsymbol{\mu}_k)(t) \tag{4}$$

The unfairness regret metric for the MOMAB takes the mean rewards of all the optimal arms into account. It looks at how many times an optimal arm is chosen compared with total times of optimal arms are chosen so far. Let $|A^*|$ be the number of optimal arms. Let $N_{i^*}(t)$ be the number of times optimal arm i^* has been selected and $N_{|A^*|}(t)$ be the number of times optimal arms in the Pareto front set A^* have been selected till time step t using the scalarization function $f_{\boldsymbol{w}}$, then the unfairness regret R_u is defined as:

$$R_u(f_{\boldsymbol{w}})(t) = \frac{1}{|A^*|} \sum_{i^* \in A^*} (N_{i^*}(t) - N_{|A^*|}(t))^2 \tag{5}$$

C. UCB1 in MOMABs

In the multi-objective multi-armed bandit MOMAB problem [6] extends the UCB1-policy to scalarized multi-objective UCB1 and it shows that this new policy can find all Pareto optimal arms in the Pareto front. Scalarized UCB1 converts a multi-objective MAB-problem into a corresponding single objective MAB-problem and then uses UCB1 [4] to trade-off between exploration and exploitation. It adds an upper confidence bound to the pulled arm i under the scalarized function (linear or Chebyshev scalarized function) $f^s_{\boldsymbol{w}}$ with scalarization s that has a predefined set of weight $\boldsymbol{w}^s$. The upper confidence bound depends on the number of times the scalarized function $f^s_{\boldsymbol{w}}$ has been selected, N^s and on the number of times the arm i has been pulled N_i^s under the scalarized function s. Firstly, the scalarized UCB1 plays each arm once and estimates the mean vector of each arm, $\hat{\mu}_i, i = 1, \cdots, |A|$. At each time step t, it pulls the optimal arm i^*_{UCB1} as follows:

$$i^*_{UCB1} = \underset{1 \leq i \leq A}{\operatorname{argmax}} \left(f^s_{\boldsymbol{w}}(\hat{\boldsymbol{\mu}}_i) + \sqrt{\frac{2\ln(N^s)}{N_i^s}} \right) \tag{6}$$

where $f^s_{\boldsymbol{w}}$ a scalarization function as before. In this paper, we use scalarized multi-objective UCB1 in the MOMAB problem with normal distributions.

D. Knowledge Gradient Policy

In the single-objective MAB problem, KG policy is an index policy that determines for each arm i the index V_i^{KG} as follows [3]:

$$V_i^{KG} = \hat{\tilde{\sigma}}_i * g\left(-\left|\frac{\hat{\mu}_i - \max\limits_{j\neq i, j\in|A|}\hat{\mu}_j}{\hat{\tilde{\sigma}}_i}\right|\right) \tag{7}$$

where $\hat{\tilde{\sigma}}_i = {}^{\hat{\sigma}_i}/_{N_i}$ is the root mean square error (RMSE) of the estimated mean $\hat{\mu}_i$ of arm i. The function $g(\zeta) = \zeta\Phi(\zeta) + \phi(\zeta)$ where $\zeta = -|{}^{(\hat{\mu}_i - \max\limits_{j\neq i, j\in|A|}\hat{\mu}_j)}/_{\hat{\tilde{\sigma}}_i}|$, $\phi(\zeta) = {}^1/_{\sqrt{2\pi}}\exp({}^{-\zeta}/_2)$ is the density and $\Phi(\zeta) = \int_{-\infty}^{\zeta}\phi(\zeta')d\zeta'$ is the cumulative distribution of the standard normal distribution $N(0,1)$. KG chooses the arm i with the largest V_i^{KG} and it prefers those arms about which comparatively little is known. These arms are the ones whose distributions around the estimate mean $\hat{\mu}_i$ have larger estimated standard deviation $\hat{\sigma}_i$. Thus, KG prefers an arm i over its alternatives if its confidence in the estimate mean $\hat{\mu}_i$ is low. The KG policy trades-off between exploration and exploitation by selecting the arm i^*_{KG} as follows:

$$i^*_{KG} = \underset{i\in|A|}{\operatorname{argmax}}\left(\hat{\mu}_i + (L-t)V_i^{KG}\right) \tag{8}$$

where L is the horizon of experiment, i.e. the total number of times an agent can play. For more details about KG policy, we refer to [3]. In [2], it is shown that the KG policy outperforms other policies on single-objective MAB problems in terms of the collected average reward and the average frequency of optimal selection performances. Moreover, the KG policy does not have any parameter to be tuned. For these reasons, we propose scalarized knowledge gradient (scalarized-KG) functions which make use of the estimated mean $\hat{\mu}$ and variance $\hat{\sigma}^2$. Scalarized-KG functions either convert the multi-dimension to one-dimension environment and then trades-off between exploration and exploitation, or vice versa.

III. Multi-Objective Knowledge Gradient Framework

In the MOMAB problem with normal distribution, at each time step t, the agent selects one arm i and receives a reward vector. The reward vector is drawn from a normal distribution $N(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$, where $\boldsymbol{\mu}_i = (\mu_i^1, \cdots, \mu_i^D)^T$ is the mean vector and $\boldsymbol{\sigma}_i^2 = (\sigma_i^{2,1}, \cdots, \mu_i^{2,D})^T$ is the diagonal covariance matrix of arm i since the reward distributions corresponding with different arms are assumed to be independent. These parameters are unknown to the agent. But by drawing arm i at time step t, the agent can update its estimates $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\sigma}}_i^2$ in each dimension d as follows [8]:

$$\begin{cases} \hat{\mu}^d_{t+1} = (1 - \frac{1}{N_{i^{t+1}}})\,\hat{\mu}^d_t + \frac{1}{N_{i^{t+1}}} r^d_{t+1}, \\ \hat{\sigma}^{2,d}_{t+1} = \frac{N_{i^{t+1}}-2}{N_{i^{t+1}}-1}\,\hat{\sigma}^{2,d}_t + \frac{1}{N_{i^{t+1}}}(r^d_{t+1} - \hat{\mu}^d_t)^2, \\ N_{i^{t+1}} = N_{i^t} + 1 \end{cases} \tag{9}$$

where $N_{i^{t+1}}$ is the updated number of times arm i has been selected, r^d_{t+1} is the collected reward from arm i in the dimension d and $\hat{\mu}^d_{t+1}$, and $\hat{\sigma}^{2,d}_{t+1}$ are the updated estimated mean and covariance of arm i for dimension d, respectively.

A. The Scalarized Multi-Objectieve KG Bandits

The pseudocode of the scalarized multi-objective, multi-armed bandit MOMAB problems is given in Fig. 1.

1. Input: length of trajectory L;type of scalarized function $y_{\boldsymbol{w}}$;set of scalarized function $S = (y^1_{\boldsymbol{w}}, \cdots, y^S_{\boldsymbol{w}})$;reward $r^d \sim N(\mu, \sigma_r^2)$.
2. Initialize: **For** $s = 1$ **to** S
   plays each arm $Initial$ steps;
   observe $(\boldsymbol{r}_i)^s$;
   update: $N^s \leftarrow N^s + 1$;
   $N^s_i \leftarrow N^s_i + 1$;
   $(\hat{\boldsymbol{\mu}}_i)^s$, $(\hat{\boldsymbol{\sigma}}_i)^s$
   End
3. **Repeat**
4. Select:a function s uniformly, randomly
5. Select:the optimal arm i^* that maximizes the scalarized function $y^s_{\boldsymbol{w}}$
6. Observe:reward vector $\boldsymbol{r}_{i^*}, \boldsymbol{r}_{i^*} = [r^1_{i^*}, \cdots, r^D_{i^*}]^T$
7. Update:the estimated mean vector $\hat{\boldsymbol{\mu}}_{i^*}$; the estimated standard deviation vector $\hat{\boldsymbol{\sigma}}_{i^*}$; $N^s_{i^*} \leftarrow N^s_{i^*} + 1$; $N^s \leftarrow N^s + 1$
8. Compute:unfairness and scalarized regret
9. **Until** L
10. Output:Unfairness and scalarized regret

Fig. 1. Algorithm:(Scalarized multi-objective function).

Given the type of the scalarized function $y_{\boldsymbol{w}}$, ($y_{\boldsymbol{w}}$ is either UCB1 scalarized functions, Section II-C or KG scalarized functions IV and the scalarized function set $(y^1_{\boldsymbol{w}}, \cdots, y^S_{\boldsymbol{w}})$ where each scalarized function $y^s_{\boldsymbol{w}}$ has different weight set, $\boldsymbol{w}^s = (w^{1,s}, \cdots, w^{D,s})$, $\sum_{d=1}^D w^{d,s} = 1$.

The algorithm in Fig. 1 plays each arm of each scalarized function $y^s_{\boldsymbol{w}}$, $Initial$ plays (step: 2). N^s is the number of times the scalarized function $y^s_{\boldsymbol{w}}$ is pulled and N^s_i is the number of times the arm i under the scalarized function $y^s_{\boldsymbol{w}}$ is pulled. $(\boldsymbol{r}_i)^s$ is the reward vector of the pulled arm i which is drawn from a normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\sigma}_r^2)$ where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\sigma}_r$ is the standard deviation vector of the reward. $(\hat{\boldsymbol{\mu}}_i)^s$ and $(\hat{\boldsymbol{\sigma}}_i)^s$ are the estimated mean and estimated standard deviation vectors of the arm i under the scalarized function s, respectively. After initial playing, the algorithm chooses *randomly at uniform* one of the scalarized function (step: 4) and selects the optimal arm i^* that maximizes the type of this scalarized function (step: 5). The algorithm simulates the selected arm i^*, and updates N^s_i, N^s, $(\hat{\boldsymbol{\mu}}_i)^s$ and $(\hat{\boldsymbol{\sigma}}_i)^s$ (step: 7). This procedure is repeated until the end of playing L steps which is the horizon of an experiment.

Note that the proposed algorithm is an adapted version from [6], but here the algorithm can be applied to both KG and UCB1 policies with normal reward distribution.

IV. Scalarized Knowledge Gradient

In this section, we introduce three instances of the scalarized knowledge gradient functions.

A. Linear Scalarized-KG across Arms

Linear scalarized-KG across arms (LS1-KG) converts immediately the multi-objective estimated mean $\hat{\boldsymbol{\mu}}_i$ and estimated variance $\hat{\boldsymbol{\sigma}}_i^2$ of each arm to one-dimension, then computes the corresponding exploration bound ExpB_i to trade-off between exploration and exploitation (*trading-off after scalarization*). We use $\hat{\boldsymbol{\sigma}}_i^2$ to refer to the estimated variance vector of arm i. At each time step t, LS1-KG weighs both the estimated mean vector, i.e. $([\hat{\mu}_i^1, \cdots, \hat{\mu}_i^D]^T)$ and estimated variance vector, i.e. $([\hat{\sigma}_i^{2,1}, \cdots, \hat{\sigma}_i^{2,D}]^T)$ of each arm i, converts the multi-dimension vectors to one-dimension values by summing the elements of each vector. Thus, we have one-dimension MAB problem. KG calculates for each arm, an exploration bound which depends on all other arms and selects the arm that has the maximum estimated mean plus exploration bound. LS1-KG is as follows:

$$\begin{cases} \widetilde{\mu}_i = f_{\boldsymbol{w}}^s(\hat{\mu}_i) = w^1\hat{\mu}_i^1 + \cdots + w^D\hat{\mu}_i^D & \forall_i, \\ \widetilde{\sigma}_i^2 = f_{\boldsymbol{w}}^s(\hat{\sigma}_i^2) = w^1\hat{\sigma}_i^{2,1} + \cdots + w^D\hat{\sigma}_i^{2,D} & \forall_i, \\ \widetilde{\widetilde{\sigma}}_i^2 = \widetilde{\sigma}_i^2/N_i & \forall_i, \\ \widetilde{v}_i = \widetilde{\widetilde{\sigma}}_i \; g\left(-|\frac{\widetilde{\mu_i} - \max\limits_{j\neq i,\; j\in A}\widetilde{\mu_j}}{\widetilde{\widetilde{\sigma}}_i}|\right) & \forall_i, \end{cases} \tag{10}$$

where $f_{\boldsymbol{w}}^s$ is a linear scalarized function that has a predefined set of weight $\boldsymbol{w}^s$. $\widetilde{\mu}_i$ and $\widetilde{\sigma}_i^2$ are the modified estimated mean and variance of an arm i, respectively which are one-dimension values. $\widetilde{\widetilde{\sigma}}_i^2$ is the RMSE of an arm i. $\widetilde{v}_i$ is the modified KG index of an arm i. The function $g(\zeta) = \zeta\Phi(\zeta) + \phi(\zeta)$ where Φ and ϕ are the cumulative distribution and the density of the standard normal density, respectively. LS1-KG selects its optimal arm $i^*_{LS_1-KG}$ according to:

$$\begin{aligned} i^*_{LS_1-KG} &= \underset{i=1,\cdots,|A|}{\operatorname{argmax}} f^s_{LS1KG} \\ &= \underset{i=1,\cdots,|A|}{\operatorname{argmax}} \left(\widetilde{\mu}_i + \widetilde{\mathrm{ExpB}_i}\right) \\ &= \underset{i=1,\cdots,|A|}{\operatorname{argmax}} \; (\widetilde{\mu}_i + (L-t) * |A|D * \widetilde{v}_i) \end{aligned} \tag{11}$$

where f^s_{LS1-KG} is a linear scalarized-KG across arms with scalarization s, $\widetilde{\mathrm{ExpB}_i}$ is the modified exploration bound of arm i, $|A|$ is the number of arms and D is the number of dimensions.

B. Linear Scalarized-KG across Dimensions

Linear scalarized-KG across dimensions (LS2-KG) computes the exploration bound vector $\mathbf{ExpB}_i$ for each arm, i.e. $\mathbf{ExpB}_i = [\mathrm{ExpB}_i^1, \cdots, \mathrm{ExpB}_i^D]$, adds the $\mathbf{ExpB}_i$ to the corresponding estimated mean vector $\hat{\boldsymbol{\mu}}_i$ to trade-off between exploration and exploitation, then converts the multi-objective problem to one-objective (*scalarization after trading-off*). At each time step t, LS2-KG computes exploration bounds for all dimensions of each arm, sums the estimated mean in each dimension with its corresponding exploration bound, weighs each dimension, then converts the multi-dimension to one-dimension value by taking the summation over each vector of each arm. LS2-KG is as follows:

$$f^s_{LS2-KG}(\hat{\mu}_i) = w^1(\hat{\mu}_i^1 + \mathrm{ExpB}_i^1) + \cdots + w^D(\hat{\mu}_i^D + \mathrm{ExpB}_i^D) \forall_i \tag{12}$$

$$\text{where, } \begin{cases} \mathrm{ExpB}_i^d = (L-t) * |A|D * v_i^d & \forall_{d\in D}, \\ v_i^d = \hat{\sigma}_i^d \; g\left(-|\frac{\hat{\mu}_i^d - \max\limits_{j\neq i,\; j\in A}\hat{\mu}_j^d}{\hat{\sigma}_i^d}|\right) & \forall_{d\in D}, \end{cases} \tag{13}$$

f^s_{LS2-KG} is a linear scalarized-KG across dimensions with scalarization s, v_i^d is the index of arm i for dimension d, $\hat{\mu}_i^d$ is the estimated mean for dimension d of arm i, $\hat{\sigma}_i^d$ is the root mean square error of arm i for dimension d, and $ExpB_i^d$ is the exploration bound of arm i for dimension d. LS2-KG selects its optimal arm $i^*_{LS_2-KG}$ that has maximum $f^s_{LS2-KG}(\hat{\mu}_i)$ as follows:

$$i^*_{LS_2-KG} = \underset{i=1,\cdots,|A|}{\operatorname{argmax}} f^s_{LS2-KG}(\hat{\mu}_i) \tag{14}$$

C. Chebyshev Scalarized-KG

Chebyshev scalarized-KG (Cheb-KG) computes the exploration bound vector of each arm in each dimension, i.e. $\mathbf{ExpB}_i = [\mathrm{ExpB}_i^1, \cdots, \mathrm{ExpB}_i^D]$ to trade-off between exploration and exploitation, then converts the multi-objective problem to one-dimension problem. Cheb-KG is as follows:

$$f^s_{Cheb-KG}(\hat{\mu}_i) = \min_{1\leq d\leq D} w^d(\hat{\mu}_i^d + \mathrm{ExpB}_i^d - z^d) \quad \forall_i \tag{15}$$

where $f^s_{Cheb-KG}$ is a Chebyshev KG-scalarized function with scalarization s, ExpB_i^d is the exploration bound of arm i for dimension d which is calculated by using Equation 13. And, $\boldsymbol{z} = [z^1, \cdots, z^D]^T$ is a reference point. For each dimension d, the corresponding reference z^d is the minimum of the current estimated means of all arms minus a small positive value, $\epsilon^d > 0$. The reference z^d for dimension d is calculated as follows:

$$z^d = \min_{1\leq i\leq |A|} \hat{\mu}_i^d - \epsilon^d, \qquad \forall_d \tag{16}$$

Cheb-KG selects its optimal arm $i^*_{Cheb-KG}$ that has maximum $f^s_{Cheb-KG}(\hat{\mu}_i)$ as follows:

$$i^*_{Cheb-KG} = \underset{i=1,\cdots,|A|}{\operatorname{argmax}} f^s_{Cheb-KG}(\hat{\mu}_i) \tag{17}$$

LS1-KG, LS2-KG and Cheb-KG balance two terms when they select their arms. First, they prefer the arms that have large estimated variance $\hat{\sigma}^2$. Second, they prefer the arms with $|\hat{\mu}_i - \max_{j\neq i}\hat{\mu}_j|$ close to 0. LS1-KG, LS2-KG and Cheb-KG reduce gradually the exploration bound of arms by multiplying the time step $(L-t)$ with the index v_i of each arm. As we get close to the end of the horizon, they select the arms with high estimated mean.

V. Experiments

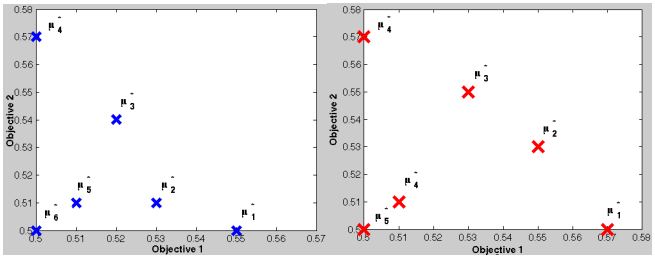
In this section, we experimentally compare the scalarized-UCB1 (linear scalarized UCB1 (LS-UCB1) and Chebyshev scalarized-UCB1 (Cheb-UCB1)), Section II-C and the scalarized-KG (linear scalarized-KG across arms (LS1-KG), linear scalarized-KG across dimensions (LS2-KG), and Chebyshev scalarized-KG (Cheb-KG)), Section IV on MOMABs with convex and non-convex (concave) mean vector arm set. The performance measures are: 1) The number of times optimal arms are pulled, i.e. the average of M experiments that optimal arms are pulled. 2) The number of times each of the optimal arms is drawn, i.e. the average of M experiments that each one of the optimal arms is pulled. 3) The average regret and the average unfairness regret, Section II-B at each time step which are the average of M experiments. The number of experiments M is 1000. The horizon of each experiment L is 1000. The weight sets $\boldsymbol{w}^s$, $s = 1, \cdots, S$ are chosen uniformly at random. For instance, for 2-objective 2-arm problem, we consider 11 weight sets $w = \{(1,0)^T, (0.9,0.1)^T, \cdots, (0.1,0.9)^T, (0,1)^T\}$. The rewards of each arm i in each dimension d, $d = 1, \cdots, D$ are drawn from normal distribution $N(\boldsymbol{\mu}_i, \boldsymbol{\sigma}^2_{i,r})$ where $\boldsymbol{\mu}_i = [\mu_i^1, \cdots, \mu_i^D]^T$ is the true mean and $\boldsymbol{\sigma}_{i,r} = [\sigma_{i,r}^1, \cdots, \sigma_{i,r}^D]^T$ is the true standard deviation of the reward. For Chebyshev scalarization (Cheb-UCB1 and Cheb-KG), ϵ is generated uniformly at random, $\epsilon \in [0, 0.1]$ as [6] and we used fixed ϵ value in all the M experiments. According to [2], the performance of KG policy increases as the standard deviation increases, therefore, the standard deviation for arms in each dimension is set to 0.01. The true means and the true standard deviations of arms are unknown parameters to the agent. Knowledge gradient KG needs the estimated standard deviation for each arm, $\hat{\sigma}_i$, therefore, each arm is played initially 10 times, $Initial = 10$. For upper confidence bounce UCB1, each arm is also played initially 10 time, i.e. $Initial = 10$ to get fairly comparison with KG policy.

A. Non-Convex Mean Vector Set

Experiment 1. We use the same example in [6], since it is simple to understand and the Pareto mean set contains values close to each others. The number of arms $|A|$ equals 6, the number of dimensions D equals 2. The true mean set vector is ($\boldsymbol{\mu}_1 = [0.55, 0.5]^T, \boldsymbol{\mu}_2 = [0.53, 0.51]^T, \boldsymbol{\mu}_3 = [0.52, 0.54]^T, \boldsymbol{\mu}_4 = [0.5, 0.57]^T, \boldsymbol{\mu}_5 = [0.51, 0.51]^T, \boldsymbol{\mu}_6 = [0.5, 0.5]^T$). Note that the Pareto optimal arm (Pareto front) set is $A^* = (a_1^*, a_2^*, a_3^*, a_4^*)$ where a_i^* refers to the optimal arm i^*. The suboptimal a_5 is not dominated by the two optimal arms a_1^* and a_4^*, but a_2^* and a_3^* dominates a_5 while a_6 is dominated by all the other mean vectors. Fig. 2 shows a set of bi-objective true means with a non-convex set.

We consider 11 weight sets, i.e. $w = \{(1,0)^T, (0.9,0.1)^T, \cdots, (0.1,0.9)^T, (0,1)^T\}$ as [6].

Table I gives the average number $\pm$ the upper and lower bounds of the confidence interval that the optimal arms are selected in column A^*, and one of the optimal arm a^* is pulled in columns a_1^*, a_2^*, a_3^*, and a_4^* using the scalarized functions in column functions.



(a) non-convex mean vectors (b) convex mean vectors

Fig. 2. Bi-objective, 6-armed. The optimal means are: $\mu_1^*, \mu_2^*, \mu_3^*$ and μ_4^*. Non-convex (concave) mean vector set is given in sub-figure a. Convex mean vector set is given in sub-figure b.

Table I shows scalarized-KG (LS1-KG, LS2-KG and Cheb-KG) is able to explore all the optimal arms, where the number of selecting the optimal arms is increased. While, scalarized-UCB1 (LS-UCB1 and Cheb-UCB1) is able to exploit the optimal arms fairly. Linear scalarized functions outperform Chebyshev scalarized functions in selecting the optimal arms, i.e. LS1-KG and LS2-KG perform better than Cheb-KG, and LS-UCB1 performs better than Cheb-UCB1 in selecting the optimal arms. In opposite, Chebyshev scalarized functions outperform linear scalarized functions in playing fairly the optimal arms, i.e. Cheb-KG performs better than LS1-KG and LS2-KG, and Cheb-UCB1 performs better than LS1-UCB1 in playing fairly the optimal arms. LS1-KG outperforms other scalarized functions in exploring the optimal arms, where the number of times optimal arms are pulled is increased with high confidence. While, Cheb-UCB1 outperforms other scalarized functions in exploiting the optimal arms.

Increasing Arms: We add another 14 additional arms in Experiment 1, resulting 20 armed. We used the same set of weights, w of Experiment 1.

Instance 1. We add 14 dominated arms by all the arms in A^*. We take $\boldsymbol{\mu}_7, \cdots, \boldsymbol{\mu}_{20} = [0.48, 0.48]^T$, leaving the Pareto optimal arm set A^* unchanged in Experiment 1. Fig. 3 gives the average scalarized and unfairness regret performances. The x-axis is the horizon of each experiment. The y-axis is either the average of the scalarized or the unfairness regret performance which is the average of 1000 experiments.

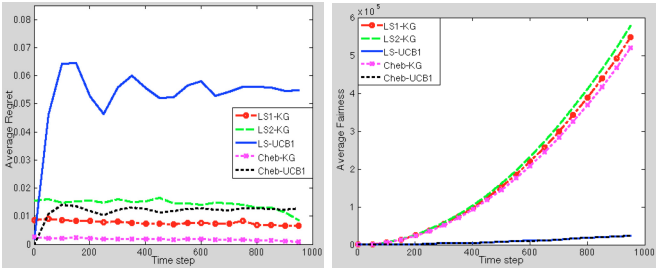
Instance 2. We add 14 arms in Experiment 1, three of them are optimal arms and 11 of them are dominated by all the arms in A^*, i.e. we increase the number of optimal arms. We take $\boldsymbol{\mu}_7 = [0.56, 0.52]^T, \boldsymbol{\mu}_8 = [0.52, 0.56]^T, \boldsymbol{\mu}_9 = [0.54, 0.54]^T, \boldsymbol{\mu}_{10} = [0.48, 0.48]^T = \cdots, \boldsymbol{\mu}_{20} = [0.48, 0.48]^T$. Pareto optimal arm set contains 7 optimal arms, i.e. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*, a_7^*, a_8^*, a_9^*)$. Fig. 4 gives the average scalarized and unfairness regret performances.

According to the scalarized regret performance, Fig. 3 and 4 show as the number of optimal arms is increased KG outperforms UCB1 in exploring the optimal arms. Cheb-KG is the best one and LS-UCB1 is the worst one. Cheb-KG performs better than LS1-KG and LS2-KG. Cheb-UCB1

TABLE I

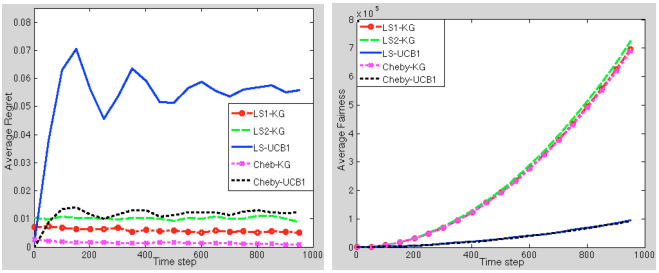
NUMBER OF TIMES OPTIMAL ARMS A^* ARE PULLED AND NUMBER OF TIMES EACH ONE OF THE OPTIMAL ARM IS PULLED PERFORMANCES ON NON-CONVEX 2-OBJECTIVE MABS WITH NUMBER OF ARMS $|A| = 6$

functions	A^*	a_1^*	a_2^*	a_3^*	a_4^*
LS1-KG	$999.9 \pm .04$	222 ± 9.7	122.6 ± 7.4	301.5 ± 14.4	353.8 ± 12.2
LS2-KG	$999.7 \pm .33$	368.2 ± 17.6	303.1 ± 18.2	96 ± 9.3	232.4 ± 8.5
Cheb-KG	$999.2 \pm .25$	279 ± 6	228.7 ± 7	264.4 ± 6	227.1 ± 4.3
LS-UCB1	$680.1 \pm .07$	$168.9 \pm .08$	$166.6 \pm .06$	$170.9 \pm .06$	$173.7 \pm .07$
Cheb-UCB1	$670.4 \pm .08$	$167.5 \pm .06$	$168.3 \pm .06$	$168.7 \pm .06$	$165.9 \pm .06$



(a) scalarized regret (b) unfairness regret

Fig. 3. 2-objective, 20-armed with concave mean vector set. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*)$. Sub-figure a shows the scalarization regret performance. Sub-figure b shows the unfairness regret performance.



(a) scalarized regret (b) unfairness regret

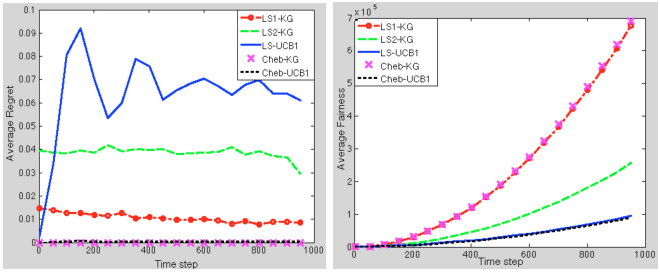
Fig. 4. 2-objective, 20-armed with concave mean vector set. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*, a_7^*, a_8^*, a_9^*)$. Sub-figure a shows the scalarization regret performance. Sub-figure b shows the unfairness regret performance.

performs better than LS-UCB1. According to the unfairness regret performance, UCB1 outperforms KG in exploiting the optimal arms. LS-UCB1 performs as same as Cheb-UCB1. Cheb-KG performs better than linear scalalized-KG, but as the number of optimal arms is increased, Cheb-KG performs as same as linear scalalized-KG (LS1-KG and LS2-KG).

Increasing dimensions: We add extra dimensions to the previous concave, 2-objective, 20-armed bandit problem in order to compare the KG and UCB1 performances on a more complex MOMAB problem, resulting in a 5-objective, 20-armed MOMAB problem. We used 11 set of weights, w.

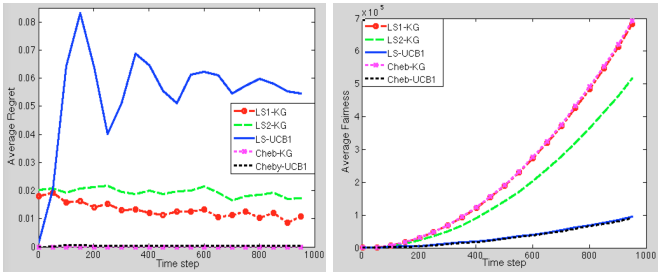
Instance 3. We add 3-objective in Instance 1, such that the Pareto optimal arm set is not changed i.e. $|A^*| = 4$, $A^* = (a_1^*, a_2^*, a_3^*, a_4^*)$. Fig. 5 gives the average scalarized and unfairness regret performances.

Instance 4. We add 3-objective in Instance 2, such that the Pareto optimal arm set is not changed i.e. $|A^*| = 7$,



(a) scalarized regret (b) unfairness regret

Fig. 5. 5-objective, 20-armed with concave mean vector set. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*)$. Sub-figure a shows the scalarization regret performance. Sub-figure b shows the unfairness regret performance.

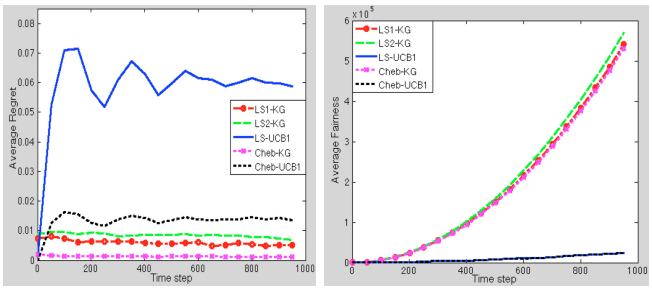


(a) scalarized regret (b) unfairness regret

Fig. 6. 5-objective, 20-armed with concave mean vector set. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*, a_7^*, a_8^*, a_9^*)$. Sub-figure a shows the scalarization regret performance. Sub-figure b shows the unfairness regret performance.

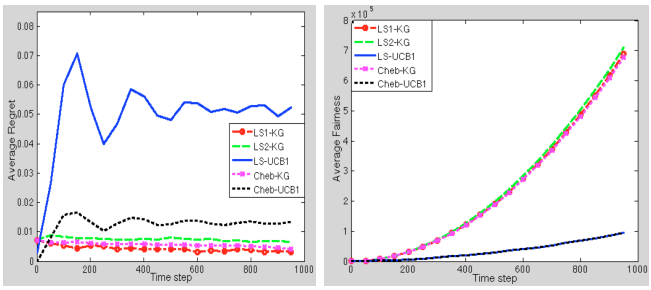
$A^* = (a_1^*, a_2^*, a_3^*, a_4^*, a_7^*, a_8^*, a_9^*)$. Fig. 6 gives the average scalarized and unfairness regret performances.

According to the scalarized regret performance, Fig. 5 and 6 show as the number of dimensions is increased, the performance of Chebyshev-scalarization (KG and UCB1) is increased. Cheb-KG and Cheb-UCB1 outperform linear-scalarization (LS1-KG, LS2-KG and LS-UCB1) in exploring the optimal arms, where Cheb-KG performs as same as Cheb-UCB1. LS-UCB1 is the worst one. According to the unfairness regret performance, Fig. 5 and 6 show as the number of dimensions is increased, the performance of LS2-KG is increased. LS2-KG performs better than Cheb-KG and LS1-KG. UCB1 outperforms KG in exploiting the optimal arms.



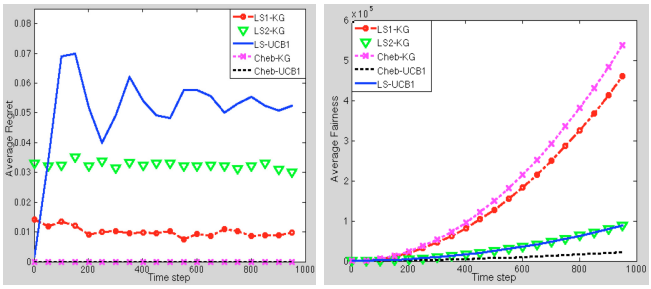
(a) scalarized regret (b) unfairness regret

Fig. 7. 2-objective, 20-armed with convex mean vector set.. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*)$. Sub-figure a shows the scalarization regret performance. Sub-figure b shows the unfairness regret performance.



(a) scalarized regret (b) unfairness regret

Fig. 8. 2-objective, 20-armed with convex mean vector set.. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*, a_7^*, a_8^*, a_9^*)$. Sub-figure a shows the scalarization regret performance. Sub-figure b shows the unfairness regret performance.



(a) scalarized regret (b) unfairness regret

Fig. 9. 5-objective, 20-armed with convex mean vector set. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*)$. Sub-figure a shows the scalarization regret performance. Sub-figure b shows the unfairness regret performance.

B. Convex Mean Vector Set

Experiment 2. With number of arms $|A|$ equals 6, number of dimensions D equals 2. The true convex mean set vector is $(\boldsymbol{\mu}_1 = [0.57, 0.5]^T, \boldsymbol{\mu}_2 = [0.55, 0.53]^T, \boldsymbol{\mu}_3 = [0.53, 0.55]^T, \boldsymbol{\mu}_4 = [0.5, 0.57]^T, \boldsymbol{\mu}_5 = [0.51, 0.51]^T, \boldsymbol{\mu}_6 = [0.5, 0.5]^T)$. Note that the Pareto optimal arm (Pareto front) set is $A^* = (a_1^*, a_2^*, a_3^*, a_4^*)$. The suboptimal a_5 is not dominated by the two optimal arms a_1^* and a_4^*, but a_2^* and a_3^* dominates a_5 while a_6 is dominated by all the other mean vectors. Fig. 2 shows a set of 2-objective true means with a convex set. We consider 11 weight sets for UCB1 and KG scalarization functions, i.e. $w = \{(1,0)^T, (0.9,0.1)^T, \cdots, (0.1,0.9)^T, (0,1)^T\}$.

Table II gives the average number $\pm$ the upper and lower bounds of the confidence interval that the optimal arms are selected in column A^*, and one of the optimal arm a^* is pulled in columns a_1^*, a_2^*, a_3^*, and a_4^* using the scalarized functions in column functions. Table II shows the performance of KG is increased when the mean vector set is a convex set. KG policy performs better than UCB1 policy in selecting (exploring) the optimal arms. UCB1 policy performs better than KG in (playing fairly) exploiting the optimal arms. LS2-KG outperforms all other scalarized-KG functions in playing fairly the optimal arms. LS-UCB1 outperforms Cheb-KG in selecting and playing fairly the optimal arms.

Increasing arms: We add another 14 arms in Experiment 2, resulting 20 armed bandits. We use the same set of weights, w of Experiment 2.

Instance 5. We add 14 dominated arms by all the arms in A^*. We take $\boldsymbol{\mu}_7, \cdots, \boldsymbol{\mu}_{20} = [0.48, 0.48]^T$, leaving the Pareto optimal arm set unchanged in *Experiment* 2. Fig. 7 gives the average scalarized and unfairness regret performances.

Instance 6. We add 14 arms in *Experiment* 2, three of them are optimal arms and 11 of them are dominated by all the arms in A^*. We take $\boldsymbol{\mu}_7 = [0.56, 0.52]^T, \boldsymbol{\mu}_8 = [0.52, 0.56]^T$, $\boldsymbol{\mu}_9 = [0.54, 0.54]^T, \boldsymbol{\mu}_{10} = [0.48, 0.48]^T, \cdots, \boldsymbol{\mu}_{20} = [0.48, 0.48]^T$. Pareto optimal arm set contains 7 optimal arms, i.e. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*, a_7^*, a_8^*, a_9^*)$. Fig. 8 gives the average scalarized and unfairness regret performances.

Fig. 7 and 8 show KG outperforms UCB1 in exploring the optimal arms while UCB1 outperforms KG in exploiting the optimal arms. Cheb-UCB1 performs better than LS-UCB1 according to the scalarized regret performance and as same as LS-UCB1 according to the unfairness regret performance. According to the unfairness regret, the performance of LS1-KG does not change with increasing the number of the optimal arms. While, the scalarized regret of LS1-KG is improved when the number of optimal armed is increased.

Increasing dimensions: We add extra dimensions to the previous convex, 2-objective, 20-armed bandit problem in order to compare the KG and UCB1 performances on a more complex MOMAB problem, resulting in a 5-objective, 20-armed MOMAB problem. We used 11 set of weights.

Instance 7. We add 3-objective in Instance 5, such that the Pareto optimal arm set is not changed i.e. $|A^*| = 4$, $A^* = (a_1^*, a_2^*, a_3^*, a_4^*)$. Fig. 9 gives the average scalarized and unfairness regret performances.

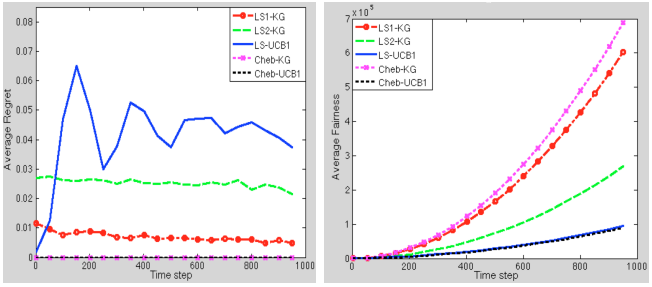
Instance 8. We add 3-objective in Instance 6, such that the Pareto optimal arm set is not changed i.e. $|A^*| = 7$, $A^* = (a_1^*, a_2^*, a_3^*, a_4^*, a_7^*, a_8^*, a_9^*)$. Fig. 10 gives the average scalarized and unfairness regret performances.

According to the scalarized regret performance, Fig. 9 and 10 show Chebyshev scalarization (using KG and UCB1) outperforms linear-scalarization. The worst scalarized regret is achieved by using LS-UCB1 and the best scalarized regret is achieved by using Cheb-KG and Cheb-UCB1. According to the unfairness regret performance, Fig. 9 and 10 show the worst unfairness regret is achieved by using Cheb-KG. As the number of optimal arms equals 4, Fig. 9 shows

TABLE II

NUMBER OF TIMES OPTIMAL ARMS A^* ARE PULLED AND NUMBER OF TIMES EACH ONE OF THE OPTIMAL ARM IS PULLED PERFORMANCES ON CONVEX BI-OBJECTIVE MABS WITH NUMBER OF ARMS $|A| = 6$

functions	A^*	a_1^*	a_2^*	a_3^*	a_4^*
LS2-KG	1000 ± 0	251.9 ± 14.3	251.2 ± 16.51	240.6 ± 15.84	256.3 ± 14.56
LS1-KG	1000 ± 0	239 ± 10.89	257.4 ± 12.26	270.5 ± 12.47	233.1 ± 10.98
Cheb-KG	1000 ± 0	267 ± 5.38	324.3 ± 5.9	321.6 ± 6.13	87.1 ± 5.69
LS-UCB1	$686.1 \pm .08$	$170.6 \pm .08$	$172.5 \pm .06$	$172.5 \pm .07$	$170.5 \pm .07$
Cheb-UCB1	$671.6 \pm .08$	$167.4 \pm .06$	$169 \pm .06$	$168.7 \pm .06$	$166.5 \pm .06$



(a) scalarized regret (b) unfairness regret

Fig. 10. 5-objective, 20-armed with convex mean vector set. $A^* = (a_1^*, a_2^*, a_3^*, a_4^*, a_7^*, a_8^*, a_9^*)$. Sub-figure a shows the scalarization regret performance. Sub-figure b shows the unfairness regret performance.

the performance of LS2-KG as same as LS-UCB1 and Cheb-UCB1 outperforms all the scalarized functions. As the number of optimal arms is increased ($A^* = 7$), Fig. 10 shows scalarized-UCB1 outperforms scalarized-KG. LS2-KG performs better than LS1-KG and Cheb-KG.

From the above results, we see that KG explores better than UCB1, while UCB1 exploits better than KG. The intuition is the exploration bonus. The exploration bonus for UCB1 depends on the time step t and the number of times N_i arm i is pulled. The exploration bonus is high if the arm i is less selected. Thus, UCB1 plays fairly the optimal arms because it selects the optimal arms that have either larger estimated mean or larger exploration bonus. In contrast, the exploration bonus for KG policy depends on the estimated mean of all other arms and on the estimated variance of arm i. The exploration bonus is large if the variance of arm i is low, or if the estimated mean of arm i exceeds in the future. Thus, KG selects more efficiently the optimal arms.

VI. CONCLUSIONS AND FUTURE WORK

We presented multi-objective, multi-armed bandit problem MOMAB, linear, and non-linear scalarized functions and the scalarized and unfairness regret measures. We also presented UCB1 policy in MOMAB and the knowledge gradient KG policy. We proposed two types of linear scalarized-KG (linear scalarized-KG across arms (LS1-KG) and linear scalarized-KG across dimensions (LS2-KG)) and Chebyshev-scalarized-KG (Cheb-KG). We studied the trade-off between exploration and exploitation in the MOMAB. The scalarized multi-objective KG bandits is either converts the multi-objective space to one-objective space then trade-off between exploration and exploitation or trade-off between exploration and exploitation directly in the multi-objective space. Finally we compared KG and UCB1 and concluded that: 1) In the MOMAB problem, KG and UCB1 policies are able to find the Pareto optimal arms set in convex and concave mean vector set. The scalarized regret is improved using KG policy, while the unfairness regret is improved using upper confidence bound (UCB1) policy. 2) Chebyshev-KG and Chebyshev-UCB1 are able to find the Pareto optimal arm set without moving the reference point. 3) According to the scalarized regret performance, Cheb-KG performs better than linear-scalarized KG (LS1-KG and LS2-KG) and LS1-KG performs better than LS2-KG. According to the unfairness regret performance, LS2-KG performs better than LS1-KG, while the performance of Cheb-KG depends on the number of armed and objectives. 4) Chebyshev-UCB1 (Cheb-UCB1) performs better than linear-scalarized UCB1 (LS-UCB1) according to the scalarized regret performance. While, LS-UCB1 performs as same as Cheb-UCB1 according to the unfairness regret performance. Future work should provide theoretical analysis for the KG in MOMAB.

REFERENCES

[1] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning.* The MIT Press, Cambridge, MA, 1998.

[2] S.Q. Yahyaa and B. Manderick, "The exploration vs exploitation trade-off in the multi-armed bandit problem: An empirical study," *European Symposium on Artificial Neural Networks (ESANN)*, pp. 549-554,2004.

[3] I.O. Ryzhov, W.B. Powell and P.I. Frazier, "The knowledge-gradient policy for a general class of online learning problems," *Operation Research*, 2011.

[4] P. Auer, N. Cesa-Bianchi and P. Fischer, "Finite-Time Analysis of the Multiarmed Bandit Problem,"*Int. J. Machine Learning*, vol. 47, no. 2-3, pp. 235-256, 2002.

[5] G. Eichfelder, editor. *Adaptive Scalarization Methods in Multiobjective Optimization*, Springer-Verlag Berlin Heidelberg, 2008.

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

[7] K. Miettinen, *Nonlinear Multiobjective Optimization.* Springer, 1999.

[8] W.B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality.* John Willey and Sons, New York, USA, 2007.

[9] S. Q. Yahyaa, M. M. Drugan and B. Manderick, " Knowledge gradient for multi-objective multi-armed bandit algorithms," *International Conference on Agents and Artificial Intelligence (ICAART), 2014.*