ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Linear Scalarized Knowledge Gradient in the Multi-Objective Multi-Armed Bandits Problem

Saba Yahyaa, Madalina M. Drugan and Bernard Manderick

Computational Modeling group, Artificial Intelligence Lab, Computer Science Department
Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
syahyaa, mdrugan, bmanderi@vub.ac.be

Abstract. The multi-objective, multi-armed bandits (MOMABs) problem is a Markov decision process with stochastic rewards. Each arm generates a vector of rewards instead of a single reward and these multiple rewards might be conflicting. The agent has a set of optimal arms and the agent's goal is not only finding the optimal arms, but also playing them fairly. To find the optimal arm set, the agent uses a linear scalarized (LS) function which converts the multi-objective arms into one-objective arms. LS function is simple, however it can not find all the optimal arm set. As a result, we extend knowledge gradient (KG) policy to LS function. We propose two variants of linear scalarized-KG, LS-KG across arms and dimensions. We experimentally compare the two variant, LS-KG across arms finds the optimal arm set, while LS-KG across dimensions plays fairly the optimal arms.

1 INTRODUCTION

The one-objective, Multi-Armed Bandits (MABs) is a sequential Markov decision process where an agent tries to optimize its decisions while improving its knowledge concerning the arms among which it has to pull. At each time step t, the agent pulls one from the available arms set A and receives a reward signal. That reward is independent from the past rewards of the pulled arm and from all other arms. The rewards from arm i are drawn from a normal distribution $N(\mu_i, \sigma_i^2)$ with mean μ_i and variance σ_i^2.

The goal of the agent is to find the best arm i^* which has the maximum mean $\mu^* = \max_{i=1,\cdots,|A|} \mu_i$ to minimize the *loss*, or *total expected regret*, R_L, $R_L = L\mu^* - \sum_{t=1}^{L} \mu_i(t)$ of not pulling the best arm i^* all the time. Where L is a fixed number of time steps and $\mu_i(t)$ is the mean of the selected arm i at time step t.

However, the mean μ_i and variance σ_i^2 parameters are unknown to the agent. Thus, by pulling each arm, the agent improves its estimates $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ of the true mean μ_i and the variance σ_i^2, respectively. To find the optimal arm as soon as possible, the agent has several policies, e.g. Knowledge Gradient (KG) policy [1]. Intuitively, KG policy finds the optimal arm by adding a bonus to the estimated mean of each arm i and selects the arm that has the maximum estimated mean plus the bonus.

In this paper, we extend KG policy to the Multi-Objective, Multi-Armed Bandits problem (MOMABs). In the Multi-Objective (MO) setting, there is a set of Pareto optimal arms that are incomparable, i.e. can not be ordered using a designed partial order relationship. The Pareto optimal arm set (Pareto front set) can be found by using Linear Scalarized (LS) function which converts the MO space to a single-objective space, i.e. the mean vectors are transformed into scalar values [2]. The LS function is simple but cannot find all the optimal arms when the mean vector set of the Pareto front

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

set is a non-convex (concave) set. As a result, we extend KG policy to be used in linear scalarization function to find the optimal arms in a concave mean vector set.

This paper is organized as follows: first, we give background information on the algorithm used (Section 2). We introduce the MOMABs, we propose two variants of the linear KG scalarization functions: linear scalarized-KG across arms and dimensions and we present scalarized multi-objective bandits. (Section 3). We describe the experimental set up followed by the experimental results (Section 4). Finally, we conclude.

2 Background

Here, we introduce LS function and regret measures for the MOMABs problem.

Let us consider MOMABs problems with number of arms $|A|$, $|A| \geq 2$ arms and with D objectives (dimensions) per arm. The mean vector of the rewards of arm i, $1 \leq i \leq |A|$, is $\boldsymbol{\mu}_i = (\mu_i^1, \cdots, \mu_i^D)^T$, where T is the transpose. Each objective d, $1 \leq d \leq D$, has a specific value and the objectives might be conflicting with each other. This means that the reward of arm i corresponding with one objective can be better but for another objective can be worse than that for another arm j.

Linear Scalarized (LS) Function converts the MO into a one-objective [2]. However, solving a MO optimization problem means finding the Pareto front set. Thus, we need a set of scalarized functions S to generate the variety of elements belonging to the Pareto front set. LS function assigns to each value μ_i^d of the mean vector $\boldsymbol{\mu}_i$ of arm i a weight w^d and the result is the sum of these weighted mean values. The LS function f is defined as:

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

where f^j is a linear function with scalarization j, $j \in S$. Each j has a different set of predefined weights $\boldsymbol{w}^j = (w^1, \cdots, w^D)$, such that $\sum_{d=1}^{D} w^d = 1$. Linear scalarization is very popular because of its simplicity. However, it cannot find all the arms in the Pareto optimal arm set A^* if its corresponding mean set is a concave set. After transforming the multi-objective problem to a single-objective one, the LS function selects the arm i^* that has the maximum function value, i.e. $i^* = \text{argmax}_{1 \leq i \leq A}\ f^j(\boldsymbol{\mu_i})$.

Regret Metrics. To measure the performance of the LS function, [3] have proposed two regret metric criteria. *The scalarized regret metric* measures the distance between the maximum value of a scalarized function and the scalarized value of an arm that is pulled at time step t. Scalarized regret is the difference between the maximum value for a LS function f^j on the set of arms A and the scalarized value for an arm k that is pulled by the scalarized f^j at time step t, $R_{scalarized^j}(t) = \max_{1 \leq i \leq A} f^j(\boldsymbol{\mu_i}) - f^j(\boldsymbol{\mu_k})(t)$.

The unfairness regret metric is related to the variance in drawing all the optimal arms which is the variance of the times the arms in A^* are pulled: $R_{unfairness}(t) = \frac{1}{|A^*|}\sum_{i^* \in A^*}(N_{i^*}(t) - N_{|A^*|}(t))^2$, where $R_{unfairness}(t)$ is the unfairness regret at time step t, $|A^*|$ is the number of optimal arms, $N_{i^*}(t)$ is the number of times an optimal arm i^* has been selected at time step t and $N_{|A^*|}(t)$ is the number of times the optimal arms, $i^* = 1, \cdots, |A^*|$ have been selected at time step t.

Knowledge Gradient (KG) Policy is an index policy that determines for each arm i

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

the index V_i^{KG} as follows [1]:

$$V_i^{KG} = \hat{\bar{\sigma}}_i * x(-|\frac{\hat{\mu}_i - \max_{j \neq i, j \in |A|} \hat{\mu}_j}{\hat{\bar{\sigma}}_i}|) \tag{2}$$

where $\hat{\bar{\sigma}}_i = {}^{\sigma_i}/_{N_i}$ is the Root Mean Square Error (RMSE) of the estimated mean $\hat{\mu}_i$ of arm i. The function $x(\zeta) = \zeta\Phi(\zeta) + \phi(\zeta)$ where Φ and ϕ are the cumulative distribution and density of the standard normal distribution $N(0, 1)$, respectively. KG chooses the arm i^*_{KG} with the largest V_i^{KG}, i.e. $i^*_{KG} = \text{argmax}_{i \in |A|} \left(\hat{\mu}_i + (L - t)V_i^{KG}\right)$ where L is the horizon of experiment. KG policy 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 variance $\hat{\sigma}_i^2$. Thus, KG prefers an arm i over its alternatives if its confidence in the estimate mean $\hat{\mu}_i$ is low. In [5], it is shown that KG policy outperforms other policies on the one-objective MABs in terms of the average frequency of optimal selection performance. Moreover, the KG-policy does not have any parameter to be tuned. For these reasons, we extend it to scalarized multi-objective KG.

3 MOMAB Framework

In this section, we present linear scalarized knowledge gradient (scalarized-KG) functions. Linear scalarized-KG functions make use of the estimated mean and variance.

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 $\boldsymbol{\sigma}_i^2 = (\sigma_{1i}^2, \cdots, \mu_{Di}^2)^T$ 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, the agent can update its estimates $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\sigma}}_i^2$ in each dimension d as follows [4]:

$$N_{i^{t+1}} = N_{i^t} + 1, \quad \hat{\mu}_{t+1}^d = (1 - \frac{1}{N_{i^{t+1}}})\,\hat{\mu}_t^d + \frac{1}{N_{i^{t+1}}} r_{t+1}^d \tag{3}$$

$$\hat{\sigma}_{d(t+1)}^2 = \frac{N_{i^{t+1}} - 2}{N_{i^{t+1}} - 1}\,\hat{\sigma}_{dt}^2 + \frac{1}{N_{i^{t+1}}}(r_{t+1}^d - \hat{\mu}_t^d)^2 \tag{4}$$

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

Linear Scalarized-KG across Arms, (LS1-KG) converts immediately the MO estimated mean $\hat{\boldsymbol{\mu}}_i$ and estimated variance $\hat{\boldsymbol{\sigma}}_i^2$ of each arm to one-dimension, then computes the corresponding bound ExpB_i. 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. $([\hat{\sigma}_{1i}^2, \cdots, \hat{\sigma}_{Di}^2]^T)$ of each arm i, converts the MO vectors to one-objective values by summing the elements of each vector. Thus, we have one-dimension, MABs. KG calculates for each arm, a bound which depends on all other arms and selects the arm that has the maximum estimated mean plus bound. The LS1-KG is as follows:

$$\widetilde{\mu}_i = f^j(\hat{\mu}_i) = w^1\hat{\mu}_i^1 + \cdots + w^D\hat{\mu}_i^D,\; \widetilde{\sigma}_i^2 = f^j(\hat{\sigma}_i^2) = w^1\hat{\sigma}_{1i}^2 + \cdots + w^D\hat{\sigma}_{Di}^2\; \forall_i$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

$$\widetilde{\widetilde{\sigma}}_i^2 = \widetilde{\sigma}_i^2/N_i,\ v_i = \widetilde{\widetilde{\sigma}}_i\ x(-|\frac{\widetilde{\mu_i} - \max_{j\neq i,\, j\in A}\widetilde{\mu_j}}{\widetilde{\widetilde{\sigma}}_i}|) \qquad \forall_i \tag{5}$$

where f^j is a LS function that has a predefined set of weight $(w^1, \cdots, w^D)^j$. $\widetilde{\mu_i}$ and $\widetilde{\sigma}_i^2$ are the modified estimated mean and variance of an arm i, respectively which are values. $\widetilde{\widetilde{\sigma}}_i^2$ is the RMSE of an arm i. v_i is the KG index of an arm i. LS1-KG selects the optimal arm i^* according to:

$$i^*_{LS_1KG} = \text{argmax}_{i\in|A|}(\widetilde{\mu_i} + \text{ExpB}_i) = \text{argmax}_{i\in|A|}(\widetilde{\mu_i} + (L-t) * |A|D * v_i)$$

where ExpB_i is the bound of arm i and D is the number of dimensions.

Linear scalarized-KG across dimensions, (LS2-KG) computes the bound vector $\mathbf{ExpB}_i$ for each arm, i.e. $\mathbf{ExpB}_i = [\text{ExpB}_i^1, \cdots, \text{ExpB}_i^D]$, adds the $\mathbf{ExpB}_i$ to the corresponding estimated mean vector $\hat{\boldsymbol{\mu}}_i$, then converts the multi-objective problem to one-objective. At each time step t, LS2-KG computes bounds for all dimensions of each arm, sums the estimated mean in each dimension with its corresponding 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:

$$f^j(\hat{\mu}_i) = w^1(\hat{\mu}_i^1 + \text{ExpB}_i^1) + \cdots + w^D(\hat{\mu}_i^D + \text{ExpB}_i^D)\ \forall_i,\ \text{where} \tag{6}$$

$$\text{ExpB}_i^d = (L-t) * |A|D * v_i^d,\ v_i^d = \hat{\bar{\sigma}}_i^d\ x(-|\frac{\hat{\mu}_i^d - \max_{j\neq i,\, j\in A}\hat{\mu}_j^d}{\hat{\bar{\sigma}}_i^d}|)\ \forall_{d\in D}$$

v_i^d, $\hat{\mu}_i^d$, $\hat{\bar{\sigma}}_i^d$, and $ExpB_i^d$ are the index, the estimated mean, the RMSE, and the bound of arm i for dimension d, respectively. LS2-KG selects the optimal arm i^* that has maximum $f^j(\hat{\mu}_i)$, i.e. $i^*_{LS_2KG} = \text{argmax}_{i=1,\cdots,|A|} f^j(\hat{\mu}_i)$.

```
1. Input:length of trajectory L;type of scalarized function f;
   set of scalarized function S = (f^1,...,f^S);reward r^d ~ N(μ,σ_r^2).
2. Initialize: For s = 1 to S
   play each arm Initial steps;observe(r_i)^s;update:N_i^s;(μ̂_i)^s;(σ̂_i)^s
              End
3. Repeat
4. Select:  a function s ∈ S uniformly, at random
5. Select:  the optimal arm i* that maximizes f^s
6. Observe: reward vector r_i*, r_i* = [r_i*^1, ..., r_i*^D]^T
7. Update:  N_i*^s;μ̂_i*;σ̂_i*
8. Compute: unfairness regret;scalarized regret
9. Until L
10. Output: Unfairness regret;scalarized regret.
```

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

The scalarized multi-objectieve bandits, The pseudocode of the scalarized MOMAB problems is given in Figure 1. Given the type of the scalarized function f, (f is either LS1-KG or LS2-KG) and the scalarized function set $(f^1, \cdots, f^S)$ where each scalarized function f^s has different weight set, $\boldsymbol{w}^s = (w^{1,s}, \cdots, w^{D,s})$.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The algorithm in Figure 1 plays each arm of each scalarized function f^s, $Initial$ plays (step: 2). N_i^s is the number of times the arm i under the scalarized function f^s 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 $\boldsymbol{\sigma}_r^2$ is the variance vector of the reward. $(\hat{\boldsymbol{\mu}}_i)^s$ and $(\hat{\boldsymbol{\sigma}}_i)^s$ are the estimated mean and standard deviation vectors of the arm i under the scalarized function s, respectively. After initial playing, the algorithm chooses uniformly at random one of the scalarized function (step: 4), selects the optimal arm i^* that maximizes the type of this scalarized function (step: 5), simulates the selected arm i^*, and updates 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. Note that the proposed algorithm is an adapted version from [3], but here the algorithm can be applied to KG with normal reward distribution.

4 Experiments

We experimentally compare the linear scalarized-KG variants across arms and dimensions of KG, Section 3 on MOMABs with concave mean vector arm set. The performance measures are: 1) The average number of times optimal arms are pulled and the average number of times each of the optimal arms is drawn which are the average of M experiments. 2) The scalarized and the unfairness average regret, Section 2 at each time step which are the average of M experiments. The number of experiments M and the horizon of each experiment L are 1000. The rewards of each arm i in each dimension $d, d \in D$ are drawn from normal distribution $N(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_{i,r}^2)$ where $\boldsymbol{\mu}_i = [\mu_i^1, \cdots, \mu_i^D]^T$ is the mean and $\boldsymbol{\sigma}_{i,r}^2 = [\sigma_{i,r}^{1,2}, \cdots, \sigma_{i,r}^{D,2}]^T$ is the variance of the reward. The variance for arms in each dimension is equal to 0.01^2. The mean for arms in each dimension is $\in [0, 1]$. The means and the variances of arms are unknown parameters to the agent. KG needs the estimated variance for each arm, $\hat{\sigma}_i^2$, therefore, each arm is played initially $Initial, Initial \geq 2$ times to estimate the variance. The number of Pareto optimal arms $|A^*|$ is unknown to the agent, therefore, $|A^*| = |A|$.

Table 1: Average number of times optimal arms are pulled and average number of times each one of the optimal arm is pulled performances on concave 2-objective, 6-armed.

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

Experiment 1*:* We use the same example in [3] that contains concave mean set with $|A| = 6$ and $D = 2$. 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. We consider 11 weight sets, i.e. $w = \{(1, 0)^T, (0.9, 0.1)^T, \cdots, (0.1, 0.9)^T, (0, 1)^T\}$.

Table 1 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 1 shows KG policy is able to explore all the optimal arms. LS1-KG

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

performs better than LS2-KG in selecting and playing fairly the optimal arms. LS1-KG prefers the optimal arm a_4^*, while LS2-KG prefers the optimal arm a_1^*.

Increasing Arms and Dimensions: In order to compare the variants linear scalarized-KG performances on a more complex MOMABs problem, We add another 14 arms and another 3 dimensions in Experiment 1, resulting 5-objective, 20-armed. The Pareto optimal set of arms A^* contains now 7 arms. We used 11 set of weights, w. Figure 2 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. Figure 2 shows LS1-KG performs better than LS2-KG according to the scalarized regret performance, while LS2-KG performs better than LS1-KG according to the unfairness regret performance.

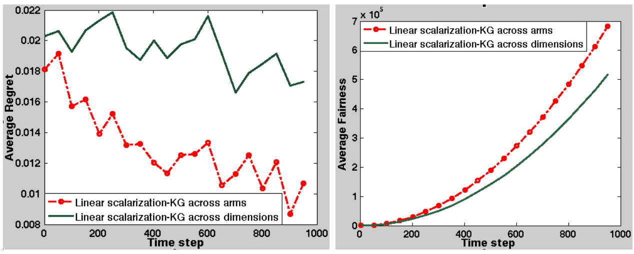

Fig. 2: Performance on 5-objective, 20-armed with $|A^*| = 7$. Left sub-figure shows the scalarized regret. Right sub-figure shows the unfairness regret.

5 Conclusions

We presented MOMABs problem, linear scalarized LS function, the scalarized regret measures and KG policy. We extended KG policy to the MOMABs. We proposed two types of LS-KG (LS-KG across arms (LS1-KG) and dimensions (LS2-KG)). Finally, we compared LS1-KG, and LS2-KG and concluded that: 1) KG policy is able to find the Pareto optimal arms set in a concave mean vector set. 2) when the number of arms and dimensions are increased LS1-KG outperforms LS2-KG according to the scalarized regret, while LS2-KG outperforms LS1-KG according to the unfairness regret.

References

[1] 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.

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

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

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

[5] S. Q. Yahyaa and B. Manderick, The exploration vs exploitation trade-off in the multi-armed bandit problem: An empirical study In M. Verleysen, editor, *proceedings of the* 20^{th} *European Symposium on Artificial Neural Networks* (ESANN 2012), d-side pub., pages 549-554, April 25-27, Belgium, 2004.