Annealing-Pareto Multi-Objective Multi-Armed Bandit Algorithm

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

Vrije Universiteit Brussel, Department of Computer Science,
Pleinlaan 2, 1050 Brussels, Belgium
{syahyaa,mdurgan,bmanderi}@vub.ac.be

The Multi-Objective Multi-Armed Bandit (MOMAB) problem is a sequential stochastic learning problem. At each time step t, an agent pulls one arm i from an available set of arms A and receives a reward vector $\boldsymbol{r}_i$ of the arm i with D objectives. The reward vector is drawn from a corresponding stationary probability distribution vector, e.g. Bernoulli distribution $\mathrm{B}(\boldsymbol{p}_i)$, where $\boldsymbol{p}_i$ is the true probability of success vector parameter of the arm i. The reward vector that the agent receives from the arm i is independent from all other arms and independent from the past reward vectors of the selected arm i. Moreover, the probability of success vector of the arm i has *independent* D distributions. We assume that the true probability of success vector parameter of each arm i is unknown parameter to the agent. Thus, by drawing each arm i, the agent maintains estimation of the true probability of success vector, $\hat{\boldsymbol{p}}_i$.

The MOMAB problem has a set of Pareto optimal arms (Pareto front) A^*, that are incomparable, i.e. can not be classified using a designed partial order relations. The agent has not to only find the optimal arms (exploring), to minimize the total Pareto loss of not pulling the optimal arms, but also has to play them fairly (exploiting), to minimize the total unfairness loss. This problem is known as the *trade-off between exploration and exploitation in the multi-objective optimization*, or the trade-off problem [1].

At each time step t, the Pareto loss (or Pareto regret) is the distance between the set mean (or the true probability of success set) of Pareto optimal arms and the mean of the selected arm (or the true probability of success vector). While, the unfairness loss (or unfairness regret) is the variance in selecting the optimal arms [2]. Thus, the total Pareto regret and the total unfairness regrets are the cumulative summation of the Pareto and unfairness regret over t time steps, respectively. Since, the total unfairness regret grows exponentially on the number of time steps and does not take into account the total number of selecting optimal arms, we propose to use the *entropy measure* to compute the unfairness regret. The entropy unfairness regret is the measurement of disarray (or disorder) on selecting the optimal arms in the Pareto front A^*.

The Pareto front A^* can be found for example, by using Pareto dominance relation which finds the Pareto front A^* by optimizing directly the Multi-Objective (MO) space. To solve the trade-off problem directly in the MO space, [2] proposed *Pareto Upper Confidence Bound* (Pareto-UCB1) and [3] proposed *Pareto Knowledge Gradient* (Pareto-KG). Both Pareto-UCB1 and Pareto-KG trade-off between exploration and exploitation by adding an exploration bound vec-

tor $\mathbf{ExpB}_i$ to the corresponding estimated mean vector (estimated probability of success vector $\hat{\boldsymbol{p}}_i$) of each arm i, $\hat{\boldsymbol{p}}_i + \mathbf{ExpB}_i$ and select the optimal arms j^* that are not dominated by all other arms $k, k \in |A|$ using Pareto dominance relations, $\hat{\boldsymbol{p}}_k + \mathbf{ExpB}_k \nprec \hat{\boldsymbol{p}}_{j^*} + \mathbf{ExpB}_{j^*}$ However, Pareto-UCB1 adds the same exploration bound to the estimated probability of success vector $\boldsymbol{p}_i$ for each objective $d, d \in D$ of an arm i, $\mathbf{ExpB}_i = [\mathrm{ExpB}_i^1 = \cdots = \mathrm{ExpB}_i^D]^T$ and this exploration bound requires only knowledge about the arm i, where T is the transpose. While, Pareto-KG adds different exploration bound to the estimated probability of success vector $\boldsymbol{p}_i$ for each objective $d, d \in D$ of an arm i, $\mathbf{ExpB}_i = [\mathrm{ExpB}_i^1, \cdots, \mathrm{ExpB}_i^D]^T$ and the exploration bound ExpB_i^d in each objective d depends on the estimated probability of success on all arms.

In this paper, we are interesting in the trade-off between exploration and exploitation by using *random* instead of adding an exploration bound to the estimated mean vectors. In the one-objective, Multi Armed Bandit (MAB) problem, Thompson Sampling [4] trades-off between exploration and exploitation by assigning to each arm i a random probability of selection P_i which is generated from Beta distribution. The random probability of selection P_i of an arm i depends on the performance of the arm i. It will be high value if the arm i has high estimated probability of success $\hat{p}_i$ value. Since, Thompson sampling is very easy to implement, empirically performs better than UCB1 policy [5] and it has been shown to be close to optimal, we will extend it to MOMABs to find the optimal arms in the MO space. *Pareto Thompson sampling* trades-off between exploration and exploitation by assigning to each arm i in each objective d a random probability of selection P_i^d which is generated from Beta distribution. Pareto Thompson sampling uses Pareto dominance relation on the random probability of selection vectors $\boldsymbol{P}$, $\boldsymbol{P} = [P_i^1, \cdots, P_i^D]^T$ to find the Pareto front A^*. We propose annealing-Pareto algorithm. The annealing-Pareto trades-off between exploration and exploitation by using a decaying parameter ϵ_t, $\epsilon_t \in (0, 1)$ in combination with the Pareto dominance relation. The ϵ_t parameter has a high value at the beginning of time step t to explore all the available arms and increase the confidence in the estimated means, but as the time step t increases, the ϵ_t parameter decreases to exploit the arms that have maximum estimated mean. To keep track on all the optimal arms in the Pareto front A^*, at each time step t, the annealing-Pareto uses Pareto dominance relation.

Annealing-Pareto algorithm has a specific mechanism to control the trade-off between exploration and exploitation. It uses an exponential decay ϵ_t, $\epsilon_t = \epsilon_{decay}^t / (|A|D)$, where ϵ_{decay} is the decay parameter and Pareto dominance relation. At the beginning of time step t, ϵ_t has a high value to explore all the available arms. As the time step t is increased, ϵ_t has a low value to exploit only the optimal arms. To keep track on all the optimal arms in the Pareto front A^*, the annealing-Pareto uses Pareto dominance relation. The decay parameter ϵ_{decay}, $\epsilon_{decay} \in (0, 1)$, when $\epsilon_{decay} = 0$ means the annealing-Pareto is a fully Pareto dominance relation and when $\epsilon_{decay} = 1$ means the annealing-Pareto uses a fixed exponential decay. The pseudocode of the annealing-Pareto is given in Algorithm 1.

Algorithm 1 (Annealing-Pareto for Bernoulli Distribution MOMAB)

1. **Input:** number of arms $|A|$; number of objectives $|D|$; selected arm set $S^d(t) = \{\ \}\ \forall_d$; decay parameter $\epsilon_{decay} \in (0,1)$.
2. **Intialize:** $\hat{p}_i^d = 0.5\ \forall_{i\in A, d\in D}$; initial ϵ-Pareto front set $A_\epsilon^*(0) = A$.
3. **For time step** $t = 1, \cdots, L$
4. Set the decay parameter $\epsilon_t = \epsilon_{decay}^t / (|A||D|)$
5. **For objective** $d = 1, \cdots, D$
6. $S^d(t) = \{\phi\}$
7. $\hat{p}^{*,d} = \max\limits_{1\le i\le A} \hat{p}_i^d$
8. **For arm** $i = 1, \cdots, A$
9. If $\hat{p}_i^d \in [\hat{p}^{*,d} - \epsilon_t, \hat{p}^{*,d}]$
10. $S^d(t) \leftarrow \{S^d(t), i\}$
11. End If
12. **End For**
13. **End For**
14. $S(t) \leftarrow S^1(t) \cup S^2(t) \cup \cdots \cup S^D(t)$
15. $S_{difference} \leftarrow A_\epsilon^*(t-1) - S(t)$
16. **For arm** $j \in S_{difference}$ do
17. If $\hat{\boldsymbol{p}}_k \not\succ \hat{\boldsymbol{p}}_j,\ \forall_k \in A$
18. $S(t) \leftarrow S(t) \cup j$
19. End If
20. **End For**
21. $A_\epsilon^*(t) \leftarrow S(t)$
22. Select an optimal arm i^* uniformly, at random from $A_\epsilon^*(t)$
23. Observe: reward vector $r_{i^*}, r_{i^*} = [r_{i^*}^1, \cdots, r_{i^*}^D]^T$; Update: $\hat{\boldsymbol{p}}_{i^*}$
24. **End For**
25. **Output:** Unfairness regret; Pareto regret

As initialization step, the estimated probability $\hat{p}_i^d$ of success for each arm i in each objective d is 0.5 and the ϵ-Pareto optimal arm set A_ϵ^* contains all the arms in the arm set A. At each time step t, the annealing-Pareto trades-off between exploration and exploitation by using the decay factor parameter ϵ_{decay}, $\epsilon_{decay} \in (0,1)$ in the exponential decay ϵ_t to (step: 4). In each objective d, $d \in D$, the annealing-Pareto detects the optimal arm in that objective $i^{*,d}$, $i^{*,d} = \text{argmax}_{i\in A}\, \hat{p}_i^d$, where $\hat{p}_i^d$ is the estimated probability of success for arm i in the objective d (step: 7). The annealing-Pareto selects all the arms in the objective d that have estimated probability of success between $[p^{*,d} - \epsilon_t, p^{*,d}]$ and includes them in the corresponding selected arm set S^d (steps: 8-12), where $p^{*,d}$, $p^{*,d} = \max_{i\in A} p_i^d$ is the probability of success of the optimal arm $i^{*,d}$ in the objective d. The annealing-Pareto constructs the total selected arm set $S(t)$ at time step t by reunion of the selected arm set S^d in each objective d (step: 14). To keep track on the Pareto front, the annealing-Pareto uses Pareto dominance relation (step: 17) on the arms j that are elements in the previous ϵ-Pareto optimal arm set $A_\epsilon^*(t-1)$ and are not element in the total selected arm set $S(t)$. If the arm j is not dominated by all other arms, then this arm will be added to the total selected arm set $S(t)$ (step: 18). The annealing-Pareto updates its

ϵ-Pareto optimal arm set $A^*_\epsilon(t)$ to be the total selected arm set $S(t)$ (step: 21). It pulls uniformly at random one of the arms i^* that is an element in the ϵ-Pareto optimal arm set $A^*_\epsilon(t)$ (step: 22), observes the corresponding reward vector $\boldsymbol{r}_{i^*}$ and updates the estimated probability of success vector $\hat{\boldsymbol{p}}_{i^*}, \hat{\boldsymbol{p}}_{i^*} = [\hat{p}^1_{i^*}, \cdots, \hat{p}^D_{i^*}]^T$ of the pulled arm i^* (step: 23). Then, it calculates the Pareto and unfairness regrets (step: 25). This procedure is repeated until the end of playing L time steps which is the horizon of an experiment.

Experimental Result We experimentally compare Pareto-UCB1, Pareto-KG, Pareto Thompson sampling and annealing-Pareto on multi-objective Bernoulli distributions. Fig. 1 gives the average cumulative Pareto and unfairness regrets performances which are the average of 1000 experiments and the horizon of each experiment is 1000 time steps. The y-axis is either the average of the cumulative Pareto or unfairness regret performance. The x-axis is the horizon of each experiment. Fig. 1 shows the annealing-Pareto is the best algorithm and Pareto-UCB1 is the worst one. Pareto-KG performs better than Pareto Thompson sampling and worse than the annealing-Pareto. Pareto Thompson sampling performs better than Pareto-UCB1 and worse than Pareto-KG. The intuition is that the annealing-Pareto explores widely the optimal arms and the near to the optimal arms not all the arms, i.e. the optimal and the non-optimal arms, while Pareto Thompson sampling explores equally all the available arms. Pareto-KG and Pareto-UCB1 add an exploration bound and this exploration bound decreases to 0 after t time step. The exploration bound by Pareto-KG depends on all the arms, therefore it explores better than Pareto-UCB1.

Conclusion The annealing-Pareto is the best performing algorithm according to both the Pareto and unfairness regret performance measures.

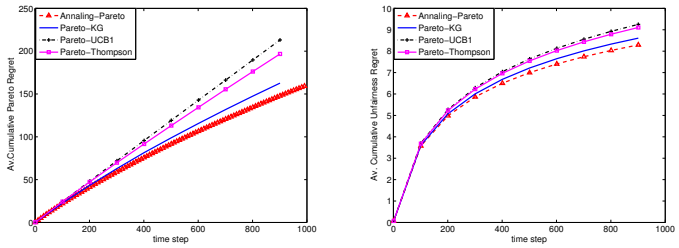

Fig. 1. Performance comparison on 5-objective, 20-armed with non-convex probability of success vector set. Left sub-figure shows the Pareto cumulative regret performance. Right sub-figure shows the unfairness regret performance.

References

1. Yahyaa, S.Q., Drugan, M.M., Manderick, M.: The Scalarized Multi-Objective Multi-Armed Bandit Problem: An Empirical Study of its Exploration vs. Exploration

Tradeoff. In: International Joint Conference on Neural Networks. (2014)
2. Drugan, M.M., Nowe, A.: Designing Multi-Objective Multi-Armed Bandits Algorithms: A study. In: International Joint Conference on Neural Networks. (2013)
3. Yahyaa, S.Q., Drugan, M.M., Manderick, B.: Knowledge Gradient for Multi-Objective Multi-Armed Bandit Algorithms. In: International Conference on Agents and Artificial Intelligence (ICAART). (2014)
4. Thompson, W.R.: On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. In: Biometrika. 25(3-4), 285–294 (1933)
5. Chapelle, O., Li, L.: An Empirical Evaluation of Thompson Sampling. In: Proc. Advances in Neural Information Processing Systems (NIPS) (2011).