Annealing-Pareto Multi-Objective Multi-Armed Bandit Algorithm

Saba Q. Yahyaa, M. Drugan and B. Manderick

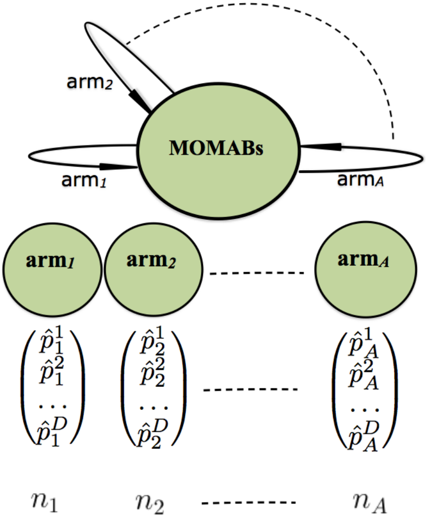
Vrije Universiteit Brussel (VUB)

September, 2014

Outline

Multi-Objective, Multi-Armed Bandits (MOMABs) Problem

The trade-off between exploration and exploitation



MOMABs Problem

The reward r_i^d, $r_i^d \in \{0, 1\}$ of an arm i in the objective d is drawn from a corresponding **Bernoulli probability distribution** .

$$
\begin{aligned}
\hat{p}_i^d(t) &= \frac{\alpha_i^d(t)}{\alpha_i^d(t) + \beta_i^d(t)} \\
\text{where} \quad \alpha_i^d(t) &= \alpha_i^d(t-1) + 1, \quad \text{if } r_i^d = 1 \\
\beta_i^d(t) &= \beta_i^d(t-1) + 1, \quad \text{if } r_i^d = 0
\end{aligned}
$$

$\alpha_i^d(t)$ is the number of successes, $\beta_i^d(t)$ is the number of failures and $\hat{p}_i^d(t)$ is the estimated probability of success of the arm i in the objective d at time step t.

- Arm i dominates or is better than j, $i \succ j$. If $\exists_d, \; i^d \succ j^d$ and $\forall_o, \; j \neq o, i^o \succeq j^o$.
- Arm i is incomparable with j, $i \parallel j$. If $i^d \succ j^d$ and $i^o \prec j^o$.
- Arm i is **not dominated** by j, $j \nsucc i$. Either $i \parallel j$ or $i \succ j$.

Using the above relationships, the Pareto front A^* set is:

- The subset of A, i.e. $A^* \subset A$.
- The set of arms that are **not dominated** by all other arms.
- The Pareto optimal arms A^* are **incomparable** with each other.

MOMABs Algorithms

Pareto-UCB1

- Play initially each arm i initial steps.
- Estimate the probability of success vector $\hat{\boldsymbol{p}}_i$, $\hat{\boldsymbol{p}}_i = [\hat{p}_i^1, \cdots, \hat{p}_i^D]^T$ for each arm i and add to each objective d an exploration bound ExpB_i^d.

$$\text{ExpB}_i^d(\text{UCB1}) = \sqrt{(2\ln(t\sqrt[4]{D|A^*|}))/N_i}$$

D is the number of objectives, $|A^*|$ is the number of optimal arms, t is the current time step, and N_i is the number of times arm i has been selected.

- Find the Pareto set A' set such that $\forall_j \in A'$, $\forall_k \notin A'$

$$\hat{\boldsymbol{p}}_k + \textbf{ExpB}_k \nsucc \hat{\boldsymbol{p}}_j + \textbf{ExpB}_j$$

- Choose uniformly at random an optimal arm i^*, $i^* \in A'$.
- Update the estimated probability of success vector $\hat{\boldsymbol{p}}_{i^*}$, and the number of times arm i^* is chosen N_{i^*}.

The exploration bound ExpB_i^d for each arm i and objective d.

$$\text{ExpB}_i^d(\text{KG}) = |A|D * (L - t) * v_i^d, \text{where}$$

$$v_i^d \triangleq \begin{cases} \frac{\alpha_i^d}{\alpha_i^d+\beta_i^d}\left(\frac{\alpha_i^d+1}{\alpha_i^d+\beta_i^d+1} - C_i^d\right), & \text{if} \frac{\alpha_i^d}{\alpha_i^d+\beta_i^d} \leq C_i^d < \frac{\alpha_i^d+1}{\alpha_i^d+\beta_i^d+1} \\ \frac{\beta_i^d}{\alpha_i^d+\beta_i^d}\left(C_i^d - \frac{\alpha_i^d}{\alpha_i^d+\beta_i^d+1}\right), & \text{if} \frac{\alpha_i^d}{\alpha_i^d+\beta_i^d+1} \leq C_i^d < \frac{\alpha_i^d}{\alpha_i^d+\beta_i^d} \\ 0 & \text{otherwise} \end{cases}$$

where

$$C_i^d = \max_{i \neq j} \alpha_j^d / (\alpha_j^d + \beta_j^d)$$

α_i^d, β_i^d, and v_i^d are the number of successes, number of failures, and the index of an arm i for dimension d, respectively. The total number of arms is $|A|$, L is the horizon of an experiment.

- Trade-off between exploration and exploitation by adding an exploration bound.
- The added exploration bound ExpB_i^d for the arm i in the objective d by Pareto-KG depends on all available arms in the objective d, each objective has different exploration bound.
- While, the added exploration bound ExpB_i^d for the arm i in the objective d by Pareto-UCB1 depends only on the arm i, each objective has the same exploration bound.

Pareto Thompson Sampling (PTS)

- Initiallly, the number of successes $\alpha_i^d = 1$ equals to the number of failures $\beta_i^d = 1$ for each arm i and objective d.
- For each arm i, in each objective d, the probability of selection P_i^d is sampled by using Beta distribution, $P_i^d = \text{Beta}(\alpha_i^d, \beta_i^d)$.
- Pareto set A' is found, such that $\forall_j \in A', \forall_k \notin A'$

$$\boldsymbol{P}_k \not\succ \boldsymbol{P}_j$$

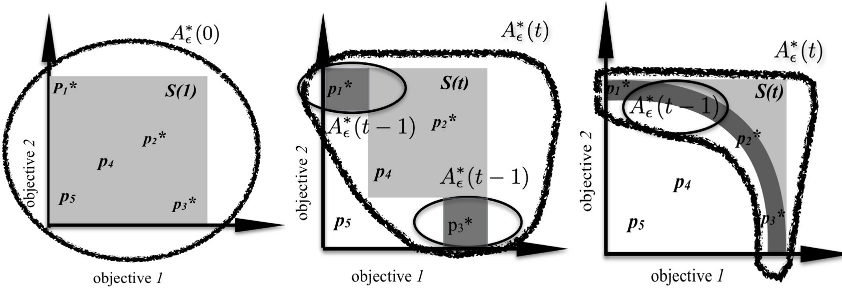
- Choose uniformly at random an optimal arm $i^*, \ i^* \in A'$.
- Observes $\boldsymbol{r}_{i^*} = [r_{i^*}^1, \cdots, r_{i^*}^D]^T$ and updates the number of successes $\alpha_{i^*}^d$ and the number of failures $\beta_{i^*}^d$ for each d.

The Annealing Pareto Algorithm

Annealing-Pareto has a specific mechanism to control the trade-off between exploration and exploitation. It uses an exponential decay ϵ_t, $\epsilon_t = \epsilon^t_{decay}/(|A|D)$ with Pareto dominance relation, where ϵ_{decay} is the decay factor parameter.

- Play initially each arm i initial steps, initialize the ϵ-Pareto front set $A^*_\epsilon(0) = A$.
- at each t
  - Set the decay parameter $\epsilon_t = \epsilon^t_{decay}/(|A||D|)$
  - For each objective $d \in D$: $i \in S(t)^d$ if $\hat{\mu}^d_i \in [\hat{\mu}^{*,d} - \epsilon_t, \hat{\mu}^{*,d}]$,
  - $S(t) \leftarrow S^1(t) \cup S^2(t) \cup \cdots \cup S^D(t)$
  - $S_{diff} \leftarrow A^*_\epsilon(t-1) - S(t)$
  - For arm $j \in S_{diff}$; If $\hat{\boldsymbol{\mu}}_k \not\succ \hat{\boldsymbol{\mu}}_j, \forall_k \in A$, then $S(t) \leftarrow S(t) \cup j$
  - $A^*_\epsilon(t) \leftarrow S(t)$
  - Select an optimal arm i^* uniformly, at random from $A^*_\epsilon(t)$; Update: $\hat{\boldsymbol{\mu}}_{i^*}$, N_{i^*}; Compute: the unfairness and Pareto regrets.

The Annealing Pareto Algorithm



a. At $t = 1$ *b*. At $t > 1$ *c*. At $t >> 1$

Figure : The dynamic of the annealing-Pareto algorithm.

Performance in MOMABs

1. **Pareto regret** measures the distance between a probability of success vector of an arm i that is pulled at time step t and the Pareto front A^*.
2. **Unfairness regret** $R_{SE}(t)$ is the Shannon entropy which is a measure of disorder on the Pareto front A^*. The higher the entropy, the higher the disorder.

We experimentally compare Pareto-UCB1, Pareto-KG, Pareto Thompson sampling, and annealing-Pareto algorithms.

The performance measures are:

1. The average Pareto and the cumulative average Pareto regret at each time step which are averaged of M experiments.
2. The average unfairness and the cumulative average unfairness regret at each time step which are averaged of M experiments.

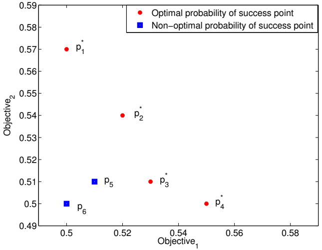
Experimental Comparison

Experiment 1: 2-objectives, 6-arms

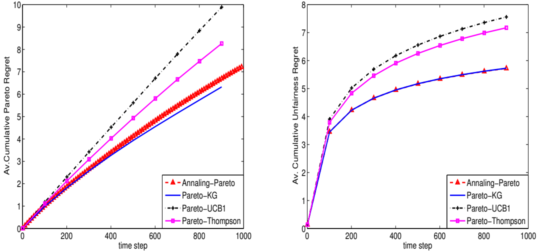
The true probability of success vector set is

$$\{\boldsymbol{p}_1 = \begin{bmatrix} .55 \\ .5 \end{bmatrix}, \boldsymbol{p}_2 = \begin{bmatrix} .53 \\ .51 \end{bmatrix}, \boldsymbol{p}_3 = \begin{bmatrix} .52 \\ .54 \end{bmatrix}, \boldsymbol{p}_4 = \begin{bmatrix} .5 \\ .57 \end{bmatrix}, \boldsymbol{p}_5 = \begin{bmatrix} .51 \\ .51 \end{bmatrix}, \boldsymbol{p}_6 = \begin{bmatrix} .5 \\ .5 \end{bmatrix}\}$$

Note that the Pareto front set is $A^* = \{a_1^*, a_2^*, a_3^*, a_4^*\}$



Experimental Comparison

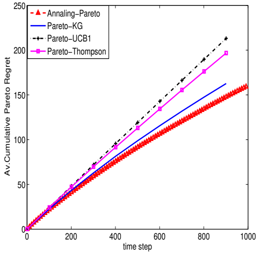


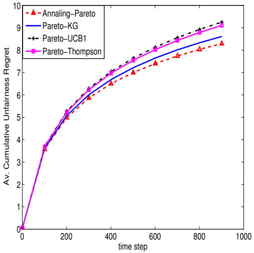
a. Pareto cumulative regret   *b*. Unfairness cumulative regret

Figure : Performance comparison on 2-objective, 6-armed with non-convex probability of success vector set. Sub-figure *a* shows the average Pareto cumulative regret. Sub-figure *b* shows the average cumulative unfairness regret.

Experimental Comparison

Experiment 2: 5-objectives, 20-arms We add extra 3 objectives and 14 arms to Experiment 1, resulting in 5-objective, 20-armed. Pareto front contains 7 optimal arms.





a. Pareto cumulative regret   *b*. Unfairness cumulative regret

Figure : Performance comparison on 5-objective, 20-armed with non-convex probability of success vector set. The average cumulative Pareto and unfairness regret performances are shown in sub-figures *a* and *b*, respectively.

Overview of the state of the Art

1. We extended Pareto Thompson sampling to the MOMAB.
2. We proposed annealing-Pareto algorithm.
3. We proposed using the entropy measure as a performance measure in the MOMAB.
4. We studied empirically the trade-off between exploration and exploitation in the MOMAB, where we compared Pareto-KG, Pareto-UCB1, Pareto Thompson sampling and the annealing-Pareto.

Questions

Thanks For Your Attention!