Annealing-Pareto Multi-Objective Multi-Armed Bandit Algorithm

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

Vrije Universiteit Brussel (VUB)

September, 2014
Outline

1. Multi-objective, Multi-armed Bandits (MOMABs)
2. Pareto Thompson Sampling
3. The Annealing Pareto Algorithm
4. Performance in MOMABs
5. Experimental Comparison
6. Contribution
Multi-Objective, Multi-Armed Bandits (MOMABs) Problem

The trade-off between exploration and exploitation

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

\[
\hat{p}_i^d(t) = \frac{\alpha_i^d(t)}{\alpha_i^d(t) + \beta_i^d(t)}
\]

where

\[
\alpha_i^d(t) = \alpha_i^d(t-1) + 1, \quad \text{if } r_i^d = 1
\]

\[
\beta_i^d(t) = \beta_i^d(t-1) + 1, \quad \text{if } r_i^d = 0
\]

$\alpha_i^d(t)$ 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 $t$. 
MOMABs Problem
The Pareto Dominance Relations

- 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 $j$, $j \not\succ i$. 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 $\hat{p}_i$, $\hat{p}_i = [\hat{p}_i^1, \ldots, \hat{p}_i^D]^T$ for each arm $i$ and add to each objective $d$ an exploration bound $\text{ExpB}_i^d$.

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

$D$ 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 $\forall j \in A'$, $\forall k \notin A'$

$$\hat{p}_k + \text{ExpB}_k \not\preceq \hat{p}_j + \text{ExpB}_j$$

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

Saba Q. Yahyaa, M. Drugan and B. Manderick (Vrije Universiteit Brussel (VUB))

Annealing-Pareto Multi-Objective Multi-Armed Bandit Algorithm

Pareto Thompson Sampling
The Annealing Pareto Algorithm
The exploration bound $\text{ExpB}^d_i$ for each arm $i$ and objective $d$.

$$\text{ExpB}^d_i(\text{KG}) = |A| D \times (L - t) \times v^d_i,$$

where

$$v^d_i = \begin{cases} 
\frac{\alpha^d_i}{\alpha^d_i + \beta^d_i} \left( \frac{\alpha^d_i + 1}{\alpha^d_i + \beta^d_i + 1} - C^d_i \right), & \text{if } \frac{\alpha^d_i}{\alpha^d_i + \beta^d_i} \leq C^d_i < \frac{\alpha^d_i + 1}{\alpha^d_i + \beta^d_i + 1} \\
\frac{\beta^d_i}{\alpha^d_i + \beta^d_i} \left( C^d_i - \frac{\alpha^d_i}{\alpha^d_i + \beta^d_i + 1} \right), & \text{if } \frac{\alpha^d_i + 1}{\alpha^d_i + \beta^d_i + 1} \leq C^d_i < \frac{\alpha^d_i}{\alpha^d_i + \beta^d_i} \\
0, & \text{otherwise}
\end{cases}$$

where

$$C^d_i = \max_{i \neq j} \frac{\alpha^d_i}{\alpha^d_i + \beta^d_i}$$

$\alpha^d_i$, $\beta^d_i$, and $v^d_i$ 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 $\text{ExpB}_i^d$ 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 $\text{ExpB}_i^d$ 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)

- Initially, the number of successes $\alpha^d_i = 1$ equals to the number of failures $\beta^d_i = 1$ for each arm $i$ and objective $d$.

- For each arm $i$, in each objective $d$, the probability of selection $P^d_i$ is sampled by using Beta distribution, $P^d_i = \text{Beta}(\alpha^d_i, \beta^d_i)$.

- Pareto set $A'$ is found, such that $\forall j \in A'$, $\forall k \notin A'$

\[
P_k \not\approx P_j
\]

- Choose uniformly at random an optimal arm $i^*$, $i^* \in A'$.

- Observes $r_{i^*} = [r^{1}_{i^*}, \cdots, r^{D}_{i^*}]^T$ and updates the number of successes $\alpha^{d}_{i^*}$ and the number of failures $\beta^{d}_{i^*}$ 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 $\epsilon_t$, $\epsilon_t = \epsilon_{\text{decay}}/(|A|D)$ with Pareto dominance relation, where $\epsilon_{\text{decay}}$ is the decay factor parameter.

- Play initially each arm $i$ initial steps, initialize the $\epsilon$-Pareto front set $A_{\epsilon}^*(0) = A$.
- at each $t$
  - Set the decay parameter $\epsilon_t = \epsilon_{\text{decay}}/(|A||D|)$
  - For each objective $d \in D$: $i \in S(t)^d$ if $\hat{\mu}_i^d \in [\hat{\mu}^*,d - \epsilon_t, \hat{\mu}^*,d]$
  - $S(t) \leftarrow S^1(t) \cup S^2(t) \cup \ldots \cup S^D(t)$
  - $S_{\text{diff}} \leftarrow A_{\epsilon}^*(t - 1) - S(t)$
  - For arm $j \in S_{\text{diff}}$: If $\hat{\mu}_k \neq \hat{\mu}_j$, $\forall k \in A$, 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: $\hat{\mu}_{i^*}, N_{i^*}$; Compute: the unfairness and Pareto regrets.
The Annealing Pareto Algorithm

Figure: The dynamic of the annealing-Pareto algorithm.

- **a.** At $t = 1$
- **b.** At $t > 1$
- **c.** At $t >> 1$
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.
Annealing-Pareto Multi-Objective Multi-Armed Bandit Algorithm

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

Multi-objective, Multi-armed Bandits (MOMABs)
Pareto Thompson Sampling
The Annealing Pareto Algorithm

Experimental Comparison

Experiment 1: 2-objectives, 6-arms

The true probability of success vector set is

\[ \{ p_1 = \begin{bmatrix} 0.55 \\ 0.5 \end{bmatrix}, p_2 = \begin{bmatrix} 0.53 \\ 0.51 \end{bmatrix}, \]

\[ p_3 = \begin{bmatrix} 0.52 \\ 0.54 \end{bmatrix}, p_4 = \begin{bmatrix} 0.5 \\ 0.57 \end{bmatrix}, \]

\[ p_5 = \begin{bmatrix} 0.51 \\ 0.51 \end{bmatrix}, p_6 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \} \]

Note that the Pareto front set is \( A^* = \{ a_1^*, a_2^*, a_3^*, a_4^* \} \)
Experimental Comparison

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.

a. Pareto cumulative regret   b. Unfairness cumulative 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.

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.

*a. Pareto cumulative regret  
*b. Unfairness cumulative regret
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.
Thanks For Your Attention!