Corporate Social Responsibility via Multi-Armed Bandits

Nwemadji Tiako Arsene Gibbs

Reinforcement Learning (RL) 2021-2022, QLS Diploma, ICTP.

Prof. Dr. Celani and Dr. Panizon

August 1, 2022

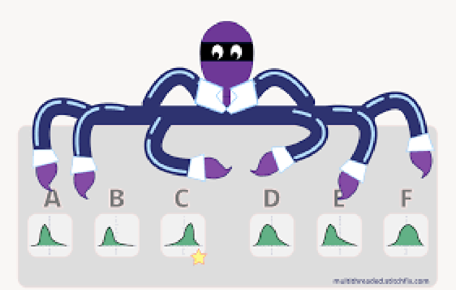
Introduction

Corporate Social Responsibility (CSR), is a form of self-regulation that reflects a business's accountability and commitment to contributing to environmental and social measures.

Figure: Coporate Social Responsability.

Multi-Armed Bandit

Multi-armed Bandits.



Setup and Notation

1. K- arms
2. T- number of time steps
3. r_t reward obtained in round t
4. $N_{i,t}$ number of times arm i is pull after t round
5. $\boldsymbol{\mu}$ reward vector, μ^* the expected reward of the optimal arm.
6. $\Delta_i = \mu^* - \mu_i$, reward gap between an arm i and the optimal arm.
7. $f : [0,1]^K \to [0,1]^K$, fairness function.
8. λ transfer cost.

Requirements and Examples of the Fairness Function

Requirements:

1. Bounded $\sum_{i=1}^{K} f(\boldsymbol{\mu})_i \leq 1$.
2. Lipschitz - $\exists L > 0 s.t. \forall \boldsymbol{\mu}^{'}$ and $\boldsymbol{\mu}, ||f(\boldsymbol{\mu}) - f(\boldsymbol{\mu}^{'}||_1 \leq L||\boldsymbol{\mu}^{'} - \boldsymbol{\mu}||_1$

Examples:

- Uniform $f^{\text{uni}}(\boldsymbol{\mu})_i := \frac{1}{K}$
- Linear $f^{\text{lin}}(\boldsymbol{\mu})_i := \frac{\mu_i}{K}$
- Step Uniform $f^{\text{stp}}(\boldsymbol{\mu}, d)_i := \frac{\theta(\mu_i - d)}{K}$
- Step Linear $f^{\text{stp}}(\boldsymbol{\mu}, d)_i := \mu_i \frac{\theta(\mu_i - d)}{K}$
- Softmax $f^{\text{sft}}(\boldsymbol{\mu}, c)_i := \frac{e^{c\mu_i}}{\sum_{j=1}^{K} e^{c\mu_j}}$.

Utility Function

With the help of the previous setting and given an algorithm ALG, the utility of the decision-maker is formally written as follow:

$$\mathcal{U}_{\lambda,f}(ALG;T) := \sum_{t=1}^{T} r_{i_t} - \lambda \sum_{i=1}^{k} \max\{Tf(\mu)_i - N_{i,T}, 0\}. \tag{1}$$

The performance of the algorithm is given by:

$$\mathcal{R}_{\lambda,f}(ALG;T) := \mathbb{E}(\mathcal{U}_{\lambda,f}(OPT;T)) - \mathbb{E}(\mathcal{U}_{\lambda,f}(ALG;T)), \tag{2}$$

where OPT is an algorithm maximizing the utility $\mathcal{U}_{\lambda,f}(OPT;T)$.

Lemma 1. *Fix an arbitrary instance* $\langle K, T, \boldsymbol{\mu}, f, \lambda \rangle$ *and let OPT be an optimal algorithm for that instance. For every sub-optimal arm* i*, if* $\Delta_i < \lambda$ *then OPT pulls* i *exactly* $Tf(\boldsymbol{\mu})_i$ *times; if* $\Delta_i > \lambda$ *OPT does not pull* i *at all. If* $\Delta_i = \lambda$*, OPT pulls arm* i *between zero and* $Tf(\boldsymbol{\mu})_i$ *times.*

Theorem 1. *Fix any arbitrary instance* $\langle K, T, \boldsymbol{\mu}, f, \lambda \rangle$ *of R-O MAB, and let* $N = 8L^{2/3}T^{2/3}\log^{1/3} T$*. Algorithm 1 has a regret of* $O(KL^{2/3}T^{2/3}\log^{1/3} T)$.

Algorithm1: Fairness-Aware-ETC

0- Input: N - number of exploration rounds

1- **for** $i = 1, \ldots, K$ **do**

2 pull arm i for N rounds

3- **for** $i = 1, \ldots, K$ **do**

4- **if** $\hat{\Delta}_i < \lambda$ **do**

5- pull arm i for $max\{Tf(\hat{\boldsymbol{\mu}}^{'})_i - N, 0\}$ rounds

6-pull an arbitrary arm from $\arg\max_{i \in [K]} \hat{\mu}$ until the execution ends.

Example

Example: Consider a multi-armed bandit for which $\boldsymbol{\mu} = (1, \frac{1}{2}, \frac{1}{3}), \lambda = 0.7$ and $f(\boldsymbol{\mu})_i = \frac{1}{3}$.

The classical ETC algorithm after discover that arm 1 is better will give:

$$T - 0.7T(\frac{1}{3} + \frac{1}{3}) = 0.533T.$$

While the optimal algorithm 1 after estimating the expected reward will gives:

$$\frac{T}{3}.1 + \frac{T}{3}.\frac{1}{2} + \frac{T}{3}.\frac{1}{3} = \frac{11}{18}T = 0.611T.$$

In this case the regret is:

$$R_{0.7} = 0.077T.$$

Algorithm 2

Theorem 2. *Fix any arbitrary instance* $\langle K, T, \boldsymbol{\mu}, f, \lambda \rangle$ *of R-O MAB, and let* $\alpha = K^{2/3} L^{2/3} T^{-1/3} \log^{1/3} T, \beta = T^{-1/3} \log^{1/3} T$. *Then Algorithm 2 has a regret of* $O(K^{5/3} L^{2/3} T^{2/3} \log^{1/3} T)$.

Algorithm2: Self-regulated Utility Maximization

0- Input: Black-box bandit algorithm ALG, allowed approximation error parameter α and β.
1- t = 1
2- Initialize arms' data $-N_i = 0, LCB(\Delta_i) = 0, UCB(\Delta_i) = 1$ for all $i \in [K]$
3- $C_1 = [0,1]^K$ // `Hyper-cube of` μ `values`
4-**While** $\exists i \in [K]$ `s.t.` $UCB(\Delta_i) > \lambda + \beta$ `and` $LCB(_i) < \lambda - \beta$ **do**// `Phase 1`
5- Pull all arms once, update t, counters, confidence bounds and C_t
6-**While** $\exists i \in [K]$ `s.t.` $\max_{\boldsymbol{\mu}' \in \boldsymbol{C_t}} f(\boldsymbol{\mu}')_i - \min_{\boldsymbol{\mu}' \in \boldsymbol{C_t}} f(\boldsymbol{\mu}')_i > \alpha$ `and` $LCB(\Delta_i) < \lambda$ **do**// `Phase 2:`
7- Pull all arms once, update t, counters, confidence bounds and C_t
8-**While** $\exists i \in [K]$ `s.t.` $LCB(\Delta_i) < \lambda$ `and` $t < T$ **do** // `Phase 3`
9- Pull all arm i the minimal number of times so $N_i \geq Tf(\hat{\boldsymbol{\mu}}')_i$, update t and counters.
10-Invoke ALG for the remaining rounds // `Phase 4`

Setup of the Simulation

- 6 arms, $\mu_i = 0.2 + (i-1)X0.15$
- Transfer costs: 0, 0.4, 0.8
- Fairness function f^{sft}
- Time Horizon: 10K, 50K, 100K and 200K
- Each experiment was repeated 50 times
- In case phase 1 or phase 2 had been made the black box algorithm is an algorithm which pull the arm with the highest expected reward
- If none exploration had been done when phase 4 start, Epsilon greedy algorithm was used as black-box algorithm.

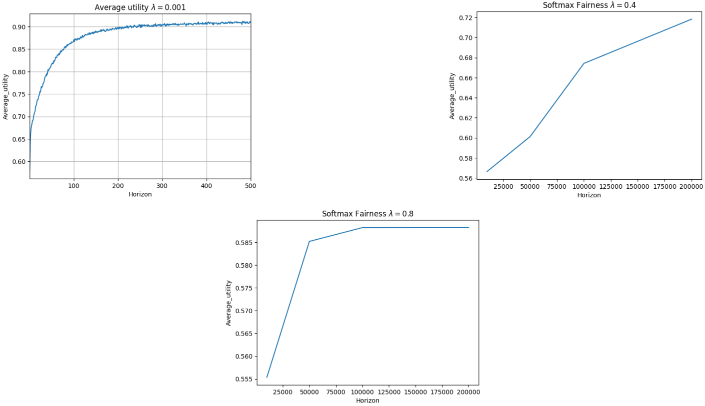


Figure: Average Utility for different transfer costs respect to the time horizon.

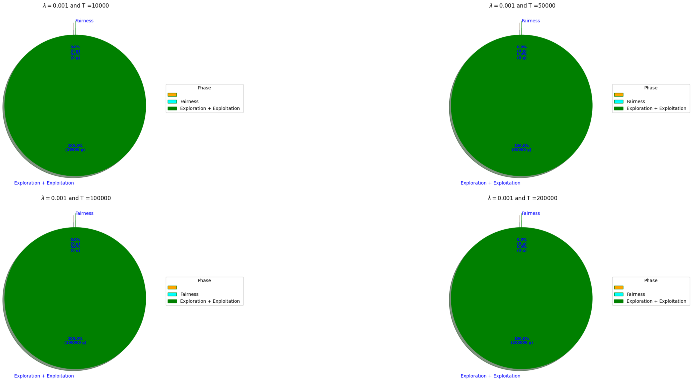


Figure: Round distribution per phase with $\lambda = 0.001$ for softmax.

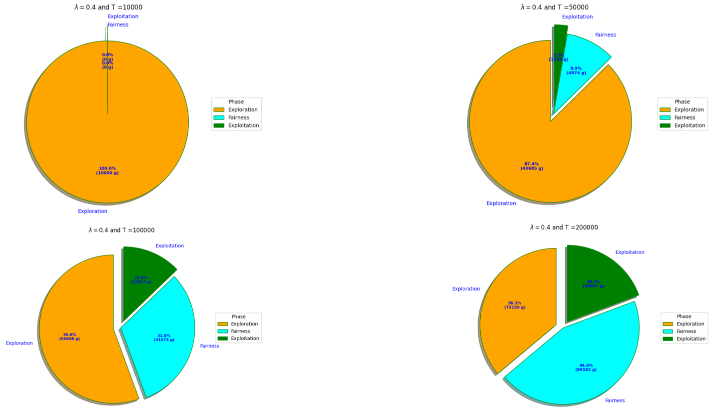


Figure: Round distribution per phase with $\lambda = 0.4$ for softmax.

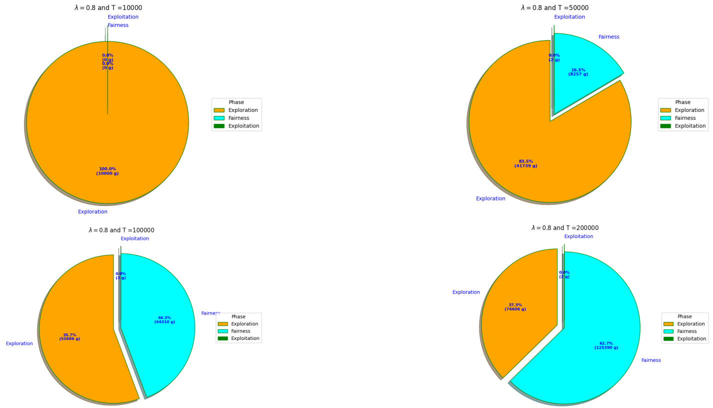


Figure: Round distribution per phase with $\lambda = 0.8$ for softmax.

Conclusion:

My goal was to recovered the numerical result obtain on the paper corporate social responsability via multi-armed bandit by Rom, Ben-Porat et Sharit 2021. Despite the fact that I did not used the exact parameter they used on the paper, I succeed to obtain most of theirs results.

Question and Discussion

QUESTION