Performance Evaluation on Classical Algorithms for Multi-Armed Bandit SI140@2021Fall final project

Cheng Peng*

January 16, 2022

Contents

*pengcheng2@shanghaitech.edu.cn ID=2020533068

1 The Overall Performance Comparison

We made a bar chart1 based on the simulation, where the length of each bar is the gap[1] between the oracle value and the average reward of the corresponding algorithm.
Generally speaking, we have TS > UCB > $\epsilon-$Greedy in the sense of average reward, when the parameter is tuned.

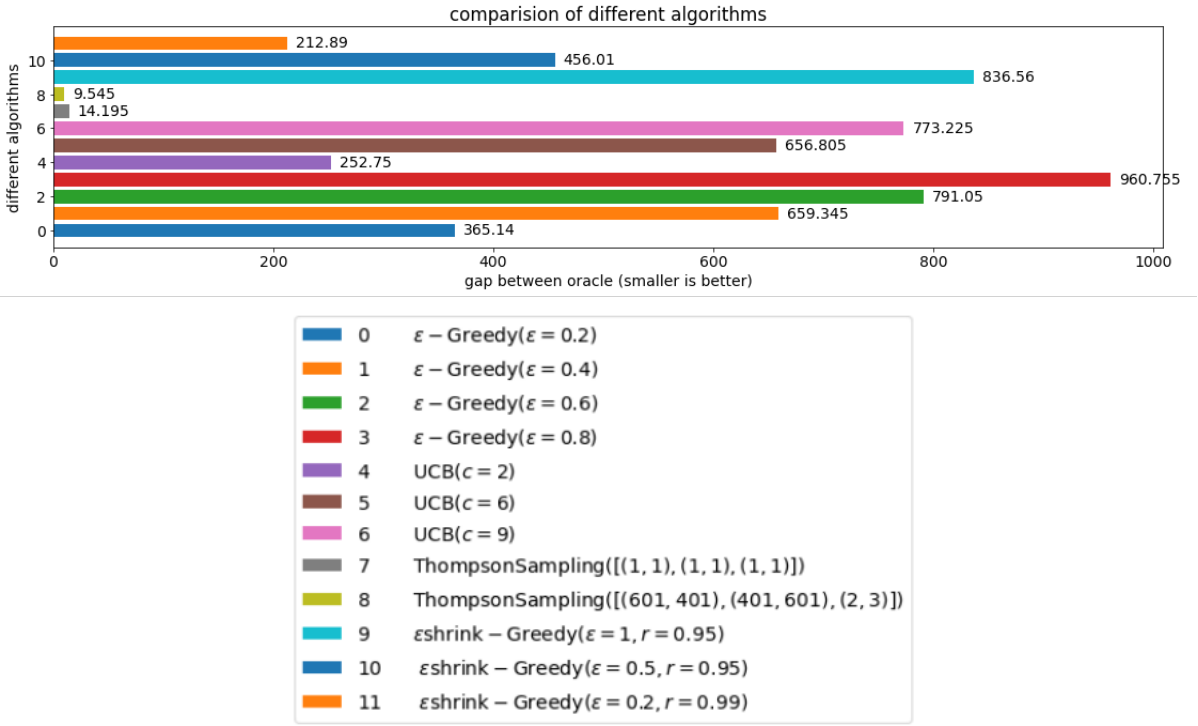

Figure 1: Bar Chart: the performance of each algorithms

To better understand the behavior of each algorithm, the knowledge used to make decision on each step is recorded. We store the estimated θ_js and organize them into scatter plots1.
We can see that *UCB* and *TS* can quickly learn the actual distribution, while ϵ-*Greedy* algorithm takes a long time to converge.
The *Simulated Annealing* algorithm failed to learn the behind distribution. We doubt that it is caused by too short explore phase i.e. ϵ shrinks too fast.

[1] This is also called *regret*, which is the goal to minimize.

Figure 2: The learning curve of different algorithms

2 The Exploration-Exploitation Trade-off and Impact of Parameter Selection

2.1 The E-E trade-off

In MAB problem settings, the agent do not have the real distribution of reward behind each arm, so the optimization and learning process have to be carried out simultaneously.
After some arm pulls, the algorithm has some knowledge on the distributions. We have decide to pull the arm that gives maximal expected reward based on acquired knowledge i.e. exploit or to pull other arms to gain more accurate information of the distributions i.e. explore.
There, we face a dilemma: To maximize the aggregated reward, we want to pull the optimal arm as much as possible. However, to find the optimal arm, we have to spend more chances pulling every arms.

Clearly, a good balance between exploration and exploitation is the key to ideal performance. We will see how the classical algorithms strives to find a proper balance.

2.2 ϵ in ϵ-Greedy algorithm

The ϵ-Greedy algorithm, which is shown in 2.2, is the most straight-forward way to balance exploration and exploitation.

Algorithm 1 $\epsilon-$greedy Algorithm

Initialize $\hat{\theta}(j) = 0, \text{count}(j) = 0, j \in \{1, 2, 3\}$

1: **for** $t = 1, 2 \ldots, N$ **do**

2:

$$I(t) \leftarrow \begin{cases} \underset{j\in\{1,2,3\}}{\arg\max}\ \hat{\theta}(j) & w.p.\ 1-\epsilon \\ \text{randomly chosen from}\{1,2,3\} & w.p.\ \epsilon \end{cases}$$

3: $\text{count}(I(t)) \leftarrow \text{count}(I(t)) + 1$

4: $\hat{\theta}(I(t)) \leftarrow \hat{\theta}(I(t)) + \frac{1}{\text{count}(I(t))}\left[r_{I(t)} - \hat{\theta}(I(t))\right]$

5: **end for**

Roughly, we make $N\epsilon$ trials to learn the distribution and $N(1-\epsilon)$ pulls to maximize the rewards.
With larger ϵ, the algorithm tends to explore more while it prefers to make exploitative decisions.
Finding a proper value of ϵ is somewhat hard in real-world applications. This makes the algorithm less flexible since it can not change the balance between exploration and exploitation adaptively.

We can enhance the algorithm by combining ϵ-Greedy with Simulated Annealing, a traditional algorithm framework for discrete/continuous optimization problems.
We make ϵ decrease as we gain more information of the distributions, this is implemented by making $\epsilon_t = \epsilon_0 r^t$ where r is a real number between 0 and 1. Typically, we choose $r = 0.95$ or $r = 0.99$.
The parameter r controlls the time that the algorithm enter a pure-exploitation phase. The agent can learn more about the distribution when having smaller r.

2.3 c in Upper Confidence Bound algorithm

UCB algorithm, first introduce in [ACF02], can be described by the following pseudo code2.3.

Algorithm 2 UCB Algorithm

for $t = 1, 2, 3$ **do**
    $I(t) \leftarrow t$
    $\text{count}(I(t)) \leftarrow 1$
    $\hat{\theta}(I(t)) \leftarrow r_{I(t)}$
end for
for $t = 4, \ldots, N$ **do**

$$I(t) \leftarrow \underset{j \in \{1,2,3\}}{\arg\max} \left(\hat{\theta}(j) + c \cdot \sqrt{\frac{2 \log t}{\text{count}(j)}} \right)$$

    $\text{count}(I(t)) \leftarrow \text{count}(I(t)) + 1$
    $\hat{\theta}(I(t)) \leftarrow \hat{\theta}(I(t)) + \frac{1}{\text{count}(I(t))} \left[r_{I(t)} - \hat{\theta}(I(t)) \right]$
end for

UCB algorithm employes interval estimation to find the parameter θ_js. We can use Hoeffding bound to find the confidence interval once the confidence level is fixed.
In time slot t, $\hat{\theta}_j + c\sqrt{\frac{2 \log t}{\text{count}_j}}$ is the upper bound of the confidence interval $[\hat{\theta} - \delta, \hat{\theta} + \delta]$. The parameter c is determined by the confidence level.

Consider the scenario where arm k has lower $\hat{\theta}_k$ than other arms while count_k is small. This indicate we haven't gain much information on that arm, we can pull arm k for exploration. In this case, $\sqrt{\frac{2 \log t}{\text{count}_k}}$ is lager than every other arm, allowing UCB algorithm to pick k with greater chances.
As time proceeds and the confidence interval shrinks, the algorithm has enough knowledge on the distributions of arms and makes decisions mainly based on $\hat{\theta}$ i.e. tends to exploit the best arm. This is the way that UCB resolves E-E dilemma.

c correspondes to the confidence level. With larger c, we can enable the algorithm to make more exploration even t is relatively large. This cause the algorithm to converge slowly as 1 shows.
In constrast, with smaller c, we are more confident for the estimation so we tend to exploit more. In the learning history curve, we can observe that $\hat{\theta}$ converges to θ rapidly.

2.4 α_i, β_is in Thompson Sampling algorithm

We found a paper [AG12] that covers the theoretical analysis of Thompson Sampling algorithm2.4.

Algorithm 3 Thompson sampling Algorithm

Initialize Beta parameter $(\alpha_j, \beta_j), j \in \{1, 2, 3\}$

1: **for** $t = 1, 2 \ldots, N$ **do**
2: *# Sample model*
3: **for** $j \in \{1, 2, 3\}$ **do**
4: Sample $\hat{\theta}(j) \sim \text{Beta}(\alpha_j, \beta_j)$
5: **end for**
6: *# Select and pull the arm*

$$I(t) \leftarrow \arg\max_{j \in \{1,2,3\}} \hat{\theta}(j)$$

7: *# Update the distribution*

$$\alpha_{I(t)} \leftarrow \alpha_{I(t)} + r_{I(t)}$$
$$\beta_{I(t)} \leftarrow \beta_{I(t)} + 1 - r_{I(t)}$$

8: **end for**

The main idea of Thompson Sampling is quite simple: use beta-binomial conjugacy to learn and estimate the parameter θ_js. The (α_j, β_j)s stands for the pseudo-count priors of each arm.

Intuitively, if we run TS for enough iterations, the $\hat{\theta}_j = \frac{a_j}{a_j + b_j}$ should eventually converge to θ_j. This can be justified by the learning curve 1. As we can see, the curves plateau and stays around the true value of θ.

However, the prior knowledge still play a crucial role in short runs. For some j, if the prior $\alpha_j + \beta_j$ is large, then we rely more on the prior knowledge rather than the actual rewards. We call this a strong prior.
When we have a strong prior that is far from the true distribution, Thompson Sampling takes great amount of time to learn the actual value of θ. See 1 TS with $\alpha_1, \beta_1 = 601, 401$.
When we have a prior that is close the true distribution, TS performs best. In that case, TS converges at a significant spped and outperforms every other algorithsm. See 1 the first plot for TS.

3 Extension: Dependent Arms

Now we drop the assumption that the reward distribution of arms are independent, which is not realistic. We found several research papers that cover this variant, one with most citation is [PCA07]. We will use the same model

In this variant, we face a MAB with N arms which are grouped into k clusters, where dependencies are allow within a group.

Let $[i]$ be the cluster of arm i and $C[i]$ be the cluster containing i. Let $s_i(t)$ be the number of times arm i gives raeward when pulled "success", and $f_i(t)$ be the number of "failures". We will have

$$s_i(t) \mid \theta_i \sim \text{Bin}(s_i(t) + f_i(t), \theta_i) \quad \theta_i \sim \eta(\pi_{[}i])$$

Where η is the probability distribution of θ and the parameter $\pi_{[}i]$ contains the information of the dependencies within $C[i]$.

We can apply a two-level UCB to exploit the dependencies. In each step, we compute the expected reward and the variance of each group and use UCB to decide the group. Then, we apply UCB again to pick the optimal arms within that group.

To estimate the reward and variance of one group, we take mean among all the arms within that group.

Our two-level UCB is better than plain UCB since it avoids exploring arms in the same group. This gives more "effective" pulls and should generate better performance.

Not Having enough time, I can not implement and run benchmark for this variant.

4 Extension: Bounded Cost

To make the MAB model more practical, we introduce bounded random cost on each arm pull. Multiple research projects have digged into this variant, we picked a popular one [Din+13]. We will use the cost model in this paper and demonstrate the strategy developed from it.

This variant is call MAB-BV[2]. In this problem pulling arm i in time slot t gives reward $r_i(t) \sim \text{Bern}(\theta_i)$ and cost $c_i(t) \sim \text{DUnif}(\frac{1}{m}, \frac{2}{m} \dots \frac{m}{m})$. Moreover, the agent has a limited budget B, which satisfies $m \mid B$.

We modify the UCB algorithm to take the cost and budget into account. In time step t before out-of-budget, we pick

$$\arg\max_j \left(\hat{\theta}_j + \frac{1 + \frac{1}{\lambda}\sqrt{\frac{\log t}{\text{count}_j}}}{\lambda - \sqrt{\frac{\log t}{\text{count}_j}}} \right)$$

where $\lambda \leq \min_j \mathbb{E}(c_j)$ charaterize the lower bound of expected cost.

UCB-BV drops the parameter c in orginal UCB. Instead we now use λ, which corresponds to the average cost, to balance exploration and exploitation.
If we are expecting high cost on each pull i.e. large λ, then we tend to exploit the best arm. On the contary, if we assume that the cost for pulling one arm is low, then we prefer to pull an arm that has less count.

Not Having enough time, I can not implement and run benchmark for this variant.

[2] BV stands for *budget constraint and variable costs*

reference

[ACF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2 (2002), pp. 235–256.

[PCA07] Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal. "Multi-armed bandit problems with dependent arms". In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 721–728.

[AG12] Shipra Agrawal and Navin Goyal. "Analysis of thompson sampling for the multi-armed bandit problem". In: *Conference on learning theory*. JMLR Workshop and Conference Proceedings. 2012, pp. 39–1.

[Din+13] Wenkui Ding et al. "Multi-armed bandit with budget constraint and variable costs". In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*. 2013.