Multi-Armed Bandits

Alexander Mitchell

November 2018

1 Introduction

Multi-armed bandits can be approximated as a simple form of reinforcement learning. The aim is to maximise the cumulative reward from unknown one-armed bandits during a known number of iterations T. The bandits give a reward of 1 or 0. Four algorithms are analysed in three separate experimental cases.

The four algorithms are uniform, ϵ-greedy, UCB-1 and Thompson sampling. All the algorithms explore then exploit the multi-armed bandits in different ways. The three experiments are stationary, adversarial and stochastic with lower bound probability of reward. Stationary bandits have a fixed time invariant probability of a reward. Adversarial bandits return a reward with two different probabilities, which alternate for an unknown number of iterations. The lower-bound probabilities are set such that the bandits have probabilities p, while another has $p^* = p - 1/\sqrt{T}$, where T is the total number of iterations.

2 Algorithms

The empirical accuracy of the algorithms is estimated using the algorithmic regret.

$$R(T) = \mu^* - \sum_{t=1}^{T} \mu_t \tag{1}$$

This the optimal mean reward for pulling the correct bandit minus the mean reward from for the bandit which was actually pulled. Theoretical lower bounds for regret or estimated regret are presented in [2].

2.1 Uniform Algorithm

Uniform algorithm explore all the arms equally before choosing the one with the highest mean reward. This is the simplest of the explore and exploit algorithms, and explores $N = (\frac{2T\sqrt{2logT}}{k})^{\frac{2}{3}}$ time steps. This requires knowledge of the total number of iterations to optimally explore. The regret is of order $R(t) \leq T^{\frac{2}{3}} O(k \log(T))^{\frac{1}{3}}$.

2.2 e-greedy

ϵ-greedy chooses from a Bernoulli distribution and if returns 1, sample from a uniform distribution and pull this bandit's arm. If the distribution returns 0, sample from the arm with the best mean. The probability of pulling an arm is ϵ. This is either set constant or decays. The experiments outline below, $\epsilon = t^{-\frac{1}{3}}$. Hence, the algorithm continues to explore and exploit. The regret is of order $\mathbb{E}(R(t)) \leq t^{\frac{2}{3}} O(k \log(t))^{\frac{1}{3}}$.

2.3 UCB-1

This algorithm chooses a bandit based upon the maximisation of both and exploitation term and exploration. Hence, if one of the bandits' estimated mean reward is higher than the other bandits, the algorithm will select this bandit. If a bandit is pulled, but gives a low mean reward, the overall cost of the bandit reduces. Hence this bandit is less likely to be picked. The arm to

select is $a \in argmax\{\hat{\mu}_a(t) + r_t(a)\}$, where $\hat{\mu}_a(t)$ is the estimated mean reward of this bandit, and $r_t(a) = \sqrt{\frac{2\log(T)}{N_t(a)}}$. The regret is of order $\mathbb{E}(R(t)) \leq O(\sqrt{kt \log T})$.

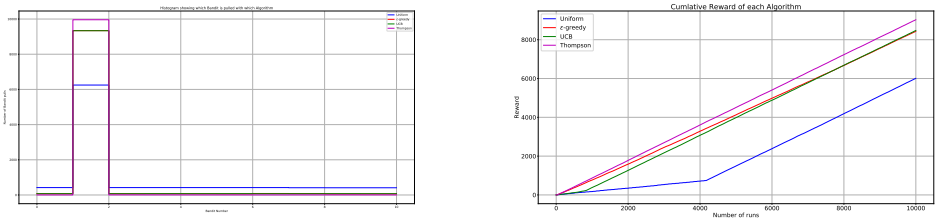
2.4 Thompson Sampling

Thompson sampling implicitly applies a prior to each bandit. The beta-distribution $\mathbb{P}(r|\alpha, \beta)$is sampled and modified as bandit rewards are acquired. Initially if no information about the bandits is known apriori, the beta distribution tends to a uniform distribution since $\alpha = 1$, and $\beta = 1$. Every time a bandit gives a reward, α is updated by 1, while every time a failure occurs β is updated by one. Hence exploration happens based upon the rewards given by the bandits.

3 Experiments

3.1 Stationary Bandits

Each bandit gives a reward 1 or 0 given from a Bernoulli distribution with a fixed time-invariant mean.



(a) Histogram for Stationary Case

(b) Cumlative Reward for Stationary Case

Figure 1: Stationary Bandits

Figure 1b reveals the exploration of the uniform algorithm. Once all the bandits are sufficiently pulled and an estimate of the mean reward for the bandits is calculated, the bandit with the highest estimated mean is pulled henceforth. This is visible in figure 2. The constant slope occurs during the exploration phase where all the bandits are pulled. In this stationary case, the bandit with the highest estimated mean is the optimal bandit and the regret does not increase. This is shown by the flat line of the uniform regret line in figure 2.

UCB-1 algorithm explores all the arms roughly equally before choosing the one with the highest mean reward.

The best bound for Thompson sampling regret is that is asymptotically converges [1]. This is seen in 2 where the regret is roughly flat as the optimal bandit has been found. However, the regret does slowly increase as the algorithm continues to explore rather than just exploit the bandit with the best mean reward.

In this experiment, the ϵ-greedy algorithm's regret is consistent with the theoretical lower bound, see figure 2. However, the other regrets show a linear regret and then a flat area once the optimal bandit is selected.

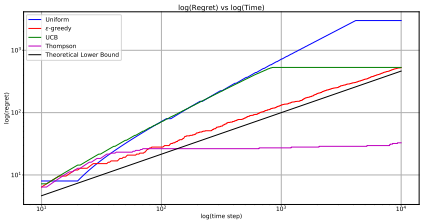
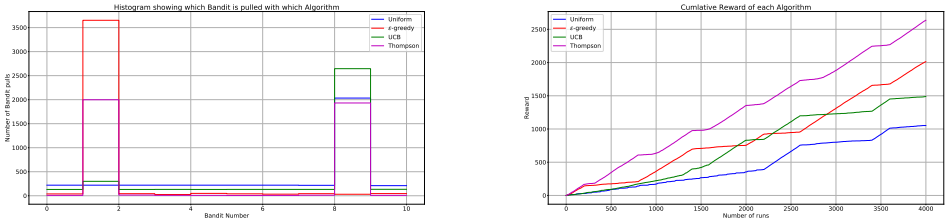


Figure 2: Stationary Regret over time

3.2 Adversarial Bandits

This experiment with adversarial bandits the two sets of mean values are $[0.1, 0.9, 0.1, ..., 0.1]$ and $[0.1, 0.1, ..., 0.9, 0.1]$. Hence there is a bandit which is clearly optimal at each time step.



(a) Histogram for Adversarial Case

(b) Cumlative Reward for Adversarial Case

Figure 3: Adversarial Bandits

Oscillations in the cumulative reward 5b are due to changes in the bandit with the optimal mean. The histogram in figure 3a shows how the stationary algorithms favour one of the bandits especially as time evolves and hence exploration does not continue. Once the optimal bandit no longer returns the best reward, Thompson sampling begins to reduce the likelihood of sampling from it and begins to explore. This is evident in figure 4 of the cumulative regret, where the regret jumps up then becomes stationary before repeating. The jump is due to the change in optimal bandit, and can be seen as an exploratory phase for this algorithm. The adversarial regrets are flat once the optimal bandit is pulled, and then there is a jump once the bandits change their means.

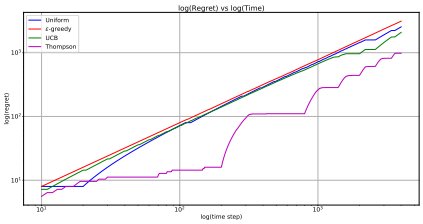


Figure 4: Adversial Regret over time

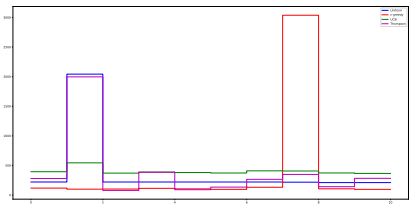
3.3 Lower-bound Means

All the bandits are stationary with means set to 0.5 except for one bandit which has mean $0.5 + \epsilon$. ϵ is set to $1/\sqrt{\frac{k}{T}}$ [2], which is the upper bound for the algorithms to be able to find the optimal bandit. However, experimental results do not reflect this, and the limit for the algorithms to find the optimal bandit. The limiting value of ϵ is set to $0.5+0.07$, at which point the uniform algorithm selects the correct bandit at the end of its exploration phase.

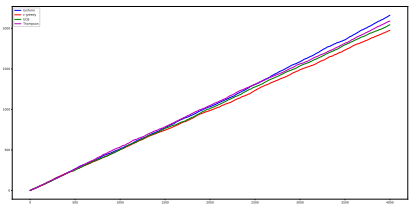
The regret for the uniform algorithm, flattens out once the optimal bandit is estimated. Neither of the other algorithms manage to estimate this correctly with this value of ϵ.

4 Conclusion

Thompson sampling provides the highest reward consistently. However, there is no bound on the regret of this algorithm which may be a problem if the run time of the algorithm is unknown [1]: the regret asymptotically tends to a value. Based upon the success or failures of each arm, it is possible for Thompson sampling to explore even if the one of the bandits' estimated means indicates that this bandit is optimal. Hence, why the regret curves for this bandit is never horizontal. UCB-1 and uniform algorithms stop exploring and only exploit one bandit after a time. This is sufficient in the stationary case but is detrimental if the total number of iterations is unkown of if the bandits are adversarial.



(a) Histogram for Lower Bound



(b) Cumlative Reward for Lower Bound

Figure 5: Lower Bound

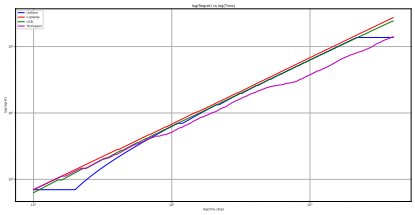


Figure 6: Lower Bound Regret

References

[1] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011.

[2] Aleksandrs Slivkins. *Introduction to Multi-Armed Bandits*. 2018.