An introduction to stochastic multi-armed bandits

Sheel Shah, 19D070052, IIT Bombay

What are bandits?

Stochastic MABs

1. K arms, each having some (unknown) reward distribution X_i.
2. The player wants to maximize his net reward and thus he must, in the long run, pull the arm with the highest mean.
3. The rewards are random, and hence the player can never be completely sure of what the optimal arm is.
4. Exploration/exploitation tradeoff.

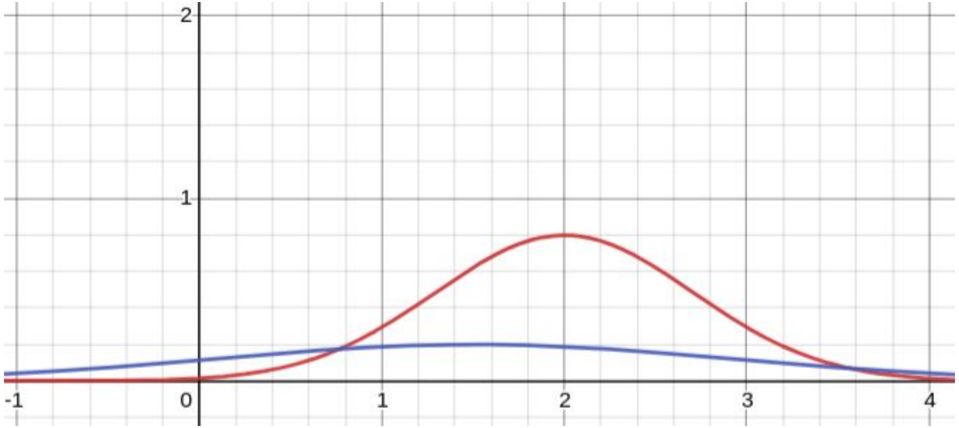
Algorithms

The Upper Confidence bound

1. The driving principle of this algorithm is optimism under uncertainty.

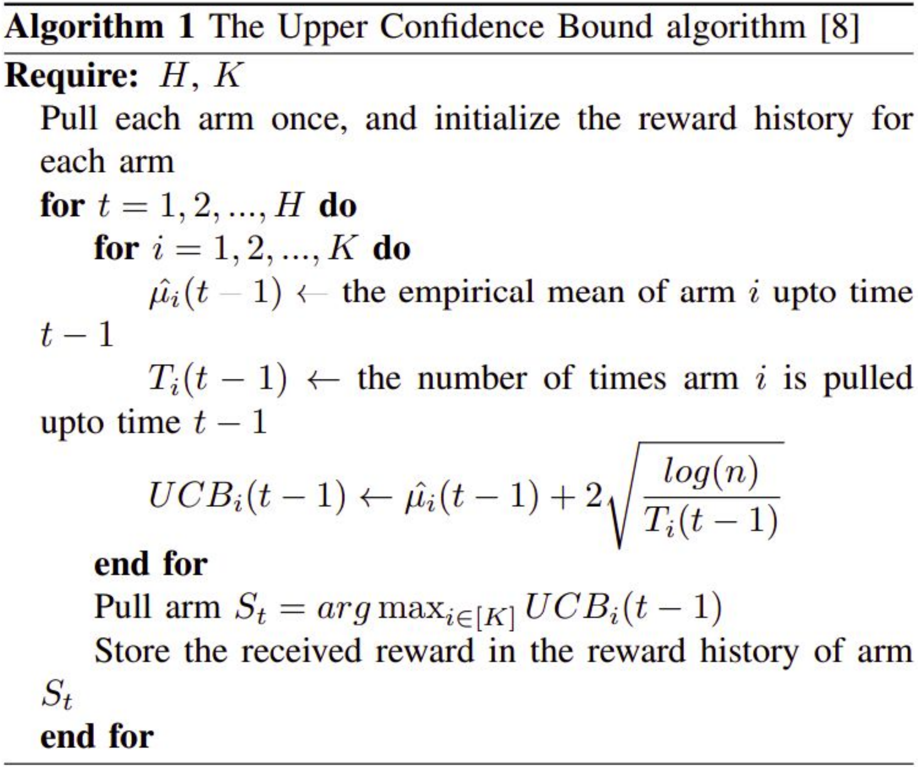
2. Given some history of an arm, one can use the Chernoff-Hoeffding inequality [10] to find an interval within which the mean of an arm lies with high probability.

3. UCB assumes the mean of each arm is the upper extreme of its confidence interval.



The blue and red curves represents the estimate of the mean of an arm after 10 pulls and 40 pulls, respectively.

The Upper Confidence bound algorithm



Algorithm 1 The Upper Confidence Bound algorithm [8]

```
Require: $H$, $K$
  Pull each arm once, and initialize the reward history for each arm
  for $t = 1, 2, ..., H$ do
    for $i = 1, 2, ..., K$ do
      $\hat{\mu}_i(t-1) \leftarrow$ the empirical mean of arm $i$ upto time $t-1$
      $T_i(t-1) \leftarrow$ the number of times arm $i$ is pulled upto time $t-1$
      $UCB_i(t-1) \leftarrow \hat{\mu}_i(t-1) + 2\sqrt{\frac{log(n)}{T_i(t-1)}}$
    end for
    Pull arm $S_t = arg\max_{i \in [K]} UCB_i(t-1)$
    Store the received reward in the reward history of arm $S_t$
  end for
```

Thompson Sampling

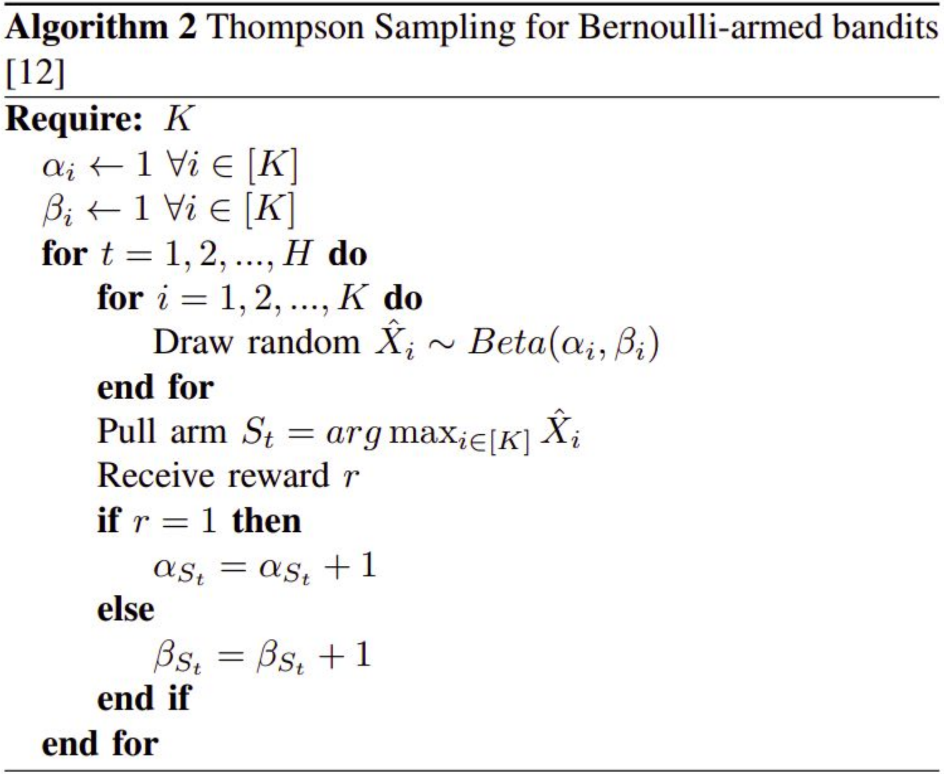
1. Thompson sampling (TS) is a Bayesian algorithm, meaning it maintains a ”belief” of the means of each arm, and this belief is updated whenever the player encounters new ”evidence”.

2. In the specific case of rewards being Bernoulli random variables, a Beta belief is used by Thompson Sampling.

3. Each evidence is the reward of an arm-pull (0 or 1).

$$posterior = \begin{cases} Beta(\alpha_i, \beta_i + 1), \text{if reward of 0 is received} \\ Beta(\alpha_i + 1, \beta_i), \text{if reward of 1 is received} \end{cases}$$

The TS algorithm



Algorithm 2 Thompson Sampling for Bernoulli-armed bandits [12]

```
Require: $K$
  $\alpha_i \leftarrow 1 \ \forall i \in [K]$
  $\beta_i \leftarrow 1 \ \forall i \in [K]$
  for $t = 1, 2, ..., H$ do
      for $i = 1, 2, ..., K$ do
          Draw random $\hat{X}_i \sim Beta(\alpha_i, \beta_i)$
      end for
      Pull arm $S_t = arg\max_{i \in [K]} \hat{X}_i$
      Receive reward $r$
      if $r = 1$ then
          $\alpha_{S_t} = \alpha_{S_t} + 1$
      else
          $\beta_{S_t} = \beta_{S_t} + 1$
      end if
  end for
```

Performance of these algorithms

Definition of regret	$R = \sum_{t=1}^{H}(E[X^*] - E[X_{S_t}])$ $where\ X^* = arg \max_{i\in[K]} E[X_i]$
Lower bound by Lai and Robbins [6]	$\liminf_{H\to\infty} \frac{R}{ln\ H} \geq \sum_{i:E[X_i]<E[X^*]} \frac{E[X^*] - E[X_i]}{D(X_i, X^*)}$
Upper bound on regret of UCB	$\lim_{H\to\infty} \frac{R}{ln\ H} \leq \sum_{i:E[X_i]<E[X^*]} \frac{16}{E[X^*] - E[X_i]}$
Upper bound on regret of TS	$\lim_{H\to\infty} \frac{R}{ln\ H} \leq \sum_{i:E[X_i]<E[X^*]} \frac{2}{E[X^*] - E[X_i]}$

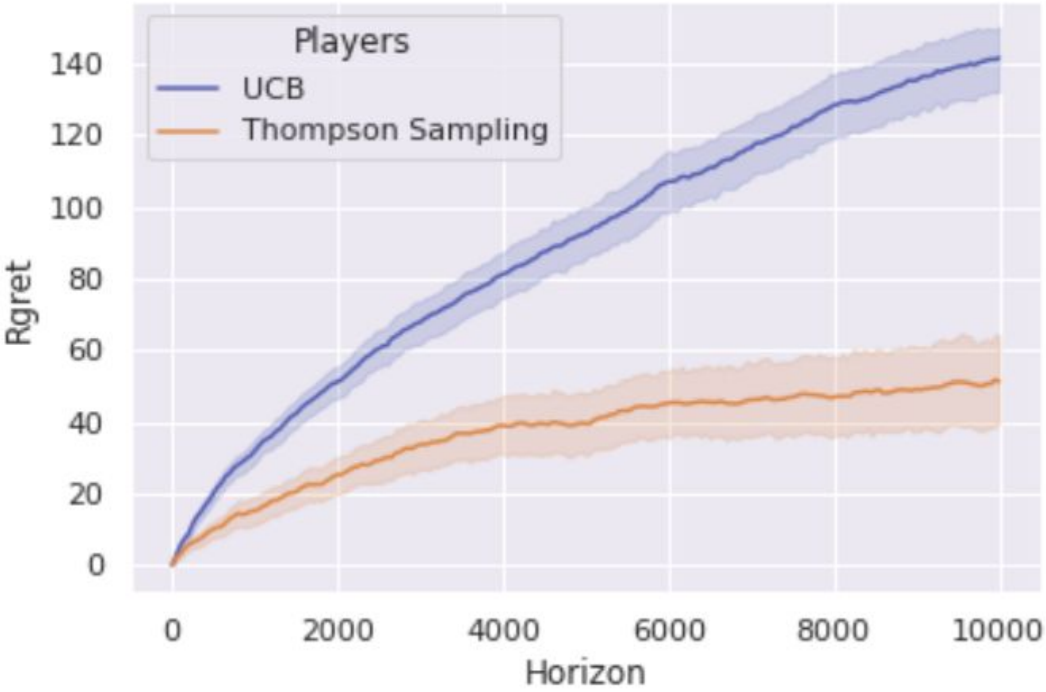


Figure 1. The regret versus horizon plots for UCB and TS. The algorithms were tested against a Bernoulli-armed bandit with three arms of means 0.5, 0.67, 0.7. A total of 100 simulations were run per algorithm to abate the effect of randomization. The logarithmic nature of the regret is clearly visible, and UCB's regret being a constant factor worse is also apparent.

Conclusion

This presentation has discussed the problem of stochastic MABs, and two popular algorithms used to solve this problem. There is a plethora of algorithms for this setting, variations of this setting, and algorithms for these variations. For brevity, these have not been discussed here, but a brief overview has been provided in the accompanying report of this presentation.

References

[1] Peter Auer et al. “The nonstochastic multiarmed bandit problem”. In: SIAM Journal on Computing 32 (Jan. 2003), pp. 48–77.

[2] Theodore Colton. “A Model for Selecting One of Two Medical Treatments”. In: Journal of the American Statistical Association 58.302 (1963), pp. 388–400. ISSN: 01621459. URL: http://www.jstor.org/stable/2283274.

[3] Pierre-Arnaud Coquelin and Remi Munos. “Bandit ´ Algorithms for Tree Search”. In: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence. UAI’07. Vancouver, BC, Canada: AUAI Press, 2007, pp. 67–74. ISBN: 0974903930.

[4] Michael O. Duff. “Optimal learning: Computational procedures for Bayes -adaptive Markov decision processes”. English. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-11- 09. PhD thesis. 2002, p. 247. ISBN: 978-0-493-52573- 0. URL: https : / / www . proquest . com / dissertations - theses / optimal - learning - computational - procedures - bayes/docview/251665294/se-2?accountid=27542.

[5] Tim van Erven and Peter Harremos. “Renyi Divergence ´ and Kullback-Leibler Divergence”. In: IEEE Transactions on Information Theory 60.7 (2014), pp. 3797– 3820. DOI: 10.1109/TIT.2014.2320500.

[6] T.L Lai and Herbert Robbins. “Asymptotically Efficient Adaptive Allocation Rules”. In: Adv. Appl. Math. 6.1 (Mar. 1985), pp. 4–22. ISSN: 0196-8858. DOI: 10.1016/ 0196-8858(85)90002-8. URL: https://doi.org/10.1016/ 0196-8858(85)90002-8.

[7] Tor Lattimore and Csaba Szepesvari. Bandit Algorithms. [Cambridge University Press], 2020, pp. 10–13.

[8] Tor Lattimore and Csaba Szepesvari. Bandit Algorithms. [Cambridge University Press], 2020, pp. 101–110.

[9] Lihong Li et al. “A Contextual-Bandit Approach to Personalized News Article Recommendation”. In: Proceedings of the 19th International Conference on World Wide Web. WWW ’10. Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, pp. 661– 670. ISBN: 9781605587998. DOI: 10 . 1145 / 1772690 . 1772758. URL: https : / / doi . org / 10 . 1145 / 1772690 . 1772758.

[10] Jeff M. Phillips. Chernoff-Hoeffding Inequality and Applications.

[11] Herbert Robbins. “Some aspects of the sequential design of experiments”. In: Bulletin of the American Mathematical Society 58.5 (1952), pp. 527–535. DOI: bams/1183517370. URL: https://doi.org/.

[12] Daniel J. Russo et al. “A Tutorial on Thompson Sampling”. In: Foundations and Trends® in Machine Learning 11.1 (2018), pp. 1–96. ISSN: 1935-8237. DOI: 10. 1561/ 2200000070. URL: http:// dx. doi. org/ 10. 1561/ 2200000070.