COMP3702/7702 Artificial Intelligence
Semester 2, 2020
Tutorial 10 - Sample Solutions

Exercise 10.1

a) In one instance of the MAB, the actions taken and rewards received for the first six trials are given in the table below:

Trial	Action	Reward
1	A_1	2.66
2	A_2	1.25
3	A_1	3.21
4	A_2	2.34
5	A_1	1.87
6	A_1	1.69

Using ϵ-greedy, which action is most likely to be chosen next?

- Assuming that the agent wishes to maximise its reward, under the ϵ-greedy strategy, it will choose the are with the highest the mean reward with probability ϵ.
- Further, assume $\epsilon > 0.5$.
- Estimated rewards:
  - $\hat{v}_1 = 2.3575$
  - $\hat{v}_2 = 1.795$
- So A_1 is selected with probability ϵ.

b) Given the same sample information, now consider UCB1 with upper bounds given by:

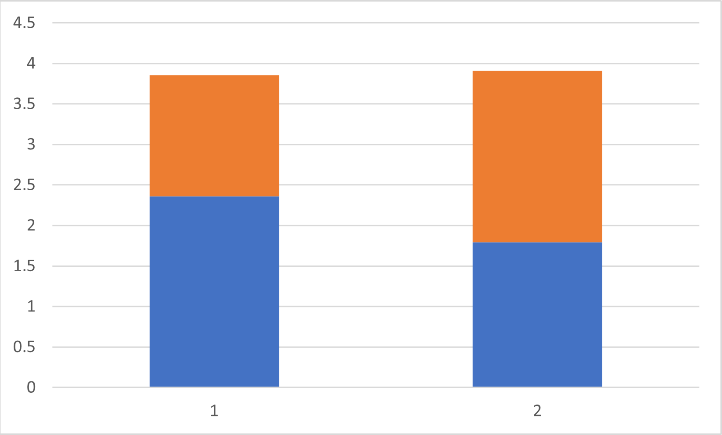
$$UCB1_a = \hat{v}_a + \sqrt{\frac{C \ \ln(N)}{n_a}}$$

Set the tunable parameter C to 5. Using this UCB algorithm, which action is chosen next?

Using the formula above:

- the interval for A_1 is 1.4965 and the upper confidence bound for A_1 is 3.8541
- the interval for A_2 is 2.116 and the upper confidence bound for A_2 is 3.9115
- So A_2 is chosen next, as it has the greater upper confidence bound.

A_2 has been sampled fewer times than A_1. This implies that there is typically greater uncertainty in its mean estimate, which is reflected in the larger interval size; and this pushes its upper confidence bound slightly above that of A_1.



To see the link to UCT and MCTS, consider this restatement of the UCB1 upper bound expression:

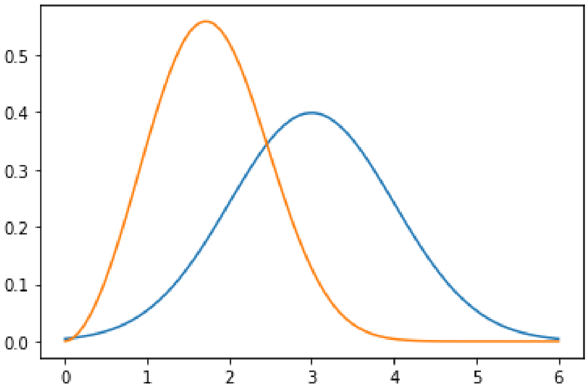
$$UCB1_a = \hat{Q}(a) + \sqrt{\frac{C\ \ln(N)}{n_a}}$$

Because a MAB has only one state, we can drop the state argument. However, if we explicitly include it, the connection to MCTS becomes apparent:

$$\begin{aligned} UCT_{a,s} &= \hat{Q}(a,s) + \sqrt{\frac{C\ \ln(N_s)}{n_{a,s}}} \\ &= \frac{R(a,s)}{n_{a,s}} + \sqrt{\frac{C\ \ln(N_s)}{n_{a,s}}} \end{aligned}$$

where now $\hat{Q}$ is a table of (a, s) values indexed by the state node, s, from which the sample and roll-out simulation is taken, and the action chosen, a.

c) Plot the distributions of rewards from each arm. If the agent wishes to maximise its cumulative reward over time and knew these distributions, which would be the optimal arm to pull?



Explanation:

- The mean of the normal distribution is given as $\mu = 3$.

- The mean of the Weibull distribution is a complicated function of a and b, but this case it is equal to $\sqrt{\pi/2} \approx 1.253$.

d) Set up a MAB instance with two arms described above, and consider the ϵ-greedy exploration strategy with random sampling parameter set to $\epsilon = 0.1$, and the UCB bound as described in b) above. For each strategy, plot their cumulative rewards over 1000 arm trials in an MAB instance. **Questions**: Which performs better initially? Which performs better in the long run?

It is difficult to say which does better initially, as performance depends on the random realisation of rewards.

However, in the long run, we expect the UCB1 algorithm to outperform ϵ-greedy. This is because ϵ-greedy continues to sample the non-optimal arm(s) with probability ϵ even long after the mean reward estimates for all have converged to very close to their true values. In contrast, UCB1 reduces the chance of exploring away from the highest-value arm by adjusting the confidence interval based on the number of samples taken.