Reinforcement Learning	May 22, 2019
Recitation 10	
Lecturer:Yishay Mansour	*TA: Lee Cohen*

Multi Arm Bandits

The word "bandit" refers to "one-armed bandits," another name for slot machines. The problem is that initially we don't know the reward distribution of any of the bandits, and we can only try them one at a time, so based on the outcomes so far we must choose to exploit the current bandit or explore others. For this problem, we'll think about rewards instead of losses and assume the distributions doesn't change with time (i.i.d.). At time t, our reward from bandit i is a random variable r_i^t.

Definition 1. The regret can be written as:

$$R_T = \left(\max_i \sum_{t=1}^{T} r_i^t \right) - \sum_{t=1}^{T} r_{I_t}^t, \quad I_t \in 1, 2, ..., N \tag{1}$$

The problem is that r_i^t is drawn from a distribution, which makes the regret problematic to compute. Instead, we can use the expected value rather than the sum; only this is too hard to compute exactly, so we will use the pseudo regret instead.

Definition 2. The pseudo regret is defined as:

$$\bar{E}[R_T] = \max_i E[\sum_{t=1}^{T} r_i^t] - E[\sum_{t=1}^{T} r_{I_t}^t], \quad I_t \in 1, 2, ..., N \tag{2}$$

Exercise 1.
Prove that -

$$\bar{E}[R_T] = \sum_{i=1}^{N} E[N_i^t] \Delta_i \tag{3}$$

Where,

N_i^T - is the number of times arm i is pulled in T time steps.
$\mu^* = \max_j \mu_j$.
$\Delta_i = \mu^* - \mu_i$.

Solution 1.

$$
\begin{aligned}
&\bar{E}[R_T] = \max_i E[\textstyle\sum_{t=1}^{T} r_i^t] - E[\sum_{t=1}^{T} r_{I_t}^t] = \\
&= \max_i \textstyle\sum_{t=1}^{T} E[r_i^t] - E[\sum_{t=1}^{T} r_{I_t}^t] = \\
&= \max_i (T\mu_i) - E[\textstyle\sum_{t=1}^{T} r_{I_t}^t] = \\
&= T\mu^* - E[\textstyle\sum_{t=1}^{T} \mu_{I_t}] = \\
&= T\mu^* - \textstyle\sum_{t=1}^{T} E[N_i^T]\mu_i = \\
&= \textstyle\sum_{t=1}^{T} E[N_i^T]\mu^* - \sum_{t=1}^{T} E[N_i^T]\mu_i = \\
&= \textstyle\sum_{t=1}^{T} E[N_i^T](\mu^* - \mu_i) = \\
&= \textstyle\sum_{t=1}^{T} E[N_i^T]\Delta_i
\end{aligned}
$$

Theorem 1. Hoeffding's Inequality
Given independent random variables $X_1, ..., X_m$ where $a_i \leqslant X_i \leqslant b_i$ almost surely (with probability 1) we have:

$$P\left(\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} EX_i \geq \epsilon\right) \leqslant exp\left(\frac{-2\epsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right) \tag{4}$$

UCB (Upper Confidence Bounds)

Simple and efficient allocation strategies based on upper confidence bounds for a bandit problem with any reward distribution with known bounded support. The algorithms demonstrate logarithmic regret performance uniformly over time, not just asymptotically.

To implement the UCB algorithm, we need both $\hat{\mu}_{i,T_i}$, the empirical reward estimate of arm i after it has been pulled T_i many times, and an upper confidence bound on that estimate. To calculate our empirical reward estimate, we simply average the observed rewards over all rounds where we pull arm i. In the below expression, we assume $T_i = |\{t : I_t = i\}|$. That is, we have progressed through an appropriate number of rounds where arm i has been pulled T_i many times: $\hat{\mu}_{i,T_i} = \frac{\Sigma_{t:I_t=i} X_t}{T_i}$

In addition to our empirical reward estimates, we need an upper confidence bound to describe the largest plausible mean of each arm. Using Hoeffding's Inequality and Chernoff Bounds, we can construct such a confidence interval. With probability at least $1 - t^{-\alpha}$, the empirical mean $\hat{\mu}_{i,T_{i,t}}$ will differ from the true mean by at most $\epsilon = \sqrt{\frac{\alpha log t}{2T_i}}$. The UCB algorithm chooses the largest such upper bound:

$$UCB_{i,t} = I_t = \underset{i\in[n]}{argmax}\left(\hat{\mu}_{i,T_{i,t}} + \sqrt{\frac{\alpha log(t)}{2T_i}}\right) \tag{5}$$

We see that our confidence bound, $\sqrt{\frac{\alpha log t}{2T_i}}$, grows slowly as we play for more rounds (as t increases), ensuring that we never stop playing any given arm. The confidence bound for arm i shrinks quickly as we pull the arm (as T_i increases). The pseudocode may be found in Algorithm 1.

Theorem 2. Regret bound for the UCB algorithm for $T \geqslant 1$

$$R(T) \leqslant \sum_{i:\Delta_i>0} 4\alpha\Delta_i^{-1} log(T) + \frac{2\alpha}{\alpha - 1}\Delta_i \tag{6}$$

Proof. Suppose, without loss of generality, that arm 1 is optimal. Then, arm $i \neq 1$ will only be played in two cases: either arms 1 and i have been sampled insufficiently to distinguish between their means, or the upper confidence bound given by Hoeffding's inequality fails

Algorithm 1 UCB

procedure UCB($\{1, 2, \ldots, n\}, T$) ▷ Arms 1 through n, max steps T
  for $1 \leq t \leq n$ **do**
    $I_t \leftarrow t$ ▷ Play each arm once
  end for
  for $n + 1 \leq t \leq T$ **do**
    $I_t = \arg\max_{i \in \{1, \cdots, n\}} \text{UCB}_{i,t-1}$
    Observe reward $X_{T_t,t}$
  end for
end procedure

for either arm 1, or arm i. We begin by bounding the chance that we pull a suboptimal arm due to insufficient sampling. Suppose that we have the following two events A_t, B_t,
A_t) $\hat{\mu}_{i,T_i} \leqslant \mu_i + \sqrt{\frac{\alpha log(t)}{2T_i}}$
B_t) $\hat{\mu}_{1,T_1} \geqslant \mu_1 - \sqrt{\frac{\alpha log(t)}{2T_1}}$
We wish to bound the probabilities of the complements of events A_t and B_t occurring. We will apply Hoeffding's inequality (Theorem 1). A_t fails when -

$$\hat{\mu}_{i,T_i} - \mu_i > \sqrt{\frac{\alpha log(t)}{2T_i}} \tag{7}$$

By Theorem 1 we have:

$$\mathbb{P}\left(\hat{\mu}_{i,T_i} - \mu_i > \sqrt{\frac{\alpha log(t)}{2T_i}}\right) \leqslant exp\left(\frac{-2\epsilon^2 t^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right) = exp\left(\frac{-2\epsilon^2 t^2}{\sum_{i=1}^{m}(1-0)^2}\right) = exp(-2\epsilon^2 t) \tag{8}$$

We plug in our bounding value $\epsilon = \sqrt{\frac{\alpha log(t)}{2T_i}}$

$$\mathbb{P}\left(\hat{\mu}_{i,T_i} - \mu_i > \sqrt{\frac{\alpha log(t)}{2T_i}}\right) \leqslant exp\left(\frac{-2t\alpha \log(t)}{2T_i}\right) = \tag{9}$$

$$= exp\left(\frac{-t\alpha \log(t)}{T_i}\right) \leqslant exp\left(\frac{-t\alpha \log(t)}{t}\right) = e^{-\alpha \log(t)} = t^{-\alpha} \tag{10}$$

The statement and justification is identical for the complement of event B_t and we return to the task of bounding the number of suboptimal arm pulls. A suboptimal arm i is only played if its upper confidence bound exceeds that of arm 1, meaning that,

$$\hat{\mu}_{i,T_i} + \sqrt{\frac{\alpha log(t)}{2T_i}} > \hat{\mu}_{1,T_1} + \sqrt{\frac{\alpha log(t)}{2T_1}} \tag{11}$$

Suppose that both A_t and B_t both hold. In this case, suboptimal arm i is pulled due to insufficient sampling up to this point. Since A_t has been assumed to be true, we generate the following bound:

$$\mu_i + 2\sqrt{\frac{\alpha log(t)}{T_i}} \geq \hat{\mu}_{i,T_i} + \sqrt{\frac{\alpha log(t)}{2T_i}} \tag{12}$$

Next, we use our assumption of B_t being true to upper-bound the right hand side of inequality (11):

$$\hat{\mu}_{1,T_1} + \sqrt{\frac{\alpha log(t)}{2T_1}} \geq \mu_1 \tag{13}$$

Chaining equations (11) and (12) we have:

$$\mu_i + 2\sqrt{\frac{\alpha log(t)}{T_i}} \geq \mu_1 \tag{14}$$

Rearranging we have:

$$\sqrt{\frac{\alpha log(t)}{T_i}} \geq \frac{\mu_1 - \mu_i}{2} \tag{15}$$

Now, recall our definition of the optimality gap of an arm, $\Delta_i = max_j \mu_j - \mu_i$. Since we know arm 1 is optimal, this becomes $\Delta_i = \mu_1 \mu_i$. Our inequality becomes

$$\sqrt{\frac{\alpha log(t)}{T_i}} \geq \frac{\Delta_i}{2} \tag{16}$$

Solving for the number of times T_i that an arm has been played, we arrive at

$$T_i \leq 4\Delta_i^{-2} \alpha \log(t) \leq 4\Delta_i^{-2} \alpha \log(T) \tag{17}$$

Thus when A_t and B_t hold, we only play suboptimal arm i at most $4\Delta_i^{-2} \alpha \log(T)$ times. Recall that I_t can only be equal to i if either it has been sampled insufficiently (fewer than $4\Delta_i^{-2} \alpha \log(T)$ times) or either event A_t or B_t fails. For any arm i, the expected number of times it is played up to round T under UCB is:

$$E[T_i] = \sum_{t=1}^{T} E[1(I_t = i)] \leq 4\Delta_i^{-2}\alpha \log(T) + \sum_{t=1}^{T} E[1(A_t^c \cup B_t^c)] \tag{18}$$

$$\leq 4\Delta_i^{-2}\alpha \log(T) + \sum_{t=1}^{T} E[1(A_t^c) + 1(B_t^c)] \tag{19}$$

$$\leq 4\Delta_i^{-2}\alpha \log(T) + \sum_{t=1}^{T} (t^{-\alpha} + t^{-\alpha}) \tag{20}$$

$$= 4\Delta_i^{-2}\alpha \log(T) + 2\sum_{t=1}^{T} t^{-\alpha} \tag{21}$$

In order to bound the second term on the right hand side, we note:

$$\sum_{t=1}^{T} t^{-\alpha} \leq 1 + \int_1^{\infty} x^{-\alpha} dx = 1 + \frac{-1}{1-\alpha} = \frac{-\alpha}{1-\alpha} \tag{22}$$

Therefore we have:

$$E[T_i] \leqslant 4\Delta_i^{-2}\alpha \log(T) + \frac{2\alpha}{\alpha - 1} \tag{23}$$

The desired result follows from summing over all suboptimal arms:

$$R(T) = \sum_{i \neq 1} \Delta_i E[T_i] = \sum_{i \neq 1} 4\Delta_i^{-1}\alpha \log(T) + \frac{2\alpha}{\alpha - 1}\Delta_i \tag{24}$$

□

References:

1. Statistical Techniques in Robotics (16-831, F11) http://www.cs.cmu.edu/~16831-f12/notes/F14/16831_lecture23_ndo_hanbyulj.pdf
2. CSE599i: Online and Adaptive Machine Learning https://courses.cs.washington.edu/courses/cse599i/18wi/resources/lecture3/lecture3.pdf