Machine Learning Theory (CS 6783)

Lecture 19: UCB Algorithm for Stochastic Multi-armed Bandit

1 Upper Confidence Bound (UCB) Algorithm

In the stochastic multi-armed bandit setting we consider the problem where losses $\ell_1, \ldots, \ell_n$ are drawn iid from some fixed distribution $\mathcal{D}$ over $[-1,1]^K$. Let us define $L_i = \mathbb{E}_{\ell \sim \mathcal{D}}[\ell[i]]$ as the expected loss of the i'th arm. Let $I_t \in [K]$ be the arm picked by the learning algorithm on round t. For arm i define

$$\hat{L}_{i,t} = \frac{1}{n_{i,t}} \sum_{s \in [t]: I_s = i} \ell_t[i]$$

where $n_{i,t} = |\{s \in [t] : I_s = i\}|$. That is the number of times arm i has been picked up to time t. The algorithm we consider is the following.

For $i = 1$ to K % First K rounds play each arm once

  Pick $I_i = i$

End For

Set $n_{i,K} = 1$ for all i

For $t = K+1$ to n

  Pick $I_t = \operatorname{argmin}_{i \in [K]} \left(LCB_{i,t-1} := \hat{L}_{i,t-1} - \sqrt{\frac{\log(t-1)}{n_{i,t-1}}} \right)$

  Receive loss $\ell_t[I_t]$

  Update $n_{I_t,t} = n_{I_t,t} + 1$

  Update $\hat{L}_{i,t}$ for all i

End For

The high level intuition is super simple. First, note that if we consider the expected regret, we have the expression:

$$\mathbb{E}\left[\frac{1}{n} \sum_{t=1}^{n} \ell_t[I_t] - \min_{i \in [K]} \frac{1}{n} \sum_{t=1}^{n} \ell_t[i] \right] = \frac{1}{n} \sum_{j=1}^{K} \mathbb{E}[n_{j,n}] \Delta_j \tag{1}$$

where we define $\Delta_j = (L_j - \min_{i \in [K]} L_i)$ the difference in the expected losses of arm j and optimal arm. This is clear because for each time we play a sub-optimal arm, we pay in expectation the sub-optimality gap of the arm. Hence in expectation we get the above expression. This shows that all we need to do to complete the proof is to bound expected number of times each arm is pulled.

Lemma 1. *At any time t and any arm j,*

$$P\left(I_{t+1} = j \mid |n_{i,t} \geq \frac{4\log t}{\Delta_j^2}\right) \leq 4t^{-2}$$

Proof. Next note that $\mathbb{E}\left[\hat{L}_{i,t}\right] = L_i$ since its an unbiased estimate of the loss of arm i. However by Hoeffding's inequality, we have that

$$P\left(\left|\hat{L}_{i,t} - L_i\right| > \epsilon\right) \leq 2\exp\left(-2\epsilon^2 t\right)$$

Plugging in $\epsilon = \sqrt{\frac{\log t}{n_{i,t}}}$ we get,

$$P\left(\left|\hat{L}_{i,t} - L_i\right| > \epsilon\right) \leq 2\exp\left(-\frac{2t\log(t)}{n_{t,i}}\right) \leq 2\exp\left(-2t\log(t)\right) \leq 2t^{-2}$$

Now let i^* be an optimal arm. Note that for any arm j, by the bound above, with probability at least $1 - 2/t^2$,

$$LCB_{j,t} = \hat{L}_{t,j} - \sqrt{\frac{\log(t)}{n_{j,t}}} \geq L_j - 2\sqrt{\frac{\log(t)}{n_{j,t}}}$$

Hence if $n_{j,t} > \frac{4\log t}{\Delta_j^2}$ we will have that

$$LCB_{j,t} < L_j - \Delta_j = L_{i^*}$$

But by Hoeffding bound again with probability at least $1 - 2/t^2$, $L_{i^*} \geq LCB_{i^*,t}$ and so by union bound, we have that when for any j, when $n_{j,t} > \frac{4\log t}{\Delta_j^2}$, then with probability at least $1 - 4/t^2$,

$$LCB_{j,t} > LCB_{i^*,t}$$

Thus we can conclude that when $n_{j,t} > \frac{4\log t}{\Delta_j^2}$ for all sub-optimal j's with high probability the UCB algorithm will pick the optimal arm instead. More specifically,

$$P\left(I_{t+1} = j \mid |n_{i,t} \geq \frac{4\log t}{\Delta_j^2}\right) \leq 4t^{-2}$$

□

Lemma 2. *For any arm j, we have that:*

$$\mathbb{E}\left[n_{i,n}\right] \leq \frac{4\log(n)}{\Delta_i^2} + 8$$

Proof. Note that:

$$\begin{aligned}
\mathbb{E}\left[n_{i,n}\right] &= 1 + \mathbb{E}\left[\sum_{t=K+1}^{n} \mathbf{1}\{I_t = i\}\right] \\
&= 1 + \mathbb{E}\left[\sum_{t=K+1}^{n} \mathbf{1}\{I_t = i, n_{i,t} < \frac{4\log(t)}{\Delta_i^2}\}\right] + \mathbb{E}\left[\sum_{t=K+1}^{n} \mathbf{1}\{I_t = i, n_{i,t} \geq \frac{4\log(t)}{\Delta_i^2}\}\right] \\
&= 1 + \mathbb{E}\left[\sum_{t=K+1}^{n} \mathbf{1}\{I_t = i, n_{i,t} < \frac{4\log(t)}{\Delta_i^2}\}\right] + \sum_{t=K+1}^{n} P\left(I_t = i, n_{i,t} \geq \frac{4\log(t)}{\Delta_i^2}\right) \\
&\leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{n} \mathbf{1}\{I_t = i, n_{i,t} < \frac{4\log(t)}{\Delta_i^2}\}\right] + \sum_{t=K+1}^{n} P\left(I_t = i \mid n_{i,t} \geq \frac{4\log(t)}{\Delta_i^2}\right) \\
&\leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{n} \mathbf{1}\{I_t = i, n_{i,t} < \frac{4\log(t)}{\Delta_i^2}\}\right] + \sum_{t=K+1}^{n} \frac{4}{t^2} \\
&\leq 8 + \mathbb{E}\left[\sum_{t=K+1}^{n} \mathbf{1}\{I_t = i, n_{i,t} < \frac{4\log(n)}{\Delta_i^2}\}\right]
\end{aligned}$$

Now say $\mathbf{1}\{I_t = i, n_{i,t} < \frac{4\log(n)}{\Delta_i^2}\}$ was switched on more than $\frac{4\log(n)}{\Delta_i^2}$ number of times, then automatically, we would have a contradiction since $n_{i,t}$ becomes larger than the condition in the indicator. Hence we can conclude that, $\sum_{t=K+1}^{n} \mathbf{1}\{I_t = i, n_{i,t} < \frac{4\log(n)}{\Delta_i^2}\} \leq \frac{4\log(n)}{\Delta_i^2}$. Hence. we get the overall bound of

$$\mathbb{E}\left[n_{i,n}\right] \leq 8 + \frac{4\log(n)}{\Delta_i^2}$$

□

Using the above lemma's result with Eq 1 we conclude the following main theorem.

Theorem 3. *For the LCB Algorithm we have the following bound on expected regret:*

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \ell_t[I_t] - \min_{i\in[K]} \frac{1}{n}\sum_{t=1}^{n} \ell_t[i]\right] \leq \frac{1}{n} \sum_{j\in[K]:\Delta_j>0} \left(\frac{4\log(n)}{\Delta_j} + 8\Delta_j\right)$$

If we further notice that if arms that are better than $\sqrt{K/n}$ sub-optimal are pulled even n times, we still have a good regret bound and then use the above result, we can also get the following corollary that shows that the same algorithm enjoys the right worst case bound as well.

Corollary 4. *For any $n > K$, the expected regret achieved by UCB algorithm is bounded as*

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} \ell_t[I_t] - \min_{i\in[K]} \frac{1}{n}\sum_{t=1}^{n} \ell_t[i]\right] \leq 5\sqrt{\frac{K\log n}{n}} + \frac{8K}{n}$$