Mathematical Foundations of Machine Learning (CS 4783/5783)

Lecture 15: Stochastic Multi-armed Bandits

1 Lower Confidence Bound (LCB) 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 $\mathbf{D}$ over $[-1,1]^K$. Let us define $L_i = \mathbb{E}_{\ell \sim \mathbf{D}}[\ell[i]]$ 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 LCB 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:*

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

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

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

Proof Sketch. Basically we use the proof of the previous theorem. Except we divide arms into two groups. First group consists of arms i for which $\Delta_i < \sqrt{\frac{K\log n}{n}}$ and second group consists of arms

i for which $\Delta_i \geq \sqrt{\frac{K \log n}{n}}$. Now note that by Eq. 1,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell_t[I_t] - \min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\ell_t[i]\right] \\
&= \frac{1}{n}\sum_{j=1}^{K}\mathbb{E}\left[n_{j,n}\right]\Delta_j \\
&= \frac{1}{n}\left(\sum_{j\in[K]:\Delta_j<\sqrt{\frac{K\log n}{n}}}\mathbb{E}\left[n_{j,n}\right]\Delta_j + \sum_{j\in[K]:\Delta_j\geq\sqrt{\frac{K\log n}{n}}}\mathbb{E}\left[n_{j,n}\right]\Delta_j\right) \\
&\leq \frac{1}{n}\left(\sqrt{\frac{K\log n}{n}}\sum_{j\in[K]:\Delta_j<\sqrt{\frac{K\log n}{n}}}\mathbb{E}\left[n_{j,n}\right] + \sum_{j\in[K]:\Delta_j\geq\sqrt{\frac{K\log n}{n}}}\mathbb{E}\left[n_{j,n}\right]\Delta_j\right) \\
&\leq \frac{1}{n}\left(\sqrt{Kn\log n} + \sum_{j\in[K]:\Delta_j\geq\sqrt{\frac{K\log n}{n}}}\mathbb{E}\left[n_{j,n}\right]\Delta_j\right) \\
&\leq \sqrt{\frac{K\log n}{n}} + \frac{1}{n}\sum_{j\in[K]:\Delta_j\geq\sqrt{\frac{K\log n}{n}}}\mathbb{E}\left[n_{j,n}\right]\Delta_j \\
&\leq \sqrt{\frac{K\log n}{n}} + \frac{1}{n}\sum_{j\in[K]:\Delta_j\geq\sqrt{\frac{K\log n}{n}}}\left(\frac{4\log(n)\sqrt{n}}{\sqrt{K\log n}} + 8\right) \\
&\leq 5\sqrt{\frac{K\log n}{n}} + \frac{8K}{n}
\end{aligned}
$$

□

This proves the theorem.

2 Non-Stochastic Bandit

While one can obtain algorithms for the adaptive adversary that picks losses for arms as they go based on random choices of the learning algorithms so far, for this section we will restrict ourself to the so called oblivious adversary. That is, an adversary that picks losses for the K arms and n rounds in advance but with knowledge of the learning algorithm. Specifically we can think of the protocol as follows:

Adversary picks $\ell_1, \ldots, \ell_n \in [0,1]^K$

For $t = 1$ to n

Learner picks distribution over arms $q_t \in \Delta([K])$

Learner draws arm $I_t \sim q_t$ for that round and suffers loss $\ell_t[I_t]$ (and only $\ell_t[I_t]$ is revealed to the learner).

End For

Our goal is to minimize expected regret given as:

$$\mathbb{E}\left[\text{Reg}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell_t(I_t)\right] - \min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\ell_t(i)$$

In the above, note that I have assumed losses are between 0 and 1 rather than -1 and 1 but we can do this without loss of generality since we can add 1 to every loss and this would leave regret unaltered. IF we further divide losses by 2 it only scales down regret by a factor of 2. So this translation can alway be done without affecting our results.

High level idea: On every round, given learners choice q_t , the fact that $I_t \sim q_t$ and the observation of the number $\ell_t(I_t)$, we compute a vector $\tilde{\ell}_t$ such that

$$\mathbb{E}_{I_t\sim q_t}\left[\tilde{\ell}_t\right] = \ell_t$$

Why can we find such a vector? Well think about the following estimate given q_t and the fact that $I_t \sim q_t$:

$$\tilde{\ell}_t = \frac{\ell(I_t)}{q_t(I_t)}e_{I_t}$$

This is the importance weighted estimate or the inverse propencity scoring. Why does this work? Well note that:

$$\mathbb{E}_{I_t\sim q_t}\left[\tilde{\ell}_t\right] = \sum_{i=1}^{K} q_t(i) \times \frac{\ell(i)}{q_t(i)}e_i = \sum_{i=1}^{K}\ell(i)e_i = \ell$$

Hence the vector $\tilde{\ell}_t$ which at round t puts on coordinate I_t the value of observed loss divided by probability of choosing I_t is indeed an unbiased estimate of ℓ_t. Now given this, we have the following observation:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell_t(I_t)\right] - \min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\ell_t(i) &= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{i\sim q_t}\left[\ell_t(i)\right]\right] - \min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\ell_t(i) \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{i\sim q_t}\left[\mathbb{E}_{I_t\sim q_t}\left[\tilde{\ell}_t(i)\right]\right]\right] - \min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{I_t\sim q_t}\left[\tilde{\ell}_t(i)\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right]\right] - \min_{i\in[K]}\frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n}\tilde{\ell}_t(i)\right]
\end{aligned}$$

Using the fact that expected min is smaller than min expected,

$$\leq \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right]\right] - \mathbb{E}\left[\min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_t(i)\right]$$
$$= \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right] - \min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_t(i)\right] \tag{2}$$

That is, we have shown that regret of our algorithm is bounded by expected regret of the algorithm when the losses are the estimates $\tilde{\ell}_t$ on each round t, and since $\tilde{\ell}_t$ vector can be computed at round t, w.r.t. the $\tilde{\ell}_t$ losses, we are in the full information or non-bandit setting. Hence we can conclude that all we need is access to a full information online learning algorithm to which we can feed the estimated losses.

We already know of such an algorithm, the exponential weights algorithm. IF we treat each arm as a model, then the exponential weights algorithm using the estimated losses would be given by

$$q_{t+1}(i) = \frac{e^{-\eta\sum_{s=1}^{t}\tilde{\ell}_s(i)}}{\sum_{k=1}^{K}e^{-\eta\sum_{s=1}^{t}\tilde{\ell}_s(k)}}$$

So on every round t, we draw $I_t \sim q_t$ and at the end of the round knowing $\ell_t(I_t)$ we compute $\tilde{\ell}_t = \frac{\ell(I_t)}{q_t(I_t)}e_{I_t}$ and use this to update q_{t+1}. Using the observation in Eq. 2, for this algorithm we obtain the following theorem.

Theorem 5. *For the exponential weights algorithm run using estimates $\tilde{\ell}_t$ mentioned above and using step-size $\eta = \sqrt{\frac{2\log(K)}{nK}}$, we have the following bound on expected regret:*

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \leq \sqrt{\frac{2K\log(K)}{n}}$$

Proof Sketch. Using the reduction to full information algorithm we have:

$$\mathbb{E}\left[\mathrm{Reg}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\ell_t(I_t)\right] - \min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\ell_t(i)$$
$$\leq \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right] - \min_{i\in[K]}\frac{1}{n}\sum_{t=1}^{n}\tilde{\ell}_t(i)\right]$$
$$\leq \frac{\log(K)}{n\eta} + \frac{1}{n\eta}\sum_{t=1}^{n}\mathbb{E}\left[\log\left(\mathbb{E}_{i\sim q_t}\left[e^{-\eta(\tilde{\ell}_t(i)-\mathbb{E}_{i\sim q_t}[\tilde{\ell}_t(i)])}\right]\right)\right] \tag{3}$$

where the last line I am using the bound we proved at the end of proof of Claim 1 in Lecture 10 where we analyzed exponential weights algorithm. Next we use a consequence of Taylor's theorem which I am stating here without proof. Look at Lemma A.4 of the "Prediction, Learning and Games" Book by Nicolo Cesa-Bianchi and Gabor Lugosi for a very short proof.

For any 0 mean random variable X, such that X is upper bounded by 1 (lower bound can even by $-\infty$), we have that

$$\log(\mathbb{E}\left[e^X\right]) \leq (e-2)\mathbb{E}\left[X^2\right]$$

Take the random variable X to be $-\eta(\tilde{\ell}_t(i) - \mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right])$ and assume $\eta < 1$. In this case, I claim that X is upper bounded by 1. Why, well note that $\tilde{\ell}_t(i) \geq 0$ and $\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right] = \ell_t(I_t) \leq 1$. Hence, $-\eta(\tilde{\ell}_t(i) - \mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right]) \leq \eta \leq 1$. Hence, using this, we can conclude that for each t,

$$\begin{aligned}
\log\left(\mathbb{E}_{i\sim q_t}\left[e^{-\eta(\tilde{\ell}_t(i)-\mathbb{E}_{i\sim q_t}[\tilde{\ell}_t(i)])}\right]\right) &\leq (e-2)\eta^2\mathbb{E}_{i\sim q_t}\left[\left(\tilde{\ell}_t(i) - \mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right]\right)^2\right] \\
&= (e-2)\eta^2\left(\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)^2\right] - \left(\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)\right]\right)^2\right) \\
&\leq (e-2)\eta^2\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)^2\right] \\
&\leq \frac{\eta^2}{2}\mathbb{E}_{i\sim q_t}\left[\tilde{\ell}_t(i)^2\right] \\
&= \frac{\eta^2}{2}\frac{\ell_t(I_t)^2}{q_t(I_t)}
\end{aligned}$$

where in the last line we used that definition of $\tilde{\ell}_t$ in that on the I_t coordinate it has $\ell_t(I_t)/q_t(I_t)$ and 0 everywhere else. Using this above in Eq. 3 we have:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Reg}_n\right] &\leq \frac{\log(K)}{n\eta} + \frac{\eta}{2n}\sum_{t=1}^{n}\mathbb{E}\left[\frac{\ell_t(I_t)^2}{q_t(I_t)}\right] \\
&= \frac{\log(K)}{n\eta} + \frac{\eta}{2n}\sum_{t=1}^{n}\mathbb{E}\left[\sum_{i=1}^{K} q_t(i)\frac{\ell_t(i)^2}{q_t(i)}\right] \\
&= \frac{\log(K)}{n\eta} + \frac{\eta}{2n}\sum_{t=1}^{n}\sum_{i=1}^{K}\ell_t(i)^2 \\
&= \frac{\log(K)}{n\eta} + \frac{\eta K}{2}
\end{aligned}$$

where in the last line we used the fact that $\ell_t(i)^2 \leq 1$. Using $\eta = \sqrt{\frac{2\log(K)}{nK}}$ we get

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \leq \sqrt{\frac{2K\log(K)}{n}}$$

□