IEOR 8100-001: Learning and Optimization for Sequential Decision Making 01/27/16

Lecture 3: UCB Algorithm, Worst-Case Regret Bound

Instructor: Shipra Agrawal Scribed by: Karl Stratos (= Jang Sun Lee)

1 UCB

1.1 Algorithm

The mechanics of the upper confidence bound (UCB) algorithm is simple. At each round, we simply pull the arm that has the highest empirical reward estimate up to that point *plus* some term that's inversely proportional to the number of times the arm has been played. More formally, define $n_{i,t}$ to be the number of times arm i has been played up to time t. Define $r_t \in [0,1]$ to be the reward we observe at time t. Define $I_t \in \{1 \dots N\}$ to be the choice of arm at time t. Then the empirical reward estimate of arm i at time t is:

$$\hat{\mu}_{i,t} = \frac{\sum_{s=1:\ I_s=i}^{t} r_s}{n_{i,t}} \tag{1}$$

UCB assigns the following value to each arm i at each time t:

$$\text{UCB}_{i,t} := \hat{\mu}_{i,t} + \sqrt{\frac{\ln t}{n_{i,t}}}$$

The UCB algorithm is given below:

UCB
Input: N arms, number of rounds $T \geq N$

1. For $t = 1 \dots N$, play arm t.
2. For $t = N+1 \dots T$, play arm

$$I_t = \arg\max_{i \in \{1 \dots N\}} \text{UCB}_{i,t-1}.$$

Note that we're assuming (at least in this formulation) that we will play for at least N times. Also, we're implicitly updating our empirical estimate (1) whenever we play an arm. Observe that at time t, the algorithm uses the $\text{UCB}_{i,t-1}$, which can be computed using observations made until time $t-1$.

At an intuitive level, the additional term $\sqrt{\frac{\ln t}{n_{i,t}}}$ helps us avoid always playing the same arm without checking out other arms. This is because as $n_{i,t}$ increases, $\text{UCB}_{i,t}$ decreases. Take the 2-arm example: arm 1 with a fixed reward 0.25 and arm 2 with a 0-1 reward following a Bernoulli distribution $\pi = 0.75$. Recall that the greedy strategy (i.e., selecting $\arg\max_{i \in \{1 \dots N\}} \hat{\mu}_{i,t}$) incurs linear regret $R(T) = O(T)$ with constant probability: with probability 0.25, arm 2 yields reward 0, upon which we will always select arm 1 and never revisit arm 2. If we track UCB in this situation, we see that we don't have this problem.

- $(t = 1)$ Arm 1 is played: $\hat{\mu}_{1,1} = 0.25$.

- ($t = 2$) Arm 2 is played: $\hat{\mu}_{2,2} = 0$ (with probability 0.25 this occurs).
- ($t = 3$) Arm 1 is played, because $\text{UCB}_{1,2} = 0.25 + \sqrt{\ln 2} > \text{UCB}_{2,2} = 0 + \sqrt{\ln 2}$
- ($t = 4$) **Arm 2 is played, because** $\text{UCB}_{1,3} = 0.25 + \sqrt{\frac{\ln 3}{2}} \approx 0.9912 < \text{UCB}_{2,3} = 0 + \sqrt{\ln 3} \approx 1.0481$

1.2 Instance-Dependent Regret Analysis

But there is a more fundamental reason for the choice of the term $\sqrt{\frac{\ln t}{n_{i,t}}}$. It is a high confidence *upper bound* on the empirical error of $\hat{\mu}_{i,t}$. Specifically, for each arm i at time t, we must have[1]

$$|\hat{\mu}_{i,t} - \mu_i| < \sqrt{\frac{\ln t}{n_{i,t}}} \tag{2}$$

with probability at least $1 - 2/t^2$. There are two useful bounds we can immediately take from (2):

1. **A lower bound for $\text{UCB}_{i,t}$.** With probability at least $1 - 2/t^2$,

$$\text{UCB}_{i,t} > \mu_i \tag{3}$$

2. **An upper bound for $\hat{\mu}_{i,t}$ with many samples.** Given that $n_{i,t} \geq \frac{4 \ln t}{\Delta_i^2}$, with probability at least $1 - 2/t^2$,

$$\hat{\mu}_{i,t} < \mu_i + \frac{\Delta_i}{2} \tag{4}$$

(3) states that the UCB value is probably as large as the true reward: in this sense, the UCB algorithm is *optimistic*. (4) states that given enough (specifically, at least $\frac{4 \ln t}{\Delta_i^2}$) samples, the reward estimate probably doesn't exceed the true reward by more than $\Delta_i/2$. These bounds can be used to show that UCB quickly figures out a suboptimal arm:

Lemma 1.1. *At any point t, if a suboptimal arm i (i.e., $\mu_i < \mu^*$) has been played for $n_{i,t} \geq \frac{4 \ln t}{\Delta_i^2}$ times, then $UCB_{i,t} < UCB_{I^*,t}$ with probability at least $1 - 4/t^2$. Therefore, for any t,*

$$P\left(I_{t+1} = i \middle| n_{i,t} \geq \frac{4 \ln t}{\Delta_i^2}\right) \leq \frac{4}{t^2}$$

[1] This is derived from the Chernoff/Hoeffding bound, which states that for iid samples $x_1 \ldots x_n \in [0, 1]$ with $\mathbf{E}[x_i] = \mu$,

$$P\left(\left|\frac{\sum_{i=1}^n x_i}{n} - \mu\right| \geq \delta\right) \leq 2e^{-2n\delta^2}$$

Since we use $n_{i,t}$ iid samples of the reward of arm i at time t, we can apply this bound with $\delta = \sqrt{\frac{\ln t}{n_{i,t}}}$ and get (2).

Proof. If both (3) and (4) hold,

$$\begin{aligned}
\mathrm{UCB}_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{\ln t}{n_{i,t}}} &\leq \hat{\mu}_{i,t} + \frac{\Delta_i}{2} && \text{since } n_{i,t} \geq \tfrac{4\ln t}{\Delta_i^2} \\
&< \left(\mu_i + \frac{\Delta_i}{2}\right) + \frac{\Delta_i}{2} && \text{by (4)} \\
&= \mu^* && \text{since } \Delta_i := \mu^* - \mu_i \\
&< \hat{\mu}_{i^*,t} + \sqrt{\frac{\ln t}{n_{i^*,t}}} && \text{by (3)} \\
&= \mathrm{UCB}_{i^*,t}
\end{aligned}$$

The probability of (3) or (4) not holding is at most $4/t^2$ by the union bound. □

The lemma is useful because once $\mathrm{UCB}_{i,t} < \mathrm{UCB}_{i^*,t}$, we will stop playing arm i and prevent it from causing further regret. This is formalized in the following bound on the expected number of pulls of a suboptimal arm i.

Lemma 1.2. *Let $n_{i,T}$ be the number of times arm i is pulled by UCB algorithm run on instance $\Theta = \{\nu_1, \mu_1, \ldots, \nu_N, \mu_N\}$ of the stochastic IID multi-armed bandit prbolem. Then, for any arm i with $\mu_i < \mu^*$,*

$$\mathbb{E}[n_{i,T}] \leq \frac{4\ln T}{\Delta_i} + 8.$$

Proof. Let $\mathbb{1}(A)$ denote indicator of an event, i.e., $\mathbb{1}(A)$ is 1 if event A is true and 0 otherwise. For any arm i, the expected number of times it is played up to round T under UCB is

$$\begin{aligned}
\mathbf{E}[n_{i,T}] &= 1 + \mathbb{E}[\sum_{t=N}^{T} \mathbb{1}(I_{t+1} = i)] \\
&= 1 + \mathbb{E}[\sum_{t=N}^{T} \mathbb{1}\left(I_{t+1} = i, n_{i,t} < \frac{4\ln t}{\Delta_i^2}\right)] + \mathbb{E}[\sum_{t=N}^{T} \mathbb{1}\left(I_{t+1} = i, n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right) \\
&\leq \frac{4\ln T}{\Delta_i^2} + \mathbb{E}[\sum_{t=N}^{T} \mathbb{1}\left(I_{t+1} = i, n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right)] \\
&= \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} P\left(I_{t+1} = i, n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right) \\
&= \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} P\left(I_{t+1} = i \;\middle|\; n_{i,t} \geq \frac{4\ln T}{\Delta_i^2}\right) P\left(n_{i,t} \geq \frac{4\ln t}{\Delta_i^2}\right) \\
&\leq \frac{4\ln T}{\Delta_i^2} + \sum_{t=N}^{T} \frac{4}{t^2} \\
&\leq \frac{4\ln T}{\Delta_i^2} + 8
\end{aligned}$$

1 was added in the first equality to account for 1 initial pull of every arm by the algorithm. For the first inequality, suppose for contradiction that the indicator $\mathbb{1}(I_{t+1} = i, n_{i,t} < L)$ takes value 1 at more than $L-1$ time steps, where $L := \frac{4\ln(T)}{\Delta_i^2}$. Let τ be the time step at which this indicator is 1 for the $(L-1)^{th}$ time. Then, arm i has been pulled at least L times until time τ (including the one initial pull), and for all $t > \tau$, $n_{i,t} \geq L$ which implies

$n_{i,t} \geq \frac{\ln(t)}{\Delta_i^2}$, since $t \leq T$. Thus, the indicator cannot be 1 for any $t > \tau$, contradicting the assumption that the indicator takes value 1 for more than $L-1$ times. This bounds $1 + \mathbb{E}[\sum_{t=N}^{T} \mathbb{1}\left(I_{t+1} = i, n_{i,t} < \frac{4\ln t}{\Delta_i^2}\right)]$ by L.

For second inequality we use Lemma 1.1 to bound the first conditional probability term, and the fact that probabilities are at most 1 to bound the second probability term.

□

Theorem 1.3. *Let $R(T,\Theta)$ denote the regret of UCB algorithm in time T for instance $\Theta = \{\nu_1, \mu_1, \ldots, \nu_N, \mu_N\}$ of the stochastic IID multi-armed bandit prbolem. For all instances Θ, and all $T \geq N$, the expected regret of UCB algorithm is bounded as:*

$$\boldsymbol{E}[R(T,\Theta)] \leq \sum_{i:\ \mu_i < \mu^*} \frac{4\ln T}{\Delta_i} + 8\Delta_i,$$

where $\Delta_i = \mu^* - \mu_i$.

Proof. Using previous lemma, the expected total regret up to round T is:

$$\mathbf{E}[R(T,\Theta)] = \sum_{i:\ \mu_i < \mu^*} \mathbf{E}[n_{i,T}]\Delta_i \leq \sum_{i:\ \mu_i < \mu^*} \frac{4\ln T}{\Delta_i} + 8\Delta_i$$

□

1.3 Instance-Independent Regret Analysis

Theorem 1.3 gives an upper bound on $\mathbf{E}[R(T,\Theta)]$ that is *logarithmic* in T. This is in an optimal form: recall from the last lecture that any reasonable algorithm must suffer $\ln T$ expected total regret, no matter what instance Θ it's given.

However, note that Theorem 1.3 is dependent on a specific instance of arms, parametrized by $\Delta_1 \ldots \Delta_N$. Such bounds are called "instance-dependent" or "problem-dependent bounds". This bound does directly imply a very good worst-case bound: for instance with $\Delta_i = \ln T/T$, then the bound is linear in T which is as bad as the naive ϵ-greedy algorithm.

But a simple trick can be applied on Theorem 1.3 to obtain the following "instance-independent" (aka"problem-independent" or "worst-case") regret bound.

Theorem 1.4. *For all $T \geq N$, the expected total regret achieved by the UCB algorithm in round T is*

$$\boldsymbol{E}[R(T)] = 5\sqrt{NT\ln T} + 8N$$

Proof. For the analysis purposes only, divide the arms into two groups:

1. Group 1 contains "almost optimal" arms with $\Delta_i < \sqrt{\frac{N}{T}\ln T}$.
2. Group 2 contains arms with $\Delta_i \geq \sqrt{\frac{N}{T}\ln T}$.

The total regret is the sum of the regret of each group. The maximum total regret incurred due to pulling arms in Group 1 is bounded by

$$\sum_{i \in \text{Group 1}} n_{i,T}\Delta_i \leq \left(\sqrt{\frac{N}{T}\ln T}\right) \sum_{i \in \text{Group 1}} n_{i,T} \leq T \cdot \sqrt{\frac{N}{T}\ln T} = \sqrt{NT\ln T}$$

where used that $\Delta_i \leq \sqrt{\frac{N}{T}\ln T}$ for all i in group 1, and the trivial bound $\sum_i n_{i,T} \leq T$ on total number of pulls. Next, we apply Lemma 1.2 on every arm in Group 2 to bound the expected regret by

$$\begin{aligned}\sum_{i\in\text{Group 2}} \mathbb{E}[n_{i,T}]\Delta_i \leq \sum_{i\in\text{Group 2}} \frac{4\ln T}{\Delta_i} + 8\Delta_i &\leq \sum_{i\in\text{Group 2}} 4\sqrt{\frac{T\ln T}{N}} + 8 \\ &\leq 4\sqrt{NT\ln T} + 8N\end{aligned}$$

where in the first inequality we used that for all $i \in$ Group 2, $\sqrt{\frac{N}{T}\ln T} \leq \Delta_i \leq 1$. Summing the two inequalities gives the desired result. □