STATS 710 – Seq. Dec. Making with mHealth Applications Sep 15, 2016

Lecture 4: Proof the finite-time upper bound of expected regret for Upper Confidence Bound (UCB)

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Tim NeCamp*

1 Clarifications from Lecture 3

Where do we use the bound $0 \leq d \leq \min_{a:\mu_a<\mu_\star} \Delta_a$? In our proof, we had

$$\begin{aligned}\mathbb{P}(A_t = a) &\leq \frac{\epsilon_t}{K} + 2x_0 e^{-x_0/5} + \frac{4}{\Delta_a^2} e^{-\Delta_a^2 \lfloor x_0 \rfloor/2} \\ &\leq \frac{c}{d^2 t} + \frac{2c}{d^2} \log\left(\frac{(t-1)d^2 e}{2cK}\right)\left(\frac{2cK}{(t-1)d^2 e}\right)^{c/(5d^2)} + \frac{4}{d^2}\left(\frac{2cK}{(t-1)d^2 e}\right)^{c/2}.\end{aligned}$$

The first inequality required no bound on the gap, however to obtain the second inequality we utilized two things:

- To eliminate x_0 we use: $x_0 := \frac{\sum_1^t \epsilon_i}{2K} \geq \frac{c}{d^2}\log(\frac{td^2 e}{2cK})$
- To eliminate Δ_a we use: $0 \leq d \leq min_{a:\mu_a<\mu_\star} \Delta_a$

2 UCB algorithm and finite time regret bound

The algorithm and its analysis are both from [ACBF02].

Algorithm 1 UCB

for $t \leq K$ **do**
  Pick action a when $t = a$.
end for
for $t > K$ **do**
  For each $a \in \mathcal{A}$, compute $\bar{R}_t^a$ and $N_t(a)$.
  Pick action $a = \text{argmax}\, \bar{R}_t^a + \sqrt{\frac{2\log(t)}{N_t(a)}}$.
end for

Theorem 1. *Assume that all distributions $(D_a, \forall a)$ have support in $[0,1]$. Then, under the UCB algorithm:*

$$\mathcal{R}_T(\mathcal{L}_{UCB}, (D_a)_{a\in\mathcal{A}}) \leq \sum_{a:\mu_a<\mu_*} \frac{8\log T}{\Delta_a} + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{a\in\mathcal{A}} \Delta_a\right)$$

3 Questions

3.1 Why is it called "upper confidence bound"?

We have optimism in the face of uncertainty. We can think of a confidence interval for μ_a as $\bar{R}_t^a \pm \sqrt{\frac{2\log(t)}{N_t(a)}}$. We assume that μ_a takes the value of the upper value (optimism) of it's confidence interval, and pick an action accordingly.

3.2 What if the reward distributions have support outside of $[0, 1]$?

As with ϵ-greedy, we really only need sub-Gaussian distributions to use a concentration inequality. The expected regret bound would of course change if we use sub-Gaussian distributions.

4 Proof strategy

For the ϵ-greedy algorithm, we bounded the probability of selecting a suboptimal action at time T. For this proof, we will obtain the result by instead bound $\mathbb{E}[N_T(a)]$. Specifically we will obtain:
$\mathbb{E}[N_T(a)] \leq \frac{8 \log T}{\Delta_a^2} + 1 + \frac{\pi^2}{3}$.

5 Proof of theorem

Let $c_{t,n} := \sqrt{\frac{2 \log t}{n}}$. For arbitrary positive integer l (we will choose its value later), at time T, and for a s.t. $\Delta_a > 0$, we have

$$\begin{aligned}
N_T(a) &= 1 + \sum_{t=K+1}^{T} \mathbb{1}(A_t = a) \\
&\leq l + \sum_{t=K+1}^{T} \mathbb{1}\left(\{A_t = a\} \cap \{N_{t-1}(a) \geq l\}\right) \\
&\leq l + \sum_{t=K+1}^{T} \mathbb{1}\left(\left\{\bar{R}^*_{N^*_{t-1}} + c_{t-1,N^*_{t-1}} \leq \bar{R}^a_{N_{t-1}(a)} + c_{t-1,N_{t-1}(a)}\right\} \cap \{N_{t-1}(a) \geq l\}\right).
\end{aligned}$$

You may want to use a concentration inequality at this point, but the inequality cannot be applied to a random time, $N_{t-1}(a)$. Continuing, we have

$$\begin{aligned}
N_T(a) &\leq l + \sum_{t=K+1}^{T} \mathbb{1}(\exists\ 1 \leq n \leq t-1, l \leq m \leq t-1 \text{ such that } \bar{R}^*_n + c_{t-1,n} \leq \bar{R}^a_m + c_{t-1,m}) \\
&\leq l + \sum_{t=K+1}^{T} \sum_{n=1}^{t-1} \sum_{m=l}^{t-1} \mathbb{1}(\bar{R}^*_n + c_{t-1,n} \leq \bar{R}^a_m + c_{t-1,m}) \\
&\leq l + \sum_{t=1}^{T} \sum_{n=1}^{t} \sum_{m=l}^{t} \mathbb{1}(\bar{R}^*_n + c_{t,n} \leq \bar{R}^a_m + c_{t,m}).
\end{aligned}$$

Thus,

$$N_T(a) \leq l + \sum_{t=1}^{T} \sum_{n=1}^{t} \sum_{m=l}^{t} \mathbb{1}(\bar{R}^*_n + c_{t,n} \leq \bar{R}^a_m + c_{t,m}). \tag{1}$$

Lemma 2. *Let $E = \{\bar{R}^*_n + c_{t,n} \leq \bar{R}^a_m + c_{t,m}\}$ (from last sum above), $F = \{\bar{R}^*_n \leq \mu_* - c_{t,n}\}$, $G = \{\bar{R}^a_m \geq \mu_a + c_{t,m}\}$, $H = \{\mu_* < \mu_a + 2c_{t,m}\}$. Then $E \subset (F \cup G \cup H)$.*

Proof. Assume ¬ F ∩ ¬ G ∩ ¬ H then:

$$\begin{aligned}
\bar{R}^*_n &> \mu_* - c_{t,n} \qquad \text{(since } F \text{ is false)} \\
&\geq \mu_a + 2c_{t,m} - c_{t,n} \qquad \text{(since } H \text{ is false)} \\
&\geq \bar{R}^a_m - c_{t,m} + 2c_{t,m} - c_{t,n} \qquad \text{(since } G \text{ is false)} \\
&= \bar{R}^a_m + c_{t,m} - c_{t,n} \qquad \text{(algebra)}
\end{aligned}$$

So $\bar{R}^*_n + c_{t,n} > \bar{R}^a_m + c_{t,m}$, which shows that (¬ F ∩ ¬ G ∩ ¬ H) ⊂ ¬ E. □

Note that H is non-random, so we choose l to make H false. Specifically, set $l = \left\lceil \frac{8 \log T}{\Delta_a^2} \right\rceil$. Then H is false because

$$\begin{aligned}
\mu_* - \mu_a - 2c_{t,m} = \mu_* - \mu_a - 2\sqrt{\frac{2\log t}{m}} &\geq \mu_* - \mu_a - 2\sqrt{\frac{2\log t}{l}} \\
&\geq \mu_* - \mu_a - 2\sqrt{\frac{2\log T}{l}} \geq \mu_* - \mu_a - \Delta_a = 0.
\end{aligned}$$

Thus, with $l = \left\lceil \frac{8 logT}{\Delta_a^2} \right\rceil$, from (1), we obtain

$$\mathbb{E}\left[N_T(a)\right] \leq l + \sum_{t=1}^{T}\sum_{n=1}^{t}\sum_{m=l}^{t} \mathbb{E}\left[\mathbb{1}(\bar{R}^*_n + c_{t,n} \leq \bar{R}^a_m + c_{t,m})\right] = l + \sum_{t=1}^{T}\sum_{n=1}^{t}\sum_{m=l}^{t} \mathbb{P}\left(\bar{R}^*_n + c_{t,n} \leq \bar{R}^a_m + c_{t,m}\right).$$

Applying Lemma 2, we have

$$\begin{aligned}
\mathbb{E}\left[N_T(a)\right] &\leq l + \sum_{t=1}^{T}\sum_{n=1}^{t}\sum_{m=l}^{t} \left[\mathbb{P}\left(\bar{R}^*_n \leq \mu_* - c_{t,n}\right) + \mathbb{P}\left(\bar{R}^a_m \geq \mu_a + c_{t,m}\right) + \mathbb{P}\left(\mu_* < \mu_a + 2c_{t,m}\right)\right] \\
&= l + \sum_{t=1}^{T}\sum_{n=1}^{t}\sum_{m=l}^{t} \left[\mathbb{P}\left(\bar{R}^*_n \leq \mu_* - c_{t,n}\right) + \mathbb{P}\left(\bar{R}^a_m \geq \mu_a + c_{t,m}\right)\right] \qquad (H \text{ has probability } 0) \\
&\leq l + \sum_{t=1}^{T}\sum_{n=1}^{t}\sum_{m=l}^{t} \left[e^{-2nc_{t,n}^2} + e^{-2mc_{t,m}^2}\right] \qquad \text{(Hoeffding-Azuma)} \\
&\leq l + \sum_{t=1}^{T}\sum_{n=1}^{t}\sum_{m=l}^{t} \left[\frac{1}{t^4} + \frac{1}{t^4}\right] \qquad \text{(definition of } c_{t,n}) \\
&\leq l + \sum_{t=1}^{T} \frac{2t^2}{t^4} \\
&\leq l + \sum_{t=1}^{\infty} \frac{2}{t^2} \\
&= l + \frac{\pi^2}{3}
\end{aligned}$$

because $\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$. Now substitute $l = \left\lceil \frac{8 \log T}{\Delta_a^2} \right\rceil$ to get

$$\mathbb{E}\left[N_T(a)\right] \leq \frac{8 \log T}{\Delta_a^2} + 1 + \frac{\pi^2}{3}.$$

Thus, our regret at time T is

$$\begin{aligned}\mathcal{R}_T(\mathcal{L}, (D_a)_{a\in\mathcal{A}}) &= \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}\left[N_T(a)\right] \\ &\leq \sum_{a\in\mathcal{A}} \Delta_a \left(\frac{8\log T}{\Delta_a^2} + 1 + \frac{\pi^2}{3}\right) \\ &= \sum_{a\in\mathcal{A}} \frac{8\log T}{\Delta_a} + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{a\in\mathcal{A}} \Delta_a\right).\end{aligned}$$

6 Discussion on only having concern for expected regret

We discussed concerns about only focusing on the expected regret, especially in a mobile health setting. For example, even if our algorithms have small expected regret, the variance of the regret could be large. Then it is not unlikely that there will be a few people for whom the regret is extremely large. This might be unethical in medical applications.

References

[ACBF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.