STATS 710 – Seq. Dec. Making with mHealth Applications Nov 22 2016

Lecture 20: PAC(Probably Approximately Correct) learning II

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Aniket Deshmukh*

1 Review

In last lecture we talked about:

- PAC Learning: we want our solution/estimates to be approximately correct with high probability at the end of learning.
- Idea of an algorithm about PAC learning for Multi-Armed Bandits
- Median Elimination (ϵ, δ) algorithm [EDMM02]

In this lecture, we continue our discussion on median elimination (ϵ, δ) algorithm. Specifically, we talk about the lemma that we need, to prove that median elimination(ϵ, δ) is (ϵ, δ)-PAC.

2 Important Lemma

A key lemma is that, with high probability, the drop in the optimality of the best arm within the remaining set of arms is not severe in going from one phase to the next.

Lemma 1. *We have* $\mathbb{P}(\max_{a \in S_l} \ \mu_a \leq \max_{a \in S_{l+1}} \ \mu_a + \epsilon_l) \geq 1 - \delta_l$

Proof. w.l.o.g let us look at the first round. We want to show that

$$P(\mu^* \leq \max_{a \in s_2} \mu_a + \epsilon_1) \geq 1 - \delta_1$$

Let's look at the complement,

$$P(\mu^* > \max_{a \in s_2} \mu_a + \epsilon_1)$$

$$\begin{aligned}
\mu^* > \max_{a \in s_2} \mu_a + \epsilon_1 \quad &\iff \quad \forall a \in s_2, \text{a is not } \epsilon_1 \text{ optimal} \\
&\implies \quad a^* \notin s_2, \text{that means } a^* \text{ was eliminated} \\
&\implies \quad a \in s_2, \hat{\mu}_a^1 \geq \hat{\mu}_{a^*}^1 \\
&\implies \quad \text{that means following set is large enough } (\geq \frac{K}{2}) \\
&\qquad b_1 = \{a \in \mathcal{A} : \mu_a < \mu^* - \epsilon_1 \quad \text{and} \quad \hat{\mu}_a^1 \geq \hat{\mu}_{a^*}^1\}
\end{aligned}$$

Let bad event be the number of arms which are not ϵ-optimal but are empirically better than the best arm. Let $E_1 = \{\hat{\mu}_{a^*}^1 < \mu^* - \frac{\epsilon_1}{2}\}$

$$
\begin{aligned}
P(\text{bad event}) &\leq P(b_1 \geq \frac{K}{2}) \\
&\leq P(E_1)P(b_1 \geq \frac{K}{2} \mid E_1) + P(E_1^c)P(b_1 \geq \frac{K}{2} \mid E_1^c) \\
&\leq P(b_1 \geq \frac{K}{2} \mid E_1^c) + P(E_1) \qquad (1)
\end{aligned}
$$

Now let's consider two terms in (1) seperatly. First let's bound the second term,

$$
\begin{aligned}
P(E_1) &= \exp\left(-\left(\frac{\epsilon_1}{2}\right)^2 \left(\frac{1}{\epsilon/2}\right)^2 \log(\frac{3}{\delta_1}) \right) \qquad \text{(Chernoff Hoefdding)} \\
&= \frac{\delta_1}{3} \qquad (2)
\end{aligned}
$$

Now for the first term,

$$
\begin{aligned}
P(b_1 \geq \frac{K}{2} \mid E_1^c) &\leq \frac{\mathbb{E}[b_1 \mid E_1^c]}{K/2} \qquad \text{(Markov inequality)} \\
&\leq \frac{K\delta_1/3}{K/2} = \frac{2\delta_1}{3} \qquad \text{(We will prove this next)} \qquad (3)
\end{aligned}
$$

Now all we need to show is $\mathbb{E}[b_1 \mid E_1^c] \leq \frac{K\delta_1}{3}$

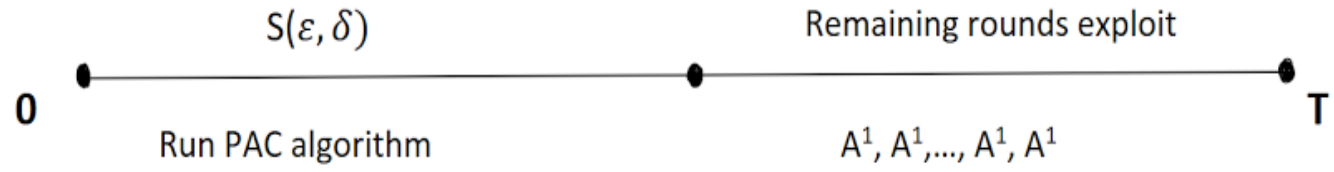
$$
\begin{aligned}
\mathbb{E}[b_1 \mid E_1^c] &= \mathbb{E}[\sum_{a \in A, \text{ a isn't } \epsilon_1 \text{ optimal}} \mathbb{1}_{[\hat{\mu}_a^1 \geq \hat{\mu}_{a^*}^1 | E_1^c]}] \\
&= \sum_{a \in A, \text{ a isn't } \epsilon_1 \text{ optimal}} P(\hat{\mu}_a^1 \geq \hat{\mu}_{a^*}^1 \mid E_1^c) \qquad (4)
\end{aligned}
$$

Now consider an action which isn't ϵ_1 optimal.

$$
\begin{aligned}
P(\hat{\mu}_a^1 \geq \hat{\mu}_{a^*}^1 \,\Big|\, E_1^c) &= P(\hat{\mu}_a^1 \geq \hat{\mu}_{a^*}^1 \mid \hat{\mu}_{a^*}^1 \geq \mu^* - \frac{\epsilon_1}{2}) \\
&\leq P(\hat{\mu}_a^1 \geq \mu^* - \frac{\epsilon_1}{2} \mid \hat{\mu}_{a^*}^1 \geq \mu^* - \frac{\epsilon_1}{2}) \\
&\leq P(\hat{\mu}_a^1 \geq \mu_a + \frac{\epsilon_1}{2} \mid \hat{\mu}_{a^*}^1 \geq \mu^* - \frac{\epsilon_1}{2}) \\
&= P(\hat{\mu}_a^1 \geq \mu_a + \frac{\epsilon_1}{2}) \\
&\leq \exp\left(-\left(\frac{\epsilon_1}{2}\right)^2 \left(\frac{1}{\epsilon_1/2}\right)^2 \log\left(\frac{3}{\delta_1}\right) \right) \\
&= \frac{\delta_1}{3} \qquad (5)
\end{aligned}
$$

Using (5) and (4) we can conclude (3). Finally from (2) and (3) we can complete the proof of lemma. □

Figure 1: PAC MAB algorithm



3 Relation between PAC setting and regret bound

Suppose we are given a (ϵ, δ) PAC algorithm with sample complexity $S(\epsilon, \delta)$. We could run this algorithm until $S(\epsilon, \delta)$ and then start exploiting the best arm (let's call it A^1) until remaining time $T - S(\epsilon, \delta)$.

Let R_T be the total regret until time T.

$$R_T = S(\epsilon, \delta) + \epsilon(T - S(\epsilon, \delta))(1 - \delta) + \delta(T - S(\epsilon, \delta)) \tag{6}$$

If $S(\epsilon, \delta) = \frac{k \log(\frac{1}{\delta})}{\epsilon^2}$ and $\delta = \frac{1}{T}$, then

$$\begin{aligned} R_T &= S(\epsilon, \delta) + \epsilon(T - S(\epsilon, \delta))(1 - \delta) + \delta(T - S(\epsilon, \delta)) \\ &\leq \frac{k \log(\frac{1}{\delta})}{\epsilon^2} + (1 - \delta)\epsilon T + \delta T \\ &= \frac{k \log(T)}{\epsilon^2} + \epsilon T + 1 \end{aligned} \tag{7}$$

by tuning ϵ we get total regret of the order $O(T^{2/3})$.

References

[EDMM02] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.