EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 8 - 2/11/2021

Lecture 8 : Wrapping Thompson Sampling and Majority Vote Algorithm

Lecturer: Jiantao Jiao *Scribe: Isaac Meza / Tianjun Zhang*

1 Thompson Sampling

We first finish our discussion on Thompson sampling.

Lemma 2. $\sum_{t=1}^{T} P(i(t)=i, \overline{\mathbb{E}_i^{\mu}(t)}) \leq \frac{1}{D(X_i||\mu_i)} + 1$

Remark This shows that the probability of $\mathbb{E}_i^{\mu}(t)$ doesn't hold is upper bounded by the quantity of $\frac{1}{D(X_i||\mu_i)} + 1$.

Lemma 3. $\sum_{t=1}^{T} P(i(t)=i, \overline{\mathbb{E}_i^{\theta}(t)}, \mathbb{E}_i^{\mu}(t)) \leq L_i(T) + 1$

Remark This shows that the probability of $\mathbb{E}_i^{\theta}(t)$ doesn't hold and $\mathbb{E}_i^{\mu}(t)$ holds is upper bounded by the quantity of $L_i(T)+1$.

Lemma 4. *Let τ_j be the time step that we pull arm 1 the j^{th} time, then:*
$\mathbb{E}[\frac{1}{P_{i,\tau_j+1}}] \leq \begin{cases} 1+\frac{3}{\Delta_i}, if j < \frac{8}{\Delta_i'}, \\ 1+\Theta(e^{-(\Delta_i')^2 j/2} + \frac{1}{(j+1)(\Delta_i')^2} e^{-D_i j} + \frac{1}{e^{(\Delta_i')^2 j/4}-1}), o.w. \end{cases}$
Here, $\Delta_i' = \mu_1 - \frac{1}{i}$ and $D_i = D(y_i||\mu_1)$, $P_{i,t} = P(\theta_1, t) > y_i|F_{t-1}$, $\mu_i < x_i < y_i < \mu_1$

Remark This illustrates that $P_{i,t}$ is close to 1, so we can bound the expectation by an upper bound. Here we consider τ_j instead of time t since we cannot tell how many times we have pulled arm 1 in t. Note that the bound is decaying very fast.

Now given that $\tau_0 = 0$ and τ_j is the j^{th} time arm 1 being pulled, we have:

$$\begin{aligned} \mathbb{E}[N_i(T+1)] &= \sum_{t=1}^{T} P(i(t)=i) && (1) \\ &= \sum_{t=1}^{T} P(i(t)=i, \mathbb{E}_i^{\theta}(t), \mathbb{E}_i^{\mu}(t)) + P(i(t)=i, \overline{\mathbb{E}_i^{\theta}(t)}, \mathbb{E}_i^{\mu}(t)) + P(i(t)=i, \overline{\mathbb{E}_i^{\mu}(t)}) && (2) \end{aligned}$$

Now we have:

$$\begin{aligned} \sum_{t=1}^{T} P(i(t)=i, \mathbb{E}_i^{\theta}(t), \mathbb{E}_i^{\mu}(t)) &\leq \sum_{t=1}^{T} \mathbb{E}[\frac{1-P_{i,t}}{P_{i,t}} \mathbb{I}(i(t)=1, \mathbb{E}_i^{\theta}(t), \mathbb{E}_i^{\mu}(t))] && (3) \\ &\leq \sum_{t=0}^{T-1} \mathbb{E}[\frac{1-P_{i,T_j+1}}{P_{i,T_j+1}} \sum_{t=\tau_j+1}^{\tau_{j+1}} \mathbb{I}(i(t)=1)] && (4) \\ &\leq \mathbb{E}[\frac{1}{P_{i,\tau_j+1}} - 1] && (5) \end{aligned}$$

Now we are ready to prove the $\sqrt{KT\log T}$ bound.

Proof Since $\mu_i < x_i = \mu_i + \frac{\Delta_i}{3} < y_i = \mu_1 - \frac{\Delta_i}{3} < \mu_i$, now we have some qualities: $\Delta_i' = \mu_1 - y_i\frac{\Delta_i}{3}, D(x_i||\mu_i) \geq \Delta_i^2, D(x_i||y_i) \geq \Delta_i^2, D(P||Q) \geq \frac{1}{2}TV^2(P,Q), L_i(T) = \frac{\log T}{D(x_i||y_i)} \leq \frac{\log T}{\Delta_i^2}, \frac{1}{D(x_i||\mu_i)} \leq \frac{1}{\Delta_i^2}$. We also want to bound the Eq. (5). We can rewrite:

$$\sum_{j=0}^{T-1} e^{-(\Delta_i')^2 j/2} + \frac{1}{(j+1)(\Delta_i')^2} e^{-D_i j} + \frac{1}{e^{(\Delta_i')^2 j/4} - 1} \tag{6}$$

$$\leq \sum_{j=0}^{T-1} \frac{1}{\Delta_i^2} + \frac{1}{\Delta_i^2(j+1)} + \frac{1}{j\Delta_i^2} \tag{7}$$

Given that $\sum_{j=1}^{T-1}\frac{1}{j+1} \leq \log T$, we can then apply the results we obtained in UCB algorithm to get the bound we want.

□

2 Full-feedback model

In the full-feedback setting, in the end of each round, we observe the outcome not only for the chosen arm, but for all other arms as well.

In general, the problem can be casted as playing a game with an adversary:

2.1. *Full-feedback & adversarial cost*

For each round $t \in [T]$:

1. *Adversary chooses cost $c_t(a)$ for each arm $a \in [K]$*
2. *Algorithm picks arm a_t*
3. *Algorithm receives cost $c_t(a_t)$ for the chosen arm*
4. *Algorithm observes $c_t(a)$ for all arms.*

An adversary is called oblivious if the chosen costs at each round t do not depend on the algorithm's choices before round t. Otherwise, it is called *adaptive*.

Now we explore another algorithm for making sequential decision making in this full-feedback setting. The setting is one in which each time we choose A/B in an adversarial environment. In particular, the adversary knows the strategy but not the concrete observations. Then chooses the whole sequence of correct labels for each step. The goal is to perform as well as the best expert in hindsight given a family of experts.

Suppose we need to predict labels for observations, and we are assisted with a committee of experts. In each round, a new observation arrives, and each expert predicts a correct label for it. We listen to the experts, and pick an answer to respond.

Theorem 1. *Given a pool of N experts, assume the best makes L mistakes. Then*

(i) There is an efficient deterministic algorithm that can guarantee at most $2(1+\epsilon)L + 2\frac{\log N}{\epsilon}$ mistakes

(ii) There is an efficient randomized algorithm such that

$$\mathbb{E}[M] \le (1+\epsilon)L + \frac{\log N}{\epsilon}$$

where M is the number of mistakes.

Intuitively the algorithm follows the majority of experts weighted by their accuracy in the past - We update the weights over time, decreasing the weight of a given expert whenever he makes a mistake. We call this the *weighted majority vote algorithm.*

Algorithm 1: Weighted Majority

```
parameter: $\epsilon \in [0,1]$
Set $w_1(i) \leftarrow 1$ , $\forall\, i \in [N]$
for each round $t$ do
    Define $S_t(j)$ the set of experts that chose $j \in \{A, B\}$;
    Define $W_t(j) = \sum_{i \in S_t(j)} w_t(i)$, for $j \in \{A, B\}$;
    Predict based on the weighted majority: $a_t \leftarrow \begin{cases} A & W_t(A) \ge W_t(B) \\ B & \text{o.w.} \end{cases}$;
    for each expert $i$ do
        $w_{t+1}(i) \leftarrow \begin{cases} w_t(i) & \text{if correct} \\ w_t(i)(1-\epsilon) & \text{o.w.} \end{cases}$;
```

Proof

(i) Let $W_t = W_t(A) + W_t(B) = \sum_{i \in [N]} W_t(i)$, and define i^* to be the best expert. Observe that $W_t > w_t(i^*) = (1-\epsilon)^L$, and $W_1 = N$. Each time the algorithm makes a mistake, the weighted majority also makes a mistake, meaning that at least half of the experts also made a mistake, so at least half of the total weight decreases by a factor of $(1-\epsilon)$. Suppose that at round t the correct label was A, and the algorithm makes a mistake:

$$\begin{aligned} W_{t+1} = \sum_{i \in [N]} W_{t+1}(i) &= \sum_{i \in S_t(A)} w_t(i) + \sum_{i \in S_t(B)} (1-\epsilon) w_t(i) \\ &= W_t - \epsilon \sum_{i \in S_t(B)} w_t(i) \\ &\le W_t - \frac{\epsilon}{2} W_t \\ &= W_t(1 - \frac{\epsilon}{2}) \end{aligned}$$

where the last inequality follows since the incorrect prediction must have the majority vote.

Then,

$$\frac{(1-\epsilon)^L}{N} \le \frac{W_T}{W_1} = \prod_{i=1}^{T-1} \frac{W_{t+1}}{W_t} \le (1 - \frac{\epsilon}{2})^M$$

with M the number of mistakes of the algorithm.

Taking logarithms and using the inequality $\ln(1-x) < -x$ for $x \in (0,1)$ gives

$$L \ln(1-\epsilon) - \ln(N) < M \ln(1 - \frac{\epsilon}{2}) < M(-\frac{\epsilon}{2})$$

and after rearrenging,

$$M < L\frac{2}{\epsilon}\ln(\frac{1}{1-\epsilon} + \frac{2}{\epsilon}\ln N < \frac{2}{1-\epsilon}L + 2\frac{\log N}{\epsilon} \leq 2(1+\epsilon)L + 2\frac{\log N}{\epsilon}$$

□

References

[1] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on learning theory.* JMLR Workshop and Conference Proceedings, 2012, pp. 39–1.

[2] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on thompson sampling," *arXiv preprint arXiv:1707.02038*, 2017.

[3] S. Agrawal and N. Goyal, "Further optimal regret bounds for thompson sampling," in *Artificial intelligence and statistics.* PMLR, 2013, pp. 99–107.

[4] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *International Conference on Machine Learning.* PMLR, 2014, pp. 100–108.

[5] B. Kveton, C. Szepesvari, S. Vaswani, Z. Wen, T. Lattimore, and M. Ghavamzadeh, "Garbage in, reward out: Bootstrapping exploration in multi-armed bandits," in *International Conference on Machine Learning.* PMLR, 2019, pp. 3601–3610.

[6] A. Slivkins, "Introduction to multi-armed bandits," *CoRR*, vol. abs/1904.07272, 2019. [Online]. Available: http://arxiv.org/abs/1904.07272