EE675 - IIT Kanpur, 2022-23-II

QUIZ 1 : SOLUTIONS

01.02.23

Lecturer: Prof. Subrahmanya Swamy Peruru *Scribe: Hemant Dhakar and Visha Kumari*

Question 1

0.1

given: a(t) be the arm played in round t
and Δ(a(t)) = μ(a$*$) $-$ μ(a(t))
then the regret R(T) will be

$$R(T) = \sum_{t=1}^{\boldsymbol{T}} \Delta(a(t))$$

taking Expectation both side

$$E[R(T)] = E[\sum_{t=1}^{\boldsymbol{T}} \Delta(a(t))] = \sum_{t=1}^{\boldsymbol{T}} E[\Delta(a(t))]$$

0.2

In exploration phase of ϵ - greedy algorithm: In each round, the ϵ - greedy algorithm will randomly select an arm with probability ϵ . i.e., The probability of selecting a particular arm 'a' due to this exploratory behavior is ϵ / k
i.e.

$$P(a(t) = a) \geq \epsilon/k \tag{1}$$

since with 1-ϵ probability also the the arm can be selected if it is the arm with highest sample average so far.
Indicator to check if arm a is picked in j^{th} round is

$$n_t(a) = \sum_{j=1}^{\boldsymbol{t}} 1_{(a(j)=a)}$$

taking expectation both side

$$E[n_t(a)] = E[\sum_{j=1}^{\boldsymbol{t}} 1_{(a(j)=a)}]$$

$$= \sum_{j=1}^{\boldsymbol{t}} E[1_{(a(j)=a)}] = \sum_{j=1}^{\boldsymbol{t}} P(a(j) = a)$$

Using Equation (1) we get

$$E[n_t(a)] \geq \sum_{j=1}^{t} \epsilon/k = \epsilon t/k$$

0.3

$$\Delta(\tilde{a}(t)) = \mu(a^*) - \mu(\tilde{a}(t))$$

Assuming confidence intervals correct

$$\leq \left(\bar{\mu}_{t-1}(a^*) + \sqrt{\frac{ln(T)}{n_{t-1}(a^*)}}\right) - \left(\bar{\mu}_{t-1}(\tilde{a}(t)) - \sqrt{\frac{ln(T)}{n_{t-1}(\tilde{a}(t)}}\right)$$

$$= \underbrace{(\bar{\mu}_{t-1}(a^*) - \bar{\mu}_{t-1}(\tilde{a}(t)))}_{\leq 0 \text{ (because } \tilde{a}(t) \text{ has best sample average)}} + \left(\sqrt{\frac{ln(T)}{n_{t-1}(a^*)}} + \sqrt{\frac{ln(T)}{n_{t-1}(\tilde{a}(t)}}\right)$$

$$\Delta(\tilde{a}(t) \leq \left(\sqrt{\frac{ln(T)}{n_{t-1}(a^*)}} + \sqrt{\frac{ln(T)}{n_{t-1}(\tilde{a}(t)}}\right) \tag{2}$$

we have assumed that confidence intervals of both the arms is correct while deriving this bound. Now let us find out what is the probability of that happening.
From Hoeffding inequality, we have
$\forall a$,

$$P\left(\mid \bar{\mu}_{t-1}(a) - \mu(a) \mid \geq \sqrt{\frac{ln(T)}{n_{t-1}(a)}}\right) \leq \frac{2}{T^2}$$

i.e., for any 'a'
Prob(confidence interval going wrong) $\leq \frac{2}{T^2}$
Prob(confidence interval going wrong for either $\tilde{a}(t)$ or a^*) $\leq \frac{4}{T^2}$
Prob(both our confidence intervals are correct) $\geq 1 - \frac{4}{T^2}$
using this along with Equation (2) we get

$$P\left(\Delta(\tilde{a}(t) \leq \sqrt{\frac{ln(T)}{n_{t-1}(a^*)}} + \sqrt{\frac{ln(T)}{n_{t-1}(\tilde{a}(t)}}\right) \geq 1 - \frac{4}{T^2}$$

0.4

$$E[\Delta(\tilde{a}(t))] = E[\Delta(\tilde{a}(t))|\text{confidence intervals correct}]\text{P(correct confidence intervals)}$$
$$+E[\Delta(\tilde{a}(t))|\text{confidence interval wrong}]\text{P(wrong confidence interval)} \tag{3}$$

From previous question (0.3), we know that
(i) if confidence intervals are correct.

$$\Delta(\tilde{a}(t) \leq \left(\sqrt{\frac{ln(T)}{n_{t-1}(a^*)}} + \sqrt{\frac{ln(T)}{n_{t-1}(\tilde{a}(t)}} \right)$$

(ii) worst case, if confidence interval is wrong

$$\Delta(\tilde{a}(t) \leq 1$$

(iii)

$$\text{P(confidence intervals correct)} \geq 1 - \frac{4}{T^2}$$

(iv)

$$\text{P(confidence interval wrong)} \leq \frac{4}{T^2}$$

Using these four observations in equation (3) we get

$$E[\Delta(\tilde{a}(t)] \leq \left(\sqrt{\frac{ln(T)}{n_{t-1}(a^*)}} + \sqrt{\frac{ln(T)}{n_{t-1}(\tilde{a}(t)}} + \frac{4}{T^2} \right)$$

0.5

-In ϵ - greedy , we play the arm with the best average so far with a probability of 1 - ϵ .
i.e., with probability 1- ϵ , we play $\tilde{a}(t)$.
-We play a random arm with probability ϵ.
$\Rightarrow$ a(t) = $\tilde{a}(t) \longrightarrow$ with probability 1 - ϵ
a(t) = random arm $\longrightarrow$ with probability ϵ
$\Rightarrow \Delta(a(t))$ = $\Delta(\tilde{a}(t)) \longrightarrow$ with probability 1 - ϵ
worst case
$\Delta(a(t)) \leq 1 \longrightarrow$ with probability ϵ
$\Rightarrow$ E[$\Delta(a(t))$] = (1-ϵ)E[$\Delta(\tilde{a}(t)$] + ϵ.1

0.6

$$E[R(T)] = \sum_{t=1}^{\boldsymbol{T}} E[\Delta(a(t))] \leq 1 + \sum_{t=2}^{\boldsymbol{T}} E[\Delta(a(t))]$$

From question (0.5)

$$E[R(T)] \leq 1 + \sum_{t=2}^{\boldsymbol{T}} \left(\epsilon + (1-\epsilon) \left[\sqrt{\frac{ln(T)}{n_{t-1}(a^*)}} + \sqrt{\frac{ln(T)}{n_{t-1}(\tilde{a}(t)}} + \frac{4}{T^2} \right] \right)$$

$$E[R(T)] \leq 1 + \epsilon T + \frac{4}{T} + \sum_{t=2}^{\boldsymbol{T}} \left(\sqrt{\frac{ln(T)}{n_{t-1}(a^*)}} + \sqrt{\frac{ln(T)}{n_{t-1}(\tilde{a}(t)}} \right)$$

Using the assumption $n_{t-1}(a) \approx$ E$[n_{t-1}(a)]$ and question (0.2) which says E$[n_t(a)] \geq \epsilon t/k$.

$$E[R(T)] \leq 1 + \epsilon T + \frac{4}{T} + \sum_{t=2}^{\boldsymbol{T}} \left(\sqrt{\frac{ln(T)}{\epsilon(t-1)/k}} + \sqrt{\frac{ln(T)}{\epsilon(t-1)/k}} \right)$$

Now substituting $\epsilon = k^{1/3}T^{-1/3}$ gives us

$$E[R(T)] \leq 1 + (k^{1/3}T^{-1/3})T + \frac{\sqrt{kln(T)}}{\sqrt{k^{1/3}T^{-1/3}}} \left(\underbrace{\sum_{t=2}^{\boldsymbol{T}} \frac{2}{\sqrt{t-1}}}_{\leq 4\sqrt{T} \text{(using below relation)}} \right)$$

$$\because \sum_{t=2}^{\boldsymbol{T}} \frac{2}{\sqrt{t-1}} \longrightarrow \sum_{t=1}^{\boldsymbol{T}} \frac{1}{\sqrt{t}} \leq \int_{t=1}^{T} \frac{1}{\sqrt{t}}\, dt$$

$$E[R(T)] \leq k^{1/3}T^{2/3} + \frac{k^{1/3}\sqrt{ln(T)}4\sqrt{T}}{T^{-1/6}}$$

$$E[R(T)] = O(k^{1/3}T^{2/3}\sqrt{lnT}$$

QUESTION II

If X is Beta distributed with Beta(α, β), then the pdf of X is given by

$$f_x(x; \alpha, \beta) = Cx^{\alpha-1}(1-x)^{\beta-1}$$

where C =$(\alpha+\beta-1)!\,(\alpha-1)!\,(\beta-1)!$ is a normalizing constant. Here the notation $\alpha!$ represents the factorial of $\alpha!$. Answer the following questions:

1. Consider a Bernoulli reward distribution whose prior is given to be Beta(α, β). Show that after observing one sample reward $\boldsymbol{r}$, the posterior is given by Beta($\alpha + r$, $\beta + 1 - r$). Note that you have to derive the expression clearly (including the normalization constant).

Solution:

We have a Bernoulli reward distribution with parameter p, and we assume a Beta prior distribution for p with parameters α and β, denoted as $p \sim$ Beta(α, β). The pdf of the prior distribution is given by:

$$f(p; \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha, \beta)},$$

where B(α, β) is the normalizing constant:

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

and $\Gamma(\cdot)$ is the gamma function.

The likelihood of r given p is Bernoulli distributiin given as:

$$f(\text{seeing reward r}|\mu = p) = p^r(1-p)^{1-r}.$$

where $r \in [0, 1]$

Posterior is: (writing$f_p(p)$ as $P(\mu = p)$ and interchangeably using $\int$ or $\sum$):

$$= P(\mu = p|\text{seen a sample reward r})$$
$$= \frac{P(\mu = p, \text{seen a reward r})}{P(\text{seen a reward r})}.$$

Using Bayes' theorem, we have:

$$= \frac{P(\text{seen a reward r}, \mu = p)P(\mu = p)}{\sum_{p\in[0,1]} P(\text{seen a reward r}, \mu = p)P(\mu = p)}.$$
$$= \frac{p^r(1-p)^{1-r}Cp^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^r(1-p)^{1-r}Cp^{\alpha-1}(1-p)^{\beta-1}dp},$$

Simplifying, we get:

$$= \frac{p^{\alpha+r-1}(1-p)^{\beta-r}}{\int_0^1 p^{\alpha+r-1}(1-p)^{\beta-r}dp},$$
$$= Cp^{(\alpha+r)-1}(1-p)^{(\beta+1-r)-1}$$

which is the pdf of a Beta distribution with parameters $\alpha + r$ and $\beta + 1 - r$. Therefore, the posterior distribution of p given the reward r is:

$$p|r \sim \text{Beta}(\alpha + r, \beta + 1 - r).$$

QUESTION III

Consider a simple generalization of Bernoulli reward distribution called the categorical distribution, which takes rewards $\in \{0, 1, 2\}$ instead of just $\{0, 1\}$. The categorical distribution is specified by giving the values of p_0, p_1, p_2, which are nothing but the probabilities of seeing reward as $0, 1, 2$ respectively.

Similarly, consider a prior distribution called the Dirichlet distribution, which is a generalization of the Beta distribution. Specifically, the Dirichlet distribution represented by Dirichlet $(\alpha_0, \alpha_1, \alpha_2)$ specifies a probability distribution on the parameters of the reward distribution $\{p_0, p_1, p_2\}$ whose pdf is given by

$$f(x_0, x_1, x_2; \alpha_0, \alpha_1, \alpha_2)$$
$$= Cx_0^{\alpha_0-1}x_1^{\alpha_1-1}x_2^{\alpha_2-1}, \quad \forall \{(x_0, x_1, x_2) \in [0, 1] \mid x_0 + x_1 + x_2 = 1\}$$

where C is a normalizing constant. Now answer the following questions.

1. Let D^a be the reward distribution of $\text{arm}\, a$, which is a categorical distribution with parameters p_0^a, p_1^a, p_2^a. Compute $\mathbb{E}\left[D^a\right]$.
2. Verify that Dirichlet $(1, 1, 1)$ corresponds to uniform prior.
3. If Dirichlet $(\alpha_0, \alpha_1, \alpha_2)$ is the prior distribution, compute the posterior after seeing one sample reward r. You can use indicator functions $1_{\{r=0\}}$, $1_{\{r=1\}}$, etc., to represent the posterior.
4. **[Bonus question]** : Write the Thompson Sampling algorithm for K arms, assuming that the reward distributions are categorical distributions and the prior distribution is Dirichlet $(1, 1, 1)$.

Hint: Use the fact that in Thompson sampling, if there are two arms a and b, we should play $\text{arm}\, a$ with a probability of $\text{arm}\, a$ being the best arm given the prior distributions, i.e., with probability $\mathbb{P}\left(E\left[D^a\right] > E\left[D^b\right]\right)$ given that D^a and D^b is sampled from the posteriors of arms a and b.

Solution

1. In categorical distribution of an arm a,

$$P(reward = 0) = P_0^a$$
$$P(reward = 1) = P_1^a$$
$$P(reward = 2) = P_2^a$$

To compute the expected value of a categorical distribution with parameters given parameters P_0^a, P_1^a, P_2^a, we use the formula:

$$E[D^a] = 0.P_0^a + 1.P_1^a + 2.P_2^a$$

Under the Dirichlet distribution

Therefore, we have:

$$E[D^a] = \frac{\alpha_1}{\alpha_0 + \alpha_1 + \alpha_2} + 2 \cdot \frac{\alpha_2}{\alpha_0 + \alpha_1 + \alpha_2} = \frac{\alpha_1 + 2\alpha_2}{\alpha_0 + \alpha_1 + \alpha_2}$$

Therefore, the expected reward for arm a under the Dirichlet prior is $\frac{\alpha_1+2\alpha_2}{\alpha_0+\alpha_1+\alpha_2}$.

2. To verify that Dirichlet(1, 1, 1) corresponds to a uniform prior, we need to show that the pdf of Dirichlet(1, 1, 1) is constant.

For Dirichlet (α_0 ,α_1, α_2)

$$f_x(x_0, x_1, x_2; \alpha_0, \alpha_1, \alpha_2) = C x_0^{\alpha_0 - 1} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \forall \{(x_0, x_1, x_2) | \sum x_i = 1\}$$

$f_x(x_0, x_1, x_2; 1, 1, 1) = C x_0^{1-1} x_1^{1-1} x_2^{1-1} = C$ (Uniform Distribution over$\{(x_0, x_1, x_2) | \sum x_i = 1\}$,where C is the normalizing constant.

3. The proof is similar to the Question II we solved for Beta Distribution.
   The posterior will turn out to be:

$$\text{Dirichlet}(\alpha_0 + \mathbf{1}_{\gamma=\mathbf{0}}, \alpha_1 + \mathbf{1}_{\gamma=\mathbf{1}}, \alpha_2 + \mathbf{1}_{\gamma=\mathbf{2}})$$

   Just like for Beta Distribution we had

$$\text{Beta}(\alpha + \mathbf{1}_{\gamma=\mathbf{1}}, \beta + \mathbf{1}_{\gamma=\mathbf{0}})$$

4. Let Dirichlet ($\alpha_0^a(t)$,$\alpha_1^a(t)$, $\alpha_2^a(t)$) denote the posterior of arm a at time 't'.
   Since our prior is Dirichlet(1,1,1), we have

$$\alpha_0^a(0) = \alpha_1^a(0) = \alpha_2^a(0) = 1 \forall a$$

   Thompson Sampling (Dirichlet prior categorical rewards)

   $Given : +(\alpha_0^a(0), \alpha_1^a(0), \alpha_2^a(0)) \forall a$ where $\{\alpha_0^a(0), \alpha_1^a(0), \alpha_2^a(0)\}$are prior distribution parameters.

   for each round t ≥ 1 :

   - for each a in A :
     sample ($\tilde{P}_0^a$,$\tilde{P}_1^a$, $\tilde{P}_2^a$) from the posterior at t-1, i.e,
     Dirichlet($\alpha_0^a(t-1)$,$\alpha_1^a(t-1)$, $\alpha_2^a(t-1)$)

   - play arm a(t) = $\arg\max_a E[\text{D}_a \mid \tilde{P}_0^a$,$\tilde{P}_1^a$, $\tilde{P}_2^a$] , i.e, a(t) = $\arg\max_a(0.+1.+2.)$
     If r is the reward obtained, and lets say a(t) = a ,then update posterior of arm a as
     Dirichlet(α_0^a (t-1) + $1_{\gamma=0}$,α_1^a (t-1) + $1_{\gamma=1}$, α_2^a (t-1) + $1_{\gamma=2}$)