EE675 - Introduction to Reinforcement Learning

Lecture 5: Thompson Sampling

23rd January 2023

Lecturer: Subrahmanya Swamy Peruru *Scribes: Gaurang Dangayach , Rick Ghosh*

From the last lecture -

- UCB algorithm & its regret analysis
- Methods used for Bandit problem :
  - Point Estimate based methods : $\mu(a) \approx \bar{\mu}_t(a)$
  - Confidence Interval based methods : $\mu(a) \in [\bar{\mu}_t(a) - \epsilon_t(a), \bar{\mu}_t(a) - \epsilon_t(a)]$
  - Probability Distribution based methods (Bayesian methods) : $P_t(\mu(a) = \theta)$

In today's lecture -

- Thompson Sampling for general reward distribution
- Regret results for Thompson Sampling

1 Thompson Sampling

Given - Prior distribution for each arm i.e. $P_0(\mu(a) = \theta)_{a \in A}$

for round $t \geq 1$, do the following :
  for each arm a :
    Sample $\tilde{\theta}_t(a)$ from $P_{t-1}(\mu(a) = \theta)$ distribution
  Play arm a(t) = $\arg\max_a \tilde{\theta}_t(a)$
  Update the posterior of arm a(t)=a based on the reward 'r' obtained
    $P_t(\mu(a) = \theta) \propto P_{t-1}(R_t = r|\mu(a) = \theta) \times P_{t-1}(\mu(a) = \theta)$

1.1 Bernoulli Reward with Beta Prior

Prior : $\{Beta(\alpha_0(a), \beta_0(a))\}_{a \in A}$

At round t :
  for each arm a :
    Sample $\tilde{\theta}_t(a)$ from $Beta(\alpha_{t-1}(a), \beta_{t-1}(a))$
  Play arm a(t) = $\arg\max_a \tilde{\theta}_t(a)$
  Update the posterior of arm a(t)=a based on the reward 'r' obtained
    $Beta(\alpha_{t-1}(a) + r, \beta_{t-1}(a) + 1 - r)$

1.2 Gaussian Reward with Gaussian Prior

If Prior follows $N(0,1)$ and Reward follows $N(\mu(a),1)$ the it can be shown that Posterior will follow $N(\bar{\mu}_t(a), \frac{1}{\eta_t(a)+1})$ *[Proof as an exercise]*

$P_{t-1}(\mu(a)=\theta) = N(\bar{\mu}_{t-1}(a), \frac{1}{\eta_{t-1}(a)+1})$
$a(t) = a$ and $R_t = r$
$P_t(\mu(a)=\theta) = N(\bar{\mu}_t(a), \frac{1}{\eta_t(a)+1})$ where $\bar{\mu}_t(a) = \frac{\eta_{t-1}(a)\times\ \bar{\mu_{t-1}}(a)+r}{\eta_{t-1}+1}$ and $\eta_t = \eta_{t-1}+1$

Follow-up Exercise: Given Prior and reward distribution, write algorithm for Thompson Sampling.

1.3 Thompson Sampling for General Reward Distributions

- Till now we have used Thompson sampling when reward distribution belong to some special cases such as Bernoulli or Gaussian
- **Question** : Can we use Thompson sampling when the underlying distributions of the arms is not known/doesn't belong to special cases like Bernoulli/Gaussian? **YES** We Can
- If we take a look at the Thompson Sampling algorithm derived for the Gaussian case, the only information we needed to run the algorithm was $\bar{\mu}_t(a), n_t(a)$ for all the arms
- If we want, we can ignore the fact that the algorithm was designed for Gaussian case and blindly want to apply for general distributions, it is feasible to implement
- However, since we are "wrongly"/"blindly" applying the Gaussian algorithm to general distributions, we have to answer questions like:
  (i) Will it work well?
  (ii) Are there any guarantees we can give on regret?

2 Regret Analysis of Thompson Algorithm

Let us understand a few terminologies before we start the regret bounds

2.1 Environment:

- A Bandit environment for a general case is completely specified by giving the distributions of all the arms, i.e.,$\{D_a\}_{a\in A}$ where D_a is the underlying distribution of arm 'a'
- For example, for a Bernoulli case, since $\mu(a)$ is the only parameter required to specify the distribution, the environment can be specified completely by $\{\mu(a)\}_{a\in A}$
- If it is a general Gaussian case, it will have $Env = \{\mu(a), \sigma(a)\}_{a\in A}$

- If we assume unit variance gaussians,

$$Env = \{\mu(a)\}_{a \in A} \qquad [\because \text{Variance = 1 is already given}]$$

2.2 KL - Divergence: (Kullback - Leibler divergence)

- It is a metric which measures how close two distributions are.

Discrete Case

Let P and Q be two probability distributions defined on the same sample space Ω. Then

$$D_{KL}(P||Q) = \Sigma_{x \in \Omega} P(x) log(\frac{P(x)}{Q(x)})$$

For example, consider an experiment of throwing a die which has 3 faces: {1,2,3}
Let P, Q be two distributions on this 3-faced die experiment.

$$Example : \qquad P = \begin{cases} P(1) = 1/3 \\ P(2) = 1/3 \\ P(3) = 1/3 \end{cases} \qquad Q = \begin{cases} Q(1) = 1/4 \\ Q(1) = 1/4 \\ Q(1) = 2/4 \end{cases}$$

Then

$$D_{KL}(f||g) = \Sigma_{x \in \{1,2,3\}} P(x) log(\frac{P(x)}{G(x)}) \ dx$$

Continous Case

Let $f(x)$ and $g(x)$ be two probability distribution functions on some sample space Ω

Then

$$D_{KL}(f||g) = \int_{x \in \Omega} f(x) log(\frac{f(x)}{g(x)}) \ dx$$

$Exercise$:

Let $f_a(x)$ ***be Gaussian*** $N(\mu_a, 1)$
Let $f_b(x)$ ***be Gaussian*** $N(\mu_b, 1)$
Show that $KL(f_a||f_b)$ = $\frac{1}{2}(\mu_a - \mu_b)^2$

2.3 KL Divergence of Bernoulli

Consider Bernoulli (μ_a), Bernoulli (μ_b)

Let $KL(Ber(\mu_a)||Ber(\mu_b))$ be denoted as $KL_{Ber}(\mu_a, \mu_b)$. Then we have the following result:

$$2(\mu_a - \mu_b)^2 \leq KL_{Ber}(\mu_a, \mu_b) \leq \frac{(\mu(a)-\mu(b))^2}{\mu_b(1-\mu_b)}$$
$$\because 0 \leq \mu_b \leq 1 \Rightarrow \mu_b(1-\mu_b) \leq 1/4$$
$$\Rightarrow 2\triangle^2(a) \leq KL(\mu_a, \mu*) \leq 4\triangle^2(a)$$
$$\therefore KL(\mu(a), \mu*) \sim O(\triangle^2(a))$$

2.4 Regret for Thompson Sampling (Bernoulli reward, Beta prior)

For any Bernoulli $Env = \{\mu(a)\}_a$, the expected regret of Thompson Sampling satisfies

$$E[R(T; env)] \leq O(logT)\, \Sigma_{a \neq a*} \frac{\triangle(a)}{KL(\mu(a), \mu*)} \quad \text{, for any Bernoulli env}$$

Since $KL(\mu(a), \mu*) \sim O(\triangle^2(a))$ the above bound gives

$$\leq \frac{O(KlogT)}{\triangle}\text{, where } \triangle = min_a \triangle(a)$$

This is an Instance dependent Bound because it involves $'\triangle'$ term

Similarly instance-independent bound of Thompson Sampling(for Bernoulli case) is

$$R(T) = max_{env \in \{Bernoulli\}} R(T, env) \qquad [\because \text{ it should be applicable for all Bernoulli env}]$$

Which satisfies the following bound

$$R(T) \leq O(\sqrt{KTlogT})$$

NOTE:
Similar results can be shown for Gaussian Thompson Sampling. when rewards are Gaussian distributed

More importantly, Similar regret bounds can be shown for a Bandit problem with general distributions although we blindly use "Gaussian version of Thompson Sampling".

2.5 Bayesian Regret

- We have motivated Thompson Sampling by saying it is very useful if we have some pair information on what the underlying true means are.
- If we already have some information on what values of true means are more likely for us to encounter, it dosen't make sense to design algorithms that work for all possible Environments.

- It makes sense to measure the performance of an algorithm with emphasis on the Environments which are most likely (according to our prior distribution informations).

- For instance we have discussed that for a gambler (lottery machine), we are more likely to see true means according to some prior like this

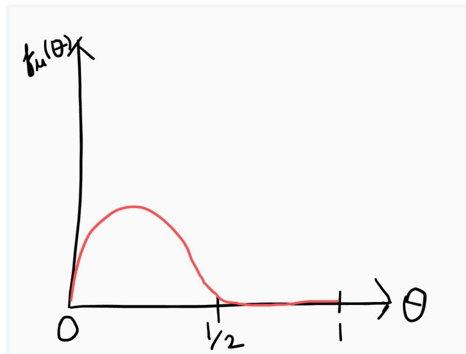

Figure 1: Gambler Prior Distribution

Then a Bayesian regret like $E_{env \sim Prior}[R(T; env)]$ makes more sense than the worst case regret like $sup_{env} R(T; env)$.

- Hence for algorithms like Thompson Sampling, Bayesian regret analysis is widely used.

References:

[1] Shipra Agarwal Notes, Lecture 4