EE675A - IITK, 2022-23-II

Lecture 5: Thompson Sampling for General distributions

23th January 2023

Lecturer: Subrahmanya Swamy Peruru *Scribe: Saurabh | Onkar Dasari*

1 Introduction

Thompson Sampling is a widely used algorithm in the field of Machine Learning and Reinforcement Learning. It is a probabilistic approach that provides a way to balance exploration and exploitation by selecting actions based on the estimated distribution of rewards. The algorithm can be applied to various distributions, making it a general method for sequential decision-making problems. In addition, it has been shown to have regret bounds, which quantify the algorithm's performance compared to an optimal strategy. This lecture aims to explore the concepts of Thompson Sampling for general distributions and the associated regret bounds.

1.1 Recap from last lecture

The last lecture focused on Upper Confidence Bound (UCB) and Thompson Sampling for prior Beta distributions. The UCB algorithm was given as a way to balance exploration and exploitation by adding a confidence constraint to the expected reward of each action.

Basic idea of UCB:

1. Play each arm a $\in \mathcal{A}$ once.
2. For each round t, Play with arm a(t) $\in \mathcal{A}$ such that a(t) = $argmax_a UCB_t(a)$,
   where $UCB_t(a) = \bar{\mu}_t(a) + \epsilon_t(a)$ and $\epsilon_t(a)$ = $\sqrt{\frac{2logT}{n_t(a)}}$

Thompson Sampling was then introduced as a Bayesian alternative to UCB, where the algorithm samples from the posterior distribution of the rewards for each action. Both methods were shown to be effective in balancing exploration and exploitation, but Thompson Sampling was shown to have better performance in some cases, especially when dealing with prior Beta distributions.

Basic idea of Thompson sampling:

1. At $t = 0$, we have prior Beta distributions for $\mu(a)$ for each arm $a \in A$ as: $B(\alpha_0(a), \beta_0(a))$
2. For($t = 1$ to T) do the following:
   (a) For each arm $a \in A$, sample $\widetilde{\theta}_t(a) \sim B(\alpha_{t-1}(a), \beta_{t-1}(a))$

(b) Play arm $a(t) = \underset{a}{} \{\widetilde{\theta}_t(a)\}$

(c) Update the posterior of the arm $a(t)$ based on the reward $r_t \in \{0, 1\}$ received in round t: $\alpha_t(a(t)) = \alpha_{t-1}(a(t)) + r_t, \beta_t(a(t)) = \beta_{t-1}(a(t)) + (1 - r_t)$.

The lecture concluded with a demonstration of how these algorithms can be used to solve the multi-armed bandit problem and the differences in their exploration-exploitation trade-off.

2 Thompson Sampling

2.1 For Bernoulli Rewards and Beta Prior

In the case of Bernoulli rewards, the probability of success for each arm is modeled as a Bernoulli random variable, taking values of 0 or 1 with a certain probability of success, denoted by θ. The prior belief over θ for each arm is modeled as a Beta distribution with parameters α and β, representing the number of successes and failures, respectively, before observing any data.[3]

A Beta prior is commonly used when the rewards are binary and follow a Bernoulli distribution.

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

Algorithm 1 :

```
for each arm $a \in \mathcal{A}$ do
    Sample $\widetilde{\theta}_t(a)$ from $B(\alpha_{t-1}(a), \beta_{t-1}(a))$
Play $argmax_a \widetilde{\theta}_t(a)$
Update posterior for arm $a(t) = a$ to Beta($\alpha_{t-1}(a) + r$, $\beta_{t-1}(a) + 1 - r$)
```

Where 'r' is the reward observed and 'Posterior' refers to the probability distribution over the parameters of a Beta distribution, updated with new data, that represents our current beliefs about these parameters after taking into account the observed data.

2.2 For Gaussian Reward

$$\text{Reward Distribution} = N(\mu(a), 1)$$
$$\text{Prior N}(0, 1) = P_0(\mu(a) = \theta)$$

When the reward follows a Gaussian distribution and the prior is also a Gaussian distribution, the posterior distribution is also a Gaussian distribution, with an updated mean and variance that reflect the new information from the observed data. This is known as Bayes' theorem, which states that the posterior is proportional to the product of the prior and the likelihood.

$$\mathrm{P}_t(\mu(a) = \theta) = \mathrm{N}(\bar{\mu}_{t(a)}, \frac{1}{n_t(a) + 1})$$

By having a Gaussian posterior, the Thompson Sampling algorithm can sample from the posterior distribution efficiently and make decisions in real-time without having to perform computationally intensive numerical methods.

Thompson Sampling for Gaussian reward and N(0,1) prior is a Bayesian approach to solve multi-arm bandit problems. In this case, the reward for each arm is assumed to be Gaussian distributed and the prior belief over the expected reward for each arm is modeled as a normal distribution with mean 0 and standard deviation 1.

Algorithm 2 :

```
Set $\mu_0(a) = 0$ $\forall$ a$\in \mathcal{A}$
for each t $\geq 1$ do
    for each arm $a \in \mathcal{A}$ do
        Sample $\widetilde{\theta}_t(a)$ from $N(\bar{\mu}_{t-1}(a), \frac{1}{n_{t-1}(a)+1})$
    Play $a(t) = argmax_a \widetilde{\theta}_t(a)$
    If $a(t) = a$, update $\bar{\mu}_t(a)$ based on reward observed
```

2.3 For General Distributions

Applying Thompson Sampling algorithm to Gaussian distributions only requires the number of trials for an arm "a", the calculated mean of arm "a", and the Gaussian distribution assumption. However, if we ignore the fact that the algorithm was designed specifically for Gaussian distributions and try to apply it to other general distributions, additional steps may be necessary.This is possible, but we must first define a few terms before stating the regret bounds.

2.3.1 Environment

A bandit environment for a general case is defined by the distributions of all the arms, denoted as $\{D_a\}_{a \in \mathcal{A}}$ where D_a represents the underlying distribution of arm a. This means that the environment is fully specified by the information about the reward distributions for each arm.

For a Bernoulli case we only need $\mu(a)$ to specify the distributions so the environment can be specified completely by $\{\mu_a\}_{a \in \mathcal{A}}$.

For a general Gaussian case, it will be Env = $\{\mu_a, \sigma(a)\}_{a\in\mathcal{A}}$ and if we assume unit variance it becomes Env = $\{\mu_a\}_{a\in\mathcal{A}}$.

2.3.2 KL-Divergence(Kullback-Leibler Divergence)

Kullback-Leibler Divergence (KL-Divergence) is a measure of the difference between two probability distributions. It is a non-symmetric and non-negative metric that quantifies the amount of information lost when approximating one distribution with another. KL-Divergence is commonly used to measure the distance between the true distribution of rewards and the estimated distribution in a bandit algorithm.

1. For a discrete case :
   Let P and Q be two probability distributions defined on the same sample space Ω . Then

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

2. For a continuous case :
   Let $f(x)$ and $g(x)$ be two probability distribution functions on the same sample space Ω . Then

$$D_{KL}(f \parallel g) = \int_{x\in\Omega} f(x) \log \frac{P(x)}{Q(x)}\, dx$$

2.3.3 KL-Divergence of Bernoulli Distribution

Let Ber(μ_a) and Ber(μ_b) be two Bernoulli Distributions . Now we will caculate their KL-Divergence. Denote KL(Ber(μ_a) $\parallel$ Ber(μ_b) as KL$_{Ber}(\mu_a, \mu_b)$. We will use the following result^2 to get the required inequality.

$$KL(f_a \parallel f_b) = \frac{1}{2}(\mu_a - \mu_b)^2$$

where $f_a(x)$ and $f_b(x)$ be distributions with unit variance.

$$\implies 2(\mu_a - \mu_b)^2 \leq KL(\mu_a \parallel \mu_b) \leq \frac{(\mu_a - \mu_b)^2}{\mu_b(1-\mu_b)}$$

$$\because 0 \leq \mu_b \leq 1 \implies \mu_b(1-\mu_b) \leq \frac{1}{4}$$

put $\Delta = (\mu_a - \mu_b)$ and $\max{(\mu_a, \mu_b)} = \mu^*$

$$\implies 2\Delta^2(a) \leq KL(\mu(a), \mu^*) \leq 4\Delta^2(a)$$

$$\therefore KL(\mu(a), \mu^*) \sim O\Delta^2(a)$$

The above eqaution gives us an estimate on KL-Divergence of Bernoulli Distributions.

3 Regret For Thompson Sampling

3.1 For Bernoulli Reward, Beta Prior

For any Bernoulli Env = $\mu(a)_{a \in \mathcal{A}}$, the expected regret satisfies:

$$E[R(T; env)] \leq O(logT) \sum_{a \neq a^*} \frac{\Delta(a)}{KL(\mu(a), \mu^*}$$

For any Bernoulli environment.[1]

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

$$E[R(T; env)] \leq \frac{O(KlogT)}{\Delta}$$

Where $\Delta = min_a \Delta(a)$

Similarly, instance-dependent bound of Thompson Sampling (For Bernoulli Case) is

$$R(T) = \text{max}_{env \in \{Bernoulli\}} R(T, env)$$

Satisfies the following bound

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

Similar regret bounds can be shown for Gaussian Thompson Sampling when rewards are Gaussian distributed.

3.2 Bayesian Regret

We have motivated TS by stating that it is extremely advantageous to have prior knowledge of the underlying true means. If we already know which true mean values are more likely to be encountered, it makes little sense to develop algorithms that function in all potential settings. It makes sense to evaluate the performance of an algorithm with an emphasis on the most probable settings (based on prior distribution knowledge).

References

[1] S. Agrawal. Multi-armed bandits and reinforcement learning lecture 5.

[2] R. S. S. A. G. Barto. *Reinforcement Learning: An Introduction*. Bradford Book, 2018.

[3] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.