**DS 102 Data, Inference, and Decisions** **Spring 2020**

# Lecture 21: Thompson Sampling

*Lecturer: Gireeja Ranade*

## 21.1 Recap of Multi-armed bandits

In the previous lecture we introduced the multi-armed bandit problem. We saw examples of how this problem arises in many different scenarios from search recommendation, advertising, and markets. We also discussed the idea that an algorithm should trade off the exploration of the different options available, and exploitation of its current knowledge of these options. Finally, we looked at a *frequentist* algorithm, the upper confidence bound (UCB) algorithm. We saw that this algorithm is "optimal' in the sense that it achieves a sublinear regret, meaning that it learns and makes a decreasing number of mistakes as time grows.

In keeping with the previous topics we have covered in the class, in these notes we discuss the *Bayesian* approach to solving a multi-armed bandit algorithm: Thompson Sampling.

## 21.2 Multi-Armed Bandit Setup

To begin, we first redefine the multi-armed bandit problem:

We consider a decision-maker who is given $K$ options from which to choose. We refer to these options as *arms*. Associated with each arm is a probability distribution over rewards which is initially unknown to the decision-maker. The decision-maker chooses an arm, usually referred to as *pulling* an arm, and receives a reward sampled from the corresponding reward distribution. This process is repeated over and over again.

### 21.2.1 Mathematical Setup

We now introduce the formal mathematical setup of multi-armed bandits.

Let $\mathcal{A}$ denote a set of $K$ arms. We will denote the reward distribution for arm $a \in \mathcal{A}$ by $P_a$. We denote the choice of arm at time $t$ by $A_t$, and we define $X_{t,A_t}$ as the reward received at time $t$, which is a random sample from distribution $P_{A_t}$.

Let $n$ denote the total number of rounds. Then, our total reward is equal to

$$\sum_{t=1}^{n} X_{t,A_t}.$$

The goal is to find the arm $a$ for which the mean of the corresponding distribution $P_a$ is highest. Informally, we will refer to this mean as "the mean of arm $a$", and we will denote it by

$$\mu_a := \mathbb{E}_{Z \sim P_a}[Z].$$

The highest mean will be denoted by $\mu^* := \max_a \mu_a$. The arm with the highest mean (that is $\mu^*$) will be denoted $a^* := \arg\max_a \mu_a$.

We formally define regret as:

$$R_n = n\mu^* - \mathbb{E}[\sum_{t=1}^{n} X_t].$$

In words, this is the difference between the best possible reward we could get if we knew which arm was the best one, and the expected regret we actually incur.

We also define the suboptimality gap for arm $a$ as:

$$\Delta_a = \mu^* - \mu_a.$$

This is the difference between the mean of the best arm, and a fixed arm $a$.

Finally, we let $T_a(t)$ denote the number of times arm $a$ is selected up to time $t$:

$$T_a(t) = \sum_{s=1}^{t} \mathbf{1}\{A_s = a\}.$$

## 21.3 Thompson Sampling

Given this setup, we will now discuss the Thompson Sampling algorithm for stochastic multi-armed bandits. The algorithm was first introduced in 1933 (!), and remains very popular in practice. In contrast to UCB, Thompson Sampling explicitly allows us to make use of prior information. Indeed, since it is a Bayesian algorithm, it makes use of posterior distributions as opposed to confidence intervals.

The algorithm is as follows: given a prior over the mean of each arm $\pi_a(\mu_a)$, at each round $t = 1, 2...$, the algorithm computes the posterior probability $p_{a,t}$ that arm $a \in \mathcal{A}$ has the highest mean reward:

$$p_{a,t} = \mathbb{P}\left(\mu_a = \max_{a'} \mu_{a'} \;\middle|\; X_{1,A_1}, ..., X_{t-1,A_{t-1}}\right).$$

The choice of arm is then randomly sampled from the distribution $p_t$ over $\mathcal{A}$, where each arm $a \in \mathcal{A}$ has probability $p_{a,t}$:

$$A_t \sim p_t$$

In practice, since the posterior distribution over each arm having the maximum mean is often hard to compute, we often implement a simpler algorithm that nevertheless accomplishes the same task.

**Thompson Sampling Algorithm:** At each round $t = 1, 2....$, you keep track of the posterior distribution over $\mu_a$, for each arm $a \in \{1, ..., K\}$, given all the samples you have observed from that arm $X_{1,a}, ..., X_{T_a(t-1),a}$:

$$P_{a,t} = \mathbb{P}(\mu_a | X_{1,a}, ..., X_{T_a(t-1)}, a).$$

You take one sample from $P_{a,t}$ and choose the arm with the highest sample:

1. $\mu_{a,t} \sim P_{a,t}$ for $a \in \{1, ..., K\}$.
2. Choose arm: $A_t = \underset{a \in \{0,1,...,K-1\}}{\text{argmax}} \mu_{a,t}$

Notice that $\mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} \mu_{a'} | X_1, ..., X_t\right) = \mathbb{P}\left(a = \arg\max_{a' \in \mathcal{A}} \mu_{a',t} | X_1, ..., X_t\right)$ by definition.

**Remark 21.1.** Note that since Thompson Sampling makes use of posterior distributions that you need to sample from, it is mostly used when the posterior has a closed form or is a known distribution which is easy to sample from. In previous lectures we investigated the properties of conjugate priors. Thompson Sampling is one scenario where conjugate priors are extremely useful because the posterior is always guaranteed to remain in the same family given that the samples are from a certain class of distributions and the prior is the conjugate prior for that family.

### 21.3.1 Regret of Thompson Sampling

We now analyze the pseudo-regret of Thompson Sampling when compared to UCB in a simulated multi-armed bandit problem. The results are available in the Thompson Sampling demo posted on the course website, and the solutions for Lab 8. We present them again below. To begin, we see that Thompson Sampling displays sublinear regret and vastly outperforms UCB when the priors are 'good' (in the sense that they reflect the correct ordering of the arms' means).

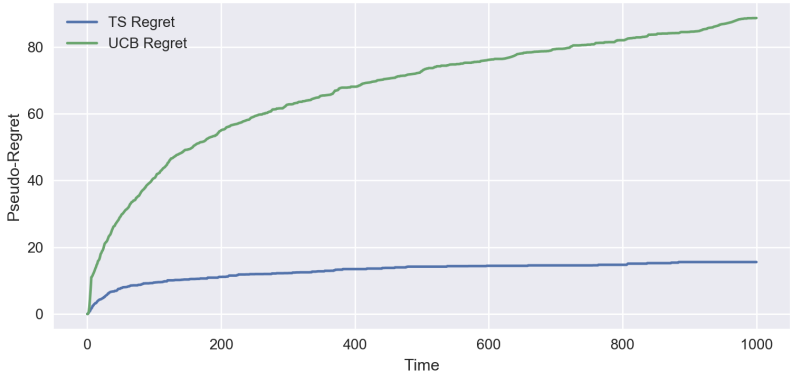


Figure 21.1: We see the pseudo-regret of Thompson Sampling on the Gaussian bandits from Lab 8 when the priors have the correct ordering of the means. Thompson Sampling in this case vastly outperforms UCB (and clearly demonstrates sublinear regret.)

When the priors are 'bad' (in the sense that they have the complete opposite ordering of the arms'

means), however, we see that Thompson sampling has a higher regret than UCB (though we empirically observe that it is still sublinear).

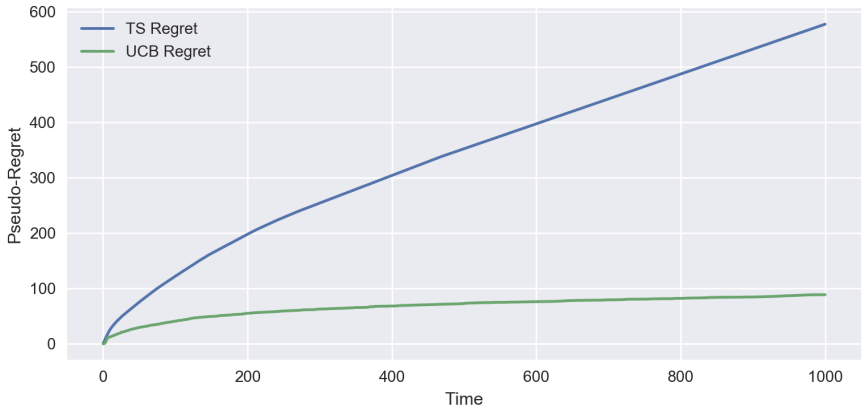


Figure 21.2: We see the pseudo-regret of Thompson Sampling on the Gaussian bandits from Lab 8 when the priors have the incorrect ordering of the means. Thompson Sampling in this case incurs higher regret than UCB (but still demonstrates sublinear regret.)

When the priors are the same for each arm and do not encode any information about the relative ordering of the arms' means we see that Thompson Sampling outperforms UCB once again:

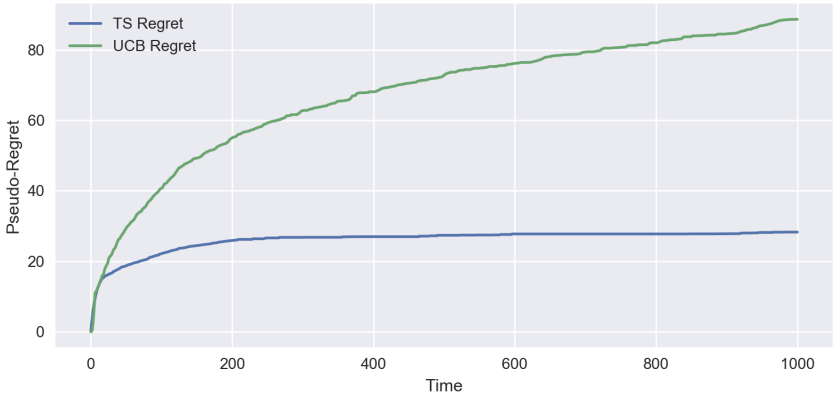


Figure 21.3: We see the pseudo-regret of Thompson Sampling on the Gaussian bandits from Lab 8 when the priors are all the same for the arms. Thompson Sampling in this case outperforms UCB (and clearly demonstrates sublinear regret.)