E0 204: Theory of Multi-Armed Bandits

Naman Mishra

August 2025

Contents

Lectures

The course

Instructor: Shubhada Agrawal

Lecture 1.
Wednesday
August 6

Resources:

(1) Tor Lattimore and Csaba Szepesvári. Bandit Algorithms

(2) Aleksandrs Slivkins. Introduction to Multi-Armed Bandits

Evaluation:

(50%) Homework

(30%) Final exam

(20%) Literature review + presentations

Multi-armed bandits is a special case of reinforcement learning. The setting is the following. We are the learner. There are k arms, labelled $1, 2, \ldots, k$. These have associated unknown probability distributions $P_1, P_2, \ldots, P_k$. At each time step t, we choose one of the arms to pull, and receive a *reward* $X_t \sim P_{\text{chosen arm}}$. The choice may depend on the previous rewards.

This problem was first studied in the 1930s by Thompson in an attempt to dynamize clinical trials—if one drug is shown to be adverse quickly, why put more patients through suffering?

Examples.

- **News websites**. Whenever a user visits a website, the action is choosing a header to display. The reward is 1 if the user clicks and 0 otherwise.
- **Dynamic pricing.** The action is to decide a price p to place. The reward is p if a purchase is made and 0 otherwise.

- **Investments.** Choose a stock to invest in. The reeward is the change in stock price.

In general, there is a trade-off between exploration and exploitation. At each time step, we could repeatedly turn the arm that seems best till now, but we must try different options many times to get to know the best arms.

The problem is theoretically rich, with connections to probability theory, concentration of measure, complexity theory, algorithm design, everything.

Relaxations

Feedback

The above described model may not model the feedback in all scenarios. That is the *bandit feedback* setting. But in the dynamic pricing example, we learn about prices other that the price p chosen.

- If the customer does not make a purchase, we learn that they wouldn't have at all higher prices.
- If the customer does make a purchase, we learn that they would have made a purchase at all lower prices.

This is a *partial feedback* setting.

In the investments example, we get feedback for all stocks. This is a *full feedback* setting, but is largely out of scope for this course and is more closely related to online learning.

Contextual bandits

The algorithm could be receiving some additional information at each time step. In the example of news websites, each user comes with a demographic profile.

Time homogeneity

In the above model, the probability distributions $P_1, P_2, \ldots, P_k$ are independent of time. Choosing from the same arm at different time steps gives iid rewards.

We could instead assume that the distributions change over time in a Markovian manner, but independent of each other.

We might have some structural assumptions on the distributions, such as in the dynamic pricing example.

Constraints

We may be attempting to maximize positive medical results subject to the constraint that nobody fucking dies.

We'll be spending roughly half the course each on the regret minimization probelm and the best arm identification problem, respectively.