Multi-Arm Bandit for Recommendation Systems

Algorithm :

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\left[R - Q(A)\right]$

- Based on the MAB algorithm defined above, we need to first define the actions and rewards for our problem setting.
- Let's assume we have a movie recommendation system, where we need to recommend movies to the user based on their preferences.
- Actions being in our case is to select a optimal action of preferred movie for the user. Each set of actions can represent a collection of movies and each movie can have a corresponding reward distribution.
- The reward can be defined based on user interactions such as clicks, ratings, or watch time after a movie is recommended.
- Our Goal is to maximize the overall reward that is for the system to recommend movies that can engage the user.
- We could model the system/bandit to have an exploit - explore factor based on a probability epsilon.