Thompson Sampling: A Detailed Explanation

Introduction

Thompson Sampling is a Bayesian approach to the multi-armed bandit problem, which is a classic problem in reinforcement learning. The goal of the multi-armed bandit problem is to maximize the total reward by balancing the trade-off between exploring new actions and exploiting known actions. In the context of advertisement click-through rates (CTR), this means optimizing which ads to show to maximize clicks.

Theory Behind Thompson Sampling

Thompson Sampling is a method for making decisions under uncertainty by using probability distributions. The key idea is to maintain a distribution over the parameters that govern the rewards for each action, and to sample from these distributions to decide which action to take.

Bayesian Inference

Thompson Sampling relies on Bayesian inference to update beliefs about the reward probabilities. The process involves:

1. **Prior Distribution**: Initially, a prior distribution is assumed for the success probability of each action. This prior reflects our initial belief about the actions before any data is observed.
2. **Likelihood**: As actions are taken and outcomes are observed (e.g., an ad is clicked or not), the likelihood function updates the prior distribution.
3. **Posterior Distribution**: Using Bayes' theorem, the prior is updated to a posterior distribution that incorporates the observed data.

Beta Distribution

For binary outcomes (e.g., click/no-click), the Beta distribution is commonly used as the prior distribution. The Beta distribution is defined by two parameters, alpha (α) and beta (β), which represent the number of successes and failures, respectively.

- **Initial Prior**: Typically, we start with a uniform prior, Beta(1, 1), indicating no prior knowledge.
- **Updating**: Each time an ad is clicked, α is incremented; if it is not clicked, β is incremented.

Decision Making

At each time step, Thompson Sampling performs the following steps:

1. **Sampling**: For each action, sample a success rate from the corresponding Beta distribution.
2. **Selection**: Choose the action with the highest sampled success rate.
3. **Observation**: Observe the outcome of the chosen action (e.g., whether the ad was clicked).
4. **Update**: Update the Beta distribution for the chosen action based on the observed outcome.

Mathematical Formulation

Let's formalize the Thompson Sampling algorithm:

1. **Initialization**:
   - For each action a , initialize the parameters of the Beta distribution: $\alpha_a = 1$ and $\beta_a = 1$
2. **Loop** (for each time step t):
   - **Sampling**: For each action a , sample θ_a from $\mathrm{Beta}(\alpha_a, \beta_a)$
   - **Selection**: Choose the action $a_t = \mathrm{argmax}_a \theta_a$
   - **Observation**: Observe the reward r_t (e.g., 1 if the ad is clicked, 0 otherwise).
   - **Update**: Update the parameters:
     - If $r_t = 1$: $\alpha_{a_t} \leftarrow \alpha_{a_t} + 1$
     - If $r_t = 1$: $\beta_{a_t} \leftarrow \beta_{a_t} + 1$

Advantages of Thompson Sampling

- **Balancing Exploration and Exploitation**: By sampling from the posterior distributions, Thompson Sampling naturally balances the exploration-exploitation trade-off.
- **Adaptivity**: The algorithm adapts over time based on observed data, making it suitable for dynamic environments.
- **Simplicity**: It is conceptually simple and easy to implement with just a few lines of code.