Thompson Sampling

Thompson Sampling is a popular algorithm used in reinforcement learning for solving multi-armed bandit problems. It is named after William R. Thompson, who first proposed it in 1933.

The basic idea behind Thompson Sampling is to choose the action that has the highest probability of being the optimal action, given the current estimate of the rewards of all the actions. In other words, it is a probability matching algorithm that chooses an action in proportion to its estimated probability of being optimal.

The algorithm works by maintaining a distribution over the rewards of each action. Initially, the distribution is assumed to be uniform or flat, meaning that all actions are equally likely to be the best one. As the algorithm interacts with the environment and receives feedback, it updates the distributions using Bayesian inference. Specifically, it uses the feedback to update the prior distribution and compute a posterior distribution over the rewards.

The algorithm then samples a value from each of the posterior distributions and selects the action with the highest sampled value as the next action to take. This process is repeated for each time step, with the distributions being updated after each action.

Thompson Sampling is useful in situations where the optimal action is not known a priori, and the algorithm must learn from experience which action is best. It is commonly used in online advertising, where advertisers must decide which ad to show to a user based on their previous behavior.

For example, imagine that an online retailer wants to determine which product to show to a user based on their browsing history. Each product corresponds to an action, and the reward is the probability that the user will make a purchase. The Thompson Sampling algorithm can be used to choose which product to show to the user in real time, based on their previous behavior.

Thompson Sampling can outperform other popular algorithms like Epsilon Greedy and Upper Confidence Bound, especially in situations where the optimal action is not clear or changes over time.

In summary, Thompson Sampling is a powerful algorithm for solving multi-armed bandit problems in reinforcement learning. It works by maintaining a distribution over the rewards of each action and sampling from these distributions to choose the next action to take. It is widely used in online advertising, recommendation systems, and other applications where the optimal action is not known a priori.