Thompson Sampling

The Thompson sampling algorithm

Abstract

An abstract

1 Thompson Sampling

Thompson sampling is a strategy that takes a Bayesian approach to selecting optimal actions in a multi-armed bandit setting. Given an assumed family of reward distributions R_θ (i.e. $R_1, \ldots, R_k \sim R_\theta$), and a prior $p(\theta)$ over parameters θ, the Thompson sampling algorithm is as follows:

Algorithm 1 Thompson sampling

for $t = 1, 2, \ldots$ **do**
  Sample $\hat{\theta} \sim p$

  $a_t \leftarrow \text{argmax}_{a \in A} \mathbb{E}_{R_{\hat{\theta}}}[r_t | a_t = a]$
  Execute action a_t, observe reward r_t

  $p \leftarrow p(\theta | a_t, r_t)$

Breaking down a single iteration of the algorithm, we see that we start by sampling a value $\hat{\theta}$ from our current prior p. $\hat{\theta}$ is a fixed parameter value with which we parametrize the reward distribution $R_{\hat{\theta}}$. From here we find the action with parametrized reward distribution having the greatest expected value, and execute that action. We then collect some reward r_t, and update our beliefs about θ given the new information by setting p to the posterior $p(\theta | a_t, r_t)$.

Thompson sampling is a prominent algorithm used in practice, and is much better at exploration and quick convergence than a number of other options. This is largely due to the fact that actions are selected with a probability equal to the likelihood the action is optimal (based on our current observations). Thompson sampling thus selects actions in accordance with how useful the additional data is expected to be in improving the understanding of the action space. Thompson sampling also affords a number of theoretical guarantees, namely bounds on the expected regret.

We can form a bound on the expected regret using information theory. We first define the information ratio

$$\Gamma_t = \frac{(\mathbb{E}[\mu(a^*, \theta) - \mu(a_t, \theta)])^2}{I(a^*; (a_t, r_t) | \mathbb{H}_{t-1})}$$

Intuitively, Γ_t gives an expected "cost" per bit of information acquired; it indicates the degree of tradeoff being made between lost immediate reward and information gained from exploration. We then let $\hat{\Gamma} = \max_{t \in \{1,\dots,T\}} \Gamma_t$, the largest likelihood ratio observed up until time T, and $\mu(a, \theta)$ be the expected reward for action a under parameter θ:

$$\mu(a, \theta) = \mathbb{E}[R_\theta(a_t)|\theta]$$

Then the expected regret can be bounded as follows:

$$\begin{aligned}
\mathbb{E}[\text{Regret}(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mu(a^*, \theta) - \mu(a_t, \theta)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mu(a^*, \theta) - \mu(a_t, \theta)\right] \\
&= \sum_{t=1}^{T} \sqrt{\Gamma_t I(a^*; (a_t, y_t)|\mathbb{H}_{t-1})} && \textit{(Expanding definition of } \Gamma_t\textit{)} \\
&\leq \sqrt{\hat{\Gamma} T \sum_{t=1}^{T} I(a^*; (a_t, y_t)|\mathbb{H}_{t-1})} && \textit{(Jensen's inequality)} \\
&\leq \sqrt{\hat{\Gamma} T H(a^*)} && \textit{(Chain rule of mutual information)}
\end{aligned}$$

Thus, we have a closed form bound for the expected regret based on the total rounds T, the maximum information ratio $\hat{\Gamma}$, and the uncertainty (entropy) about the optimal action a^*, $H(a^*)$.