Thompson Sampling for Bandits

David S. Rosenberg

1 General setup

A k-armed stochastic bandit is described by a probability distribution over a reward vector $R = (R(1), \ldots, R(k)) \in \mathbb{R}^k$. For notational simplicity, we will assume that the distribution of R comes from a parametric family of distributions $p(r \mid q)$ with parameter $q \in \mathcal{Q}$. We'll write q_* for the true, but unknown, parameter corresponding to the distribution of R. We'll write $\mathbb{E}_q$ for expectations taken with respect to $p(r \mid q)$. The expected reward for action a under $p(r \mid q)$ will be of great interest to us, so let us define

$$\mu_a(q) = \mathbb{E}_q \left[R(a) \right],$$

If we knew the true parameter value q_*, then the optimal action would always be

$$a_* = \arg\max_a \mu_a(q_*) = \arg\max_a \mathbb{E}_{q_*} \left[R(a) \right].$$

The reward vectors $R, R_1, R_2, \ldots, R_t$ are generated i.i.d., though we only observe one entry of each reward vector per round. At the beginning of round t, we've collected some partial reward observations, which we'll write as

$$\mathcal{D}_t = \left((A_1, R_1(A_1)), \ldots, (A_{t-1}, R_{t-1}(A_{t-1})) \right).$$

We can use $\mathcal{D}_t$ to help us decide how to choose the action A_t in round t.

2 Going Bayesian

Thompson sampling is a Bayesian approach to choosing the action in every round. When we go Bayesian, we change all unknown parameters into random elements. In our context, the unknown reward parameter $q_* \in \mathcal{Q}$ is replaced by the random element $Q \in \mathcal{Q}$, with some prior distribution that we choose. We'll denote the prior by $p(q)$. At this point, the full rewards distribution is given by

$$\begin{aligned} Q &\sim p(q) \\ R_i \mid Q &\sim p(r_i \mid Q)\ \forall i, \end{aligned}$$

where $R_1, R_2, \ldots$ are conditionally independent given Q. At this point, there is no more "statistics" to do in the usual frequentist sense: there are no parameters to estimate. Everything is just probability theory from now on, where the main operation will be to find the conditional distributions and/or expectations of unknown random variables given the observed data $\mathcal{D}_t$. These are referred to as "posterior distributions" or "posterior means", since they represent our beliefs after (or "posterior" to) seeing the data.

We wish we could choose actions as

$$a = \arg\max_a \mu_a(Q) = \arg\max_a \mathbb{E}\left[R(a) \mid Q\right],$$

but we don't observe Q, so we can't do this. However, at time t we have some information about Q that we can glean from the data $\mathcal{D}_t$, and we can use that to update our expected rewards. A purely exploitative action choice would be select the action for which the posterior mean reward is the largest:

$$a = \arg\max_a \mathbb{E}\left[R(a) \mid \mathcal{D}_t\right].$$

While we could implement this strategy for action selection, it's probably not making a good tradeoff between exploration and exploitation.

3 Thompson sampling

As noted above, ideally we'd choose action a for which $a = \arg\max_a \mu_a(Q)$. Since we don't observe Q, we cannot compute $\mu_a(Q)$. However, with a distribution on Q, we can compute the probability that $a = \arg\max_a \mu_a(Q)$, for each a.

The key idea in Thompson sampling is to select action a with probability $p_a := \mathbb{P}\left(a = \arg\max_{a'} \mu_{a'}(Q) \mid \mathcal{D}_t\right)$, where p_a is the probability that $R(a)$ has the largest expectation under the posterior distribution $p(q \mid \mathcal{D}_t)$. Ties in the $\arg\max$ can be resolved arbitrarily, but consistently (e.g. ties goes to the smallest action).

In general, actually calculating p_a may be difficult or intractable. We can always approximate using Monte Carlo. However, it turns out that you can get a sample from exactly the right distribution, without ever computing the p_a's. I call this the "Thompson sampling trick":

1. Let $Q_t \sim p(q \mid \mathcal{D}_t)$ be a draw from the posterior distribution on Q.

2. Choose action to be $A_t = \arg\max_a \mu_a(Q_t)$.

Now note that

$$\begin{aligned}
\mathbb{P}(A_t = a) &= \mathbb{P}\left(a = \arg\max_{a'} \mu_{a'}(Q_t)\right) \\
&= \mathbb{P}\left(a = \arg\max_{a'} \mu_{a'}(Q) \mid \mathcal{D}_t\right) \\
&= p_a,
\end{aligned}$$

which is exactly the distribution we wanted for A_t.

There are a few important things to recognize about Thompson sampling at this point. First, you should see that Thompson sampling is making a particular tradeoff between exploration and exploitation. Second, note that Thompson sampling is just a heuristic for making this tradeoff. While it does enjoy some optimality properties (see [LS20, Ch 36], and references therein), it's by no means the only reasonable thing to do in the Bayesian setting. 3) The choice of prior can be important for Thompson sampling: if the prior indicates that a particular action has reward that is much lower than it actually is, the corresponding action may never get played under a finite time horizon.

References

[LS20] Tor Lattimore and Csaba Szepesvári, *Bandit algorithms*, Cambridge University Press, 2020.