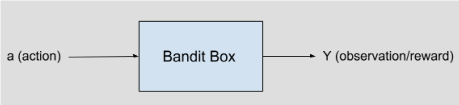
EE194 Lecture 20: Thompson Sampling Regret Bounds

Murtaza Dalal, Russell Mendonca

30 October 2018

1 Review of Thompson Sampling

We have a prior on the best action A^*. Recall that in Thompson Sampling, we draw the action a_t according to $P(A^*|A_1, Y_1, A_2, Y_2....A_{t-1}, Y_{t-1})$. The regret is defined as $\mathbb{E}[Regret(T)] = \mathbb{E}\sum_{t=1}^{T} R(Y_{t,A^*}) - R(Y_{t,A_t})$



2 Linearly Parameterized Bandits

Consider a setting where the rewards observed (Y) are given by

$$Y = a^T\theta^* + \mathcal{N}(0, 1)$$

a is the action taken, and we have $a \in A$, where A is a set of points in R^d
θ^* is d dimensional, and we assume it has a prior $\mathcal{N}(\mu_0, \Sigma_0)$

Algorithm 1 Linearly Parametrized Bandits

```
Initialize $\mu = \mu_0$ , $\Sigma = \Sigma_0$
for iteration $t \in \{1, \dots, T\}$ do
    Draw $\theta_t \sim \mathcal{N}(\mu, \Sigma)$
    Compute $a_t = \text{argmax}_{a \in A}\, a^T\theta_t$
    Plug $a_t$ to obtain $Y_t$
    Set $\mu = \mathbb{E}[\theta|Y_t]$ , $\Sigma = \mathbb{E}[(\theta - \mu)(\theta - \mu)^T|Y_t]$
end for
```

The above algorithm can be used to approximate θ^*. Note that since Y and θ are jointly Gaussian random variables (by definition), $\mathbb{E}[\theta|Y_t] = L[\theta|Y_t]$. where $L[\theta|Y_t] = \mathbb{E}[X] + \frac{cov(X,Y_t)}{var(Y_t)}(Y_t - \mathbb{E}[Y_t])$

3 Towards a Regret Bound for T.S.

Continuing from last lecture, we wish to bound the expected regret of Thompson Sampling in the case where the information ratio (Γ) is bounded. We can write the expected regret as follows:

$$\mathbb{E}[Regret(T)] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[R(Y_{t,A^*}) - R(Y_{t,A_t})\right]\right] \ (1)$$

Where A^* is defined as the optimal bandit/arm and A_t is the Thompson Sampled Action. Here $\mathbb{E}_t$ denotes an expectation conditioned on all past information, i.e. the sequence $((A_0, Y_{0,A_0}), (A_1, Y_{1,A_1}), ..., (A_{t-1}, Y_{t-1,A_{t-1}}))$.
Now recall from last lecture that we defined the information ratio (Γ) as such:

$$\Gamma_t = \frac{\left(\mathbb{E}_t[R(Y_{t,A^*}) - R(Y_{t,A_t})]\right)^2}{I_t(A^*;(A_t, Y_{t,A_t}))} \ (2)$$

Where Γ_t is the information ratio at time t and $I_t(A^*; (A_t, Y_{t,A_t}))$ is the mutual information between the optimal action A^* and the joint distribution of the Thompson sampled action and its corresponding observation. Recall from last lecture that the information ratio measures how much the incremental regret squared changes with the information gained on the optimal action by choosing the particular arm. Essentially we normalize the regret of Thompson sampling by what you learned which quantifies the tradeoff between learning and paying a regret cost. If you will learn a lot relative to your regret, your information ratio will be low, and conversely if you learn little from the action, while paying a lot of regret, your information ratio will be high. We will bound the regret by using an upper bound on this quantity.

Given the above definitions, we make the following claim about the expected regret of Thompson Sampling:

If the information ratio at time t has an upper bound, that is to say if we have $\Gamma_t \leq \Gamma$, then we have the following bound on the expected regret:

$$\mathbb{E}[Regret(T)] \leq \sqrt{T * \Gamma * \mathcal{H}(A^*)} \ (3)$$

Here $\mathcal{H}(A^*)$ is the entropy of the optimal action.

Proof:

$$\begin{aligned}
\mathbb{E}[Regret(T)] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[R(Y_{t,A^*}) - R(Y_{t,A_t})|\left((A_0, Y_{0,A_0}), (A_1, Y_{1,A_1}), ..., (A_{t-1}, Y_{t-1,A_{t-1}})\right)\right]\right] \\
&= \mathbb{E}\sum_{t=1}^{T}\sqrt{\Gamma_t * I_t(A^*; (A_t, Y_{t,A_t}))} \\
&\leq \sqrt{\Gamma}\,\mathbb{E}\sum_{t=1}^{T}\sqrt{I_t(A^*; (A_t, Y_{t,A_t}))} \\
&\leq \sqrt{\Gamma}\sqrt{T}\,\mathbb{E}\sum_{t=1}^{T}\sqrt{I_t(A^*; (A_t, Y_{t,A_t}))} \\
&\leq \sqrt{T * \Gamma * \mathcal{H}(A^*)}
\end{aligned}$$

We expand the expectation of regret using the law of iterated expectation, where we condition on past observations and actions in the inner expectation. From the first line to the second, we use the definition of Γ_t to substitute $\sqrt{\Gamma_t * I_t(A^*; (A_t, Y_{t,A_t}))}$ for $\mathbb{E}_t[R(Y_{t,A^*}) - R(Y_{t,A_t})]$. From the second to the third line, simply plug in the bound $\Gamma_t \leq \Gamma$. For the next two lines in the proof, we will introduce two definitions that will prove to be quite useful:

Cauchy Schwartz Inequality for Random Vectors X,Y:

$$\mathbb{E}[X^TY] \leq \sqrt{\mathbb{E}[X^TX]\mathbb{E}[Y^TY]} \text{ (4)}$$

Mutual Information for Random Variables X,Y :

$$I(X, Y) = \mathcal{H}(X) - \mathcal{H}(X|Y) \text{ (5)}$$

To go from the third line to the fourth line in the proof, we will utilize the Cauchy Schwartz inequality defined above where we define X to be the random vector such that $X_t = \sqrt{I_t(A^*; (A_t, Y_{t,A_t}))}$ and Y to be the random vector of all ones. Then using the Cauchy Schwartz inequality, we $\mathbb{E}[X^TX] = \sum_{t=1}^{T} I_t(A^*; (A_t, Y_{t,A_t}))$ and $\mathbb{E}[Y^TY] = \sum_{t=1}^{T} 1 = T$. Using these, we obtain the fourth line of the proof.

Now we can decompose the mutual information $I_t(A^*; (A_t, Y_{t,A_t}))$ using the definition of Mutual Information as $H(A^*) - H(A^*|A_1, Y_{1,A_1}) + H(A^*|A_1, Y_{1,A_1}) - H(A^*|A_2, Y_{2,A_2}) + ... = H(A^*) - H(A^*|A_{t-1}, ...) \leq H(A^*)$. From here, the last line of the proof follows.

As you can see, if we can bound our information ratio, we can obtain a regret bound that is sublinear in T.

4 Regret Bound for Full Information Feedback Case (Prediction)

Now that we have a general bound for the expected regret, we will derive the bound on the information ratio (step 5 from the previous lecture) as follows:

Proof :

$$\begin{aligned}
\mathbb{E}_t[R(Y_{t,A^*}) - R(Y_{t,A_t})] &= \sum_a \mathbb{P}_t[A^* = a](\mathbb{E}[R(Y_{t,A*})|A^* = a]) - \mathbb{E}[R(Y_{t,A_t})] \\
&\leq \sum_a \mathbb{P}_t[A^* = a]\sqrt{2 * \mathbb{D}_{KL}(\mathbb{P}_t(Y|A^* = a)||\mathbb{P}_t(Y))} \\
&\leq \sqrt{2\sum_a \mathbb{P}_t[A^* = a] * \mathbb{D}_{KL}(\mathbb{P}_t(Y|A^* = a)||\mathbb{P}_t(Y))} \\
&\leq \sqrt{I_t(A^*;Y)} \\
\implies \Gamma_t &\leq \frac{2I_t(A^*;Y)}{I_t(A^*;(Y_{t,A_t},A_t)}
\end{aligned}$$

We obtain the first line through the law of iterated expectations. Then using the bound on $\mathbb{E}[R(Y_{t,A*})|A^* = a]) - \mathbb{E}[R(Y_{t,A_t})$ that we derived last lecture, the next line follows. Now recall the following inequality:
Jensen's Inequality:

$$\text{if a function f is convex, then } f(\mathbb{E}[X]) \leq \mathbb{E}(f(X))$$

Using this inequality, and taking f to be the square root function, which is concave, line 3 follows from line 2 of the proof.
To obtain the final line of the proof, we use the fact that the mutual information between two random variables is given by the following expression:

$$I(X,Y) = \textstyle\sum_x \mathbb{P}[X = x]\mathbb{D}_{KL}(\mathbb{P}[Y|X] \mid\mid \mathbb{P}[Y])$$

Using this equality, if we choose X to be A^* and Y as Y, then we obtain the last line of the proof.