CMSC 858G: Bandits, Experts and Games 09/12/16

Lecture 5: Regret Bounds for Thompson Sampling

Instructor: Alex Slivkins Scribed by: Yancy Liao

1 Regret Bounds for Thompson Sampling

For each round t, we defined a posterior distribution over arms a as:

$$p_t(a) = \mathbb{P}(a = a^*|H_t)$$

where a^* is the best arm for the problem instance and H_t is the history of rewards up to round t. By definition, the Thompson algorithm draws arm a_t independently from this distribution p_t. If we consider a^* to be a random variable dependent on the problem instance, a_t and a^* are identically distributed. We will use this fact later on:

$$a_t \sim a^* \text{ when } H_t \text{ is fixed}$$

Our aim is to prove an upper bound on Bayesian regret for Thompson sampling. Bayesian regret is defined as:

$$\mathop{\mathbb{E}}_{\mu \sim prior}[\mathop{\mathbb{E}}_{rewards}[\, R(t) \mid \mu \,]\,]$$

Notice that Bayesian regret is simply our usual regret placed inside of an expectation conditioned over problem instances. So a regular regret bound (holding over all problem instances) implies the same bound on Bayesian regret.

1.1 Lower/Upper Confidence Bounds

Recall our definition of the lower and upper confidence bounds on an arm's expected rewards at a certain time t (given history):

$$\begin{aligned} r_t(a) &= \sqrt{2 * \tfrac{\log(T)}{n_t(a)}} \\ UCB_t &= \bar{\mu}_t(a) + r_t(a) \\ LCB_t &= \bar{\mu}_t(a) - r_t(a) \end{aligned}$$

Here T is the time horizon, $n_t(a)$ is the number of times arm a has been played so far, and $\bar{\mu}_t(a)$ is the average reward from this arm so far. The quantity $r_t(a)$ is called the *confidence radius*. As we've seen before, $\mu(a) \in [LCB_t(a),\, UCB_t(a)]$ with high probability.

Now we want to generalize this formulation of upper and lower confidence bounds to be explicit functions of the arm a and the history H_t up to round t: respectively, $U(a,\, H_t)$ and $L(a,\, H_t)$. There are two properties we want these functions to have, for some $\gamma > 0$ to be specified later:

$$\forall a,\ t,\quad \mathbb{E}[\,[U(a,\ H_t) - \mu(a)\,]^-] \leq \frac{\gamma}{T \cdot K} \tag{1}$$

$$\forall a,\ t,\quad \mathbb{E}[\,[\mu(a) - L(a,\ H_t)]^-] \leq \frac{\gamma}{T \cdot K}. \tag{2}$$

As usual, K denotes the number of arms. As a matter of notation, x^- is defined to be the negative portion of the number x, that is, 0 if x is non-negative, and $|x|$ if x is negative.

Intuitively, the first property says that the upper bound U does not exceed the mean reward by too much *in expectation*, and the second property makes a similar statement about the lower bound L.

The confidence radius is then defined as $r(a,\ H_t) = \frac{U(a,\ H_t) - L(a,\ H_t)}{2}$.

Lemma 1.1. *Assuming we have lower and upper bound functions that satisfy those two properties, the Bayesian Regret of Thompson sampling can be bounded as follows:*

$$BR(T) \leq 2\gamma + 2\sum_{t=1}^{T} \mathbb{E}[r(a_t,\ H_t)] \tag{3}$$

Proof. First note that:

$$\mathbb{E}[\,U(a^*,\ H_t) \mid H_t\,] = \mathbb{E}[\,U(a_t,\ H_t) \mid H_t\,] \tag{4}$$

This is because, as noted previously, $a^* \sim a_t$ for any fixed H_t, and U is deterministic.

Fix round t. Then the Bayesian Regret for that round is:

$$\begin{aligned}
BR_t(T) &= \mathbb{E}[R(t)] && \text{expectation over everything} \\
&= \mathbb{E}[\mu(a^*) - \mu(a_t)] && \text{instantaneous regret for round t} \\
&= \mathop{\mathbb{E}}_{H_t}[\,\mathbb{E}[\mu(a^*) - \mu(a_t)] \mid H_t]\,] && \text{bring out expectation over history} \\
&= \mathop{\mathbb{E}}_{H_t}[\,\mathbb{E}[U(a_t, H_t) - \mu(a_t) + \mu(a^*) - U(a^*, H_t)] \mid H_t]\,] \\
& && \text{by the equality in equation 4 above} \\
&= \underbrace{\mathbb{E}[U(a_t, H_t) - \mu(a_t)]}_{\text{Part 1}} + \underbrace{\mathbb{E}[\mu(a^*) - U(a^*, H_t)]}_{\text{Part 2}}.
\end{aligned}$$

We will use properties (1) and (2) to bound Part 1 and Part 2. Note that we cannot *immediately* use these properties because they assume a fixed arm a, whereas both a_t and a^* are random variables. (The best arm a^* is a random variable because we take an expectation over everything, including the problem instance.)

Let's handle Part 2:

$$\begin{aligned}
\mathbb{E}[\mu(a^*) - U(a^*, H_t)] &\leq \mathbb{E}[(\mu(a^*) - U(a^*, H_t))^+] \\
&\leq \mathbb{E}[\sum_{\text{arms } a} (\mu(a) - U(a, H_t))^+] \\
&= \mathbb{E}[\sum_{\text{arms } a} (\mu(a) - U(a, H_t))^+] \\
&= \sum_{\text{arms } a} \mathbb{E}[(U(a, H_t) - \mu(a))^-] \\
&\leq k * \frac{\gamma}{T * k} && \text{by 1 over all arms} \\
&= \frac{\gamma}{T}
\end{aligned}$$

Let's handle Part 1:

$$\begin{aligned}
\mathbb{E}[U(a_t, H_t) - \mu(a_t)] &= \mathbb{E}[2r(a_t, H_t) + L(A_t, H_t) - \mu(a_t)] && \text{from definition of r} \\
&= \mathbb{E}[2r(a_t, H_t)] + \mathbb{E}[L(A_t, H_t) - \mu(a_t)]
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}[L(a_t, H_t) - \mu(a_t)] &\leq \mathbb{E}[(L(a_t, H_t) - \mu(a_t))^+] \\
&\leq \mathbb{E}[\sum_{\text{arms } a} ((L(a, H_t) - \mu(a))^+] \\
&= \sum_{\text{arms } a} \mathbb{E}[(\mu(a) - L(a, H_t))^-] \\
&\leq k * \frac{\gamma}{T * k} && \text{by 2 over all arms} \\
&= \frac{\gamma}{T}
\end{aligned}$$

Putting parts 1 and 2 together, (for fixed t) $BR_t(T) \leq \frac{\gamma}{T} + \frac{\gamma}{T} + \mathbb{E}[2r(a_t, H_t)]$.

Summing up over t, $BR(T) \leq 2\gamma + 2\sum_{t=1}^{T} \mathbb{E}[r(a_t,\ H_t)]$ as desired. □

Remark 1.2. Lemma 1.1 holds regardless of what the U and L functions are, so long as they satisfy properties (1) and (2). Furthermore, Thompson Sampling does not need to know what U and L are, either.

Remark 1.3. While Lemma 1.1 does not assume any specific structure of the prior, it embodies a general technique: it can be used to upper-bound Bayesian regret of Thompson Sampling for arbitrary priors, and it also for a particular class of priors (e.g., priors over linear reward functions) whenever one has "nice" confidence bounds U and L for this class.

Let us use Lemma 1.1 to prove a $\mathcal{O}(\sqrt{KT \log T})$ upper bound on regret, which matches the best possible result for Bayesian regret of K-armed bandits.

Theorem 1.4. *Over k arms, $BR(T) = \mathcal{O}(\sqrt{kT \log(T)})$*

Proof. Let the confidence radius be $r(a, H_t) = \sqrt{\frac{2\log T}{n_t(a)}}$ and $\gamma = 2$ where $n_t(a)$ is the number of rounds arm a is played up to time t. Then, by lemma above,

$$BR(T) \leq a + 2\sum_{t=1}^{T} \mathbb{E}[\sqrt{\frac{2\log T}{n_t(a)}}] = \mathcal{O}(\sqrt{\log T}) \sum_{t=1}^{T} \mathbb{E}[\sqrt{\frac{1}{n_t(a)}}]$$

Additionally,

$$\begin{aligned}
\sum_{t=1}^{T} \sqrt{\frac{1}{n_t(a)}} &= \sum_{\text{arms } a} \sum_{t:a_t=a} \frac{1}{\sqrt{n_t(a)}} \\
&= \sum_{\text{arms } a} \sum_{j=1}^{n(a)} \frac{1}{\sqrt{j}} && \text{n(a) is total times arm a is picked} \\
&= \sum_{\text{arms } a} \mathcal{O}(\sqrt{n(a)}) && \text{by taking an integral}
\end{aligned}$$

So,

$$BR(T) \leq \mathcal{O}(\sqrt{\log T}) \sum_{\text{arms } a} \sqrt{n(a)} \leq \mathcal{O}(\sqrt{\log T}) \sqrt{k \sum_{\text{arms } a} n(a)} = \mathcal{O}(\sqrt{kT \log T})),$$

where the last inequality is using the fact that the arithmetic mean is less than (or equal to) the quadratic mean. □

1.2 Thompson Sampling with no prior

We can use Thompson Sampling for the "basic" bandit problem that we have studied before: the bandit problem with IID rewards in $[0, 1]$, but without priors on the problem instances.

We can treat a prior just as a parameter to Thompson Sampling (rather than a feature of reality). This way, we can consider an arbitrary prior (we'll call it a "fake prior), and it will give a well-defined algorithm for IID bandits without priors. We This approach makes sense as long as this algorithm performs well.

Prior work considered two such "fake priors":

(i) independent, uniform priors and 0-1 rewards,

(ii) independent, Gaussian priors and Gaussian unit-variance rewards (so each reward is distributed as $\mathcal{N}(\mu(a), 1)$, where $\mu(a)$ is the mean).

To fully specify approach (i), we need to specify how to deal with rewards r that are neither 0 or 1; this can be handled very easily: flip a random coin with expectation r, and pass the outcome of this coin flip as a reward to Thompson Sampling. In approach (ii), note that the prior allows the realized rewards to be arbitrarily large, whereas we assume the rewards are bounded on $[0, 1]$; this is OK because the algorithm is still well-defined.

We will state the regret bounds for these two approaches, without a proof.

Theorem 1.5. *Consider IID bandits with no priors. For Thompson Sampling with both approaches (i) and (ii) we have:* $\mathbb{E}[R(T)] \leq \mathcal{O}(\sqrt{kT \log T})$.

Theorem 1.6. *Consider IID bandits with no priors. For Thompson sampling with approach (i),*

$$\mathbb{E}[R(T)] \leq (1+\epsilon)(\log T) \underbrace{\sum_{\substack{\text{arms } a \\ \text{s.t.} \Delta(a)>0}} \frac{\Delta(a)}{KL(\mu(a),\mu^*)}}_{(*)} + \frac{f(\mu)}{\epsilon^2},$$

for all $\epsilon > 0$. Here $f(\mu)$ depends on the reward function μ, but not on the ϵ, and $\Delta(a) = \mu(a^*) - \mu(a)$.

The (*) is the optimal constant: it matches the constant in the logarithmic lower bound which we have seen before. So this regret bound gives a partial explanation for why Thompson Sampling is so good in practice. However, it is not quite satisfactory because the term $f(\mu)$ can be quite big, as far as they can prove.

1.3 Bibliographic notes

The results in Section 1.1 are from Russo and Roy (2014). (The statement of Lemma 1.1 is a corollary of the result proved in Russo and Roy (2014) which makes the technique a little more transparent.) Russo and Roy (2014) also use this approach to obtain improved upper bounds for some specific classes of priors, including: priors over linear reward functions, priors over "generalized linear" reward functions, and priors given by a Gaussian Process.

The prior-independent results in Section 1.2 are from (Agrawal and Goyal, 2012; Kaufmann et al., 2012; Agrawal and Goyal, 2013). Specifically, the first "prior-independent" regret bound for Thompson Sampling has appeared in Agrawal and Goyal (2012) (a weaker version of Theorem 1.6). Theorem 1.5 is from Agrawal and Goyal (2013), and Theorem 1.6 is from (Kaufmann et al., 2012; Agrawal and Goyal, 2013).[1] For Thompson Sampling with Gaussian priors (approach (ii)), Agrawal and Goyal (2013) achieve a slightly stronger version of Theorem 1.5 in which the $\log(T)$ factor is replaced with $\log(K)$, and prove a matching lower bound for Bayesian regret of this algorithm.

References

Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. In *25nd Conf. on Learning Theory (COLT)*, 2012.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *16th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pages 99–107, 2013.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *23rd Intl. Conf. on Algorithmic Learning Theory (ALT)*, pages 199–213, 2012.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Math. Oper. Res.*, 39 (4):1221–1243, 2014.

[1] As stated, Theorem 1.6 is from Agrawal and Goyal (2013). The earlier paper (Kaufmann et al., 2012) proves a slightly weaker version in which $\ln(T)$ is replaced with $\ln(T) + \ln\ln(T)$.