MS&E338 Reinforcement Learning Lecture 15 - May 21, 2018

Regret Analysis in Active Learning

Lecturer: Ben Van Roy *Scribe: Raunak Bhattacharyya, Andrea Zanette*

1 Regret Bounds for Ucb

We now sketch the regret decomposition for UCB which is the main starting point for all analyses based on optimism. Precisely, to derive a high probability regret upper bound we need two ingredients:

- Optimism
- Concentration of the confidence intervals

The key idea is that by using optimism we can avoid examining the agent's chosen sequence of actions to compute the regret and instead we recast the problem as a concentration of the confidence intervals. Precisely, for UCB we can write the per-step regret as:

$$\text{REGRET} = \overline{r}(x^*, \theta) - \overline{r}(\overline{x}_t, \theta) = \overline{r}(x^*, \theta) - u_t(\overline{x}_t) + u_t(\overline{x}_t) - \overline{r}(\overline{x}_t, \theta) \tag{1}$$

$$\leq \underbrace{\overline{r}(x^*, \theta) - u_t(x^*)}_{\substack{\text{Pessimism} \\ \leq 0 \text{ with high prob}}} + \underbrace{u_t(\overline{x}_t) - \overline{r}(\overline{x}_t, \theta)}_{\text{Width}}] \tag{2}$$

where u_t is an upper confidence bound function which is deterministic when conditioned on the filtration, i.e., on the history experienced by the agent (u_t typically depends on the visit counts to actions before the start of round t among other things). When conditioned on the filtration, UCB chooses the action deterministically.

Choice of the confidence intervals Assume the rewards are bounded $r \in [0, 1]$. This assumption is non-restrictive since an appropriate rescaling of the problem would achieve this as long as the rewards are compactly supported. Let $\hat{r}(x)$ be the empirical mean of the rewards corresponding to action $x \in \mathcal{X}$. Define the confidence interval $[l_t(x), u_t(x)]$ as follow:

$$l_t(x) = \max\left\{\hat{r}(x) - \sqrt{\frac{2 + 6 \ln T}{n_t(x)}}, 0\right\} \tag{3}$$

$$u_t(x) = \min\left\{\hat{r}(x) + \sqrt{\frac{2 + 6 \ln T}{n_t(x)}}, 1\right\}. \tag{4}$$

In the above expression, $n_t(x)$ is the visit counts to action x prior to the beginning of timestep t, $\hat{r}(x)$ is the empirical mean of the reward and T is the total timestep for which we wish to run the algorithm. Therefore, to use these confidence intervals we need to know in advance for how many timesteps T we need to tun the algorithm; there are however variations of this that do not require knowledge of the time horizon in advance.

We mention that the above specified confidence interval follows from using Hoeffding inequality so that we can guarantee with probability at least $1 - \frac{1}{T}$ that the confidence interval remains valid at all timesteps $t \in 1, 2, \ldots, T$ and for all actions $x \in X$. Notice that this is a tighter requirement than just requiring one empirical mean to fall within its confidence interval with high probability. In other words, we require that the probability that there exist a timestep and action that violates the confidence intervals is low:

$$\mathbb{P}\left(\bigcup_{t=1}^{T} \{\hat{r}(x) \notin [l_t(x), u_t(x)]\}\right) \leq \frac{1}{T}. \tag{5}$$

High Probability Regret Bound for Ucb With probability at least $1 - \frac{1}{T}$ the regret of Ucb with the confidence intervals defined in the previous paragraph is upper bounded by:

$$\textsc{Regret} \leq 2\min\{|\mathcal{X}|, T\} + 4\sqrt{|\mathcal{X}|T(2+6\ln T)}. \tag{6}$$

Some remarks are in order:

- this is a worst-case regret bound. For a specific bandit problem the analysis can be improved
- term $2\min\{|\mathcal{X}|, T\}$ is a really a minimum regret that we could expect: at least all the actions have to be tried once, and if $T < |\mathcal{X}|$ then we won't incur more than T regret since the rewards are bounded in $[0, 1]$
- the leading expression on the regret depends on the time T and the number of actions $|\mathcal{X}|$. We can ignore logarithmic factors as they grow more slowly than $\sqrt{T}$.

For the average regret to be less than ϵ we need:

$$\frac{\textsc{Regret(T)}}{T} \sim \frac{\sqrt{|\mathcal{X}|T}}{T} \sim \sqrt{\frac{|\mathcal{X}|}{T}} \sim \epsilon \tag{7}$$

which yields:

$$T \gtrsim \frac{|\mathcal{X}|}{\epsilon^2} \tag{8}$$

to achieve an average per-round regret less than ϵ.

2 Regret Bounds for Thompson Sampling

We now examine Thompson sampling and derive a Bayesian Regret bound of such algorithm. We begin with an important observation that relates Thompson sampling with UCB:

Lemma 1 (Thompson Sampling Lemma)**.** *Let $u_t(\cdot)$ denote some upper confidence bound function that is deterministic conditioned on the history up to time t. Let the actions x_t be selected according to Thompson sampling. Then,*

$$\mathbb{E}[u_t(x_t)] = \mathbb{E}[\mathbb{E}[u_t(x_t) \mid \mathcal{H}_t]] = \mathbb{E}[\mathbb{E}[u_t(x^*) \mid \mathcal{H}_t]] = \mathbb{E}[u_t(x^*)] \tag{9}$$

Proof. Conditioned on the observed samples (i.e., the filtration $\mathcal{H}_t$) Thompson sampling selects the action with the probability that it is optimal, i.e., conditioned on $\mathcal{H}_t$ we have that x_t and x^* are identically distributed. Further, since u_t is deterministic conditioned on $\mathcal{H}_t$, we have $\mathbb{E}[u_t(x_t) \mid \mathcal{H}_t] = \mathbb{E}[u_t(x^*) \mid \mathcal{H}_t]$ a.s. Thus, the lemma follows. □

Using the above result we can decompose the Bayesian regret of Thompson sampling:

$$\begin{aligned}
\textsc{BayesRegret} = \mathbb{E}[\overline{r}(x^*, \theta) - \overline{r}(x_t, \theta)] &= \mathbb{E}[\overline{r}(x^*, \theta) - u_t(x_t) + u_t(x_t) - \overline{r}(x_t, \theta)] &(10)\\
&= \mathbb{E}[\underbrace{\overline{r}(x^*, \theta) - u_t(x^*)}_{\text{Pessimism}} + \underbrace{u_t(x_t) - \overline{r}(x_t, \theta)}_{\text{Width}}] &(11)
\end{aligned}$$

Since $u_t(\cdot)$ is arbitrary, we can choose the same upper confidence bound that we choose for Ucb in section 1, immediately yielding a Bayesian regret bound of:

$$\textsc{BayesRegret} \leq 2\min\{|\mathcal{X}|, T\} + 4\sqrt{|\mathcal{X}|T(2+6\ln T)} \tag{12}$$

This notion is weaker than the frequentist bound of Ucb as an expectation is taken over the prior of θ.

3 Linear Bandit

The linear bandit is a generalization of the multi-armed bandit problem. It introduces features $\{x \in \mathcal{X}\}$ and assumes the model is linear, i.e., the agent observes:

$$r_t = \theta^\top x_t + w_t \tag{13}$$

where $\theta \sim P(d\theta)$ is the model parameter, possibly constrained in Θ_C, x_t is the action / feature selected and $w_t \sim \mathcal{N}(0, \sigma^2)$ is some noise that corrupts the observation. Upon rescaling the features and the distribution support $P(d\theta)$ we can assume $\|\theta\|_2 \leq 1$, $\|x\|_2 \leq 1$.

To tackle this problem, we define elliptic confidence intervals for the unknown parameter θ:

$$\Theta_k = \{\theta \mid (\theta - \widehat{\theta}_k)^\top \widehat{\Sigma}_k^{-1} (\theta - \widehat{\theta}_k) \leq \beta_k\} \tag{14}$$

where $\widehat{\theta}_k$ and $\widehat{\Sigma}_k$ are the maximum likelihood estimates for the empirical mean and the covariance matrix, respectively and β_k is a width parameter. The UCB-variant that efficiently learns this problem selects actions by solving:

$$x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \max_{\hat{\theta} \in \Theta_C \cap \Theta_t} \hat{\theta}^\top x \tag{15}$$

while Thompson sampling samples an instance of the true model parameter from the posterior:

$$\hat{\theta} \sim P(d\theta \mid \mathcal{H}_t) \tag{16}$$

and chooses the best action with respect to that model:

$$x_t = \operatorname*{argmax}_{x \in \mathcal{X}} \hat{\theta}^\top x. \tag{17}$$

If the set of actions is constrained, the optimization problem that UCB tries to solve may be NP-hard; for Thompson sampling, instead, it may be difficult to update the posterior distribution exactly.

One can show that under some conditions, UCB and TS achieve

$$\textsc{BayesRegret} = \tilde{O}(d\sqrt{T}) \tag{18}$$

where d is the dimension of the feature space. This implies that

$$\sim \frac{d^2}{\epsilon^2} \tag{19}$$

samples are need to achieve an average regret below ϵ.

4 Tabular RL

A classical algorithm for performing optimistic exploration in tabular RL is UCRL2. Here we mention a variant for episodic RL, which assumes bounded rewards $r(s,a) \in [0,1]$ for every state-action (s,a) pair. UCRL2 looks for an optimistic MDP by constructing confidence intervals around the empirical system dynamics $\hat{p}(s,a)$ and rewards $\hat{r}(s,a)$, precisely $\|p(s,a) - \hat{p}(s,a)\|_1 \leq \beta(n)$ and $|r(s,a) - \hat{r}(s,a)| \leq \beta(n)$ where $\beta(n) = \sqrt{\frac{28S\ln(SAHL)}{\max(1, n(s,a))}}$ is the confidence interval width. Here S, A, H, L represent the cardinality of the state and action space, the episode length and number of episodes, respectively, while $n(s,a)$ is the number of visits to the s,a pair. UCRL2 computes an optimist policy via extended value iteration and enjoys a high probability regret upper bound of $\tilde{O}(HS\sqrt{AT})$. Using these confidence sets, one can derive a similar regret bound for PSRL.