DDA4230 Reinforcement learning Thompson sampling

Lecture 8

Lecturer: Baoxiang Wang *Scribe: Baoxiang Wang*

1 Goal of this lecture

To understand algorithms based Thompson sampling (TS), in terms of the regret analysis and underlying the Bayesian perspective. For applications, students should also gain some intuition about different algorithms' advantages.

Suggested reading: Chapter 36 of *Bandit algorithms*; *A tutorial on Thompson sampling* by Russo, van Roy, Kazerouni, Osband, and Wen; *Further optimal regret bounds for Thompson sampling* by Agrawal and Goyal; *An information-theoretic analysis of Thompson sampling* by Russo and van Roy;

2 Recap: ε-greedy, ETC, and UCB

For ε-greedy, by choosing $\varepsilon_t = \min\{1, Ct^{-1}\Delta_{\min}^{-2}m\}$ for some absolute constant C, the regret satisfies

$$\overline{R}_T \le C' \sum_{i\ge 2} \left(\Delta_i + \frac{\Delta_i}{\Delta_{\min}^2} \log \max\left\{ e, \frac{T\Delta_{\min}^2}{m} \right\} \right), \tag{1}$$

where C' is an absolute constant.

For ETC under 2-armed bandits, when $T \ge 4\sqrt{2\pi e}/\Delta^2$, By choosing $k = \lceil \frac{2}{\Delta^2} W(\frac{T^2\Delta^4}{32\pi}) \rceil$, the regret satisfies

$$\overline{R}_T \le O(\frac{1}{\Delta}\log T\Delta^2) + o(\log T) + \Delta, \tag{2}$$

where $W(y)\exp(W(y)) = y$ denotes the Lambert function.

For UCB, by setting $\delta = T^{-2}$, the regret satisfies

$$\overline{R}_T \le 3\sum_{i=1}^{m} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log T}{\Delta_i}.$$

This result is followed by a series of improvements.

3 Recap: Bayesian statistics and Bernoulli-Beta conjugate

Recall that the reward $r(i)$ of arm i follows some distribution. Assume that the reward distributions of arms belong to the same family with respective parameters, which writes

$$r(i) \sim p(x \mid \theta_i)\,.$$

Recall that when estimating θ, the posterior is

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int_{\theta'} p(x \mid \theta')p(\theta')d\theta'}\,.$$

In Bayesian probability theory, if the posterior distributions $p(\theta \mid x)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(x \mid \theta)$. Some infamous conjugate priors are Gaussian-Gaussian, Bernoulli-Beta, Poisson-Gamma, categorical-Dirichlet. Conjugate priors are convenient in analyses.

The Bernoulli-Beta is important for Thompson sampling for Bernoulli bandits. Recall that the Beta distribution $\mathrm{Beta}(\alpha, \beta)$ with parameter $\theta = \{\alpha, \beta\}$ follows the probability density function of

$$p(x) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,,$$

where $\Gamma(z) = \int_0^\infty x^{z-1}\exp(-x)dx$, $z \in \mathbb{C}$ is the Gamma function. When $p(\theta) \sim \mathrm{Beta}(\alpha_0, \beta_0)$ and we observe $x_1, \ldots, x_{\alpha'+\beta'} \sim x$ i.i.d. with α' ones and β' zeros, then

$$\begin{aligned} p(\theta \mid x_1, \ldots, x_{\alpha'+\beta'}) &= \frac{p(x_1, \ldots, x_{\alpha'+\beta'} \mid \theta)p(\theta)}{\int_{\theta'} p(x_1, \ldots, x_{\alpha'+\beta'} \mid \theta')p(\theta')d\theta'} \\ &= \frac{\binom{\alpha'+\beta'}{\alpha'}\theta^{\alpha'}(1-\theta)^{\beta'}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}}{\int_{\theta'} p(x_1, \ldots, x_{\alpha'+\beta'} \mid \theta')p(\theta')d\theta'} \\ &= \frac{\binom{\alpha'+\beta'}{\alpha'}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}{\int_{\theta'} p(x_1, \ldots, x_{\alpha'+\beta'} \mid \theta')p(\theta')d\theta'}\theta^{\alpha_0+\alpha'-1}(1-\theta)^{\beta_0+\alpha'-1} \\ &\sim \mathrm{Beta}(\alpha_0+\alpha', \beta_0+\alpha')\,. \end{aligned}$$

This implies that if our current belief of $\mu_i = \theta$ is $\mathrm{Beta}(\alpha, \beta)$ and we observe a new data $x \sim \mathrm{Ber}(\theta)$, then we update $\alpha \mathrel{+}= 1$ or $\beta \mathrel{+}= 1$ when $x = 1$ or $x = 0$, respectively.

4 The Thompson sampling algorithms

4.1 The first algorithm

We return to where it all began, in bandits, to the first algorithm proposed by Thompson in 1933. The idea is a simple one. Before the game starts, the learner chooses a prior over a set of possible bandit environments. In each round, the learner samples an environment from the posterior and acts according to the optimal action in that environment.

The exploration in Thompson sampling comes from the randomisation. If the posterior is poorly concentrated, then the fluctuations in the samples are expected to be large and the policy will likely explore. On the other hand, as more data is collected, the posterior concentrates towards the true environment and the rate of exploration decreases. We discuss finite-armed stochastic bandits, but Thompson sampling has been extended to all kinds of models (see Chapter 36 of the book).

Randomisation is crucial for adversarial bandit algorithms and can be useful in stochastic settings (see Chapters 23 and 32 of the book for examples). We should be wary, however, that injecting noise into our algorithms might come at a cost in terms of variance. What is gained or lost by the randomisation in Thompson sampling is still not clear, but we leave this cautionary note as a suggestion to the reader to think about some of the costs and benefits.

Algorithm 1: Thompson sampling (Bernoulli bandits)

```
Input: Prior $\alpha_0$, $\beta_0$
Output: $a_t, t \in [T]$
Initialize $\alpha_i = \alpha_0$, $\beta_i = \beta_0$, for $i \in [m]$
while $t \leq T - 1$ do
    Sample $\theta_i(t) \sim \mathrm{Beta}(\alpha_i, \beta_i)$ independently for $i \in [m]$
    $a_t = \arg\max_{i \in [m]} \theta_i(t)$ with arbitrary tiebreaker
    If $r_t = 1$ then $\alpha_{a_t}$ += 1; If $r_t = 0$ then $\beta_{a_t}$ += 1;
```

A general TS algorithm works on any conjugate priors. When the family of the underlying reward distribution is unknown, a Gaussian-Gaussian conjugate (the non-informative prior) can be useful.

Algorithm 2: Thompson sampling

```
Input: Prior $\theta_0$
Output: $a_t, t \in [T]$
Initialize $\theta_i = \theta_0$, for $i \in [m]$
while $t \leq T - 1$ do
    Sample independently for $i \in [m]$
        $\theta_i(t) \sim p(\theta \mid \{r_{t'}\}_{\mathbb{1}\{a_{t'}=i, t' \leq t-1\}})$
    $a_t = \arg\max_{i \in [m]} \theta_i(t)$ with arbitrary tiebreaker
    Update $\theta_{a_t}(t+1)$ by
        $p(\theta_{a_t}(t+1) \mid \{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}}) = \frac{p(\{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}} \mid \theta) p(\theta)}{\int_{\theta'} p(\{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}} \mid \theta') p(\theta') d\theta'}$
```

4.2 Analysis of Thompson sampling

Theorem 1 *Assume the rewards of arms are μ_i-Bernoulli. The regret under TS (Bernoulli bandits) is at most*

$$\overline{R}_T \leq \sum_{i:\Delta_i > 0} \frac{\mu_1 - \mu_i}{d_{KL}(\mu_1 \parallel \mu_i)} \log T + o(\log T)\,,$$

where the Kullback-Leibler divergence

$$d_{KL}(\mu_1 \parallel \mu_i) = \mu_1 \log(\frac{\mu_1}{\mu_i}) + (1-\mu_1)\log(\frac{1-\mu_1}{1-\mu_i})\,.$$

As is similar to ETC and UCB, instance-independent regret bound of $O(\sqrt{mT\log T})$ can be obtained.

The proof of the regret bound can be obtained by either using probability or using techniques in information theory. We refer the proofs to the papers listed in suggested reading.

4.3 Advantages of different bandit algorithms

TS admits good practical performance in general and is one of the first-to-try algorithms in bandits. This extends to problems in machine learning like feature selection, automatic machine learning, and active learning, etc.

TS is effective across a broad range of problems, but there are contexts in which TS leaves a lot of value on the table, including

1. Problems that do not require exploration or problems where deterministic actions incur sufficient exploration (consider that TS is a randomized algorithm which introduces additional uncertainty);

2. Problems that do not require exploitation, like pure exploration problems (e.g. Bayesian optimization);

3. Time-sensitive problems with small T;

4. Problems requiring careful assessment of information gain, where TS is not choosing the action that maximizes the gain (see 8.2.4 of the tutorial for more examples). This also connects to the fact that TS is not converging as fast as pure exploration methods.

4.4 Applications

Students should implement ε-greedy, ETC, UCB, and TS and try them on synthetic and real datasets to gain some intuition about their behavior.

Acknowledgement

This lecture notes partially use material from *Bandit algorithms*.