Lecture 7 - Thompson sampling

Guiliang Liu

The Chinese University of Hong Kong, Shenzhen

DDA4230: Reinforcement Learning
Course Page: [Click]

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

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

Conjugate distributions: The posterior distributions $p(\theta \mid x)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$.

Recap: Bayesian statistics and Bernoulli-Beta conjugate

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

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

where $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) dx$, $z \in \mathbb{C}$ is the Gamma function.

Recap: Bayesian statistics and Bernoulli-Beta conjugate

When $p(\theta) \sim \text{Beta}(\alpha_0, \beta_0)$ and we observe $x_1, \ldots, x_{\alpha'+\beta'} \sim x$ i.i.d. with α' ones and β' zeros (observe a new data $x \sim \text{Ber}(\theta)$), then the posterior should follow:

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

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Thompson sampling algorithms

- Before the game starts, the learner sets a prior over 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 randomization, i.e., whether the posterior concentrates or not.

Algorithm 1: Thompson sampling (Bernoulli bandits)

Input: Prior α_0, β_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 $\alpha_{a_t} \mathrel{+}= 1$; If $r_t = 0$ then $\beta_{a_t} \mathrel{+}= 1$;

Thompson sampling algorithms

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 θ_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 the posterior probability distribution of $\theta_{a_t}(t+1)$ by

$$p(\theta_{a_t}(t+1) \mid \{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}}) = \frac{p(\{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}} \mid \theta)p(\theta)}{\int_{\theta'} p(\{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}} \mid \theta')p(\theta')d\theta'}$$

（深 圳）

of Hong Kong, Shenzhen

The Regret of Thompson sampling Algorithms

Theorem

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

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

where the Kullback-Leibler divergence:

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

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Question and Answering (Q&A)