

# 1 Intro

## 2 Model

To keep it simple, we will start with an integer-valued reward that increases by 1 at each round in which a reward is not observed. So the outcome  $y \in \mathcal{Y}$  conditioned on the action  $a \in \mathcal{A}$  is distributed  $y|a \sim \text{geometric}(p_a)$  for some unobserved  $p_a \in [0, 1]$ . We would like to find  $a^* = \arg \max_{a \in \mathcal{A}} \mathbb{E}[r(y)|a]$ , where we start by considering  $r(y) = y$  for simplicity. In the Bayesian scenario, we need to put a prior on each  $p_a$ , and we choose  $p_a \sim f_{p_a}(p_a) = \text{beta}(\alpha_a, \beta_a)$  prior for conjugacy.

## 3 Notation

We denote the full history for  $N$  rounds by  $\mathcal{Z}^{(N)} = \{(a_i, y_i)\}_{i=1}^N$ , where  $a_i$  is the chosen arm pull and  $y_i$  is the outcome as well as the length of the delay before the outcome is revealed to us. We denote the subset of rounds at which arm  $a$  was pulled by  $I_a^{(N)} = \{i : i \in [N], a_i = a\}$ , and we denote the subset of rounds at which arm  $a$  was pulled and an outcome has yet to be observed as  $U_a^{(N)} = \{i : i \in I_a^{(N)}, i + y_i > N\}$ .

## 4 Algorithm

We want to derive a Thompson sampling algorithm for the model described in the prequel. To get the algorithm, we need to derive the posterior distribution for this model and look at it as a function of some sufficient statistic that we can calculate from the rewards we observe or do not observe at each round.

We must be cautious when considering the contribution to the likelihood of arm pulls for which rewards have not been observed. Suppose that, on the first round, we have pulled arm  $a$ , and we do not immediately receive a reward. Then, after that round, the contribution of the first arm pull to the likelihood of arm  $a$  is the probability that the reward will manifest after the first round,  $P(y > 1|a) = 1 - P(y \leq 1|a)$ , i.e. one minus the cdf of  $\text{geometric}(p_a)$ . For a pull of arm  $a$  at round  $j$  with reward still unobserved after round  $k > j$ , the contribution is  $P(y > k - j|a)$ . Overall, the posterior at round  $N$  is given by

$$P(p_a|U_a^{(N)}, I_a^{(N)}, \mathcal{Z}_a^{(N)}) = \frac{L(U_a^{(N)}, I_a^{(N)}, \mathcal{Z}_a^{(N)}|p_a) \times f_{p_a}(p_a)}{\int_{[0,1]} L(U_a^{(N)}, I_a^{(N)}, \mathcal{Z}_a^{(N)}|p_a) \times f_{p_a}(p_a) dp_a}.$$

where  $L(\mathcal{Z}_a|p_a)$  is the geometric likelihood of the sequence of observed and unobserved rewards resulting from pulls to arm  $a$ , and  $f_{p_a}(p_a) = \text{beta}(\alpha_a, \beta_a)$  is the prior distribution on  $p_a$ . This likelihood has the form

$$\begin{aligned} L(U_a^{(N)}, I_a^{(N)}, \mathcal{Z}_a^{(N)}|p_a) &= \left( \prod_{i \in I_a \setminus U_a} (1 - p_a)^{y_i - 1} p_a \right) \left( \prod_{i \in U_a} (1 - p_a)^{N - i} \right) \\ &= \left( (1 - p_a)^{\left( \sum_{i \in I_a^{(N)} \setminus U_a^{(N)}} y_i \right) - |I_a^{(N)} \setminus U_a^{(N)}|} p_a^{|I_a^{(N)} \setminus U_a^{(N)}|} \right) \left( (1 - p_a)^{|U_a| \times N - \sum_{i \in U_a} i} \right) \\ &= (1 - p_a)^{|U_a^{(N)}| \times N + \left( \sum_{i \in I_a^{(N)} \setminus U_a^{(N)}} y_i \right) - \left( \sum_{j \in U_a^{(N)}} j \right) - |I_a^{(N)} \setminus U_a^{(N)}|} p_a^{|I_a^{(N)} \setminus U_a^{(N)}|}, \end{aligned}$$

and the prior is a  $\text{beta}(\alpha_a, \beta_a)$  distribution, which has the form

$$f_{p_a}(p_a) = \frac{\Gamma(\alpha_a) \Gamma(\beta_a)}{\Gamma(\alpha_a + \beta_a)} p_a^{\alpha_a - 1} (1 - p_a)^{\beta_a - 1}.$$

Then, removing the constant multiplicative factors to examine the kernel of the posterior

$$P(p_a|U_a^{(N)}, I_a^{(N)}, \mathcal{Z}^{(N)}) \propto (1-p_a)^{|U_a^{(N)}| \times N + \left(\sum_{i \in I_a^{(N)} \setminus U_a^{(N)}} y_i\right) - \left(\sum_{j \in U_a^{(N)}} j\right) - |I_a^{(N)} \setminus U_a^{(N)}|} p_a^{|I_a^{(N)} \setminus U_a^{(N)}| + \alpha_a - 1},$$

we find that this is the kernel of a *beta*  $\left(\alpha_a^{(N)}, \beta_a^{(N)}\right)$ , where

$$\begin{aligned}\alpha_a^{(N)} &= |I_a^{(N)} \setminus U_a^{(N)}| + \alpha_a \\ \beta_a^{(N)} &= |U_a^{(N)}| \times N + \left(\sum_{i \in I_a^{(N)} \setminus U_a^{(N)}} y_i\right) - \left(\sum_{j \in U_a^{(N)}} j\right) - |I_a^{(N)} \setminus U_a^{(N)}| + \beta_a.\end{aligned}$$

From this observation, we can conclude that, since we are integrating over the domain of the *beta* distribution, the integral must evaluate to the reciprocal of  $\frac{\Gamma(\alpha_a^{(N)})\Gamma(\beta_a^{(N)})}{\Gamma(\alpha_a^{(N)} + \beta_a^{(N)})}$ . After some further cancelations and reciprocations, we get that the posterior at time  $N$  has the form *beta*  $\left(\alpha_a^{(N)}, \beta_a^{(N)}\right)$ , which gives the parameter update for the posterior from which our Thompson sampling scheme takes samples.