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 $y|a \sim geometric(p_a)$ 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 $p_a \sim f_{p_a}(p_a) = beta(\alpha_a, \beta_a)$ prior for conjugacy.

3 Notation

We denote the full history for N rounds by $\mathcal{Z} = \{(a_i, y_i)\}_{i=1}^N$, 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 subsequence of rounds at which arm a was pulled by $I_a = \{i\}_{i=1,\dots,N:a_i=a} = \left\{i_j^{(a)}\right\}_{j=1}^{|I_a|}$ (this notation could use some work).

4 Algorithm

To get an algorithm, we need to derive the posterior distribution for the Bayesian model described in the prequel 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 $geometric(p_a)$. For a pull of arm a at round j with reward still unobserved after round $k > j$, the contribution is $P(y > k - j|a)$.

For a sequence of N arm pulls on arm a, $\mathcal{Z}_a = \{(a_i, y_i) : a_i = a\}_{i=1}^N$ where a_i is the arm that is pulled at time i, and y_i is the resulting reward, and the length of the delay before the reward is revealed, we denote $U = \{(a_i, y_i) : y_i + i > N, i \in [N]\} \subset \mathcal{Z}$, the set of arm pulls for which a reward have not been observed. Overall, the posterior at round N is given by

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

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 $f_{p_a}(p_a) = beta(\alpha_a, \beta_a)$ is the prior distribution on p_a. This likelihood has the form

$$L(\mathcal{Z}_a|p_a) \;=\; \left(\prod_{(a',y)\in\mathcal{Z}\setminus U} (1-p_a)^y p\right)\left(\prod_{(a',y)\in U} \left(1-(1-p_a)^{y+1}\right)\right)$$