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 $a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[r(y)|a\right]$, 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 $p_a \sim beta(\alpha_a, \beta_a)$ prior for conjugacy.

3 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 $P(y > 1|a) = 1 - P(y \leq 1|a)$ that the reward will manifest after the first round.