Fitting Thompson Sampling to Behavior

Data

$\mathbf{x} = \{x_1, x_2, \cdots, x_T\}$
$\mathbf{r} = \{r_1, r_2, \cdots, r_T\}$

Initialize

$\alpha^{(0)} = \{\alpha_1^{(0)}, \cdots, \alpha_K^{(0)}\}$
$\beta^{(0)} = \{\beta_1^{(0)}, \cdots, \beta_K^{(0)}\}$

Environment + Hyperparameters

$K :=$ # of arms
$N :=$ # of posterior samples
$\gamma_\alpha := \alpha$ learning rate
$\gamma_\beta := \beta$ learning rate

Model Fitting Procedure

Thompson Sampling

For $t \in [1, T]$:

1. Estimate $p(x_t|\alpha^{(t-1)}, \beta^{(t-1)})$
   (a) For $n \in [1, N]$:
      i. For $k \in [1, K]$:
         A. sample $b_k^{(n)} = \text{Beta}(\alpha_k^{(t-1)}, \beta_k^{(t-1)})$
      ii. $\hat{x}_t^{(n)} = \arg\max_{k \in [1,K]} b_k^{(n)}$
   (b) $p(x_t|\alpha^{(t-1)}, \beta^{(t-1)}) \approx \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(\hat{x}_t^{(n)} = x_t)$
2. Append the loss $\mathcal{L}_t = -\log p(x_t|\alpha^{(t-1)}, \beta^{(t-1)})$
3. Update parameters of the beta distribution:
   (a) $\alpha_{x_t}^{(t)} \leftarrow \alpha_{x_t}^{(t-1)} + \gamma_\alpha r_t$
   (b) $\beta_{x_t}^{(t)} \leftarrow \beta_{x_t}^{(t-1)} + \gamma_\beta (1 - r_t)$