Exploration-Exploitation-Engagement in Multi-Armed Bandits with Abandonment

Zixian Yang

EECS, University of Michigan, Ann Arbor

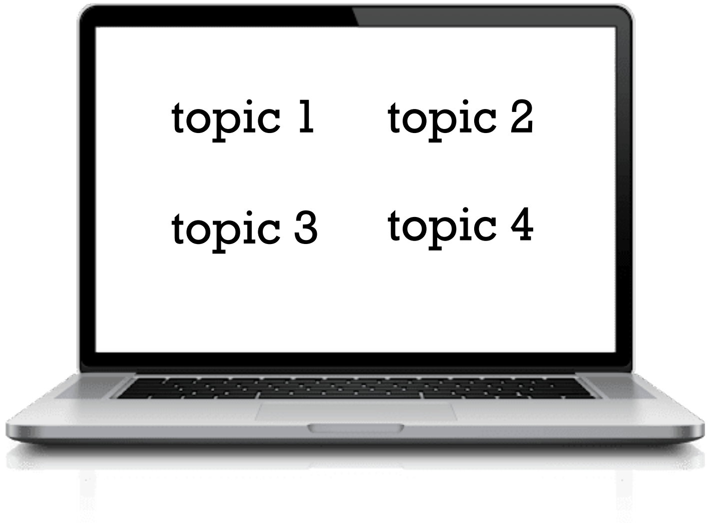
zixian@umich.edu

Joint work with Prof. Xin Liu (Shanghai Tech) and my advisor Prof. Lei Ying (Michigan)

Motivation and Applications

Short video
recommendations

Content recommendations
in online education

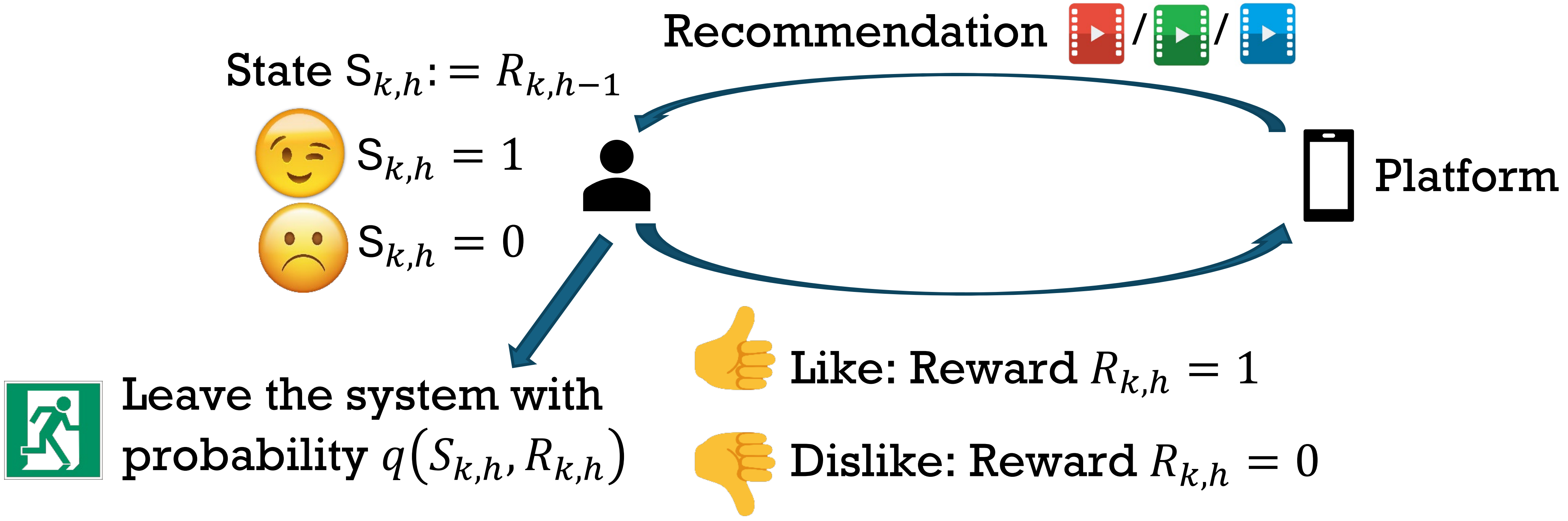


Model: Multi-Armed Bandit (MAB)

The MAB model overlooks **user abandonment**.

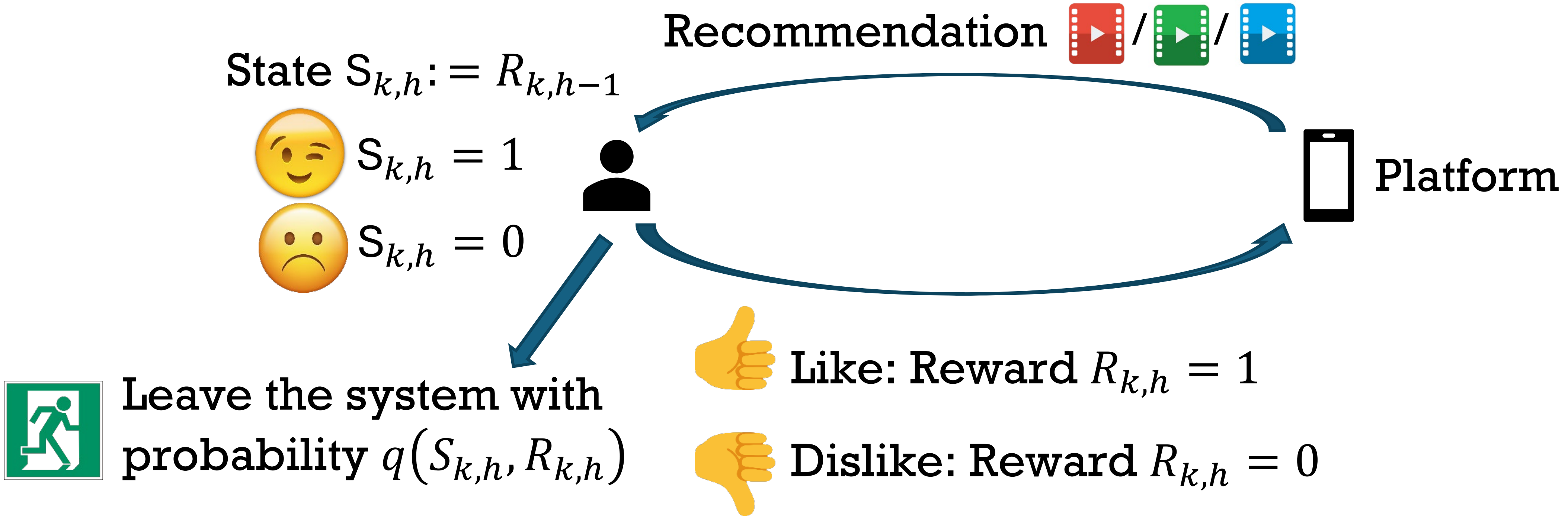
Exploration-Exploitation-Engagement: A Simple Model

- M arms $\{a_1, a_2, \cdots, a_M\}$
- Consider K episodes. State at step h of the kth episode is $S_{k,h} \in \{0, 1\}$
- Bernoulli rewards with mean $\mu(a_i)$



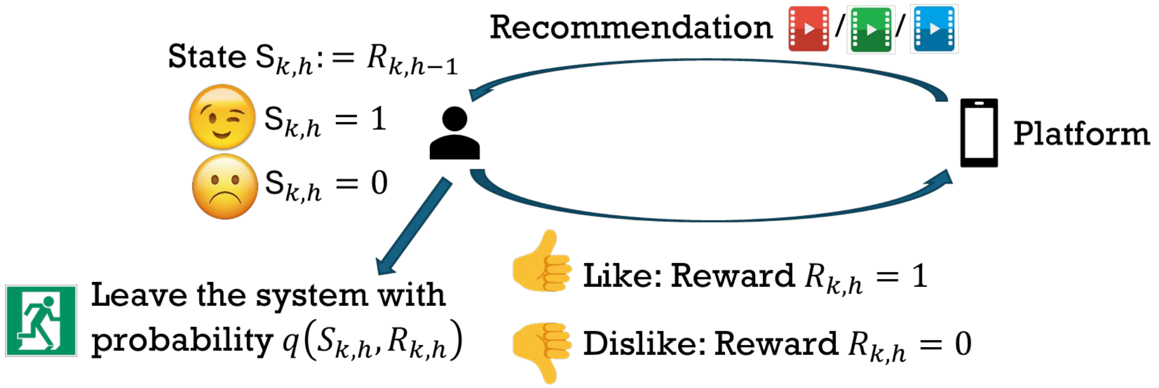
Assumption

- The user is less likely to abandon the system when getting higher reward $q(S_{k,h}=s, R_{k,h}=r) \leq q(S_{k,h}=s', R_{k,h}=r')$ if $s+r > s'+r'$.

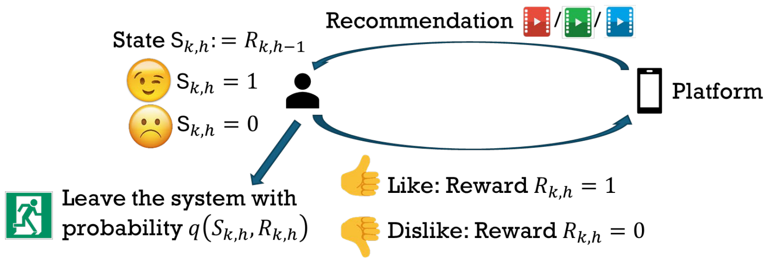


Problem Definition

- Baseline π^* : A genie-aided, optimal policy is always pulling the arm with the highest mean
- Regret for a given policy: the difference between the expected total reward achieved by the optimal policy, π^*, and that achieved by the given policy.
- Goal: minimize the regret



Upper and Lower Confidence Bound (ULCB)



- ULCB --- state-dependent bonus/penalty terms
  - When state is 0, $\tilde{\mu}_t(a) = \bar{\mu}_t(a) - \sqrt{\dfrac{\log t + 4 \log\log t}{2N_t(a)}}$ discourage exploration

    sample mean — number of times for which arm a has been pulled

  - When state is 1, $\tilde{\mu}_t(a) = \bar{\mu}_t(a) + \sqrt{\dfrac{\log t + 4 \log\log t}{2N_t(a)}}$ encourage exploration
  - Choose arm $a \in \text{argmax}_a\ \tilde{\mu}_t(a)$
- KL-ULCB --- use KL divergence instead of Euclidean distance.

Main Results

- Theoretically, KL-ULCB is asymptotically optimal. (number of episodes $K \to \infty$)
- Empirically, KL-ULCB performs significantly better than other algorithms.