Adapting multi-armed bandits to real-life: Flexible models and approximate inference

**In many problems in science, engineering and medicine, one observes the world and must sequentially:**
- Decide which action to take next,
- Based on previous interactions with the world,
- In order to maximize future returns.

**Randomized Controlled Trial Vs Multi-Armed Bandits**
- Different potential actions to take (arms)
- Stochastic rewards: e.g., success/failure
- After observing previous actions and rewards
  - If parameters are known, pick optimal action
  - If parameters are unknown, exploration-exploitation tradeoff

**Thompson sampling (TS) in practice:**
- Pick (randomly) best arm, according to learned model
- Bayesian parametric modeling of the world
- Update model based on observed actions and rewards
- Draw a sample parameter from updated model
- Pick the optimal arm for such sample (“believe”)

**We propose 3 novel improvements**

1. **Dynamic-categorical rewards**
   - Models beyond stationary and Bernoulli rewards needed
   - User ignores the recommended movie, clicks on the trailer, or watches the movie
   - Users’ preferences evolve over time
   - We propose:
     - Categorical rewards via the softmax function
     - Dynamics via a general linear model on parameters
     - Sequential Monte Carlo combined with TS
       - Approximations to posterior accurate enough
       - Attain competitive regret performance

2. **Continuous-context dependent rewards**
   - State of the art:
     - TS for context-dependent continuous rewards
     - Based on linear-Gaussians distributions
   - We propose:
     - TS for complex scenarios, with unknown distributions
     - Nonparametric Gaussian mixture reward models:
       - A Bayesian generative process
       - Naturally aligned with the multi-armed bandit setting
       - It accommodates a very flexible set of distributions
     - Implementation of an efficient and flexible TS:
       - The nonparametric model autonomously determines its complexity in an online fashion,
       - As new rewards are observed for the played arms.
       - MCMC based inference via a Gibbs sampler

3. **Sequentially observed rewards**
   - In practice, a learning agent can only rely on
     - Partially observed sequences of rewards, e.g.,
       - after a movie is recommended,
       - the user ignores it or clicks on the trailer,
       - but the end-goal is whether she watches it.
   - We propose:
     - A Bayesian generative model for TS
     - Rewards are observed at different scales
       - Observations at scales \( s = \{1, \ldots, S\} \)
     - We consider a sequential and causal dependency
     - The reward at scale \( S \) is the reward to maximize:
       - How to maximize final reward,
       - as partial sequential observations are acquired

**Approximate Bayesian Inference symposium @ NIPS2018**

**Bayesian nonparametrics workshop @ NIPS2018**
https://arxiv.org/abs/1808.02932

**RL under partial observability workshop @ NIPS2018**