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

Foundations of Data Science
DATA SCIENCE INSTITUTE

Iñigo Urteaga and Chris H. Wiggins
Applied Physics and Applied Mathematics
Fu Foundation School of Engineering, Columbia University

Data Science Institute
COLUMBIA UNIVERSITY

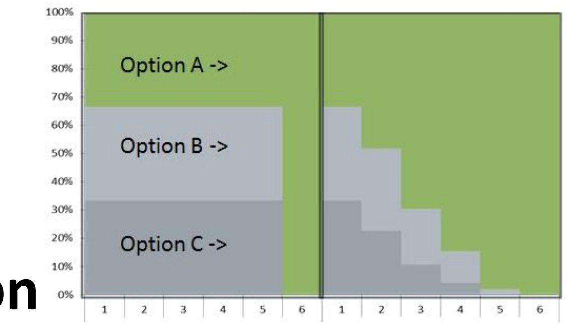
Sequential Decision Making

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 (1933)

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”)

Thompson sampling (TS) in practice:

Good: easy to implement, generalizable to include context and continuous arms

Bad: simple models of the word assumed, for computational and theoretical reasons

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

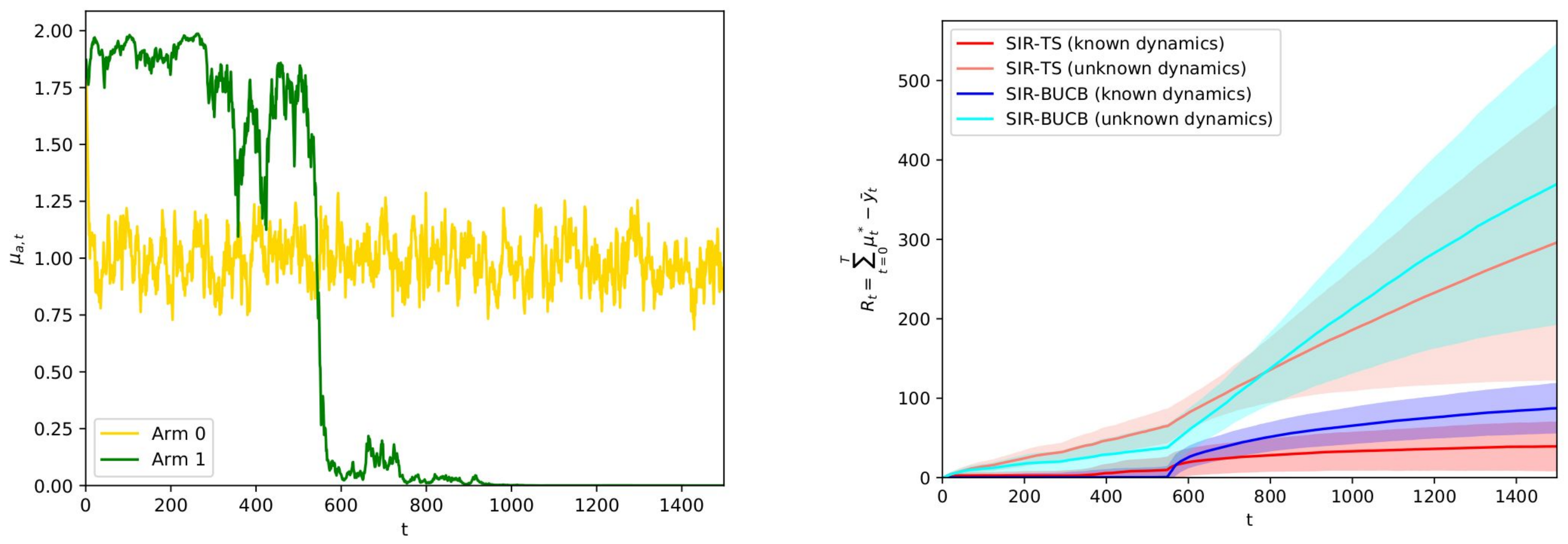


Figure 1. Dynamic 3-categorical rewards for2-armed bandit.

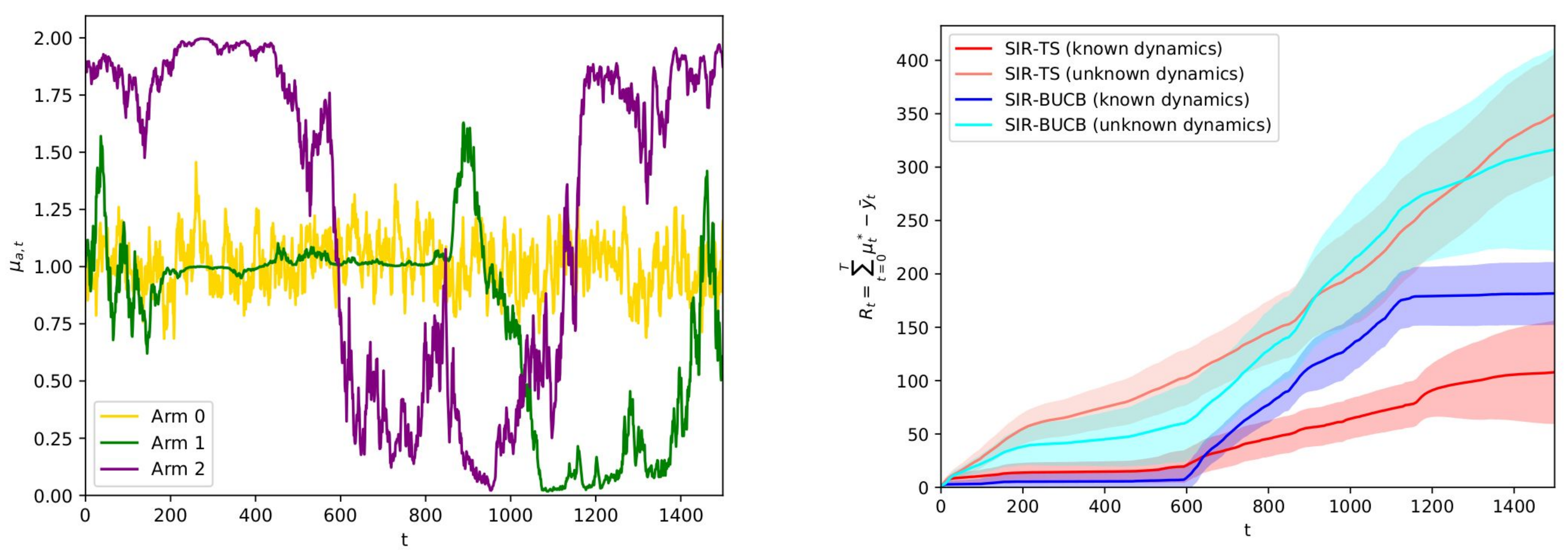


Figure 2. Dynamic 3-categorical rewards for 3-armed bandit.

Approximate Bayesian Inference symposium @ NIPS2018

https://arxiv.org/abs/1808.02933

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 accomodates 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

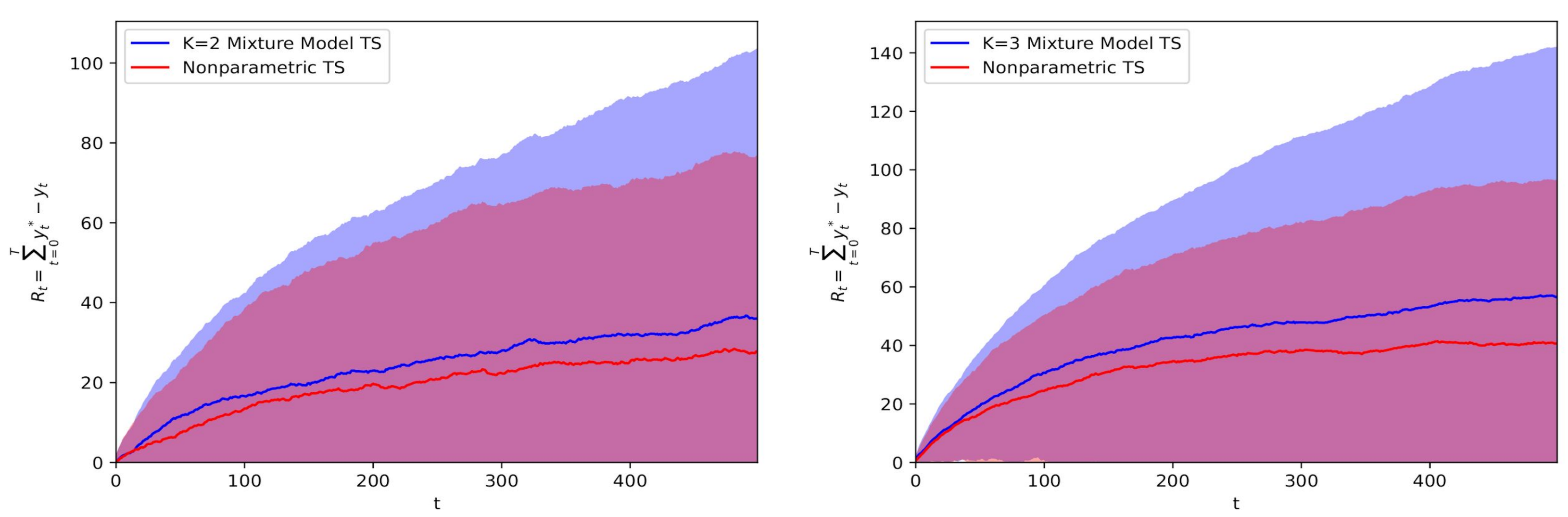


Figure 3. Performance in simulated mixture-model dataset.

		May 4th		May 5th	
		CTR	Normalized CTR	CTR	Normalized CTR
Model	Logistic	0.0451 +/- 0.0068	1.0855 +/- 0.1794	0.0462 +/- 0.0054	1.0472 +/- 0.1486
	Nonparametric mixture model	0.0474 +/- 0.0044	1.1413 +/- 0.1381	0.0483 +/- 0.0038	1.0932 +/- 0.1098

Figure 4. Performance of approach in the Yahoo dataset.

Bayesian nonparametrics workshop @ NIPS2018

https://arxiv.org/abs/1808.02932

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, . . . , 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

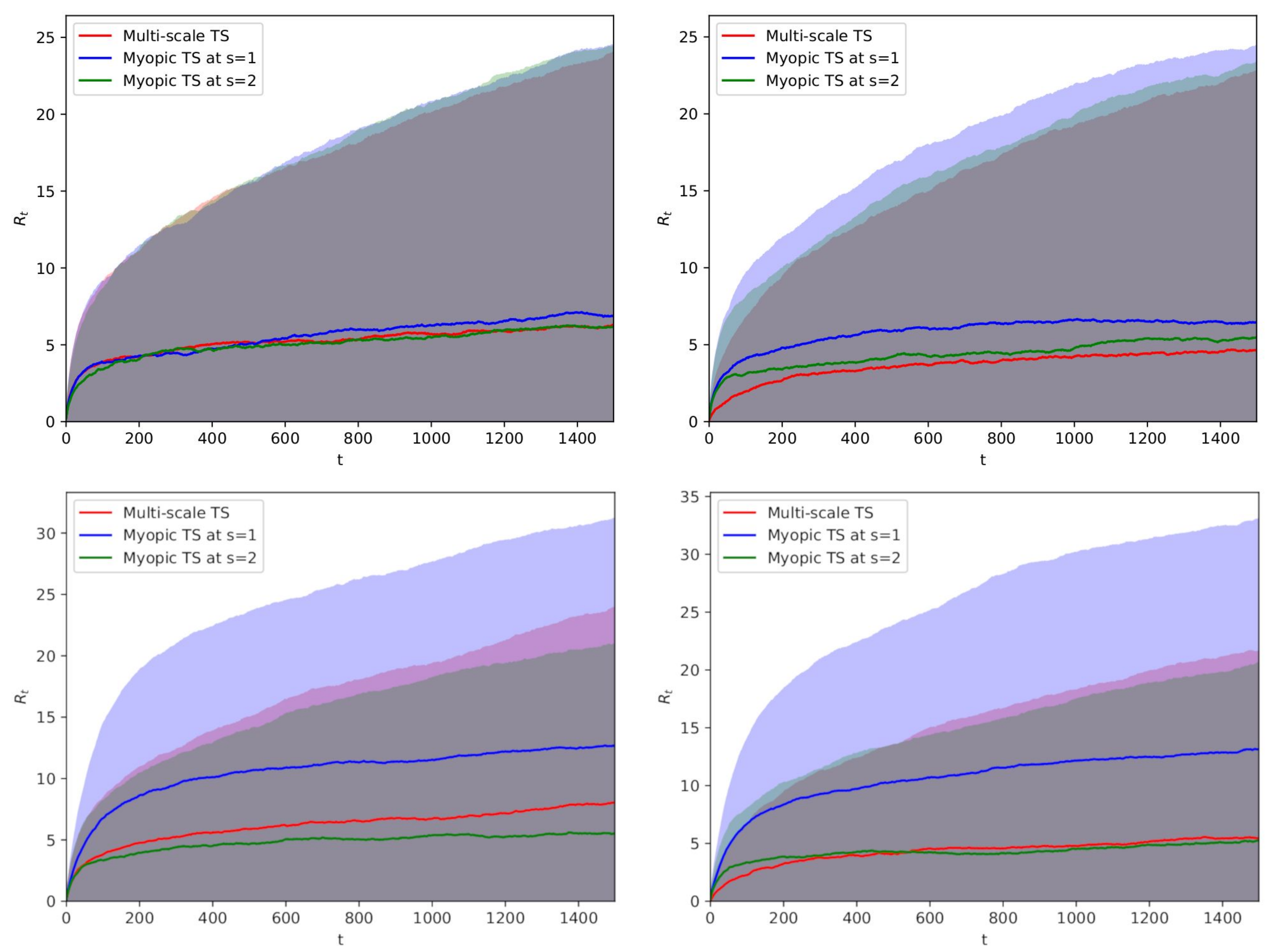


Figure 5. Performance on two-scale Bernoulli 2-armed bandits.

RL under partial observability workshop @ NIPS2018