# Project Proposal

## CSC 696H - Advanced Topics in Probabilistic Graphical Models

Yang Hong
hong1@arizona.edu

University of Arizona

October 12, 2022

# Problem

- Question: Is there a better way to do posterior sampling than existing Thompson sampling approach?

- Problem: Extend general Thompson sampling approach to the Multi-armed Bernoulli Bandit setting (constrained)

# Motivation

- Reinforcement learning is an active research area
- Used to model wide range and variety of real world problems
- Scalability of Thompson sampling
- Balancing the Exploration-Exploitation tradeoff

# Bandit problem

- One-armed bandit
- A set of arms/actions $M := \{m_1, m_2, \ldots, m_n\}$
- A finite number of tries/rounds/money T
- Dist. of the payouts corresp. to $m_i$ is unknown
- Goal is to maximize the cumulative reward $r : M \to \mathbb{R}$
  - Minimize the total regret $\mathcal{R}(T)$
  - Tradeoff of exploration-exploitation


# Multi-armed Bandit problem

- A set of arms $M := \{m_1, m_2, \ldots, m_n\}$
- T rounds
- On choosing $m_t$ , the reward $r(m_t)$ is observed
- $t^*$ := optimal choice for round t
- $$\mathscr{R}(T) = \sum_{t=1}^{T} r(m_{t^*}) - \sum_{t=1}^{T} r(m_t)$$

# More variants

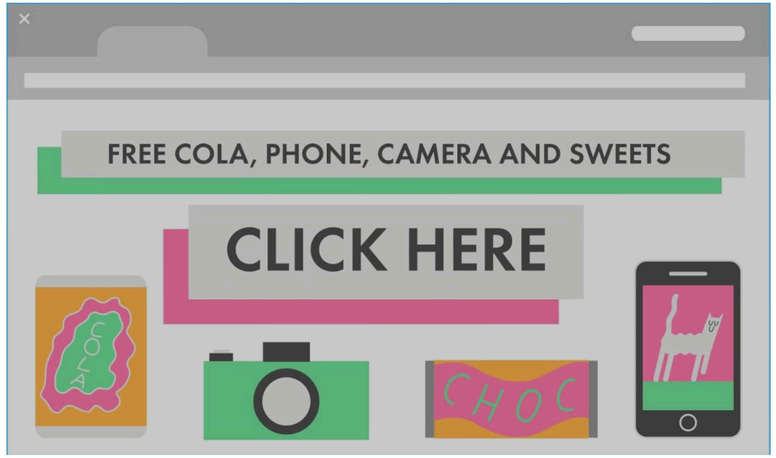
- Contextual Multi-armed Bandit
  - A context set $\mathcal{X}$ and a set of policies $\Pi$
  - Choose $m_t$ upon seeing a context/feature vector $x$
- Contextual Semi-Bandit
  - Restricted bandit feedback
  - A set of costs $C := \{c_1, c_2, \ldots, c_n\}$
  - Total/per round budget $B$

# And more…

- Stochastic bandit - reward is drawn i.i.d from an unknown distribution
- Cascade model
- Position-based model
- Low-rank bandit
  - Rank-1 bandit
- …

# Real world applications

- Internet advertising
  - Bernoulli model: a success/failure$(1/0)$ is associated with a click on the ad
  - Position-based model: item-position matrix
    - attraction and examination are i.i.d Bernoulli random variables
- Stock investment
  - Each day, you have a finite budget to invest in a few stocks
- Medical trials
  - You have many drugs to choose from for a health problem
- Dynamic pricing - Amazon
- Article/song/movie recommendation system
- Trip planning/Online shortest path, game theory, wireless networks…




# Proposed Approach

- Posterior density approximation by
  - Variational Inference
  - Expectation Propagation
  - Ensemble sampling

# Thompson sampling

- … is not a posterior sampling approach
- An algorithm
  - Tractable -> exact Bayesian inference
  - In practical
    - models more complex
    - Posterior intractable
    - -> posterior approximation methods
  - MH-TS, Gibbs-TS, Laplace-TS, Hamiltonian-TS, PG-TS…

# Greedy vs. Thompson sampling

**Algorithm 3** Greedy$(\mathcal{X}, p, q, r)$

**for** $t = 1, 2, \ldots$ **do**
  #estimate model:
  $\hat{\theta} \leftarrow \mathbb{E}_p[\theta]$

  #select and apply action:
  $x_t \leftarrow \text{argmax}_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x]$
  Apply $x_t$ and observe $y_t$

  #update distribution:
  $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot | x_t, y_t)$
**end for**

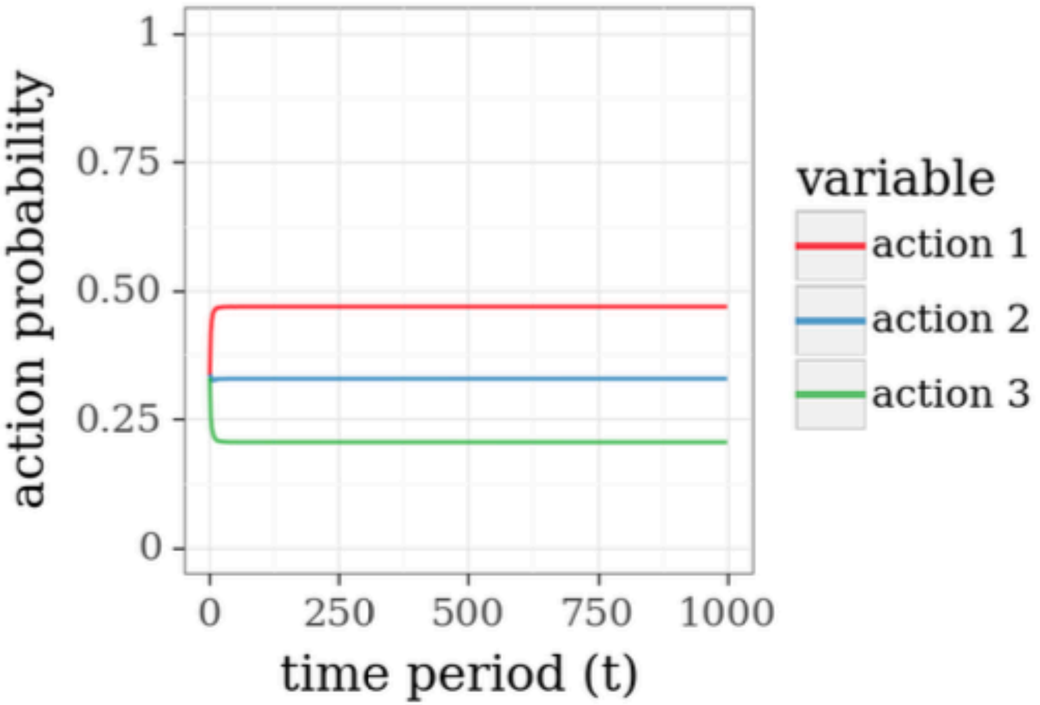
**Algorithm 4** Thompson$(\mathcal{X}, p, q, r)$

**for** $t = 1, 2, \ldots$ **do**
  #sample model:
  Sample $\hat{\theta} \sim p$

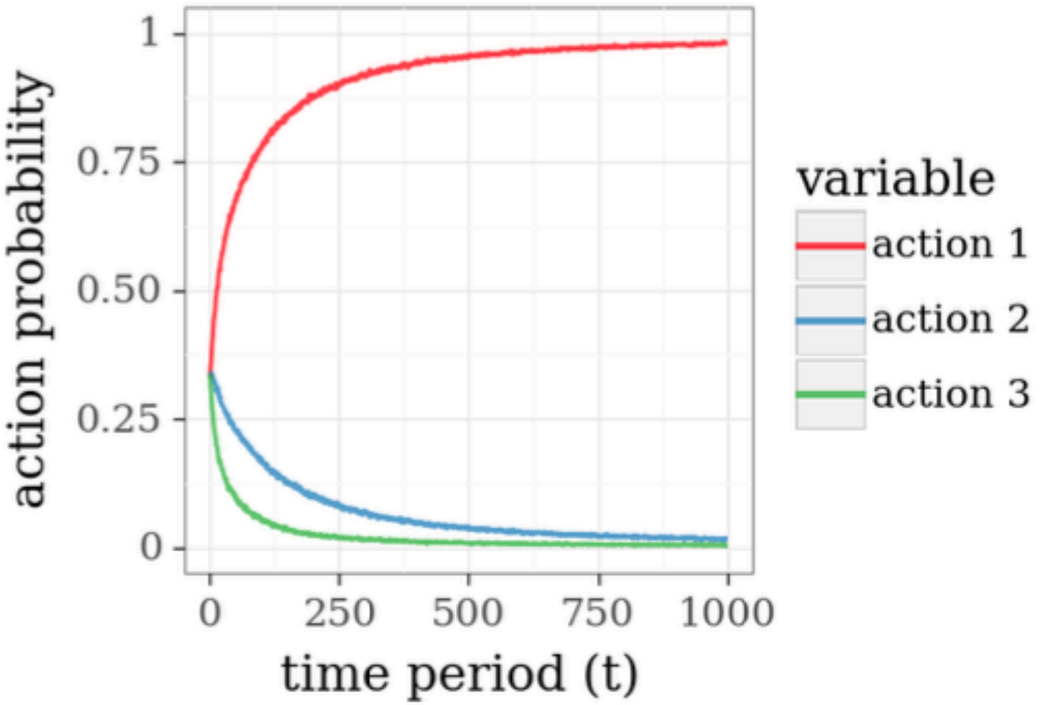
  #select and apply action:
  $x_t \leftarrow \text{argmax}_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x]$
  Apply $x_t$ and observe $y_t$

  #update distribution:
  $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot | x_t, y_t)$
**end for**

# Existing work and results

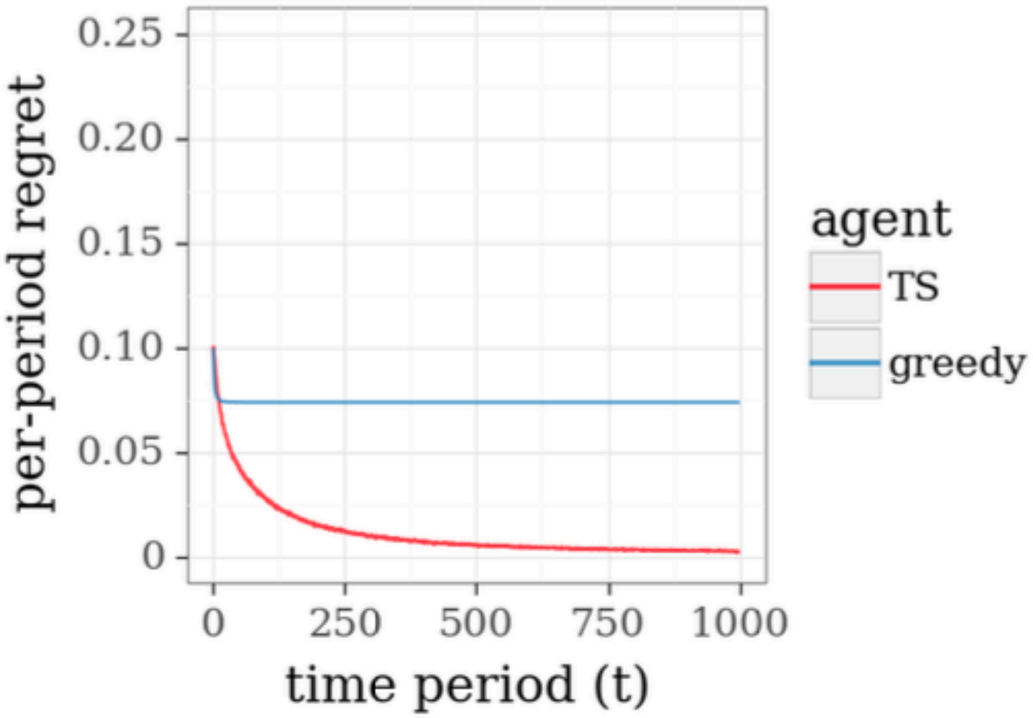



(a) greedy algorithm

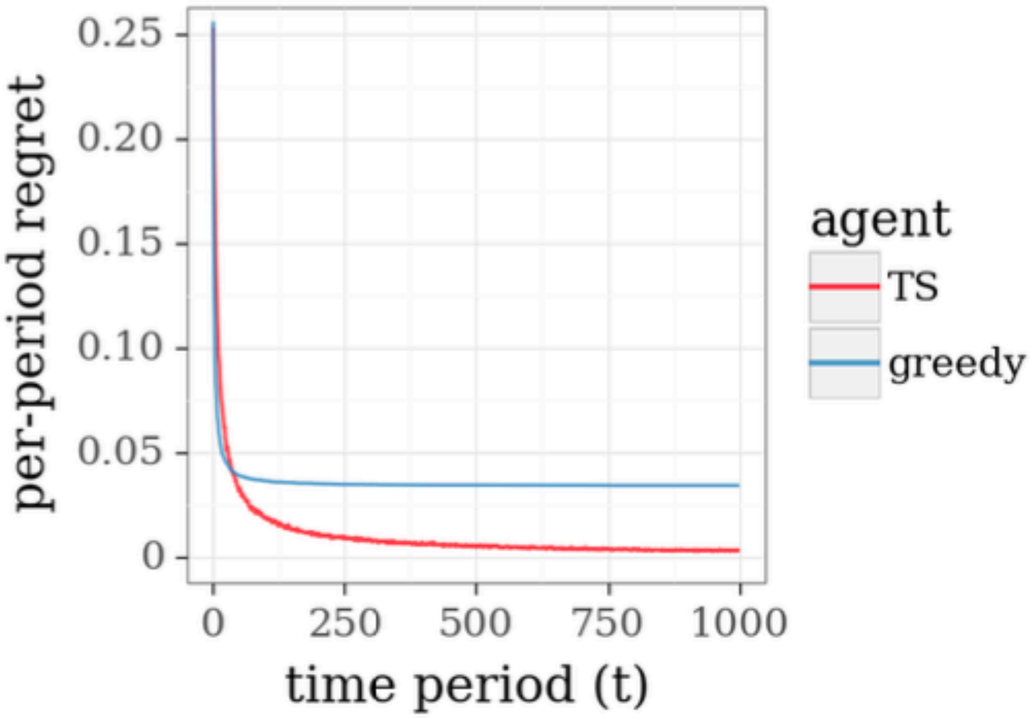



(b) Thompson sampling

# Existing work and results

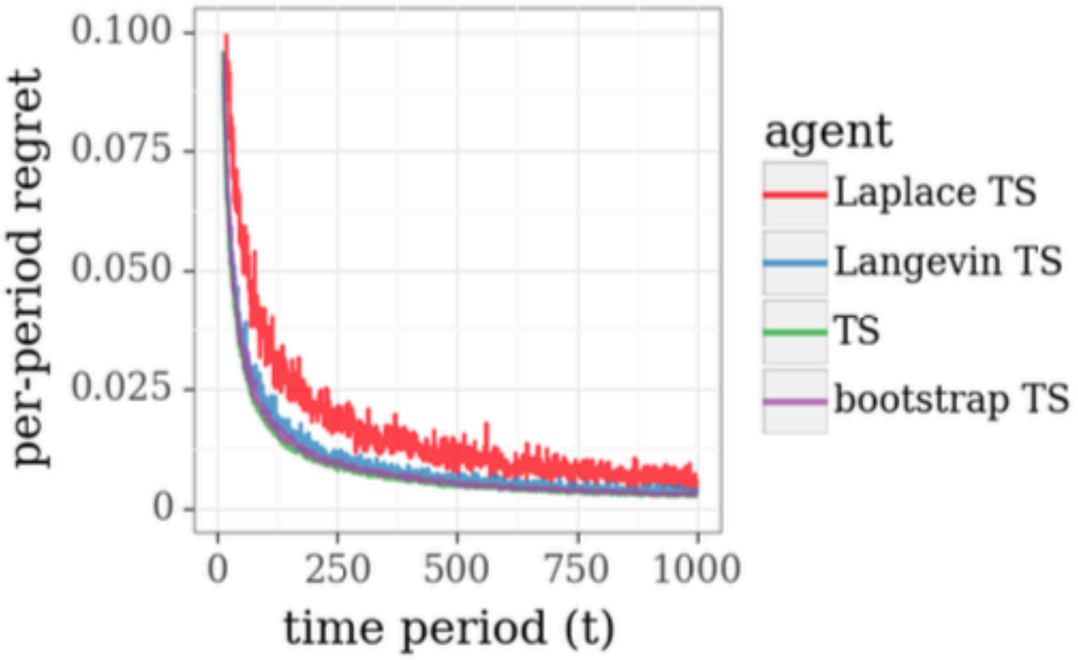



**(a)** $\theta = (0.9, 0.8, 0.7)$

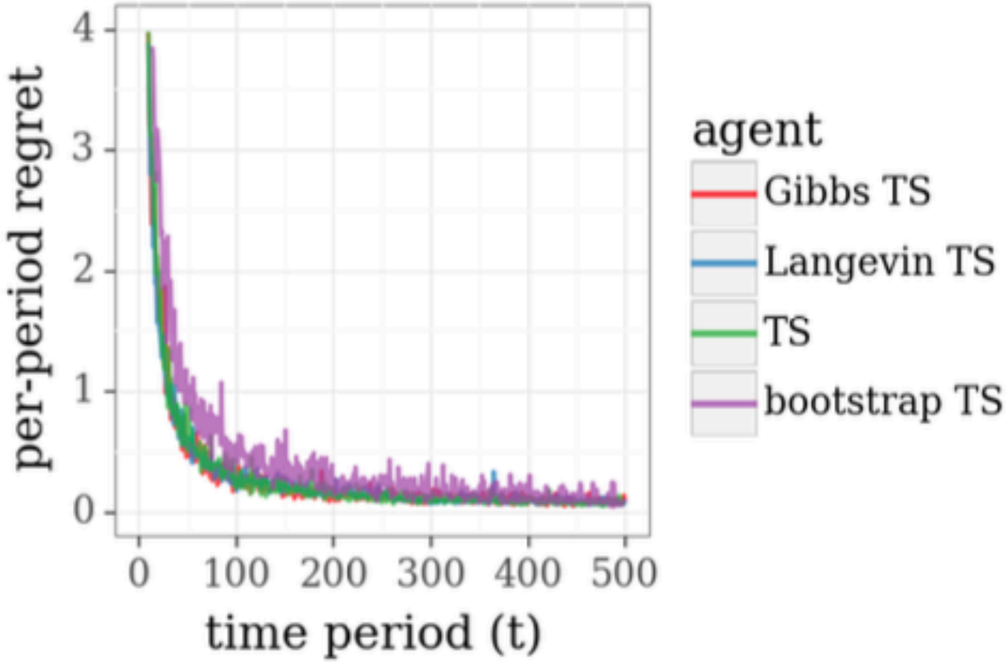



**(b)** average over random $\theta$

# Existing work and results




**(a)** Bernoulli bandit




**(b)** online shortest path

$$(\mu_e, \sigma_e^2) \leftarrow \left( \frac{\frac{1}{\sigma_e^2}\mu_e + \frac{1}{\tilde{\sigma}^2}\left(\ln(y_{t,e}) + \frac{\tilde{\sigma}^2}{2}\right)}{\frac{1}{\sigma_e^2} + \frac{1}{\tilde{\sigma}^2}}, \frac{1}{\frac{1}{\sigma_e^2} + \frac{1}{\tilde{\sigma}^2}} \right).$$

# Validation plan

- Keep the bandit problem setting the same
- Develop a (VI)-TS algorithm
- Theoretically, mathematical correctness
- Empirically, Greedy as upper bound, TS as lower bound
- Compare action probability, should converge to optical action/arm
- Compare regret over time period, should be in-between the two