**Optimistic Thompson Sampling-Based Algorithms for Episodic Reinforcement Learning**

Bingshan Hu (University of Alberta; Amii), Tianyue H. Zhang (University of British Columbia), Nidhi Hegde (University of Alberta; Amii), Mark Schmidt (University of British Columbia; Amii)

---

**Stochastic Multi-Armed Bandits (MAB)**

A stochastic MAB instance: \( \Theta = ([K]; \mu_1, \mu_2, \ldots, \mu_K) \)

1. Environment generates a reward vector: \( \left(X_1(t), X_2(t), \ldots, X_K(t)\right) \)
2. Simultaneously, Learner pulls an arm \( J_t \in [K] \)
3. Environment reveals \( X_k(t) \); Learner observes and obtains \( X_k(t) \)

Regret can be expressed as:

\[
R(T; \theta) = \sum_{t=1}^{T} \mathbb{E} \left[ \max_{j \in [K]} \mu_j - \mu_{J_t} \right] = \sum_{t=1}^{T} \mathbb{E} \left[ \Delta_{J_t} \right]
\]

Mean reward of optimal action: \( \mu_* = \max_{j \in [K]} \mu_j \)

Mean reward gap of sub-optimal action: \( \Delta_j = \mu_* - \mu_j \)

**Empirical MAB instance: \( \Theta_t = ([K]; \hat{\mu}_1(t-1), \hat{\mu}_2(t-1), \ldots, \hat{\mu}_K(t-1)) \)**

---

**Optimistic Thompson Sampling (O-TS)**

- **Sampled parameters are guaranteed to be better than empirical parameters**
- **Reshape posterior distributions in an optimistic way**

---

**Key Challenge: Exploitation vs Exploration Trade-Off**

**Exploitation**

Take actions with high empirical reward to gain pay-off.

**Exploration**

Take less observed actions to gather information.

---

**Vanilla Stochastic Bandit Algorithms**

- **UCB:** Optimism in face of uncertainty
  - Maintain posterior distributions for the mean rewards
  - Randomized
  - Construct confidence intervals
- **TS:**
  - Draw random posterior samples
  - \( \pi_t(J) \) is \( \hat{\mu}_j(t-1) + \sqrt{\frac{\ln(T)}{T_j(t)}} \)
  - \( J_t = \arg \max_{j \in [K]} \pi_t(J) \)
  - TS with Gaussian Priors: \( \hat{\mu}_j(t) \sim N\left(\tilde{\mu}_j(t-1), \frac{1}{b}\right) \)
  - \( J_t = \arg \max_{j \in [K]} \tilde{\mu}_j(t) \)

---

**Contributions and Related Work**

- **O-TS-MDP** enjoys an elegant theoretical analysis, avoiding bounding the absolute value of approximation error. O-TS-MDP* can be viewed as a randomized version of UCB-VI [Azar et al., 2017].

- **O-TS** for bandits was originally proposed and empirically evaluated in Chapelle and Li [2011]. May et al. [2012]. O-TS+ for bandits can be viewed as a randomized version of UCB1 [Auer et al., 2002].

---

**Acknowledgements**

This work was supported by Alberta Machine Intelligence Institute (Amii), the Canada CIFAR AI Chair Program, and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants.