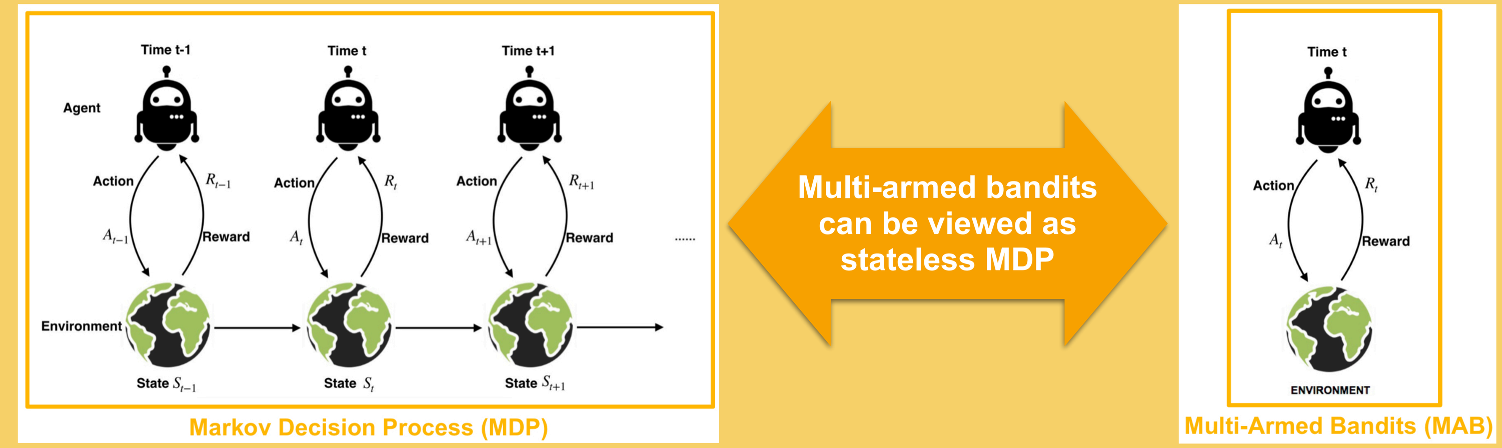
uai2023

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)

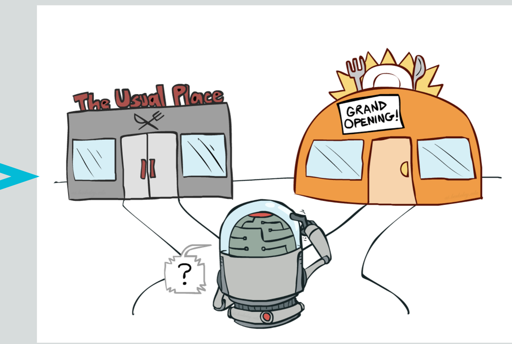
- Real-world environments are **complex and uncertain**
- Training data is **expensive**

Fast RL algorithms allow an agent to use less samples to learn a good policy



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

Stochastic Multi-Armed Bandits (MAB)

A stochastic MAB instance: $\Theta := ([K]; \mu_1, \mu_2, \ldots, \mu_K)$
Learning protocol: in every round $t = 1, 2, \ldots, T$

1. Environment generates a reward vector $\left(X_1(t), \ldots, \underbrace{X_j(t)}_{\sim \mathrm{Ber}(\mu_j)}, \ldots, X_K(t) \right)$
2. Simultaneously, Learner pulls an arm $J_t \in [K]$
3. Environment reveals $X_{J_t}(t)$; Learner observes and obtains $X_{J_t}(t)$

Regret can be expressed as

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

Mean reward of optimal action $\quad \mu_* = \max_{j \in [K]} \mu_j$

Mean reward gap of sub-optimal action $\quad \Delta_j = \mu_* - \mu_j$

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

Vanilla Stochastic Bandit Algorithms

UCB:
- Optimism in face of uncertainty
- Deterministic
- Construct confidence intervals

TS:
- Maintain posterior distributions for the mean rewards
- Randomized
- Draw random posterior samples

UCB: $\quad \overline{\mu}_j(t) = \widehat{\mu}_j(t-1) + \sqrt{\frac{1.5 \ln(t)}{O_j(t-1)}}, \quad J_t = \arg\max_{j \in [K]} \overline{\mu}_j(t)$

TS with Gaussian Priors: $\quad \widetilde{\mu}_j(t) \sim \mathcal{N}\left(\widehat{\mu}_j(t-1), \frac{1}{O_j(t-1)} \right), \quad J_t = \arg\max_{j \in [K]} \widetilde{\mu}_j(t)$

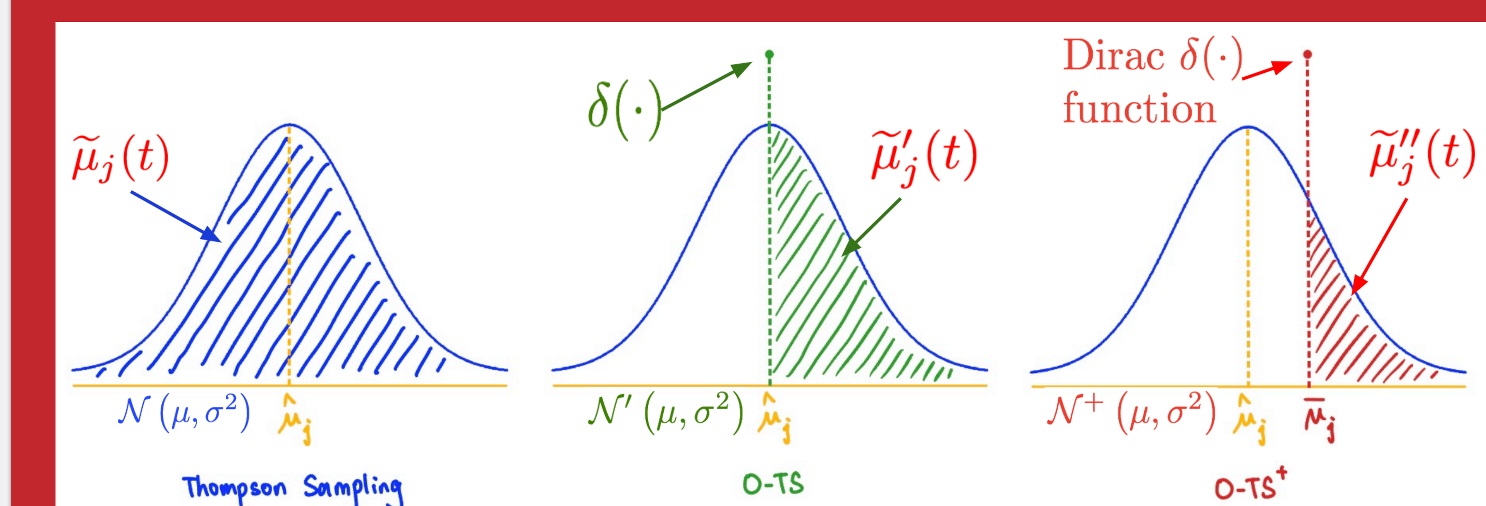
O-TS and O-TS+

	Bandit Regret Bounds
UCB1	$O\left(\sqrt{KT \ln(T)}\right)$
TS	$O\left(\sqrt{KT \ln(K)}\right)$
O-TS	$O\left(\sqrt{KT \ln(K)}\right)$
O-TS$^+$	$O\left(\sqrt{KT \ln(T)}\right)$

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.

Optimistic Thompson Sampling (O-TS)

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



Contributions and Related Work

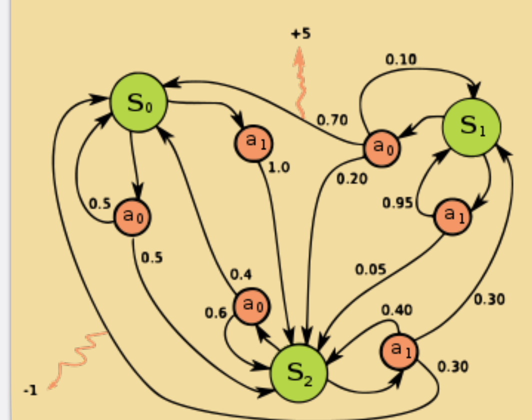
- O-TS-MDP enjoys an elegant theoretical analysis, avoiding bounding the absolute value of approximation error. O-TS-MDP+ can be viewed as a randomized version of UCB-VI [Azar et al., 2017].
- O-TS for bandits was originally proposed and empirically evaluated in Chapelle and Li [2011], May et al. [2012]. O-TS+ for bandits can be viewed as a randomized version of UCB1 [Auer et al., 2002].

Episodic Markov Decision Processes (MDP)

An MDP instance $M = ([S], [A], H, \vec{p}, \mu, T, p_0)$

- $(s, a) \in [S] \times [A]$: state-action pair
- $\vec{p}_{s,a,t}$: transition probability distribution for (s, a) in round t
- $\mu_{s,a,t}$: mean reward for (s, a) in round t
- p_0: initial state distribution
- H: number of rounds in an episode
- T: number of episodes

Goal of learner:
Visit a sequence of state-action pairs to accumulate as much reward as possible over T episodes (in total HT rounds)



Policy $\pi = (\pi(\cdot, 1), \pi(\cdot, 2), \ldots, \pi(\cdot, H))$: a sequence of functions, where each $\pi(\cdot, t) : \mathcal{S} \to \mathcal{A}$ takes a state s as input and outputs an action a that will be taken in that round t

Regret can be expressed as

$$\mathcal{R}(T; M) = \sum_{k=1}^{T} \mathbb{E}\left[V_1^{\pi_*}(s_1^k) - V_1^{\pi_k}(s_1^k) \right]$$

V_t^{π} : value function for policy π in round t

Empirical MDP instance: $M_k := ([S], [A], H, \widehat{p}_{k-1}, \widehat{\mu}_{k-1}, T, p_0)$

O-TS-MDP and O-TS-MDP+

	MDP Regret Bounds		
UCB-VI	$\widetilde{O}\left(\sqrt{ASH^3T}\right)$	Model-based: $\overline{\mu}, \widehat{p}$	Deterministic
RLSVI	$\widetilde{O}\left(\sqrt{AS^2H^4T}\right)$	Model-free	Randomized
O-TS-MDP	$\widetilde{O}\left(\sqrt{AS^2H^4T}\right)$	Model-based: $\widetilde{\mu}', \widehat{p}$	Randomized
O-TS-MDP$^+$	$\widetilde{O}\left(\sqrt{ASH^3T}\right)$	Model-based: $\widetilde{\mu}'', \widehat{p}$	Randomized

MDP Experiments

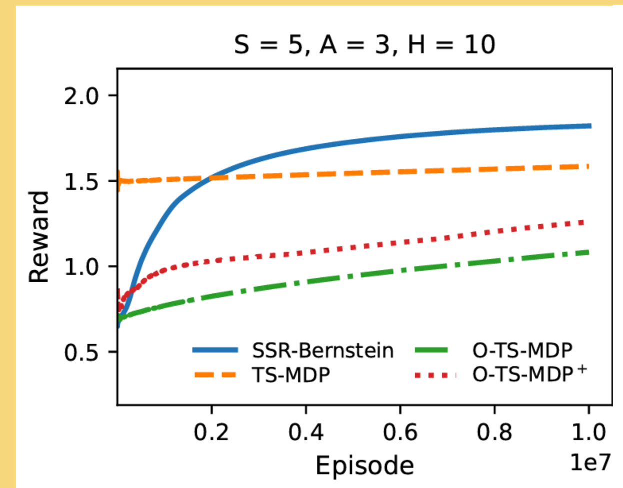


Figure 1: Empirical performance for 5 states

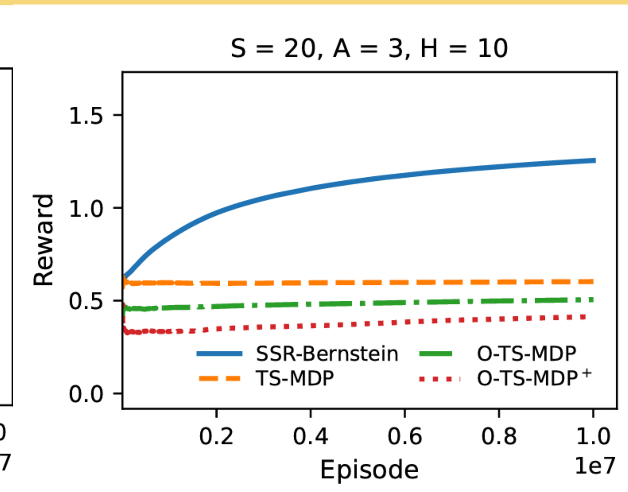


Figure 2: Empirical performance for 20 states

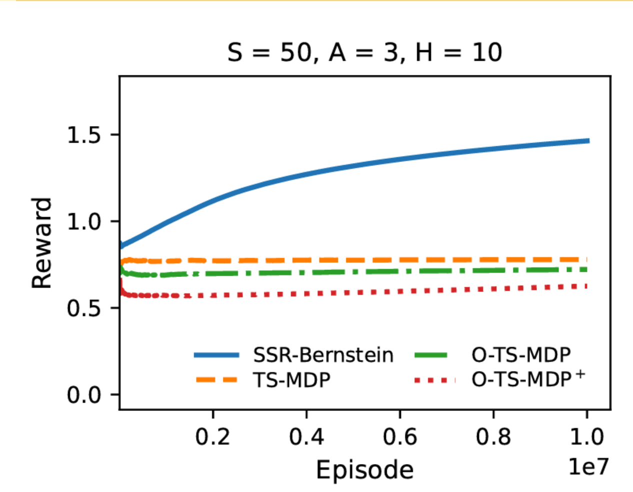


Figure 3: Empirical performance for 50 states