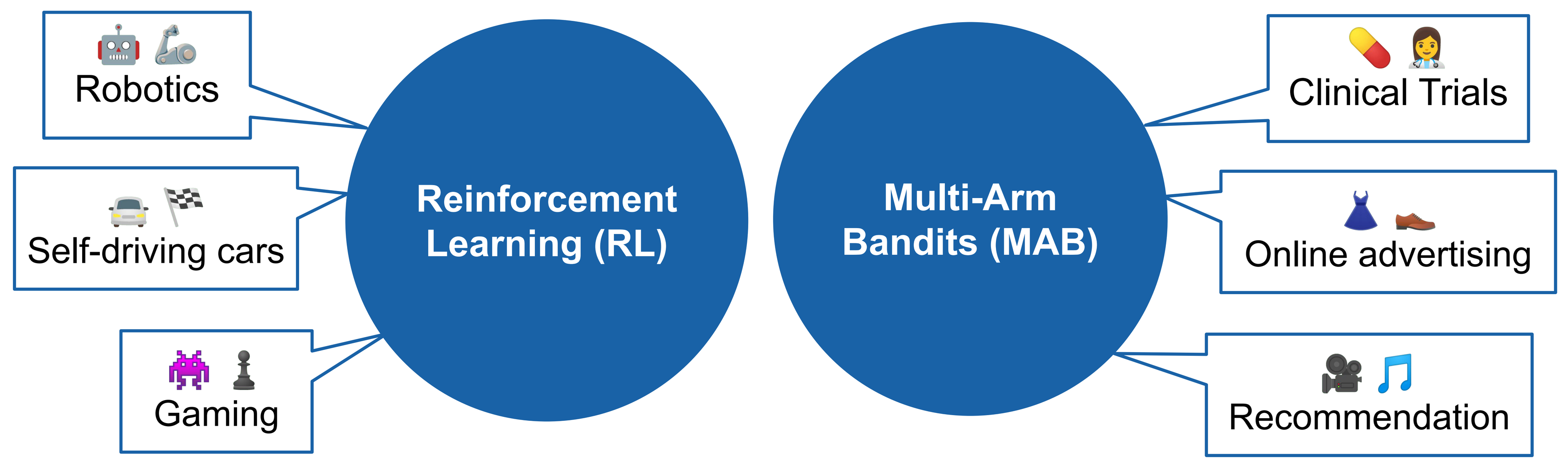
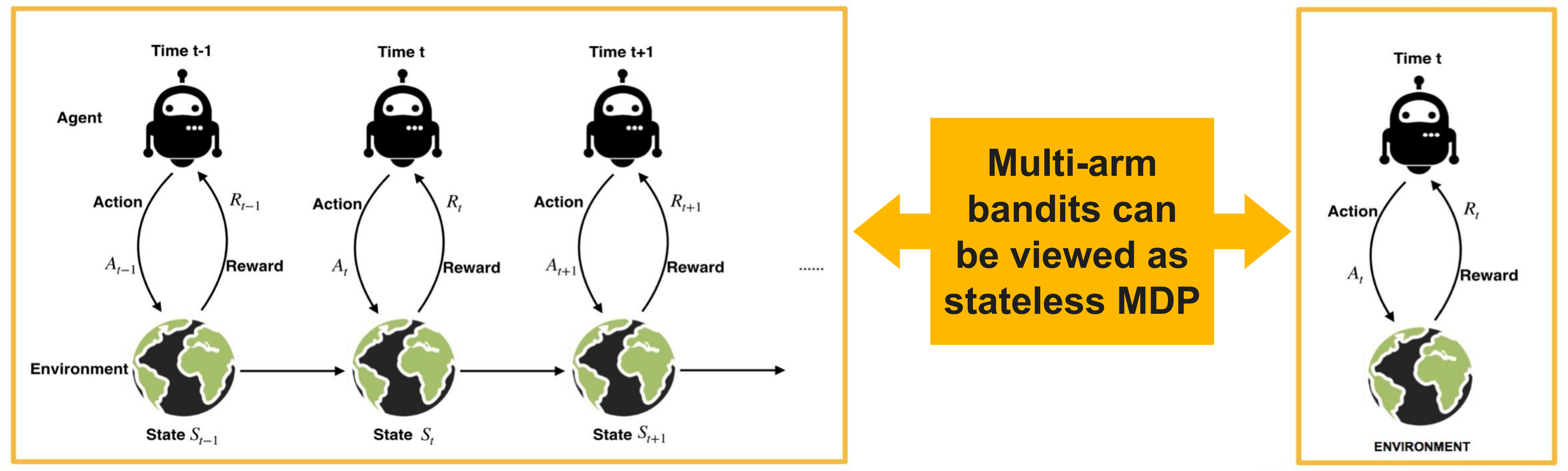
Optimistic Thompson Sampling-based Algorithms for Reinforcement Learning

Bingshan Hu (U-Alberta, Amii), Tianyue H. Zhang (UBC), Nidhi Hegde (U-Alberta, Amii), Mark Schmidt (UBC, Amii)

UNIVERSITY OF ALBERTA · THE UNIVERSITY OF BRITISH COLUMBIA · amii

Motivation: RL and Bandits





Markov Decision Process (MDP)

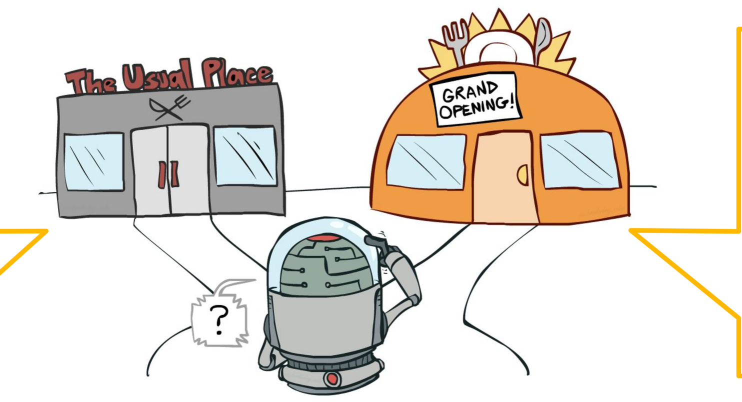
Multi-arm Bandits (MAB)

Check out our work at UAI 2023 (Oral presentation)!!!

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 $\Theta := ([K]; \mu_1, \mu_2, \ldots, \mu_K)$
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 $X_{J_t}(t)$

Goal: pull arms sequentially to maximize cumulative reward

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

Episodic Markov Decision Processes (MDP)

An MDP instance $M := \left(T, H, [S], [A], \{\mu\}_{[S]\times[A]\times[H]}, \{\vec{P}\}_{[S]\times[A]\times[H]}, p_0\right)$

. Number of episodes: T
. Number of rounds in each episode: H
. Mean reward function: $\{\mu_{s,a,t}\}$
. Transition probability distribution function: $\{\vec{p}_{s,a,t}\}$
. Deterministic initial state distribution: p_0
. State space: $[S]$
. Action space: $[A]$

Policy: $\pi = (\pi(\cdot,1), \pi(\cdot,2), \ldots, \pi(\cdot,H))$ with each $\pi(\cdot,t) : \mathcal{S} \to \mathcal{A}$ taking a state s_t as input and outputs an action a_t that will played in that state

Goal: play a sequence of policies $\pi_1, \pi_2, \ldots, \pi_k, \ldots, \pi_K$ to accumulate as much reward as possible

Regret: $\mathcal{R}(T;M) = \mathbb{E}\left[\sum_{k=1}^{K}\left(V_1^{\pi_*}(s_1^k) - V_1^{\pi_k}(s_1^k)\right)\right]$, where $V_t^{\pi}(s)$ is the value function and π_* is the optimal policy

UPPER Confidence BOUND (UCB) vs Thompson Sampling (TS) in Bandits

Unknown parameters: $(\mu_1, \mu_2, \ldots, \mu_K)$
Empirical parameters: $(\hat{\mu}_{1,O_1(t-1)}, \hat{\mu}_{2,O_2(t-1)}, \ldots, \hat{\mu}_{K,O_K(t-1)})$

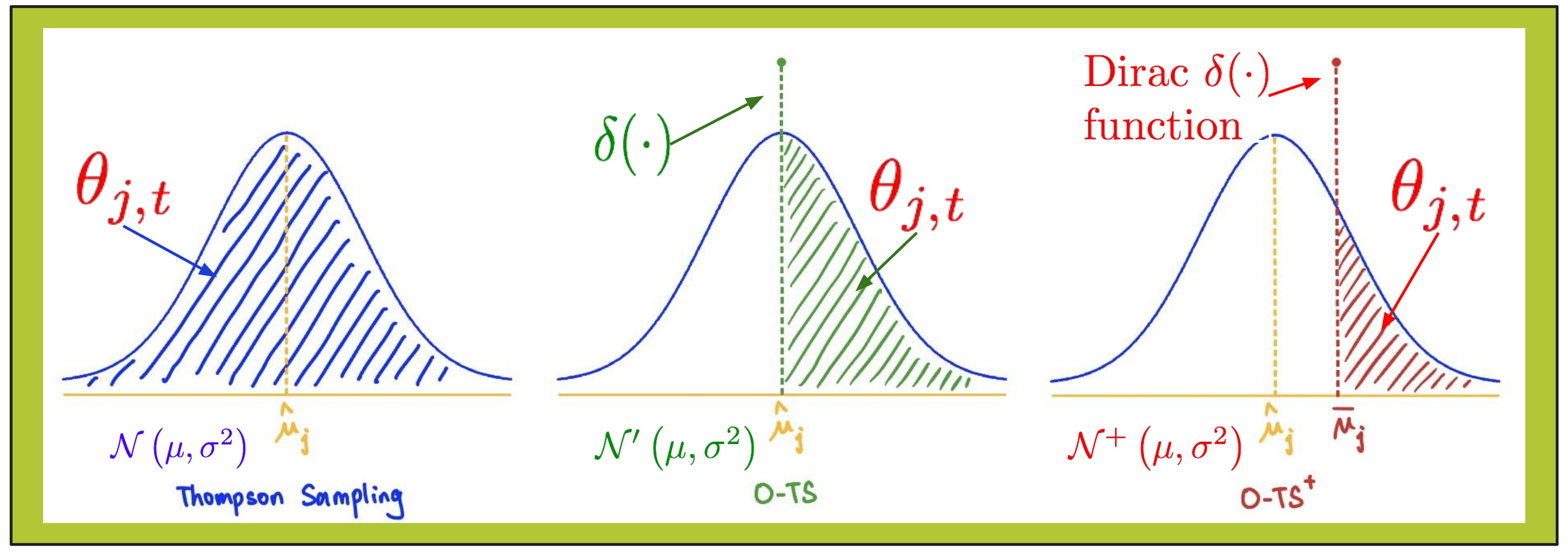
UPPER confidence BOUND (UCB): $\bar{\mu}_{j,t} = \hat{\mu}_{j,O_j(t-1)} + \sqrt{\frac{2\log(t)}{O_j(t-1)}}$ Pull arm $J_t = \arg\max \bar{\mu}_{j,t}$

Thompson Sampling (TS): $\theta_{j,t} \sim \mathcal{N}\left(\hat{\mu}_{j,O_j(t-1)}, \frac{1}{O_j(t-1)}\right)$

Optimistic TS (O-TS): $\theta_{j,t} \sim \mathcal{N}'\left(\hat{\mu}_{j,O_j(t-1)}, \frac{1}{O_j(t-1)}\right)$

Optimistic TS$^+$ (O-TS$^+$): $\theta_{j,t} \sim \mathcal{N}^+\left(\hat{\mu}_{j,O_j(t-1)}, \frac{1}{O_j(t-1)}\right)$

$J_t = \arg\max \theta_{j,t}$



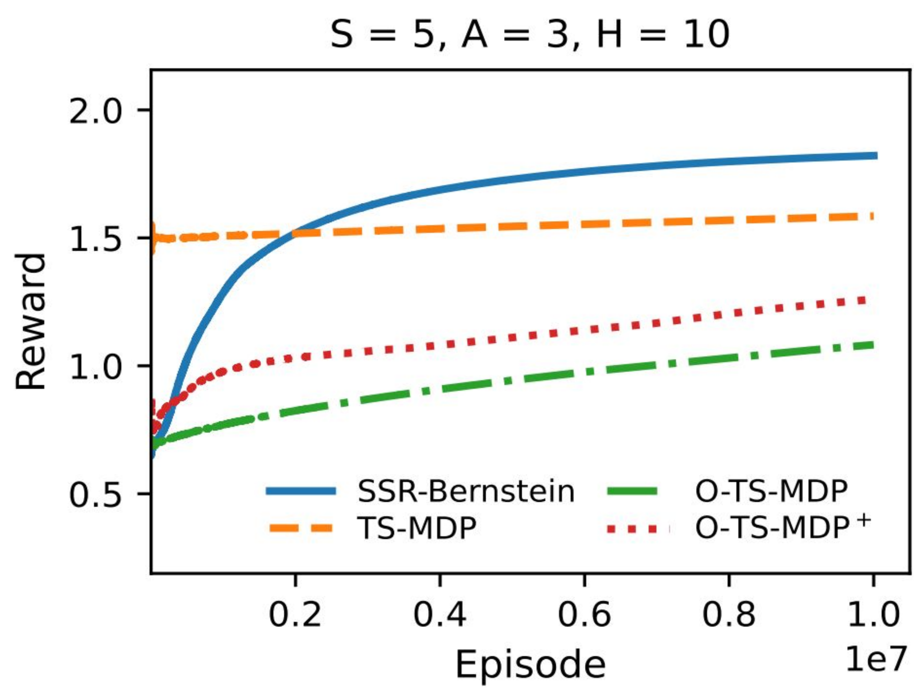
Regret UPPER BOUND

$O\left(\sqrt{KT\ln(T)}\right)$

More Optimistic Distributions!!

- O-TS for bandits was originally proposed and empirically evaluated in Chapelle and Li [2011], May et al. [2012].
- Key idea: sampled parameters are always better than empirical parameters!

Experiments for MDP



O-TS-MDP vs O-TS-MDP$^+$ in MDPs

Unknown parameters: $\mu_{s,a,t}, \vec{P}_{s,a,t}$

Empirical parameters: $\hat{\mu}_{s,a,t}^{k-1}, \hat{P}_{s,a,t}^{k-1}$

Model-based:
. Construct a model M^k in each episode k
. Find the best policy π_k for M^k

UCB-VI: $M^k = \left\{[S], [A], H, \bar{\mu}^k, \hat{P}^{k-1}\right\}$,
where $\bar{\mu}_{s,a,t}^k = \hat{\mu}_{s,a,t}^{k-1} + \tilde{O}\left(\sqrt{\frac{H^2}{O_{s,a,t}^{k-1}}}\right)$

O-TS-MDP: $M^k = \left\{[S], [A], H, \theta^k, \hat{P}^{k-1}\right\}$,
where: $\theta_{s,a,t}^k \sim \mathcal{N}'\left(\hat{\mu}_{s,a,t}^{k-1}, \tilde{O}\left(\frac{H^3 S}{O_{s,a,t}^{k-1}}\right)\right)$

O-TS-MDP$^+$: $M^k = \left\{[S], [A], H, \theta^k, \hat{P}^{k-1}\right\}$,
where: $\theta_{s,a,t}^k \sim \mathcal{N}^+\left(\hat{\mu}_{s,a,t}^{k-1}, \tilde{O}\left(\frac{H^2}{O_{s,a,t}^{k-1}}\right)\right)$

Regret UPPER BOUND

O-TS : $\tilde{O}\left(\sqrt{AS^2H^4T}\right)$
O-TS$^+$: $\tilde{O}\left(\sqrt{ASH^3T}\right)$

- O-TS-MDP enjoys an elegant theoretical analysis, avoiding bounding the absolute value of approximation error.
- O-TS-MDP$^+$ has the same regret bound as UCB-VI [Azar et al., 2017] and can be viewed as a randomized version of UCB-VI.

Acknowledgement

This work was supported by Alberta Machine Intelligence Institute (Amii) and the Canada CIFAR AI Program.