Society of Agents: Regret Bounds of Concurrent Thompson Sampling

Yan Chen (joint with)

Perry Dong, Qinxun Bai, Maria Dimakopoulou, Wei Xu, Zhengyuan Zhou

DUKE FUQUA SCHOOL OF BUSINESS

Berkeley UNIVERSITY OF CALIFORNIA

Horizon Robotics

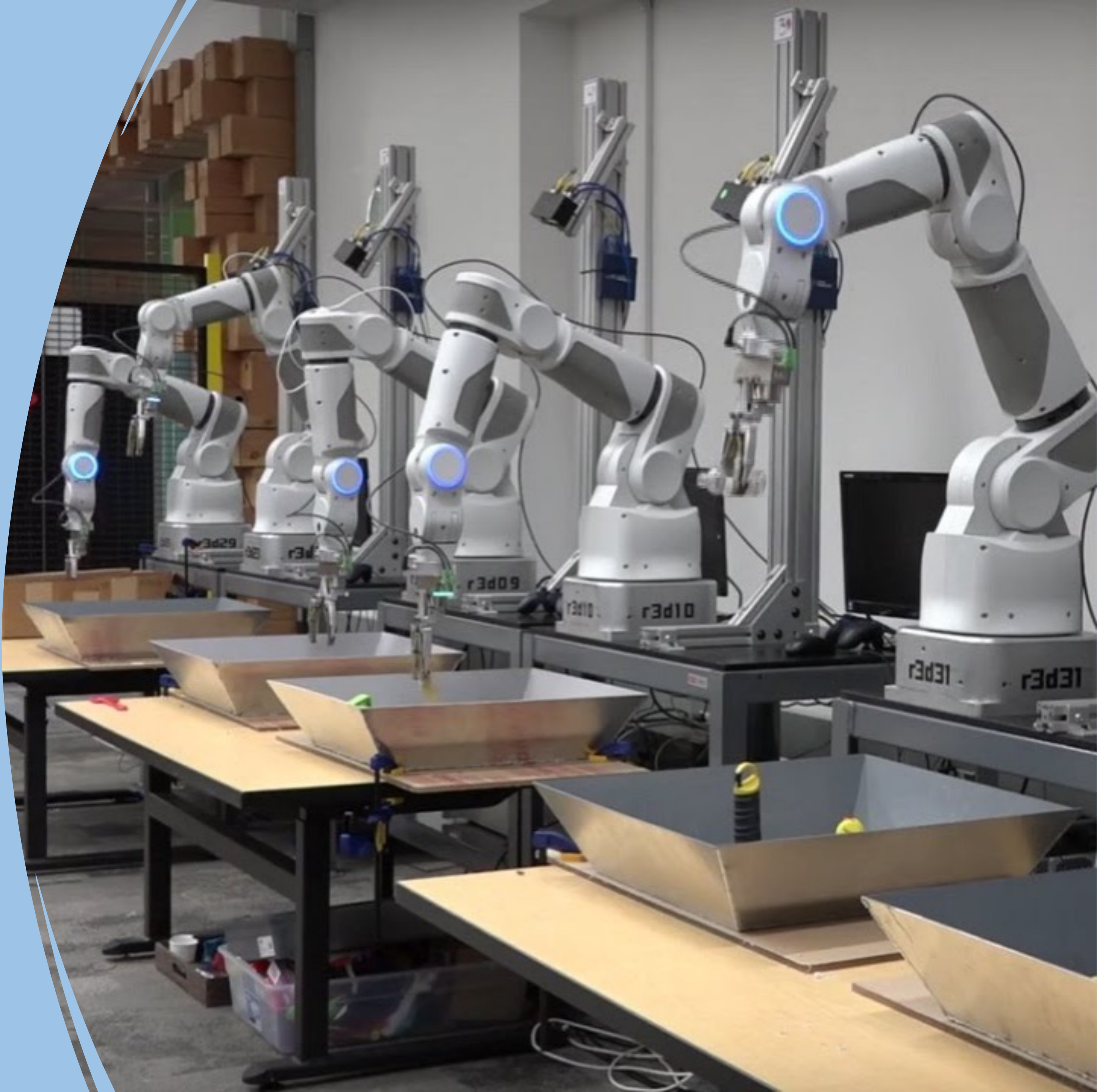
Spotify

NYU STERN

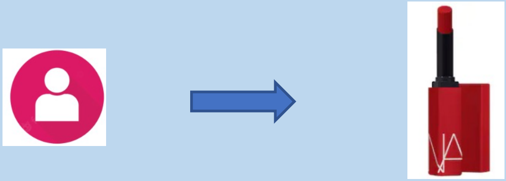
Arena

Google AI Robot Farm

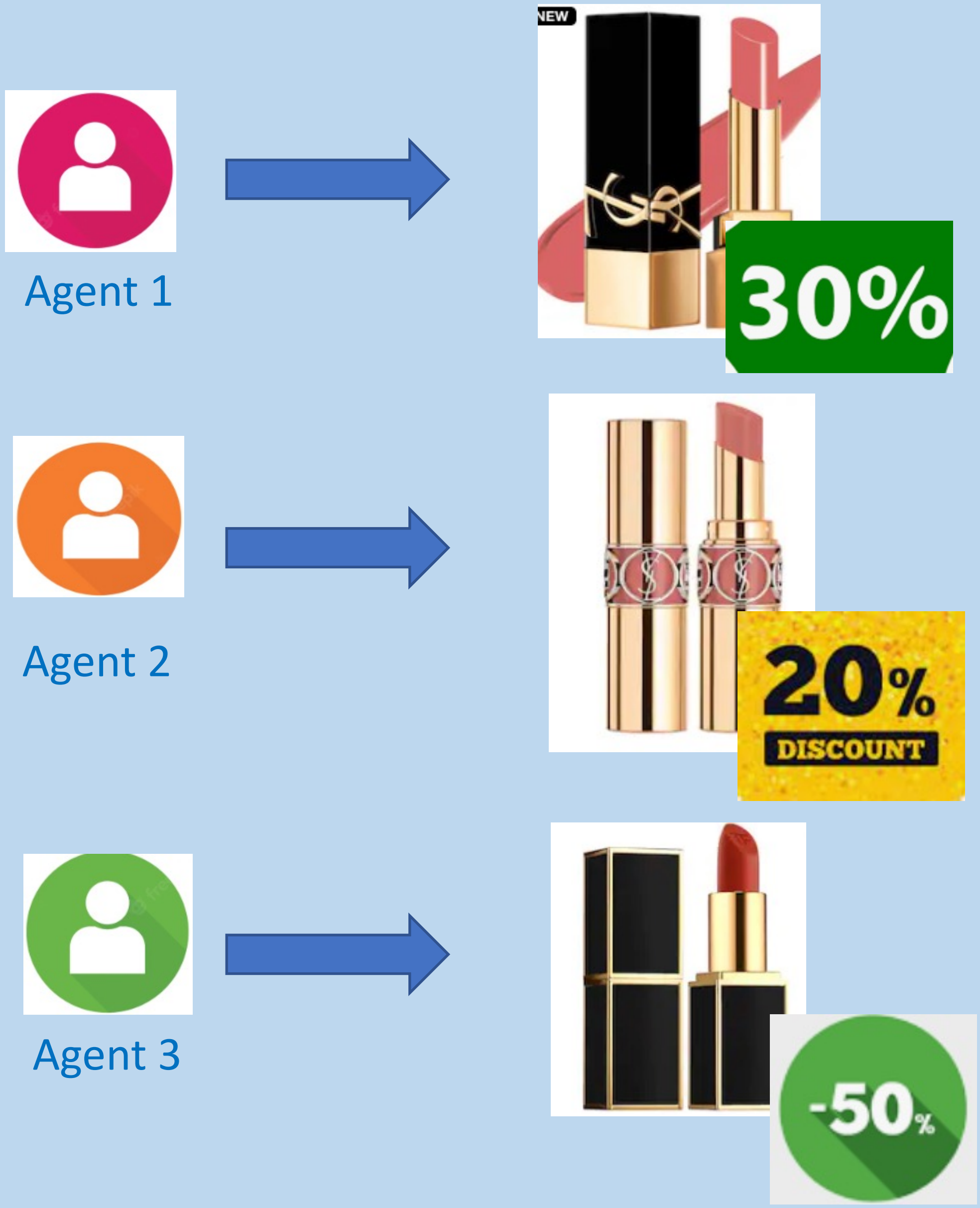
>



Personalized Promotions



Arena



Agent 1

Agent 2

Agent 3

Exploration

Concurrent UCRL

- **Same behavior of agents**
- NO DIVERSITY

- Upper Confidence Bounds

e.g. Guo et.al 2015, Pazis et.al 2016

Concurrent PSRL

- **Different behaviors of agents**
- DIVERSIFIED

- Posterior Sampling

e.g. Dimakopoulou et.al 2018, Dimakopoulou&Van Roy 2018

Motivation

- **Concurrent Posterior Sampling**

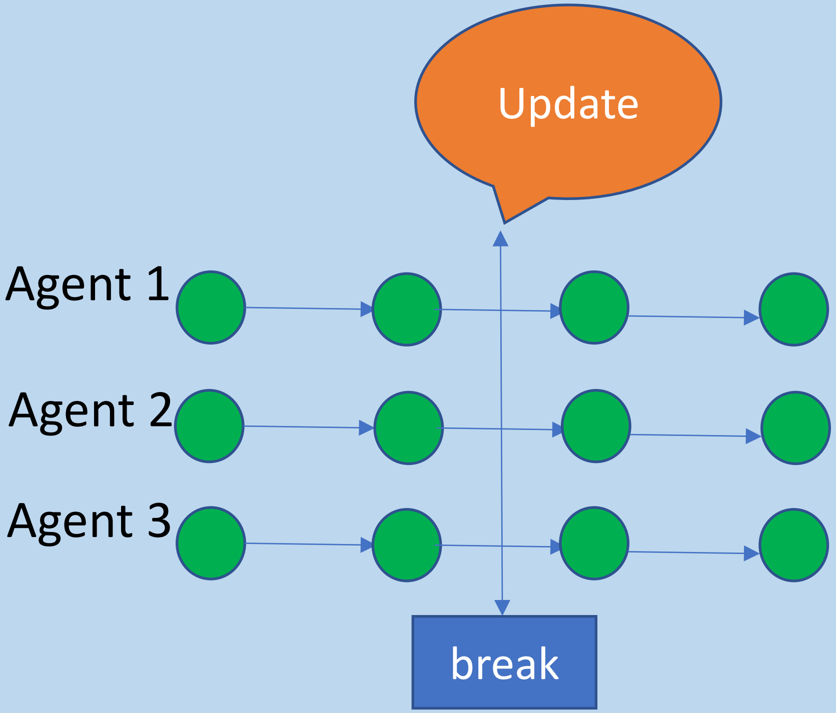
  - Empirical Evidence:
  - (e.g. Dimakopoulou et.al 2018, Dimakopoulou&Van Roy 2018)

  - **Theory:**
  - **?**

Our Contribution

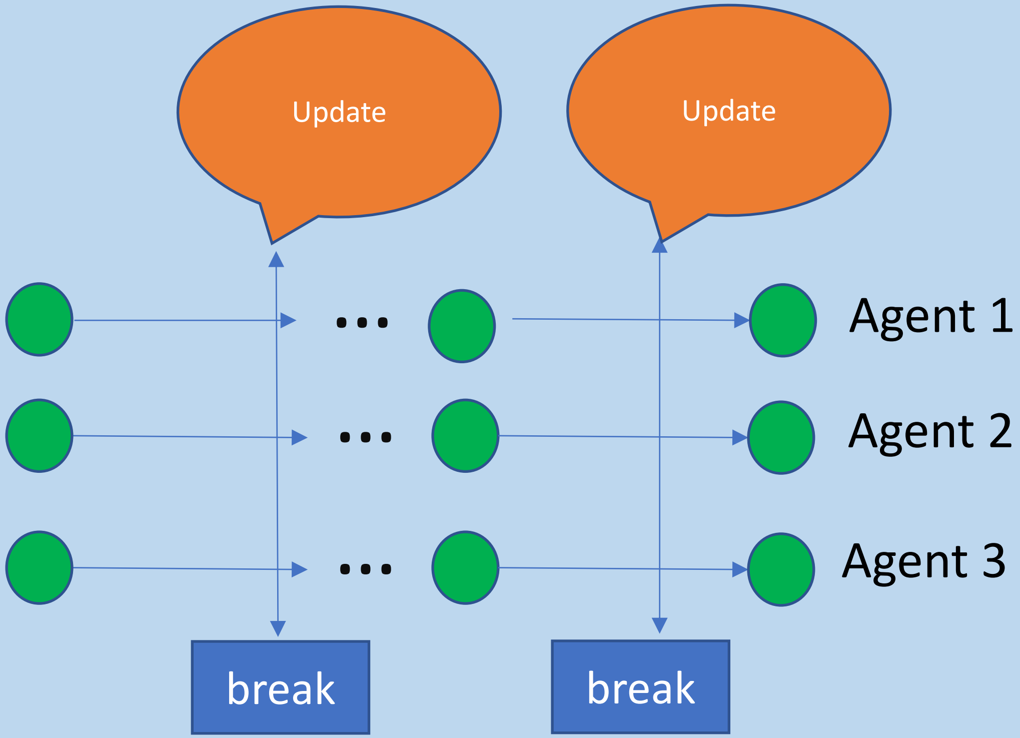
- **First Regret bounds** on simple-but natural concurrent PSRL
- **Finite-Horizon** & **Infinite-Horizon**

Models



Finite-horizon

2-episode, 2-horizon, 3 agents



Infinite-horizon

n=3 *(double epoch)*

Result Overview:

Per-Agent Bayesian Regret Bounds

Finite-horizon & infinite-Horizon

S: state space size;
n: number of agents

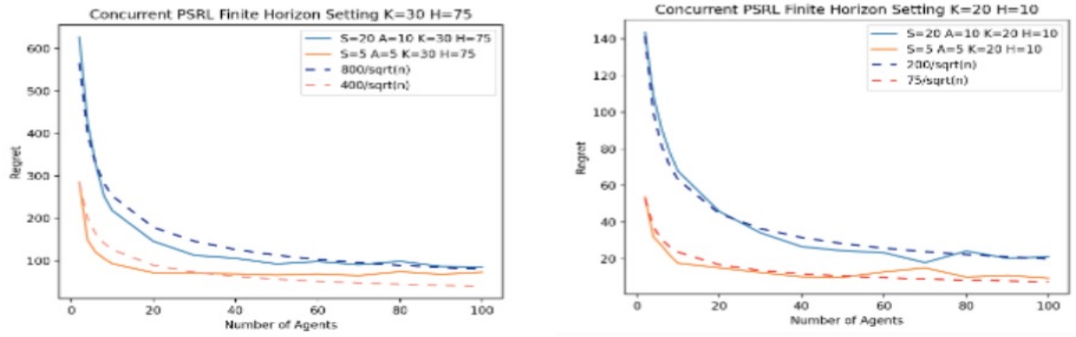
- General Prior: $\tilde{\boldsymbol{O}}\left(\frac{\boldsymbol{S}}{\sqrt{\boldsymbol{n}}}\right)$
- Dirichlet Prior: $\tilde{\boldsymbol{O}}(\sqrt{\boldsymbol{S}/\boldsymbol{n}})$

Numerical Results

$\widetilde{O}(1/\sqrt{n})$ **per-agent**

Bayesian Regret

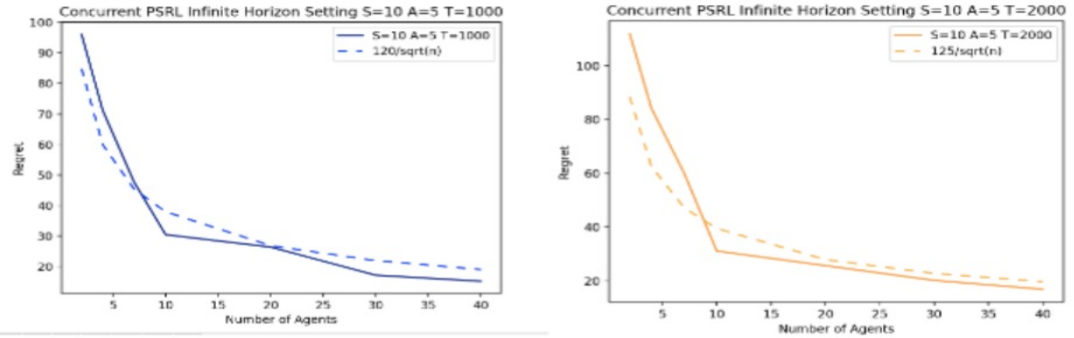
Finite-Horizon



(a) K=30, H=75

(b) K=20, H=10

Infinite-Horizon



(a) S=10, A=5, T=1000

(b) S=10, A=5, T=2000

Literature

- [1] Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.
- [2] Maria Dimakopoulou, Ian Osband, and Benjamin Van Roy. Scalable coordinated exploration in concurrent reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.
- [3] Maria Dimakopoulou and Benjamin Van Roy. Coordinated exploration in concurrent reinforcement learning. In *International Conference on Machine Learning*, pages 1271–1279. PMLR, 2018.
- [4] Zhaohan Guo and Emma Brunskill. Concurrent pac rl. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2015.
- [5] Jason Pazis and Ronald Parr. Efficient pac-optimal exploration in concurrent, continuous state mdps with delayed updates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- [6] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.
- [7] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.