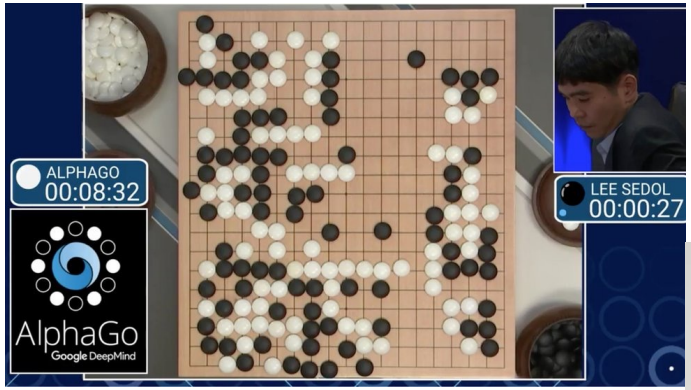
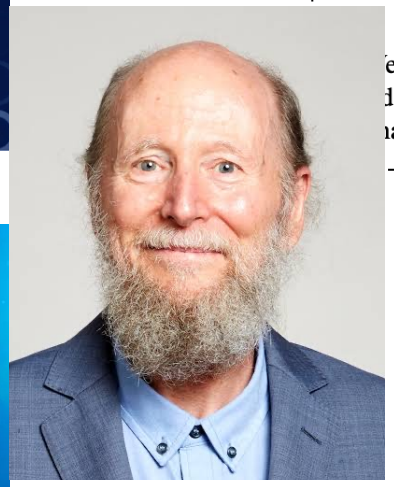
Efficient and Adaptive RL in Dynamic and Multi-Agent Environments: Theory and Applications

Shuai Li

Shanghai Jiao Tong University

Motiovation



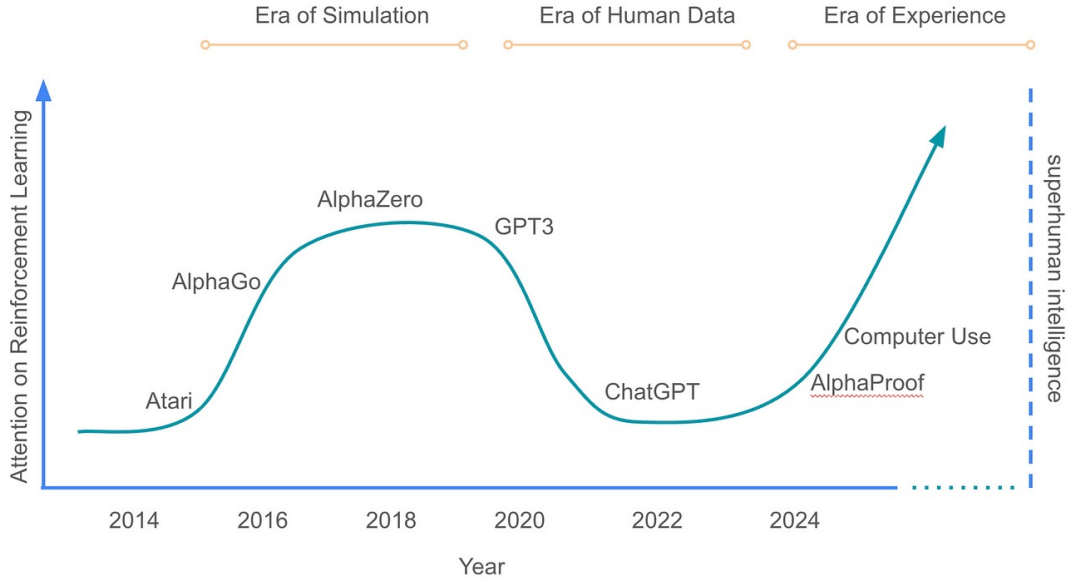


Welcome to the Era of Experience

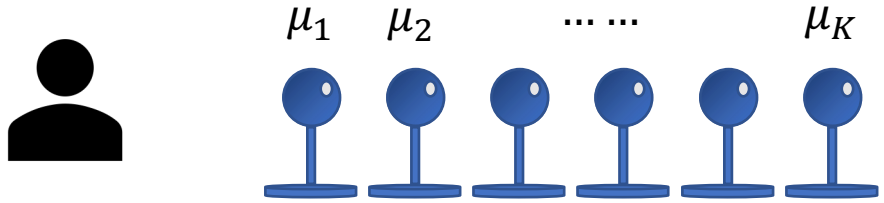
David Silver, Richard S. Sutton*

Abstract

e stand on the threshold of a new era in artificial intelligence that promises to achieve an unprece-
l level of ability. A new generation of agents will acquire superhuman capabilities by learning pre-
antly from experience. This note explores the key characteristics that will define this upcoming era.



Background



- Bandit algorithms
  - At each time t
    - The learning agent selects an action a_t
    - Receives reward X_t with mean $\mu(a_t)$
  - Objective: Minimize the cumulative regret

$$R(T) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu(a_t)\right]$$

  - Design algorithms with good regret guarantees
- Reinforcement learning algorithms
  - At each time with a state transition (Markov decision process)

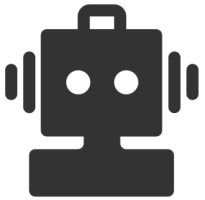
Challenges

Environment

Action

Feedback

Player



Challenge 1: Structured Action Space

- Dependence in the sequential structure
- Exponential size for the combinatorial structure

ICML'16,AAAI'18,COCOON'18,NeurIPS'18, ICML'19,NeurIPS'20,NeurIPS'21,AAMAS'23 ICML'24

Challenge 2: Changing Environment

- Adversarial environment destabilize algorithms
- Corrupted environment interfere learning

NeurIPS'21,AAAI'22,AISTATS'22,ICML'22, COLT'23,NeurIPS'23,AAAI'25

Challenge 3: Multi-players' Game

- Multiple players affect observations
- Strategic interactions impede rewards

IJCAI'22,CIKM'22,SODA'23,AAAI'24,TCS'24,NeurIPS' 24, ICLR'25,IJCAI'25survey

Structured Action Space – Online Learning to Rank

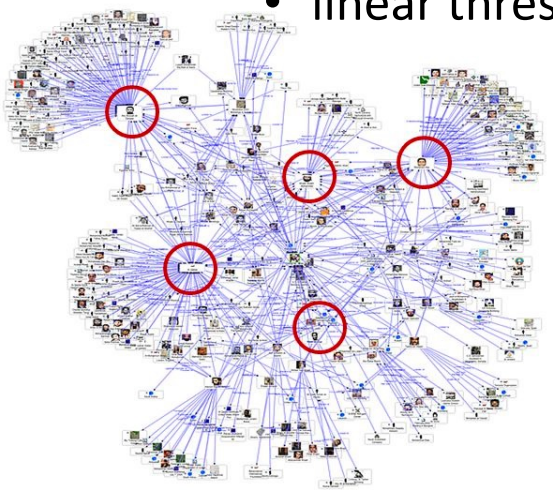
	Context	Click Model	Regret
[KSWA, 2015]	-	CM	$O(\frac{L}{\Delta}\log(T))$
[LWZC, ICML'2016]	Linear	CM	$O(\frac{d}{p^*}\sqrt{TK}\log(T))$
[LZ, AAAI'2018]	GL	CM	$O(d\sqrt{TK}\log(T))$
[KKSW, 2016]	-	DCM	$O(\frac{L}{\Delta}\log(T))$
[LLZ, COCOON'2018]	GL	DCM	$O(dK\sqrt{TK}\log(T))$
[LVC, 2016]	-	PBM with β	$O(\frac{L}{\Delta}\log(T))$
[ZTGKSW, 2017]	-	General	$O(\frac{K^3L}{\Delta}\log(T))$
[LKLS, NIPS'2018]	-	General	$O\left(\frac{KL}{\Delta}\log(T)\right)$ $O\left(\sqrt{K^3LT\log(T)}\right)$ $\Omega\left(\sqrt{KLT}\right)$
[LLS, ICML'2019]	Linear	General	$O(K\sqrt{dT\log(LT)})$

✗
✗
✓
?
?

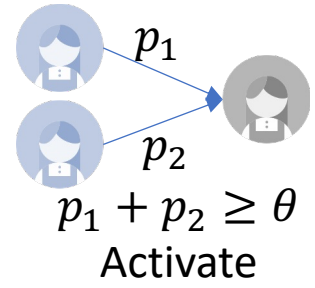
Collaborated with DeepMind, Adobe Research, University of Alberta, MSRA, etc.

Structured Action Space - Combinatorial Structure

- Influence maximization is one of the most representative NP-hard problem
- Jon Kleinberg brings up two common propagation models (KDD'03, 10k+ citations)
  - independent cascade model
  - linear threshold model

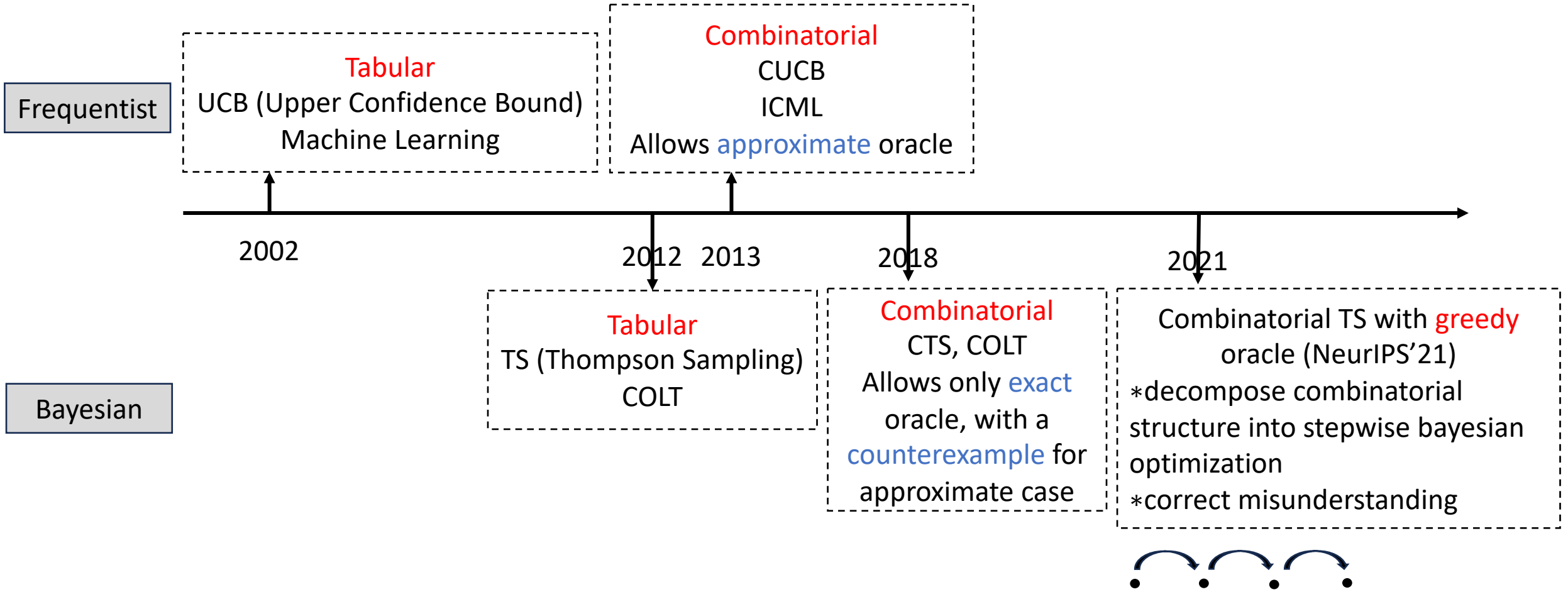


- Wei Chen (ACM/IEEE Fellow) solves online influence maximization only under IC model (ICML'13)
- Linear threshold model characterizes herd behavior
- Dependence causes challenges for analysis!
- We provided a group continuity property for LT model, then first present the convergence result (NeurIPS'20, after 7 years of IC)



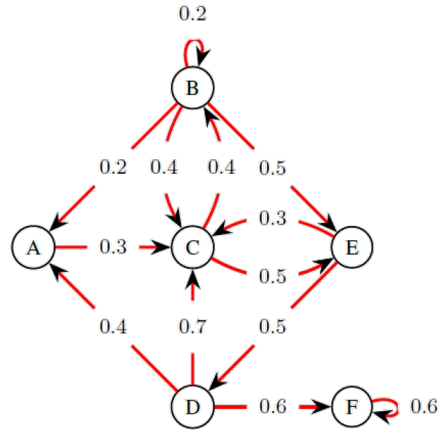
$$|r(S,w') - r(S,w)| \leq \mathbb{E}\left[\sum_{v\in V\setminus S}\sum_{u\in V_{S,v}}\sum_{\tau=\tau_1(u)}^{\tau_2(u)-1}\left|\sum_{e\in E_\tau(u)}(w'(e)-w(e))\right|\right]$$

Structured Action Space - Combinatorial Structure

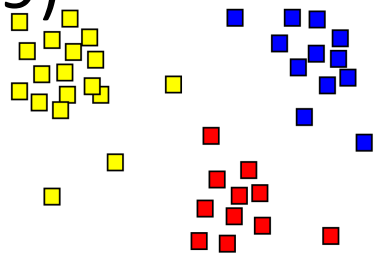


Structured Action Space - Graph Structure

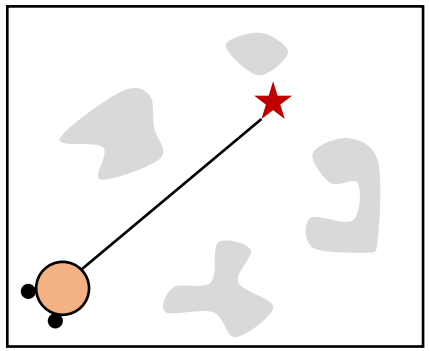
- Graph feedback infers feedback of relevant items
- We first consider the setting of probabilistic graph feedback, which infer relevant feedbacks with probability (AAAI'20)



- Online clustering structure
- Automatically adapt to finest clustering structure (IJCAI'19)



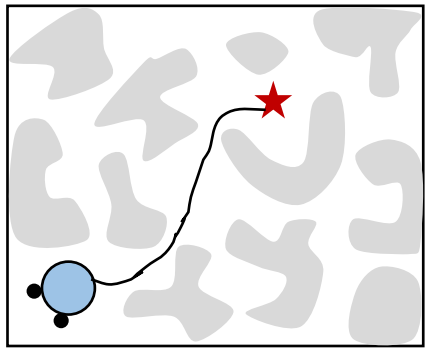
Dynamic Environment



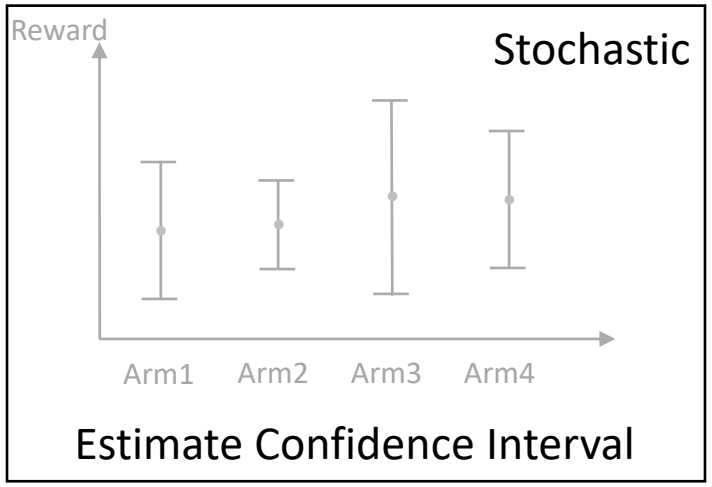
Stochastic Environment

Assumption Change

Algorithm Fail

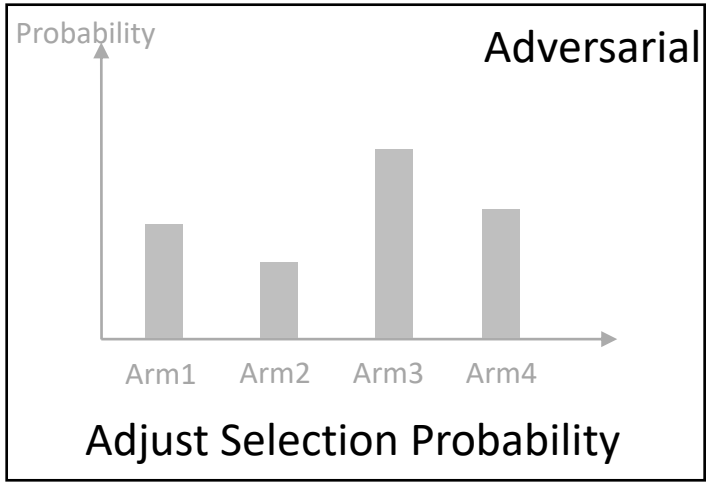


Adversarial Environment



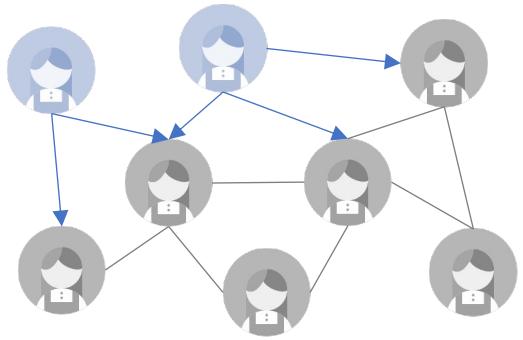
Combine

?

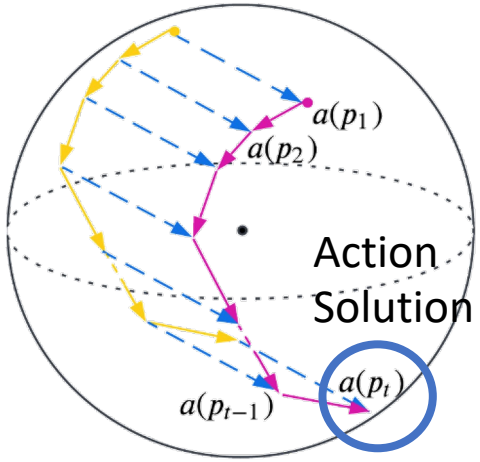


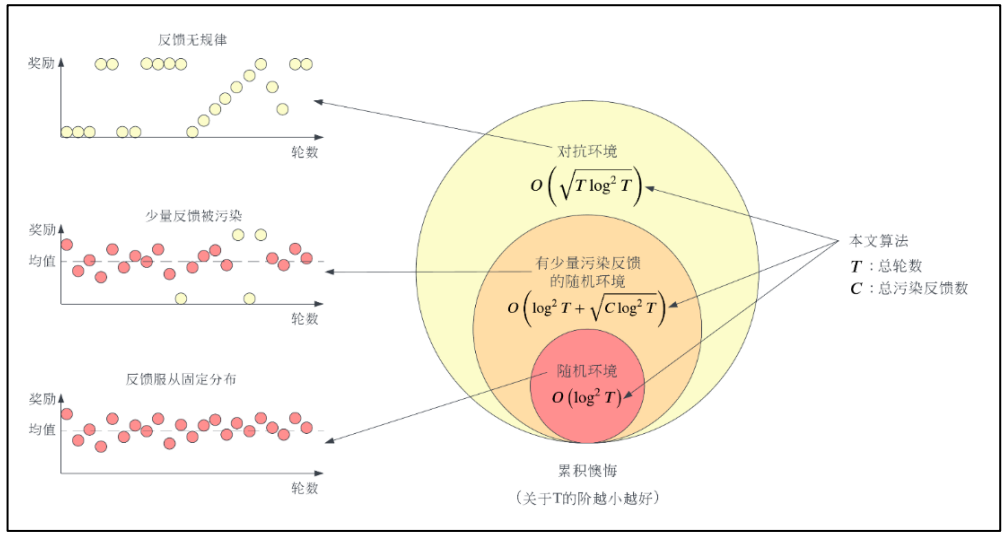
Dynamic Environment - Switch & Optimization

- Graph feedback
- Effective exploration relies on the graph structure
- For the general graph, design monitor-switch algorithm. Once detected environmental change, switch the algorithm [ICML'22 spotlight]



- Linear structure
- Use FTRL (follow-the-regularized-leader) structure to automatically adapt to different settings
- [COLT'23] First COLT paper in SJTU

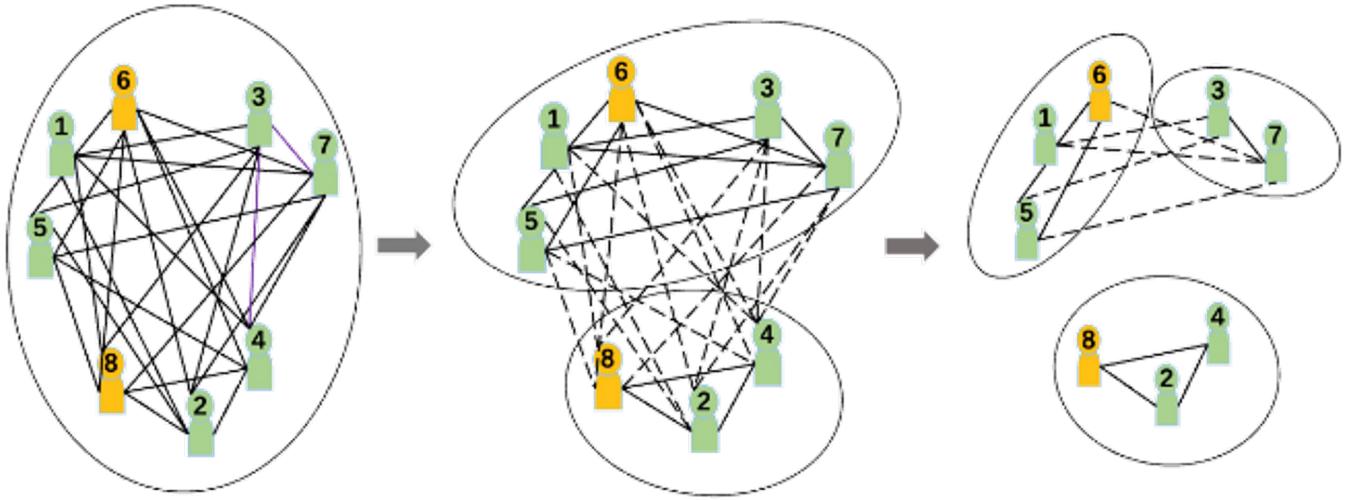




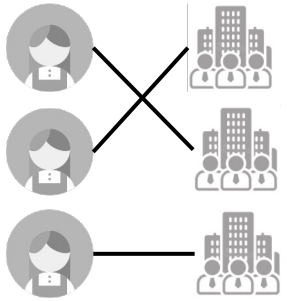
Dynamic Environment - Attack & Defense

- Ranking structure
- Design attack strategy (add noise to the returned feedback) to fail existing algorithms
- [NeurIPS'23a]

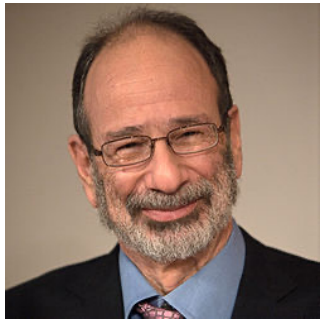
- Clustering structure
- Design defense strategy to enhance algorithm's exploration s.t. it can split out corrupted parts and stay robust
- [NeurIPS'23bc, AAAI'25 (oral)]



Multi-players' Game - Matching



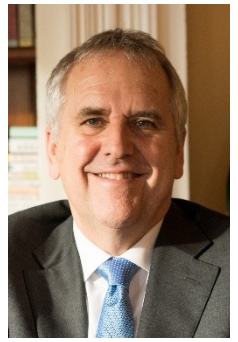
Matching Markets



Alvin E. Roth

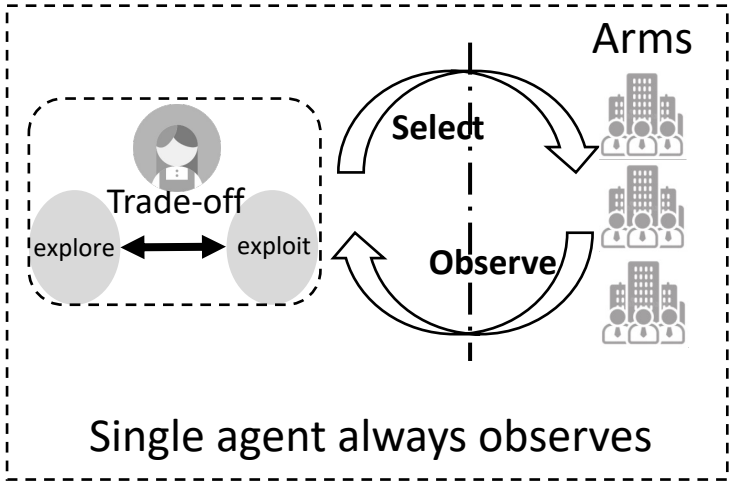
Lloyd Shapley

Stable Matching, 2012 Nobel Prize in Economics

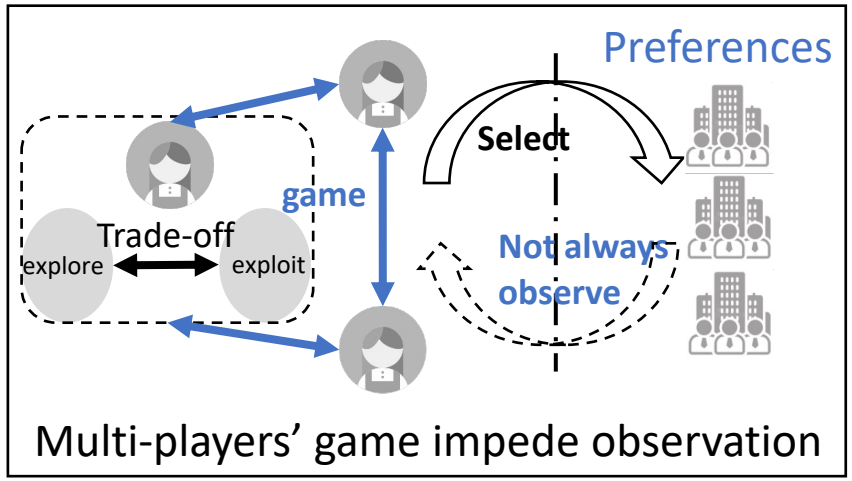


Michael Jordan
ACM Fellow
IEEE Fellow

Learning in matching markets with unknown preferences



Single agent always observes

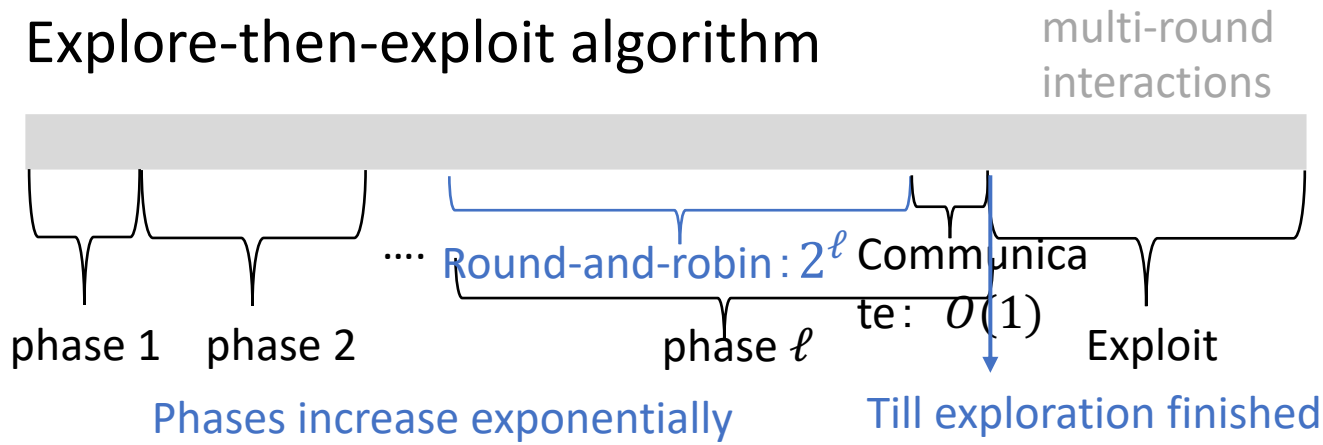


Multi-players' game impede observation

Multi-players' Game - Matching

- Existing works only borrow single-agent algorithm, but can't guarantee enough exploration
- Best lower bound
$$\Omega\left(\frac{N\log T}{\Delta^2}+\frac{K\log T}{\Delta}\right)$$
- Best upper bound player-pessimal stable regret $O\left(\frac{NK\log T}{\Delta^2}\right)$
- $N=$ #of players, $K=$ # of arms

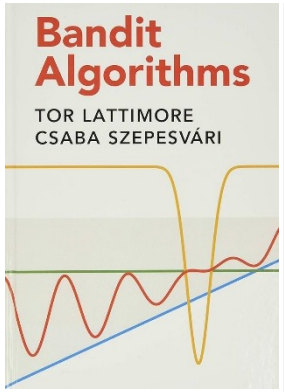
- Explore-then-exploit algorithm



- First player-optimal stable regret $O\left(\frac{K\log T}{\Delta^2}\right)$ [SODA'23]
- Single-agent self-determines exploration-exploitation, further improve learning efficiency [NeurIPS'24]
- Many-to-one matching markets [CIKM'22, AAAI'24]
- Indifference [ICLR'25]

Academic Influences

The algorithm for online learning to rank is included as a chapter in the textbook《Bandit Algorithms》



- Cambridge University Press
- Fundamental textbook in RL theory
- Reference textbook for graduate course in Oxford、Columbia、Washington U
- Citation: 3.7k

Apply online learning algorithm to combinatorial optimization problems, like SAT, win the bronze medal and Huawei spark award

	Solver	Authors	PAR-2	Solved
1	PRS-parallel	Zhihan Chen, Xindi Zhang, Yuhang Qian and Shaowei Cai	1143.56	143
2	Mallob64	Dominik Schreiber	1505.21	137
3	pKisDS-step	Zhihui Xie, Xu Liu, Wanqian Luo, Junhua Huang, Hui-Ling Zhen, Xijun Li, Mingxuan Yuan and Shuai Li	1496.61	132

Bronze Medal

Accelerate and solve 132 instances in competition

First apply online algorithms to parallel reasoning for SAT

- AAMAS-2024 tutorial on bandit learning in matching markets
- AAMAS-2025 tutorial on Markov games (twice, only Chinese)
- IJCAI-2025 tutorial on bandit learning in matching markets
- IJCAI-2025 survey on bandit learning in matching markets

AAMAS 2024 AUCKLAND NZ

T3. Bandit Learning in Mechanism Design: Matching Markets and Beyond

Presenters:

Shuai Li (Shanghai Jiao Tong University), Fang Kong (Shanghai Jiao Tong University)

Contact email: shuaili8@sjtu.edu.cn

DETROIT aamas2025

T3. Theoretical Foundations for Markov Games (link)

Presenters: Shuai Li (Shanghai Jiao Tong University), Canzhe Zhao (Shanghai Jiao Tong University)

Applications

Adaptive Merchant-Centric Risk Control by Ant Group

- Decrease the risk exposure rate by 33%, and merchant complaints 69%, marking a notable improvement in both risk mitigation and user satisfaction

AAAI-IAAI Deployed Application Award

《基于实时反馈数据的智能决策系统研究》产学研课题合作证明

上海交通大学:

我司和上海交通大学于 2024 年签署了合同编号为：2024011962803 的【基于实时反馈数据的智能决策系统研究】项目合同（下称"项目"）进行合作，贵单位的李帅老师作为该项目的【项目 PI】，现阶段，该项目的产出成果之一"基于无偏决策和动态优化的商户风控方案"已在我司内部业务场景上线使用，并带来了一定的指标提升。

支付宝（杭州）信息技术有限公司

2025 年 5 月

Online Test at Tencent

- Weed out 30.45% strategies during experimentation and save about 10% experimental time

Tencent 腾讯

基于老虎机算法的在线序列化最优策略检测方法

应用证明

腾讯微信实验平台和上海交通大学李帅老师"基于老虎机算法的在线序列化最优策略检测方法"的科研合作项目的联合研究产出之一"基于 Elimination 的流量调整方案"已经在微信实验平台中上线使用。

相比于产品已有算法，带来以下业务指标提升:

1. *相较于平台现有的 A/B 测试方案，基于 Elimination 的流量调整方案可以提早淘汰表现较差的策略，平均有 30.45%的待测试策略可以在实验结束前淘汰;*
2. *在得到相同实验结论的情况下，基于 Elimination 的流量调整方案相较于平台现有的 A/B 测试方案平均节省约 10%的实验时间。*

特此证明。

腾讯科技〔深圳〕有限公司

2024 年 01 月

Intelligent Decisions

- Improve the success rate by 18.3%

应用证明

项 目 名 称	水面舰 SSDK 智能决策技术研究
应 用 单 位	中国船舶集团有限公司第七二六研究所
通 讯 地 址	上海市闵行区金都路 5200 号
应用起止时间	2023 年 1 月-至今

应用情况及社会、军事效益:

中国船舶集团有限公司第七二六研究所和上海交通大学李帅老师、张伟楠老师的关于"水面舰 SSDK 智能决策技术研究"的科研合作项目的联合研究产出之一"基于 bandit 的智能决策技术研究"在仿真测试中实现了多角度拦截，提升了已有单步算法的成功率。另一联合研究产出"基于强化学习的智能决策技术研究"在多步决策仿真测试中实现了全角度拦截，并可基于 QC 数量进行泛化，成功率以及决策逻辑符合项目要求。两研究成果为新型 SSDKZZ 软件的研制提供有力支撑。

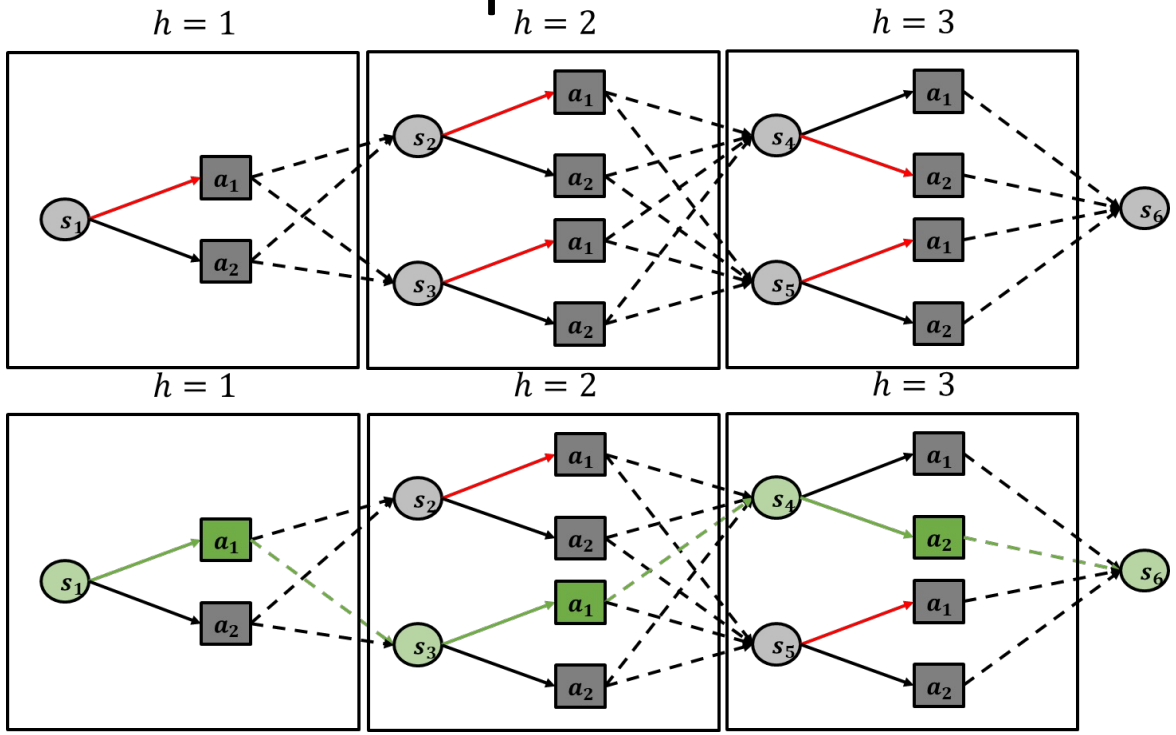
具体而言，相比于产品已有算法，研究成果带来以下业务指标提升:

1. 对于单步决策问题，基于 bandit 的算法在多角度下均能学习到有实际意义、拦截成功率高的策略。在相同的条件下，该算法的成功率相较于已有单步算法平均提升 18.3%。
2. 对于多步决策问题，基于强化学习的算法在全角度下学习到决策逻辑符合要求、拦截成功率高的策略。该算法可针对 QC 数量进行泛化，且算法成功率相对于单步决策进一步提升 4.2%。

中国船舶集团有限公司第七二六研究所

2024 年 01 月 04 日

MDP is a special case of Combinatorial Bandit

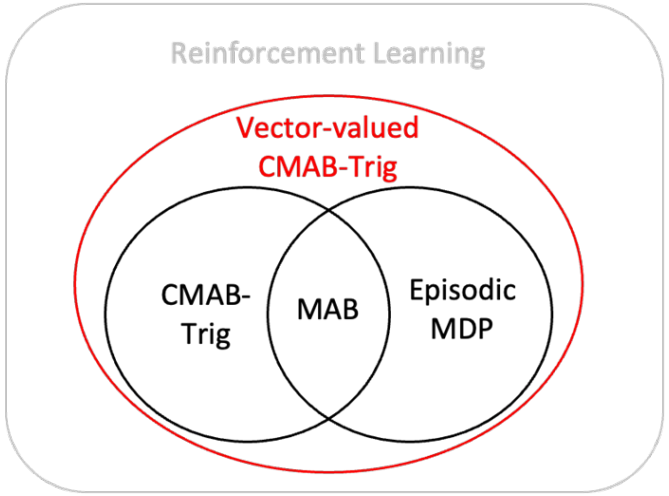


Key Observations

Policy = {state-action pairs}

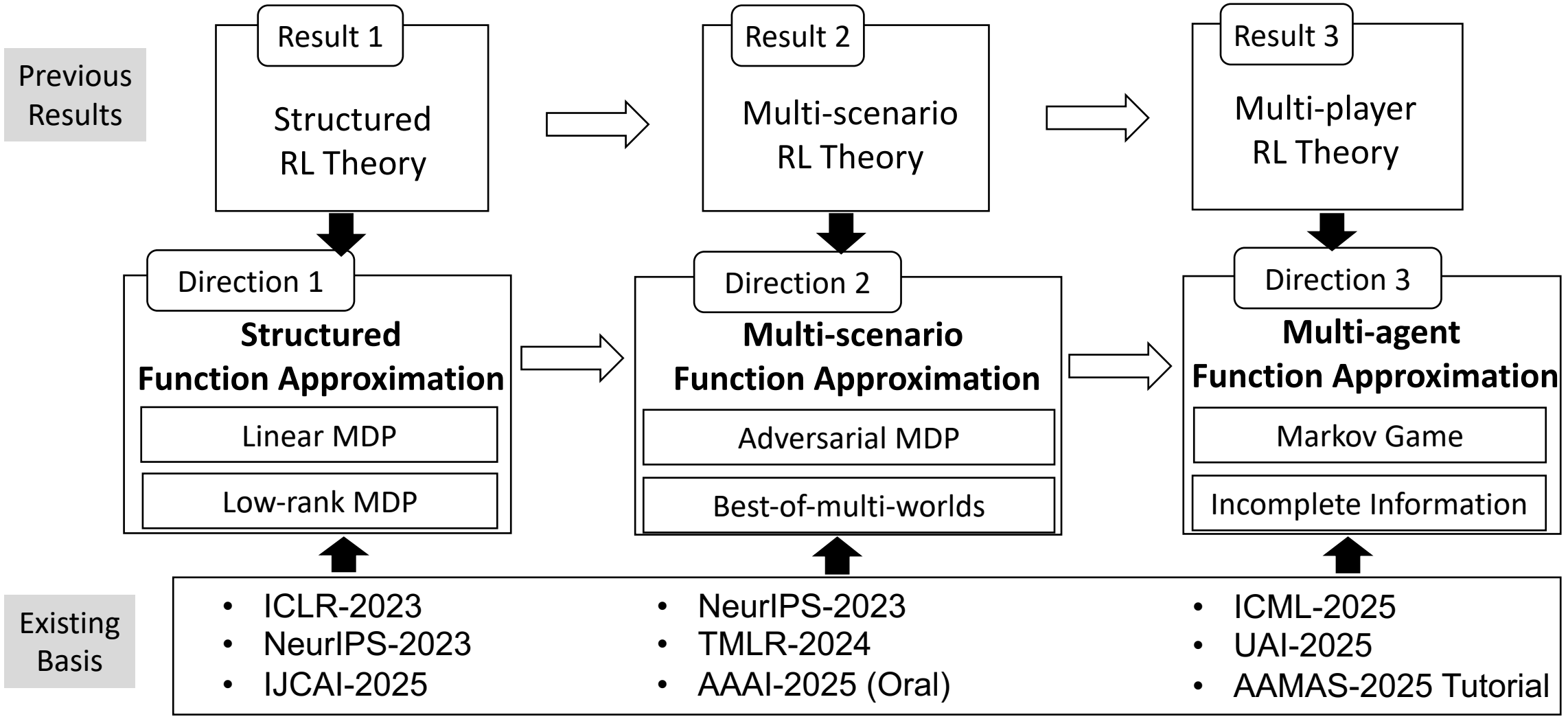
Combinatorial

state-action pair
triggered to be observed



- Regret $O\left(\sqrt{H^3SAT\log(SAHT)} + H^3S^2A\log^{3/2}(SAHT)\right)$
  - Match lower bound $\Omega\left(\sqrt{H^3SAT}\right)$ up to log factors
  - Save $O\left(\log^{5/2}(SAHT)\right)$ factor for $O(\sqrt{T})$ term compared to [Zanette and Brunskill,19]
  - As a by-product, this work could derive gap-dependent bound naturally
    [Simchowitz and Jamieson,19] use a very complicated analysis to derive gap-free bound from gap-dependent bound

Future Direction - Large-scale RL Theory



Thanks!
&
Questions?

Shuai Li

- Associate professor at Shanghai Jiao Tong University
- Research interests: RL/ML Theory
- Personal website: https://shuaili8.github.io/