Advances in Offline Reinforcement Learning and Beyond

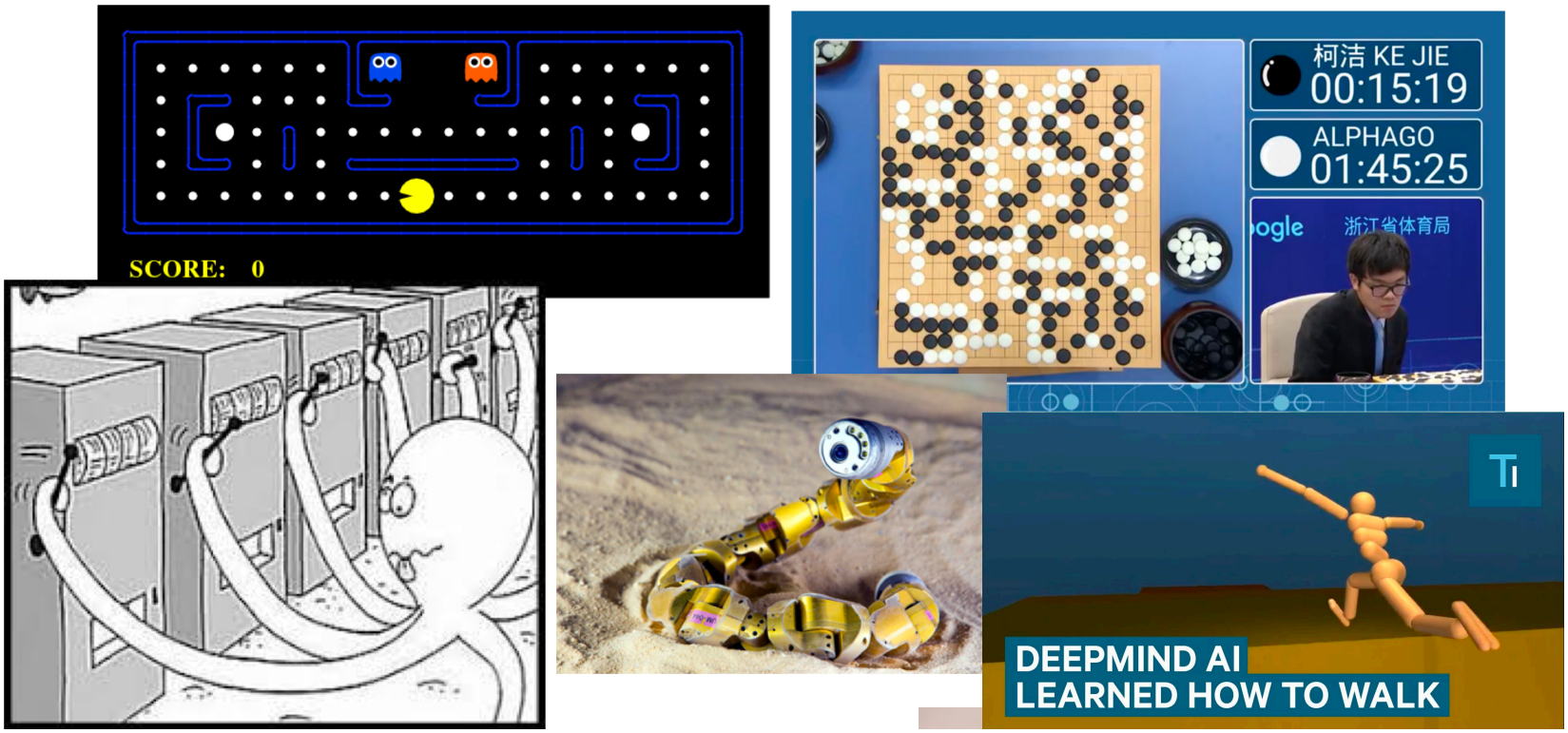
Yu-Xiang Wang

Based on the work of my student

Ming Yin: **On the job market!**

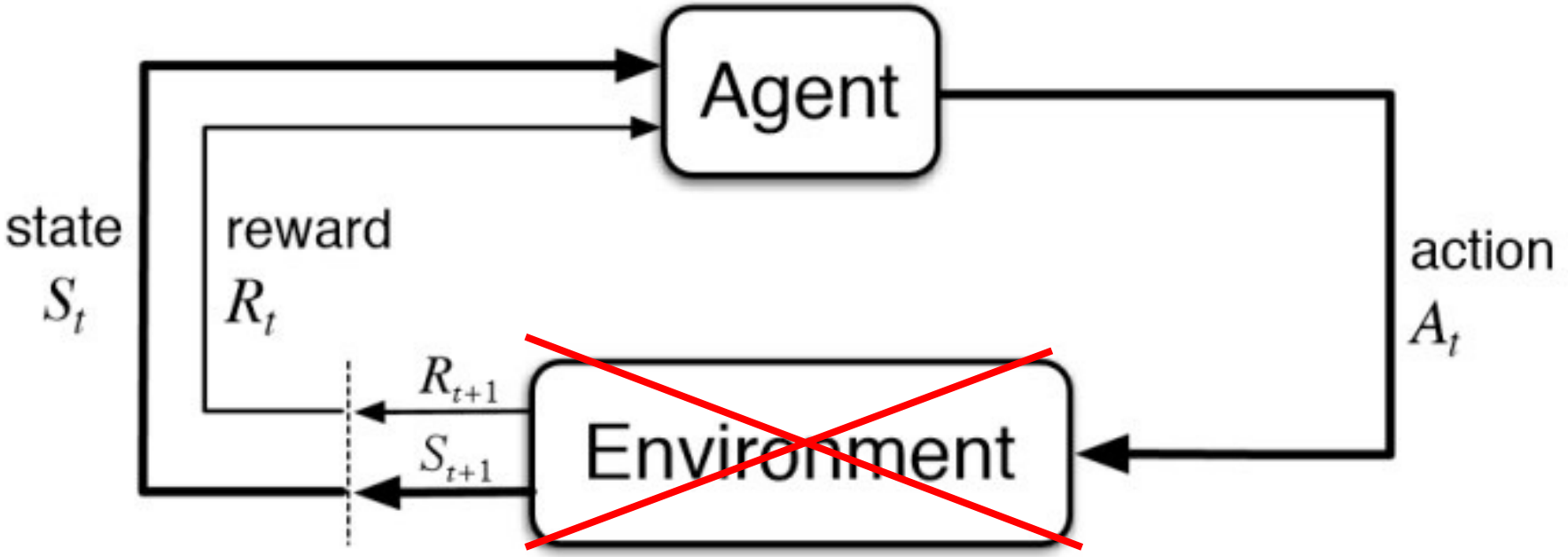
COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Reinforcement learning is among the hottest area of research in ML!



"RL" is Top 1 Keyword at NeurIPS'2021, appearing 199 times
"Deep Learning" only 129 times [source]

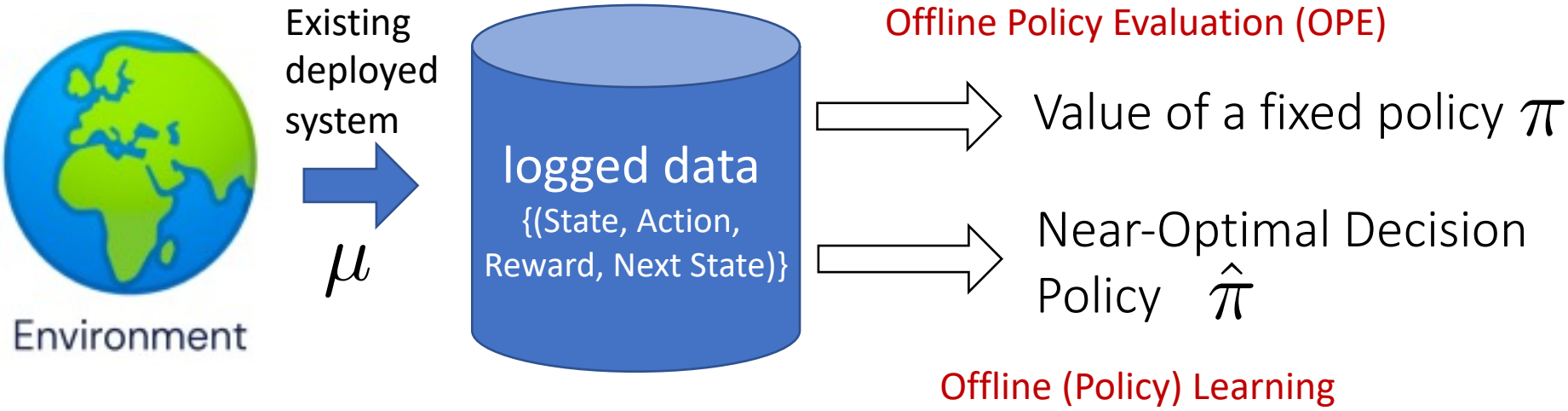
In real-life applications, we have limited access to the environment.



- Exploration is often costly, unsafe, illegal, …
- "Drive off road and crash the car to learn it's a bad idea"

RL in practice always starts with an existing dataset => Offline RL

- Two typical tasks are: OPE and Offline Learning



*Notation: $v^\pi := \mathbb{E}_\pi[\text{Total Reward}]$ $\quad$ Optimal policy $\pi^* := \arg\max_\pi v^\pi$

Example: Medical treatment

Offline Policy Evaluation (OPE)

Evaluate a new treatment plan.

Learn better treatment plan from data?

Offline (Policy) Learning

Illegal, unethical to “trial-and-error” on the fly with online RL.
But electronic patient records / and other health data from human doctors’ treatments are available.

Other applications of Offline RL

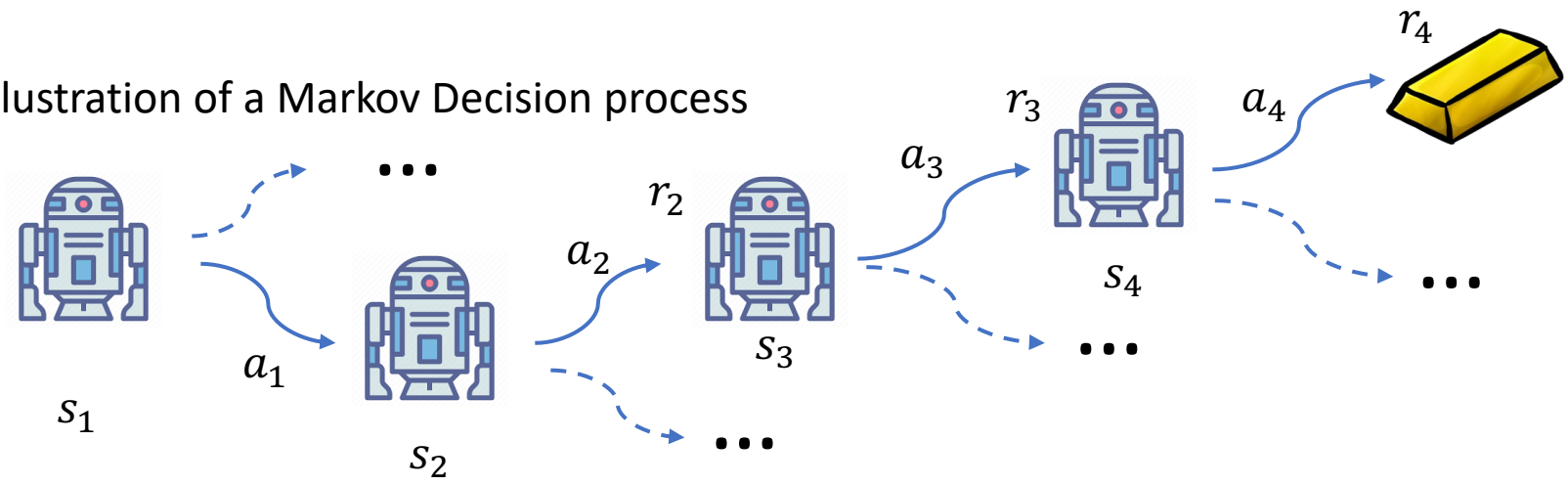
- Self-driving car:
  - Cannot deploy untested self-driving algorithms
  - Large amount data with human driver in control available.
- Ads / Recommendation systems
  - Expensive / risky to run online experiments
  - A lot of offline click-through data exists.
- New material discovery (UCSB IDEAS institute)
  - A lot of existing data (or related but different materials)
  - Easy to parallelize the experiments, but hard to have many iterations
- Computer networking (UCSB RELIEF project)
  - Bandwidth allocation on network devices
  - Quality of experience measurements.
- Many more from the INFORMS community!

Offline RL is very challenging!

- Oliver never played before, but watched hundreds of Joshua's games. Can Oliver play "optimally" the first time he tries?
- What if Joshua is a lousy player?

The fundamental difficulties of offline RL is to answer **"What if "** questions with **observational data.**

Illustration of a Markov Decision process



- Whenever Oliver plays differently he will run into different situations not seen by Joshua! But Oliver *has to* play differently to play better.

- Oliver need to infer using Joshua's data:

  "what happens if he chooses a path that Joshua did not take?"

The problem is more challenging than standard causal inference because:

- Long planning horizon
  - Naïve Importance Sampling suffer exponential variance!
- Large state-space
  - Size of the state-space is astronomically large
- Poor data coverage
  - Joshua may not visit high-reward states that the optimal policy visits
- Need to optimize rather than just estimate
  - How to handle uncertainty?

Remainder of the talk

1. Our results on offline RL
2. RL with low-switching cost

Ming Yin: **On the job market!**

Dan Qiao (2nd year PhD)

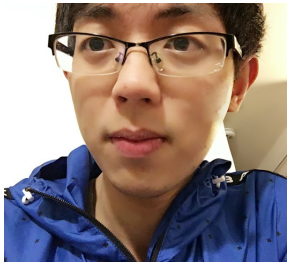
Also contributions from Mengdi Wang, Yaqi Duan and Yu Bai.

We consider finite horizon, episodic, tabular MDP model

- S states, A actions, horizon H.
  - time-inhomogeneous transitions.
  - 0< Reward < 1 (per step)
- Logging policy, Target policy, optimal policy:

  μ π π^*
- Number of trajectories: n
- Each policy induces a visitation measure $d_t^\pi(s,a)$
- Value of a policy $v^\pi = \mathbb{E}_\pi \left[\sum_{t=1}^{H} r_t \right].$

Our results on offline RL



Ming Yin

- Optimal {OPE, uniform OPE, offline RL, offline reward-free RL, offline RL with linear function approx…}

Optimal bound for OPE

$$\mathbb{E}[(\hat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi})^2] \le \frac{1}{n}\sum_{h=0}^{H}\sum_{s_h,a_h}\frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\cdot\mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)})\middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h\right] + O(n^{-1.5})$$

Or if in a simplified expression: $\left|\hat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi}\right| \asymp \sqrt{\frac{H^2}{n\, d_m^{\mu}}}$

(Xie, Ma & W., NeurIPS'19)

(Yin & W., AISTATS-20)

Optimal bound for Offline Learning via local Uniform OPE

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \hat{v}^{\pi}_{\mathrm{TMIS}}$$

$$v^{\pi^*} - v^{\hat{\pi}} \lesssim \sqrt{\frac{H^3}{n\, d_m^{\mu}}}$$

(Yin, Bai & W., AISTATS'21)

*Uniform coverage condition: $d_m^{\mu} := \min_{t,s,a} d_t^{\mu}(s,a)$

Per-instance optimal offline learning?

Ming Yin

Results under different exploration assumptions and special properties of the MDPs.

Uniform Visitation

$$\widetilde{O}\left(\sqrt{\frac{H^3}{n \cdot d_m}}\right)$$

(Yin, Bai & W., 2021)

Single Concentrability

$$\widetilde{O}\left(\sqrt{\frac{H^3 S C^*}{n}}\right)$$

(RZMJR 2021)

Adaptive Domain

$$\widetilde{O}\left(\sum_{h=1}^{H}\sqrt{\frac{\mathbb{Q}_h^*}{n \cdot d_m}}\right) + \widetilde{O}\left(\frac{H^3}{n \cdot d_m}\right)$$

(Zanette and Brunskill, 2019)

Per-instance optimal offline learning?

Ming Yin

"Pessimism is all you need" (Yin and W., NeurIPS-21)

Intrinsic Offline Learning Bound

$$\sum_{h=1}^{H}\sum_{s_h,a_h} d_h^{\pi^*}(s_h,a_h)\sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^*+r_h)}{d_h^{\mu}(s_h,a_h)}}\cdot\sqrt{\frac{1}{n}} \quad +\tilde{O}\left(\frac{1}{nd_m}\right)$$

Uniform Visitation

$$\tilde{O}\left(\sqrt{\frac{H^3}{n\cdot d_m}}\right)$$

(Yin, Bai & W., 2021)

Single Concentrability

$$\tilde{O}\left(\sqrt{\frac{H^3SC^*}{n}}\right)$$

(RZMJR 2021)

Adaptive Domain

$$\tilde{O}\left(\sum_{h=1}^{H}\sqrt{\frac{\mathbb{Q}_h^*}{n\cdot d_m}}\right)+\tilde{O}\left(\frac{H^3}{n\cdot d_m}\right)$$

(Zanette and Brunskill, 2019)

Strongest (most adaptive) result in offline RL to date!

What if the optimal policy visits states never seen in the data?

- Lazy answer: "Optimal policy not measurable"

- "Maybe we could still learn something?"

$$v^{\hat{\pi}} \geq \max_{\pi} v^{\pi} - \tilde{O}\left(\sum_{h=1}^{H} \sum_{s_h, a_h} d_h^{\pi}(s_h, a_h) \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^{\pi} + r_h)}{d_h^{\mu}(s_h, a_h)}} \cdot \sqrt{\frac{1}{n}} \right) - \tilde{O}\left(\frac{1}{n\, d_m}\right)$$

Pessimism VI

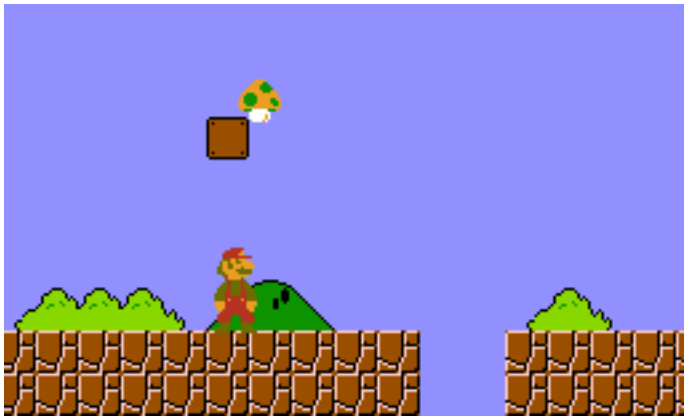
Arbitrary comparator policy

Regret / performance difference

" Learn as much as we can. Identify the best policy identifiable! "

Function approximation allows the learner to generalize to unseen states

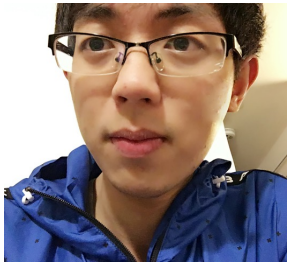
Joshua: “Red mushroom is good for me!”



Oliver a **new unseen state**:
“I haven’t seen Green Mushroom, but it must be good for me too?”

This is achieved by describing each state by a d-dimensional feature vector!

Pessimism works for linear function approximation too!



Ming Yin

- Setting: Linear MDP
- Idea: Pessimistic LSVI with variance weighting
- Coverage assumption: $\mathbb{E}_\mu \left[\phi\phi^T\right] \succ \kappa I$
- Result: instance dependent oracle inequality

$$v^{\pi} - v^{\widehat{\pi}} \leq \widetilde{O}\Big(\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\sqrt{\phi(\cdot,\cdot)^\top \Lambda_h^{-1} \phi(\cdot,\cdot)}\right]\Big) + \frac{2H^4\sqrt{d}}{K}$$

where $\Lambda_h = \sum_{k=1}^{K} \frac{\phi(s_h^k, a_h^k) \cdot \phi(s_h^k, a_h^k)^\top}{\sigma^2_{\widehat{V}_{h+1}(s_h^k, a_h^k)}} + \lambda I_d.$

Valid for any comparator policy at the same time!

Yin, Duan, Wang, W., "Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism", ICLR'22

New results on **nonlinear differentiable** function approx.

Ming Yin

Idea: Pessimistic Fitted Q-Iterations with Variance Reweighting

~~Feature vector~~ Gradients

$$v^{\pi} - v^{\widehat{\pi}} \leq \sum_{h=1}^{H} 8d \cdot \mathbb{E}_{\pi} \left[\sqrt{\nabla_{\theta}^{\top} f(\widehat{\theta}_h, \phi(s_h, a_h)) \Lambda_h^{-1} \nabla_{\theta} f(\widehat{\theta}_h, \phi(s_h, a_h))} \right] \cdot \iota + \widetilde{O}(\frac{\bar{C}_{\text{hot}}}{K}),$$

~~Covariance~~ "Fisher information"

Yin, Wang, W., "Offline Reinforcement Learning with Differentiable Function Approximation is Provably Efficient", Arxiv: https://arxiv.org/abs/2210.00750

Check point: Our solutions to the challenges in offline RL!

- Long planning horizon
  - Naïve Importance Sampling suffer exponential variance!
- Large state-space
  - Size of the state-space is astronomically large
- Poor data coverage
  - Logging policy may not visit high-reward states that the optimal policy visits
- Need to optimize rather than just estimate
  - How to handle uncertainty?

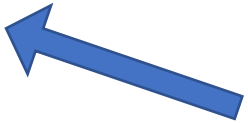
Marginalized IS achieves optimal rate.

(Xie, Ma & W., NeurIPS'19)

(Yin & W., AISTATS-20)

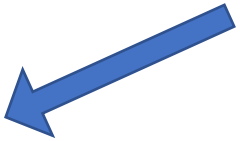
Function approximation: represent state by features.

(Yin, Duan, Wang, W., ICLR'22)
(Yin, Wang, W., 22)



"Pessimism is all you need"

(Yin and W., NeurIPS-21)
(Yin, Duan, Wang, W., ICLR'22)



Remainder of the talk

1. Advances in offline RL
2. RL with low-switching cost

Ming Yin: **On the job market!**

Dan Qiao (2nd year PhD)

Also contributions from Mengdi Wang, Yaqi Duan and Yu Bai.

Online RL vs Offline RL, revisited

	Online RL	Offline RL
Sample Complexity	$\tilde{O}\left(\frac{H^3SA}{\epsilon^2}\right)$	$\tilde{O}\left(\frac{H^3}{\epsilon^2 d_m}\right)$ or "Best effort learning" when d_m too small

Algorithmically enforce "Good Exploration"

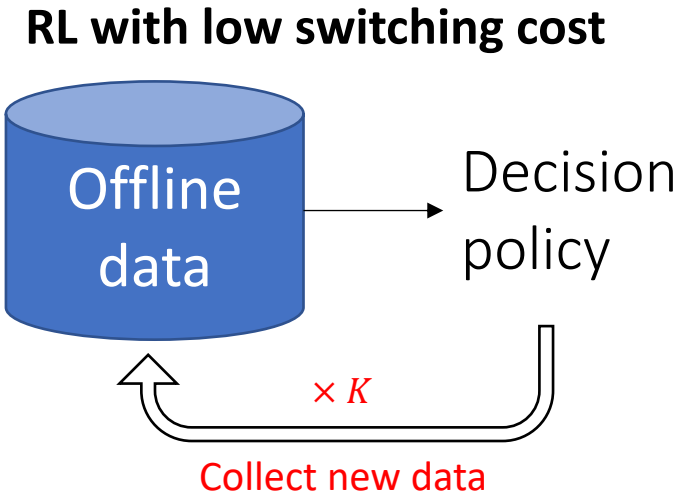
Assume "Good Exploration" or **weaken goal**

T rounds of adaptivity.
One per iteration!

1 rounds of adaptivity.

Anything in between?

Emerging new setting between online and offline RL



Can we solve exploration with a small number of policy changes?

Why does switching cost matter?

- Deployment of a new policy is costly!
  - We cannot afford to reflash the firmware too frequently
  - Cheaper to schedule experiments in parallel.
- Need extensive testing / approvals to actually deploy each new policy
  - A/B testing
  - IRB approval / ethics approval
- Low-switching cost is a desirable property for running experiments in parallel.

New result: K = O(HSAlog log T) is sufficient and necessary!

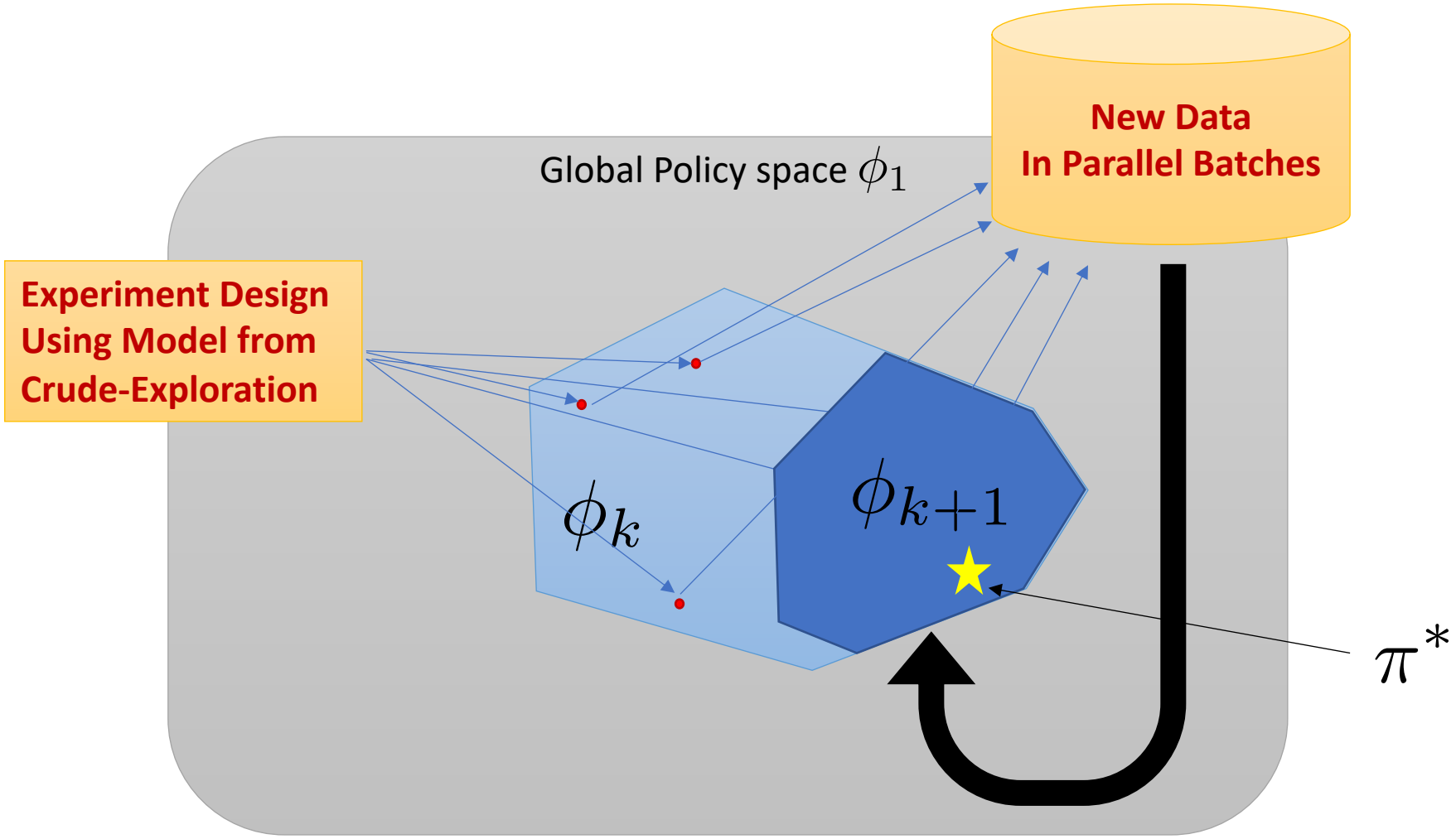
(among algorithms with sqrt(T) regret.)

Algorithms for regret minimization	*Regret*	*Switching cost*
UCB2-Bernstein [Bai et al., 2019]	$\widetilde{O}(\sqrt{H^3SAT})$	Local: $O(H^3SA\log T)$
UCB-Advantage [Zhang et al., 2020c]	$\widetilde{O}(\sqrt{H^2SAT})$	Local: $O(H^2SA\log T)$
Algorithm 1 in [Gao et al., 2021] *	$\widetilde{O}(\sqrt{d^3H^3T})$	Global: $O(dH\log T)$
APEVE (Our Algorithm 1)	$\widetilde{O}(\sqrt{H^4S^2AT})$	Global: $O(HSA\log\log T)$
Explore-First w. LARFE (Our Algorithm 4)	$\widetilde{O}(T^{2/3}H^{4/3}S^{2/3}A^{1/3})$	Global: $O(HSA)$
Lower bound (Our Theorem 4.2)	if $\widetilde{O}(\sqrt{T})$ ("Optimal regret")	Global: $\Omega(HSA\log\log T)$
Lower bound (Our Theorem 4.3)	if $o(T)$ ("No regret")	Global: $\Omega(HSA)$

- First of its kind with log-log T switching cost
- Information-theoretically optimal!
- No need to "assume" exploration

Qiao, Yin, Min, W., "Sample-Efficient Reinforcement Learning with loglog(T) Switching Cost", ICML'2021

Main challenge: Can't use optimism. need to solve exploration differently!



Summary of the main results

- Real-life RL imposes constraints on exploration.
- RL with offline data
  - Asymptotic efficiency for OPE using plug-in.
  - Pessimistic value iteration works even if the coverage is poor.. It competes with any policy (optimal or not!)
- RL with low-switching cost
  - Emerging new setting
  - Algorithmically enforce “exploration”, but retain deployment efficiency!

Future work and open problems

- RL with low switching cost / deployment efficiency in the function approximation regime
  - We have released a recent work under linear MDP
- Combine offline RL and low-adaptive exploration
  - Using offline data as a “LaunchPad”
- The compatibility of pessimism and efficient exploration? Can we get both?

Thank you for your attention!

References and co-authors:

Yin and W. **Asymptotically Efficient Off-Policy Evaluation for Tabular Reinforcement Learning**. In AISTATS 2020.

Yin, Bai and W. **Near Optimal Provable Uniform Convergence in Offline Policy Evaluation for Reinforcement Learning**. In AISTATS 2021

Yin and W. **Optimal Uniform OPE and Model-based Offline Reinforcement Learning in Time-Homogeneous, Reward-Free and Task-Agnostic Settings**. In NeurIPS 2021.

Yin and W. **Towards Instance-Optimal Offline Reinforcement Learning with Pessimism**. In NeurIPS 2021.

Yin, Duan, Wang, W., **Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism**, In ICLR 2022

Qiao, Yin, Min, W., **"Sample-Efficient Reinforcement Learning with loglog(T) Switching Cost",** ICML'2021

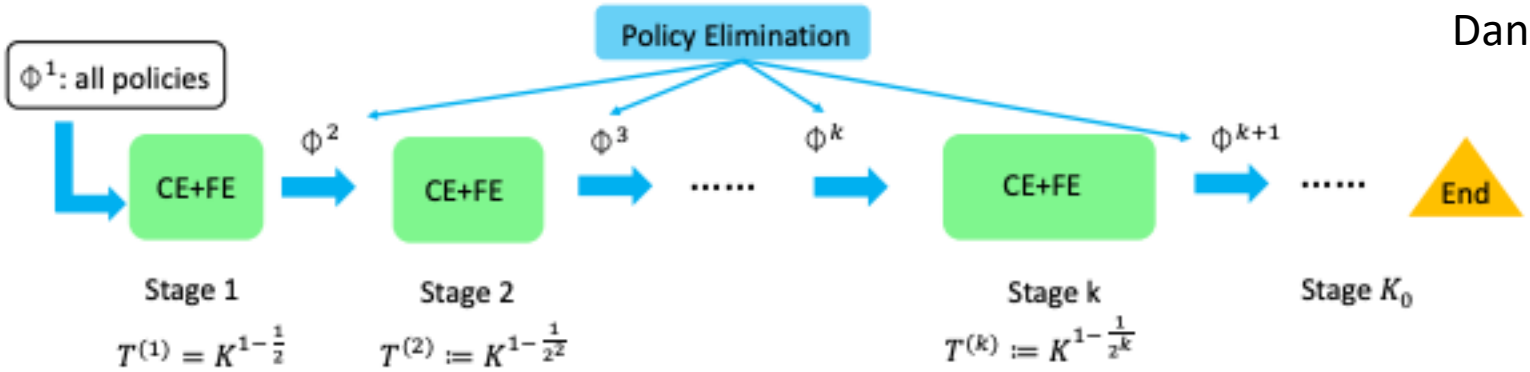




Supplementary Slides

Our approach: Policy Elimination With two-stage exploration

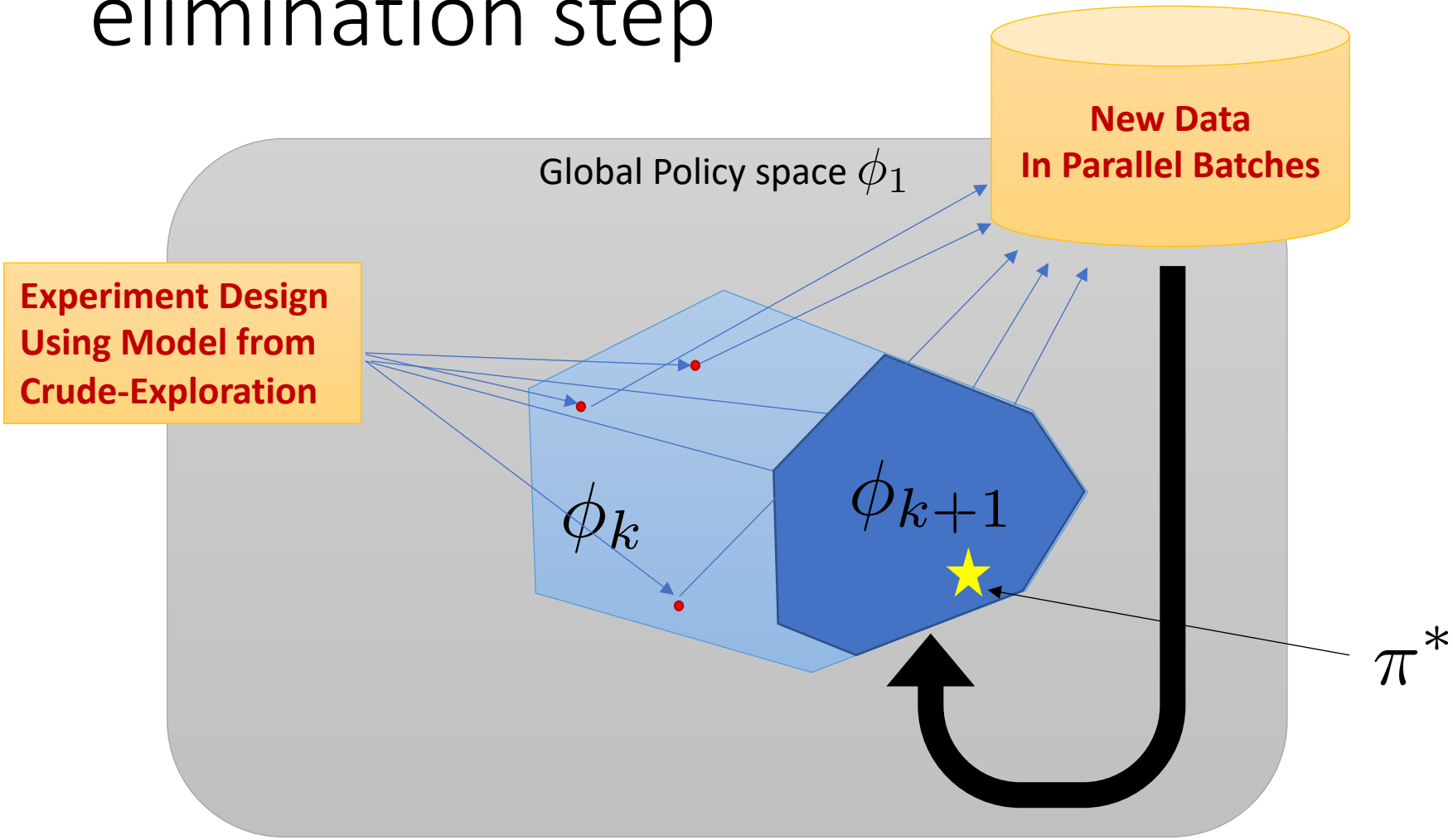
Dan Qiao



- Crude layer-wise exploration
  - Get a "ball park" approximation of the model
- Fine stagewise exploration
  - Use the crude model to identify "representative" policies
- Policy elimination
  - Disqualify policies that are certifiably suboptimal.

Qiao, Yin, Min, W., "Sample-Efficient Reinforcement Learning with loglog(T) Switching Cost", ICML'2021

Illustration of the policy elimination step



Tight Confidence bound

Results: Optimal Switching Cost with nearly optimal regret

Algorithms for regret minimization	*Regret*	*Switching cost*
UCB2-Bernstein [Bai et al., 2019]	$\widetilde{O}(\sqrt{H^3SAT})$	Local: $O(H^3SA\log T)$
UCB-Advantage [Zhang et al., 2020c]	$\widetilde{O}(\sqrt{H^2SAT})$	Local: $O(H^2SA\log T)$
Algorithm 1 in [Gao et al., 2021] *	$\widetilde{O}(\sqrt{d^3H^3T})$	Global: $O(dH\log T)$
APEVE (Our Algorithm 1)	$\widetilde{O}(\sqrt{H^4S^2AT})$	Global: $O(HSA\log\log T)$
Explore-First w. LARFE (Our Algorithm 4)	$\widetilde{O}(T^{2/3}H^{4/3}S^{2/3}A^{1/3})$	Global: $O(HSA)$
Lower bound (Our Theorem 4.2)	if $\widetilde{O}(\sqrt{T})$ ("Optimal regret")	Global: $\Omega(HSA\log\log T)$
Lower bound (Our Theorem 4.3)	if $o(T)$ ("No regret")	Global: $\Omega(HSA)$

- First of its kind with log-log T switching cost
- Information-theoretically optimal!
- No need to "assume" exploration