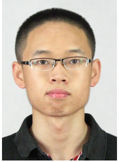
Statistical and Algorithmic Foundations of Reinforcement Learning

Yuting Wei

Statistics & Data Science, Wharton
University of Pennsylvania

JCSDS, Yunnan, 2024

Our wonderful collaborators



Gen Li

UPenn $\rightarrow$ CUHK

Shicong Cen

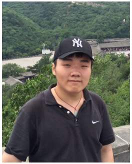
CMU

Laixi Shi

CMU $\rightarrow$ Caltech

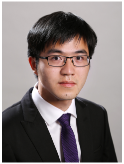
Chen Cheng

Stanford



Yuling Yan

Princeton $\rightarrow$ MIT



Changxiao Cai

UPenn $\rightarrow$ UMich

Matthieu Geist

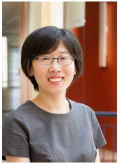
Google $\rightarrow$ Cohere

Jianqing Fan

Princeton

Yuxin Chen

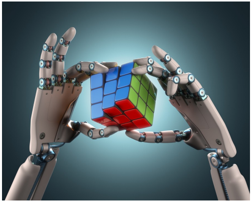
UPenn

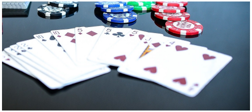


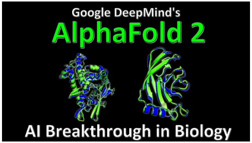
Yuejie Chi

CMU

Recent successes in reinforcement learning (RL)







RL holds great promise in the next era of artificial intelligence.

Recap: Supervised learning

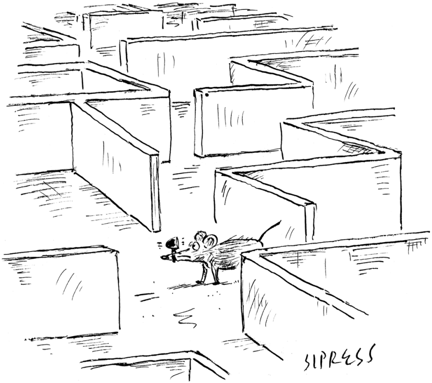
Given i.i.d training data, the goal is to make prediction on unseen data:

— *pic from internet*

Reinforcement learning (RL)

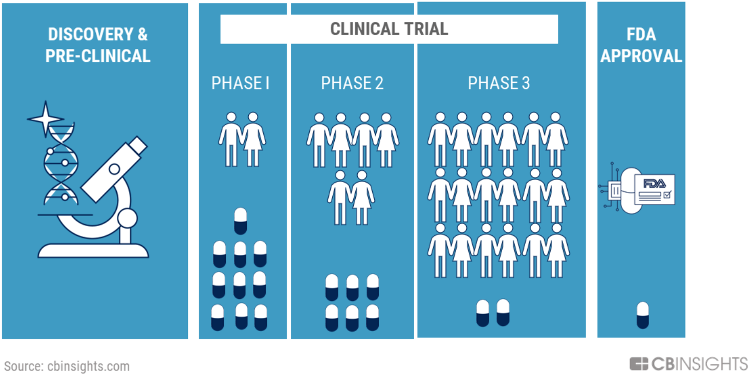
In RL, an agent learns by interacting with an environment.

- no training data
- trial-and-error
- maximize total rewards
- delayed reward



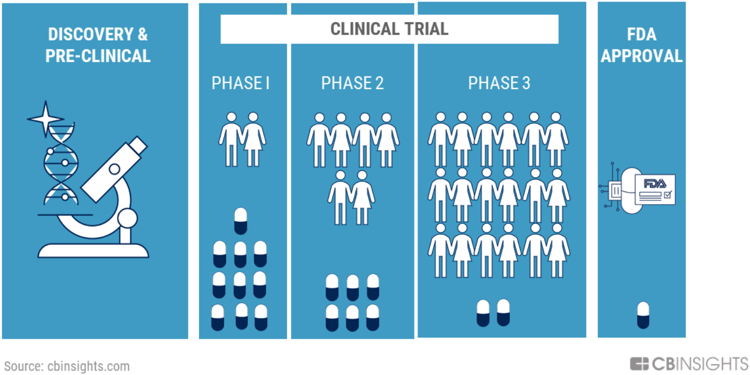
"Recalculating ... recalculating ..."

Sample efficiency



- prohibitively large state & action space
- collecting data samples can be expensive or time-consuming

Sample efficiency



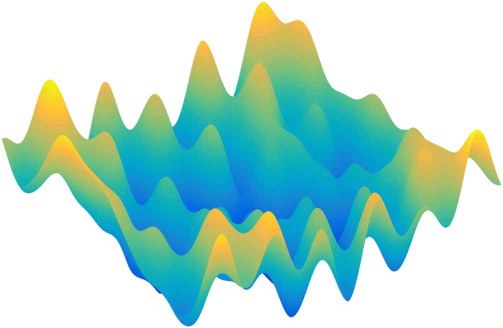
- prohibitively large state & action space
- collecting data samples can be expensive or time-consuming

Challenge: design sample-efficient RL algorithms

Computational efficiency

Running RL algorithms might take a long time ...

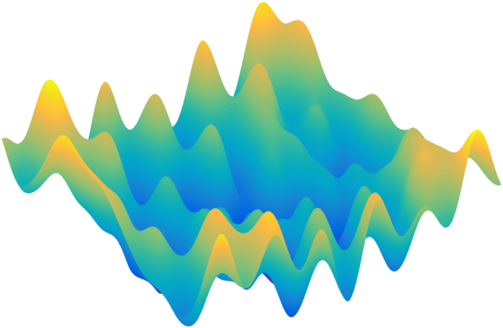
- enormous state-action space
- nonconvexity



Computational efficiency

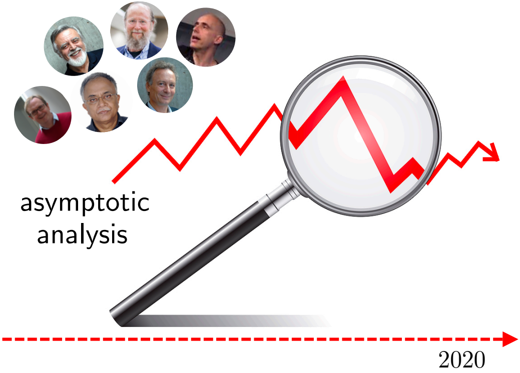
Running RL algorithms might take a long time . . .

- enormous state-action space
- nonconvexity

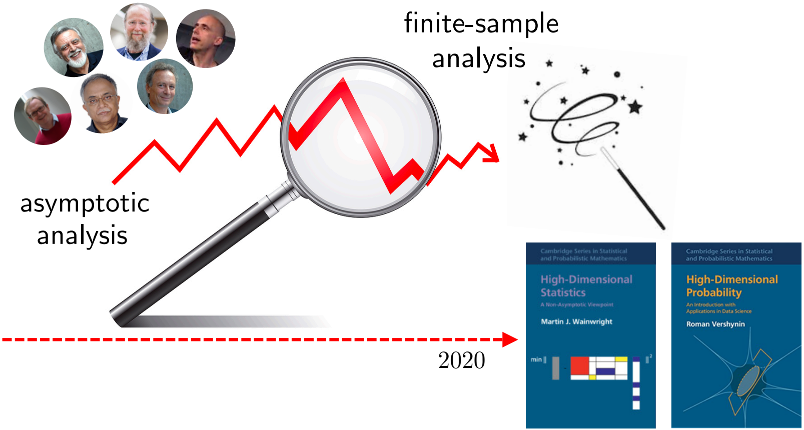


Challenge: design computationally efficient RL algorithms

Theoretical foundation of RL

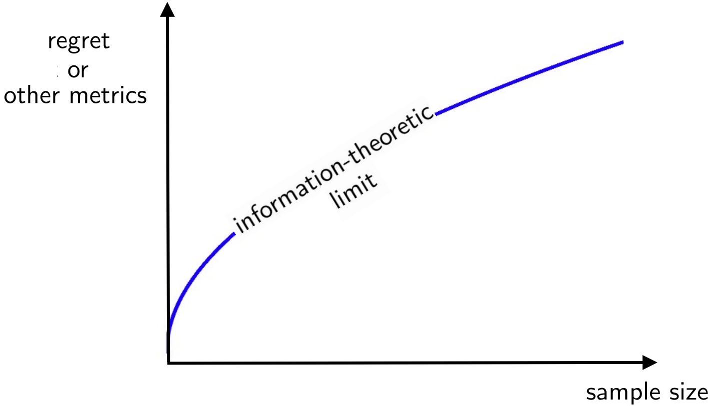


Theoretical foundation of RL

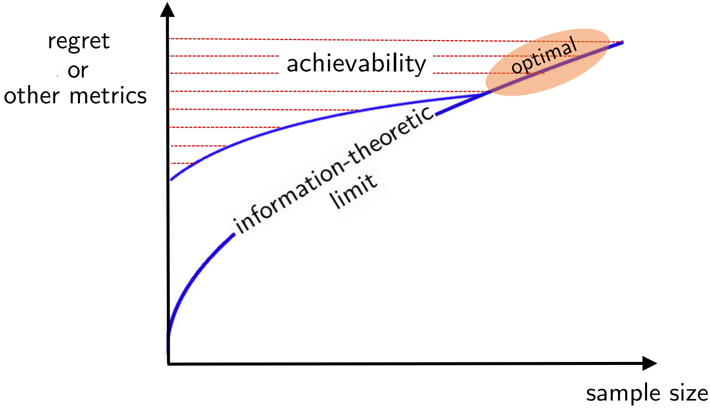


Understanding sample efficiency of RL requires a modern suite of non-asymptotic analysis tools

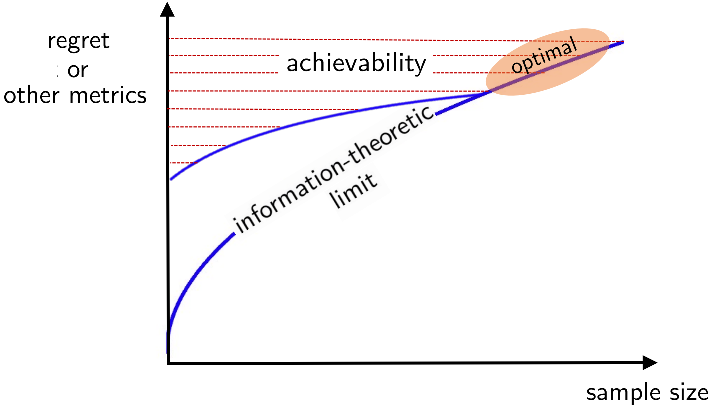
Sample complexity issues that permeate state-of-the-art RL theory



Sample complexity issues that permeate state-of-the-art RL theory

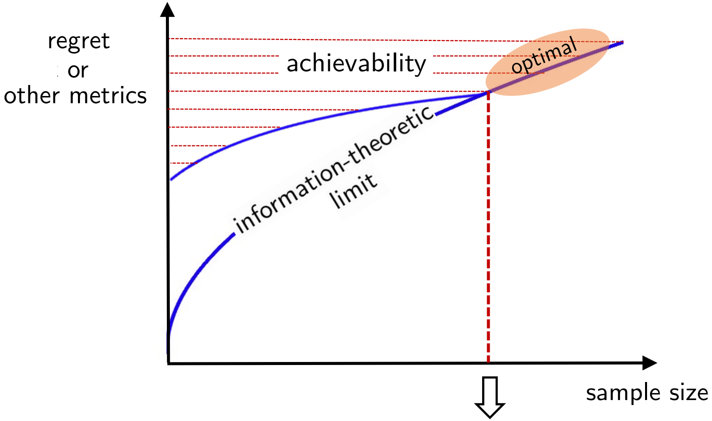


Sample complexity issues that permeate state-of-the-art RL theory



high-dimensional statistics ⟺ large-sample theory

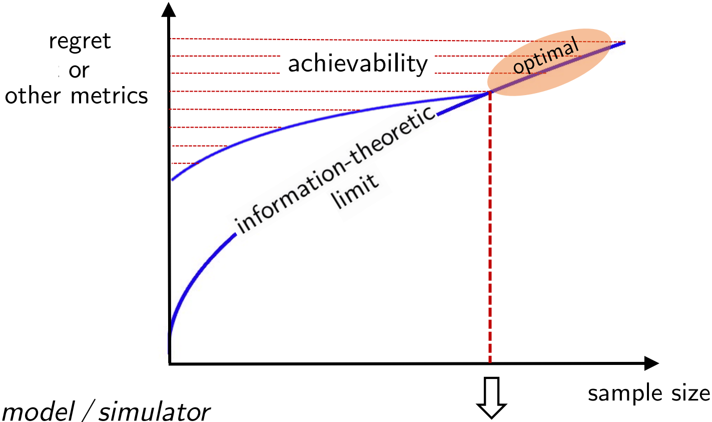
Sample complexity issues that permeate state-of-the-art RL theory



huge burn-in cost!

high-dimensional statistics ⟺ large-sample theory

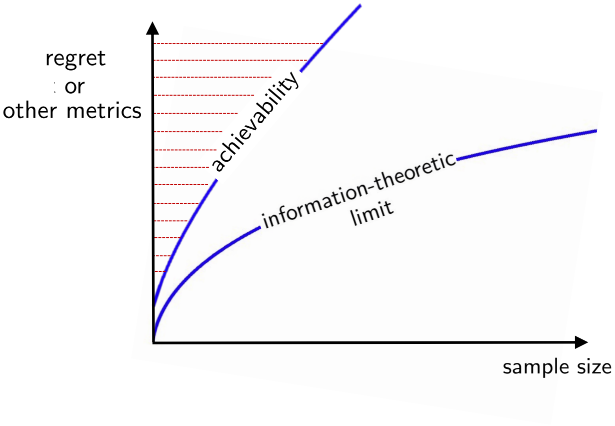
Sample complexity issues that permeate state-of-the-art RL theory



huge burn-in cost!

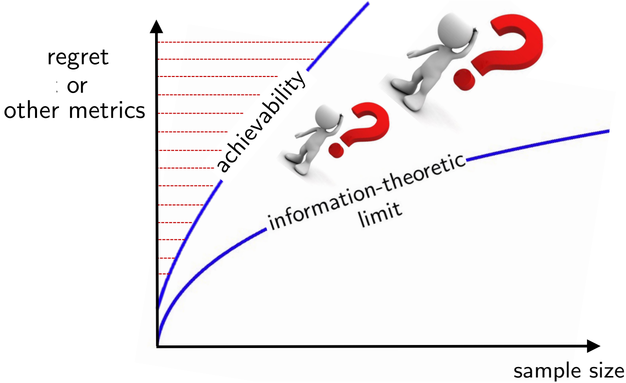
- *generative model / simulator*
- *online RL*
- *offline RL*
- ...

Sample complexity issues that permeate state-of-the-art RL theory



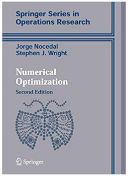
- *multi-agent RL*
- *partially observable MDPs*
- ...

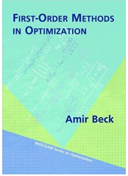
Sample complexity issues that permeate state-of-the-art RL theory

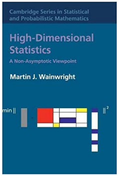


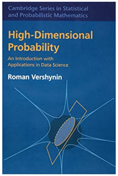
- *multi-agent RL*
- *partially observable MDPs*
- ...

This tutorial







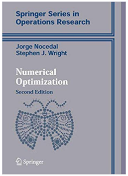


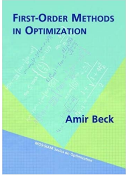
(large-scale) optimization

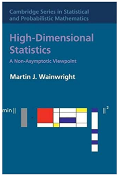
(high-dimensional) statistics

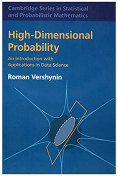
Design sample- and computationally-efficient RL algorithms

This tutorial









(large-scale) optimization

(high-dimensional) statistics

Design sample- and computationally-efficient RL algorithms

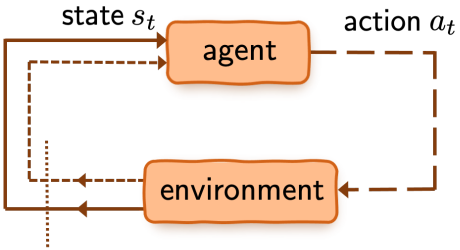
Part 1. basics, RL w/ a generative model

Part 2. online / offline RL, multi-agent / robust RL

Part 1

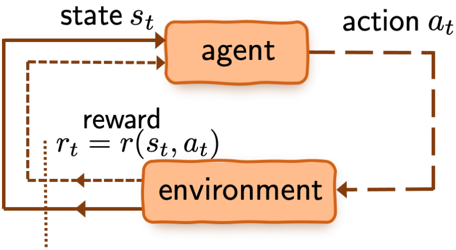
1. Basics: Markov decision processes
2. RL w/ a generative model (simulator)
   - model-based algorithms (a “plug-in” approach)
   - model-free algorithms

Markov decision process (MDP)



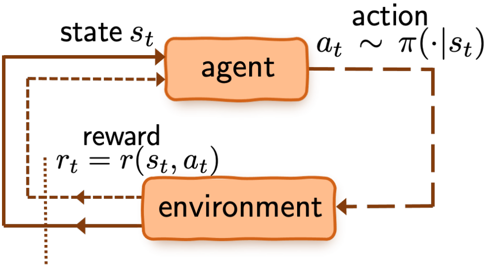
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

Markov decision process (MDP)



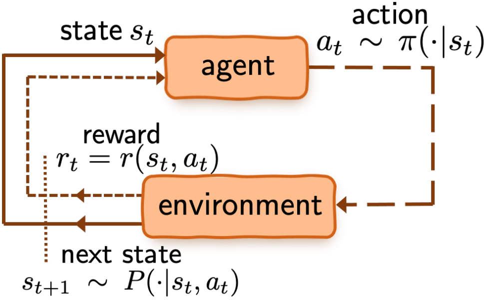
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

Infinite-horizon Markov decision process



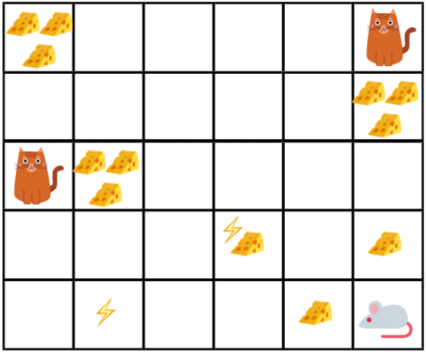
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

Infinite-horizon Markov decision process

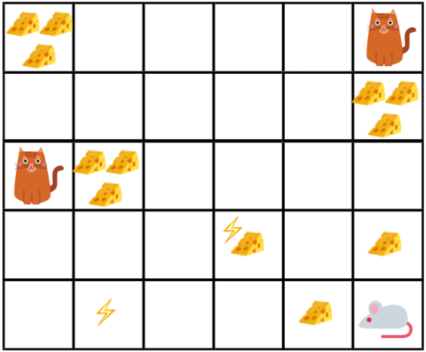


- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)
- $P(\cdot|s,a)$: unknown transition probabilities

Help the mouse!

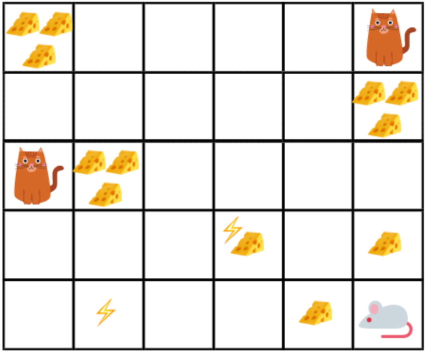


Help the mouse!



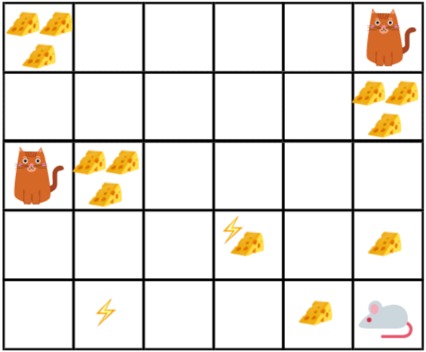
- state space $\mathcal{S}$: positions in the maze

Help the mouse!



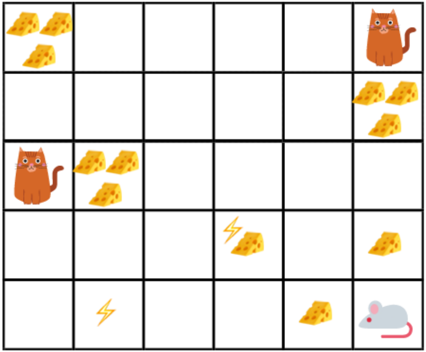
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right

Help the mouse!



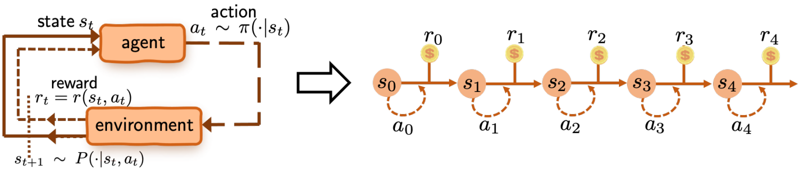
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right
- immediate reward r: cheese, electricity shocks, cats

Help the mouse!



- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right
- immediate reward r: cheese, electricity shocks, cats
- policy $\pi(\cdot|s)$: the way to find cheese

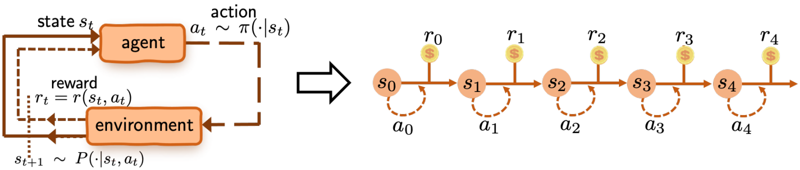
Value function



Value of policy π: cumulative discounted reward

$$\forall s \in \mathcal{S} : \quad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$$

Value function



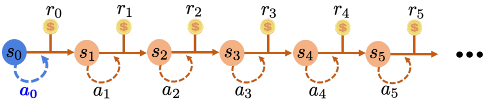
Value of policy π: cumulative discounted reward

$$\forall s \in \mathcal{S}: \quad V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\Big|\, s_0 = s\right]$$

- $\gamma \in [0, 1)$: discount factor
  - take $\gamma \to 1$ to approximate long-horizon MDPs
  - **effective horizon**: $\frac{1}{1-\gamma}$

Q-function (action-value function)

$Q^\pi(s_0, a_0)$

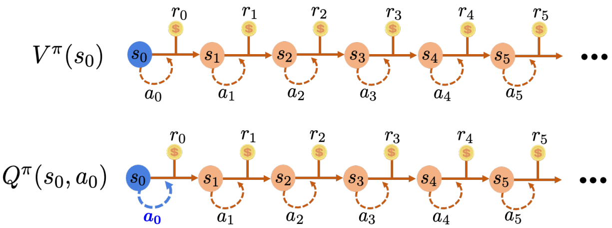


Q-function of policy π:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, \textcolor{red}{a_0 = a}\right]$$

- $(\cancel{a_0,}\ s_1, a_1, s_2, a_2, \cdots)$: induced by policy π

Q-function (action-value function)

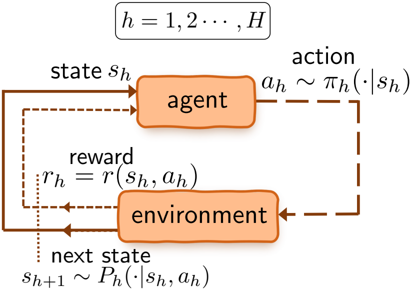


Q-function of policy π:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

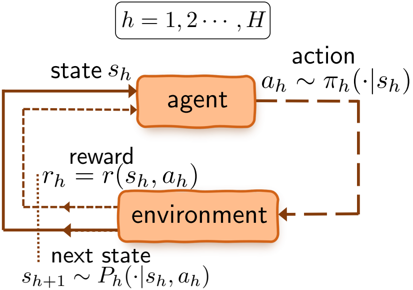
- $(\cancel{a_0}, s_1, a_1, s_2, a_2, \cdots)$: induced by policy π

Finite-horizon MDPs



- H: horizon length
- $\mathcal{S}$: state space with size S
- $\mathcal{A}$: action space with size A
- $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h
- $\pi = \{\pi_h\}_{h=1}^{H}$: policy (or action selection rule)
- $P_h(\cdot \mid s, a)$: transition probabilities in step h

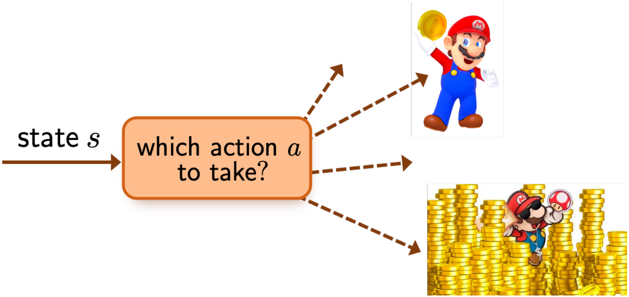
Finite-horizon MDPs



value function: $V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^{H} r_h(s_h, a_h) \,\middle|\, s_h = s\right]$

Q-function: $Q_h^\pi(s, a) := \mathbb{E}\left[\sum_{t=h}^{H} r_h(s_h, a_h) \,\middle|\, s_h = s, a_h = a\right]$

Optimal policy and optimal value



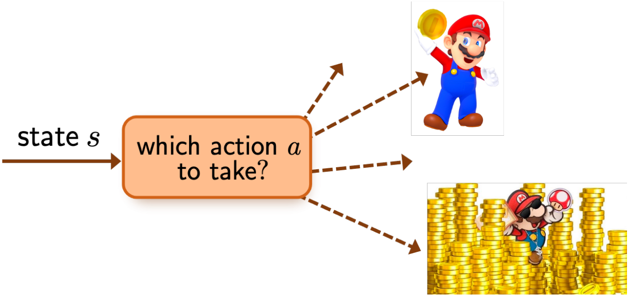
- **optimal policy** $\pi^\star$: maximizing value function $\max_\pi V^\pi$

Proposition (Puterman'94)

For infinite horizon discounted MDP, there always exists a deterministic policy $\pi^\star$*, such that*

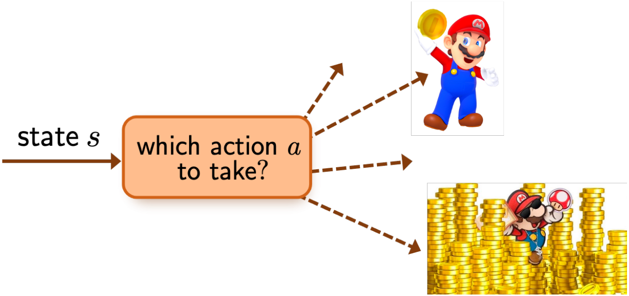
$$V^{\pi^\star}(s) \geq V^\pi(s), \quad \forall s, \text{ and } \pi.$$

Optimal policy and optimal value



- **optimal policy** $\pi^\star$: maximizing value function $\max_\pi V^\pi$
- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$

Optimal policy and optimal value



- **optimal policy** $\pi^\star$: maximizing value function $\max_\pi V^\pi$
- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$
- How to find this $\pi^\star$?

Basic dynamic programming algorithms when MDP specification is known

Policy evaluation: Given MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \gamma)$ and policy $\pi : \mathcal{S} \mapsto \mathcal{A}$, how good is π? (i.e., how to compute $V^\pi(s), \ \forall s$?)

Policy evaluation: Given MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \gamma)$ and policy $\pi : \mathcal{S} \mapsto \mathcal{A}$, how good is π? (i.e., how to compute $V^\pi(s), \ \forall s$?)

Possible scheme:

- execute policy evaluation for each π
- find the optimal one

Policy evaluation: Bellman's consistency equation

- $V^{\pi}\,/\,Q^{\pi}$: value / action-value function under policy π

Policy evaluation: Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\big[Q^\pi(s,a)\big]$$

$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}}\Big]$$

Richard Bellman

Policy evaluation: Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$
$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}}\Big]$$

- one-step look-ahead

Richard Bellman

Policy evaluation: Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$

$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}}\Big]$$

- one-step look-ahead
- let P^π be the state-action transition matrix induced by π:

$$Q^\pi = r + \gamma P^\pi Q^\pi \quad \Longrightarrow \quad Q^\pi = (I - \gamma P^\pi)^{-1} r$$

Richard Bellman

Optimal policy $\pi^\star$: Bellman's optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\left[\max_{a' \in \mathcal{A}} Q(s',a')\right]}_{\text{next state's value}}$$

- one-step look-ahead

Optimal policy $\pi^\star$: Bellman's optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\left[\max_{a' \in \mathcal{A}} Q(s',a')\right]}_{\text{next state's value}}$$

- one-step look-ahead

Bellman equation: $Q^\star$ is *unique* solution to

$$\mathcal{T}(Q^\star) = Q^\star$$

γ-contraction of Bellman operator:

$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$$

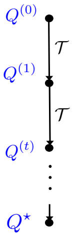
Richard Bellman

Two dynamic programming algorithms

Value iteration (VI)

For $t = 0, 1, \ldots,$

$$Q^{(t+1)} = \mathcal{T}(Q^{(t)})$$

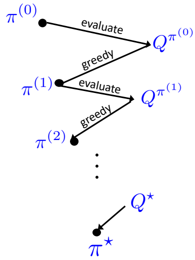


Policy iteration (PI)

For $t = 0, 1, \ldots,$

policy evaluation: $Q^{(t)} = Q^{\pi^{(t)}}$

policy improvement: $\pi^{(t+1)}(s) = \arg\max_{a \in \mathcal{A}} Q^{(t)}(s, a)$



Iteration complexity

Theorem (Linear convergence of policy/value iteration)

$$\|Q^{(t)} - Q^\star\|_\infty \le \gamma^t \|Q^{(0)} - Q^\star\|_\infty$$

Iteration complexity

Theorem (Linear convergence of policy/value iteration)

$$\|Q^{(t)} - Q^{\star}\|_{\infty} \leq \gamma^{t} \|Q^{(0)} - Q^{\star}\|_{\infty}$$

Implications: to achieve $\|Q^{(t)} - Q^{\star}\|_{\infty} \leq \varepsilon$, it takes no more than

$$\frac{1}{1-\gamma} \log \left(\frac{\|Q^{(0)} - Q^{\star}\|_{\infty}}{\varepsilon} \right) \quad \text{iterations}$$

Iteration complexity

Theorem (Linear convergence of policy/value iteration)

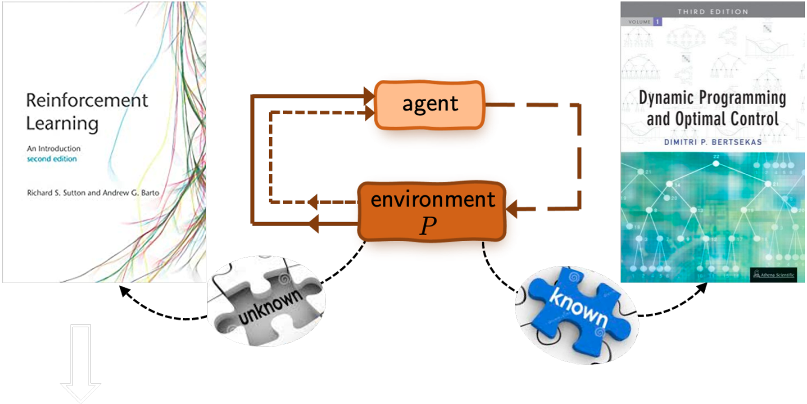
$$\|Q^{(t)} - Q^\star\|_\infty \le \gamma^t \|Q^{(0)} - Q^\star\|_\infty$$

Implications: to achieve $\|Q^{(t)} - Q^\star\|_\infty \le \varepsilon$, it takes no more than

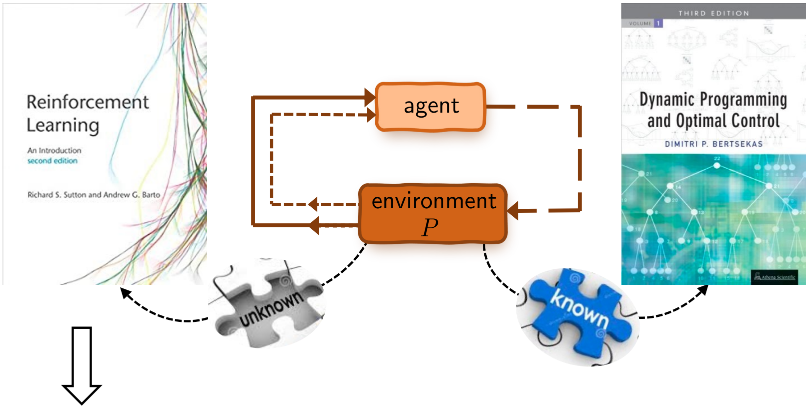
$$\frac{1}{1-\gamma} \log\left(\frac{\|Q^{(0)} - Q^\star\|_\infty}{\varepsilon}\right) \quad \text{iterations}$$

Linear convergence at a **dimension-free** rate!

When the model is unknown . . .

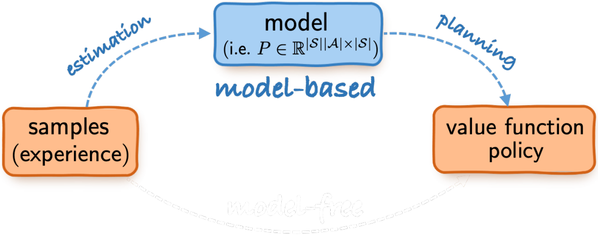


When the model is unknown . . .



Need to learn optimal policy from samples w/o model specification

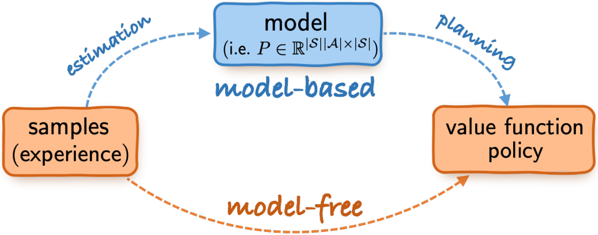
Two approaches



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

Two approaches



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

Model-free approach

— learning w/o estimating the model explicitly

Sampling mechanisms

1. RL w/ a generative model (a.k.a. simulator)
   - can query arbitrary state-action pairs to draw samples

Sampling mechanisms

1. RL w/ a generative model (a.k.a. simulator)
   - can query arbitrary state-action pairs to draw samples

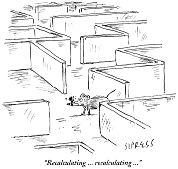
2. online RL
   - execute MDP in real time to obtain sample trajectories

Sampling mechanisms

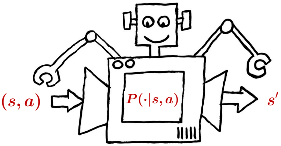
1. RL w/ a generative model (a.k.a. simulator)
   - can query arbitrary state-action pairs to draw samples
2. online RL
   - execute MDP in real time to obtain sample trajectories
3. offline RL
   - use pre-collected historical data

Exploration vs exploitation

offline RL



online RL

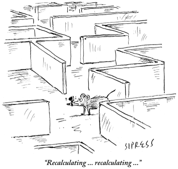


generative model

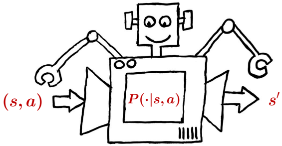
Exploration vs exploitation

Exploration

offline RL



online RL



generative model

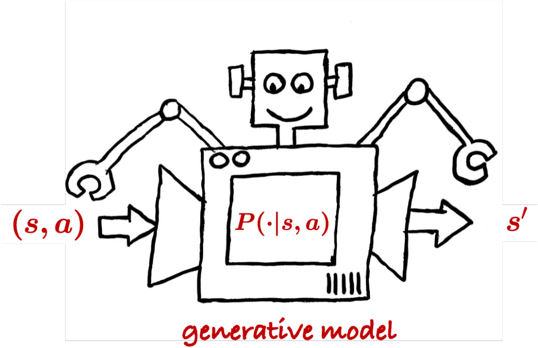
Varying levels of trade-offs between exploration and exploitation.

Part 1

1. Basics: Markov decision processes
2. RL w/ a generative model (simulator)
   - model-based algorithms (a “plug-in” approach)
   - model-free algorithms

A generative model / simulator

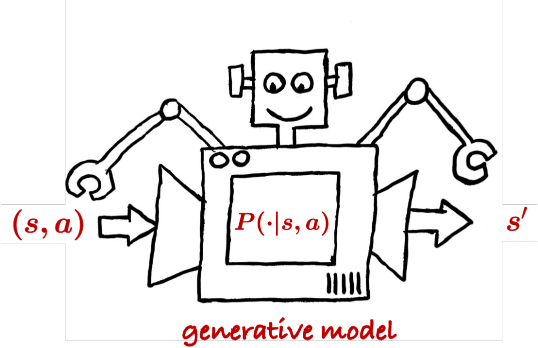
— *Kearns and Singh, 1999*



- **sampling:** for each (s,a), collect N samples $\{(s,a,s'_{(i)})\}_{1\le i\le N}$

A generative model / simulator

— *Kearns and Singh, 1999*



- **sampling:** for each (s, a), collect N samples $\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}$
- construct $\widehat{\pi}$ based on samples (in total $|\mathcal{S}||\mathcal{A}| \times N$)

ℓ_∞-**sample complexity:** how many samples are required to learn an $\underbrace{\varepsilon\text{-optimal policy}}_{\forall s:\ V^{\hat{\pi}}(s) \geq V^{\star}(s) - \varepsilon}$?

An incomplete list of works

- Kearns and Singh, 1999
- Kakade, 2003
- Kearns et al., 2002
- Azar et al., 2012
- Azar et al., 2013
- Sidford et al., 2018a, 2018b
- Wang, 2019
- Agarwal et al., 2019
- Wainwright, 2019a, 2019b
- Pananjady and Wainwright, 2019
- Yang and Wang, 2019
- Khamaru et al., 2020
- Mou et al., 2020
- Cui and Yang, 2021
- ...

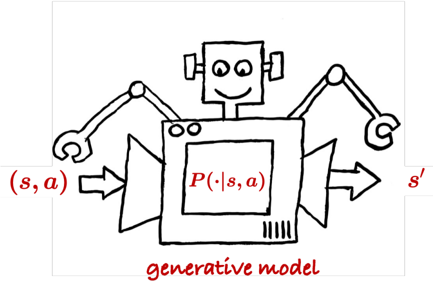
An even shorter list of prior art

algorithm	sample size range	sample complexity	ε-range
Empirical QVI Azar et al., 2013	$\big[\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$\big(0, \frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}}\big]$
Sublinear randomized VI Sidford et al., 2018b	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$	$\big(0, \frac{1}{1-\gamma}\big]$
Variance-reduced QVI Sidford et al., 2018a	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, 1]$
Randomized primal-dual Wang 2019	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$	$\big(0, \frac{1}{1-\gamma}\big]$
Empirical MDP + planning Agarwal et al., 2019	$\big[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\big)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$\big(0, \frac{1}{\sqrt{1-\gamma}}\big]$

important parameters $\Longrightarrow$

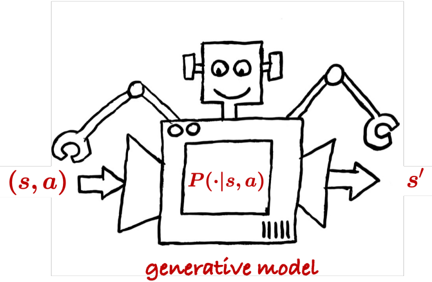
- # states $|\mathcal{S}|$, # actions $|\mathcal{A}|$
- the discounted complexity $\frac{1}{1-\gamma}$
- approximation error $\varepsilon \in (0, \frac{1}{1-\gamma}]$

Model estimation



Sampling: for each (s, a), collect N ind. samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$

Model estimation



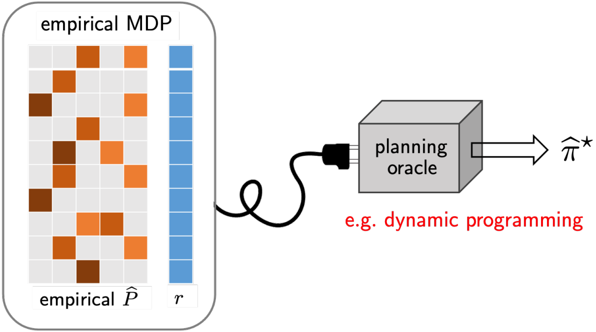
Sampling: for each (s, a), collect N ind. samples $\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}$

Empirical estimates:

$$\widehat{P}(s'|s, a) = \underbrace{\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{s'_{(i)} = s'\}}_{\text{empirical frequency}}$$

Empirical MDP + planning

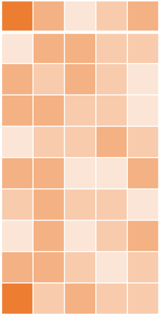
— Azar et al., 2013, Agarwal et al., 2019



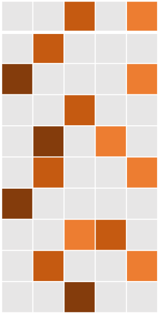
Find policy based on the empirical MDP (*empirical maximizer*)

using, e.g., policy iteration — $(\widehat{P}, r)$

Challenges in the sample-starved regime



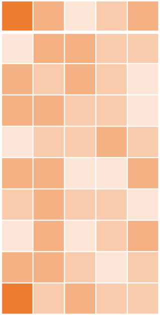
truth: $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$



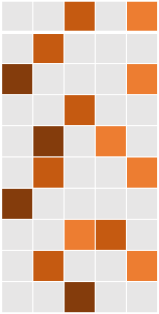
empirical estimate: $\widehat{P}$

- Can't recover P faithfully if sample size $\ll |\mathcal{S}|^2|\mathcal{A}|$!

Challenges in the sample-starved regime



truth: $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



empirical estimate: $\widehat{P}$

- Can't recover P faithfully if sample size $\ll |\mathcal{S}|^2|\mathcal{A}|$!
- Can we trust our policy estimate when reliable model estimation is infeasible?

ℓ_∞-based sample complexity

Theorem (Agarwal, Kakade, Yang '19)

For any $0 < \varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$, *the optimal policy* $\widehat{\pi}^\star$ *of empirical MDP achieves*

$$\|V^{\widehat{\pi}^\star} - V^\star\|_\infty \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

ℓ_∞-based sample complexity

Theorem (Agarwal, Kakade, Yang '19)

For any $0 < \varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$*, the optimal policy* $\widehat{\pi}^\star$ *of empirical MDP achieves*

$$\|V^{\widehat{\pi}^\star} - V^\star\|_\infty \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- matches minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ when $\varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$ (equivalently, when sample size exceeds $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$) Azar et al., 2013

ℓ_∞-based sample complexity

Theorem (Agarwal, Kakade, Yang '19)

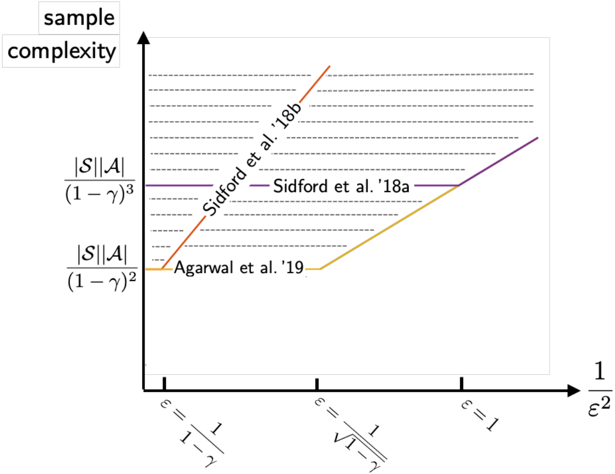
For any $0 < \varepsilon \le \frac{1}{\sqrt{1-\gamma}}$, *the optimal policy* $\widehat{\pi}^\star$ *of empirical MDP achieves*

$$\|V^{\widehat{\pi}^\star} - V^\star\|_\infty \le \varepsilon$$

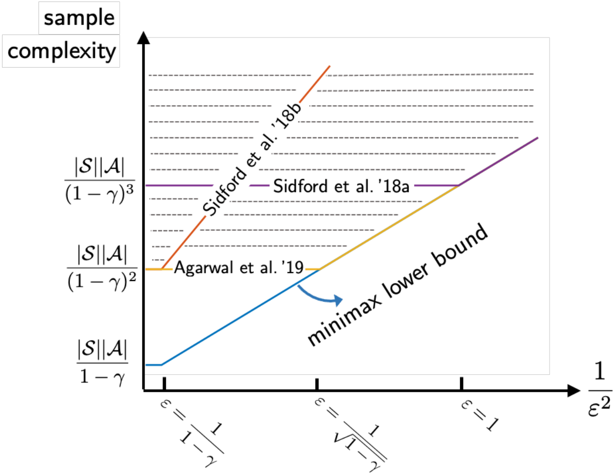
with high prob., with sample complexity at most

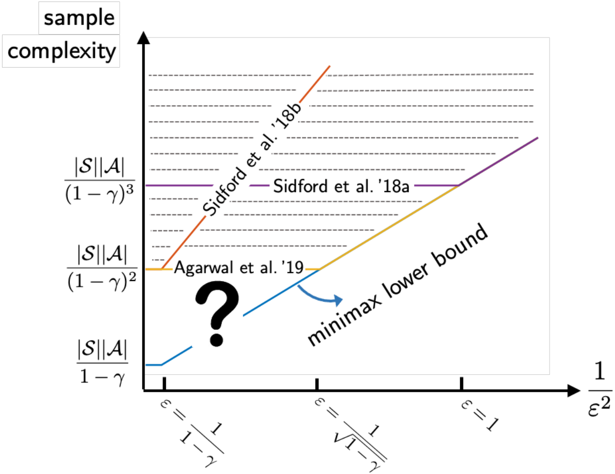
$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- matches minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ when $\varepsilon \le \frac{1}{\sqrt{1-\gamma}}$ (equivalently, when sample size exceeds $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$) Azar et al., 2013
- established upon leave-one-out analysis framework

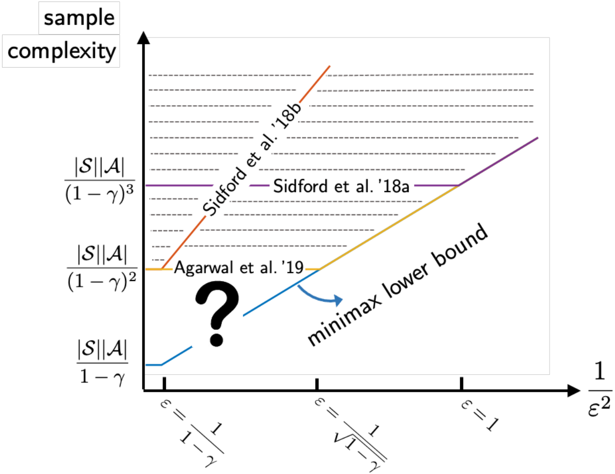


sample
complexity
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$
Sidford et al. '18b
Sidford et al. '18a
Agarwal et al. '19
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$
$\varepsilon = 1$
$\frac{1}{\varepsilon^2}$


sample
complexity
Sidford et al. '18b
Sidford et al. '18a
Agarwal et al. '19
minimax lower bound
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$
$\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$
$\varepsilon = 1$



Agarwal et al., 2019 still requires a burn-in sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

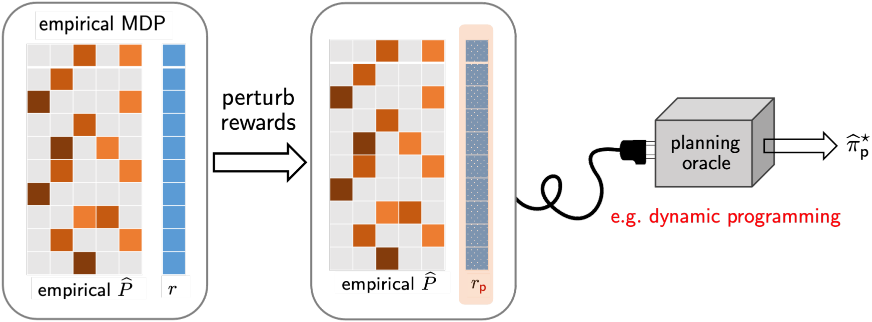


Agarwal et al., 2019 still requires a burn-in sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

Question: is it possible to break this sample size barrier?

Perturbed model-based approach (Li et al. '20)

—*Li et al., 2020*



Find policy based on the empirical MDP with slightly perturbed rewards

Optimal ℓ_∞-based sample complexity

Theorem (Li, Wei, Chi, Chen '20)

For any $0 < \varepsilon \le \frac{1}{1-\gamma}$*, the optimal policy* $\widehat{\pi}_{\mathrm{p}}^\star$ *of perturbed empirical MDP achieves*

$$\|V^{\widehat{\pi}_{\mathrm{p}}^\star} - V^\star\|_\infty \le \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

Optimal ℓ_∞-based sample complexity

Theorem (Li, Wei, Chi, Chen '20)

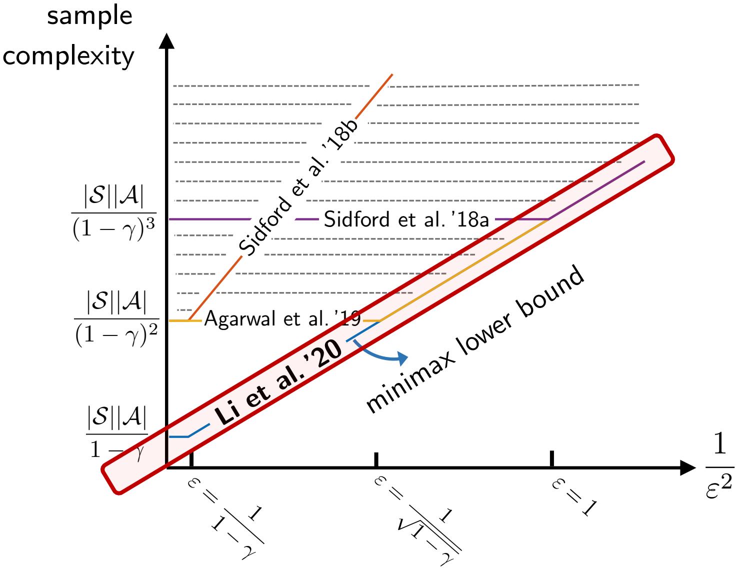
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *the optimal policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of perturbed empirical MDP achieves*

$$\|V^{\widehat{\pi}_{\mathrm{p}}^{\star}} - V^{\star}\|_{\infty} \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- matches minimax lower bound: $\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$ Azar et al., 2013
- full ε-range: $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$ $\longrightarrow$ no burn-in cost
- established upon more refined leave-one-out analysis and a perturbation argument



sample complexity
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$
$\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$
Sidford et al. '18b
Sidford et al. '18a
Agarwal et al. '19
Li et al. '20
minimax lower bound
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$
$\varepsilon = 1$

A sketch of the main proof ingredients

Notation and Bellman equation

Bellman equation: $V^\pi = r_\pi + \gamma P_\pi V^\pi$

- V^π: value function under policy π
  - Bellman equation: $V^\pi = (I - \gamma P_\pi)^{-1} r_\pi$
- $\widehat{V}^\pi$: <u>empirical version</u> value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \gamma \widehat{P}_\pi)^{-1} r_\pi$

Notation and Bellman equation

Bellman equation: $V^{\pi} = r_{\pi} + \gamma P_{\pi} V^{\pi}$

- V^{π}: value function under policy π
  - Bellman equation: $V^{\pi} = (I - \gamma P_{\pi})^{-1} r_{\pi}$
- $\widehat{V}^{\pi}$: empirical version value function under policy π
  - Bellman equation: $\widehat{V}^{\pi} = (I - \gamma \widehat{P}_{\pi})^{-1} r_{\pi}$
- $\pi^{\star}$: optimal policy for V^{π}
- $\widehat{\pi}^{\star}$: optimal policy for $\widehat{V}^{\pi}$

Main steps

Elementary decomposition:

$$\begin{aligned} V^\star - V^{\widehat{\pi}^\star} &= \left(V^\star - \widehat{V}^{\pi^\star}\right) + \left(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\right) + \left(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\right) \\ &\le \left(V^{\pi^\star} - \widehat{V}^{\pi^\star}\right) + 0 + \left(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\right) \end{aligned}$$

Main steps

Elementary decomposition:

$$V^\star - V^{\widehat{\pi}^\star} = \big(V^\star - \widehat{V}^{\pi^\star}\big) + \big(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\big) + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big)$$
$$\leq \big(V^{\pi^\star} - \widehat{V}^{\pi^\star}\big) + 0 + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big)$$

- **Step 1:** control $V^\pi - \widehat{V}^\pi$ for a fixed π (called "policy evaluation")
  (Bernstein inequality + a peeling argument)

Main steps

Elementary decomposition:

$$V^\star - V^{\widehat{\pi}^\star} = \left(V^\star - \widehat{V}^{\pi^\star}\right) + \left(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\right) + \left(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\right)$$
$$\le \left(V^{\pi^\star} - \widehat{V}^{\pi^\star}\right) + 0 + \left(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\right)$$

- **Step 1:** control $V^\pi - \widehat{V}^\pi$ for a fixed π (called "policy evaluation") (Bernstein inequality + a peeling argument)
- **Step 2:** extend it to control $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$ ($\widehat{\pi}^\star$ depends on samples) (decouple statistical dependency)

Key idea 1: a peeling argument (for fixed policy)

First-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\widehat{V}^{\pi} \qquad \text{[Agarwal et al., 2019]}$$

Key idea 1: a peeling argument (for fixed policy)

First-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\widehat{V}^{\pi} \qquad \text{[Agarwal et al., 2019]}$$

Ours: higher-order expansion + Bernstein $\longrightarrow$ tighter control

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi} + \\ + \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\big(\widehat{V}^{\pi} - V^{\pi}\big)$$

Bernstein's inequality: $|\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi}| \leq \sqrt{\frac{Var[V^{\pi}]}{N}} + \frac{\|V^{\pi}\|_{\infty}}{N}$

Key idea 1: a peeling argument (for fixed policy)

First-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\widehat{V}^{\pi}$$

[Agarwal et al., 2019]

Ours: higher-order expansion + Bernstein $\longrightarrow$ tighter control

$$\begin{aligned}\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi} + \\ + \gamma^{2}\Big(\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\Big)^{2} V^{\pi} \\ + \gamma^{3}\Big(\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\Big)^{3} V^{\pi} \\ + \dots\end{aligned}$$

Bernstein's inequality: $|\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi}| \leq \sqrt{\frac{Var[V^{\pi}]}{N}} + \frac{\|V^{\pi}\|_{\infty}}{N}$

Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. *For every* $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *plug-in estimator* $\widehat{V}^{\pi}$ *obeys*

$$\|\widehat{V}^{\pi} - V^{\pi}\|_{\infty} \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big).$$

Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. *For every* $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *plug-in estimator* $\widehat{V}^{\pi}$ *obeys*

$$\|\widehat{V}^{\pi} - V^{\pi}\|_{\infty} \leq \varepsilon$$

with sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big).$$

- minimax lower bound [Azar et al., 2013, Pananjady and Wainwright, 2019]

Byproduct: policy evaluation

Theorem (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. *For every* $0 < \varepsilon \le \frac{1}{1-\gamma}$, *plug-in estimator* $\widehat{V}^{\pi}$ *obeys*

$$\|\widehat{V}^{\pi} - V^{\pi}\|_{\infty} \le \varepsilon$$

with sample complexity at most

$$\widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big).$$

- minimax lower bound [Azar et al., 2013, Pananjady and Wainwright, 2019]
- tackle sample size barrier: prior work requires sample size $> \frac{|\mathcal{S}|}{(1-\gamma)^2}$ [Agarwal et al., 2013, Pananjady and Wainwright, 2019, Khamaru et al,, 2020]

Step 2: controlling $\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}$

A natural idea: apply our policy evaluation theory + union bound

Step 2: controlling $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

A natural idea: apply our policy evaluation theory + union bound

- highly suboptimal!

Step 2: controlling $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

A natural idea: apply our policy evaluation theory + union bound

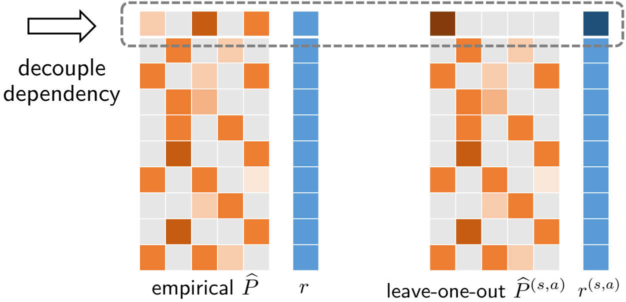
- highly suboptimal!

key idea 2: a leave-one-out argument to decouple stat. dependency btw $\widehat{\pi}$ and samples

— *inspired by [Agarwal et al., 2019] but quite different . . .*

Key idea 2: decouple dependency for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

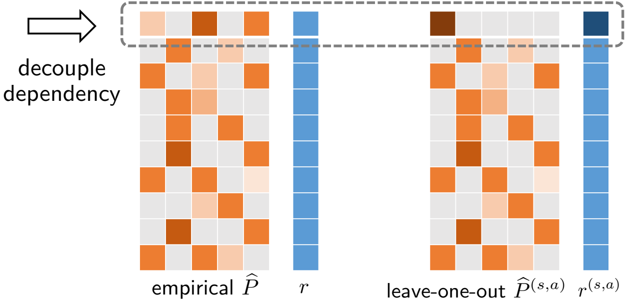
— *inspired by [Agarwal et al., 2019] but quite different ...*



- define $\widehat{\pi}^\star_{(s,a)} \xrightarrow{\text{empirical maximizer}} \big(\widehat{P}^{(s,a)}, r^{(s,a)}\big)$

Key idea 2: decouple dependency for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

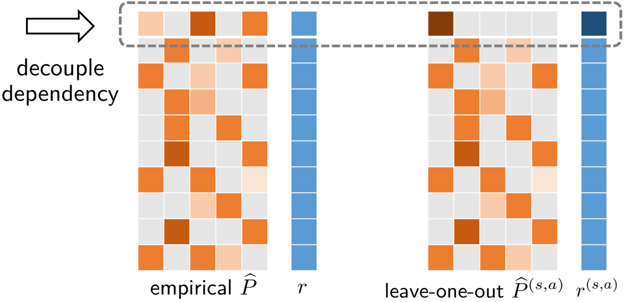
— *inspired by [Agarwal et al., 2019] but quite different ...*



- define $\widehat{\pi}^\star_{(s,a)} \xrightarrow{\text{empirical maximizer}} \big(\widehat{P}^{(s,a)}, r^{(s,a)}\big)$
  - decouple dependency by dropping randomness in $\widehat{P}(\cdot \mid s, a)$
  - scalar $r^{(s,a)}$ ensures $\widehat{Q}^\star$ and $\widehat{V}^\star$ unchanged

Key idea 2: decouple dependency for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

— *inspired by [Agarwal et al., 2019] but quite different ...*



- define $\widehat{\pi}^\star_{(s,a)} \xrightarrow{\text{empirical maximizer}} \big(\widehat{P}^{(s,a)}, r^{(s,a)}\big)$
- $\widehat{\pi}^\star_{(s,a)} = \widehat{\pi}^\star$ can be determined under separation condition

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^\star(s, \widehat{\pi}^\star(s)) - \max_{a:\, a \neq \widehat{\pi}^\star(s)} \widehat{Q}^\star(s,a) > 0$$

Key idea 3: tie-breaking via perturbation

- How to ensure the optimal policy stand out from other policies?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a: a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s, a) \geq \omega$$

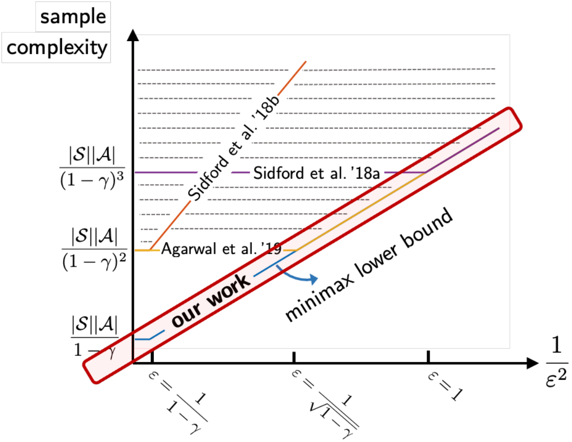
Key idea 3: tie-breaking via perturbation

- How to ensure the optimal policy stand out from other policies?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a: a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s, a) \geq \omega$$

- **Solution:** slightly perturb rewards $r \implies \widehat{\pi}^{\star}_{\mathrm{p}}$
  - ensures the uniqueness of $\widehat{\pi}^{\star}_{\mathrm{p}}$
  - $V^{\widehat{\pi}^{\star}_{\mathrm{p}}} \approx V^{\widehat{\pi}^{\star}}$

Summary of model-based RL

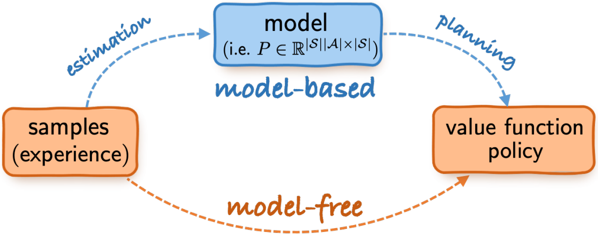


Model-based RL is minimax optimal & does not suffer from a sample size barrier!

Part 1

1. Basics: Markov decision processes
2. RL w/ a generative model (simulator)
   - model-based algorithms (a "plug-in" approach)
   - model-free algorithms

Model-based vs. model-free RL



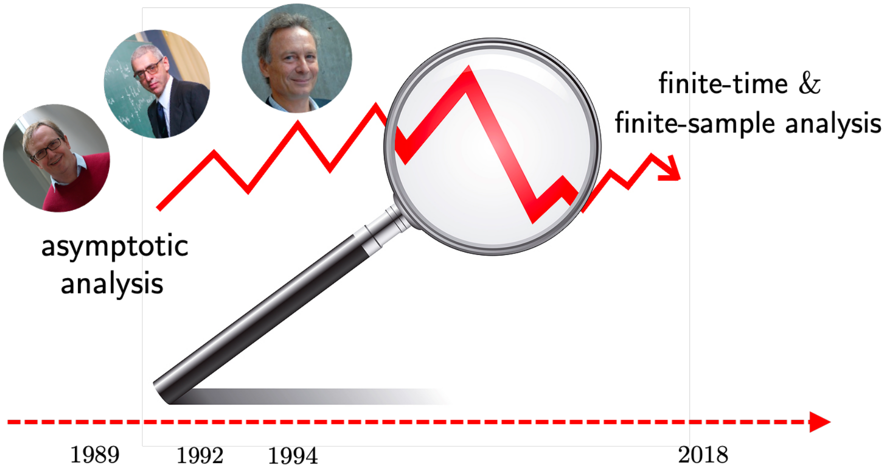
Model-based approach ("plug-in")

1. build empirical estimate $\widehat{P}$ for P
2. planning based on empirical $\widehat{P}$

Model-free / value-based approach

— learning w/o modeling & estimating environment explicitly

— memory-efficient, online, ...



Focus of this part: classical **Q-learning** algorithm and its variants

A starting point: Bellman optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\left[\max_{a' \in \mathcal{A}} Q(s',a')\right]}_{\text{next state's value}}$$

- one-step look-ahead

A starting point: Bellman optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\Big[\max_{a' \in \mathcal{A}} Q(s',a')\Big]}_{\text{next state's value}}$$

- one-step look-ahead

Bellman equation: $Q^\star$ is *unique* solution to

$$\mathcal{T}(Q^\star) = Q^\star$$

A starting point: Bellman optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\Big[\max_{a' \in \mathcal{A}} Q(s',a')\Big]}_{\text{next state's value}}$$

- one-step look-ahead

Bellman equation: $Q^\star$ is *unique* solution to

$$\mathcal{T}(Q^\star) = Q^\star$$

- **takeaway message:** it suffices to solve the Bellman equation
- **challenge:** how to solve it using stochastic samples?

Richard Bellman

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

$\underbrace{\text{Stochastic approximation}}_{\text{Robbins \& Monro, 1951}}$ for solving the **Bellman equation**

$$\mathcal{T}(Q) - Q = 0$$

where

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \Big[\underbrace{\max_{a' \in \mathcal{A}} Q(s',a')}_{\text{next state's value}} \Big].$$

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $\mathcal{T}(Q) - Q = 0$

$$\underbrace{Q_{t+1}(s,a) = Q_t(s,a) + \eta_t\big(\mathcal{T}_t(Q_t)(s,a) - Q_t(s,a)\big)}_{\text{sample transition } (s,a,s')}, \quad t \geq 0$$

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $\mathcal{T}(Q) - Q = 0$

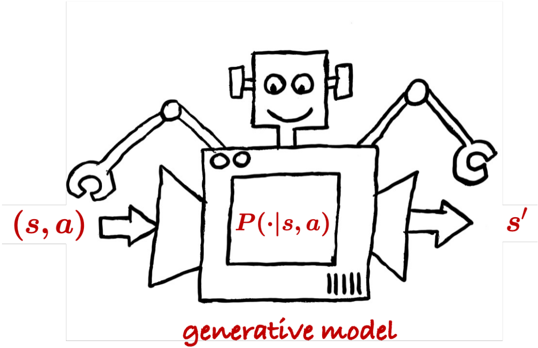
$$\underbrace{Q_{t+1}(s,a) = Q_t(s,a) + \eta_t\big(\mathcal{T}_t(Q_t)(s,a) - Q_t(s,a)\big)}_{\text{sample transition } (s,a,s')}, \quad t \geq 0$$

$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

$$\mathcal{T}(Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \big[\max_{a'} Q(s',a')\big]$$

A generative model / simulator

— *Kearns, Singh '99*



Each iteration, draw an independent sample (s, a, s') for given (s, a)

Synchronous Q-learning

Chris Watkins

Peter Dayan

for $t = 0, 1, \ldots, T$

for each $(s,a) \in \mathcal{S} \times \mathcal{A}$

draw a sample (s, a, s'), run

$$Q_{t+1}(s,a) = (1-\eta_t)Q_t(s,a) + \eta_t \Big\{ r(s,a) + \gamma \max_{a'} Q_t(s',a') \Big\}$$

synchronous: all state-action pairs are updated simultaneously

- total sample size: $T|\mathcal{S}||\mathcal{A}|$

Sample complexity of synchronous Q-learning

Theorem (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$, with sample size at most

$$\begin{cases} \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) & \text{if } |\mathcal{A}| \geq 2 \\ \widetilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right) & \text{if } |\mathcal{A}| = 1 \qquad (\textit{TD learning}) \end{cases}$$

Sample complexity of synchronous Q-learning

Theorem (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$, with sample size at most

$$\begin{cases} \widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) & \text{if } |\mathcal{A}| \geq 2 \\ \widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big) & \text{if } |\mathcal{A}| = 1 \qquad (\text{TD learning}) \end{cases}$$

- Covers both *constant* and *rescaled linear* learning rates:

$$\eta_t \equiv \frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \quad \text{or} \quad \eta_t = \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}}$$

Sample complexity of synchronous Q-learning

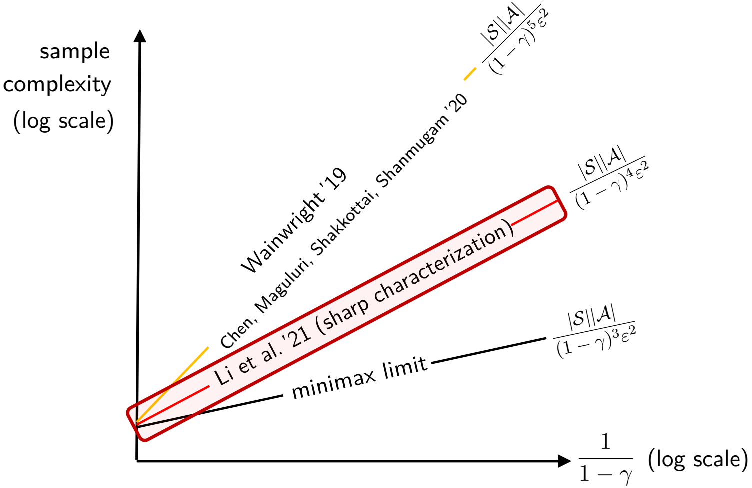
Theorem (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^{\star}\|_{\infty}] \leq \varepsilon$, with sample size at most

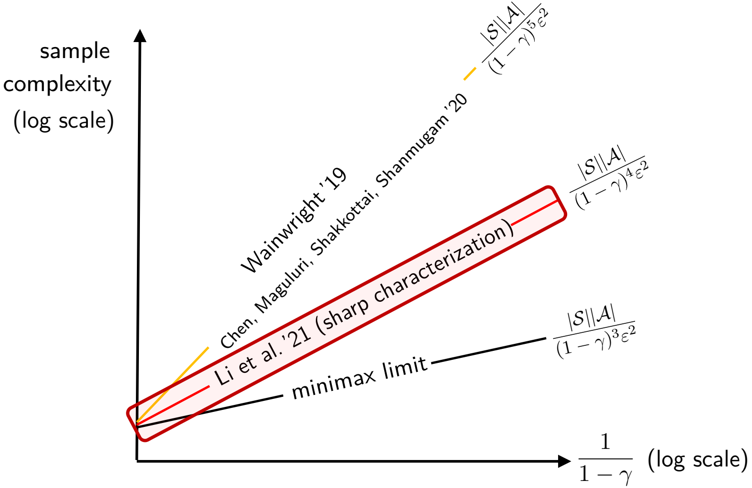
$$\begin{cases} \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) & \text{if } |\mathcal{A}| \geq 2 \qquad (?) \\ \widetilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right) & \text{if } |\mathcal{A}| = 1 \qquad \text{(minimax optimal)} \end{cases}$$

other papers	sample complexity
Even-Dar & Mansour '03	$2^{\frac{1}{1-\gamma}} \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$
Beck & Srikant '12	$\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^5\varepsilon^2}$
Wainwright '19	$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$
Chen, Maguluri, Shakkottai, Shanmugam '20	$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$

All this requires sample size at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ ($|\mathcal{A}| \geq 2$) ...



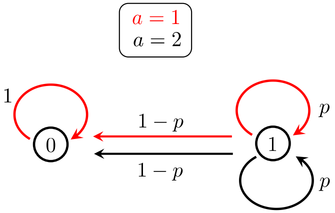
All this requires sample size at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ ($|\mathcal{A}| \geq 2$) ...



Question: *Is Q-learning sub-optimal, or is it an analysis artifact?*

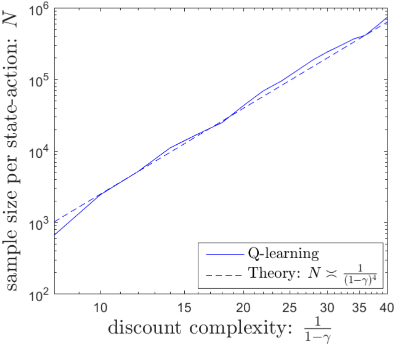
A numerical example: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ samples seem necessary . . .

— *observed in Wainwright '19*



$$p = \frac{4\gamma - 1}{3\gamma}$$

$$r(0,1) = 0, \quad r(1,1) = r(1,2) = 1$$



Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, there exists an MDP with $|\mathcal{A}| \geq 2$ such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$

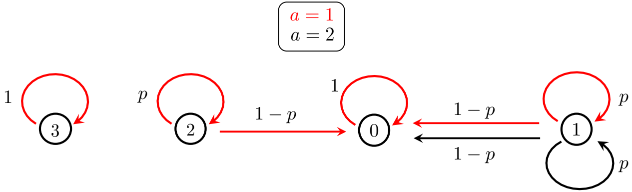
Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \le 1$, there exists an MDP with $|\mathcal{A}| \ge 2$ such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$

- Tight **algorithm-dependent** lower bound
- Holds for both constant and rescaled linear learning rates

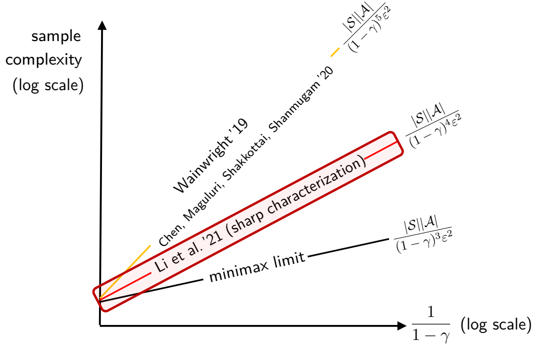


Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, there exists an MDP with $|\mathcal{A}| \geq 2$ such that to achieve $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$



Improving sample complexity via **variance reduction**

— *a powerful idea from finite-sum stochastic optimization*

Variance-reduced Q-learning updates (Wainwright '19)

— *inspired by SVRG (Johnson & Zhang '13)*

$$Q_t(s,a) = (1-\eta)Q_{t-1}(s,a) + \eta\Big(\mathcal{T}_t(Q_{t-1}) \underbrace{-\mathcal{T}_t(\overline{Q}) + \widetilde{\mathcal{T}}(\overline{Q})}_{\text{use } \overline{Q} \text{ to help reduce variability}}\Big)(s,a)$$

Variance-reduced Q-learning updates (Wainwright '19)

— *inspired by SVRG (Johnson & Zhang '13)*

$$Q_t(s,a) = (1-\eta)Q_{t-1}(s,a) + \eta\Big(\mathcal{T}_t(Q_{t-1}) \underbrace{-\mathcal{T}_t(\overline{Q}) + \widetilde{\mathcal{T}}(\overline{Q})}_{\text{use } \overline{Q} \text{ to help reduce variability}}\Big)(s,a)$$

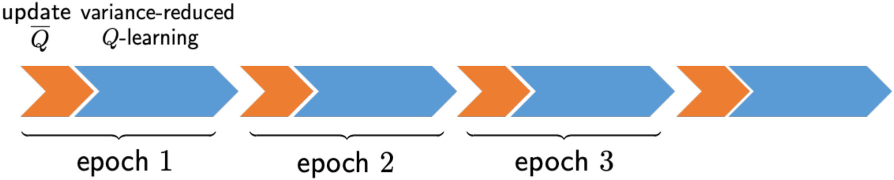
- $\overline{Q}$: some reference Q-estimate
- $\widetilde{\mathcal{T}}$: empirical Bellman operator (using a *batch* of samples)

$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

$$\widetilde{\mathcal{T}}(Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim \widetilde{P}(\cdot|s,a)} \big[\max_{a'} Q(s',a')\big]$$

An epoch-based stochastic algorithm

— *inspired by Johnson & Zhang '13*



for each epoch

1. update $\overline{Q}$ and $\widetilde{\mathcal{T}}(\overline{Q})$ (which stay fixed in the rest of the epoch)
2. run variance-reduced Q-learning updates iteratively

Sample complexity of variance-reduced Q-learning

Theorem (Wainwright '19)

For any $0 < \varepsilon \leq 1$, *sample complexity for* **variance-reduced synchronous Q-learning** *to yield* $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ *is at most*

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- allows for more aggressive learning rates

Sample complexity of variance-reduced Q-learning

Theorem (Wainwright '19)

For any $0 < \varepsilon \le 1$, *sample complexity for* **variance-reduced synchronous Q-learning** *to yield* $\|\widehat{Q} - Q^{\star}\|_{\infty} \le \varepsilon$ *is at most*

$$\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$$

- allows for more aggressive learning rates
- minimax-optimal for $0 < \varepsilon \le 1$
  - remains suboptimal if $1 < \varepsilon < \frac{1}{1-\gamma}$

Reference: general RL textbooks I

- "*Reinforcement learning: An introduction*," R. S. Sutton, A. G. Barto, MIT Press, 2018
- "*Reinforcement learning: Theory and algorithms*," A. Agarwal, N. Jiang, S. Kakade, W. Sun, 2019
- "*Reinforcement learning and optimal control*," D. Bertsekas, Athena Scientific, 2019
- "*Algorithms for reinforcement learning*," C. Szepesvari, Springer, 2022
- "*Bandit algorithms*," T. Lattimore, C. Szepesvari, Cambridge University Press, 2020

Reference: model-based algorithms I

- "*Finite-sample convergence rates for Q-learning and indirect algorithms*," M. Kearns, S. Satinder, *NeurIPS*, 1998
- "*On the sample complexity of reinforcement learning*," S. Kakade, 2003
- "*A sparse sampling algorithm for near-optimal planning in large Markov decision processes*," M. Kearns, Y. Mansour, A. Y. Ng, *Machine learning*, 2002
- "*Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model*," M. G. Azar, R. Munos, H. J. Kappen, *Machine learning*, 2013
- "*Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time*," *Mathematics of Operations Research*, 2020
- "*Near-optimal time and sample complexities for solving Markov decision processes with a generative model*," A. Sidford, M. Wang, X. Wu, L. Yang, Y. Ye, *NeurIPS*, 2018
- "*Variance reduced value iteration and faster algorithms for solving Markov decision processes*," A. Sidford, M. Wang, X. Wu, Y. Ye, *SODA*, 2018
- "*Model-based reinforcement learning with a generative model is minimax optimal*," A. Agarwal, S. Kakade, L. Yang, *COLT*, 2020

Reference: model-based algorithms II

- "*Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning*," A. Pananjady, M. J. Wainwright, *IEEE Trans. on Information Theory*, 2020
- "*Spectral methods for data science: A statistical perspective*," Y. Chen, Y. Chi, J. Fan, C. Ma, *Foundations and Trends® in Machine Learning*, 2021
- "*Breaking the sample size barrier in model-based reinforcement learning with a generative model,*" G. Li, Y. Wei, Y. Chi, Y. Chen, *Operations Research*, 2024

Reference: model-free algorithms I

- "*A stochastic approximation method,*" H. Robbins, S. Monro, *Annals of Mathematical Statistics*, 1951
- "*Robust stochastic approximation approach to stochastic programming,*" A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, *SIAM Journal on optimization*, 2009
- "*Q-learning,*" C. Watkins, P. Dayan, *Machine Learning*, 1992
- "*Learning rates for Q-learning,*" E. Even-Dar, Y. Mansour, *Journal of Machine Learning Research*, 2003
- "*The asymptotic convergence-rate of Q-learning,*" C. Szepesvari, *NeurIPS*, 1998
- "*Error bounds for constant step-size Q-learning,*" C. Beck, R. Srikant, *Systems & Control Letters*, 2012
- "*Stochastic approximation with cone-contractive operators: Sharp ℓ_∞ bounds for Q-learning,*" M. Wainwright, 2019
- "*Is Q-learning minimax optimal? a tight sample complexity analysis*," G. Li, C. Cai, Y. Chen, Y. Wei, Y. Chi, *Operations Research*, 2024
- "*Variance-reduced Q-learning is minimax optimal*," M. Wainwright, 2019

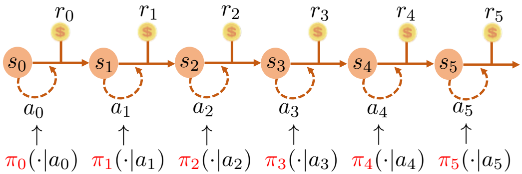
Reference: model-free algorithms II

- "*Sample-optimal parametric Q-learning using linearly additive features,*" L. Yang, M. Wang, *ICML*, 2019
- "*Asynchronous stochastic approximation and Q-learning*," J. Tsitsiklis, *Machine learning*, 1994
- "*Finite-time analysis of asynchronous stochastic approximation and Q-learning*," G. Qu, A. Wierman, *COLT*, 2020
- "*Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes*," Z. Chen, S. T. Maguluri, S. Shakkottai, K. Shanmugam, *NeurIPS*, 2020
- "*Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction*," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, *IEEE Trans. on Information Theory*, 2022

Part 2

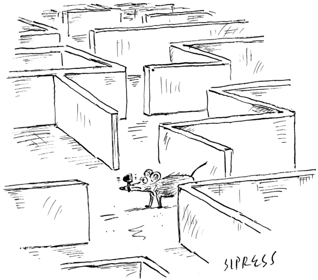
1. Online RL
2. Offline RL
3. Multi-agent RL
4. Robust RL

Online RL: interacting with real environment



exploration via adaptive policies

- trial-and-error
- sequential and online
- adaptive learning from data

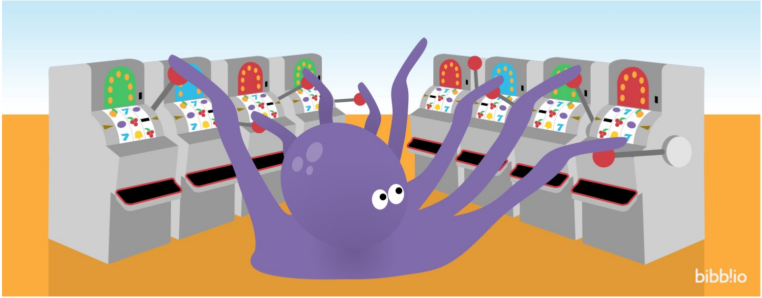


"Recalculating ... recalculating ..."

A much simpler problem: multi-arm bandit

Multi-arm bandit

Which slot machine will give me the most money?



First proposed in [Thompson'33], popularized by [Robbins'52].

Learning the best arm

Can we learn which slot machine gives the most money?

$1
$0
$0

$1
$4
$0
$2
$1
$3
$5

$1
$0
$1
$2

Formulation

We can play multiple rounds $t = 1, 2, \ldots, T$.

In each round, we select an arm i_t from a fixed set $i = 1, 2, \ldots, n$; and observe the reward r_t that the arm gives.

Arm 1

Arm 2

Arm 3

Formulation

We can play multiple rounds $t = 1, 2, \dots, T$.

In each round, we select an arm i_t from a fixed set $i = 1, 2, \dots, n$; and observe the reward r_t that the arm gives.

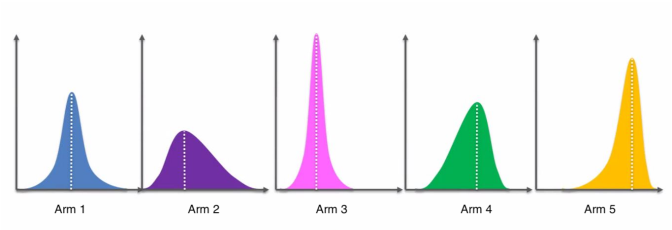
Arm 1

Arm 2

Arm 3

Objective: Maximize the total reward over time.

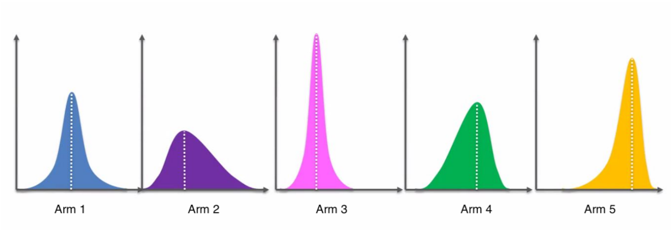
Stochastic bandit with i.i.d. rewards



- Each arm distributes rewards according to some (unknown) distribution over $[0, 1]$, with

$$\mathbb{E}[r_{i,t}] = \mu_i, \quad \forall i \in [n],\ t = 1, 2 \ldots$$

Stochastic bandit with i.i.d. rewards



- Each arm distributes rewards according to some (unknown) distribution over $[0, 1]$, with

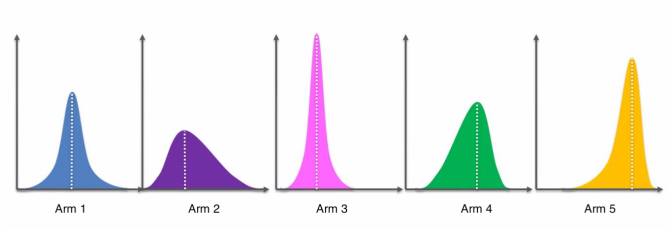
$$\mathbb{E}[r_{i,t}] = \mu_i, \quad \forall i \in [n], \ t = 1, 2 \ldots$$

- Suppose we play arm i_t at round t, and receive the reward

$$r_{i_t,t}$$

drawn i.i.d. from the arm i_t's distribution.

Stochastic bandit with i.i.d. rewards



- Each arm distributes rewards according to some (unknown) distribution over $[0, 1]$, with

$$\mathbb{E}[r_{i,t}] = \mu_i, \quad \forall i \in [n], \ t = 1, 2 \dots$$

- Suppose we play arm i_t at round t, and receive the reward

$$r_{i_t,t}$$

drawn i.i.d. from the arm i_t's distribution.

Partial information: Every round we cannot observe the reward of all arms: we just know the reward of the arm that we played.

Regret: performance metric

We design algorithms that determine the sequence $\{i_t\}$, i.e. *policies*.

How to evaluate the performance?

Definition (Expected regret)

The expected regret over T rounds is defined as

$$R_T = \max_{1 \le i \le n} \mathbb{E}\left[\sum_{t=1}^{T} (r_{i,t} - r_{i_t,t})\right] = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_{i_t,t}\right],$$

where $\mu^\star = \max_{1 \le i \le n} \mu_i$ is the highest expected reward over all arms.

Regret: performance metric

We design algorithms that determine the sequence $\{i_t\}$, i.e. *policies*.

How to evaluate the performance?

Definition (Expected regret)

The expected regret over T rounds is defined as

$$R_T = \max_{1 \le i \le n} \mathbb{E}\left[\sum_{t=1}^{T} (r_{i,t} - r_{i_t,t})\right] = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_{i_t,t}\right],$$

where $\mu^\star = \max_{1 \le i \le n} \mu_i$ is the highest expected reward over all arms.

- 1st term captures the highest cumulative reward in *hindsight*.

Regret: performance metric

We design algorithms that determine the sequence $\{i_t\}$, i.e. *policies*.

How to evaluate the performance?

Definition (Expected regret)

The expected regret over T rounds is defined as

$$R_T = \max_{1 \le i \le n} \mathbb{E}\left[\sum_{t=1}^{T} (r_{i,t} - r_{i_t,t})\right] = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_{i_t,t}\right],$$

where $\mu^\star = \max_{1 \le i \le n} \mu_i$ is the highest expected reward over all arms.

- 1st term captures the highest cumulative reward in *hindsight*.
- 2nd term captures the *actual* accumulated reward.

The UCB algorithm

[Auer et al.'02]:: the idea is to always try the best arm, where "best" includes exploration and exploitation.

1. **Initial phase:** try each arm and observe the reward.

The UCB algorithm

[Auer et al.'02]:: the idea is to always try the best arm, where "best" includes exploration and exploitation.

1. **Initial phase:** try each arm and observe the reward.
2. For each round $t = n + 1, \ldots, T$:

The UCB algorithm

[Auer et al.'02]:: the idea is to always try the best arm, where "best" includes exploration and exploitation.

1. **Initial phase:** try each arm and observe the reward.
2. For each round $t = n + 1, \ldots, T$:
   - Calculate the UCB (upper confidence bound) index for each arm i:

$$\mathsf{UCB}_{i,t} = \overline{\mu}_{i,t} + \sqrt{\frac{\log t}{T_{i,t}}},$$

     where $\overline{\mu}_{i,t}$ is the empirical average reward for arm i and $T_{i,t}$ is the number of times arm i has been played up to round t.

The UCB algorithm

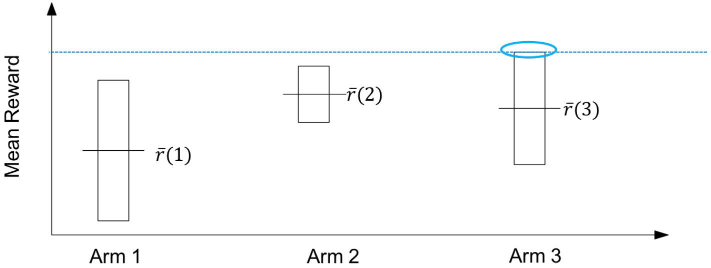
[Auer et al.'02]:: the idea is to always try the best arm, where "best" includes exploration and exploitation.

1. **Initial phase:** try each arm and observe the reward.
2. For each round $t = n + 1, \ldots, T$:
   - Calculate the UCB (upper confidence bound) index for each arm i:

     $$\mathsf{UCB}_{i,t} = \overline{\mu}_{i,t} + \sqrt{\frac{\log t}{T_{i,t}}},$$

     where $\overline{\mu}_{i,t}$ is the empirical average reward for arm i and $T_{i,t}$ is the number of times arm i has been played up to round t.
   - Play the arm with the highest UCB index and observe the reward.

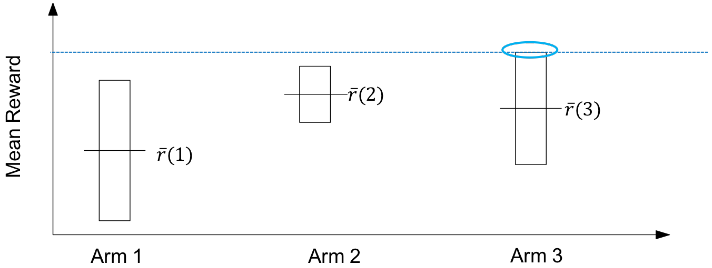
Understanding UCB



$$\text{UCB}_{i,t} = \overline{\mu}_{i,t} + \sqrt{\frac{\log t}{T_{i,t}}},$$

- Exploitation: $\overline{\mu}_{i,t}$ is the average observed reward. High observed rewards of an arm leads to high UCB index.

Understanding UCB



$$\mathrm{UCB}_{i,t} = \overline{\mu}_{i,t} + \sqrt{\frac{\log t}{T_{i,t}}},$$

- Exploitation: $\overline{\mu}_{i,t}$ is the average observed reward. High observed rewards of an arm leads to high UCB index.
- Exploration: $\sqrt{\frac{\log t}{T_{i,t}}}$ decreases as we make more observations ($T_{i,t}$ grows). Few observations of an arm leads to high UCB index.

Theory of UCB algorithm

Theorem (Worst-case regret bound of UCB)

For $T \geq n$, the expected regret of UCB algorithm is upper bounded as

$$R_T \leq 4\sqrt{nT \log T} + 8n.$$

Theory of UCB algorithm

Theorem (Worst-case regret bound of UCB)

For $T \geq n$, the expected regret of UCB algorithm is upper bounded as

$$R_T \leq 4\sqrt{nT \log T} + 8n.$$

- When $n = O(1)$, the regret scales as

$$R_T = O\big(\sqrt{T \log T}\big) = \widetilde{O}\big(\sqrt{T}\big)$$

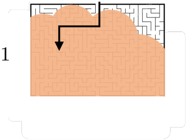
- The logarithmic factor can be shaved away [Audibert and Bubeck'09]

Back to online RL...

Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps

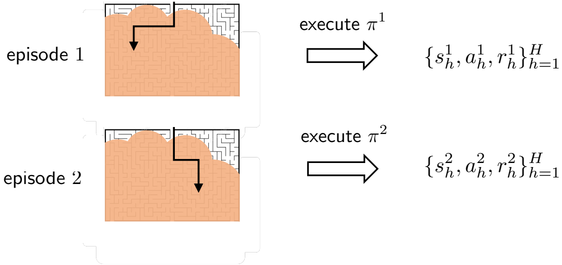
episode 1



execute π^1 $\Longrightarrow$ $\{s_h^1, a_h^1, r_h^1\}_{h=1}^H$

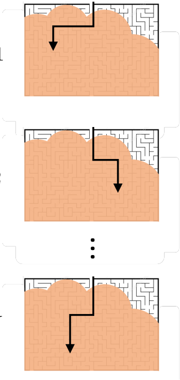
Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



episode 1 — execute π^1 $\Longrightarrow$ $\{s_h^1, a_h^1, r_h^1\}_{h=1}^H$

episode 2 — execute π^2 $\Longrightarrow$ $\{s_h^2, a_h^2, r_h^2\}_{h=1}^H$

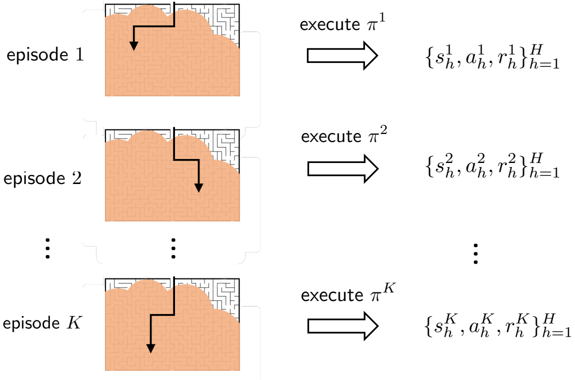
$\vdots$

episode K — execute π^K $\Longrightarrow$ $\{s_h^K, a_h^K, r_h^K\}_{h=1}^H$

Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps

— *sample size: $T = KH$*



exploration (exploring unknowns) vs. **exploitation** (exploiting learned info)

Regret: gap between learned policy & optimal policy

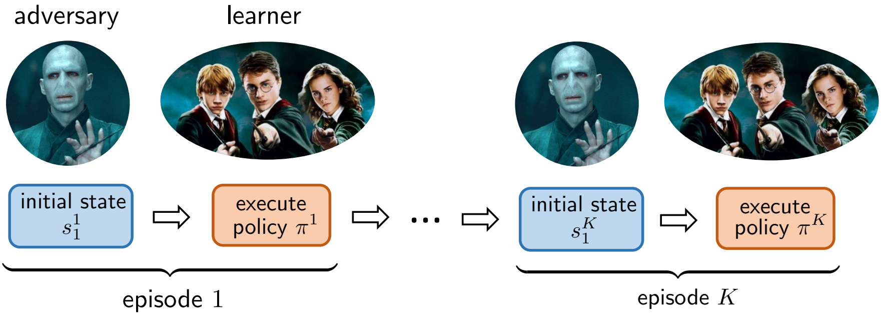
adversary

learner

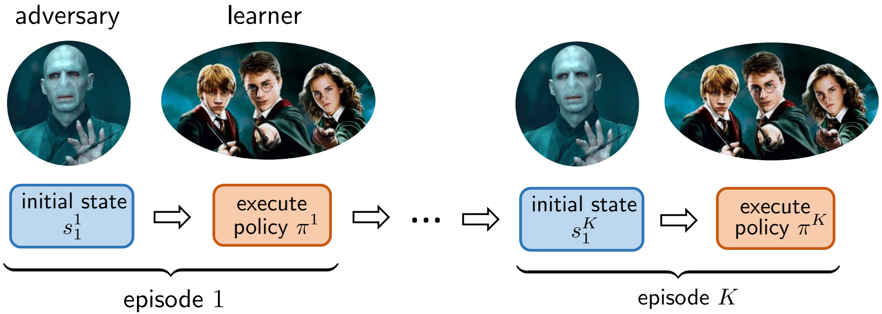
initial state s_1^1 $\Rightarrow$ execute policy π^1

episode 1

Regret: gap between learned policy & optimal policy



Regret: gap between learned policy & optimal policy



Performance metric: given initial states $\{s_1^k\}_{k=1}^K$, define

$$\text{Regret}(T) \;:=\; \sum_{k=1}^{K} \Big(V_1^{\star}(s_1^k) - V_1^{\pi^k}(s_1^k) \Big)$$

Lower bound

(Domingues et al, 2021)

$$\text{Regret}(T) \gtrsim \sqrt{H^2SAT}$$

Existing algorithms

- UCB-VI: Azar et al, 2017
- UBEV: Dann et al, 2017
- UCB-Q-Hoeffding: Jin et al, 2018
- UCB-Q-Bernstein: Jin et al, 2018
- UCB2-Q-Bernstein: Bai et al, 2019
- EULER: Zanette et al, 2019
- UCB-Q-Advantage: Zhang et al, 2020
- MVP: Zhang et al, 2020
- UCB-M-Q: Menard et al, 2021
- Q-EarlySettled-Advantage: Li et al, 2021
- (modified) MVP: Zhang et al, 2024

Lower bound

(Domingues et al, 2021)

$$\text{Regret}(T) \gtrsim \sqrt{H^2 SAT}$$

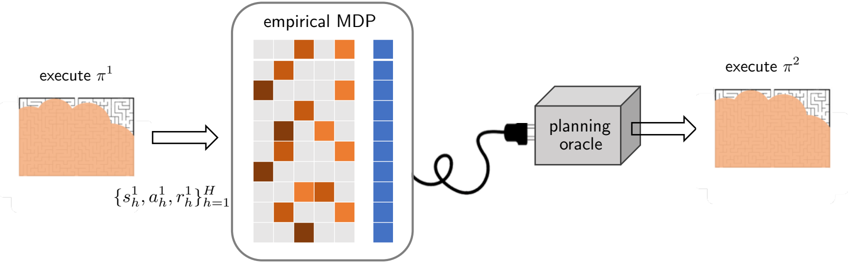
Existing algorithms

- UCB-VI: Azar et al, 2017
- UBEV: Dann et al, 2017
- UCB-Q-Hoeffding: Jin et al, 2018
- UCB-Q-Bernstein: Jin et al, 2018
- UCB2-Q-Bernstein: Bai et al, 2019
- EULER: Zanette et al, 2019
- UCB-Q-Advantage: Zhang et al, 2020
- MVP: Zhang et al, 2020
- UCB-M-Q: Menard et al, 2021
- Q-EarlySettled-Advantage: Li et al, 2021
- (modified) MVP: Zhang et al, 2024

Which online RL algorithms achieve near-minimal regret?

Model-based online RL with UCB exploration

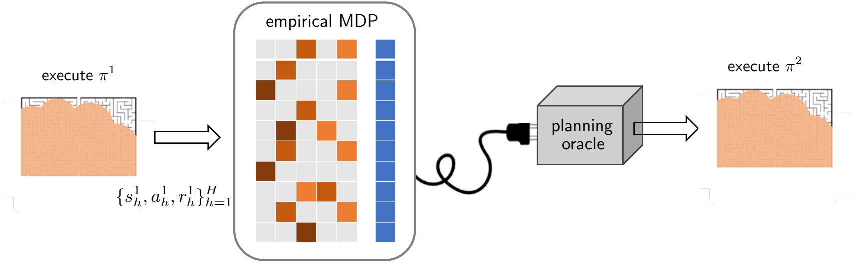
Model-based approach for online RL



repeat:

- use collected data to estimate transition probabilities
- apply planning to the estimated model to derive a new policy for sampling in the next episode

Model-based approach for online RL



repeat:

- use collected data to estimate transition probabilities
- apply planning to the estimated model to derive a new policy for sampling in the next episode

How to balance exploration and exploitation in this framework?

T. L. Lai

H. Robbins

Optimism in the face of uncertainty:

- explores based on the best optimistic estimates associated with the actions!
- a common framework: utilize $\underbrace{\text{upper confidence bounds (UCB)}}_{\text{accounts for estimates + uncertainty level}}$

T. L. Lai

H. Robbins

Optimism in the face of uncertainty:

- explores based on the best optimistic estimates associated with the actions!
- a common framework: utilize upper confidence bounds (UCB) — accounts for estimates + uncertainty level

Optimistic model-based approach: incorporates UCB framework into model-based approach

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H-1, \ldots, 1$: run **value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\widehat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H-1, \ldots, 1$: run **optimistic** **value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\widehat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1} + \underbrace{b_h(s_h, a_h)}_{\text{bonus (upper confidence width)}}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H-1, \ldots, 1$: run **optimistic value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\widehat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1} + \underbrace{b_h(s_h, a_h)}_{\text{bonus (upper confidence width)}}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

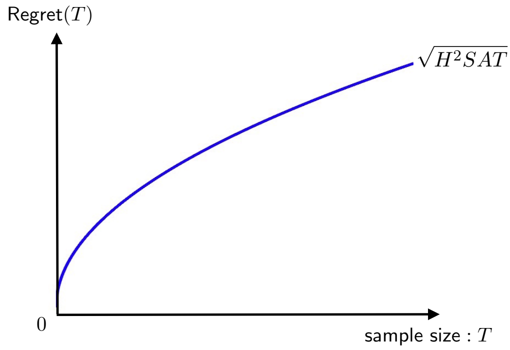
2. Forward $h = 1, \ldots, H$: take actions according to **greedy policy**

$$\pi_h(s) \leftarrow \arg\max_{a \in \mathcal{A}} Q_h(s, a)$$

to sample a new episode $\{s_h, a_h, r_h\}_{h=1}^H$

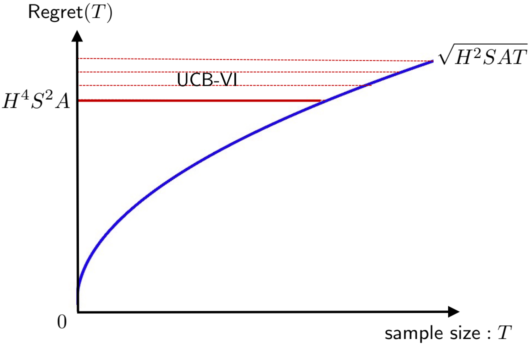
UCB-VI is asymptotically regret-optimal

— *Azar, Osband, Munos, 2017*



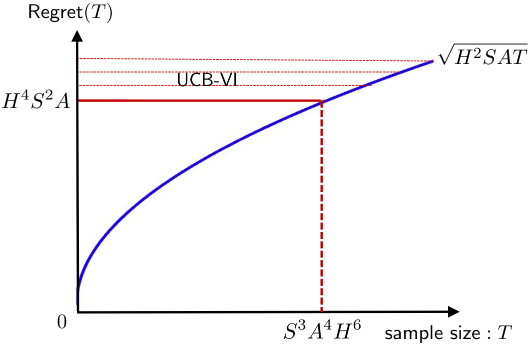
UCB-VI is asymptotically regret-optimal

— *Azar, Osband, Munos, 2017*



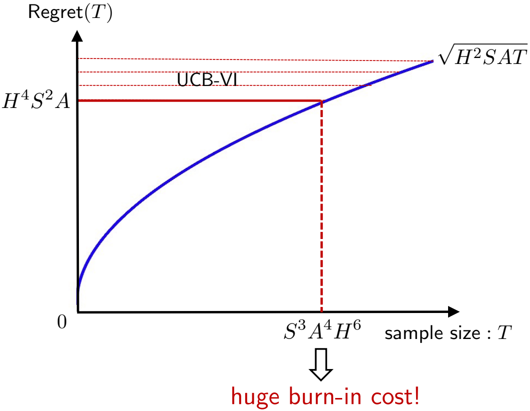
UCB-VI is asymptotically regret-optimal

— *Azar, Osband, Munos, 2017*



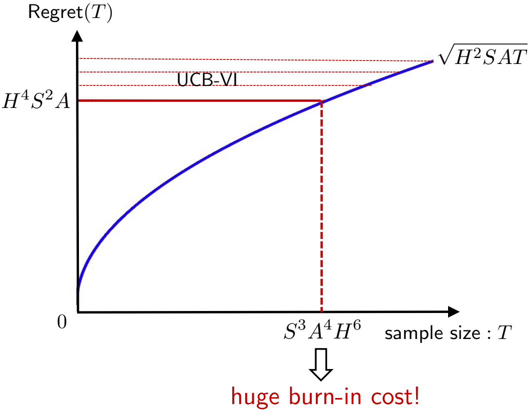
UCB-VI is asymptotically regret-optimal

— *Azar, Osband, Munos, 2017*



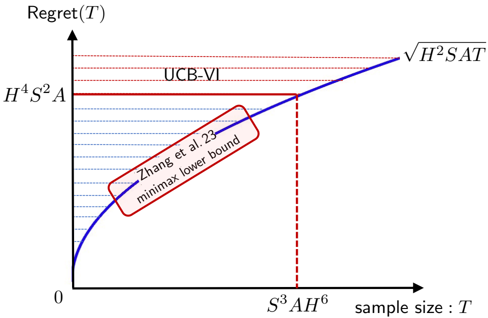
UCB-VI is asymptotically regret-optimal

— Azar, Osband, Munos, 2017



Issues: large burn-in cost

Regret-optimal algorithm w/o burn-in cost

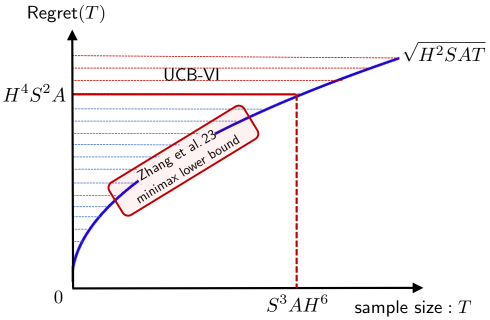


Theorem (Zhang, Chen, Lee, Du '24)

The model-based algorithm Monotonic Value Propagation achieves

$$\textit{Regret}(T) \lesssim \widetilde{O}\big(\sqrt{H^2SAT}\big)$$

Regret-optimal algorithm w/o burn-in cost



Theorem (Zhang, Chen, Lee, Du '24)

The model-based algorithm Monotonic Value Propagation achieves

$$\textit{Regret}(T) \lesssim \widetilde{O}\big(\sqrt{H^2SAT}\big)$$

- the only algorithm so far that is regret-optimal w/o burn-ins

Part 2

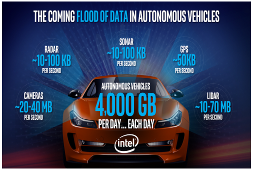
Four variants of our basics settings to illustrate the approaches so far:

- Online RL
- Offline RL
- Multi-agent RL
- Robust RL

Offline RL / batch RL

- Collecting new data might be expensive or time-consuming
- But we have already stored tons of historical data

medical records



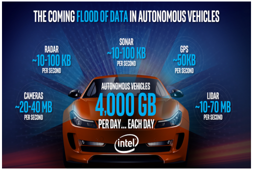
data of self-driving

clicking times of ads

Offline RL / batch RL

- Collecting new data might be expensive or time-consuming
- But we have already stored tons of historical data

medical records

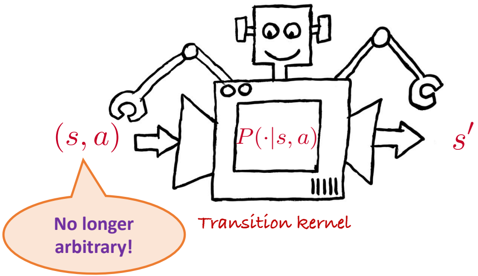


data of self-driving

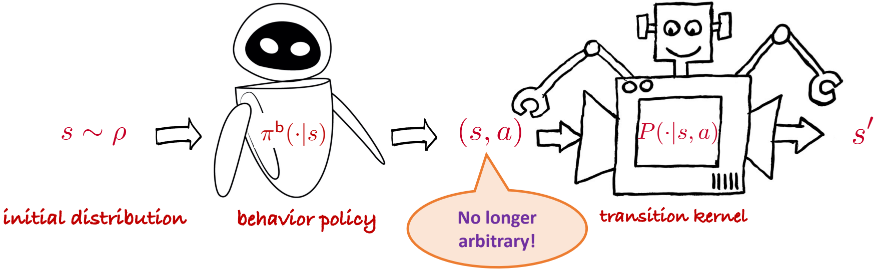
clicking times of ads

Question: Can we design algorithms based solely on historical data?

Offline RL / batch RL



Offline RL / batch RL



Offline RL / batch RL

A historical dataset $\mathcal{D} = \{(s^{(i)}, a^{(i)}, s'^{(i)})\}$**:** N independent copies of

$$s \sim \rho^{\mathsf{b}}, \qquad a \sim \pi^{\mathsf{b}}(\cdot \,|\, s), \qquad s' \sim P(\cdot \,|\, s, a)$$

for some state distribution ρ^{b} and behavior policy π^{b}

Offline RL / batch RL

A historical dataset $\mathcal{D} = \left\{\left(s^{(i)}, a^{(i)}, s'^{(i)}\right)\right\}$**:** N independent copies of

$$s \sim \rho^{\mathsf{b}}, \qquad a \sim \pi^{\mathsf{b}}(\cdot \,|\, s), \qquad s' \sim P(\cdot \,|\, s, a)$$

for some state distribution ρ^{b} and behavior policy π^{b}

Goal: given some test distribution ρ and accuracy level ε, find an ε-optimal policy $\widehat{\pi}$ based on $\mathcal{D}$ obeying

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) = \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\star}(s)\right] - \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\widehat{\pi}}(s)\right] \leq \varepsilon$$

— *in a sample-efficient manner*

Challenges of offline RL

- **Distribution shift**:

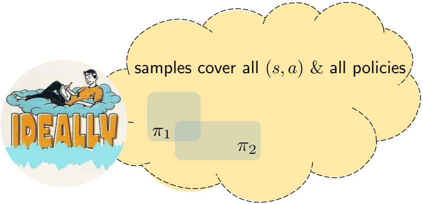
$$\text{distribution}(\mathcal{D}) \;\neq\; \text{target distribution under } \pi^{\star}$$

Challenges of offline RL

- **Distribution shift**:

$$\text{distribution}(\mathcal{D}) \neq \text{target distribution under } \pi^\star$$

- **Partial coverage of state-action space**:



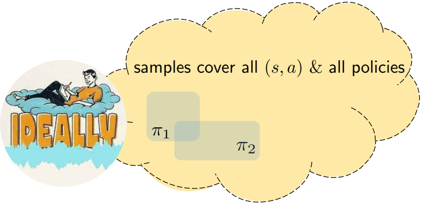
uniform coverage over entire space
(sufficiently explored)

Challenges of offline RL

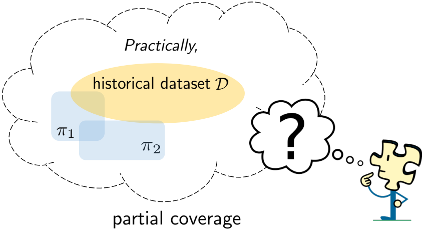
- **Distribution shift**:

$$\text{distribution}(\mathcal{D}) \;\neq\; \text{target distribution under } \pi^\star$$

- **Partial coverage of state-action space**:



uniform coverage over entire space
(sufficiently explored)



partial coverage
(inadequately explored)

How to quantify the distribution shift?

Single-policy concentrability coefficient (Rashidineiad et al.)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1$$

where $d^\pi(s,a) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}\big((s^t,a^t)=(s,a)\,|\,\pi\big)$ *is the state-action occupation density of policy* π.

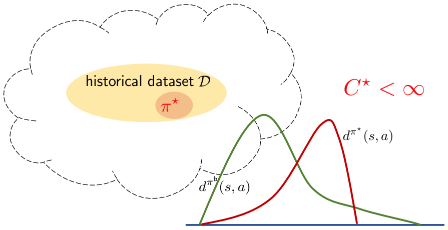
How to quantify the distribution shift?

Single-policy concentrability coefficient (Rashidineiad et al.)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^\mathsf{b}}(s,a)} \geq 1$$

where $d^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\big((s^t,a^t)=(s,a)\,|\,\pi\big)$ *is the state-action occupation density of policy* π.

- captures distribution shift
- allows for partial coverage



How to quantify the distribution shift? — a refinement

Single-policy clipped concentrability coefficient (Li et al., '22)

$$C^{\star}_{\mathsf{clipped}} := \max_{s,a} \frac{\min\{d^{\pi^\star}(s,a), 1/S\}}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1/S$$

where $d^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\big((s^t,a^t)=(s,a)\,|\,\pi\big)$ *is the state-action occupation density of policy* π*.*

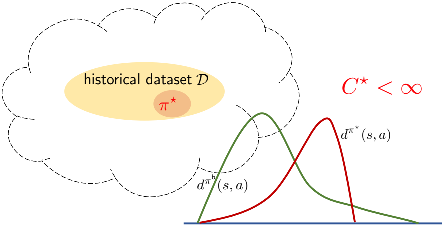
How to quantify the distribution shift? — a refinement

Single-policy clipped concentrability coefficient (Li et al., '22)

$$C^{\star}_{\text{clipped}} := \max_{s,a} \frac{\min\{d^{\pi^\star}(s,a), 1/S\}}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1/S$$

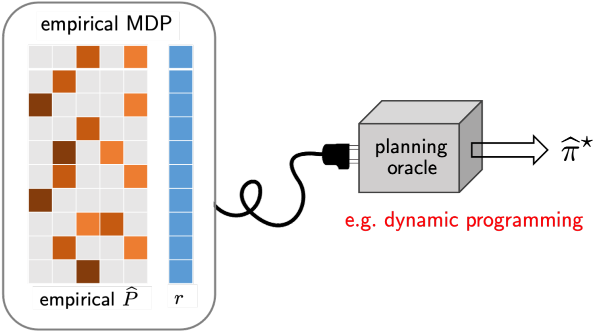
where $d^{\pi}(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}\big((s^t,a^t)=(s,a)\,|\,\pi\big)$ *is the state-action occupation density of policy* π*.*

- captures distribution shift
- allows for partial coverage
- $C^{\star}_{\text{clipped}} \leq C^{\star}$



A "plug-in" model-based approach

— (Azar et al. '13, Agarwal et al. '19, Li et al. '20)



Planning (e.g., value iteration) based on the the empirical MDP $\widehat{P}$:

$$\widehat{Q}(s,a) \leftarrow r(s,a) + \gamma \big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V} \big\rangle, \quad \widehat{V}(s) = \max_a \widehat{Q}(s,a).$$

Issue: poor value estimates under partial and poor coverage.

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*


online

upper confidence bounds

— promote exploration of under-explored (s, a)

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

online

upper confidence bounds

— promote exploration of under-explored (s, a)

offline

lower confidence bounds

— stay cautious about under-explored (s, a)

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

A model-based offline algorithm: VI-LCB

1. build empirical model $\widehat{P}$
2. **(value iteration)** for $t \le \tau_{\max}$:

$$\widehat{Q}_t(s,a) \;\leftarrow\; \Big[r(s,a) + \gamma \big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}_{t-1} \big\rangle \Big]_+$$

for all (s,a), where $\widehat{V}_t(s) = \max_a \widehat{Q}_t(s,a)$

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

A model-based offline algorithm: VI-LCB

1. build empirical model $\widehat{P}$
2. **(pessimistic value iteration)** for $t \le \tau_{\max}$:

$$\widehat{Q}_t(s,a) \leftarrow \Big[r(s,a) + \gamma \big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}_{t-1} \big\rangle - \underbrace{b(s,a;\widehat{V}_{t-1})}_{\text{penalize poorly visited } (s,a)} \Big]_+$$

for all (s,a), where $\widehat{V}_t(s) = \max_a \widehat{Q}_t(s,a)$

Key idea: pessimism in the face of uncertainty

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*

A model-based offline algorithm: VI-LCB

1. build empirical model $\widehat{P}$
2. **(pessimistic value iteration)** for $t \leq \tau_{\max}$:

$$\widehat{Q}_t(s,a) \leftarrow \Big[r(s,a) + \gamma \big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}_{t-1} \big\rangle - \underbrace{b(s,a;\widehat{V}_{t-1})}_{\text{penalize poorly visited } (s,a)} \Big]_+$$

compared w/ prior works

- no need of variance reduction
- variance-aware penalty

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *the policy* $\widehat{\pi}$ *returned by VI-LCB achieves*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right)$$

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $0 < \varepsilon \le \frac{1}{1-\gamma}$*, the policy* $\widehat{\pi}$ *returned by VI-LCB achieves*

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \le \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right)$$

- depends on distribution shift (as reflected by $C^\star_{\mathsf{clipped}}$)
- full ε-range (no burn-in cost)

Minimax optimality of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $\gamma \in [2/3, 1)$, $S \geq 2$, $C^{\star}_{\mathsf{clipped}} \geq 8\gamma/S$, *and* $0 < \varepsilon \leq \frac{1}{42(1-\gamma)}$, *there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below*

$$\widetilde{\Omega}\left(\frac{SC^{\star}_{\mathsf{clipped}}}{(1-\gamma)^{3}\varepsilon^{2}}\right).$$

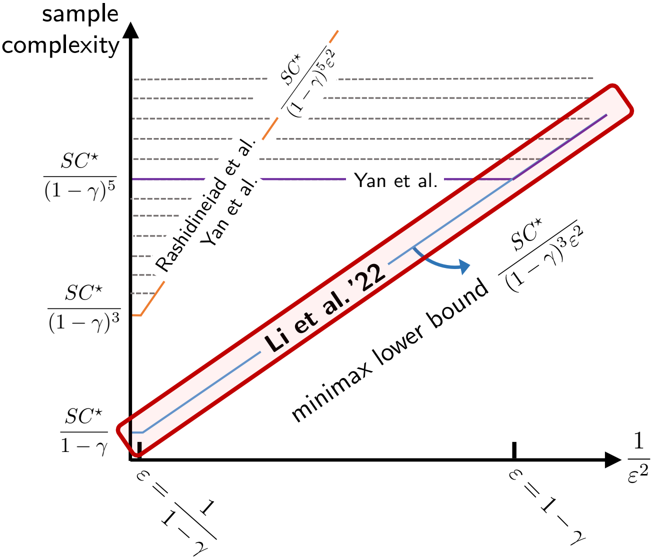
Minimax optimality of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $\gamma \in [2/3, 1)$, $S \geq 2$, $C^{\star}_{\mathsf{clipped}} \geq 8\gamma/S$, and $0 < \varepsilon \leq \frac{1}{42(1-\gamma)}$, there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below

$$\widetilde{\Omega}\left(\frac{SC^{\star}_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right).$$

- verifies the near-minimax optimality of the pessimistic model-based algorithm
- improves upon prior results by allowing $C^{\star}_{\mathsf{clipped}} \asymp 1/S$.



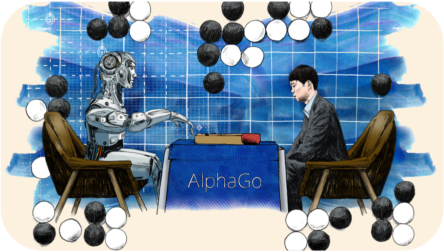
sample complexity
$\frac{SC^\star}{(1-\gamma)^5}$
$\frac{SC^\star}{(1-\gamma)^3}$
$\frac{SC^\star}{1-\gamma}$
$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$
Rashidinejad et al.
Yan et al.
Yan et al.
Li et al. '22
minimax lower bound
$\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}$
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = 1-\gamma$

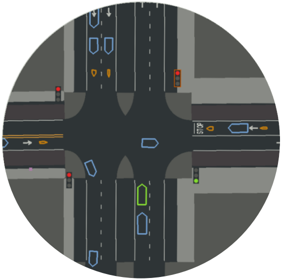
Part 2

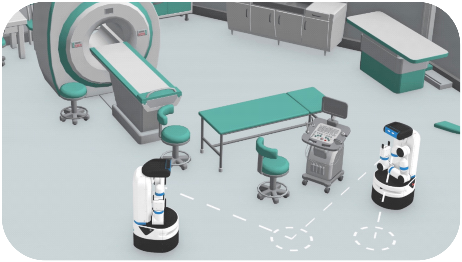
Four variants of our basics settings to illustrate the approaches so far:

- Online RL
- Offline / batch RL
- Multi-agent RL
- Robust RL

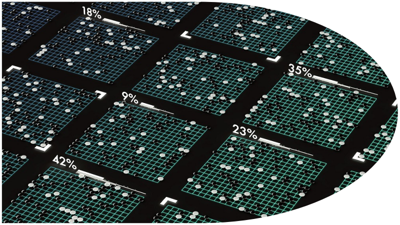
Multi-agent reinforcement learning (MARL)







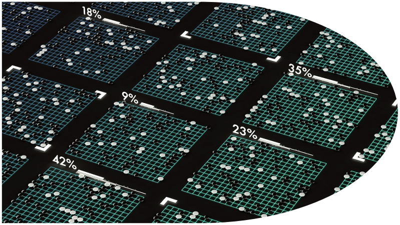
Challenges



In MARL, agents learn by probing the (shared) environment

- unknown or changing environment
- delayed feedback
- explosion of dimensionality

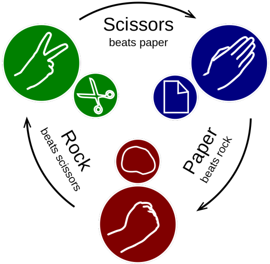
Challenges



In MARL, agents learn by probing the (shared) environment

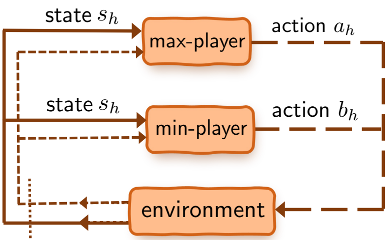
- unknown or changing environment
- delayed feedback
- explosion of dimensionality
- **curse of multiple agents**

Background: two-player zero-sum Markov games



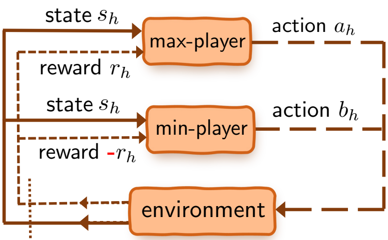
			
	0	-1	1
	1	0	-1
	-1	1	0

Two-player zero-sum Markov games



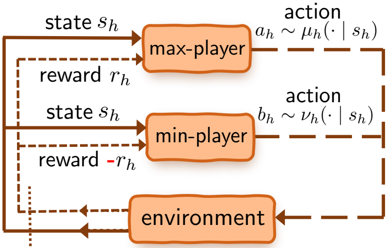
- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player

Two-player zero-sum Markov games



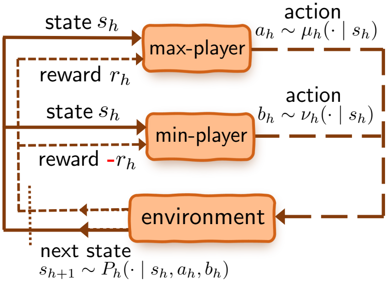
- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s, a, b) \in [0, 1]$
  min-player $-r(s, a, b)$

Two-player zero-sum Markov games



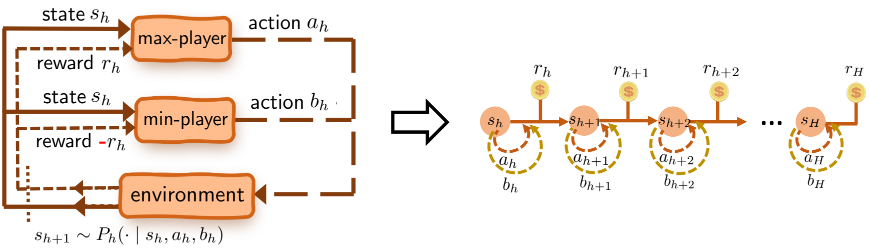
- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s, a, b) \in [0, 1]$
  min-player $-r(s, a, b)$
- $\mu : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$: policy of max-player
  $\nu : \mathcal{S} \times [H] \to \Delta(\mathcal{B})$: policy of min-player

Two-player zero-sum Markov games



- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s, a, b) \in [0, 1]$
  min-player $-r(s, a, b)$
- $\mu : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$: policy of max-player
  $\nu : \mathcal{S} \times [H] \to \Delta(\mathcal{B})$: policy of min-player
- $P_h(\cdot \mid s, a, b)$: unknown transition probabilities

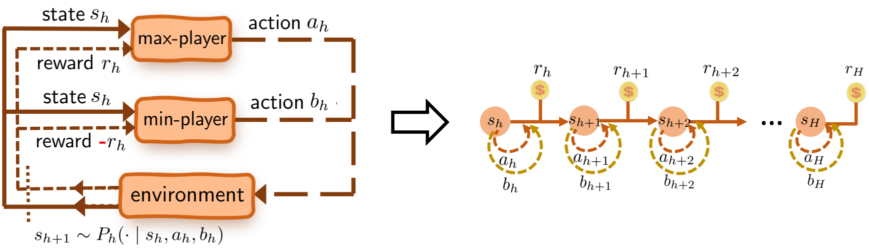
Value function & Q-function



Value function of policy pair (μ, ν):

$$V_1^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=1}^{H} r(s_t, a_t, b_t) \,\middle|\, s_1 = s\right]$$

Value function & Q-function

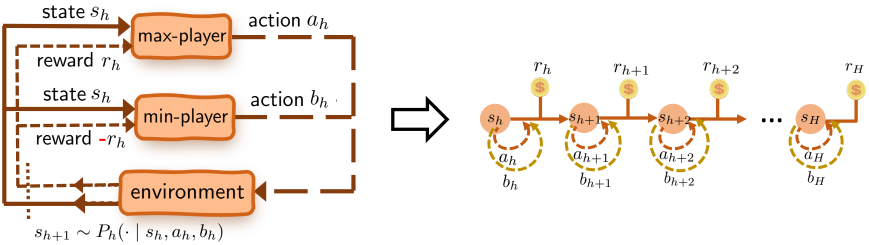


Value function of policy pair (μ, ν):

$$V_1^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=1}^{H} r(s_t, a_t, b_t) \,\middle|\, s_1 = s\right]$$

- $(a_1, b_1, s_2, \cdots)$: generated when max-player and min-player execute policies μ and ν *independently (i.e., no coordination)*

Value function & Q-function



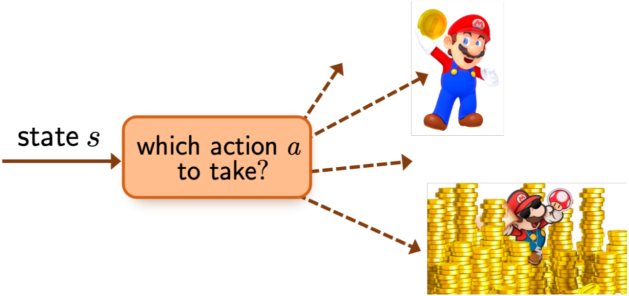
Value function and **Q function** of policy pair (μ, ν):

$$V_1^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=1}^{H} r(s_t, a_t, b_t) \,\middle|\, s_1 = s\right]$$

$$Q_1^{\mu,\nu}(s,a,b) := \mathbb{E}\left[\sum_{t=1}^{H} r(s_t, a_t, b_t) \,\middle|\, s_1 = s, a_1 = a, b_1 = b\right]$$

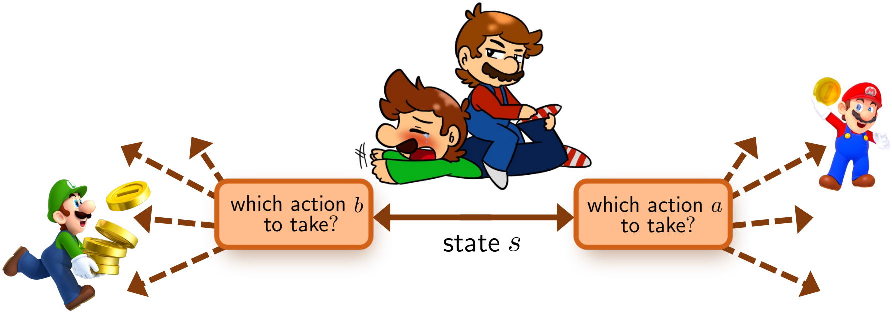
- $(a_1, b_1, s_2, \cdots)$: generated when max-player and min-player execute policies μ and ν *independently (i.e., no coordination)*

Optimal policy?



- Each agent seeks **optimal policy** maximizing her own value

Optimal policy?



- Each agent seeks **optimal policy** maximizing her own value
- But two agents have conflicting goals ...

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

$$\max_{\mu} V^{\mu,\nu^\star} = V^{\mu^\star,\nu^\star} = \min_{\nu} V^{\mu^\star,\nu}$$

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

$$\max_{\mu} V^{\mu,\nu^\star} = V^{\mu^\star,\nu^\star} = \min_{\nu} V^{\mu^\star,\nu}$$

- no unilateral deviation is beneficial

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

$$\max_{\mu} V^{\mu,\nu^\star} = V^{\mu^\star,\nu^\star} = \min_{\nu} V^{\mu^\star,\nu}$$

- no unilateral deviation is beneficial
- no coordination between two agents (they act *independently*)

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

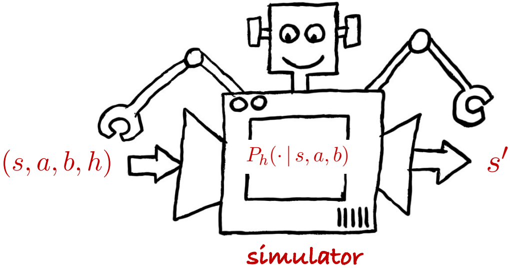
An ε-NE policy pair $(\widehat{\mu}, \widehat{\nu})$ obeys

$$\max_{\mu} V^{\mu,\widehat{\nu}} - \varepsilon \leq V^{\widehat{\mu},\widehat{\nu}} \leq \min_{\nu} V^{\widehat{\mu},\nu} + \varepsilon$$

- no unilateral deviation is beneficial
- no coordination between two agents (they act *independently*)

Sampling mechanism: a generative model / simulator

— *Kearns, Singh '99*

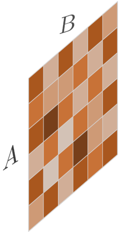


One can query generative model w/ state-action-step tuple (s, a, b, h), and obtain $s' \overset{\text{ind.}}{\sim} P_h(s' \mid s, a, b)$

Question: *how many samples are sufficient to learn an ε-Nash policy pair?*

Model-based approach w/ non-adaptive sampling

— Zhang, Kakade, Başar, Yang '20

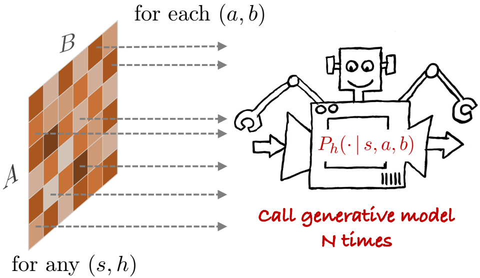


for any (s, h)

1. for each (s, a, b, h), call generative models N times

Model-based approach w/ non-adaptive sampling

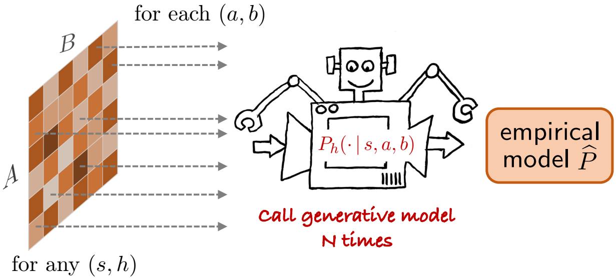
— Zhang, Kakade, Başar, Yang '20



1. for each (s, a, b, h), call generative models N times

Model-based approach w/ non-adaptive sampling

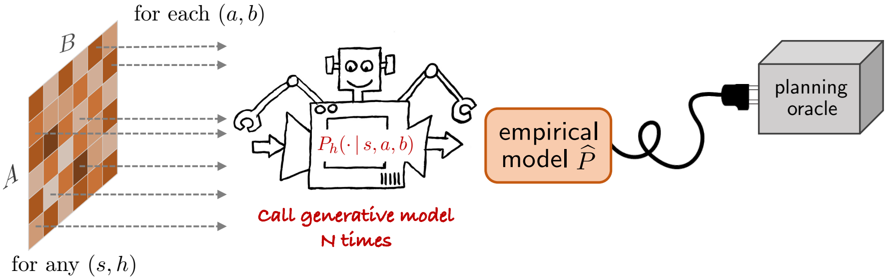
— *Zhang, Kakade, Başar, Yang '20*



1. for each (s, a, b, h), call generative models N times
2. build empirical model $\widehat{P}$

Model-based approach w/ non-adaptive sampling

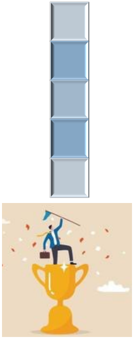
— *Zhang, Kakade, Başar, Yang '20*



1. for each (s, a, b, h), call generative models N times
2. build empirical model $\widehat{P}$, and run classical planning algorithms

sample complexity: $\frac{H^4 S A B}{\varepsilon^2}$

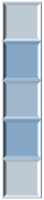
Curse of multiple agents



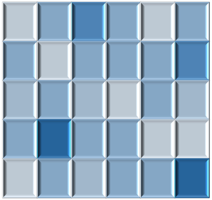
1 player: A

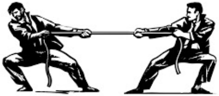
Let's look at the size of joint action space ...

Curse of multiple agents



1 player: A

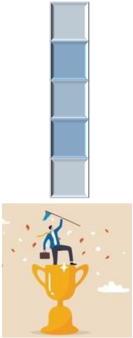




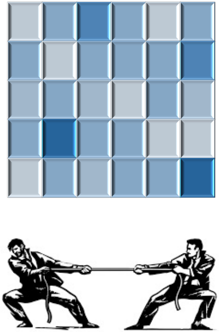
2 players: AB

Let's look at the size of joint action space . . .

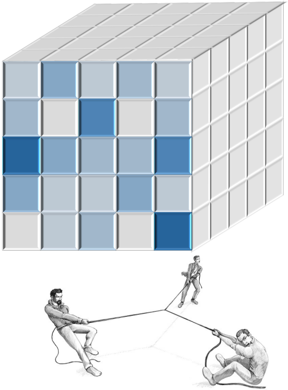
Curse of multiple agents



1 player: A



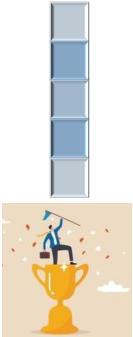
2 players: AB



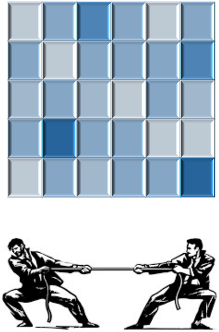
3 players: $A_1A_2A_3$

Let's look at the size of joint action space ...

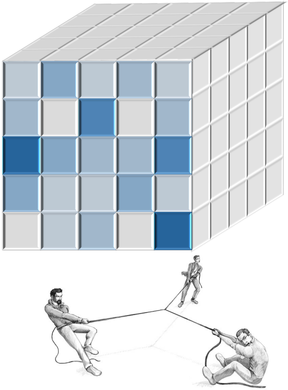
Curse of multiple agents



1 player: A



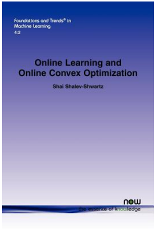
2 players: AB



3 players: $A_1 A_2 A_3$

The number of joint actions blows up geometrically in # players!

Breaking curse of multi-agents?

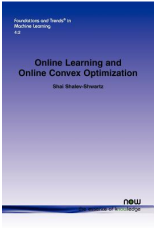


— *Song, Mei, Bai '21, Jin, Liu, Wang, Yu '21, . . .*

V-learning: overcomes curse of multi-agents in *online* RL

- estimate V-function only (much lower-dimensional than Q)

Breaking curse of multi-agents?

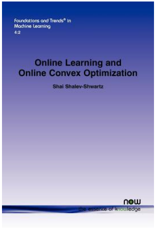


— *Song, Mei, Bai '21, Jin, Liu, Wang, Yu '21, . . .*

V-learning: overcomes curse of multi-agents in *online* RL

- estimate V-function only (much lower-dimensional than Q)
- *adaptive sampling:* take sample based on current policy iterates

Breaking curse of multi-agents?

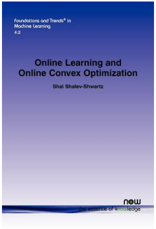


— *Song, Mei, Bai '21, Jin, Liu, Wang, Yu '21, . . .*

V-learning: overcomes curse of multi-agents in *online* RL

- estimate V-function only (much lower-dimensional than Q)
- *adaptive sampling:* take sample based on current policy iterates
- *adversarial learning subroutine*: Follow-the-Regularized-Leader

Breaking curse of multi-agents?

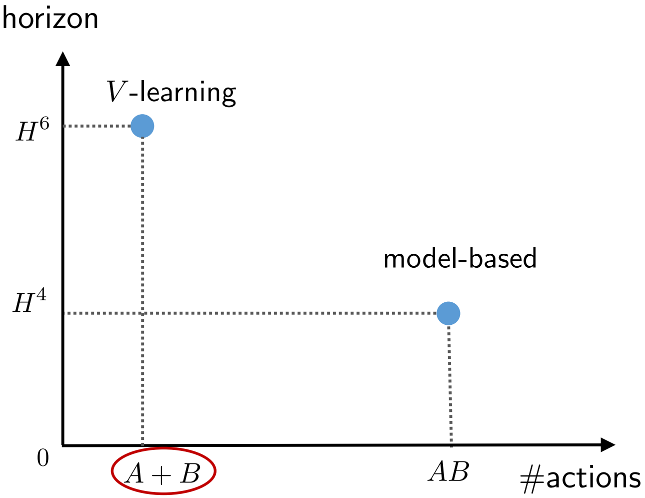


— *Song, Mei, Bai '21, Jin, Liu, Wang, Yu '21, . . .*

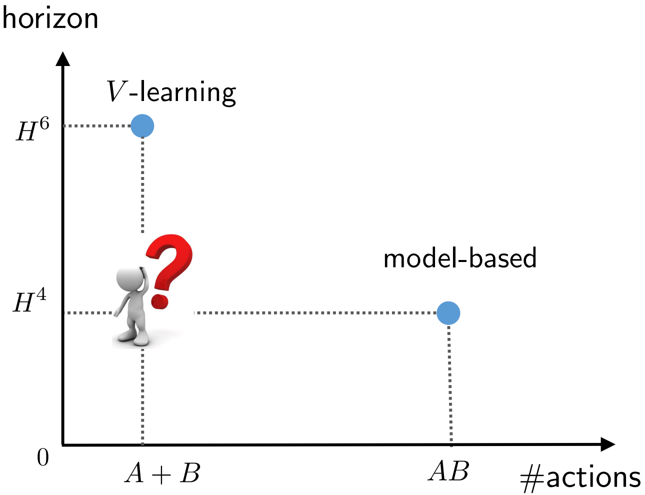
V-learning: overcomes curse of multi-agents in *online* RL

- estimate V-function only (much lower-dimensional than Q)
- *adaptive sampling:* take sample based on current policy iterates
- *adversarial learning subroutine*: Follow-the-Regularized-Leader

sample complexity: $\frac{H^6 S(A+B)}{\varepsilon^2}$ samples or $\frac{H^5 S(A+B)}{\varepsilon^2}$ episodes



horizon
V-learning
H^6
model-based
H^4
0
$A+B$
AB
#actions



Can we simultaneously overcome curse of multi-agents & barrier of long horizon?

Our algorithm

Key ingredients:

- for each player, estimate only **one-sided objects**
  - e.g. $Q(s,a)$ as opposed to $Q(s,a,b)$

Our algorithm

Key ingredients:

- for each player, estimate only **one-sided objects**
  - e.g. $Q(s,a)$ as opposed to $Q(s,a,b)$
- **adaptive sampling**
  - sampling based on current policy iterates

Our algorithm

Key ingredients:

- for each player, estimate only **one-sided objects**
  - e.g. $Q(s,a)$ as opposed to $Q(s,a,b)$
- **adaptive sampling**
  - sampling based on current policy iterates
- **adversarial learning subroutine** for policy updates
  - e.g. Follow-the-Regularized-Leader (FTRL)

Our algorithm

Key ingredients:

- for each player, estimate only **one-sided objects**
  - e.g. $Q(s, a)$ as opposed to $Q(s, a, b)$
- **adaptive sampling**
  - sampling based on current policy iterates
- **adversarial learning subroutine** for policy updates
  - e.g. Follow-the-Regularized-Leader (FTRL)
- **optimism principle** in value estimation
  - upper confidence bounds (UCB)

Main result (two-player zero-sum Markov games)

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by the proposed algorithm is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$$

Main result (two-player zero-sum Markov games)

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by the proposed algorithm is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$$

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$
- breaks curse of multi-agents & long-horizon barrier at once!

Main result (two-player zero-sum Markov games)

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by the proposed algorithm is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$$

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$
- breaks curse of multi-agents & long-horizon barrier at once!
- full ε-range (no burn-in cost)

Main result (two-player zero-sum Markov games)

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \le H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by the proposed algorithm is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$$

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$
- breaks curse of multi-agents & long-horizon barrier at once!
- full ε-range (no burn-in cost)
- other features: Markov policy, decentralized, ...

Extension: m-player general-sum Markov games

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the joint policy $\widehat{\pi}$ returned by the proposed algorithm is ε-CCE, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S \sum_i A_i}{\varepsilon^2}\right)$$

Extension: m-player general-sum Markov games

Theorem (Li, Chi, Wei, Chen '22)

For any $0 < \varepsilon \leq H$, the joint policy $\widehat{\pi}$ returned by the proposed algorithm is ε-CCE, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S \sum_i A_i}{\varepsilon^2}\right)$$

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S \max_i A_i}{\varepsilon^2}\right)$
- near-optimal when number of players m is fixed

Part 2

1. Online RL
2. Offline RL
3. Multi-agent RL
4. Robust RL

Safety and robustness in RL

(Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022;)

Training environment

$\neq$

Test environment

Safety and robustness in RL

(Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022;)

Training environment

$\neq$

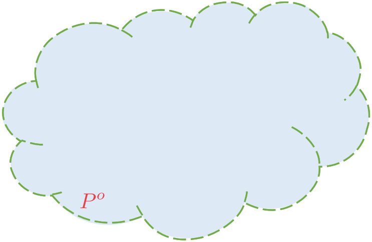
Test environment

Sim2Real Gap: Can we learn optimal policies that are robust to model perturbations?

Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

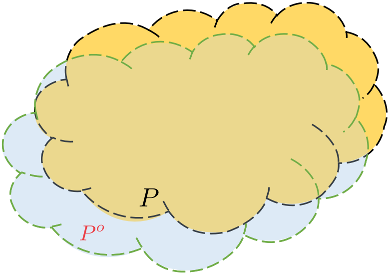
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \le \sigma\}$$

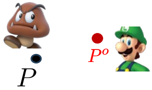


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

$$\mathcal{U}^\sigma(P^o) = \left\{P : \ \rho\left(P, P^o\right) \leq \sigma\right\}$$

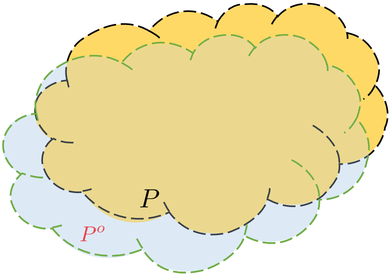


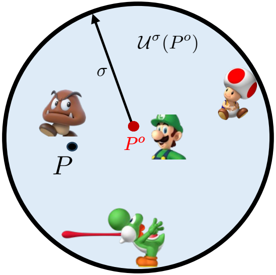


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$

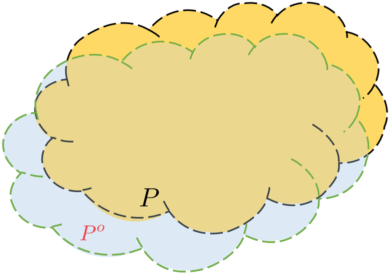


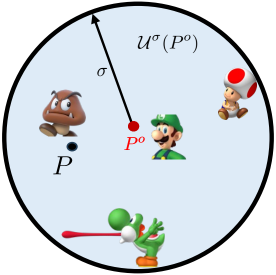


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

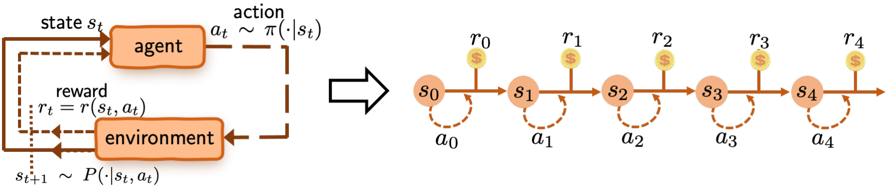
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \le \sigma\}$$





- Examples of ρ: f-divergence (TV, χ^2, KL...)

Robust value/Q function



Robust value/Q function of policy π:

$$\forall s \in \mathcal{S}: \qquad V^{\pi,\sigma}(s) := \inf_{P \in \mathcal{U}^{\sigma}(P^{o})} \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right]$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^{\pi,\sigma}(s,a) := \inf_{P \in \mathcal{U}^{\sigma}(P^{o})} \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

Measures the worst-case performance of the policy in the uncertainty set.

Distributionally robust MDP

Find the policy $\pi^{\star}$ that maximizes $V^{\pi,\sigma}$

(Iyengar. '05, Nilim and El Ghaoui. '05)

Distributionally robust MDP

Find the policy $\pi^\star$ that maximizes $V^{\pi,\sigma}$

(Iyengar. '05, Nilim and El Ghaoui. '05)

Robust Bellman's optimality equation: the optimal robust policy $\pi^\star$ and optimal robust value $V^{\star,\sigma} := V^{\pi^\star,\sigma}$ satisfy

$$Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \left\langle P_{s,a}, V^{\star,\sigma} \right\rangle ,$$

$$V^{\star,\sigma}(s) = \max_a Q^{\star,\sigma}(s,a)$$

Distributionally robust MDP

Find the policy $\pi^\star$ that maximizes $V^{\pi,\sigma}$

(Iyengar. '05, Nilim and El Ghaoui. '05)

Robust Bellman's optimality equation: the optimal robust policy $\pi^\star$ and optimal robust value $V^{\star,\sigma} := V^{\pi^\star,\sigma}$ satisfy

$$Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \left\langle P_{s,a}, V^{\star,\sigma} \right\rangle,$$

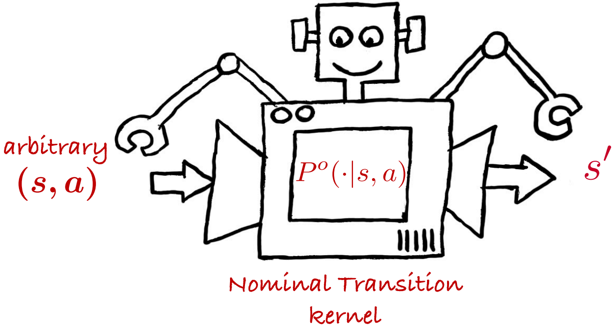
$$V^{\star,\sigma}(s) = \max_a Q^{\star,\sigma}(s,a)$$

Distributionally robust value iteration (DRVI):

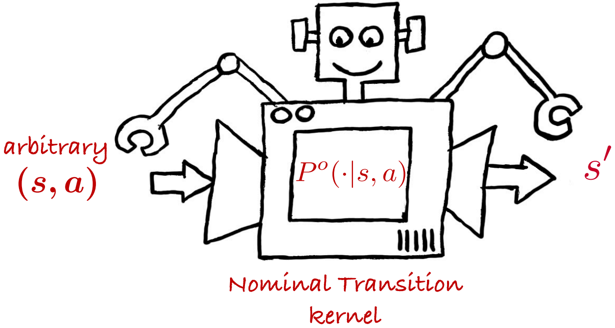
$$Q(s,a) \leftarrow r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \langle P_{s,a}, V \rangle,$$

where $V(s) = \max_a Q(s,a)$.

Learning distributionally robust MDPs



Learning distributionally robust MDPs

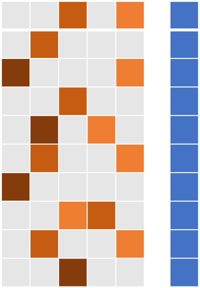


Goal of robust RL: given $\mathcal{D} := \{(s_i, a_i, s_i')\}_{i=1}^N$ from the *nominal* environment P^0, find an ε-optimal robust policy $\widehat{\pi}$ obeying

$$V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \leq \varepsilon$$

— *in a sample-efficient manner*

A curious question

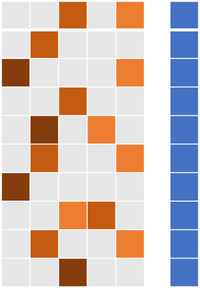


empirical MDP

Learn the optimal policy of the nominal MDP?

Learn the **robust** policy around the nominal MDP?

A curious question



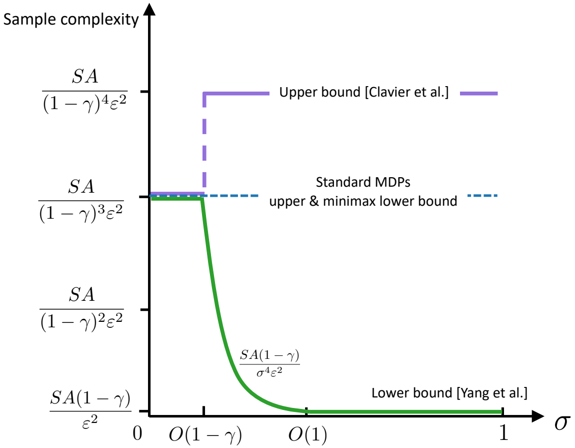
empirical MDP

Learn the optimal policy of the nominal MDP?

Learn the **robust** policy around the nominal MDP?

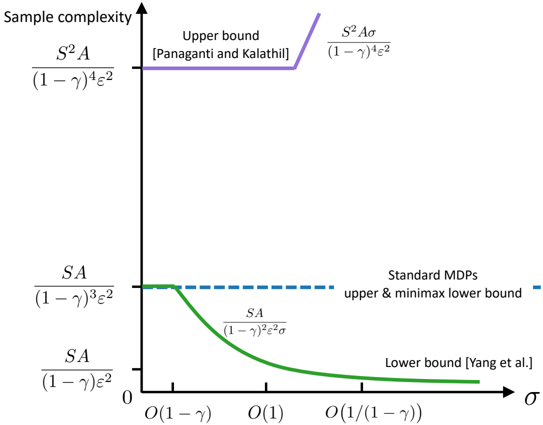
Robustness-statistical trade-off? Is there a statistical premium that one needs to pay in quest of additional robustness?

Prior art: TV uncertainty



- Large gaps between existing upper and lower bounds
- Unclear benchmarking with standard MDP

Prior art: χ^2 uncertainty



- Large gaps between existing upper and lower bounds
- Unclear benchmarking with standard MDP

Our theorem under TV uncertainty

Theorem (Shi et al., 2023)

Assume the uncertainty set is measured via the TV distance with radius $\sigma \in [0, 1)$. For sufficiently small $\varepsilon > 0$, DRVI outputs a policy $\widehat{\pi}$ that satisfies $V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \leq \varepsilon$ with sample complexity at most

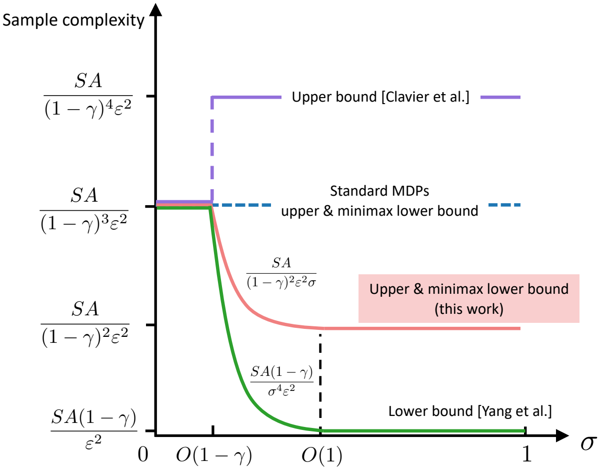
$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2}\right)$$

ignoring logarithmic factors. In addition, no algorithm can succeed if the sample size is below

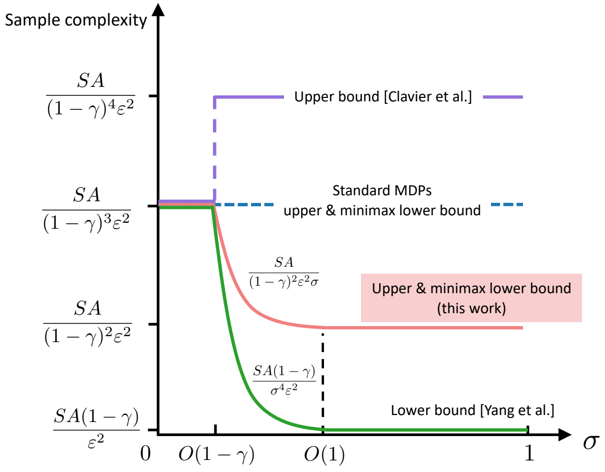
$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2}\right).$$

- Establish the minimax optimality of DRVI for RMDP under the TV uncertainty set over the full range of σ.

When the uncertainty set is TV



When the uncertainty set is TV



RMDPs are **easier** to learn than standard MDPs.

Our theorem under χ^2 uncertainty

Theorem (Upper bound, Shi et al., 2023)

Assume the uncertainty set is measured via the χ^2 divergence with radius $\sigma \in [0, \infty)$. For sufficiently small $\varepsilon > 0$, DRVI outputs a policy $\widehat{\pi}$ that satisfies $V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \leq \varepsilon$ with sample complexity at most

$$\widetilde{O}\left(\frac{SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\right)$$

ignoring logarithmic factors.

Our theorem under χ^2 uncertainty

Theorem (Upper bound, Shi et al., 2023)

Assume the uncertainty set is measured via the χ^2 divergence with radius $\sigma \in [0, \infty)$. For sufficiently small $\varepsilon > 0$, DRVI outputs a policy $\widehat{\pi}$ that satisfies $V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \leq \varepsilon$ with sample complexity at most

$$\widetilde{O}\left(\frac{SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\right)$$

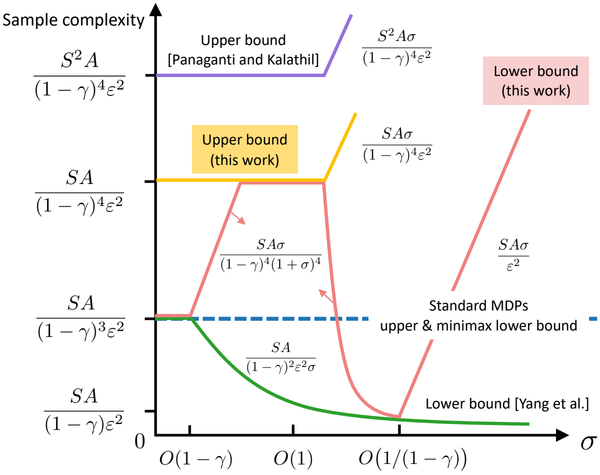
ignoring logarithmic factors.

Theorem (Lower bound, Shi et al., 2023)

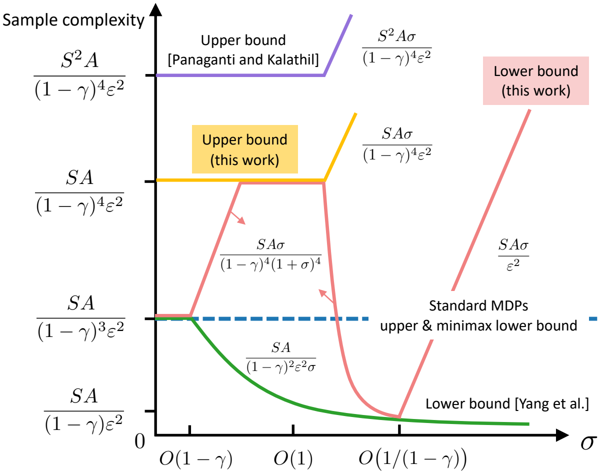
In addition, no algorithm succeeds when the sample size is below

$$\begin{cases} \widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right) & \text{if } \sigma \lesssim 1-\gamma \\ \widetilde{\Omega}\left(\frac{\sigma SA}{\min\{1,(1-\gamma)^4(1+\sigma)^4\}\varepsilon^2}\right) & \text{otherwise} \end{cases}$$

When the uncertainty set is χ^2 divergence



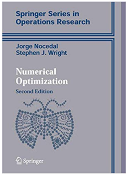
When the uncertainty set is χ^2 divergence

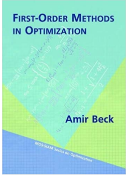


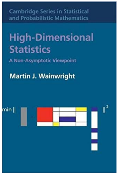
RMDPs can be **harder** to learn than standard MDPs.

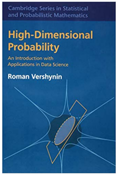
Concluding remarks

This tutorial







(large-scale) optimization

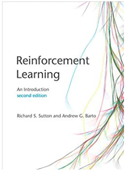
(high-dimensional) statistics

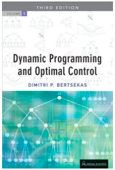
Demystify sample- and computational efficiency of RL algorithms

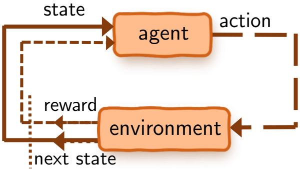
Part 1. basics, RL w/ a generative model

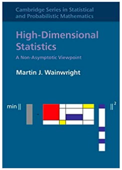
Part 2. online / offline RL, multi-agent / robust RL

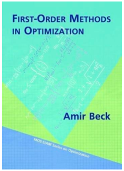
Concluding remarks











Understanding non-asymptotic performances of RL algorithms is a fruitful playground!

Beyond the tabular setting

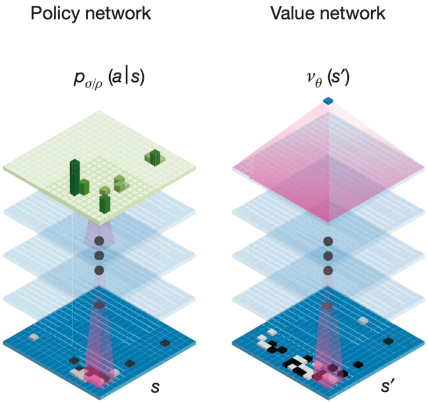
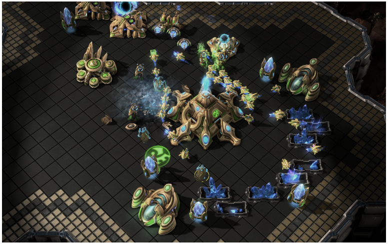


Figure credit: (Silver et al., 2016)

- function approximation for dimensionality reduction
- Provably efficient RL algorithms under minimal assumptions

(Osband and Van Roy, 2014; Dai et al., 2018; Du et al., 2019; Jin et al., 2020)

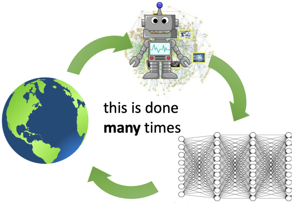
Multi-agent RL



- **Competitive setting:** finding Nash equilibria for Markov games
- **Collaborative setting:** multiple agents jointly optimize the policy to maximize the total reward

(Zhang, Yang, and Basar, 2021; Cen, Wei, and Chi, 2021)

Hybrid RL

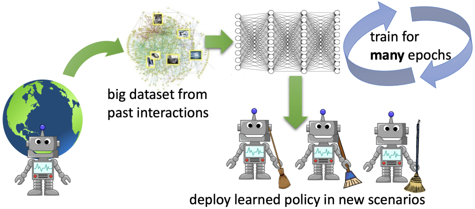


Online RL

- interact with environment
- actively collect new data

Offline/Batch RL

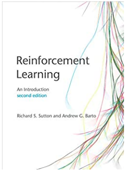
- no interaction
- data is given

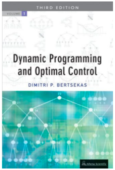


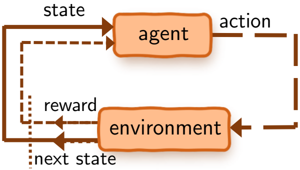
Can we achieve the best of both worlds?

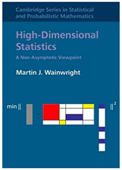
(Wagenmaker and Pacchiano, 2022; Song et al., 2022; Li et al., 2023)

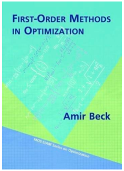
Concluding remarks











Understanding non-asymptotic performances of RL algorithms is a fruitful playground!

Promising directions:

- function approximation
- multi-agent/federated RL
- hybrid RL
- many more...

Thank you for your attention! https://yutingwei.github.io/

Reference: online RL I

- "*Asymptotically efficient adaptive allocation rules*," T. L. Lai, H. Robbins, *Advances in applied mathematics*, vol. 6, no. 1, 1985
- "*Finite-time analysis of the multiarmed bandit problem*," P. Auer, N. Cesa-Bianchi, P. Fischer, *Machine learning*, vol. 47, pp. 235-256, 2002
- "*Minimax regret bounds for reinforcement learning*," M. G. Azar, I. Osband, R. Munos, *ICML*, 2017
- "*Is Q-learning provably efficient?*" C. Jin, Z. Allen-Zhu, S. Bubeck, and M. Jordan, *NeurIPS*, 2018
- "*Provably efficient Q-learning with low switching cost*," Y. Bai, T. Xie, N. Jiang, Y. X. Wang, *NeurIPS*, 2019
- "*Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited*" O. D. Domingues, P. Menard, E. Kaufmann, M. Valko, *Algorithmic Learning Theory*, 2021
- "*Almost optimal model-free reinforcement learning via reference-advantage decomposition,*" Z. Zhang, Y. Zhou, X. Ji, *NeurIPS*, 2020

Reference: online RL II

- "*Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon,*" Z. Zhang, X. Ji, and S. Du, *COLT*, 2021
- "*Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning,*" G. Li, L. Shi, Y. Chen, Y. Gu, Y. Chi, *NeurIPS*, 2021
- "*Regret-optimal model-free reinforcement learning for discounted MDPs with short burn-in time,*" X. Ji, G. Li, *NeurIPS*, 2023
- "*Reward-free exploration for reinforcement learning*," C. Jin, A. Krishnamurthy, M. Simchowitz, T. Yu, *ICML*, 2020
- "*Minimax-optimal reward-agnostic exploration in reinforcement learning*," G. Li, Y. Yan, Y. Chen, J. Fan, *COLT*, 2024
- "*Settling the sample complexity of online reinforcement learning,*" Z. Zhang, Y. Chen, J. D. Lee, S. S. Du, *COLT*, 2024

Reference: offline RL I

- "*Bridging offline reinforcement learning and imitation learning: A tale of pessimism,*" P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, S. Russell, *NeurIPS*, 2021
- "*Is pessimism provably efficient for offline RL?*" Y. Jin, Z. Yang, Z. Wang, *ICML*, 2021
- "*Settling the sample complexity of model-based offline reinforcement learning,*" G. Li, L. Shi, Y. Chen, Y. Chi, Y. Wei, *Annals of Statistics*, vol. 52, no. 1, pp. 233-260, 2024
- "*Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity*," L. Shi, G. Li, Y. Wei, Y. Chen, Y. Chi, *ICML*, 2022
- "*The efficacy of pessimism in asynchronous Q-learning*," Y. Yan, G. Li, Y. Chen, J. Fan, *IEEE Transactions on Information Theory*, 2023
- "*Policy finetuning: Bridging sample-efficient offline and online reinforcement learning*" T. Xie, N. Jiang, H. Wang, C. Xiong, Y. Bai, *NeurIPS*, 2021

Reference: multi-agent RL I

- "*Stochastic games*," L. S. Shapley, *Proceedings of the national academy of sciences*, 1953
- "*Twenty lectures on algorithmic game theory*," T. Roughgarden, 2016
- "*Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity*," K. Zhang, S. Kakade, T. Basar, L. Yang, *NeurIPS*, 2020
- "*When can we learn general-sum Markov games with a large number of players sample-efficiently*?" Z. Song, S. Mei, Y. Bai, *ICLR*, 2021
- "*V-learning–A simple, efficient, decentralized algorithm for multiagent RL*," C. Jin, Q. Liu, Y. Wang, T. Yu, 2021
- "*Minimax-optimal multi-agent RL in Markov games with a generative model*," G. Li, Y. Chi, Y. Wei, Y. Chen, *NeurIPS*, 2022
- *When are offline two-player zero-sum Markov games solvable*?" Q. Cui, S. S. Du, *NeurIPS*, 2022
- "*Model-based reinforcement learning for offline zero-sum Markov games*," Y. Yan, G. Li, Y. Chen, J. Fan, *Operations Research*, 2024

Reference: robust RL I

- "*Robust dynamic programming*," G. Iyengar, *Mathematics of Operations Research*, 2005
- "*The curious price of distributional robustness in reinforcement learning with a generative model.*," L. Shi, G. Li, Y. Wei, Y. Chen, M. Geist, Y. Chi, *NeurIPS*, 2023
- "*Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity*," L. Shi, Y. Chi, 2022
- "*On the foundation of distributionally robust reinforcement learning*," S. Wang, N. Si, J. Blanchet, and Z. Zhou, 2023
- "*Sample complexity of robust reinforcement learning with a generative model*," K. Panaganti, D. Kalathil, *AISTATS*, 2022
- "*Sample-Efficient Robust Multi-Agent Reinforcement Learning in the Face of Environmental Uncertainty*," L. Shi, E. Mazumdar, Y. Chi, and A. Wierman, *ICML*, 2024