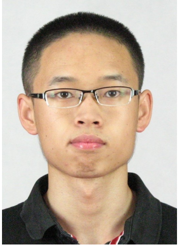
Breaking the Sample Size Barrier in Model-Based Reinforcement Learning

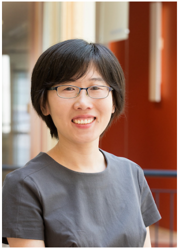
Yuting Wei

Carnegie Mellon University

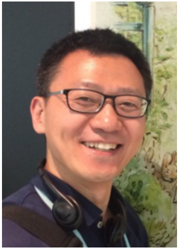
Nov, 2020



Gen Li
Tsinghua EE



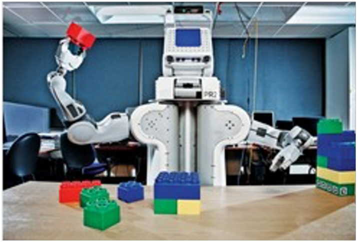
Yuejie Chi
CMU ECE



Yuantao Gu
Tsinghua EE

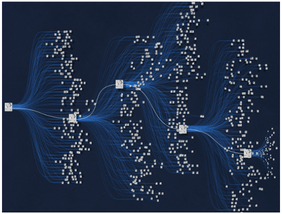
Yuxin Chen
Princeton EE

Reinforcement learning (RL)



RL challenges

- Unknown or changing environment
- Credit assignment problem
- Enormous state and action space



Provable efficiency

- Collecting samples might be expensive or impossible:
  sample efficiency
- Training deep RL algorithms might take long time:
  computational efficiency

This talk

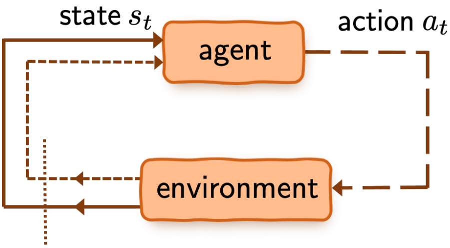
Question: can we design sample- and computation-efficient RL algorithms?

—— *inspired by numerous prior work*

[Kearns and Singh, 1999, Sidford et al., 2018a, Agarwal et al., 2019]...

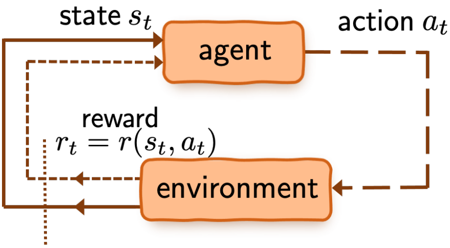
Background: Markov decision processes

Markov decision process (MDP)



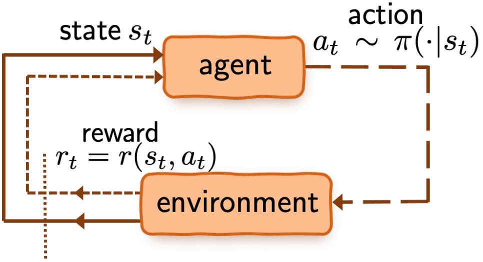
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

Markov decision process (MDP)



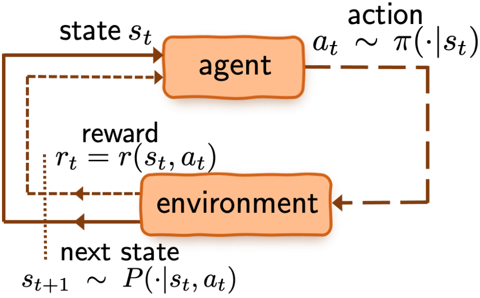
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

Markov decision process (MDP)



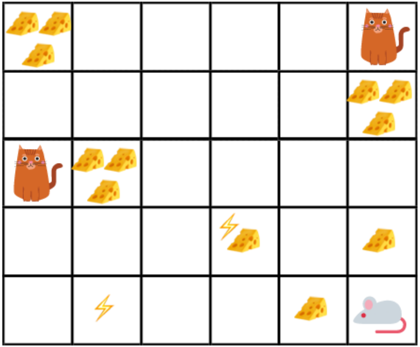
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

Markov decision process (MDP)

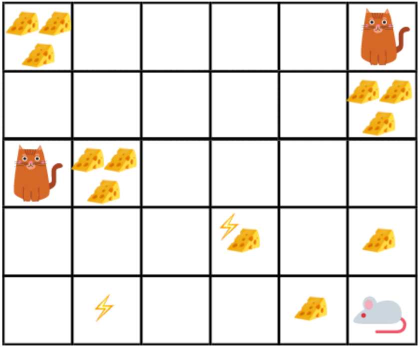


- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)
- $P(\cdot|s, a)$: unknown transition probabilities

Help the mouse!

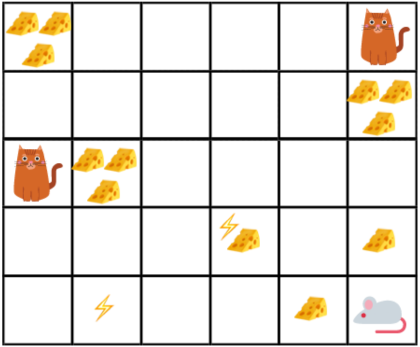


Help the mouse!



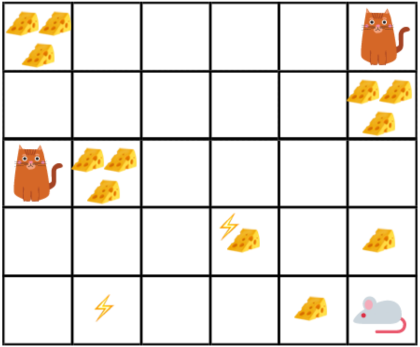
- state space $\mathcal{S}$: positions in the maze

Help the mouse!



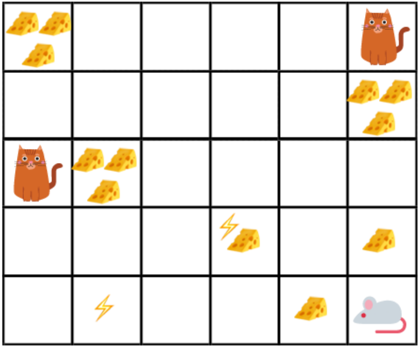
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right

Help the mouse!



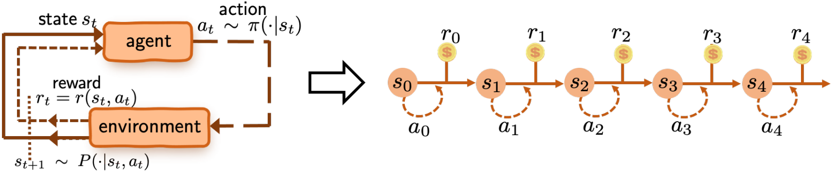
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right
- immediate reward r: cheese, electricity shocks, cats

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

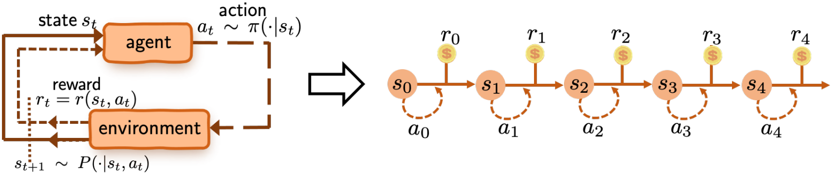
Value function



Value function of policy π: long-term discounted reward

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

Value function

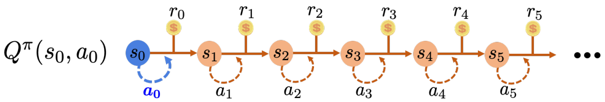


Value function of policy π: long-term discounted reward

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

- $\gamma \in [0, 1)$: discount factor
- $(a_0, s_1, a_1, s_2, a_2, \cdots)$: generated under policy π

Action-value function (a.k.a. Q-function)

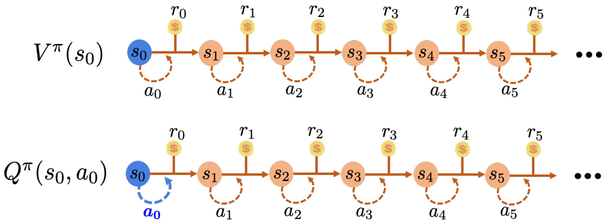


Q-function of policy π

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

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

Action-value function (a.k.a. Q-function)



Q-function of policy π

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

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

Optimal policy

Optimal policy

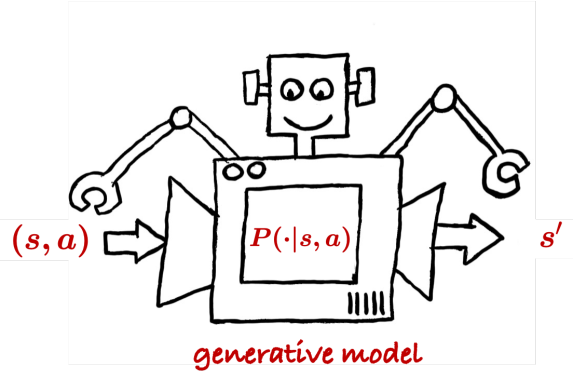
- **optimal policy** $\pi^\star$: maximizing value function

Optimal policy

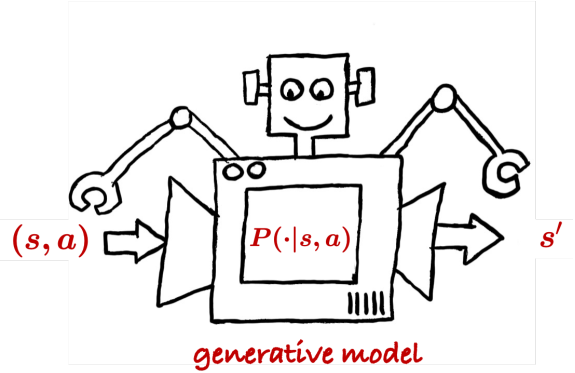
- **optimal policy** $\pi^{\star}$: maximizing value function
- optimal value / Q function: $V^{\star} := V^{\pi^{\star}}$; $Q^{\star} := Q^{\pi^{\star}}$

Practically, learn the optimal policy from data samples ...

This talk: sampling from a generative model

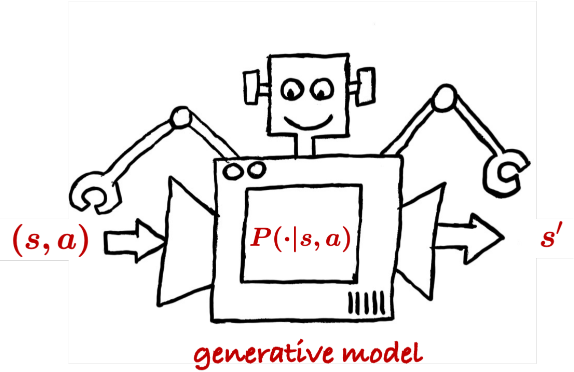


This talk: sampling from a generative model



For each state-action pair (s, a), collect N samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$

This talk: sampling from a generative model



For each state-action pair (s, a), collect N samples $\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}$

How many samples are sufficient to learn an ε-optimal policy?

An incomplete list of prior art

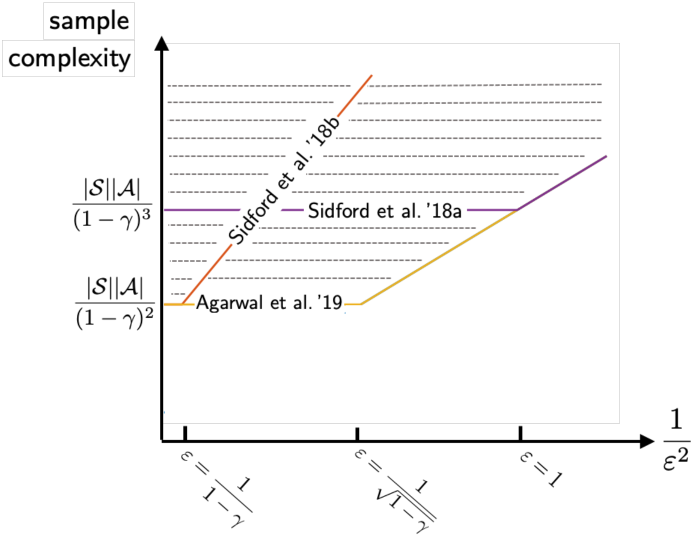
- [Kearns and Singh, 1999]
- [Kakade, 2003]
- [Kearns et al., 2002]
- [Azar et al., 2012]
- [Azar et al., 2013]
- [Sidford et al., 2018a]
- [Sidford et al., 2018b]
- [Wang, 2019]
- [Agarwal et al., 2019]
- [Wainwright, 2019a]
- [Wainwright, 2019b]
- [Pananjady and Wainwright, 2019]
- [Yang and Wang,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

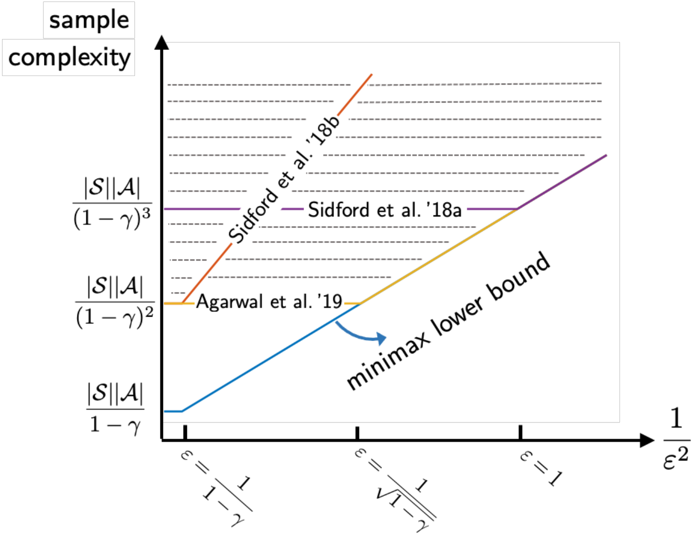
An even shorter list of prior art

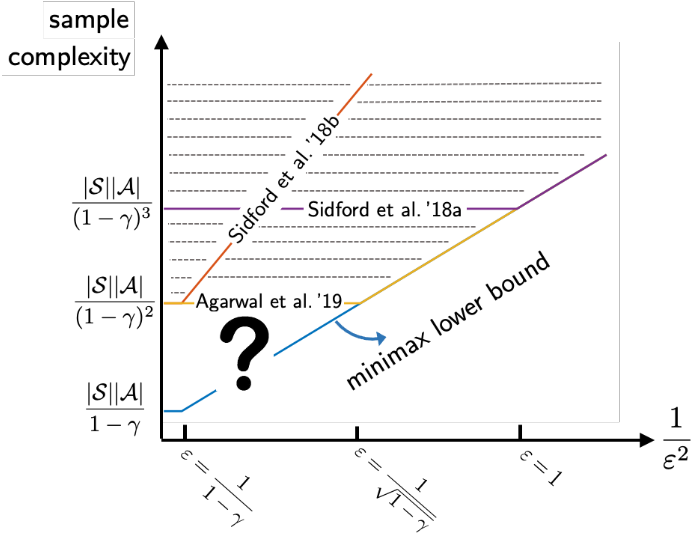
algorithm	sample size range	sample complexity	ε-range
Empirical QVI [Azar et al., 2013]	$\left[\frac{\|\mathcal{S}\|^2\|\mathcal{A}\|}{(1-\gamma)^2}, \infty\right)$	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{(1-\gamma)\|\mathcal{S}\|}}]$
Sublinear randomized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

important parameters $\Longrightarrow$

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


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
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$
$\varepsilon = 1$


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
minimax lower bound
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$
$\varepsilon = 1$

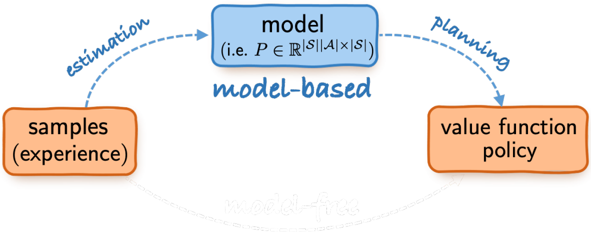


All prior theory requires sample size $> \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

sample size barrier

This talk: break the sample complexity barrier

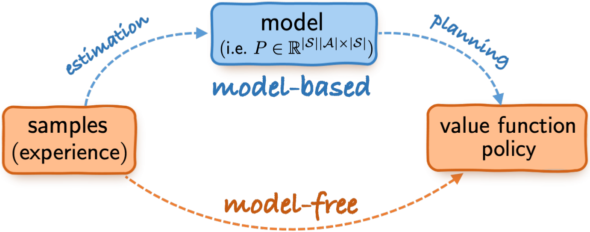
Two approaches



Model-based approach ("plug-in")

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

Two approaches



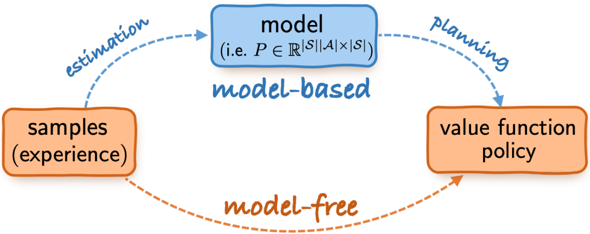
Model-based approach ("plug-in")

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

Model-free approach

— learning w/o constructing a model explicitly

Two approaches



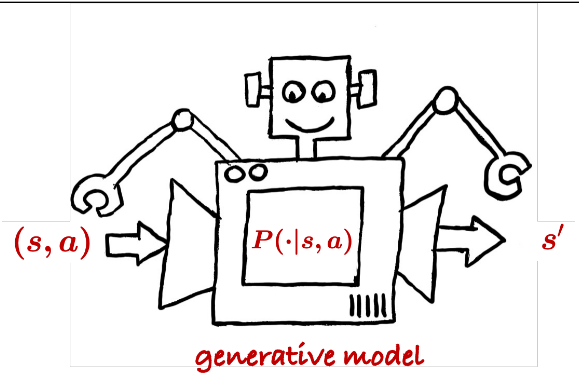
Model-based approach ("plug-in")

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

Model-free approach

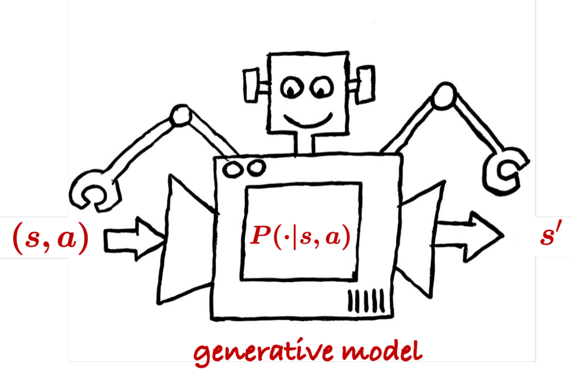
— learning w/o constructing a model explicitly

Model estimation



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

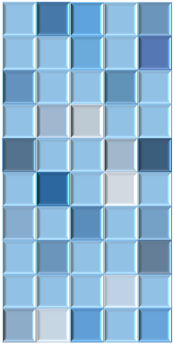
Model estimation



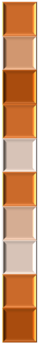
Sampling: for each (s,a), collect N ind. samples $\{(s,a,s'_{(i)})\}_{1\le i\le N}$

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

Our method: plug-in estimator + perturbation



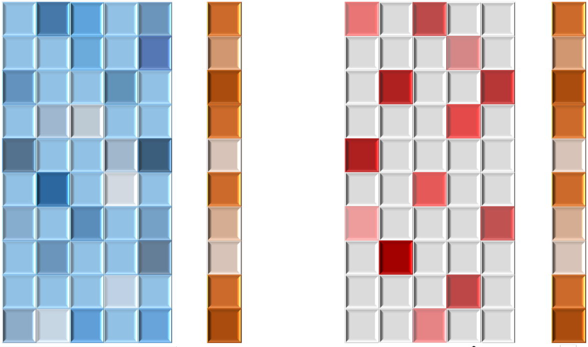
truth: $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



$r \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

original MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ $\Longrightarrow$ $\pi^\star = \arg\max_\pi V^\pi$

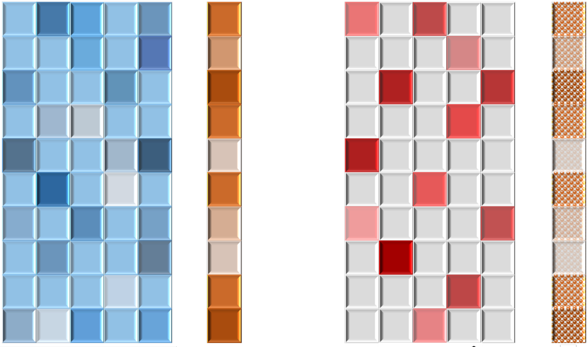
Our method: plug-in estimator + perturbation



original MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ $\Longrightarrow$ $\pi^\star = \arg\max_\pi V^\pi$

empirical MDP $(\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma)$ $\Longrightarrow$ $\widehat{\pi}^\star = \arg\max_\pi \widehat{V}^\pi$

Our method: plug-in estimator + perturbation

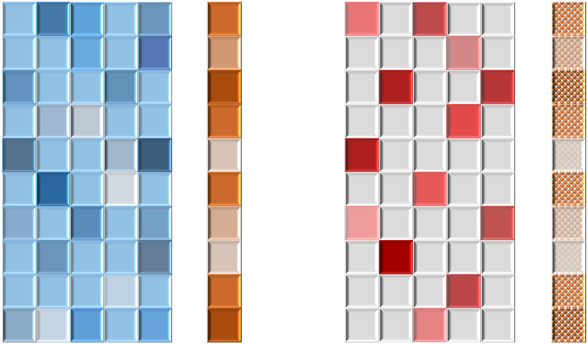


$$\text{original MDP } (\mathcal{S}, \mathcal{A}, P, r, \gamma) \implies \pi^\star = \arg\max_\pi V^\pi$$

$$\text{empirical MDP } (\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma) \implies \widehat{\pi}^\star = \arg\max_\pi \widehat{V}^\pi$$

$$\text{perturbed MDP } (\mathcal{S}, \mathcal{A}, \widehat{P}, \tilde{r}, \gamma) \implies \widehat{\pi}^\star_{\mathrm{p}} = \arg\max_\pi \widehat{V}^\pi_{\mathrm{p}}$$

Our method: plug-in estimator + perturbation



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

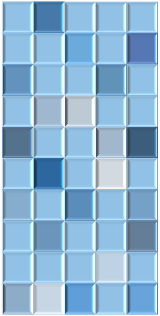
original MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ $\Longrightarrow$ $\pi^\star = \arg\max_\pi V^\pi$

empirical MDP $(\mathcal{S}, \mathcal{A}, \widehat{P}, r, \gamma)$ $\Longrightarrow$ $\widehat{\pi}^\star = \arg\max_\pi \widehat{V}^\pi$

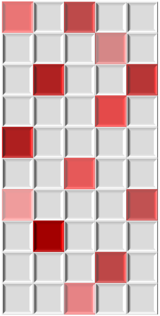
perturbed MDP $(\mathcal{S}, \mathcal{A}, \widehat{P}, \tilde{r}, \gamma)$ $\Longrightarrow$ $\widehat{\pi}^\star_{\mathrm{p}} = \arg\max_\pi \widehat{V}^\pi_{\mathrm{p}}$

planning (policy iteration, Q-value iteration, $\cdots$)

Challenges in the sample-starved regime



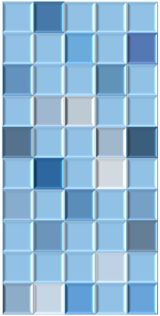
truth: $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



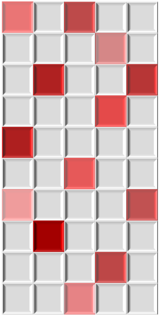
empirical estimate: $\widehat{P}$

- 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?

Main result

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

For every $0 < \varepsilon \le \frac{1}{1-\gamma}$, *policy* $\widehat{\pi}^\star_{\rm p}$ *of perturbed empirical MDP achieves*

$$\|V^{\widehat{\pi}^\star_{\rm p}} - V^\star\|_\infty \le \varepsilon \quad \textit{and} \quad \|Q^{\widehat{\pi}^\star_{\rm p}} - Q^\star\|_\infty \le \gamma\varepsilon$$

with sample complexity at most

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

Main result

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

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

$$\|V^{\widehat{\pi}_{\mathrm{p}}^{\star}} - V^{\star}\|_{\infty} \leq \varepsilon \quad \textit{and} \quad \|Q^{\widehat{\pi}_{\mathrm{p}}^{\star}} - Q^{\star}\|_{\infty} \leq \gamma\varepsilon$$

with sample complexity at most

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

- $\widehat{\pi}_{\mathrm{p}}^{\star}$: obtained by empirical QVI or PI within $\widetilde{O}\big(\frac{1}{1-\gamma}\big)$ iterations

Main result

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

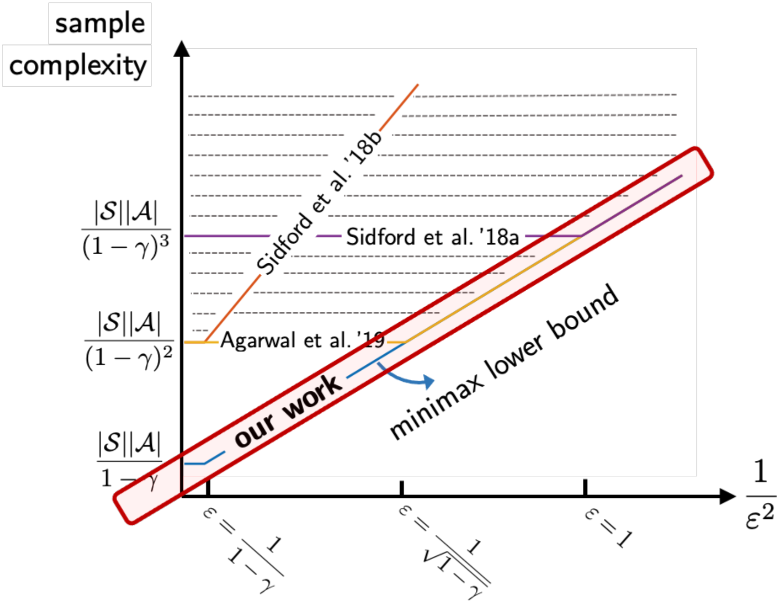
For every $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of perturbed empirical MDP achieves*

$$\|V^{\widehat{\pi}_{\mathrm{p}}^{\star}} - V^{\star}\|_{\infty} \leq \varepsilon \quad \text{and} \quad \|Q^{\widehat{\pi}_{\mathrm{p}}^{\star}} - Q^{\star}\|_{\infty} \leq \gamma\varepsilon$$

with sample complexity at most

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

- $\widehat{\pi}_{\mathrm{p}}^{\star}$: obtained by empirical QVI or PI within $\widetilde{O}\big(\frac{1}{1-\gamma}\big)$ iterations
- minimax lower bound: $\widetilde{\Omega}\big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\big)$ [Azar et al., 2013]


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
our work
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

- V^π: true value function under policy π
  - Bellman equation: $V = (I - \gamma P_\pi)^{-1} r$ [Sutton and Barto, 2018]

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V = (I - \gamma P_\pi)^{-1} r$ [Sutton and Barto, 2018]
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V} = (I - \gamma \widehat{P}_\pi)^{-1} r$

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V = (I - \gamma P_\pi)^{-1} r$ [Sutton and Barto, 2018]
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V} = (I - \gamma \widehat{P}_\pi)^{-1} r$
- $\pi^\star$: optimal policy w.r.t. true value function
- $\widehat{\pi}^\star$: optimal policy w.r.t. empirical value function

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V = (I - \gamma P_\pi)^{-1} r$ [Sutton and Barto, 2018]
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V} = (I - \gamma \widehat{P}_\pi)^{-1} r$
- $\pi^\star$: optimal policy w.r.t. true value function
- $\widehat{\pi}^\star$: optimal policy w.r.t. empirical value function
- $V^\star := V^{\pi^\star}$: optimal values under true models
- $\widehat{V}^\star := \widehat{V}^{\widehat{\pi}^\star}$: optimal values under empirical models

Proof ideas (cont.)

Elementary decomposition:

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

Proof ideas (cont.)

Elementary decomposition:

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

Proof ideas (cont.)

Elementary decomposition:

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

- **Step 1:** control $V^\pi - \widehat{V}^\pi$, for fixed π
  (**Bernstein's inequality** + **high order decomposition**)

Proof ideas (cont.)

Elementary decomposition:

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

- **Step 1:** control $V^\pi - \widehat{V}^\pi$, for fixed π
  (**Bernstein's inequality** + **high order decomposition**)
- **Step 2:** control $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$
  (**decouple statistical dependence**)

Step 1: high order decomposition

Bellman equation $V^{\pi} = (I - \gamma P_{\pi})^{-1} r$

Step 1: high order decomposition

Bellman equation $V^{\pi} = (I - \gamma P_{\pi})^{-1} r$

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

Step 1: high order decomposition

Bellman equation $V^\pi = (I - \gamma P_\pi)^{-1} r$

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

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

Step 1: high order decomposition

Bellman equation $V^\pi = (I - \gamma P_\pi)^{-1} r$

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

[ours]
$$\widehat{V}^\pi - V^\pi = \gamma \big(I - \gamma P_\pi \big)^{-1} \big(\widehat{P}_\pi - P_\pi \big) V^\pi + \\ + \gamma^2 \Big(\big(I - \gamma P_\pi \big)^{-1} \big(\widehat{P}_\pi - P_\pi \big) \Big)^2 \widehat{V}^\pi$$

Step 1: high order decomposition

Bellman equation $V^\pi = (I - \gamma P_\pi)^{-1} r$

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

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

$$+ \gamma^2\Big((I - \gamma P_\pi\big)^{-1}\big(\widehat{P}_\pi - P_\pi)\Big)^2 V^\pi$$

$$+ \gamma^2\Big((I - \gamma P_\pi\big)^{-1}\big(\widehat{P}_\pi - P_\pi)\Big)^2\big[\widehat{V}^\pi - V^\pi\big]$$

Step 1: high order decomposition

Bellman equation $V^\pi = (I - \gamma P_\pi)^{-1} r$

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

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

Step 1: high order decomposition

Bellman equation $V^\pi = (I - \gamma P_\pi)^{-1} r$

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

[ours]

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

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

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* $0 < \varepsilon \le \frac{1}{1-\gamma}$, *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]

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]
- tackle sample size barrier: prior work requires sample size $> \frac{|\mathcal{S}|}{(1-\gamma)^2}$ [Agarwal et al., 2019, 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 $\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}$

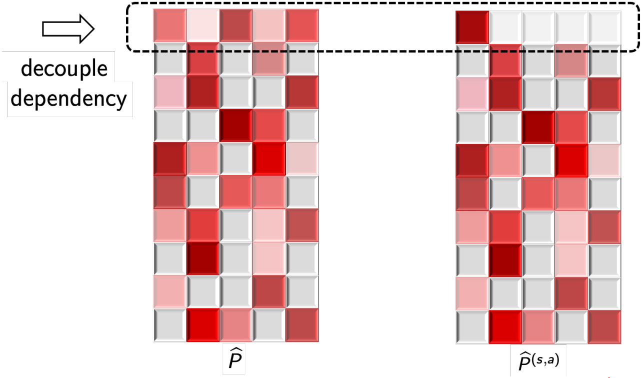
A natural idea: apply our policy evaluation theory + union bound

- highly suboptimal!

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

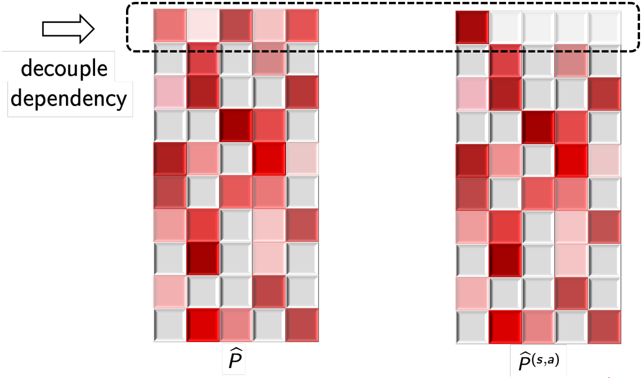
— *inspired by [Agarwal et al., 2019] but quite different . . .*

Key idea 2: leave-one-out argument



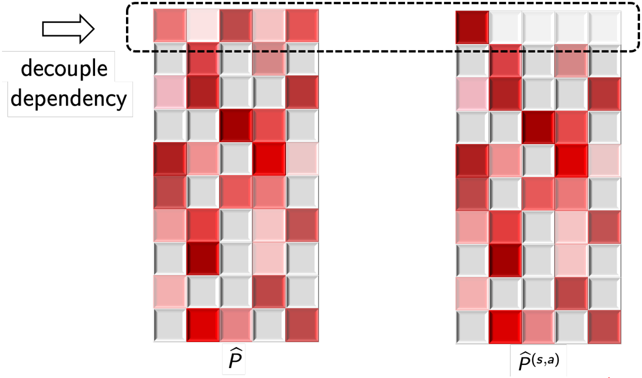
- state-action absorbing MDP for each (s, a): $(\mathcal{S}, \mathcal{A}, \widehat{P}^{(s,a)}, r, \gamma)$

Key idea 2: leave-one-out argument



- state-action absorbing MDP for each (s, a): $(\mathcal{S}, \mathcal{A}, \widehat{P}^{(s,a)}, r, \gamma)$
- $(\widehat{P} - P)_{s,a} V_{\widehat{\pi}^\star} = (\widehat{P} - P)_{s,a} V_{\widehat{\pi}^\star_{s,a}}$ ($\widehat{\pi}^\star_{s,a}$: optimal for new MDP)

Key idea 2: leave-one-out argument



- state-action absorbing MDP for each (s, a): $(\mathcal{S}, \mathcal{A}, \widehat{P}^{(s,a)}, r, \gamma)$
- $(\widehat{P} - P)_{s,a} V_{\widehat{\pi}^\star} = (\widehat{P} - P)_{s,a} V_{\widehat{\pi}^\star_{s,a}}$ ($\widehat{\pi}^\star_{s,a}$: optimal for new MDP)

Caveat: require $\widehat{\pi}^\star$ to stand out from other policies

Key idea 3: tie-breaking via perturbation

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

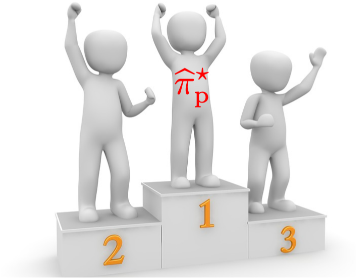
$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a: a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s, a) \geq \omega$$

Key idea 3: tie-breaking via perturbation

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

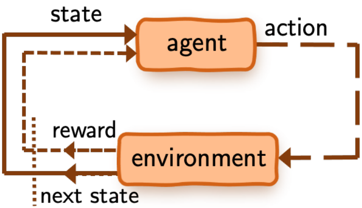
$$\forall s \in \mathcal{S}, \quad \widehat{Q}^\star(s, \widehat{\pi}^\star(s)) - \max_{a : a \neq \widehat{\pi}^\star(s)} \widehat{Q}^\star(s, a) \geq \omega$$

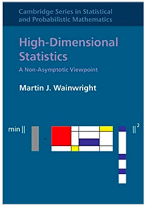
- **Solution:** slightly perturb rewards $r \implies \widehat{\pi}_{\mathrm{p}}^\star$
  - ensures the uniqueness of $\widehat{\pi}_{\mathrm{p}}^\star$
  - $V^{\widehat{\pi}_{\mathrm{p}}^\star} \approx V^{\widehat{\pi}^\star}$

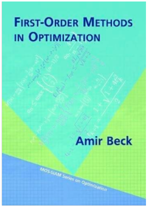


Concluding remarks

Understanding RL requires modern statistics and optimization

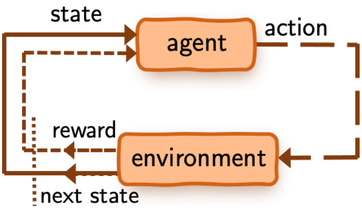


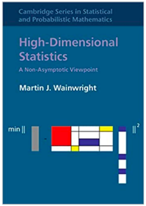


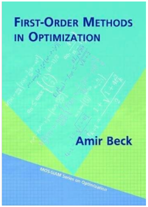


Concluding remarks

Understanding RL requires modern statistics and optimization







Future directions

- beyond the tabular setting
  [Feng et al., 2020, Jin et al., 2019, Duan and Wang, 2020]
- finite-horizon episodic MDPs
  [Dann and Brunskill, 2015, Jiang and Agarwal, 2018, Wang et al., 2020]

Paper:

"Breaking the sample size barrier in model-based reinforcement learning with a generative model," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, arxiv:2005.12900, 2020