Breaking the sample size barrier in reinforcement learning via $\underbrace{\textbf{model-based}}_{\text{“plug-in”}}$ methods

Yuxin Chen

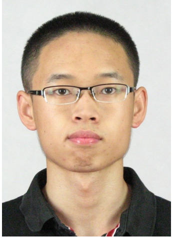
ECE, Princeton University

RETROSPECTIVE

David Blackwell, 1919–2010: An explorer in mathematics and statistics

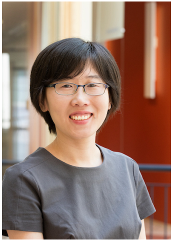
Peter J. Bickel[a,1]

Blackwell channel. He also began to work in dynamic programming, which is now called reinforcement learning. In a series of papers, Blackwell gave a rigorous foundation to the theory of dynamic programming, introducing what have become known as Blackwell optimal policies.

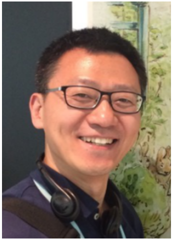


Gen Li
Tsinghua EE

Yuting Wei
CMU Statistics



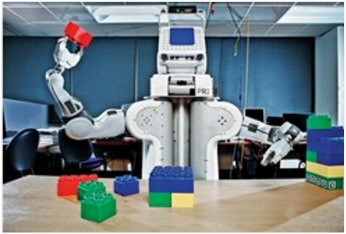
Yuejie Chi
CMU ECE



Yuantao Gu
Tsinghua EE

"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

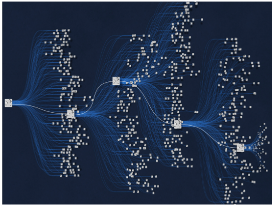
Reinforcement learning (RL)



RL challenges

In RL, an agent learns by interacting with an environment

- unknown or changing environments
- delayed rewards or feedback
- enormous state and action space
- nonconvexity



Sample efficiency

Collecting data samples might be expensive or time-consuming

clinical trials

online ads

Sample efficiency

Collecting data samples might be expensive or time-consuming

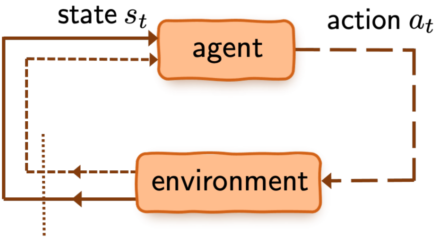
clinical trials

online ads

Calls for design of sample-efficient RL algorithms!

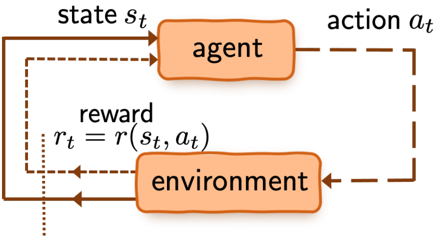
Background: Markov decision processes

Markov decision process (MDP)



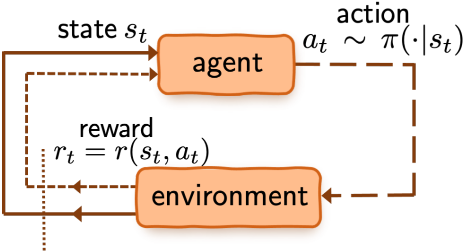
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

Markov decision process (MDP)



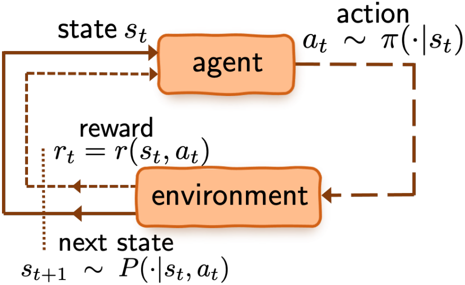
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

Markov decision process (MDP)



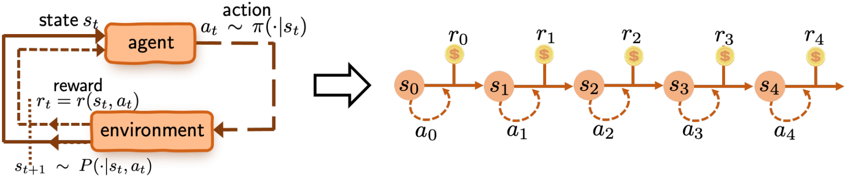
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

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

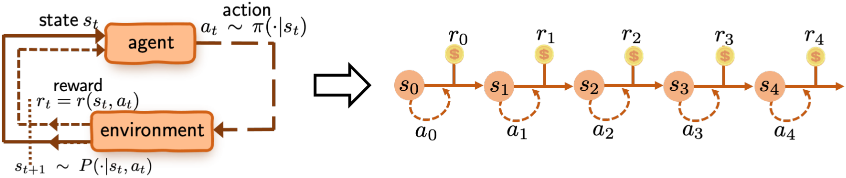
Value function



Value of policy π: long-term *discounted* reward

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

Value function

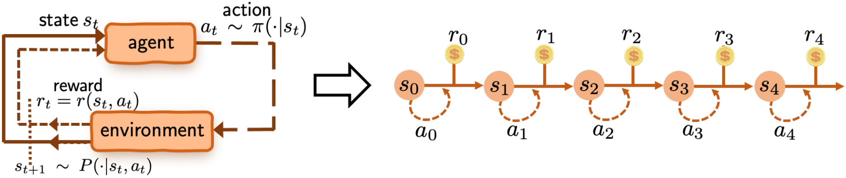


Value of policy π: long-term *discounted* reward

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

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

Value function

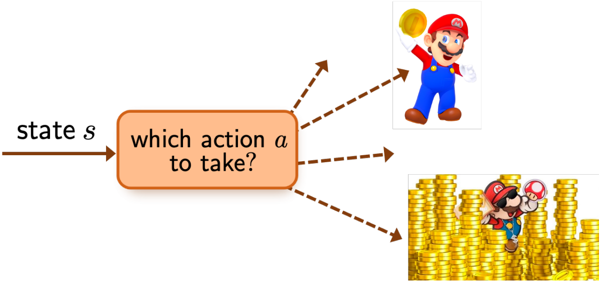


Value of policy π: long-term *discounted* reward

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

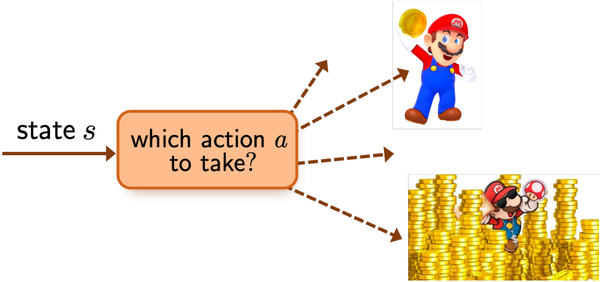
- $(a_0, s_1, a_1, s_2, a_2, \cdots)$: generated under policy π
- $\gamma \in [0, 1)$: discount factor
  - take $\gamma \to 1$ to approximate *long-horizon* MDPs

Optimal policy and optimal values



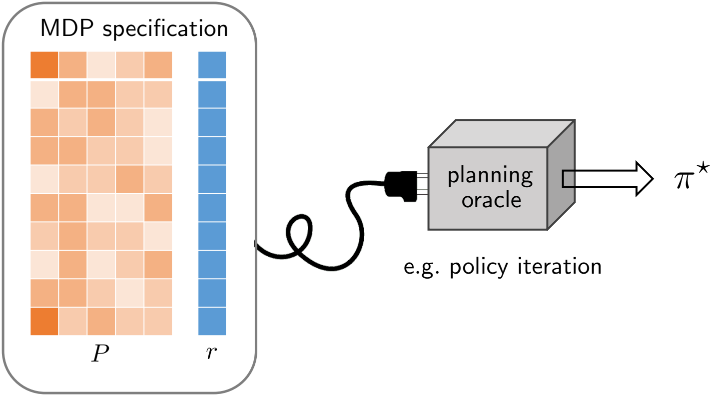
- **Optimal policy** $\pi^\star$: maximizing the value function

Optimal policy and optimal values



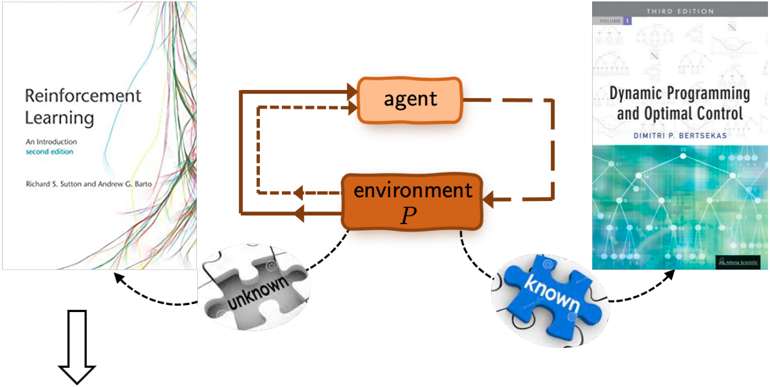
- **Optimal policy** $\pi^\star$: maximizing the value function
- Optimal values: $V^\star := V^{\pi^\star}$

When the model is known ...



Planning: computing the optimal policy $\pi^\star$ given MDP specification

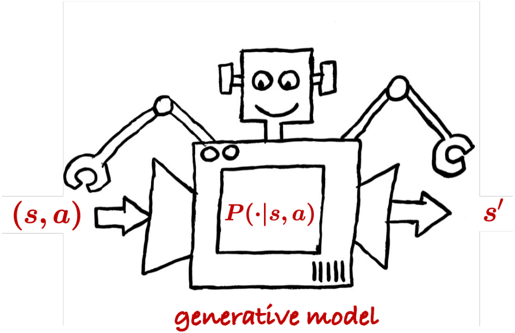
When the model is unknown . . .



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

This talk: RL with a generative model / simulator

— *Kearns, Singh '99*



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

Question: *how many samples are sufficient to learn an* $\underbrace{\varepsilon\text{-optimal policy}}$ *?*

Question: *how many samples are sufficient to learn an* $\underbrace{\varepsilon\text{-optimal policy}}_{\forall s:\ V^{\widehat{\pi}}(s) \geq V^\star(s) - \varepsilon}$ *?*

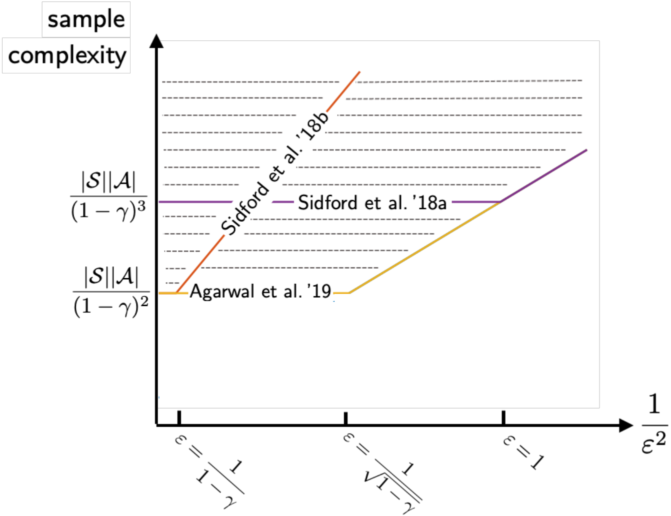
An incomplete list of prior art

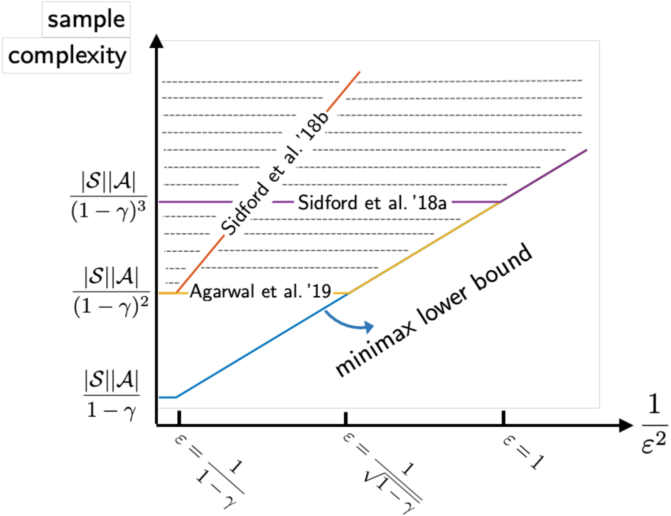
- Kearns & Singh '99
- Kakade '03
- Kearns, Mansour & Ng '02
- Azar, Munos & Kappen '12
- Azar, Munos, Ghavamzadeh & Kappen '13
- Sidford, Wang, Wu, Yang & Ye '18
- Sidford, Wang, Wu & Ye '18
- Wang '17
- Agarwal, Kakade & Yang '19
- Wainwright '19a
- Wainwright '19b
- Pananjady & Wainwright '20
- Yang & Wang '19
- Khamaru, Pananjady, Ruan, Wainwright & Jordan '20
- Mou, Li, Wainwright, Bartlett & Jordan '20
- ...

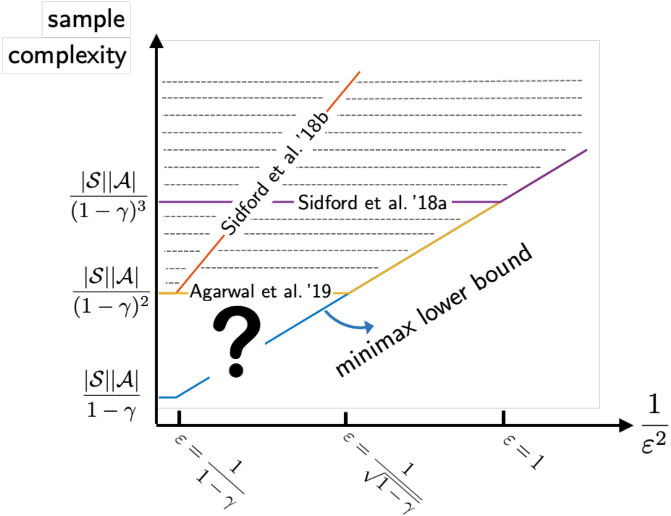
An even shorter list of prior art

algorithm	sample size range	sample complexity	ε-range
phased Q-learning Kearns and Singh '99	$\left[\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^5}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\varepsilon^2}$	$(0, \frac{1}{1-\gamma}]$
empirical QVI Azar et al. '13	$\left[\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}}]$
sublinear randomized VI Sidford et al. '18a	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$	$\left(0, \frac{1}{1-\gamma}\right]$
variance-reduced QVI Sidford et al. '18b	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, 1]$
empirical MDP + planning Agarwal et al. '19	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{1-\gamma}}]$

— *see also Wainwright '19a '19b (for estimating optimal values)*


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

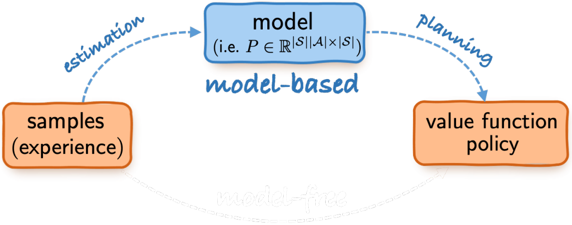

sample
complexity
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$
$\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$
Sidford et al. '18b
Sidford et al. '18a
Agarwal et al. '19
minimax lower bound
$\frac{1}{\varepsilon^2}$
$\varepsilon = \frac{1}{1-\gamma}$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



All prior theory requires $\underbrace{\color{red}{\text{sample size} > \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}}}_{\text{sample size barrier}}$

Is it possible to close the gap?

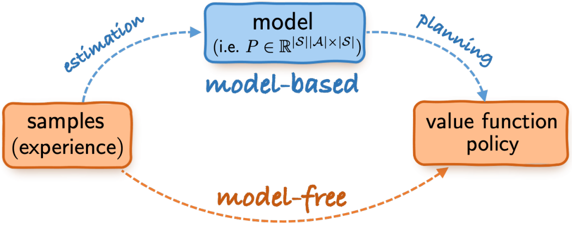
Two approaches



Model-based approach ("plug-in")

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

Two approaches



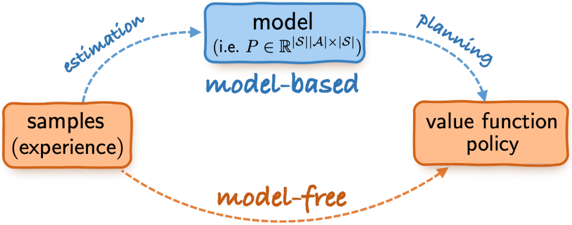
Model-based approach ("plug-in")

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

Model-free approach (e.g. Q-learning; see our other works)

— learning w/o estimating the model explicitly

Two approaches



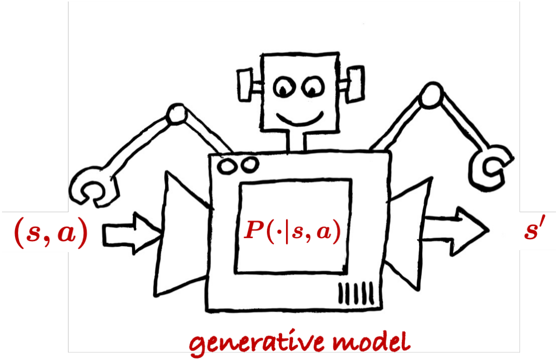
Model-based approach ("plug-in")

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

Model-free approach (e.g. Q-learning; see our other works)

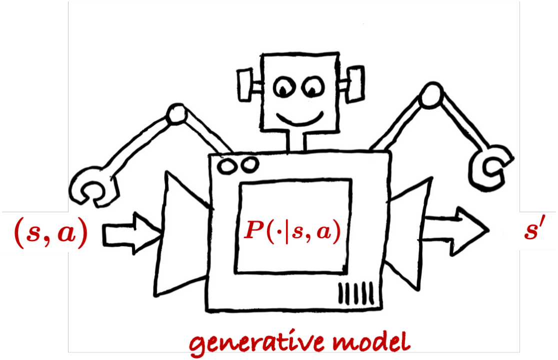
— learning w/o estimating the model explicitly

Model estimation



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

Model estimation

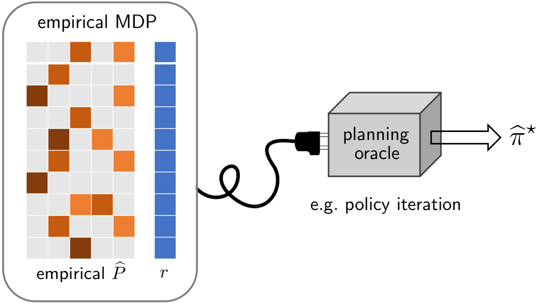


Sampling: for each (s,a), collect N ind. samples $\{(s,a,s'_{(i)})\}_{1\leq i\leq N}$

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

Model-based (plug-in) estimator

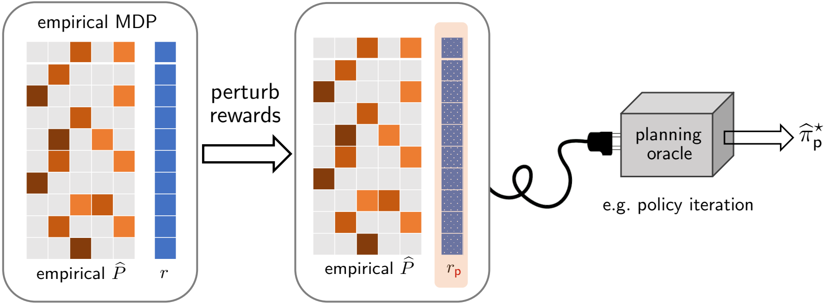
— *Azar et al. '13, Agarwal et al. '19, Pananjady et al. '20*



Planning based on the *empirical* MDP with *slightly perturbed rewards*

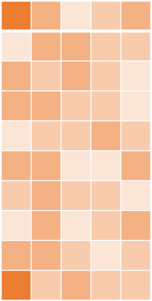
Our method: plug-in estimator + perturbation

— *Li, Wei, Chi, Gu, Chen '20*

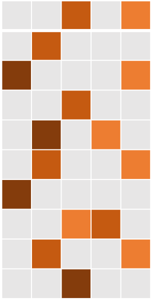


Run planning algorithms based on the *empirical* MDP

Challenges in the sample-starved regime



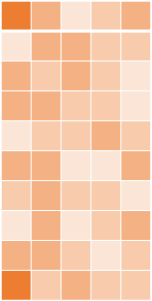
truth:
$P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



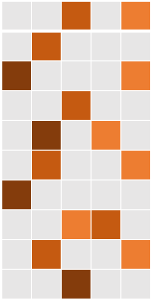
empirical estimate:
$\widehat{P}$

- 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 1 (Li, Wei, Chi, Gu, Chen '20)

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

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

with sample complexity at most

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

Main result

Theorem 1 (Li, Wei, Chi, Gu, Chen '20)

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

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

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 $\widetilde{O}\left(\frac{1}{1-\gamma}\right)$ iterations

Main result

Theorem 1 (Li, Wei, Chi, Gu, Chen '20)

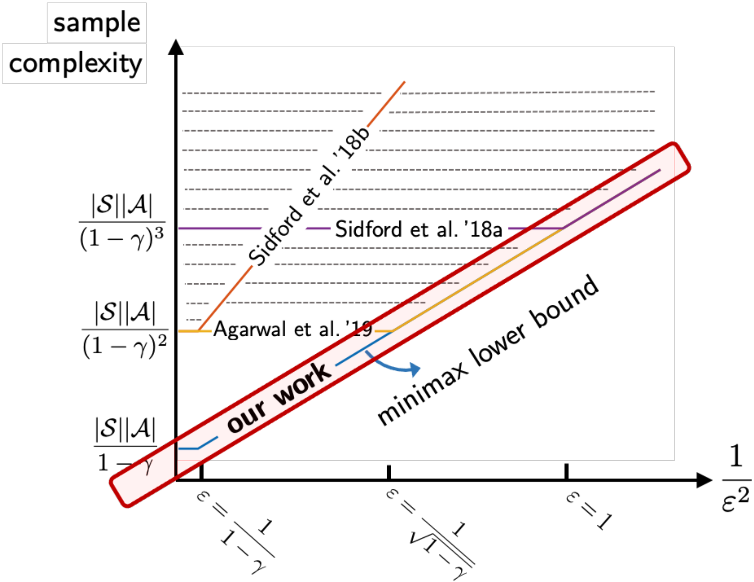
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$*, the optimal policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of the perturbed empirical MDP achieves*

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

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 $\widetilde{O}(\frac{1}{1-\gamma})$ iterations
- **Minimax lower bound:** $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ (Azar et al. '13)


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

Analysis

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V^\pi = (I - \gamma P_\pi)^{-1} r$

Notation and Bellman equation

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

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V^\pi = (I - \gamma P_\pi)^{-1} r$
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \gamma \widehat{P}_\pi)^{-1} r$
- $\pi^\star$: 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: $V^\pi = (I - \gamma P_\pi)^{-1} r$
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \gamma \widehat{P}_\pi)^{-1} r$
- $\pi^\star$: 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

Elementary decomposition:

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

Proof ideas

Elementary decomposition:

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

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

Proof ideas

Elementary decomposition:

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

- **Step 1:** control $V^{\pi} - \widehat{V}^{\pi}$ for a fixed π
  (**Bernstein inequality** + **high-order decomposition**)
- **Step 2:** extend it to control $\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}$ ($\widehat{\pi}^{\star}$ depends on samples)
  (**decouple statistical dependency**)

Step 1: improved theory for policy evaluation

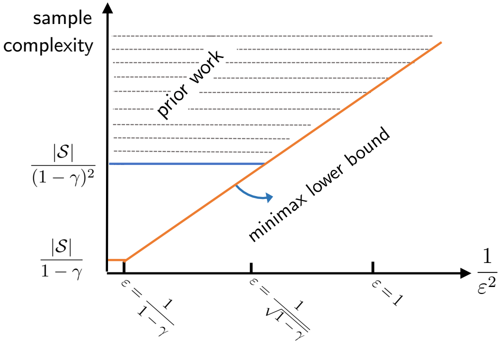
Model-based policy evaluation:

— given a fixed policy π, estimate V^{π} via the plug-in estimate $\widehat{V}^{\pi}$

Step 1: improved theory for policy evaluation

Model-based policy evaluation:

— given a fixed policy π, estimate V^{π} via the plug-in estimate $\widehat{V}^{\pi}$



- A sample size barrier $\frac{|\mathcal{S}|}{(1-\gamma)^2}$ already appeared in prior works (Agarwal et al. '19, Pananjady & Wainwright '19, Khamaru et al. '20)

Step 1: improved theory for policy evaluation

Model-based policy evaluation:

— given a fixed policy π, estimate V^{π} via the plug-in estimate $\widehat{V}^{\pi}$

Theorem 2 (Li, Wei, Chi, Gu, Chen'20)

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

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

with sample complexity at most

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

Step 1: improved theory for policy evaluation

Model-based policy evaluation:

— given a fixed policy π, estimate V^π via the plug-in estimate $\widehat{V}^\pi$

Theorem 2 (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. For $0 < \varepsilon \leq \frac{1}{1-\gamma}$, the 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 optimal for all ε (Azar et al. '13, Pananjady & Wainwright '19)

Key idea 1: a peeling argument

Agarwal et al. '19 & other prior works: 1st-order expansion

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

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

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

Key idea 1: a peeling argument

Agarwal et al. '19 & other prior works: 1st-order expansion

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

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

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

Key idea 1: a peeling argument

Agarwal et al. '19 & other prior works: 1st-order expansion

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

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

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

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

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

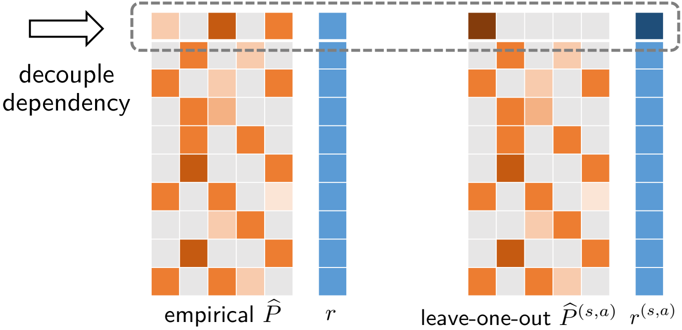
Step 2: controlling $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

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

- highly suboptimal! (there are exponentially many policies)

Key idea 2: leave-one-out analysis

Decouple dependency by introducing auxiliary state-action absorbing MDPs by dropping randomness for each (s, a)

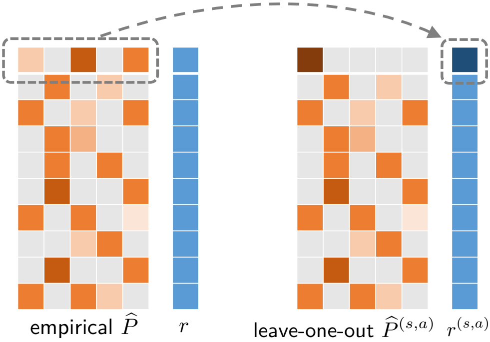


— *inspired by Agarwal et al. '19 but quite different ...*

Key idea 2: leave-one-out analysis

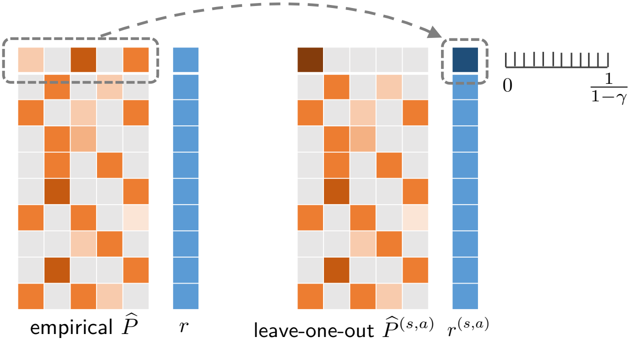
- Stein '72
- El Karoui, Bean, Bickel, Lim, Yu '13
- El Karoui '15
- Javanmard, Montanari '15
- Zhong, Boumal '17
- Lei, Bickel, El Karoui '17
- Sur, Chen, Candès '17
- Abbe, Fan, Wang, Zhong '17
- Chen, Fan, Ma, Wang '17
- Ma, Wang, Chi, Chen '17
- Chen, Chi, Fan, Ma '18
- Ding, Chen '18
- Dong, Shi '18
- Chen, Chi, Fan, Ma, Yan '19
- Chen, Fan, Ma, Yan '19
- Cai, Li, Poor, Chen '19
- Agarwal, Kakade, Yang '19
- Pananjady, Wainwright '19
- Ling '20

Key idea 2: leave-one-out analysis



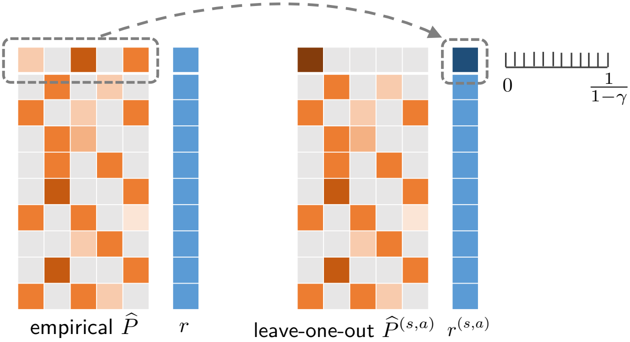
1. embed all randomness from $\widehat{P}_{s,a}$ into a single scalar (i.e. $r_{s,a}^{(s,a)}$)

Key idea 2: leave-one-out analysis



1. embed all randomness from $\widehat{P}_{s,a}$ into a single scalar (i.e. $r_{s,a}^{(s,a)}$)
2. build an ϵ-net for this scalar

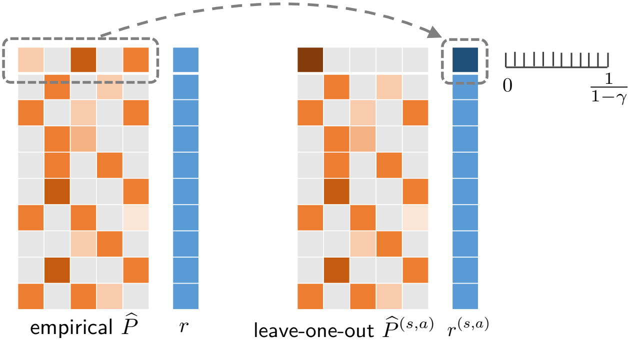
Key idea 2: leave-one-out analysis



1. embed all randomness from $\widehat{P}_{s,a}$ into a single scalar (i.e. $r_{s,a}^{(s,a)}$)
2. build an ϵ-net for this scalar
3. $\widehat{\pi}^\star$ can be determined by this ϵ-net 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 2: leave-one-out analysis



Our decoupling argument vs. Agarwal, Kakade, Yang '19

- *Agarwal et al. '19:* dependency btw value $\widehat{V}$ & samples
- *Ours:* dependency btw policy $\widehat{\pi}$ & samples

Key idea 3: tie-breaking via perturbation

- How to ensure separation between the optimal policy and others?

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

Key idea 3: tie-breaking via perturbation

- How to ensure separation between the optimal policy and others?

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

- **Solution:** *slightly perturb rewards* $r \implies \widehat{\pi}_{\mathrm{p}}^{\star}$
  - *ensures* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *can be differentiated from others*
  - $V^{\widehat{\pi}_{\mathrm{p}}^{\star}} \approx V^{\widehat{\pi}^{\star}}$

Key idea 3: tie-breaking via perturbation

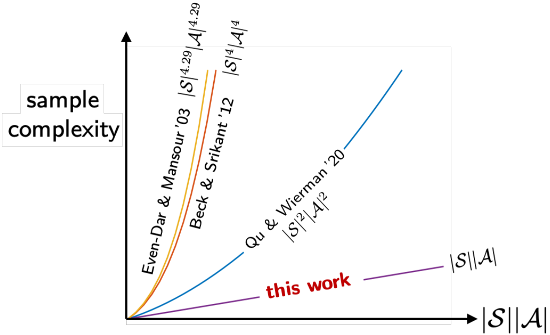
- How to ensure separation between the optimal policy and others?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a:\, a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s, a) > \frac{(1-\gamma)\varepsilon}{|\mathcal{S}|^5 |\mathcal{A}|^5}$$

- **Solution:** *slightly perturb rewards* $r \implies \widehat{\pi}_{\mathrm{p}}^{\star}$
  - *ensures* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *can be differentiated from others*
  - $V^{\widehat{\pi}_{\mathrm{p}}^{\star}} \approx V^{\widehat{\pi}^{\star}}$

Other stories: sharpened analysis of Q-learning

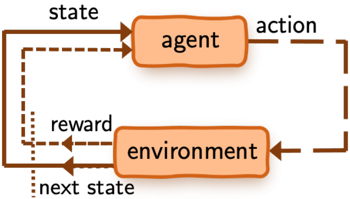
Improves existing sample complexity guarantees for asynchronous Q-learning by at least a factor of $|\mathcal{S}||\mathcal{A}|$!

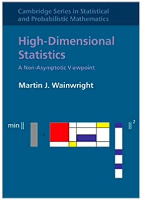


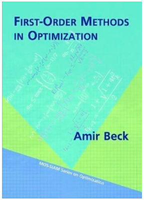
"Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, NeurIPS 2020

Concluding remarks

Understanding RL requires modern statistics and optimization

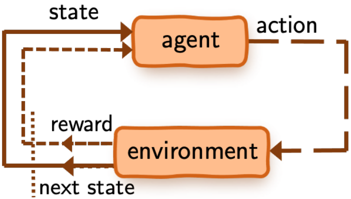


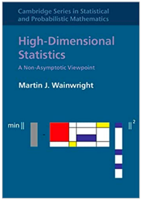


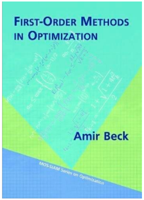


Concluding remarks

Understanding RL requires modern statistics and optimization







future directions

- finite-horizon episodic MDPs
- beyond the tabular settings
- Markov games
- exploration settings

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