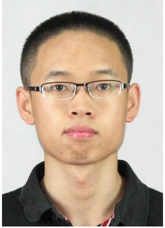
Breaking the Sample Size Barrier in Reinforcement Learning

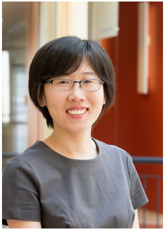
Yuting Wei

Statistics & Data Science, Wharton
University of Pennsylvania

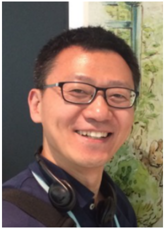
MIT Statistics Seminar, 2021



Gen Li
Tsinghua EE



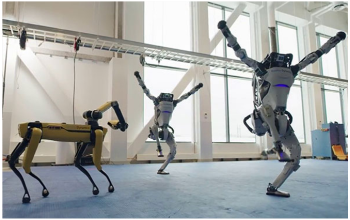
Yuejie Chi
CMU ECE



Yuantao Gu
Tsinghua EE

Yuxin Chen
Princeton EE

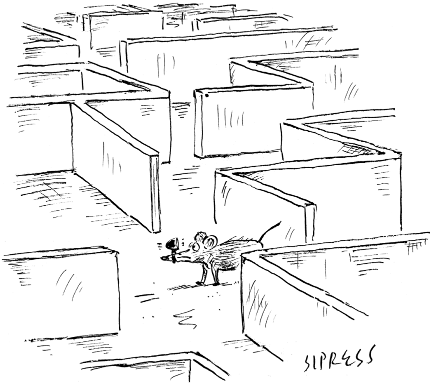
Success stories of reinforcement learning



Reinforcement learning (RL)

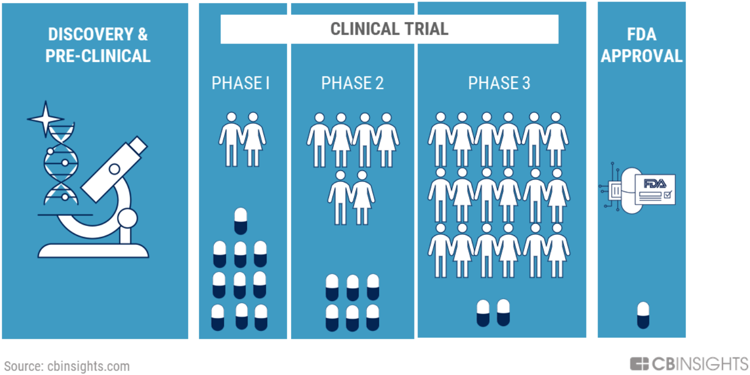
In RL, an agent learns by interacting with an environment.

- no training data
- trial-and-error
- maximize total rewards
- sequential and online



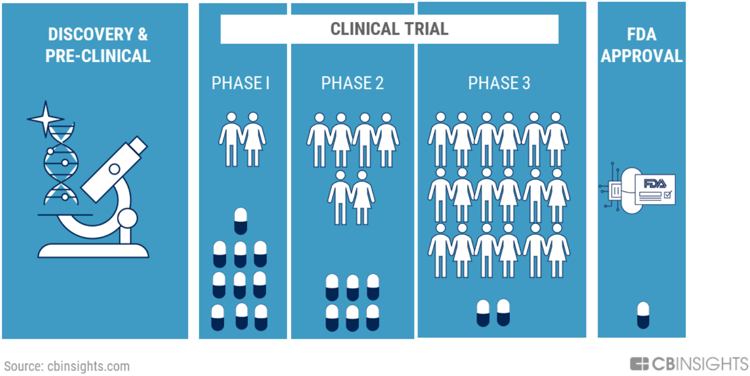
"Recalculating ... recalculating ..."

Sample efficiency



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

Sample efficiency

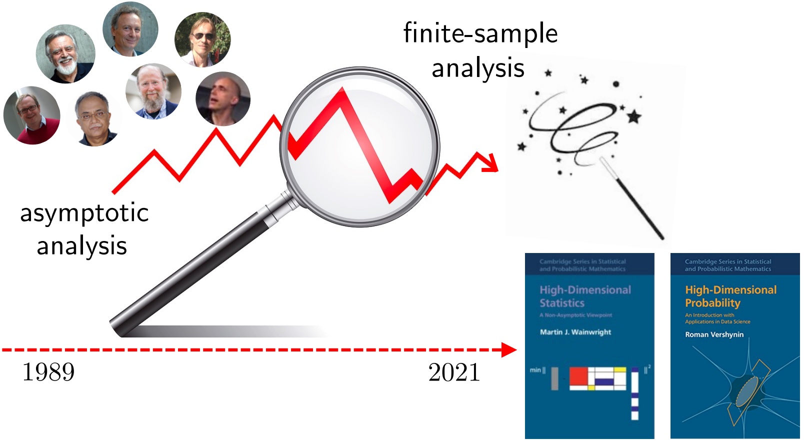


- prohibitively large state & action space
- collecting data samples can be expensive or time-consuming

Challenge: design & understand sample efficient RL algorithms

Statistical foundation of RL

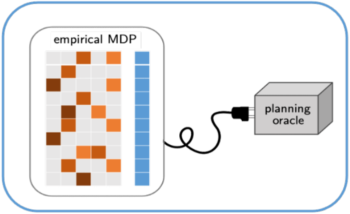
Statistical foundation of RL



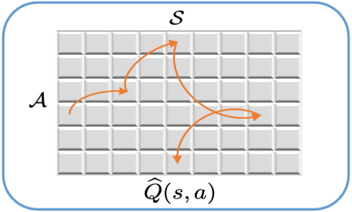
Understanding sample efficiency of RL requires a modern suite of non-asymptotic statistical tools.

Outline

- Background
- Vignette #1: model-based RL ("plug-in" approach)
- Vignette #2: model-free RL (Q-learning on Markovian samples)



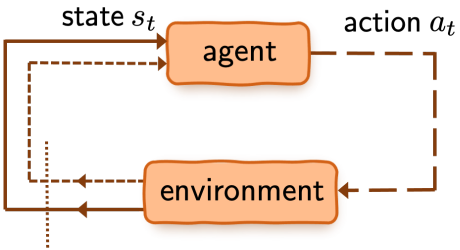
model based RL



model free RL

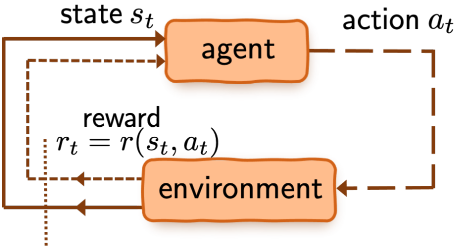
Background: Markov decision processes

Markov decision process (MDP)



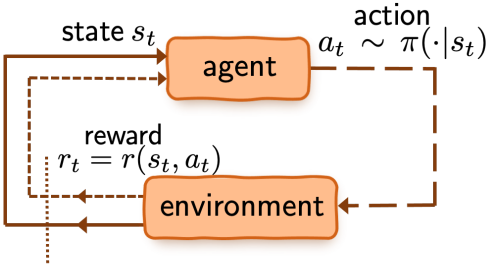
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

Markov decision process (MDP)



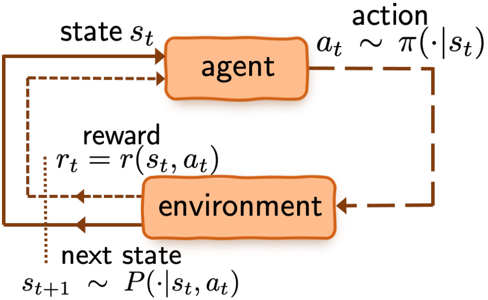
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

Markov decision process (MDP)



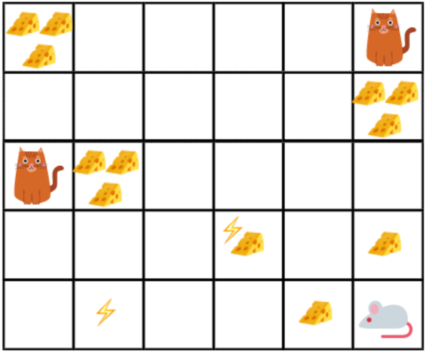
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

Markov decision process (MDP)

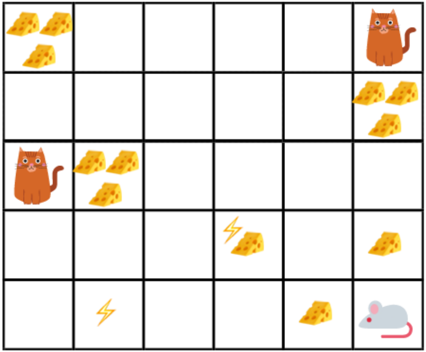


- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)
- $P(\cdot|s, a)$: unknown transition probabilities

Help the mouse!

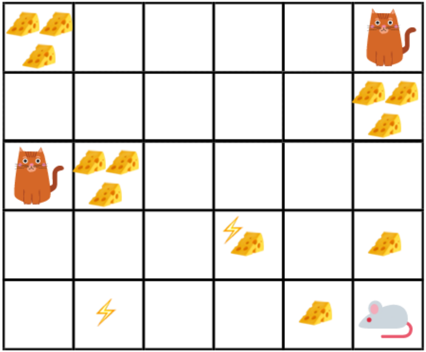


Help the mouse!



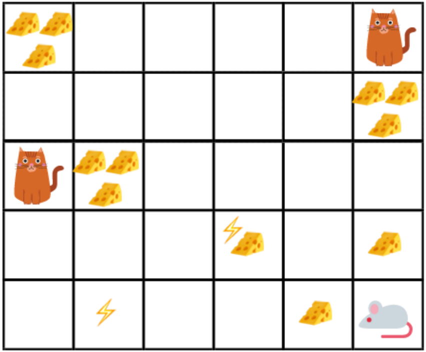
- state space $\mathcal{S}$: positions in the maze

Help the mouse!



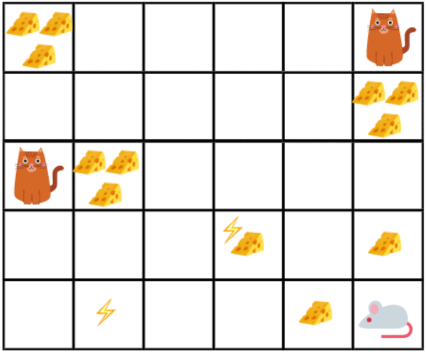
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right

Help the mouse!



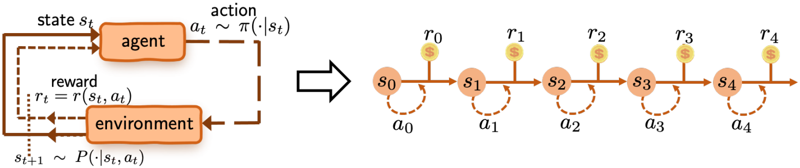
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right
- immediate reward r: cheese, electricity shocks, cats

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

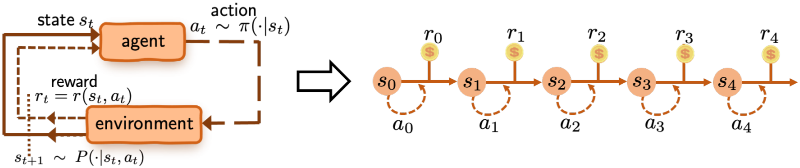
Value function



Value 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

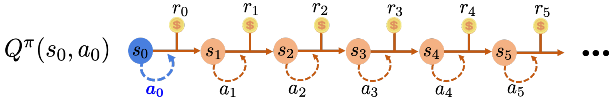


Value of policy π: cumulative discounted reward

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

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

Q-function (action-value function)

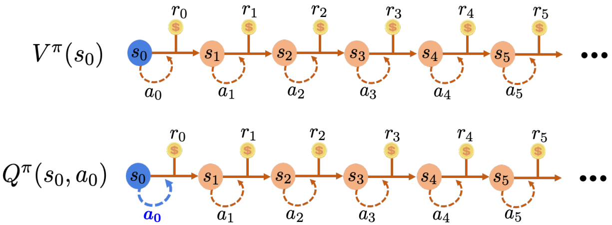


Q-function of policy π:

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

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

Q-function (action-value function)

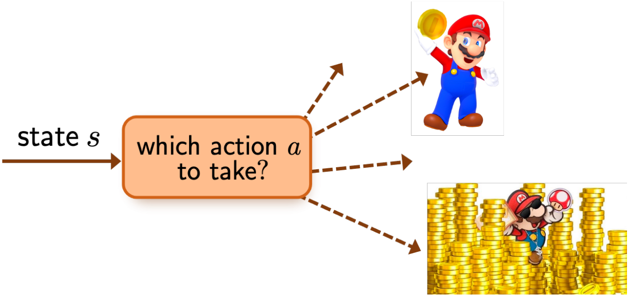


Q-function of policy π:

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

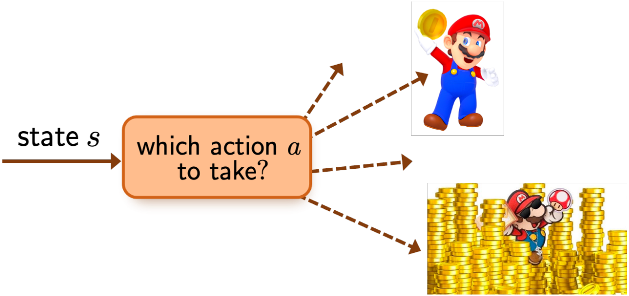
- $(\cancel{a_0,}\ s_1, a_1, s_2, a_2, \cdots)$: induced by policy π

Optimal policy and optimal value



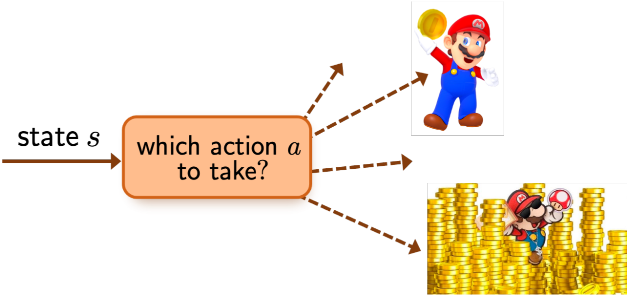
- **optimal policy** $\pi^\star$: maximizing value function $\max_{\pi} V^{\pi}(s)$

Optimal policy and optimal value



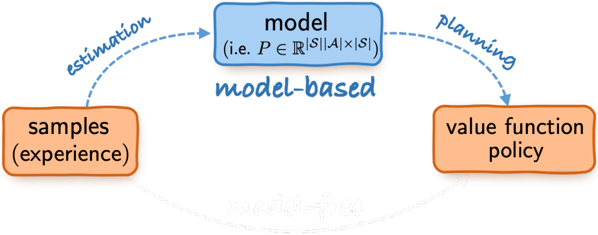
- **optimal policy** $\pi^\star$: maximizing value function $\max_{\pi} V^{\pi}(s)$
- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$

Optimal policy and optimal value



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

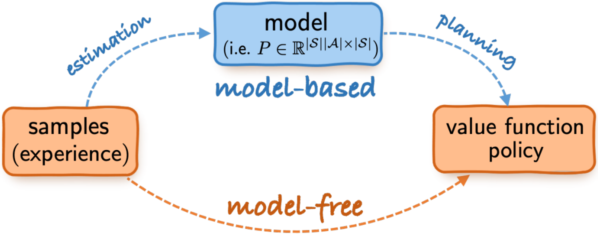
Model-based vs. model-free RL



Model-based approach ("plug-in")

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

Model-based vs. model-free RL



Model-based approach ("plug-in")

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

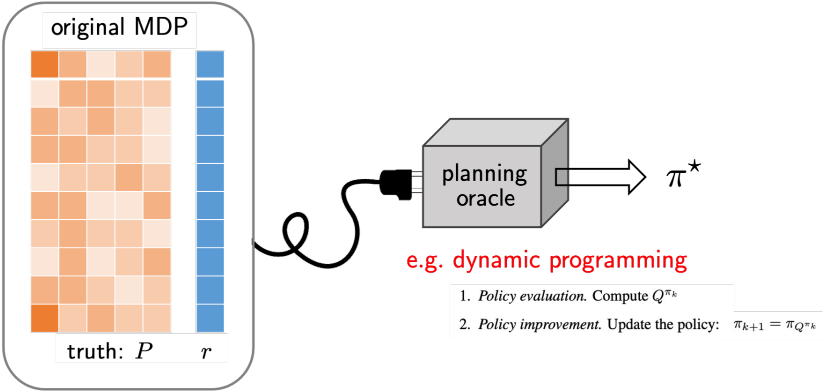
Model-free approach (e.g. Q-learning)

— learning w/o modeling & estimating environment explicitly

Vignette #1: Model-based RL (a "plug-in" approach)

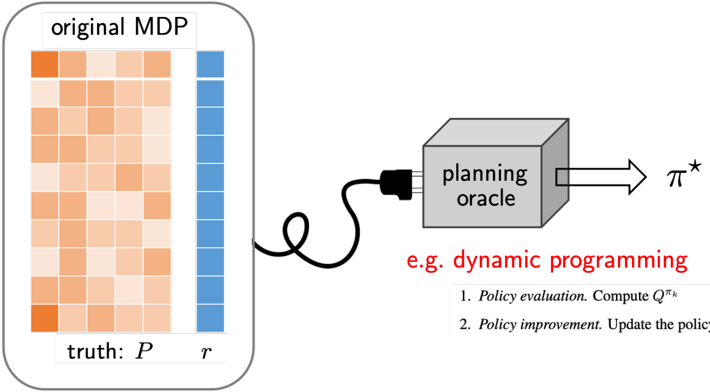
"Breaking the sample size barrier in model-based reinforcement learning with a generative model," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, NeurIPS, 2020

When the model is known ...



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

When the model is known . . .



e.g. dynamic programming

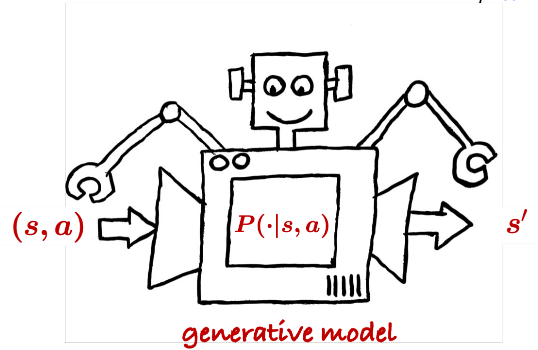
1. *Policy evaluation.* Compute Q^{π_k}
2. *Policy improvement.* Update the policy: $\pi_{k+1} = \pi_{Q^{\pi_k}}$

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

In practice, do not know transition matrix P!

This work: sampling from a generative model

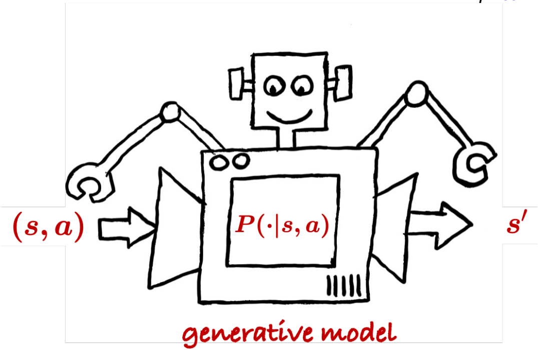
— [Kearns and Singh, 1999]



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

This work: sampling from a generative model

— [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 prior art

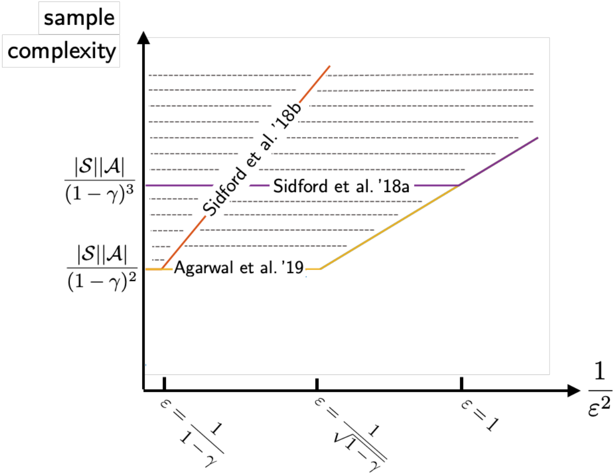
- [Kearns and Singh, 1999]
- [Kakade, 2003]
- [Kearns et al., 2002]
- [Azar et al., 2012]
- [Azar et al., 2013]
- [Sidford et al., 2018a]
- [Sidford et al., 2018b]
- [Wang, 2019]
- [Agarwal et al., 2019]
- [Wainwright, 2019a, Wainwright, 2019b]
- [Pananjady and Wainwright, 2019]
- [Yang and Wang, 2019]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

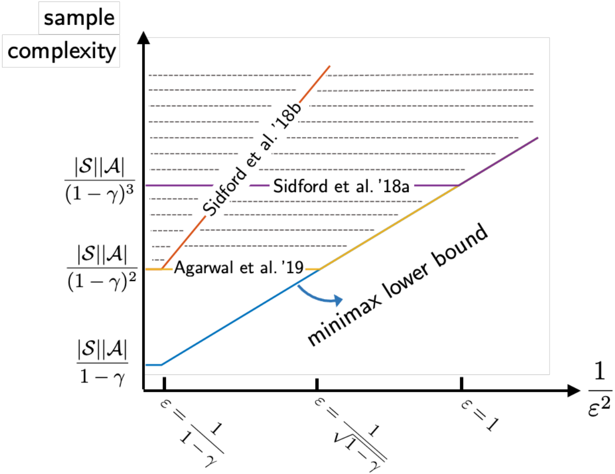
An even shorter list of prior art

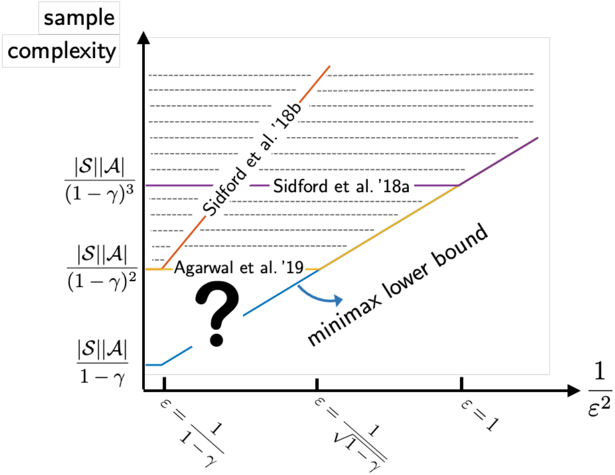
algorithm	sample size range	sample complexity	ε-range
Empirical QVI [Azar et al., 2013]	$\left[\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$\left(0, \frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}}\right]$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

important parameters:

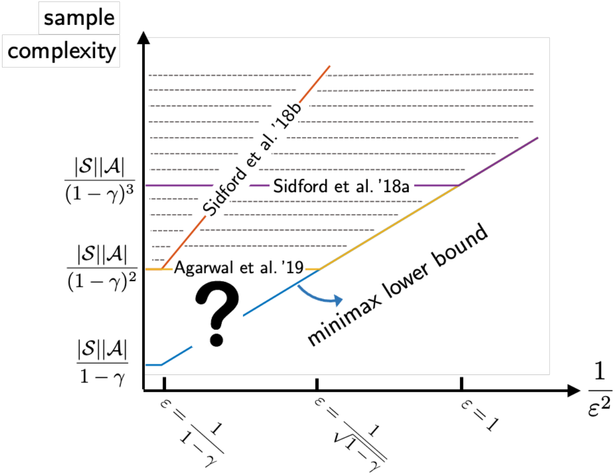
- $\lvert\mathcal{S}\rvert$: # states , $\lvert\mathcal{A}\rvert$: # actions
- $\frac{1}{1-\gamma}$: effective horizon
- $\varepsilon \in [0, \frac{1}{1-\gamma}]$: approximation error


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



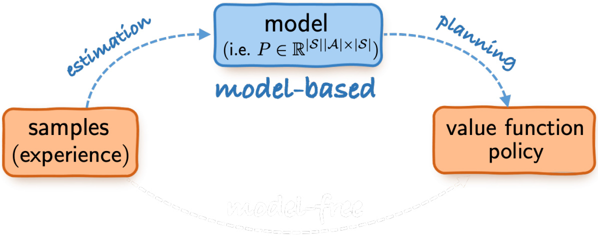
All prior theory requires sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$



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

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

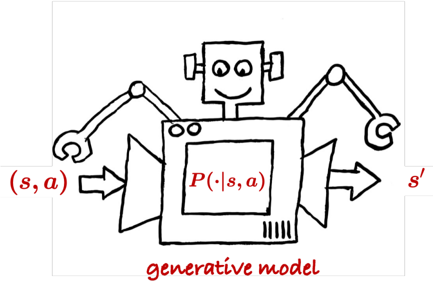
Our algorithm: model-based RL



Model-based approach ("plug-in")

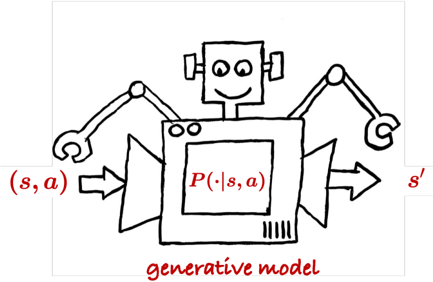
1. build an empirical estimate $\widehat{P}$ for P
2. planning based on empirical $\widehat{P}$

Model estimation



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

Model estimation

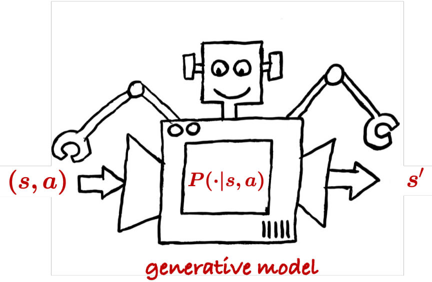


Sampling: for each (s, a), collect N ind. samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$

Empirical estimates:

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

Model estimation



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

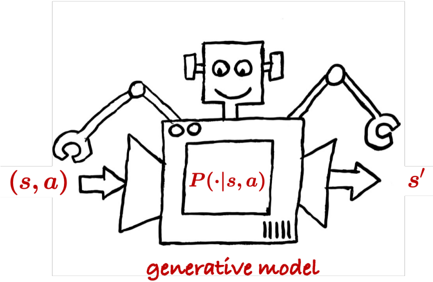
Empirical estimates:

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

Hoeffding's inequality

With probability $1-\delta$, we have $|\widehat{P}(s'|s,a) - P(s'|s,a)| \leq \sqrt{\frac{\log(1/\delta)}{N}}$

Model estimation



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

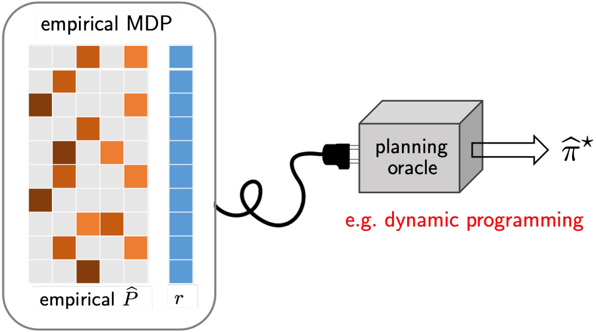
Empirical estimates:

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

If sample size $\ll |\mathcal{S}|^2|\mathcal{A}|$, then we cannot recover P faithfully.

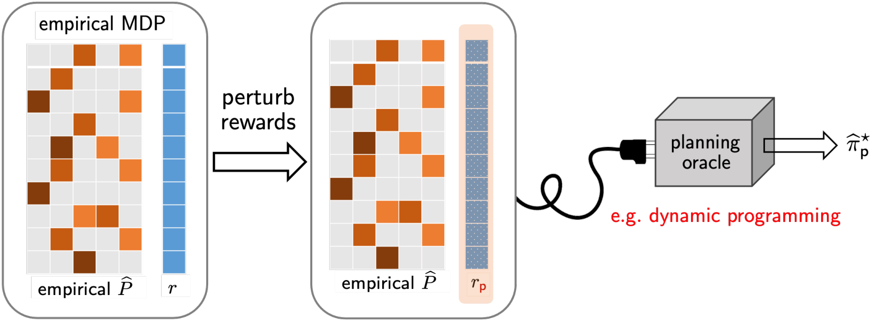
Model-based (plug-in) estimator

—[*Azar et al., 2013*, *Agarwal et al., 2019*, *Pananjady and Wainwright, 2019*]



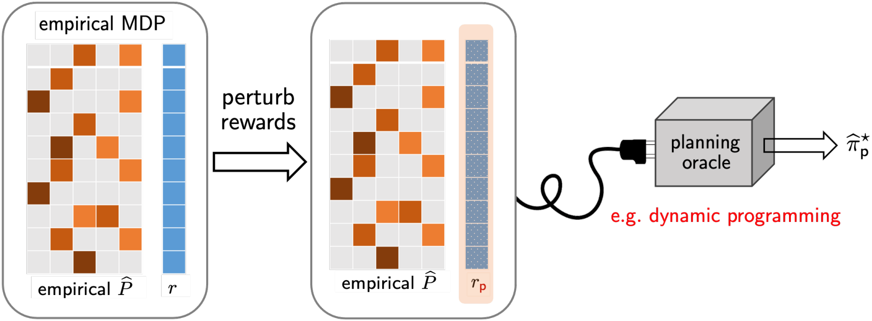
Find policy based on the empirical MDP (*empirical maximizer*)

Our method: plug-in estimator + perturbation



Find policy based on the empirical MDP with slightly perturbed rewards

Our method: plug-in estimator + perturbation



Find policy based on the empirical MDP with slightly perturbed rewards

Question: Can we trust our $\widehat{\pi}$ when $\widehat{P}$ is not accurate?

Main result: ℓ_∞-based sample complexity

Theorem (Li, Wei, Chi, Gu, 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 sample complexity at most

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

Main result: ℓ_∞-based sample complexity

Theorem (Li, Wei, Chi, Gu, 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 sample complexity at most

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

- minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ [Azar et al., 2013]

Main result: ℓ_∞-based sample complexity

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

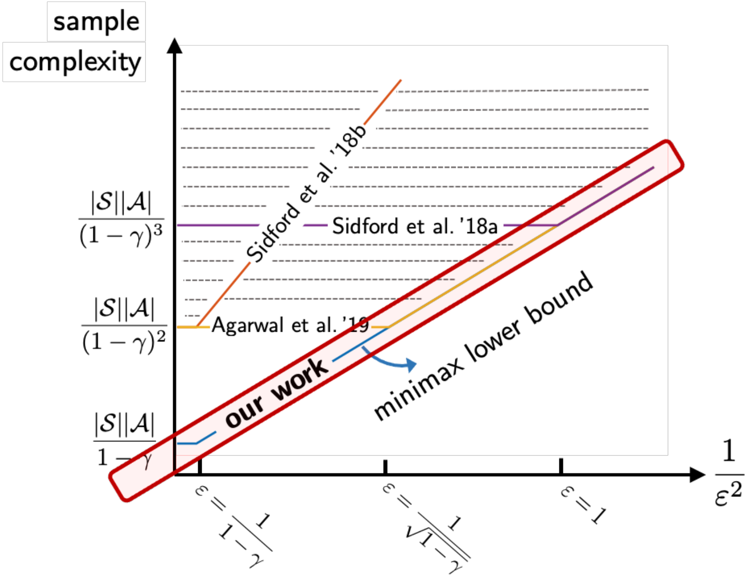
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *the optimal policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of perturbed empirical MDP achieves*

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

with sample complexity at most

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

- minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ [Azar et al., 2013]
- $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big] \quad \rightarrow \quad$ sample size range $\big[\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}, \infty\big)$


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 glimpse of the key analysis ideas

Notation and Bellman equation

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

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

Notation and Bellman equation

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

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

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

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

Main steps

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 (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)

[Agarwal et al., 2019] and prior work: first-order expansion

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

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

[Agarwal et al., 2019] and prior work: first-order expansion

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

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

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

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

[Agarwal et al., 2019] and prior work: first-order expansion

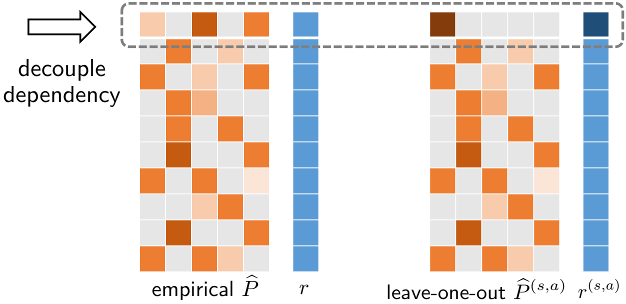
$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\widehat{V}^{\pi}$$

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

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

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

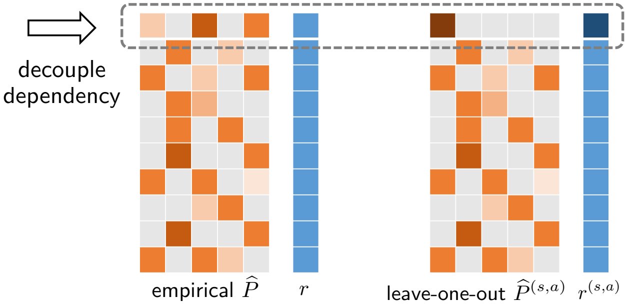
— *inspired by [Agarwal et al., 2019] but quite different . . .*



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

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

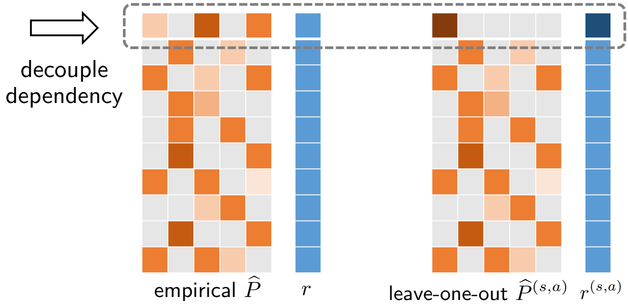
— *inspired by [Agarwal et al., 2019] but quite different ...*



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

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

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

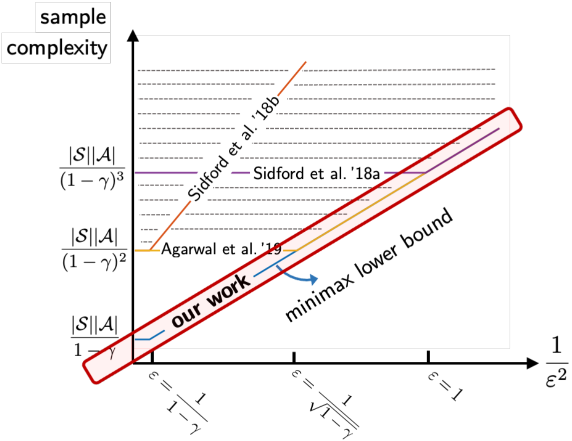
Key idea 3: tie-breaking via reward perturbation

- How to ensure separation btw 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
    - $V^{\widehat{\pi}_{\mathrm{p}}^\star} \approx V^{\widehat{\pi}^\star}$

Summary of this part



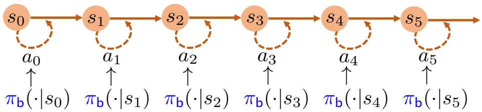
Model-based RL is minimax optimal & does not suffer from a sample size barrier!

Vignette #2: Model-free approach

"Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, IEEE Transactions on Information Theory, 2021

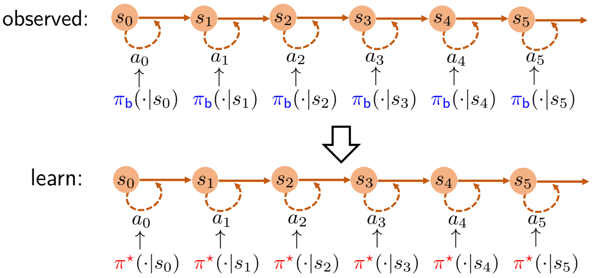
Markovian samples and behavior policy

observed:



Observed: $\underbrace{\{s_t, a_t, r_t\}_{t\geq 0}}_{\text{Markovian trajectory}}$ induced by behavior policy π_b

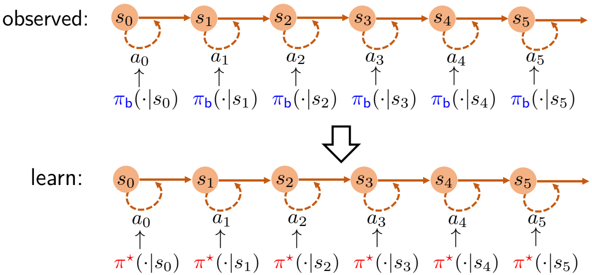
Markovian samples and behavior policy



Observed: $\underbrace{\{s_t, a_t, r_t\}_{t\geq 0}}_{\text{Markovian trajectory}}$ induced by behavior policy π_b

Goal: learn optimal value $V^\star$ and $Q^\star$ based on sample trajectory

Markovian samples and behavior policy



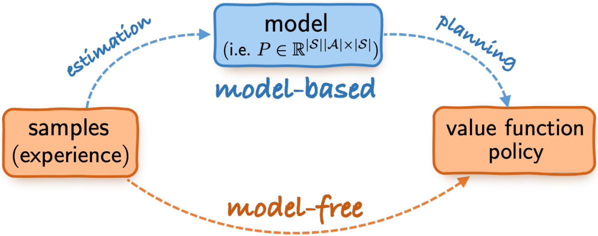
Key quantities of sample trajectory

- minimum state-action occupancy probability

$$\mu_{\mathsf{min}} := \min \quad \underbrace{\mu_{\pi_b}(s,a)}_{\text{stationary distribution}}$$

- mixing time: t_{mix}

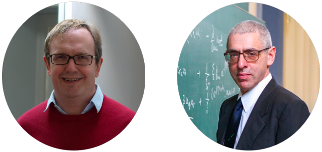
Model-based vs. model-free RL



Model-free approach (e.g. Q-learning)

— learning w/o modeling & estimating environment explicitly

Q-learning: a classical model-free algorithm



Chris Watkins *Peter Dayan*

$\underbrace{\text{Stochastic approximation}}_{\text{Robbins \& Monro '51}}$ for solving **Bellman equation** $Q = \mathcal{T}(Q)$

Aside: 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

Aside: 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

Richard Bellman

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

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

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

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

$$\underbrace{Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \eta_t(\mathcal{T}_t(Q_t)(s_t, a_t) - Q_t(s_t, a_t)),}_{\textit{only}\text{ update }(s_t,a_t)\text{-th entry}} \quad t \geq 0$$

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

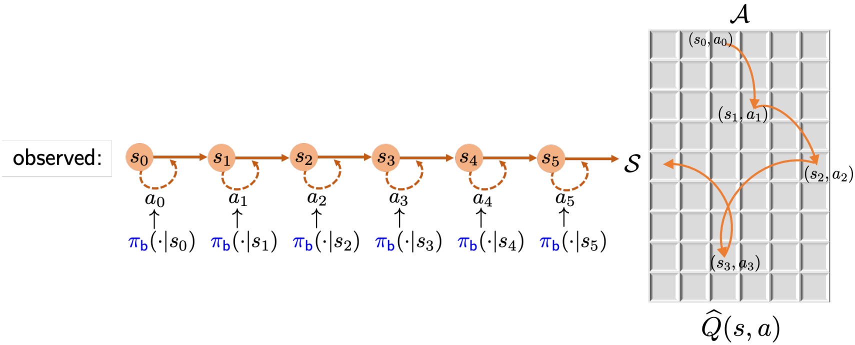
Stochastic approximation for solving Bellman equation $Q = \mathcal{T}(Q)$

$$\underbrace{Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \eta_t(\mathcal{T}_t(Q_t)(s_t, a_t) - Q_t(s_t, a_t)),}_{\textit{only}\text{ update }(s_t, a_t)\text{-th entry}} \quad t \geq 0$$

$$\mathcal{T}_t(Q)(s_t, a_t) := r(s_t, a_t) + \gamma \max_{a'} Q(s_{t+1}, a')$$

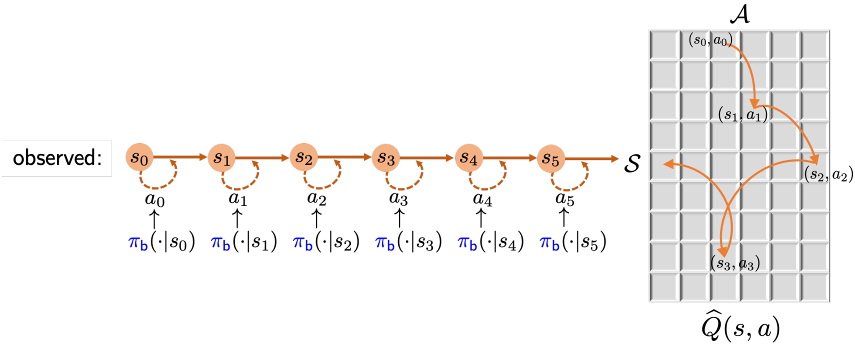
$$\mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s', a')\right]$$

Q-learning on Markovian samples



- **asynchronous:** only a single entry is updated each iteration

Q-learning on Markovian samples



- **asynchronous:** only a single entry is updated each iteration
  - resembles Markov-chain *coordinate descent*

What is sample complexity of (async) Q-learning?

A highly incomplete list of prior work

- [Watkins and Dayan, 1992]
- [Tsitsiklis, 1994]
- [Jaakkola et al., 1994]
- [Szepesvári, 1998]
- [Kearns and Singh, 1999]
- [Borkar and Meyn, 2000]
- [Even-Dar and Mansour, 2003]
- [Beck and Srikant, 2012]
- [Jin et al., 2018]
- [Shah and Xie, 2018]
- [Wainwright, 2019a]
- [Chen et al., 2019]
- [Yang and Wang, 2019]
- [Du et al., 2020]
- [Chen et al., 2020]
- [Qu and Wierman, 2020]
- [Devraj and Meyn, 2020]
- ...

Prior art: sample complexity

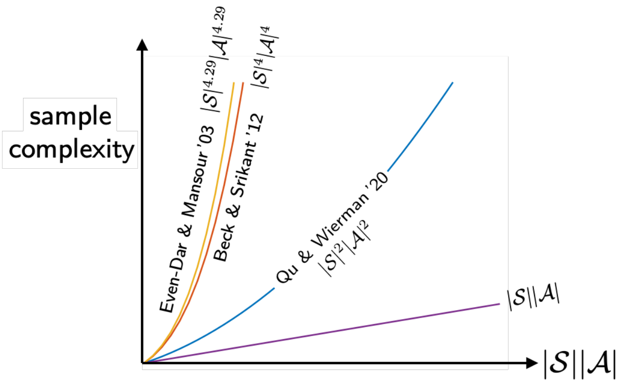
Question: how many samples are needed to ensure $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$?

paper	sample complexity	learning rate
[Even-Dar and Mansour, 2003]	$\frac{(t_{\text{cover}})^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\varepsilon^2}$	linear: $\frac{1}{t}$
[Even-Dar and Mansour, 2003]	$\big(\frac{t_{\text{cover}}^{1+3\omega}}{(1-\gamma)^4\varepsilon^2}\big)^{\frac{1}{\omega}} + \big(\frac{t_{\text{cover}}}{1-\gamma}\big)^{\frac{1}{1-\omega}}$	poly: $\frac{1}{t^\omega}$, $\omega \in (\frac{1}{2}, 1)$
[Beck and Srikant, 2012]	$\frac{t_{\text{cover}}^3 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$	constant
[Qu and Wierman, 2020]	$\frac{t_{\text{mix}}}{\mu_{\min}^2(1-\gamma)^5\varepsilon^2}$	rescaled linear

— cover time: $t_{\text{cover}} \asymp \frac{t_{\text{mix}}}{\mu_{\min}}$

Prior art: sample complexity

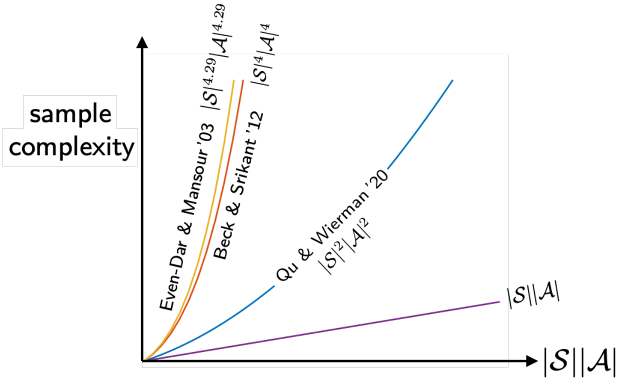
Question: how many samples are needed to ensure $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$?



if we take $\mu_{\mathsf{min}} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}$, $t_{\mathsf{cover}} \asymp \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}$

Prior art: sample complexity

Question: how many samples are needed to ensure $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$?

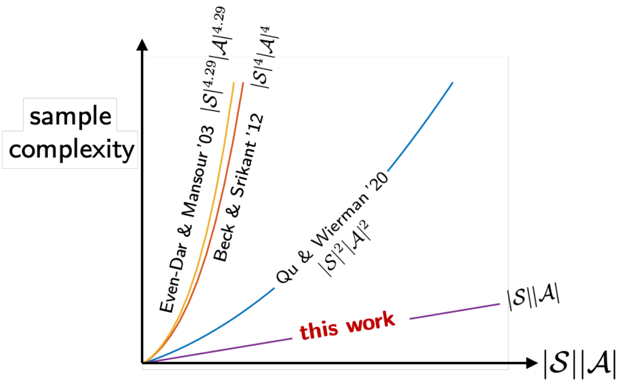


if we take $\mu_{\text{min}} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}$, $t_{\text{cover}} \asymp \frac{t_{\text{mix}}}{\mu_{\text{min}}}$

All prior results require sample size of at least $t_{\text{mix}}|\mathcal{S}|^2|\mathcal{A}|^2$!

This work: sample complexity

Question: how many samples are needed to ensure $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$?



if we take $\mu_{\mathsf{min}} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}$, $t_{\mathsf{cover}} \asymp \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}$

All prior results require sample size of at least $t_{\mathsf{mix}}|\mathcal{S}|^2|\mathcal{A}|^2$!

Main result: ℓ_∞-based sample complexity

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

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, sample complexity of async Q-learning to yield $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ is at most (up to some log factor)

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

Main result: ℓ_∞-based sample complexity

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

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, sample complexity of async Q-learning to yield $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ is at most (up to some log factor)

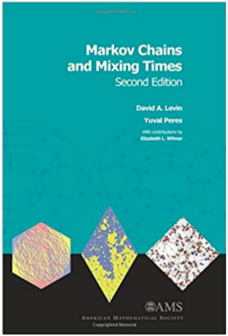
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

— *prior art:* $\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}$ ([Qu and Wierman, 2020])

- Improves upon prior art by **at least** $|\mathcal{S}||\mathcal{A}|$!

Effect of mixing time on sample complexity

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

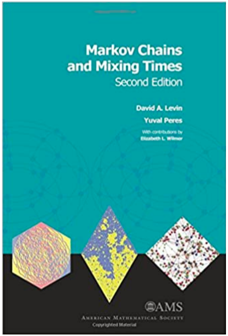


- reflects cost taken to reach steady state
- one-time expense (almost independent of ε)

  — it becomes amortized as algorithm runs

Effect of mixing time on sample complexity

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$



- reflects cost taken to reach steady state
- one-time expense (almost independent of ε)
  — it becomes amortized as algorithm runs

— *prior art:* $\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}$ ([Qu and Wierman, 2020])

Dependence on effective horizon

minimax lower bound
(Azar et al. '13)

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2}$$

asyn Q-learning
(ignoring dependency on t_{mix})

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}$$

Dependence on effective horizon

minimax lower bound (Azar et al. '13)

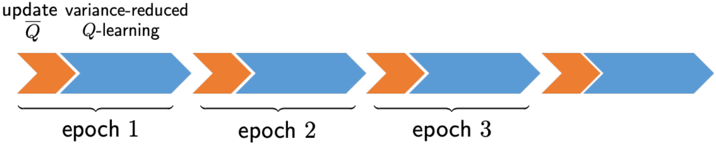
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2}$$

asyn Q-learning (ignoring dependency on t_{mix})

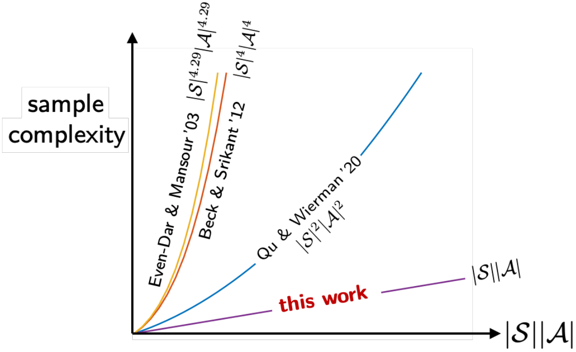
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}$$

The dependency on $\frac{1}{1-\gamma}$ can be tightened by *variance reduction*.

— *inspired by [Johnson and Zhang, 2013], [Wainwright, 2019b]*

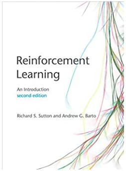


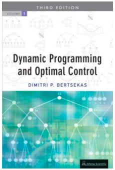
Summary of this part

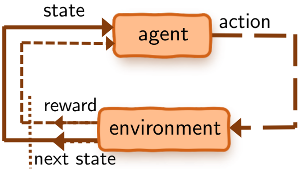


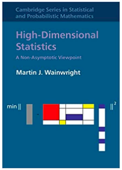
Sharper sample complexity for asyn Q-learning in terms of $|\mathcal{S}||\mathcal{A}|$ and t_{mix}!

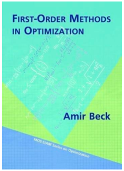
Concluding remarks











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

Future directions:

- function approximation
- multi-agent RL
- offline RL
- many more...

Thanks for your attention!

Other details

Improved theory for policy evaluation

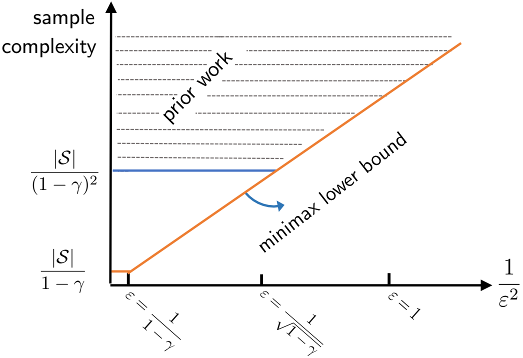
Model-based policy evaluation:

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

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 work (Agarwal et al. '19, Pananjady & Wainwright '19, Khamaru et al. '20)

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 (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*

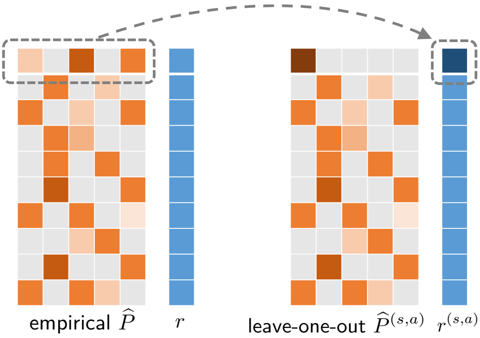
$$\|\widehat{V}^\pi - V^\pi\|_\infty \leq \varepsilon$$

with sample complexity at most

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

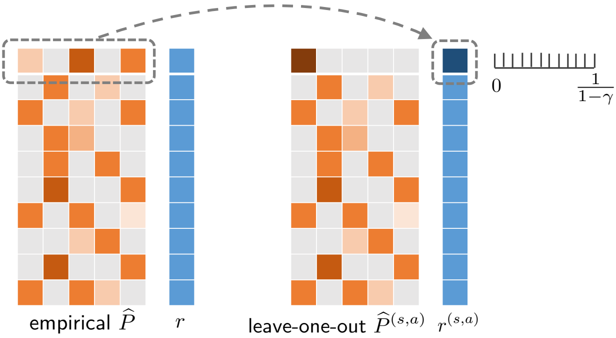
- Minimax optimal for all ε (Azar et al. '13, Pananjady & Wainwright '19)

Key idea 2: leave-one-out analysis for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$



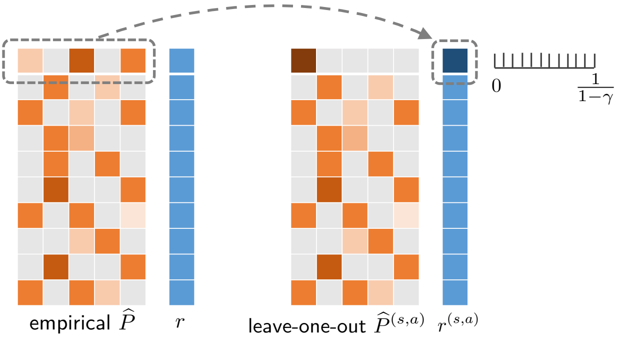
1. embed all randomness from $\widehat{P}(\cdot \mid s, a)$ into a single scalar (i.e. $r^{(s,a)}$)

Key idea 2: leave-one-out analysis for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$



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

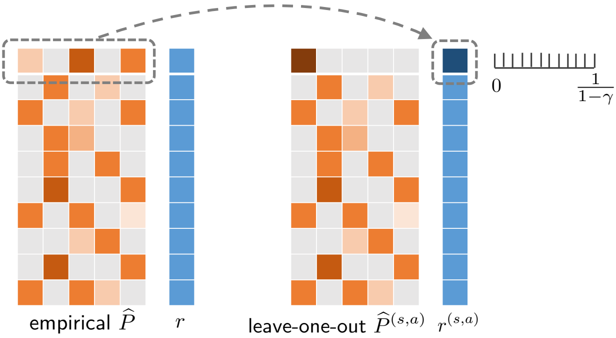
Key idea 2: leave-one-out analysis for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$



1. embed all randomness from $\widehat{P}(\cdot \mid s,a)$ into a single scalar (i.e. $r^{(s,a)}$)
2. build an ϵ-net for this scalar
3. $\widehat{\pi}^\star_{(s,a)} = \widehat{\pi}^\star$ 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 2: leave-one-out analysis for $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$



Compared to [Agarwal et al., 2019]

- [Agarwal et al., 2019]: dependency btw value $\widehat{V}$ & samples
- **Ours:** dependency btw policy $\widehat{\pi}$ & samples

Key decomposition for asyn Q-learning

Error decomposition

$$\boldsymbol{\Delta}_t = \big(\boldsymbol{I} - \boldsymbol{\Lambda}_t\big)\boldsymbol{\Delta}_{t-1} + \gamma\boldsymbol{\Lambda}_t\big(\boldsymbol{P}_t - \boldsymbol{P}\big)\boldsymbol{V}^{\star} + \gamma\boldsymbol{\Lambda}_t\boldsymbol{P}_t\big(\boldsymbol{V}_{t-1} - \boldsymbol{V}^{\star}\big)$$

Applying this relation recursively gives

$$\begin{aligned}\boldsymbol{\Delta}_t &= \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\big(\boldsymbol{P}_i - \boldsymbol{P}\big)\boldsymbol{V}^{\star} \\ &\quad + \gamma\sum_{i=1}^{t}\prod_{j=i+1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Lambda}_i\boldsymbol{P}_i\big(\boldsymbol{V}_{i-1} - \boldsymbol{V}^{\star}\big) + \prod_{j=1}^{t}\big(\boldsymbol{I} - \boldsymbol{\Lambda}_j\big)\boldsymbol{\Delta}_0\end{aligned}$$

Learning rates

constant stepsize $\eta_t \equiv \min\left\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2}, \frac{1}{t_{\mathsf{mix}}}\right\}$

- [Qu and Wierman, 2020]: rescaled linear $\eta_t = \frac{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)}}{t+\max\{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)}, t_{\mathsf{mix}}\}}$
- [Beck and Srikant, 2012] constant $\underbrace{\eta_t \equiv \frac{(1-\gamma)^4\varepsilon^2}{|\mathcal{S}||\mathcal{A}|t_{\mathsf{cover}}^2}}_{\text{too conservative}}$
- [Even-Dar and Mansour, 2003]: polynomial $\eta_t = t^{-\omega}$ $(\omega \in (\frac{1}{2}, 1])$

Adaptive learning rates

$$\eta_t = \min\left\{1, c\exp\left(\left\lfloor \log \frac{\log t}{\widehat{\mu}_{\mathsf{min},t}(1-\gamma)\gamma^2 t}\right\rfloor\right)\right\}$$

$$\widehat{\mu}_{\mathsf{min},t} = \begin{cases} \frac{1}{|\mathcal{S}||\mathcal{A}|}, & \min_{s,a} K_t(s,a) = 0; \\ \widehat{\mu}_{\mathsf{min},t-1}, & \frac{1}{2} < \frac{\min_{s,a} K_t(s,a)/t}{\widehat{\mu}_{\mathsf{min},t-1}} < 2; \\ \min_{s,a} K_t(s,a)/t, & \text{otherwise.} \end{cases}$$

One strategy: variance reduction

— *inspired by [Johnson and Zhang, 2013], [Wainwright, 2019b]*

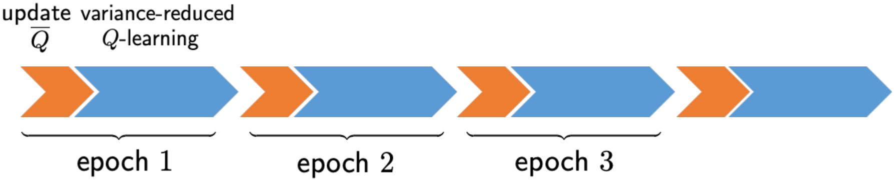
Variance-reduced Q-learning updates

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

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

Variance-reduced Q-learning

— *inspired by [Johnson and Zhang, 2013], [Wainwright, 2019b]*



for each epoch

1. update $\overline{Q}$ and $\widetilde{T}(\overline{Q})$
2. run variance-reduced Q-learning updates