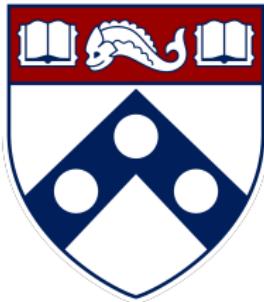


# **Reinforcement learning (Part 1): Basics and Model-based RL**

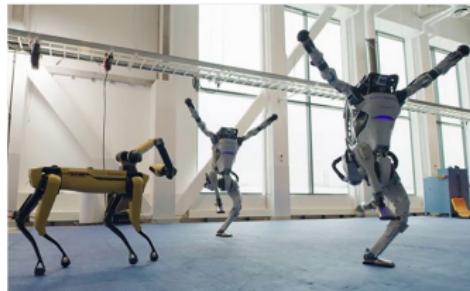


Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

# Successes of reinforcement learning (RL)

---



# Supervised learning

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



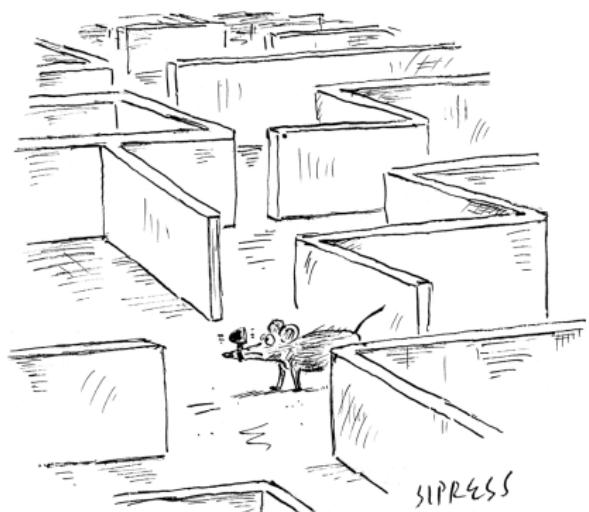
— pic from internet

# Reinforcement learning (RL)

---

In RL, an agent learns by interacting with an environment

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



"Recalculating ... recalculating ..."

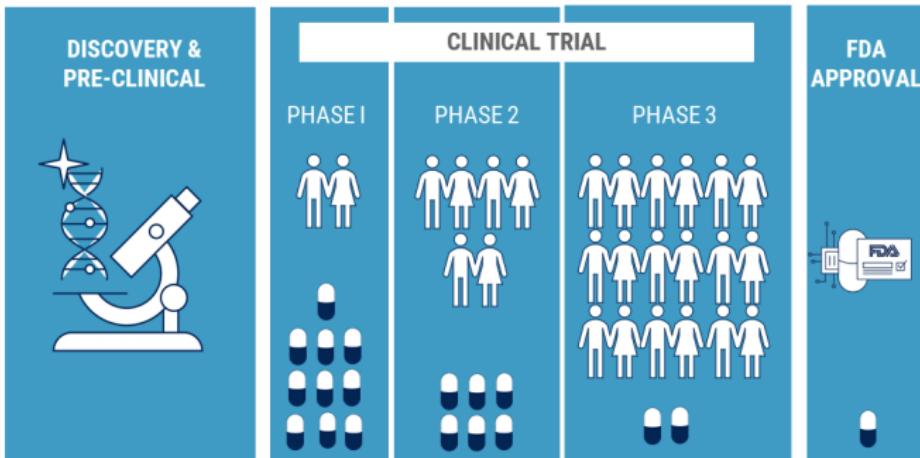
# Challenges of RL

---

- explore or exploit: unknown or changing environments
- credit assignment problem: delayed rewards or feedback
- enormous state and action space
- nonconvex optimization



# Sample efficiency

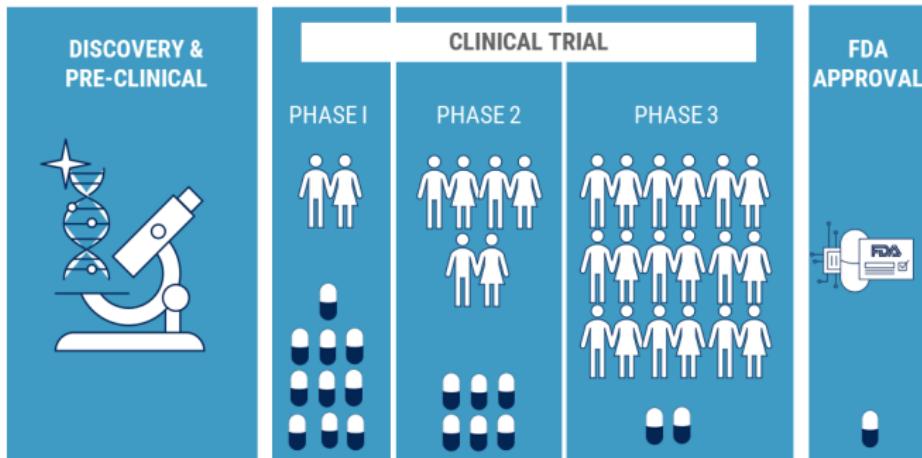


Source: cbinsights.com

CB INSIGHTS

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

# Sample efficiency



Source: cbinsights.com

CBINSIGHTS

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

**Challenge:** how to design sample-efficient RL algorithms?

# Statistical foundation of RL

---



## The Contributions of Herbert Robbins to Mathematical Statistics

Tze Leung Lai and David Siegmund

### 2. STOCHASTIC APPROXIMATION AND ADAPTIVE DESIGN

In 1951, Robbins and his student, Sutton Monro, founded the subject of stochastic approximation with the publication of their celebrated paper [26]. Consider the problem of finding the root  $\theta$  (assumed unique) of an equation  $g(x) = 0$ . In the classical

### 4. SEQUENTIAL EXPERIMENTATION AND OPTIMAL STOPPING

The well known “multiarmed bandit problem” in the statistics and engineering literature, which is prototypical of a wide variety of adaptive control and design problems, was first formulated and studied by Robbins [28]. Let  $A, B$  denote two statistical populations with finite means  $\mu_A, \mu_B$ . How should we draw a



Herbert Robbins



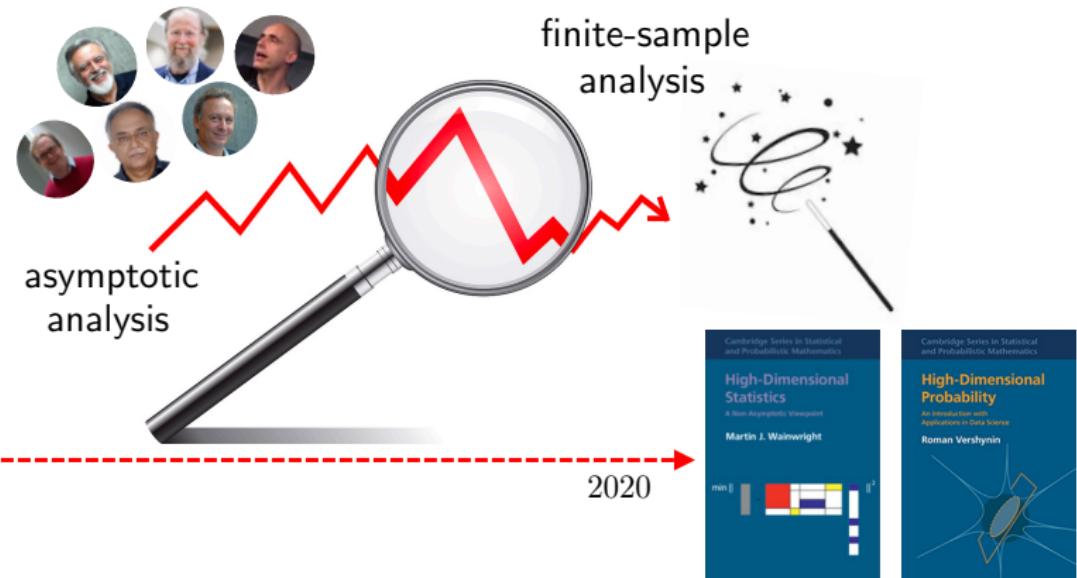
David Blackwell

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

Peter J. Bickel<sup>a,1</sup>

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.

# Statistical foundation of RL



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

# Outline (Part 1)

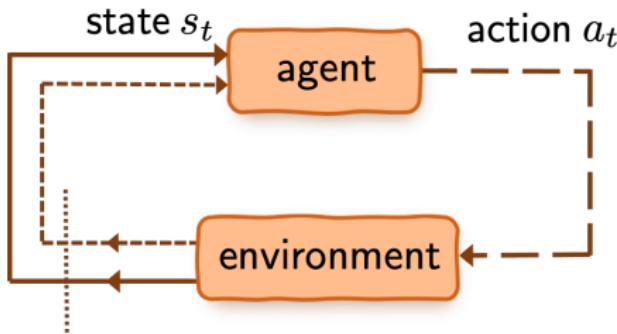
---

- Basics of Markov decision processes
- Basic algorithms for policy evaluation/maximization
- RL with a generative model

*Background: Markov decision processes*

# Markov decision process (MDP)

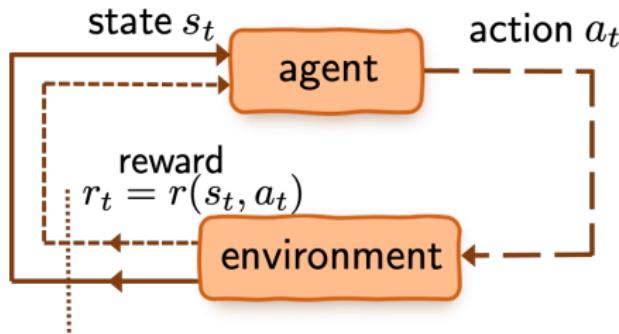
---



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

# Markov decision process (MDP)

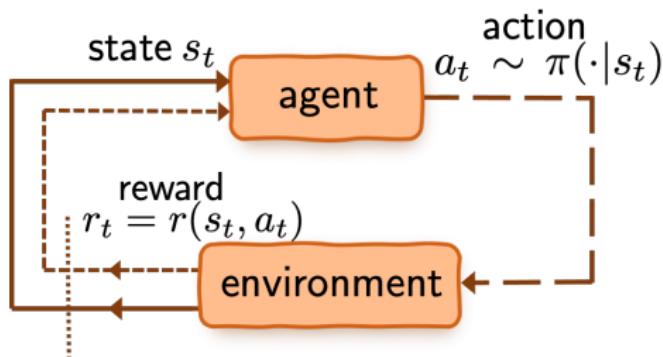
---



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

# Markov decision process (MDP)

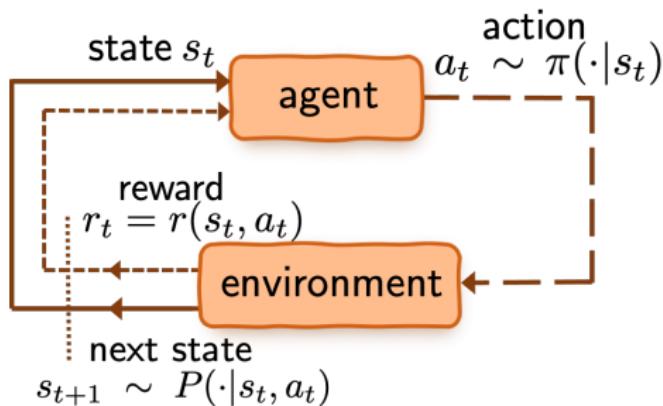
---



- $\mathcal{S}$ : 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)

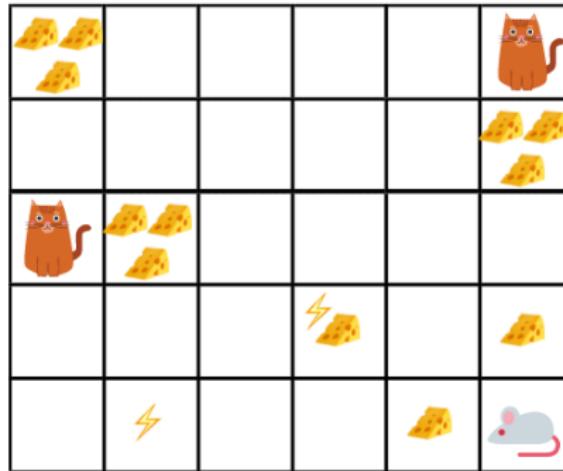
---



- $\mathcal{S}$ : 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

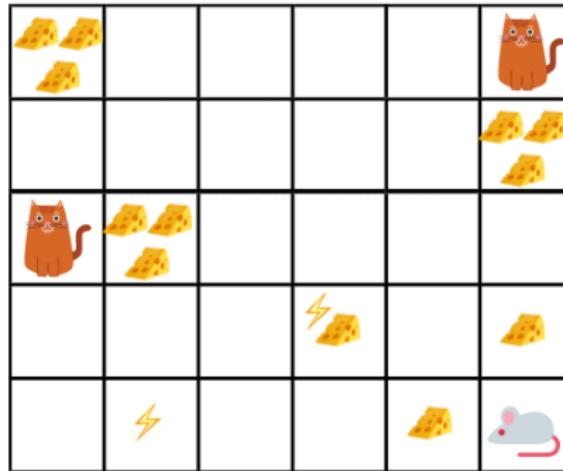
# Help the mouse!

---



# Help the mouse!

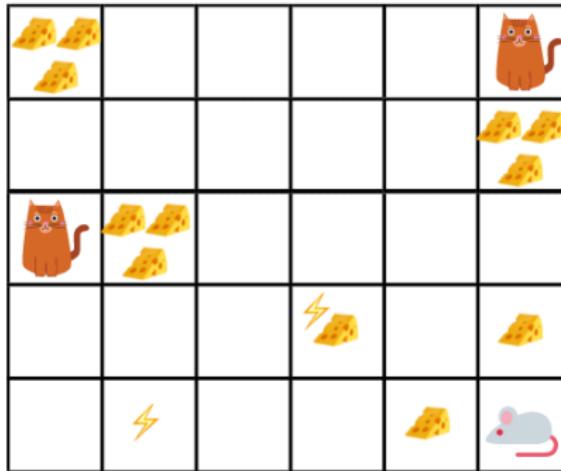
---



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

# Help the mouse!

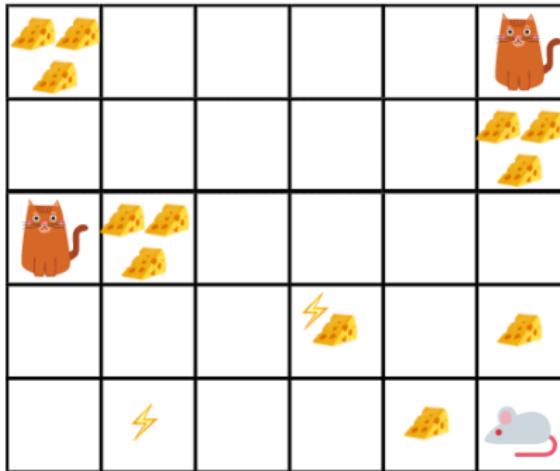
---



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

# Help the mouse!

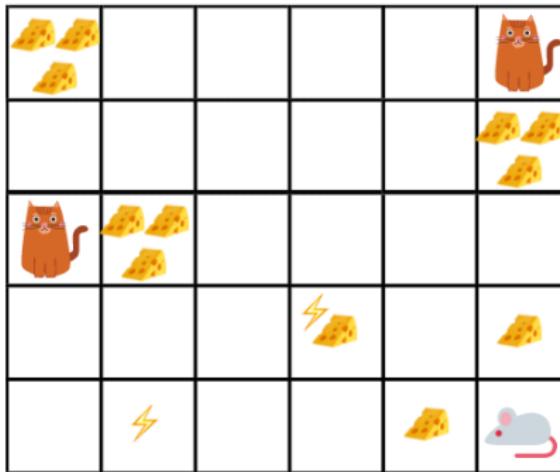
---



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

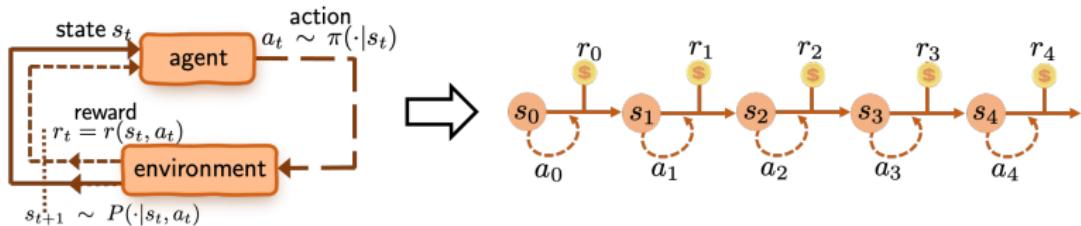
# Help the mouse!

---



- 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

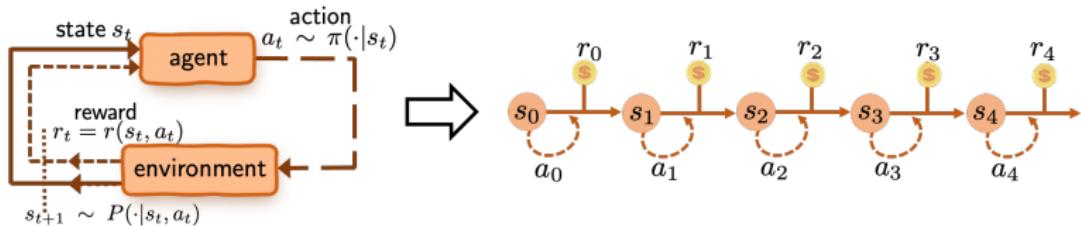
# Value function



Value of policy  $\pi$ : cumulative **discounted** reward

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

# Value function



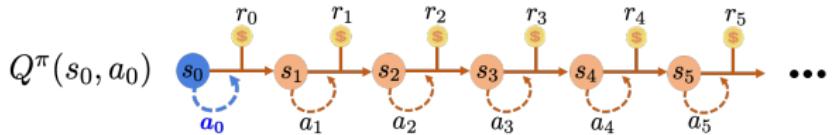
Value of policy  $\pi$ : cumulative **discounted** reward

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

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

# Q-function (action-value function)

---



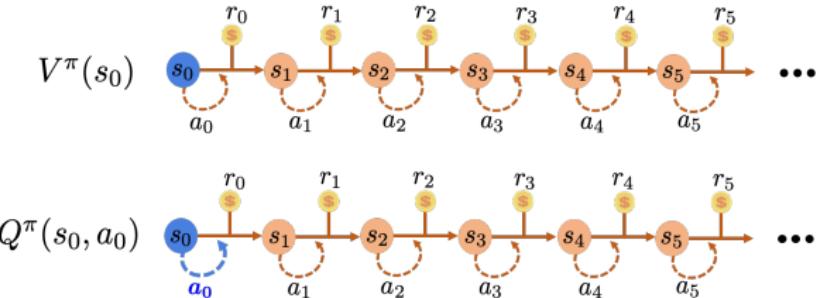
Q-function of policy  $\pi$ :

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

- $(\textcolor{red}{a_0}, s_1, a_1, s_2, a_2, \dots)$ : induced by policy  $\pi$

# Q-function (action-value function)

---



Q-function of policy  $\pi$ :

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

- $(\textcolor{red}{a_0}, s_1, a_1, s_2, a_2, \dots)$ : induced by policy  $\pi$

# Optimal policy and optimal value



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

## Theorem 1 (Puterman'94)

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

$$V^{\pi^*}(s) \geq V^{\pi}(s), \quad \forall s, \pi.$$

# Optimal policy and optimal value

---



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

# Optimal policy and optimal value

---



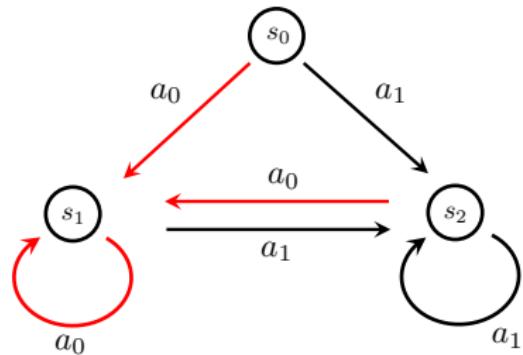
- **optimal policy**  $\pi^*$ : maximizing value function  $\max_{\pi} V^{\pi}(s)$
- optimal value / Q function:  $V^* := V^{\pi^*}$ ,  $Q^* := Q^{\pi^*}$
- How to find this  $\pi^*$ ?

# Example

---

Consider a **deterministic** MDP with 3 states & 2 actions

What is the optimal policy?



Reward:  $r(s_1, a_0) = 1, 0$  else where

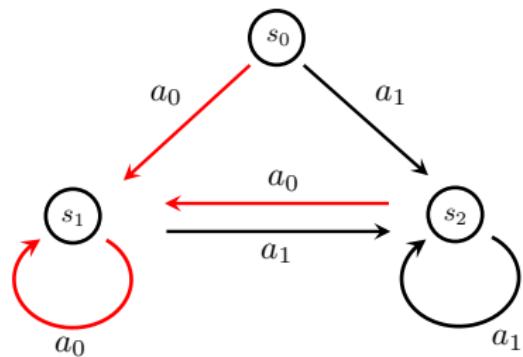
# Example

---

Consider a **deterministic** MDP with 3 states & 2 actions

What is the optimal policy?

- $\pi^*(s) = a_0, \forall s$

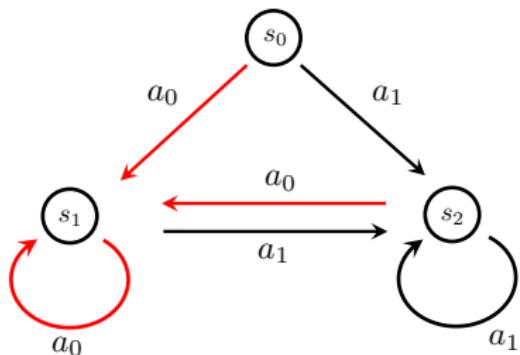


Reward:  $r(s_1, a_0) = 1, 0$  else where

# Example

---

Consider a **deterministic** MDP with 3 states & 2 actions



What is the optimal policy?

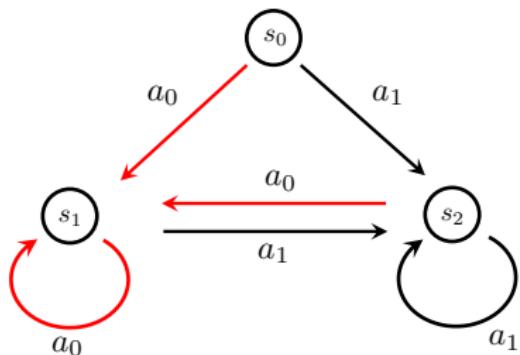
- $\pi^*(s) = a_0, \forall s$
- $V^*(s_0) = \frac{\gamma}{1-\gamma},$   
 $V^*(s_1) = \frac{1}{1-\gamma}, V^*(s_2) = \frac{\gamma}{1-\gamma}$

Reward:  $r(s_1, a_0) = 1, 0$  else where

# Example

---

Consider a **deterministic** MDP with 3 states & 2 actions



Reward:  $r(s_1, a_0) = 1, 0 \text{ else where}$

What is the optimal policy?

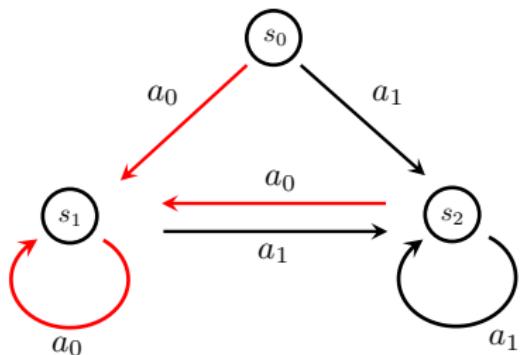
- $\pi^*(s) = a_0, \forall s$
- $V^*(s_0) = \frac{\gamma}{1-\gamma},$   
 $V^*(s_1) = \frac{1}{1-\gamma}, V^*(s_2) = \frac{\gamma}{1-\gamma}$

What is  $V^\pi$  for  $\pi(s) = a_1, \forall s$ ?

# Example

---

Consider a **deterministic** MDP with 3 states & 2 actions



Reward:  $r(s_1, a_0) = 1, 0 \text{ else where}$

What is the optimal policy?

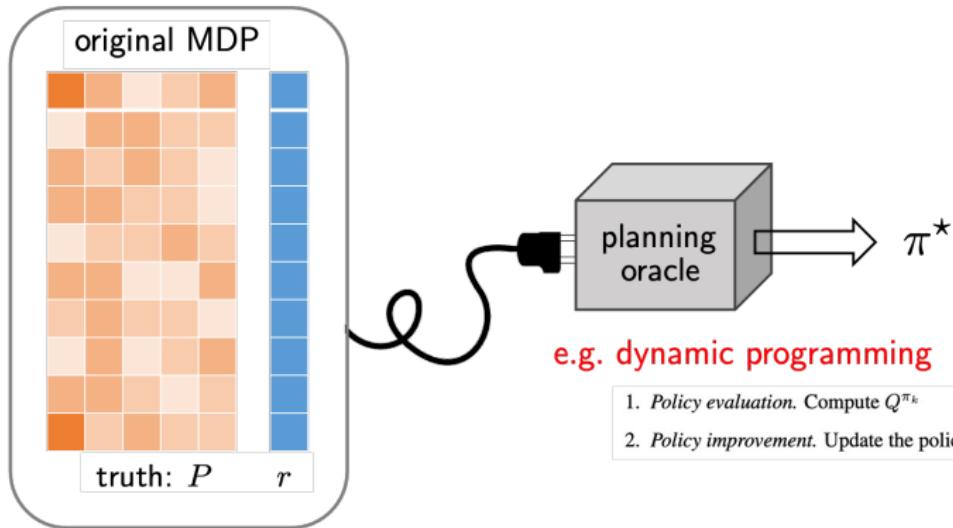
- $\pi^*(s) = a_0, \forall s$
- $V^*(s_0) = \frac{\gamma}{1-\gamma},$   
 $V^*(s_1) = \frac{1}{1-\gamma}, V^*(s_2) = \frac{\gamma}{1-\gamma}$

What is  $V^\pi$  for  $\pi(s) = a_1, \forall s$ ?

- $V^\pi(s) = 0, \forall s$

*Background: Basic dynamic programming algorithms*

# When the model is known . . .



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

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

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

*Possible scheme:*

- exact policy evaluation for each  $\pi$
- find the optimal one

## Policy evaluation: Bellman's consistency equation

---

- $V^\pi / Q^\pi$ : value / action-value function under policy  $\pi$

# Policy evaluation: Bellman's consistency equation

---

- $V^\pi / Q^\pi$ : value / action-value function under policy  $\pi$

## Bellman's consistency equation

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



*Richard Bellman*

# Policy evaluation: Bellman's consistency equation

---

- $V^\pi / Q^\pi$ : value / action-value function under policy  $\pi$

## Bellman's consistency equation

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

- one-step look-ahead



*Richard Bellman*

# Policy evaluation: Bellman's consistency equation

---

- $V^\pi / Q^\pi$ : value / action-value function under policy  $\pi$

## Bellman's consistency equation

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

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

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



Richard Bellman

# Bellman's optimality principle

---

## Bellman operator

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

- one-step look-ahead

# Bellman's optimality principle

## Bellman operator

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

- one-step look-ahead

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

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

**$\gamma$ -contraction of Bellman operator:**

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



*Richard Bellman*

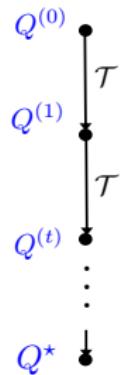
# Value iteration (VI)

---

## Value iteration (VI)

Initialize at  $Q = 0$ . For  $t = 0, 1, \dots$ ,

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



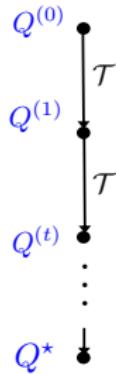
# Value iteration (VI)

---

## Value iteration (VI)

Initialize at  $Q = 0$ . For  $t = 0, 1, \dots$ ,

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



## Iterative algorithm for fix-point solution:

Initialize at 0, repeat  $x^{t+1} = f(x^t)$ . If  $f$  is a contraction mapping, then  $x^t \rightarrow x^*$ .

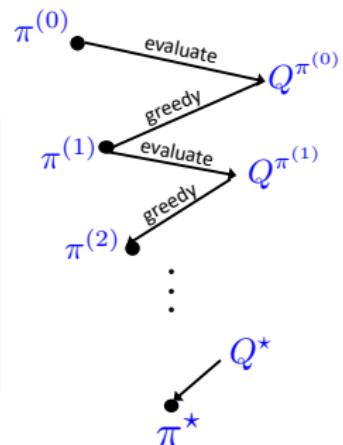
# Policy iteration (PI)

## Policy iteration (PI)

Initialize at  $Q = 0$ . For  $t = 0, 1, \dots$ ,

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

**policy improvement:**  $\pi^{(t+1)}(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q^{(t)}(s, a)$



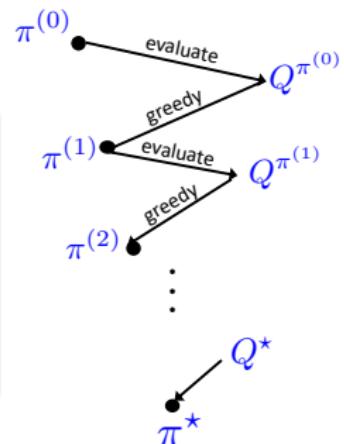
# Policy iteration (PI)

## Policy iteration (PI)

Initialize at  $Q = 0$ . For  $t = 0, 1, \dots$ ,

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

**policy improvement:**  $\pi^{(t+1)}(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q^{(t)}(s, a)$



## Monotonic improvement:

$$Q^{\pi^{t+1}}(s, a) \geq Q^{\pi^t}(s, a) \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}$$

# Iteration complexity

---

**Theorem 1 (Linear convergence of policy/value iteration)**

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

# Iteration complexity

---

**Theorem 1 (Linear convergence of policy/value iteration)**

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

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

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

# Iteration complexity

**Theorem 1 (Linear convergence of policy/value iteration)**

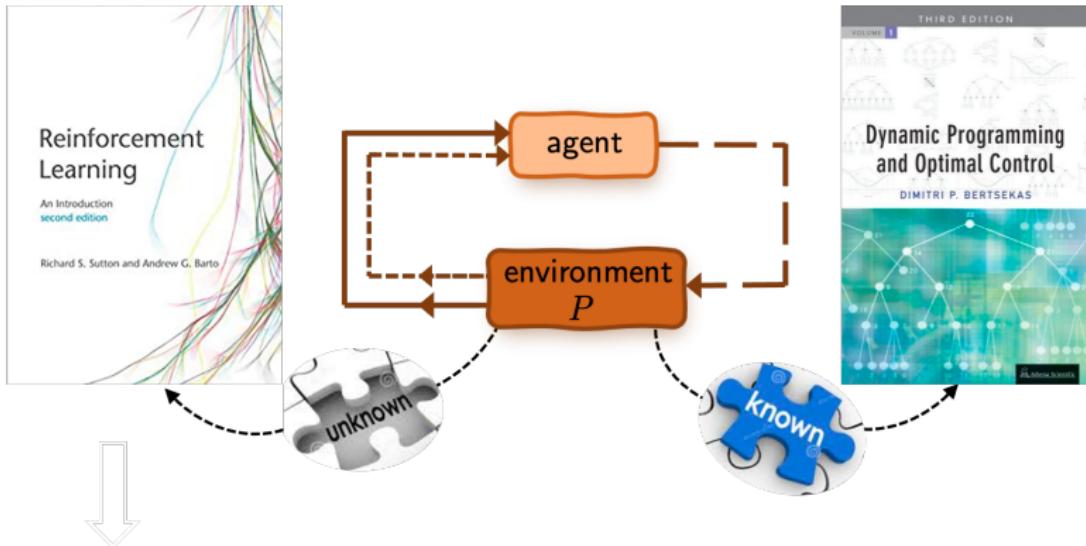
$$\|Q^{(t)} - Q^*\|_\infty \leq \gamma^t \|Q^{(0)} - Q^*\|_\infty$$

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

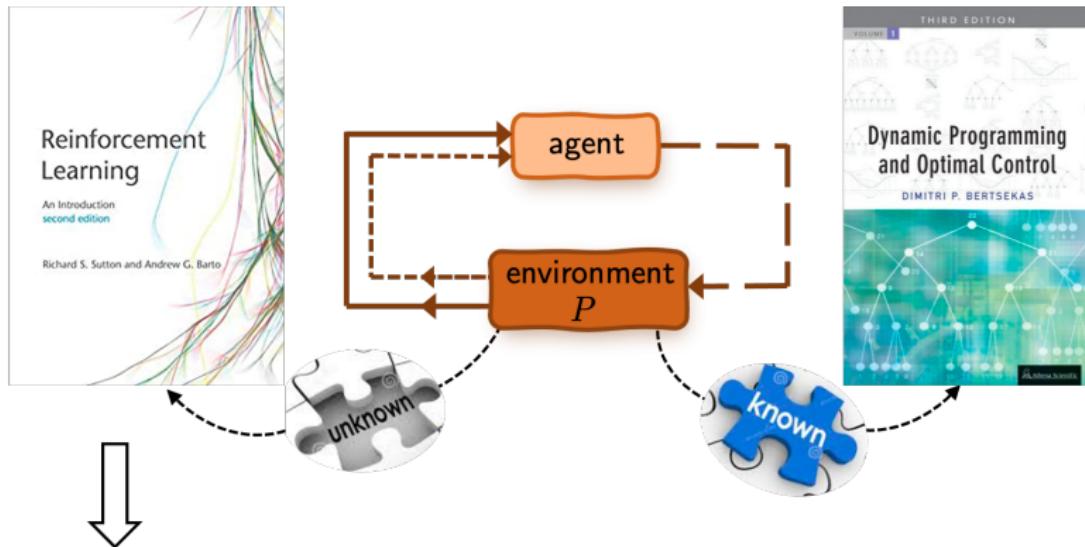
$$\frac{1}{1-\gamma} \log \left( \frac{\|Q^{(0)} - Q^*\|_\infty}{\varepsilon} \right) \text{ iterations}$$

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

# When the model is unknown ...



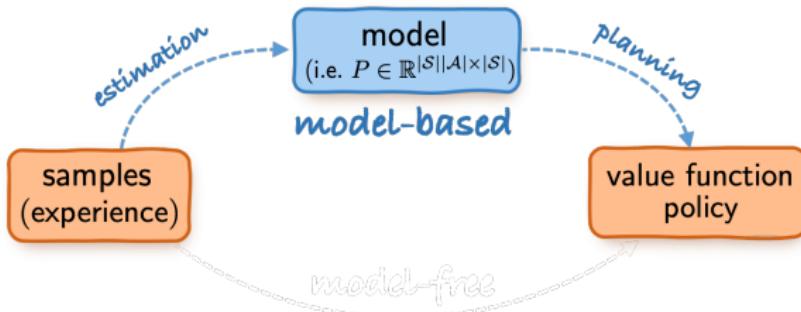
# When the model is unknown ...



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

# Two approaches

---

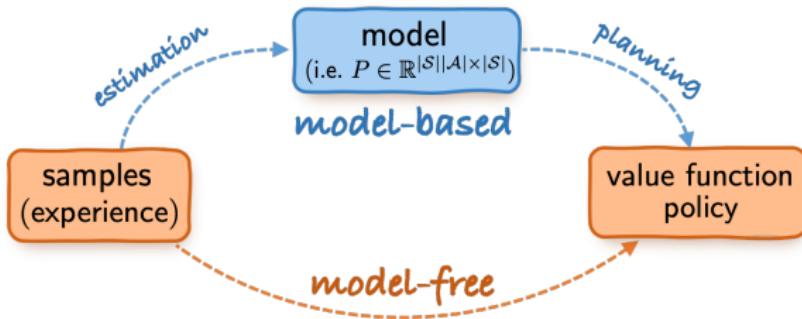


## Model-based approach (“plug-in”)

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

# Two approaches

---



## Model-based approach (“plug-in”)

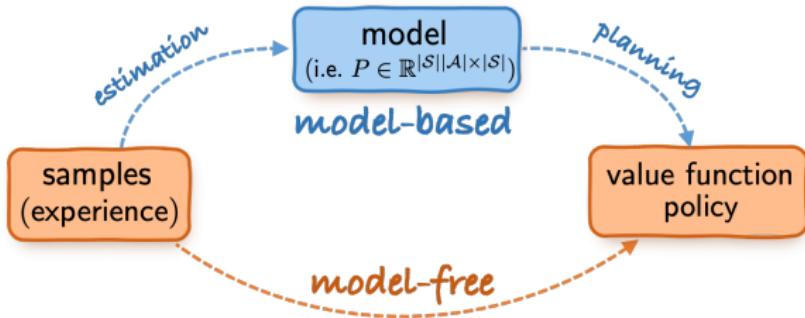
1. build an empirical estimate  $\hat{P}$  for  $P$
2. planning based on the empirical  $\hat{P}$

## Model-free approach (e.g. Q-learning; part iii)

— learning w/o estimating the model explicitly

# Two approaches

---



## Model-based approach (“plug-in”)

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

## Model-free approach (e.g. Q-learning; part iii)

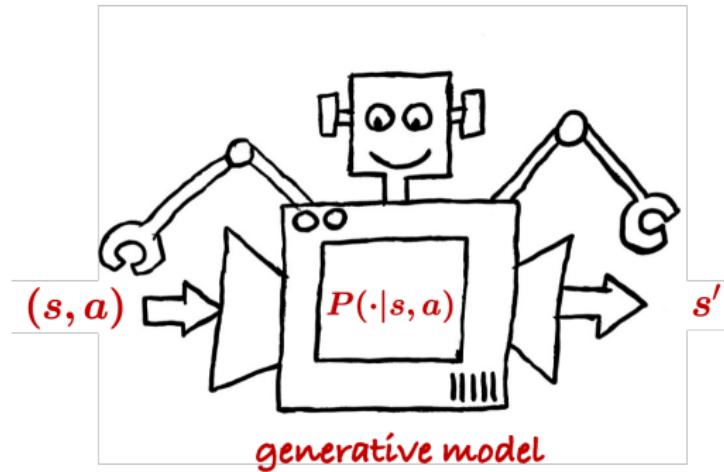
— learning w/o estimating the model explicitly

## **Model-based RL (a “plug-in” approach)**

1. Sampling from a generative model (simulator)
2. Offline RL / batch RL

# Sampling from a generative model

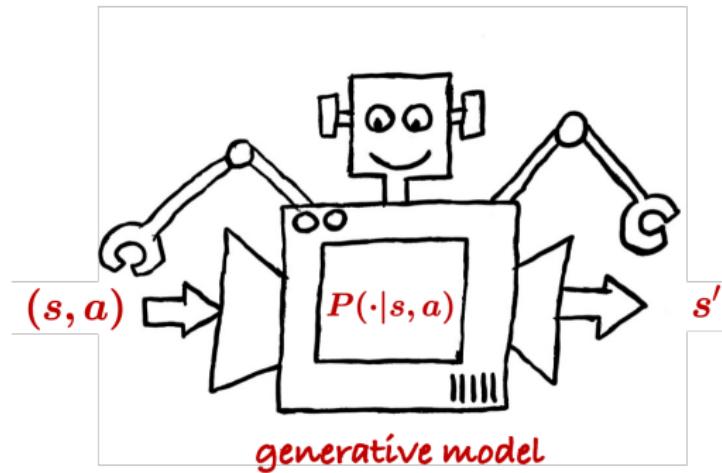
— Kearns, Singh '99



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

# Sampling from a generative model

— Kearns, Singh '99



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

$\ell_\infty$ -sample complexity: how many samples are required to  
learn an  $\varepsilon$ -optimal policy ?

$$\forall s: \hat{V^\pi}(s) \geq V^*(s) - \varepsilon$$

# An incomplete list of prior art

---

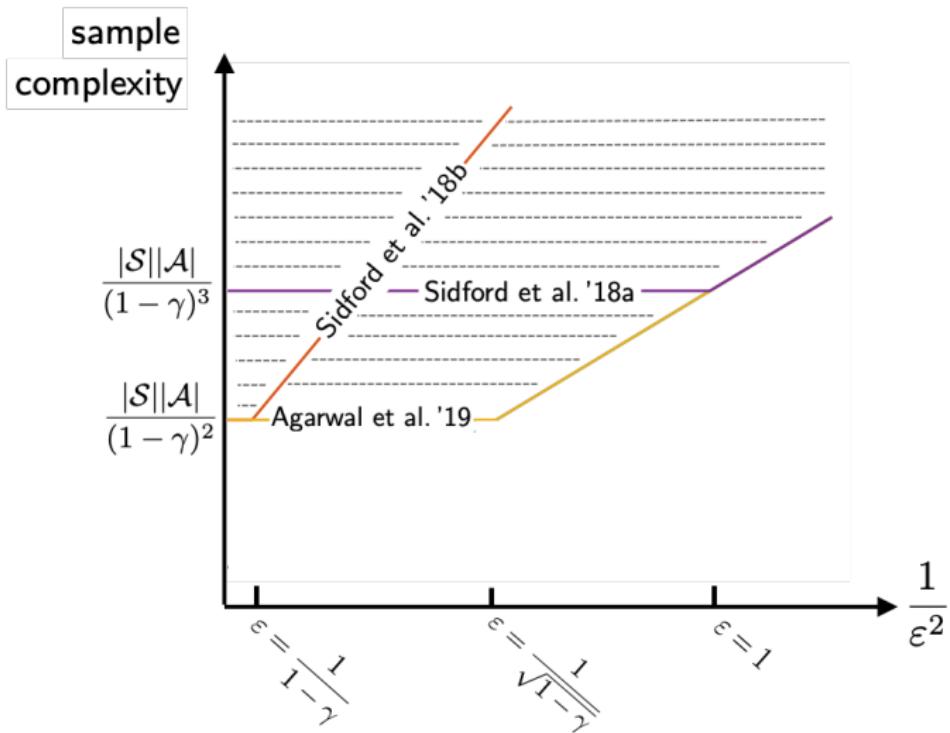
- Kearns & Singh '99
- Kakade '03
- Kearns, Mansour & Ng '02
- Azar, Munos & Kappen '12
- Azar, Munos, Ghavamzadeh & Kappen '13
- Sidford, Wang, Wu, Yang & Ye '18
- Sidford, Wang, Wu & Ye '18
- Wang '17
- Agarwal, Kakade & Yang '19
- Wainwright '19a
- Wainwright '19b
- Pananjady & Wainwright '20
- Yang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

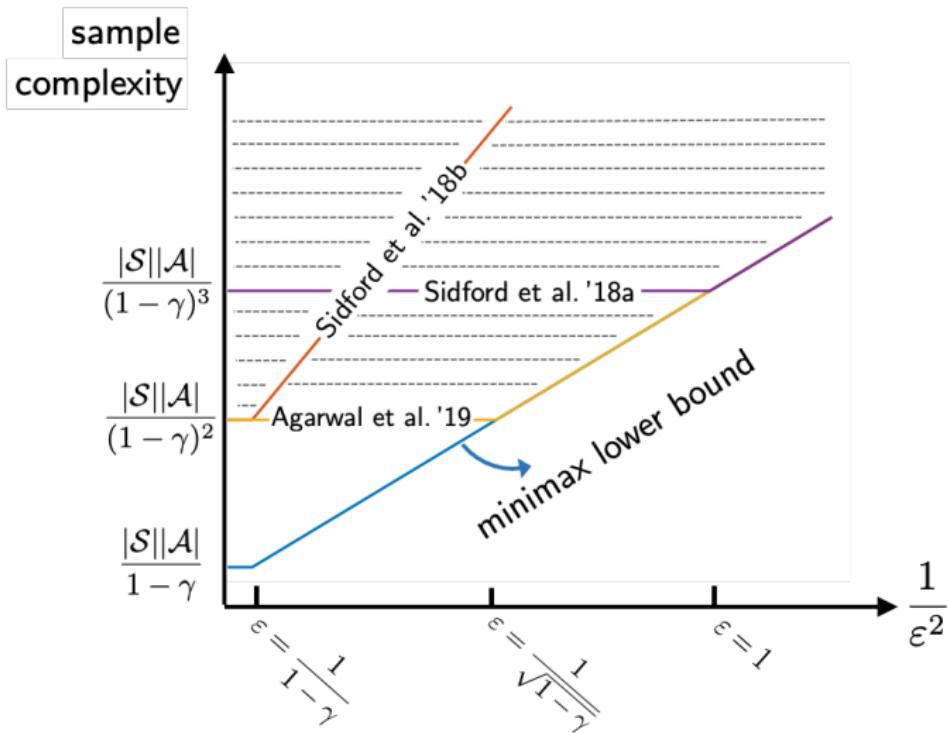
## An even shorter list of prior art

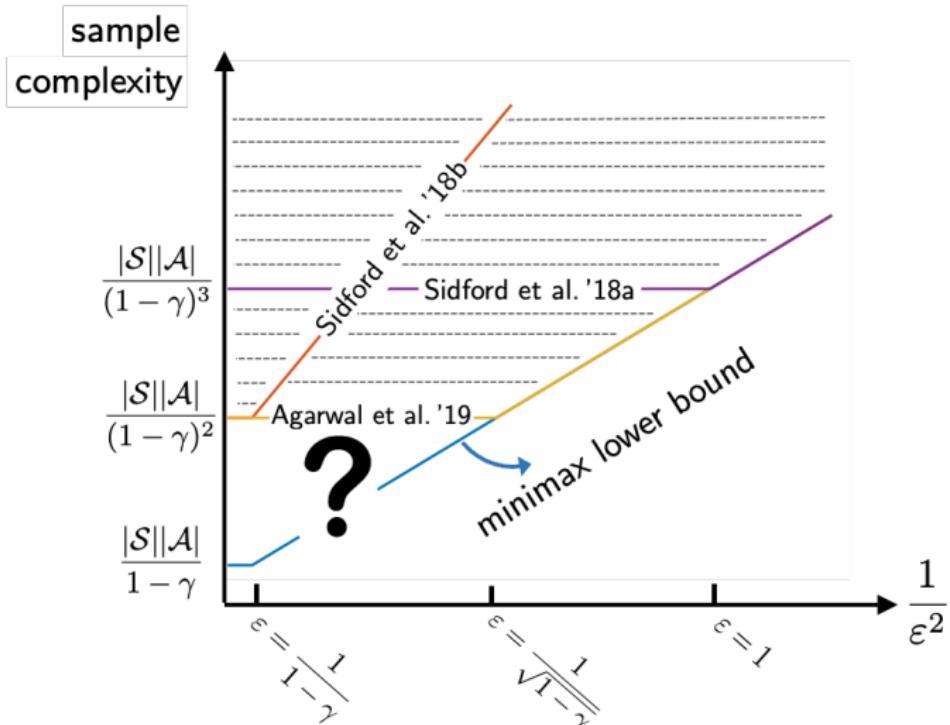
---

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

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



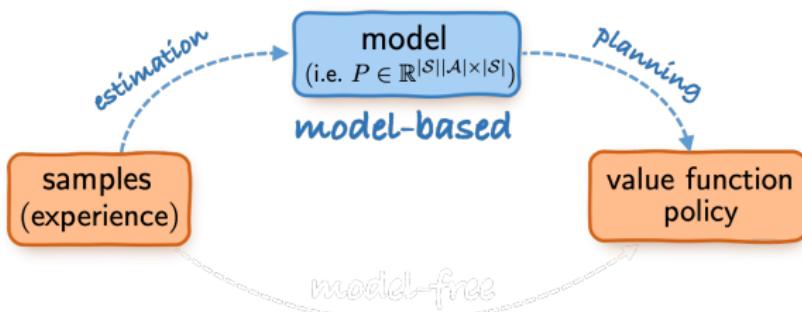




All prior theory requires sample size  $> \underbrace{\frac{|S||\mathcal{A}|}{(1 - \gamma)^2}}_{\text{sample size barrier}}$

# Our algorithm: model-based RL

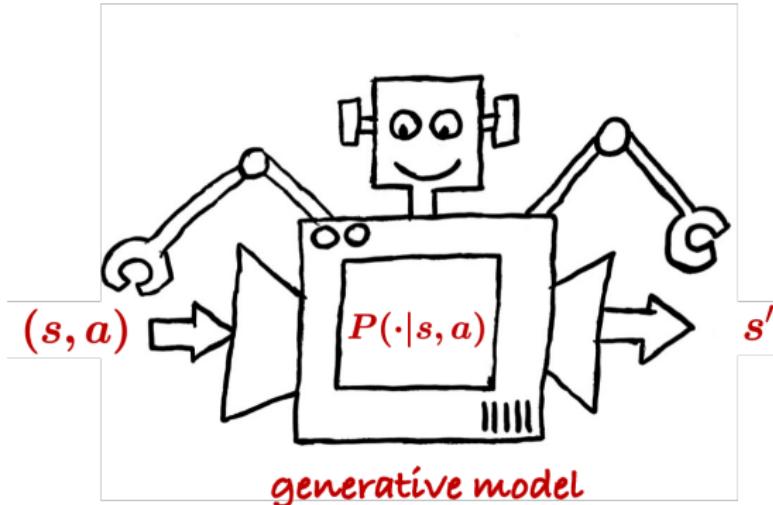
---



## Model-based approach (“plug-in”)

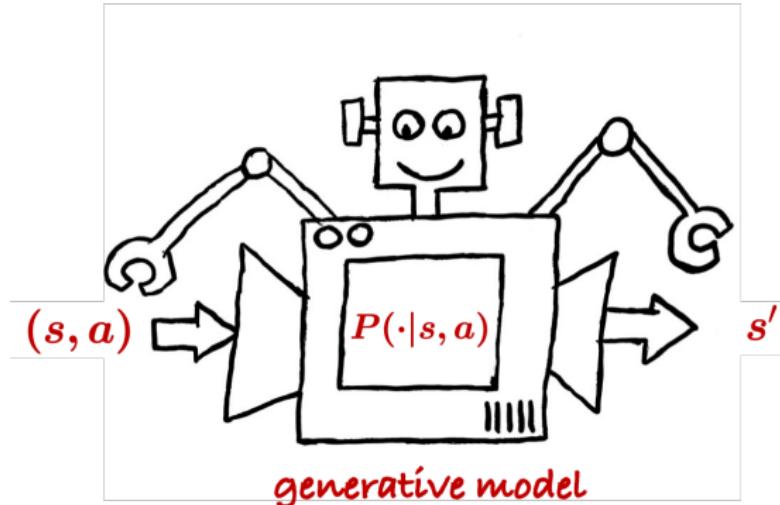
1. build an empirical estimate  $\hat{P}$  for  $P$
2. planning based on empirical  $\hat{P}$

# Model estimation



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

# Model estimation



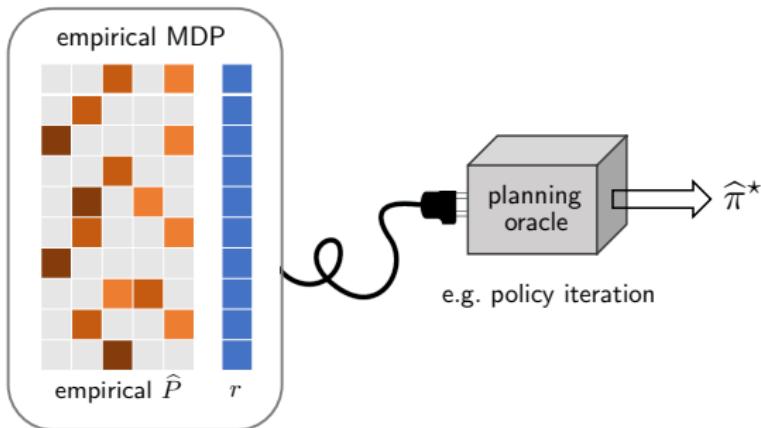
**Sampling:** for each  $(s, a)$ , collect  $N$  ind. samples  $\{(s, a, s'_{(i)})\}_{1 \leq i \leq N}$

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

# Model-based (plug-in) estimator

---

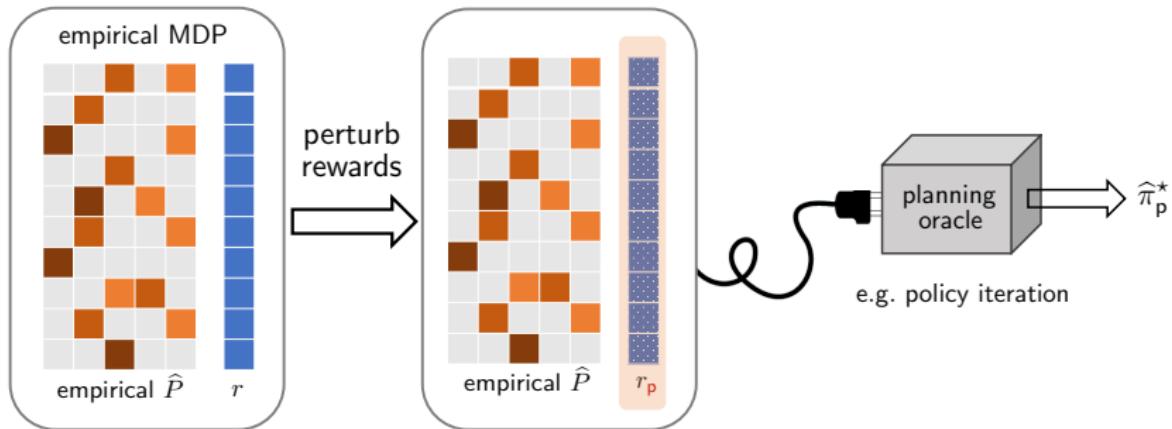
— Azar et al. '13, Agarwal et al. '19, Pananjady et al. '20



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

# Our method: plug-in estimator + perturbation

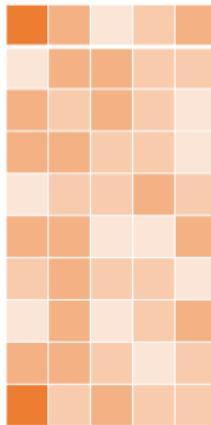
— Li, Wei, Chi, Gu, Chen '20



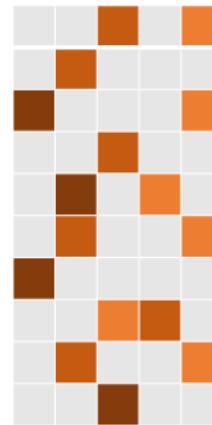
Run planning algorithms based on the *empirical* MDP

# Challenges in the sample-starved regime

---



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

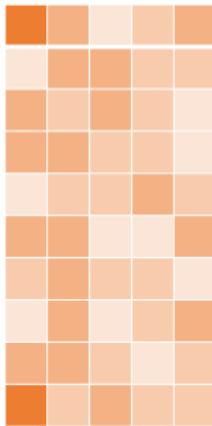


empirical estimate:  
 $\hat{P}$

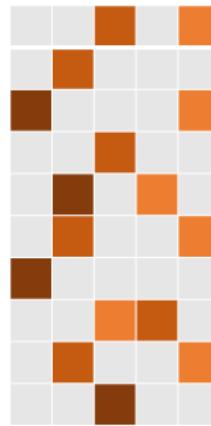
- Can't recover  $P$  faithfully if sample size  $\ll |\mathcal{S}|^2|\mathcal{A}|$ !

# Challenges in the sample-starved regime

---



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



empirical estimate:  
 $\hat{P}$

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

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

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

with high prob., with sample complexity at most

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

## Main result

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

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

$$\|V^{\hat{\pi}_p^*} - V^*\|_\infty \leq \varepsilon$$

with high prob., with sample complexity at most

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

- $\hat{\pi}_p^*$ : obtained by empirical QVI or PI within  $\tilde{O}\left(\frac{1}{1-\gamma}\right)$  iterations

# Main result

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

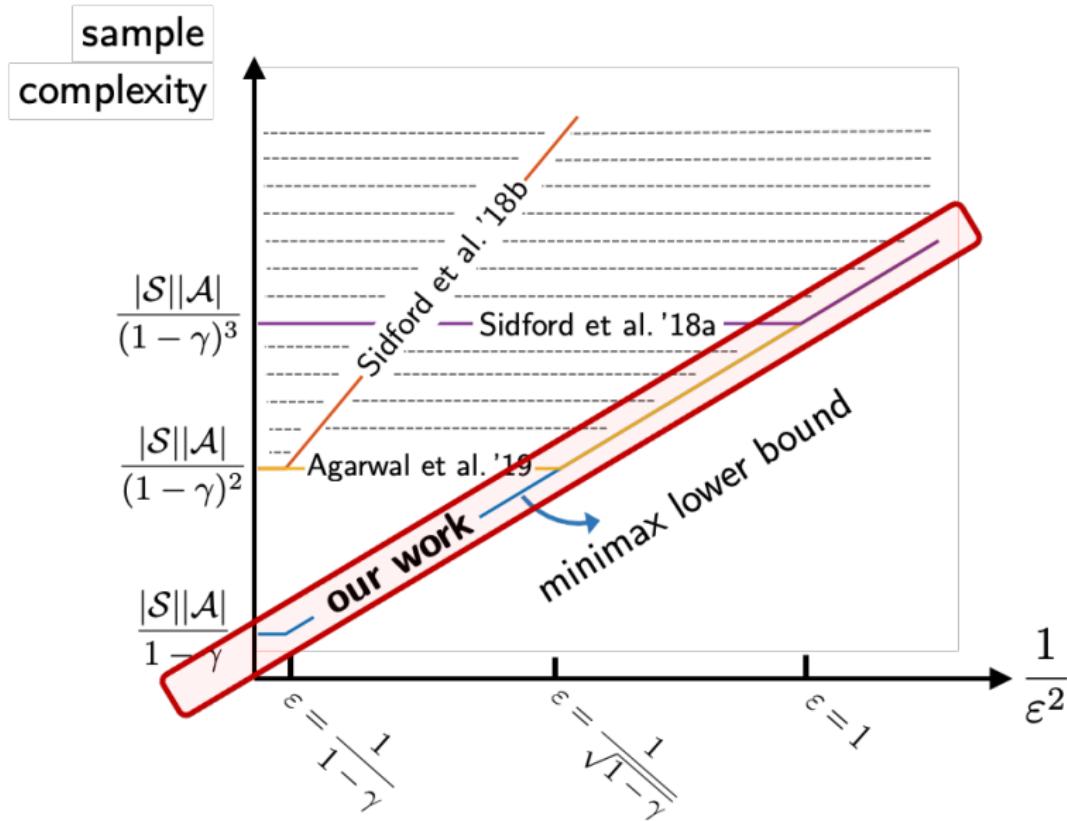
For any  $0 < \varepsilon \leq \frac{1}{1-\gamma}$ , the optimal policy  $\widehat{\pi}_p^*$  of the perturbed empirical MDP achieves

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

with high prob., with sample complexity at most

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

- $\widehat{\pi}_p^*$ : obtained by empirical QVI or PI within  $\tilde{O}\left(\frac{1}{1-\gamma}\right)$  iterations
- **Minimax lower bound:**  $\tilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$  (Azar et al. '13)



## **Model-based RL (a “plug-in” approach)**

1. Sampling from a generative model (simulator)
2. Offline RL / batch RL

# Offline RL / Batch RL

---

- Collecting new data might be expensive or time-consuming
- Having stored tons of historical data



medical records



data of self-driving



clicking times of ads

# Offline RL / Batch RL

---

- Collecting new data might be expensive or time-consuming
- Having stored tons of historical data



medical records



data of self-driving



clicking times of ads

**Can we design algorithms based solely on historical data?**

# Offline RL / Batch RL

---

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

$$s \sim \rho^b, \quad a \sim \pi^b(\cdot | s), \quad s' \sim P(\cdot | s, a)$$

for some state distribution  $\rho^b$  and behavior policy  $\pi^b$

# Offline RL / Batch RL

---

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

$$s \sim \rho^b, \quad a \sim \pi^b(\cdot | s), \quad s' \sim P(\cdot | s, a)$$

for some state distribution  $\rho^b$  and behavior policy  $\pi^b$

**Goal:** given some test distribution  $\rho$  and accuracy level  $\varepsilon$ , find an  $\varepsilon$ -optimal policy  $\hat{\pi}$  based on  $\mathcal{D}$  obeying

$$V^*(\rho) - V^{\hat{\pi}}(\rho) = \mathbb{E}_{s \sim \rho} [V^*(s)] - \mathbb{E}_{s \sim \rho} [V^{\hat{\pi}}(s)] \leq \varepsilon$$

— *in a sample-efficient manner*

# Challenges of offline RL

---

- **Distribution shift:**

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

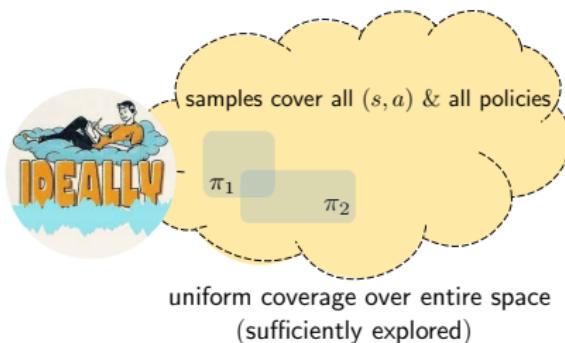
# Challenges of offline RL

---

- **Distribution shift:**

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

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

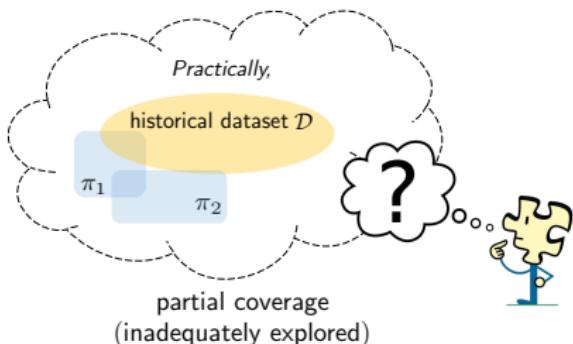
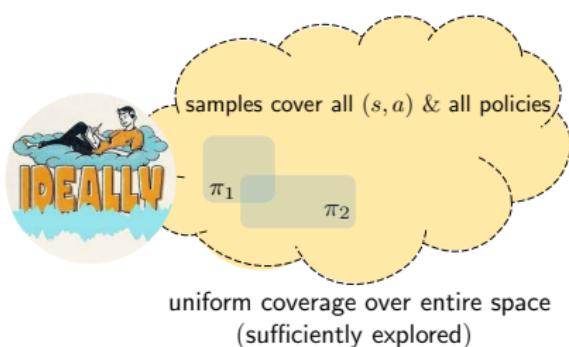


# Challenges of offline RL

- **Distribution shift:**

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

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



*How to quantify quality of historical dataset  $\mathcal{D}$  (induced by  $\pi^b$ )?*

*How to quantify quality of historical dataset  $\mathcal{D}$  (induced by  $\pi^b$ )?*

### Single-policy concentrability coefficient (Rashidinejad et al. '21)

$$C^* := \max_{s,a} \frac{d^{\pi^*}(s,a)}{d^{\pi^b}(s,a)}$$

where  $d^\pi(s,a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}((s^t, a^t) = (s, a) | \pi)$

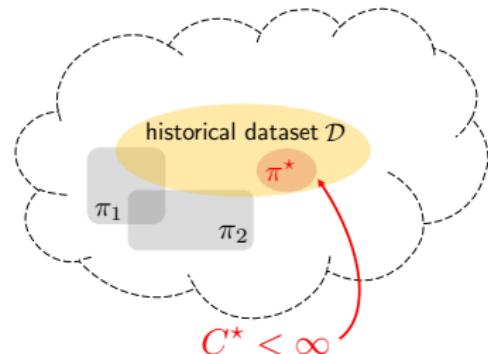
How to quantify quality of historical dataset  $\mathcal{D}$  (induced by  $\pi^b$ )?

### Single-policy concentrability coefficient (Rashidinejad et al. '21)

$$C^* := \max_{s,a} \frac{d^{\pi^*}(s,a)}{d^{\pi^b}(s,a)} = \left\| \frac{\text{occupancy density of } \pi^*}{\text{occupancy density of } \pi^b} \right\|_\infty \geq 1$$

where  $d^\pi(s,a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}((s^t, a^t) = (s, a) | \pi)$

- captures distributional shift
- allows for partial coverage



## A model-based offline algorithm: VI-LCB

---

**Pessimism in the face of uncertainty:** penalize value estimate of those  $(s, a)$  pairs that were poorly visited

# A model-based offline algorithm: VI-LCB

---

**Pessimism in the face of uncertainty:** penalize value estimate of those  $(s, a)$  pairs that were poorly visited

**Algorithm:** value iteration w/ lower confidence bounds

- compute empirical estimate  $\hat{P}$  of  $P$
- initialize  $\hat{Q} = 0$ , and repeat

$$\hat{Q}(s, a) \leftarrow \max \left\{ r(s, a) + \gamma \langle \hat{P}(\cdot | s, a), \hat{V} \rangle - \underbrace{b(s, a; \hat{V})}_{\text{Bernstein-style confidence bound}}, 0 \right\}$$

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

# Minimax optimality of model-based offline RL

## Theorem 3 (Li, Shi, Chen, Chi, Wei '22)

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

$$V^*(\rho) - V^{\hat{\pi}}(\rho) \leq \varepsilon$$

with high prob., with sample complexity at most

$$\tilde{O} \left( \frac{SC^*}{(1-\gamma)^3 \varepsilon^2} \right)$$

# Minimax optimality of model-based offline RL

## Theorem 3 (Li, Shi, Chen, Chi, Wei '22)

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

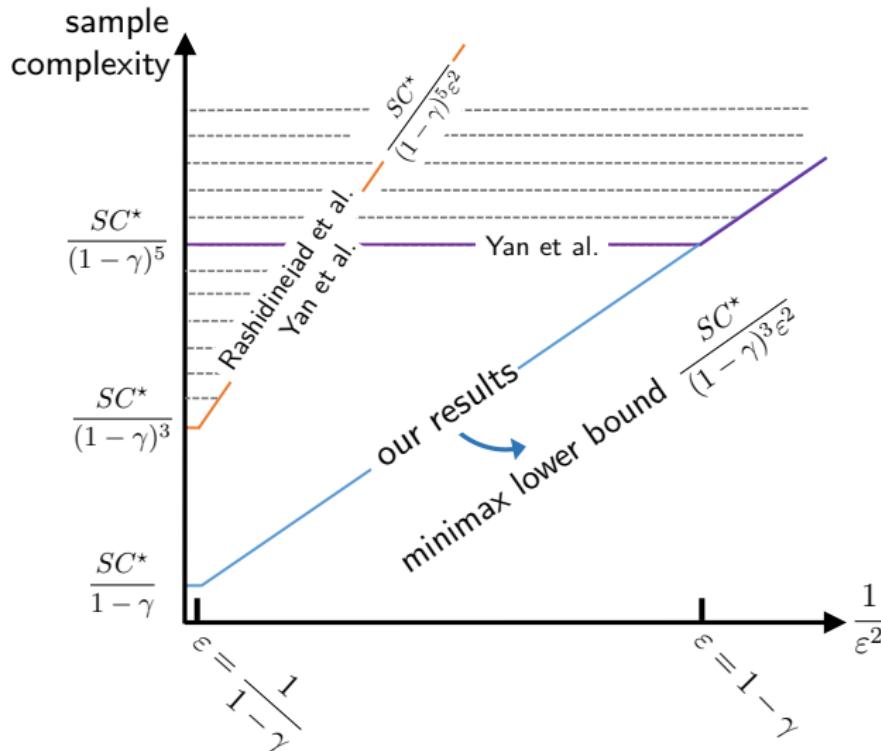
$$V^*(\rho) - V^{\hat{\pi}}(\rho) \leq \varepsilon$$

with high prob., with sample complexity at most

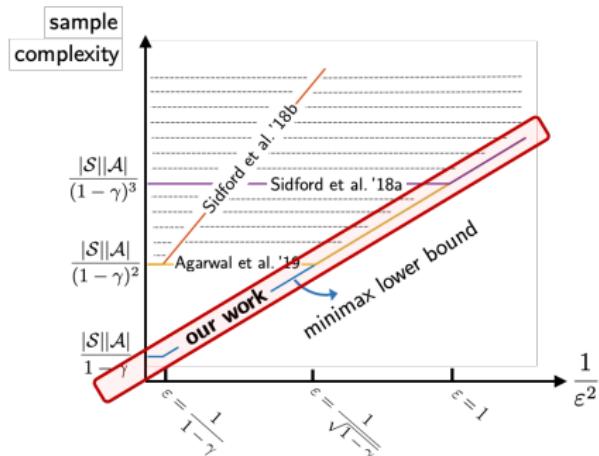
$$\tilde{O}\left(\frac{SC^*}{(1-\gamma)^3\varepsilon^2}\right)$$

- matches minimax lower bound:  $\tilde{\Omega}\left(\frac{SC^*}{(1-\gamma)^3\varepsilon^2}\right)$  (Rashidinejad et al. '21)
- depends on distribution shift (as reflected by  $C^*$ )
- full  $\varepsilon$ -range (no burn-in cost)

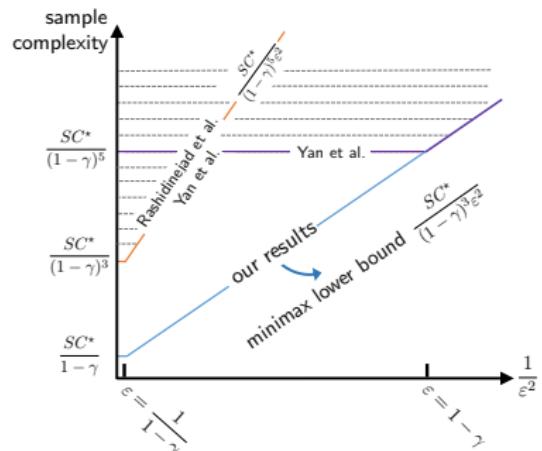
# Comparisons with prior art



# Summary of this part



generative model



offline RL

Model-based RL is minimax optimal with no burn-in cost!

# Reference I

---

- “*Reinforcement Learning: An Introduction*,” R. Sutton, A. Barto, 2018.
- “*Reinforcement Learning: Theory and Algorithms*,” A. Agarwal, N. Jiang, S. Kakade, W. Sun, in preparation.
- “*Dynamic programming and optimal control (4th edition)*,” D. Bertsekas, 2017.
- “*Finite-sample convergence rates for Q-learning and indirect algorithms*,” M. Kearns, S. Singh *NeurIPS*, 1998.
- “*Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model*,” M. Azar, R. Munos, H. J. Kappen, *Machine Learning*, vol. 91, no. 3, 2013.
- “*Near-optimal time and sample complexities for solving Markov decision processes with a generative model*,” A. Sidford, M. Wang, X. Wu, L. Yang, Y. Ye, *NeurIPS*, 2018.

## Reference II

---

- “*Model-based reinforcement learning with a generative model is minimax optimal,*” A. Agarwal, S. Kakade, L. F. Yang, *COLT*, 2020.
- “*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.
- “*Offline reinforcement learning: Tutorial, review, and perspectives on open problems,*” S. Levine, A. Kumar, G. Tucker, J. Fu, arXiv:2005.01643, 2020.
- “*Is pessimism provably efficient for offline RL?*” Y. Jin, Z. Yang, Z. Wang, *ICML*, 2021
- “*Bridging offline reinforcement learning and imitation learning: A tale of pessimism,*” P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, S. Russell, *NeurIPS*, 2021.

## Reference III

---

- “*Policy finetuning: Bridging sample-efficient offline and online reinforcement learning,*” T. Xie, N. Jiang, H. Wang, C. Xiong, Y. Bai, *NeurIPS*, 2021.
- “*Settling the sample complexity of model-based offline reinforcement learning,*” G. Li, L. Shi, Y. Chen, Y. Chi, Y. Wei, arXiv:2204.05275, 2022.