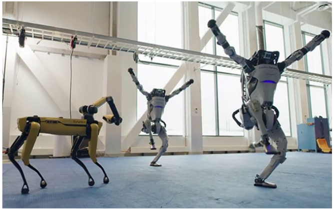
Reinforcement learning (Part 1): Basics and Model-based RL

Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

Successes of reinforcement learning (RL)



Supervised learning

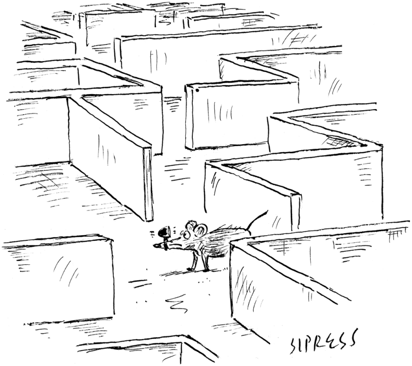
Given i.i.d. training data, the goal is to make prediction on unseen data:

— *pic from internet*

Reinforcement learning (RL)

In RL, an agent learns by interacting with an environment

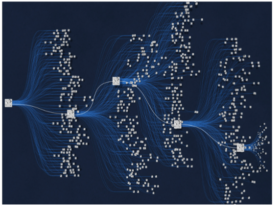
- no training data
- maximize total rewards
- trial-and-error
- sequential and online



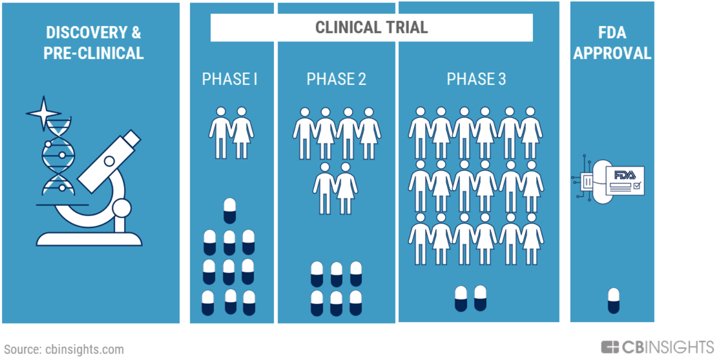
"Recalculating ... recalculating ..."

Challenges of RL

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

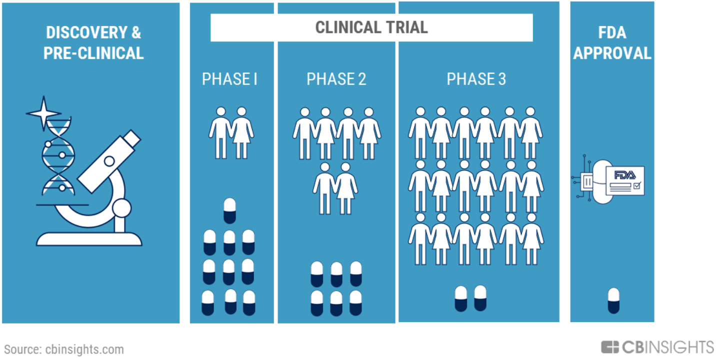


Sample efficiency



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

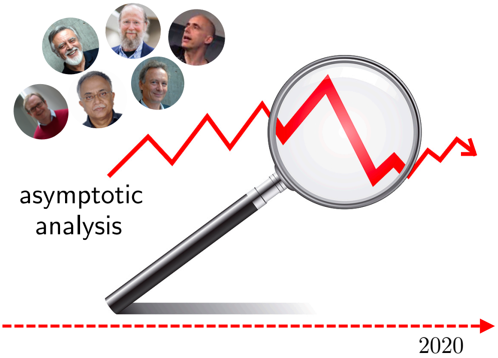
Sample efficiency



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

Challenge: how to design sample-efficient RL algorithms?

Statistical foundation of RL



Statistical Science
1986, Vol. 1, No. 2, 276–284

The Contributions of Herbert Robbins to Mathematical Statistics

Tze Leung Lai and David Siegmund

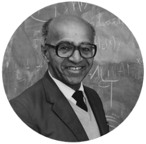
2. STOCHASTIC APPROXIMATION AND ADAPTIVE DESIGN

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 θ (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 μ_A, μ_B. How should we draw a

Herbert Robbins



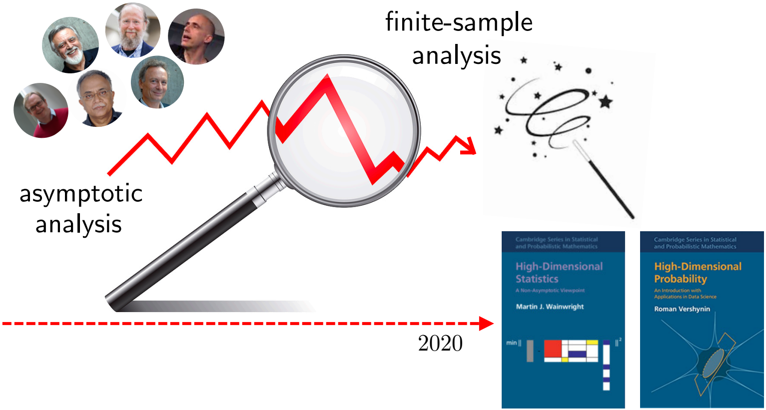
David Blackwell

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.

Statistical foundation of RL



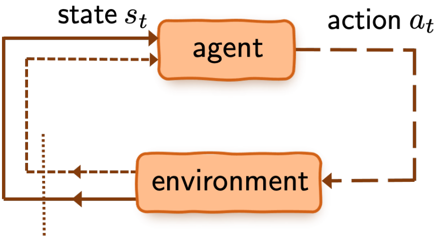
Understanding sample efficiency of RL requires a modern suite of non-asymptotic statistical tools

Outline (Part 1)

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

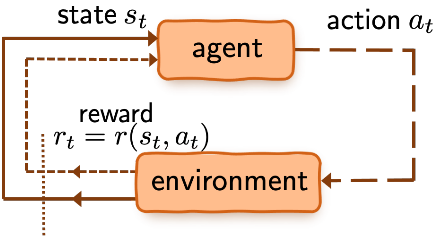
Background: Markov decision processes

Markov decision process (MDP)



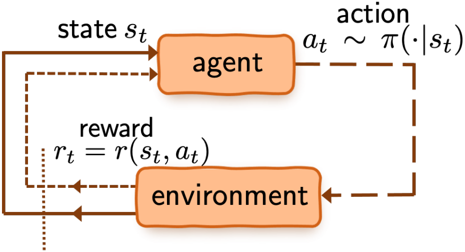
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

Markov decision process (MDP)



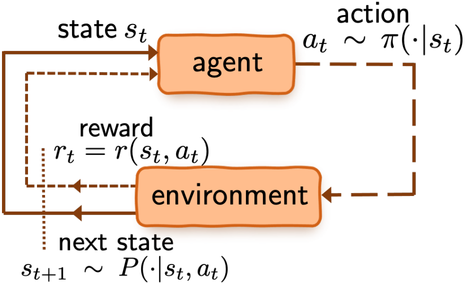
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

Markov decision process (MDP)



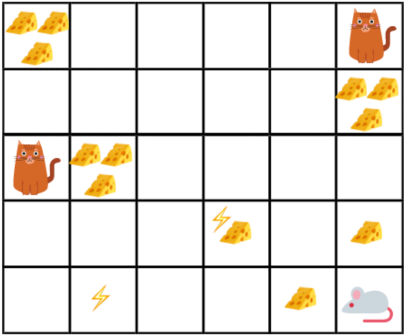
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

Markov decision process (MDP)

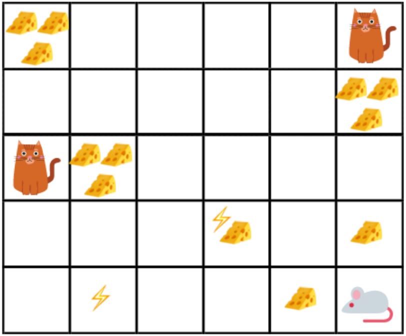


- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)
- $P(\cdot|s,a)$: unknown transition probabilities

Help the mouse!

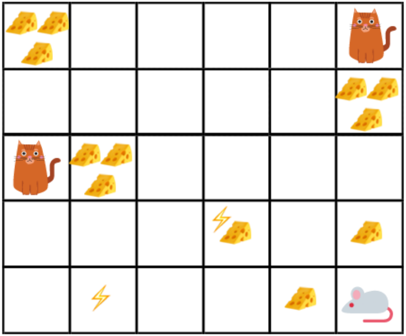


Help the mouse!



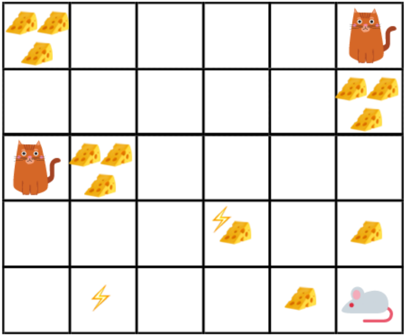
- state space $\mathcal{S}$: positions in the maze

Help the mouse!



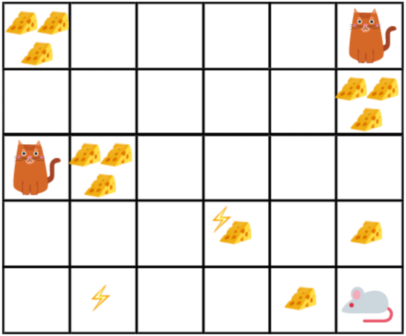
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right

Help the mouse!



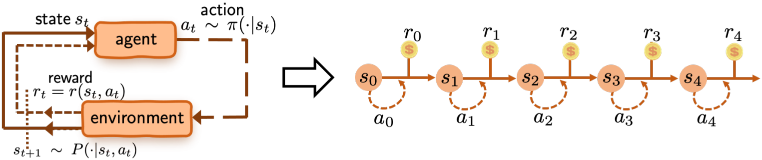
- state space $\mathcal{S}$: positions in the maze
- action space $\mathcal{A}$: up, down, left, right
- immediate reward r: cheese, electricity shocks, cats

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

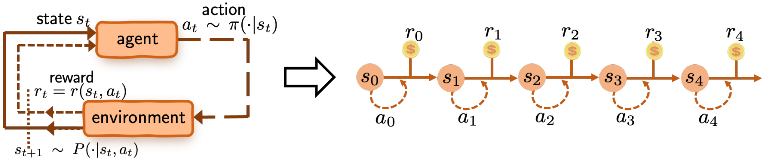
Value function



Value of policy 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

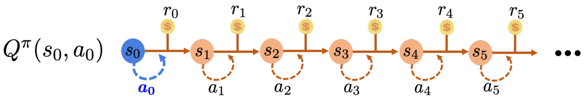


Value of policy π: cumulative discounted reward

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

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

Q-function (action-value function)

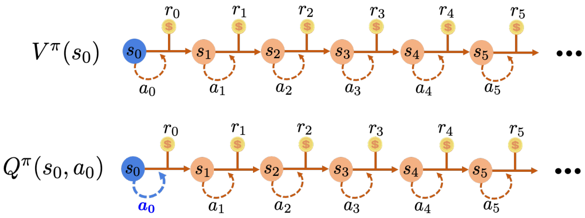


Q-function of policy π:

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

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

Q-function (action-value function)

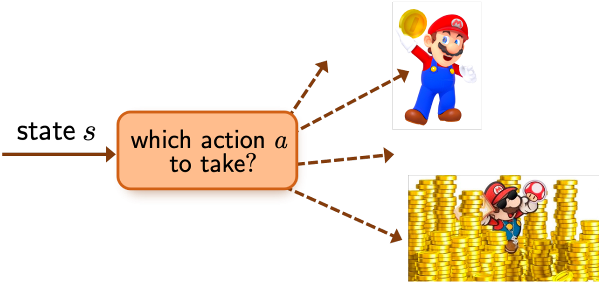


Q-function of policy π:

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

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

Optimal policy and optimal value



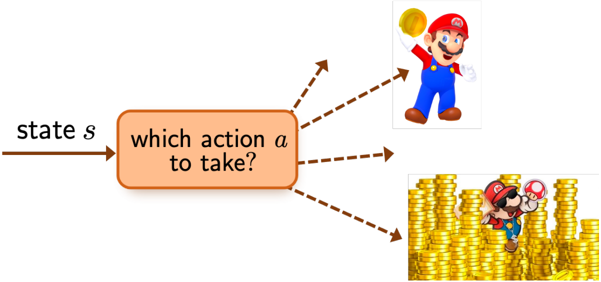
optimal policy $\pi^\star$: maximizing value function $\max_{\pi} V^{\pi}$

Theorem 1 (Puterman'94)

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

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

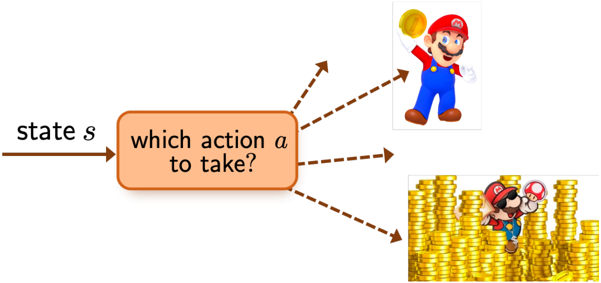
Optimal policy and optimal value



optimal policy $\pi^\star$: maximizing value function $\max_{\pi} V^{\pi}$

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

Optimal policy and optimal value

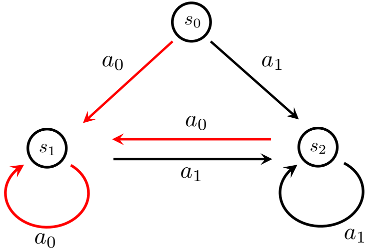


optimal policy $\pi^\star$: maximizing value function $\max_\pi V^\pi$

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

Example

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

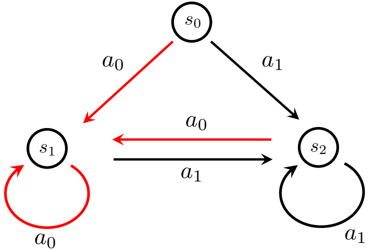


Reward: $r(s_1, a_0) = 1$, 0 else where

What is the optimal policy?

Example

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



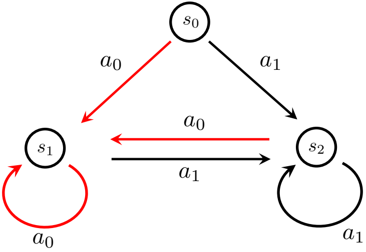
Reward: $r(s_1, a_0) = 1,$ 0 else where

What is the optimal policy?

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

Example

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



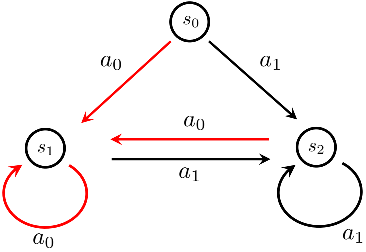
Reward: $r(s_1, a_0) = 1$, 0 else where

What is the optimal policy?

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

Example

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



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

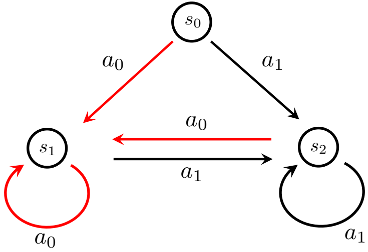
What is the optimal policy?

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

What is V^π 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 else where

What is the optimal policy?

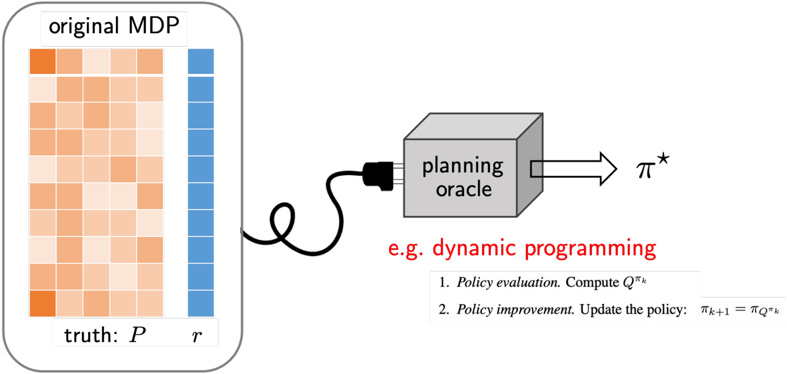
- $\pi^\star(s) = a_0, \ \forall s$
- $V^\star(s_0) = \frac{\gamma}{1-\gamma}$,
  $V^\star(s_1) = \frac{1}{1-\gamma}$, $V^\star(s_2) = \frac{\gamma}{1-\gamma}$

What is V^π 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 $\pi^\star$ 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 π? (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 π? (i.e., how to compute $V^\pi, \ \forall s$?)

Possible scheme:

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

Policy evaluation: Bellman's consistency equation

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

Policy evaluation: Bellman's consistency equation

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

Bellman's consistency equation

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

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

Richard Bellman

Policy evaluation: Bellman's consistency equation

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

Bellman's consistency equation

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

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

- one-step look-ahead

Richard Bellman

Policy evaluation: Bellman's consistency equation

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

Bellman's consistency equation

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

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

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

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

Richard Bellman

Bellman's optimality principle

Bellman operator

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

- one-step look-ahead

Bellman's optimality principle

Bellman operator

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

- one-step look-ahead

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

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

γ**-contraction of Bellman operator:**

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

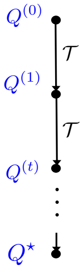
Richard Bellman

Value iteration (VI)

Value iteration (VI)

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

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

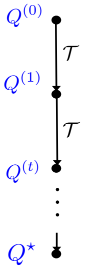


Value iteration (VI)

Value iteration (VI)

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

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



Iterative algorithm for fix-point solution:

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

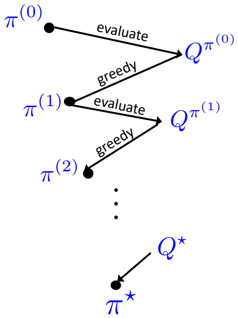
Policy iteration (PI)

Policy iteration (PI)

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

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

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



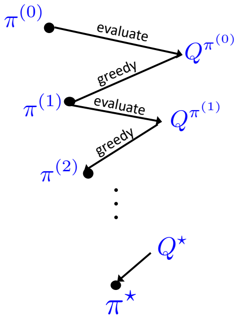
Policy iteration (PI)

Policy iteration (PI)

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

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

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



Monotonic improvement:

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

Iteration complexity

Theorem 1 (Linear convergence of policy/value iteration)

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

Iteration complexity

Theorem 1 (Linear convergence of policy/value iteration)

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

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

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

Iteration complexity

Theorem 1 (Linear convergence of policy/value iteration)

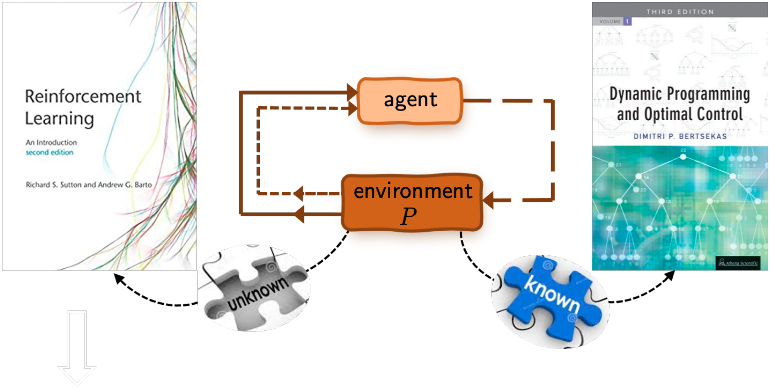
$$\|Q^{(t)} - Q^{\star}\|_{\infty} \leq \gamma^{t}\|Q^{(0)} - Q^{\star}\|_{\infty}$$

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

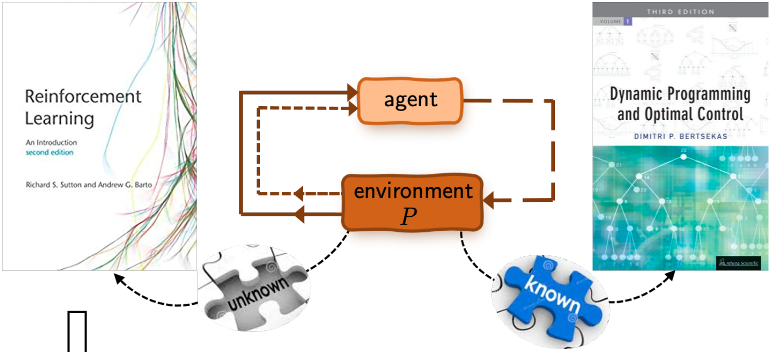
$$\frac{1}{1-\gamma}\log\left(\frac{\|Q^{(0)} - Q^{\star}\|_{\infty}}{\varepsilon}\right) \quad \text{iterations}$$

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

When the model is unknown . . .



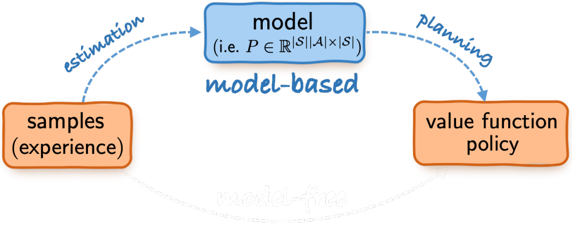
When the model is unknown . . .



$\Downarrow$

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

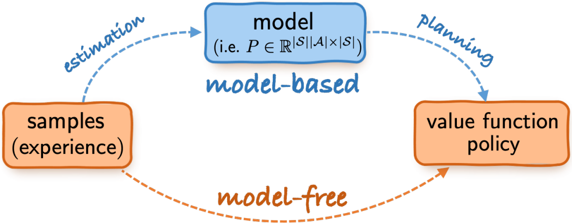
Two approaches



Model-based approach ("plug-in")

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

Two approaches



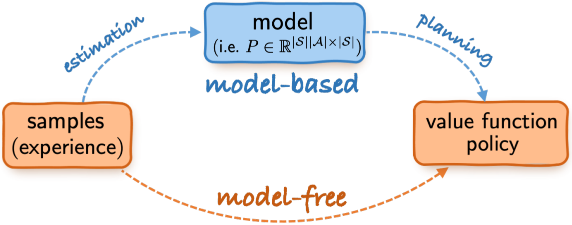
Model-based approach ("plug-in")

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

Model-free approach (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 $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

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

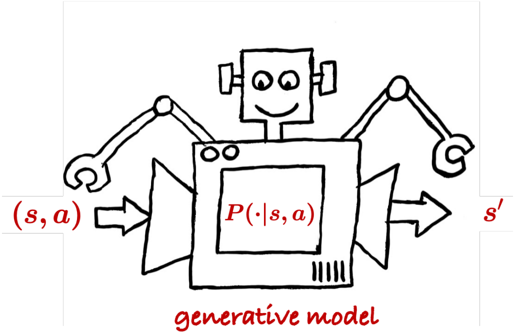
— learning w/o estimating the model explicitly

Model-based RL (a "plug-in" approach)

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

A generative model / simulator

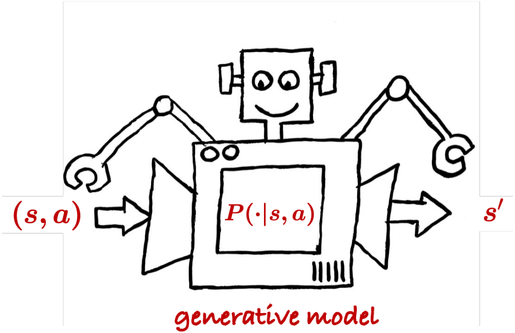
— *Kearns, Singh '99*



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

A generative model / simulator

— *Kearns, Singh '99*



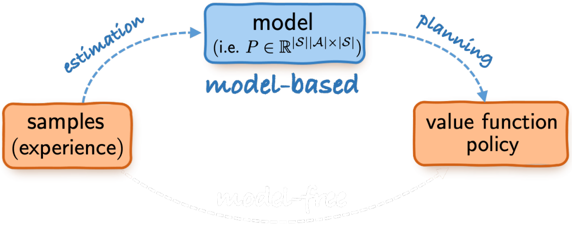
- **sampling:** for each (s, a), collect N samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$
- construct $\widehat{\pi}$ based on samples (in total $|\mathcal{S}||\mathcal{A}| \times N$)

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

An incomplete list of works

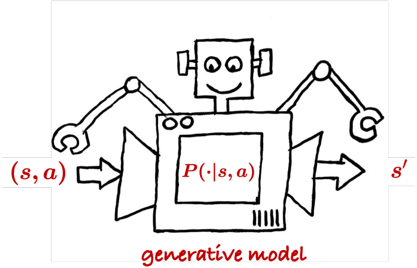
- Kearns & Singh '99
- Kakade '03
- Kearns, Mansour & Ng '02
- Azar, Munos & Kappen '12
- **Azar, Munos, Ghavamzadeh & Kappen '13**
- Sidford, Wang, Wu, Yang & Ye '18
- Sidford, Wang, Wu & Ye '18
- Wang '17
- **Agarwal, Kakade & Yang '19**
- Wainwright '19a
- Wainwright '19b
- Pananjady & Wainwright '20
- Yang & Wang '19
- Khamaru, Pananjady, Ruan, Wainwright & Jordan '20
- Mou, Li, Wainwright, Bartlett & Jordan '20
- **Li, Wei, Chi, Gu, Chen '20**
- Cui, Yang '21
- ...

Model-based approach ("plug-in")



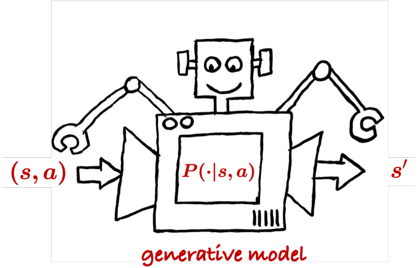
1. build an empirical estimate $\widehat{P}$ for P
2. planning based on empirical $\widehat{P}$

Model estimation



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

Model estimation



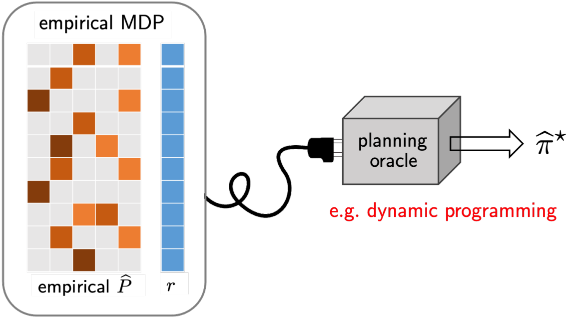
Sampling: for each (s,a), collect N ind. samples $\{(s,a,s'_{(i)})\}_{1\leq i\leq N}$

Empirical estimates:

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

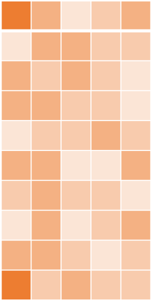
Empirical MDP + planning

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

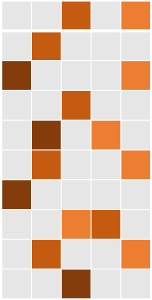


$\underbrace{\text{Find policy}}_{\text{using, e.g., policy iteration}}$ based on the $\underbrace{\text{empirical MDP}}_{(\widehat{P}, r)}$ (*empirical maximizer*)

Challenges in the sample-starved regime



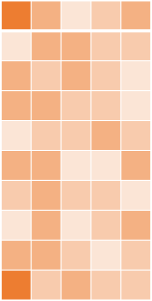
truth:
$P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



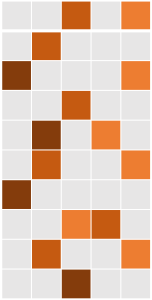
empirical estimate:
$\widehat{P}$

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

Challenges in the sample-starved regime



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



empirical estimate:
$\widehat{P}$

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

ℓ_∞-based sample complexity

Theorem 2 (Agarwal, Kakade, Yang '19)

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

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

with high prob., with sample complexity at most

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

ℓ_∞-based sample complexity

Theorem 2 (Agarwal, Kakade, Yang '19)

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

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

with high prob., with sample complexity at most

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

- matches minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ when $\varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$ (equivalently, when sample size exceeds $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$) (Azar et al. '13)

ℓ_∞-based sample complexity

Theorem 2 (Agarwal, Kakade, Yang '19)

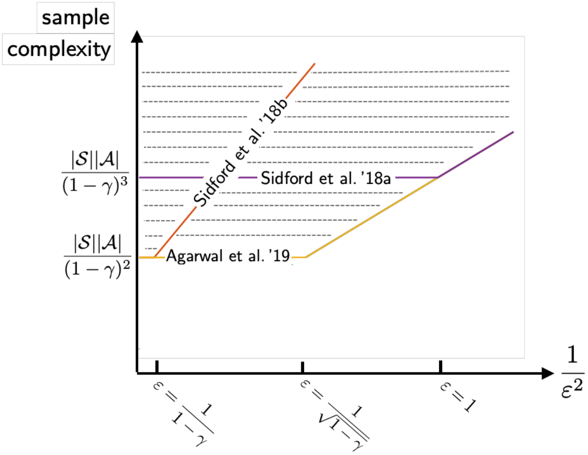
For any $0 < \varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$, *the optimal policy* $\widehat{\pi}^\star$ *of empirical MDP achieves*

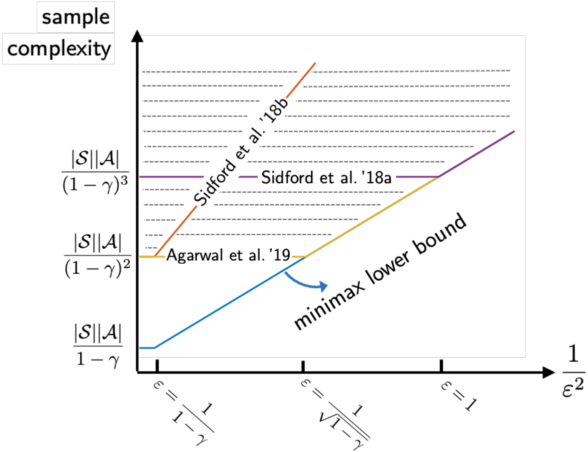
$$\|V^{\widehat{\pi}^\star} - V^\star\|_\infty \leq \varepsilon$$

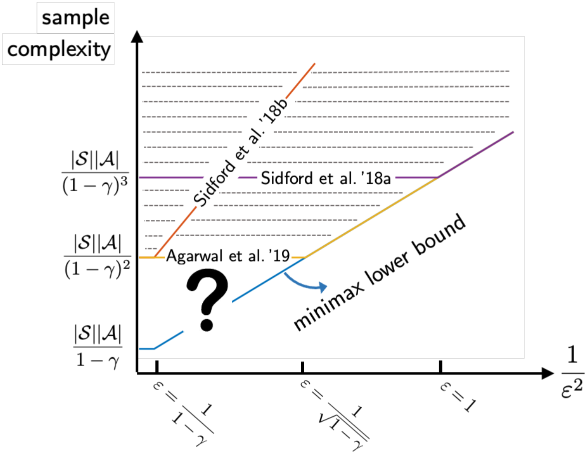
with high prob., with sample complexity at most

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

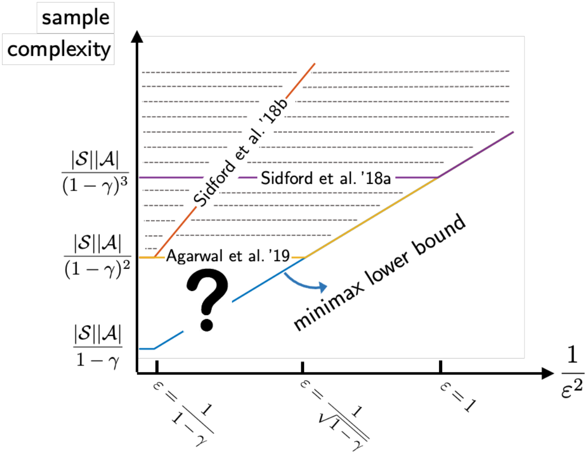
- matches minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ when $\varepsilon \leq \frac{1}{\sqrt{1-\gamma}}$ (equivalently, when sample size exceeds $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$) (Azar et al. '13)
- established upon leave-one-out analysis framework


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


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



Agarwal et al. '19 still requires a burn-in sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

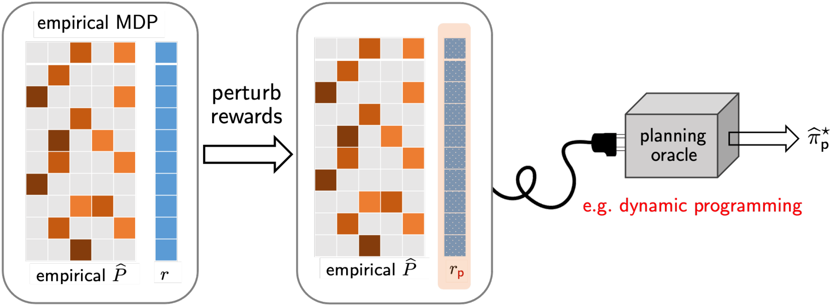


Agarwal et al. '19 still requires a burn-in sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

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

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

— *Li, Wei, Chi, Gu, Chen '20*



Find policy based on the empirical MDP with slightly perturbed rewards

Optimal ℓ_∞-based sample complexity

Theorem 3 (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 high prob., with sample complexity at most

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

Optimal ℓ_∞-based sample complexity

Theorem 3 (Li, Wei, Chi, Gu, Chen '20)

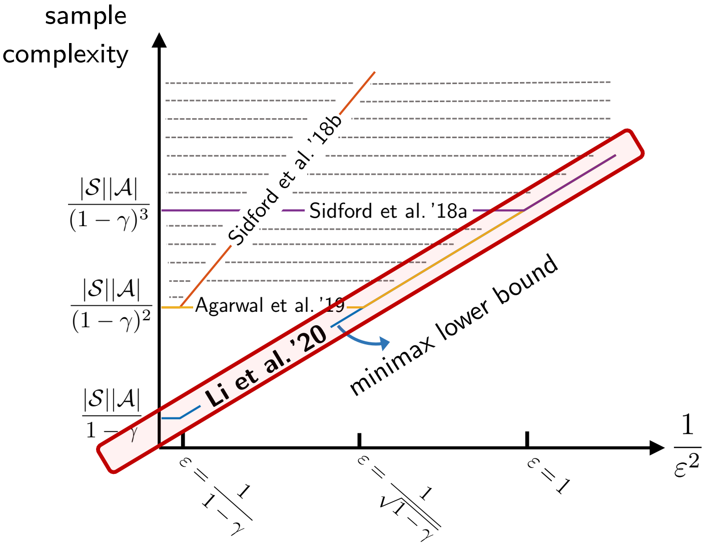
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, the optimal policy $\widehat{\pi}_{\mathrm{p}}^{\star}$ of perturbed empirical MDP achieves

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

with high prob., with sample complexity at most

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

- matches minimax lower bound: $\widetilde{\Omega}(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2})$ (Azar et al. '13)
- full ε-range: $\varepsilon \in (0, \frac{1}{1-\gamma}]$ $\longrightarrow$ no burn-in cost
- established upon more refined leave-one-analysis analysis and a perturbation argument


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

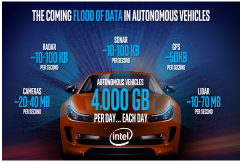
Model-based RL (a "plug-in" approach)

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
- But we have already stored tons of historical data

medical records



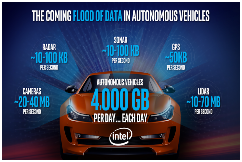
data of self-driving

clicking times of ads

Offline RL / Batch RL

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

medical records



data of self-driving

clicking times of ads

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

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

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

Offline RL / Batch RL

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

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

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

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

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

— *in a sample-efficient manner*

Challenges of offline RL

- **Distribution shift**:

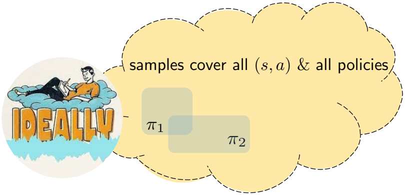
$$\text{distribution}(\mathcal{D}) \;\neq\; \text{target distribution under } \pi^{\star}$$

Challenges of offline RL

- **Distribution shift**:

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

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



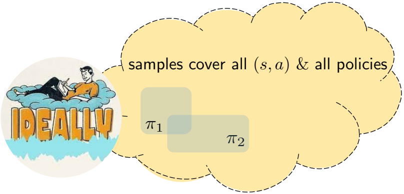
uniform coverage over entire space
(sufficiently explored)

Challenges of offline RL

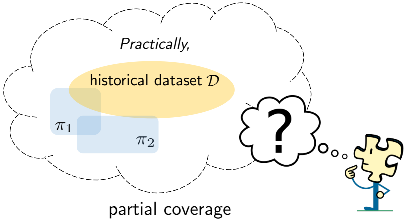
- **Distribution shift**:

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

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



uniform coverage over entire space
(sufficiently explored)



partial coverage
(inadequately explored)

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

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

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

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

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

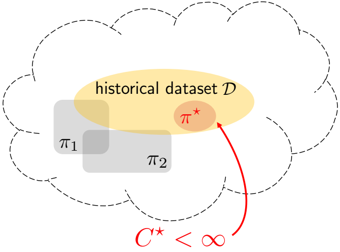
How to quantify quality of historical dataset $\mathcal{D}$ (induced by π^{b})?

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

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

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

- 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 $\widehat{P}$ of P
- initialize $\widehat{Q} = 0$, and repeat

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

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

Minimax optimality of model-based offline RL

Theorem 4 (Li, Shi, Chen, Chi, Wei '22)

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

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

with high prob., with sample complexity at most

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

Minimax optimality of model-based offline RL

Theorem 4 (Li, Shi, Chen, Chi, Wei '22)

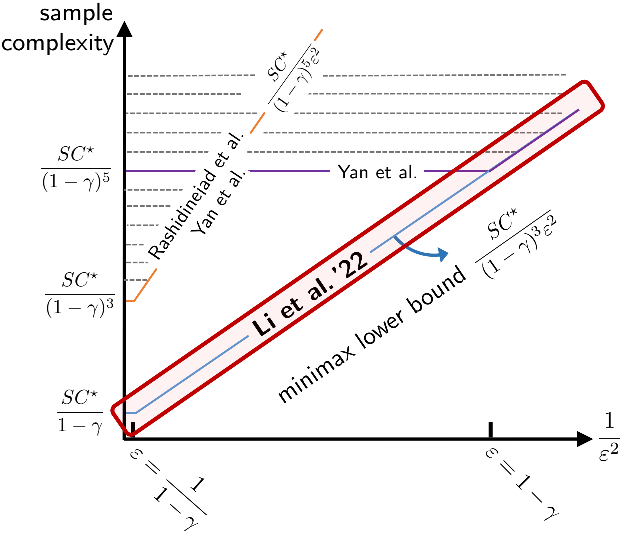
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$*, the policy* $\widehat{\pi}$ *returned by VI-LCB achieves*

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

with high prob., with sample complexity at most

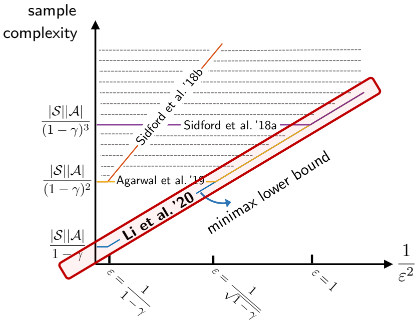
$$\widetilde{O}\left(\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}\right)$$

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

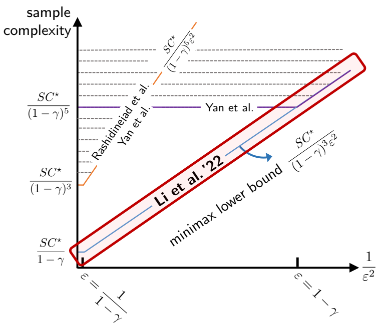


sample complexity
$\frac{SC^\star}{(1-\gamma)^5}$
$\frac{SC^\star}{(1-\gamma)^3}$
$\frac{SC^\star}{1-\gamma}$
$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$
Rashidinejad et al.
Yan et al.
Yan et al.
Li et al. '22
minimax lower bound
$\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}$
$\frac{1}{\varepsilon^2}$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

Summary of this part



generative model



offline/batch 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.