

Statistical and Algorithmic Foundations of Reinforcement Learning

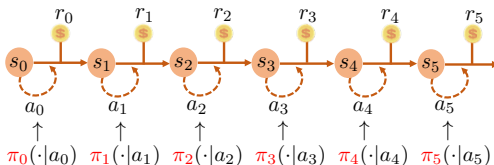
Yuejie Chi
Yale University

DeepLearn 2026
Tutorial Part 2

Part 2

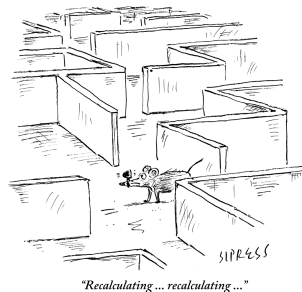
1. Online RL
2. Offline RL
3. Multi-agent RL
4. Robust RL

Online RL: interacting with real environment



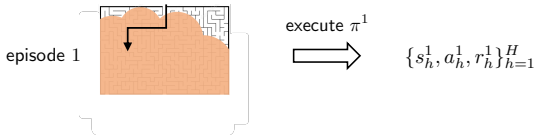
exploration via adaptive policies

- trial-and-error
- sequential and online
- adaptive learning from data



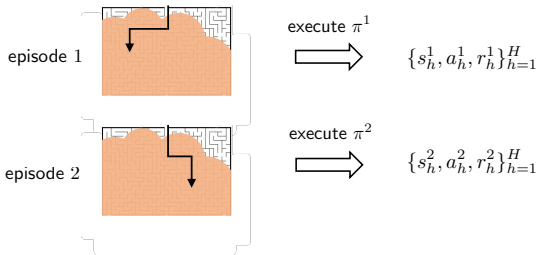
Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



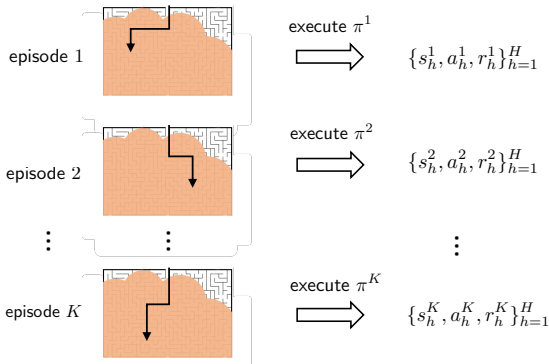
Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



Online episodic RL

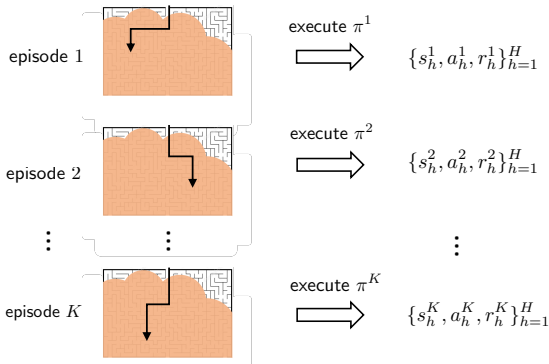
Sequentially execute MDP for K episodes, each consisting of H steps



Online episodic RL

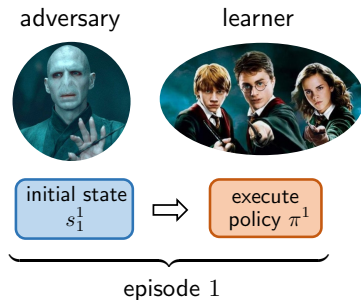
Sequentially execute MDP for K episodes, each consisting of H steps

— *sample size: $T = KH$*

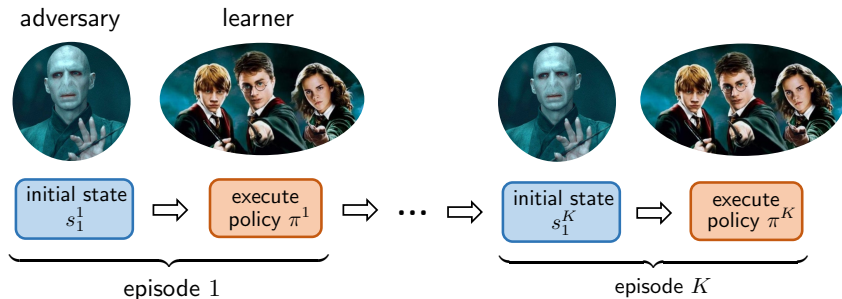


exploration (exploring unknowns) vs. **exploitation** (exploiting learned info)

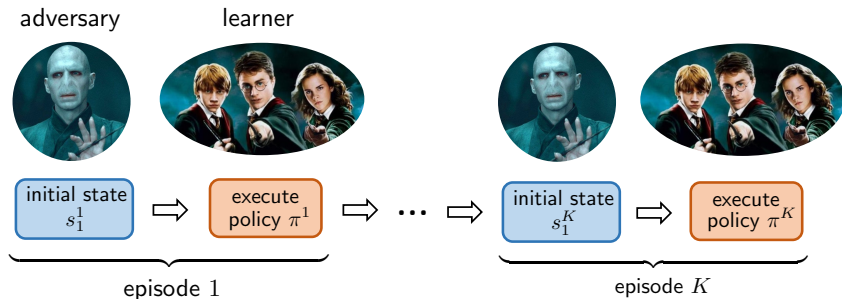
Regret: gap between learned policy & optimal policy



Regret: gap between learned policy & optimal policy



Regret: gap between learned policy & optimal policy



Performance metric: given initial states $\{s_1^k\}_{k=1}^K$, define

$$\text{Regret}(T) \quad := \quad \sum_{k=1}^K \left(V_1^{\star}(s_1^k) - V_1^{\pi^k}(s_1^k) \right)$$

Lower bound

([Domingues et al, 2021](#))

$$\text{Regret}(T) \gtrsim \sqrt{H^2 SAT}$$

Existing algorithms

- UCB-VI: [Azar et al, 2017](#)
- UBEV: [Dann et al, 2017](#)
- UCB-Q-Hoeffding: [Jin et al, 2018](#)
- UCB-Q-Bernstein: [Jin et al, 2018](#)
- UCB2-Q-Bernstein: [Bai et al, 2019](#)
- EULER: [Zanette et al, 2019](#)
- UCB-Q-Advantage: [Zhang et al, 2020](#)
- MVP: [Zhang et al, 2020](#)
- UCB-M-Q: [Menard et al, 2021](#)
- Q-EarlySettled-Advantage: [Li et al, 2021](#)
- (modified) MVP: [Zhang et al, 2024](#)

Lower bound

([Domingues et al, 2021](#))

$$\text{Regret}(T) \gtrsim \sqrt{H^2 SAT}$$

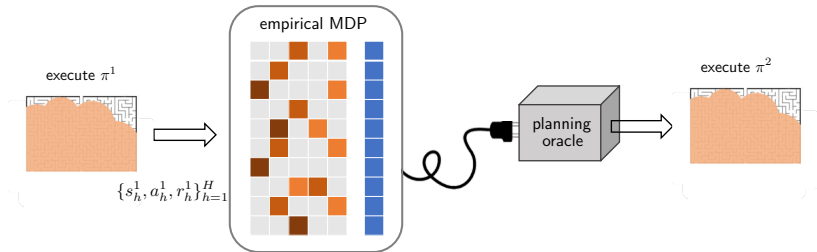
Existing algorithms

- UCB-VI: [Azar et al, 2017](#)
- UBEV: [Dann et al, 2017](#)
- UCB-Q-Hoeffding: [Jin et al, 2018](#)
- UCB-Q-Bernstein: [Jin et al, 2018](#)
- UCB2-Q-Bernstein: [Bai et al, 2019](#)
- EULER: [Zanette et al, 2019](#)
- UCB-Q-Advantage: [Zhang et al, 2020](#)
- MVP: [Zhang et al, 2020](#)
- UCB-M-Q: [Menard et al, 2021](#)
- Q-EarlySettled-Advantage: [Li et al, 2021](#)
- (modified) MVP: [Zhang et al, 2024](#)

Which online RL algorithms achieve near-minimal regret?

Model-based online RL with UCB exploration

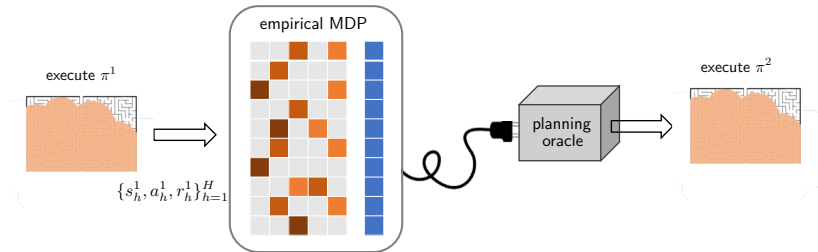
Model-based approach for online RL



repeat:

- use collected data to estimate transition probabilities
- apply planning to the estimated model to derive a new policy for sampling in the next episode

Model-based approach for online RL



repeat:

- use collected data to estimate transition probabilities
- apply planning to the estimated model to derive a new policy for sampling in the next episode

How to balance exploration and exploitation in this framework?



T. L. Lai



H. Robbins

Optimism in the face of uncertainty:

- explores based on the best optimistic estimates associated with the actions!
- a common framework: utilize upper confidence bounds (UCB)
accounts for estimates + uncertainty level



T. L. Lai



H. Robbins

Optimism in the face of uncertainty:

- explores based on the best optimistic estimates associated with the actions!
- a common framework: utilize upper confidence bounds (UCB)
accounts for estimates + uncertainty level

Optimistic model-based approach: incorporates **UCB** framework into model-based approach

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H - 1, \dots, 1$: run **value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\hat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H - 1, \dots, 1$: run **optimistic value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\hat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1} + \underbrace{b_h(s_h, a_h)}_{\text{bonus (upper confidence width)}}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H - 1, \dots, 1$: run **optimistic value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\hat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1} + \underbrace{b_h(s_h, a_h)}_{\text{bonus (upper confidence width)}}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

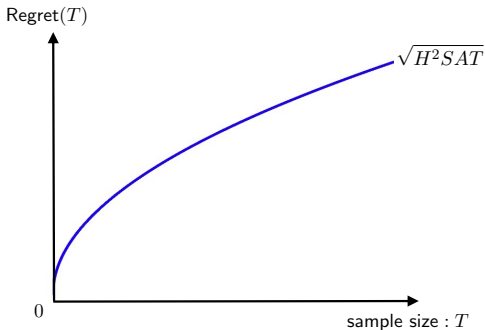
2. Forward $h = 1, \dots, H$: take actions according to **greedy policy**

$$\pi_h(s) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h(s, a)$$

to sample a new episode $\{s_h, a_h, r_h\}_{h=1}^H$

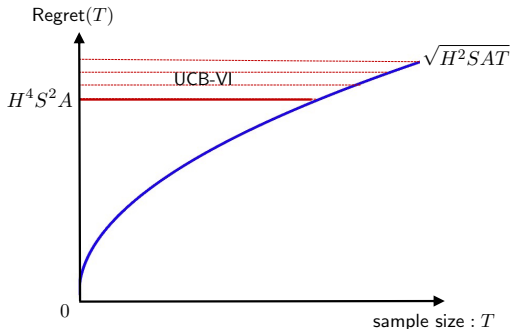
UCB-VI is asymptotically regret-optimal

— Azar, Osband, Munos, 2017



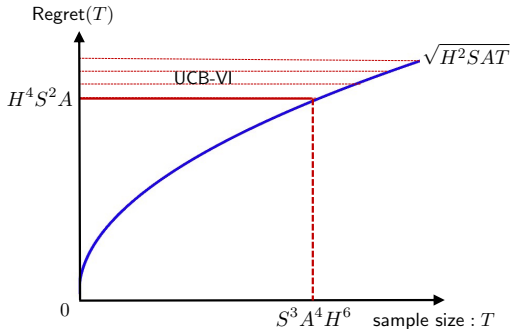
UCB-VI is asymptotically regret-optimal

— Azar, Osband, Munos, 2017



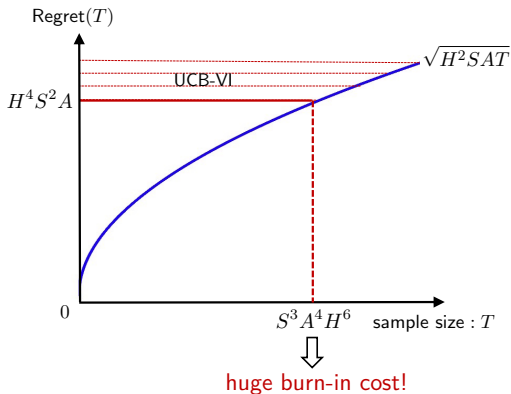
UCB-VI is asymptotically regret-optimal

— Azar, Osband, Munos, 2017



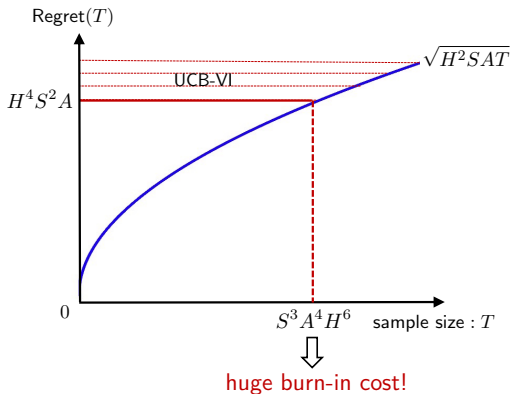
UCB-VI is asymptotically regret-optimal

— Azar, Osband, Munos, 2017



UCB-VI is asymptotically regret-optimal

— Azar, Osband, Munos, 2017



Issues: large burn-in cost

Other asymptotically regret-optimal algorithms

Algorithm	Regret upper bound	Range of K that attains optimal regret
UCBVI (Azar et al, 2017)	$\sqrt{SAH^2T} + S^2AH^3$	$[S^3AH^3, \infty)$
ORLC (Dann et al, 2019)	$\sqrt{SAH^2T} + S^2AH^4$	$[S^3AH^5, \infty)$
EULER (Zanette et al, 2019)	$\sqrt{SAH^2T} + S^{3/2}AH^3(\sqrt{S} + \sqrt{H})$	$[S^2AH^3(\sqrt{S} + \sqrt{H}), \infty)$
UCB-Adv (Zhang et al, 2020)	$\sqrt{SAH^2T} + S^2A^{3/2}H^{33/4}K^{1/4}$	$[S^6A^4H^{27}, \infty)$
MVP (Zhang et al, 2020)	$\sqrt{SAH^2T} + S^2AH^2$	$[S^3AH, \infty)$
UCB-M-Q (Menard et al, 2021)	$\sqrt{SAH^2T} + SAH^4$	$[SAH^5, \infty)$

Other asymptotically regret-optimal algorithms

Algorithm	Regret upper bound	Range of K that attains optimal regret
UCBVI (Azar et al, 2017)	$\sqrt{SAH^2T} + S^2AH^3$	$[S^3AH^3, \infty)$
ORLC (Dann et al, 2019)	$\sqrt{SAH^2T} + S^2AH^4$	$[S^3AH^5, \infty)$
EULER (Zanette et al, 2019)	$\sqrt{SAH^2T} + S^{3/2}AH^3(\sqrt{S} + \sqrt{H})$	$[S^2AH^3(\sqrt{S} + \sqrt{H}), \infty)$
UCB-Adv (Zhang et al, 2020)	$\sqrt{SAH^2T} + S^2A^{3/2}H^{33/4}K^{1/4}$	$[S^6A^4H^{27}, \infty)$
MVP (Zhang et al, 2020)	$\sqrt{SAH^2T} + S^2AH^2$	$[S^3AH, \infty)$
UCB-M-Q (Menard et al, 2021)	$\sqrt{SAH^2T} + SAH^4$	$[SAH^5, \infty)$

Can we find a regre-optimal algorithm with no burn-in cost?

Monotonic Value Propagation

UCB-VI with **doubling update rules** and **variance-aware bonus**

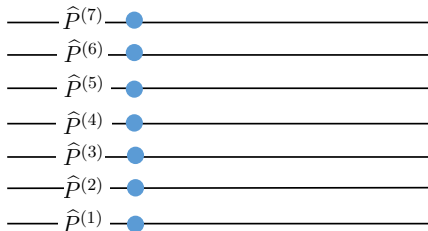
- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \dots\}$ -th time

Monotonic Value Propagation

UCB-VI with **doubling update rules** and **variance-aware bonus**

- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \dots\}$ -th time

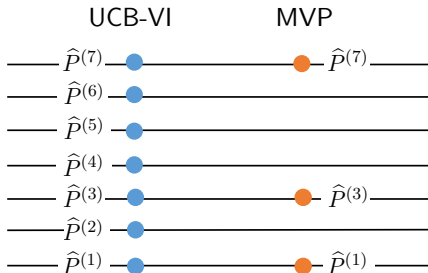
UCB-VI



Monotonic Value Propagation

UCB-VI with **doubling update rules** and **variance-aware bonus**

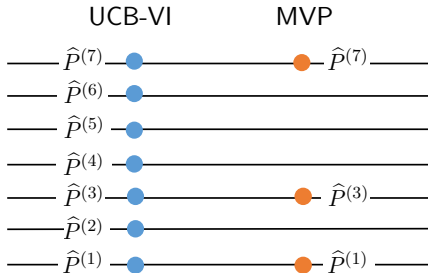
- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \dots\}$ -th time



Monotonic Value Propagation

UCB-VI with **doubling update rules** and **variance-aware bonus**

- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \dots\}$ -th time

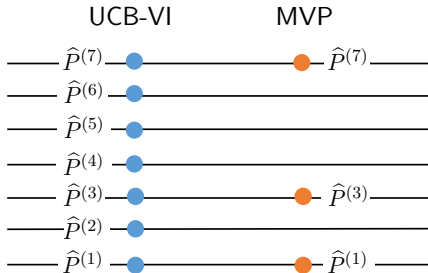


- visitation counts change much less frequently
→ reduces covering number dramatically

Monotonic Value Propagation

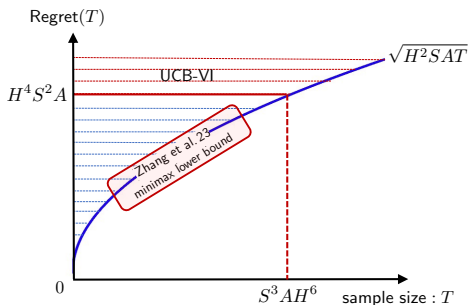
UCB-VI with **doubling update rules** and **variance-aware bonus**

- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \dots\}$ -th time



- visitation counts change much less frequently
→ reduces covering number dramatically
- data-driven bonus terms (chosen based on empirical variances)

Regret-optimal algorithm w/o burn-in cost

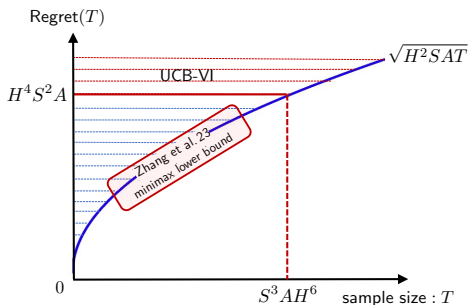


Theorem (Zhang, Chen, Lee, Du '24)

The model-based algorithm Monotonic Value Propagation achieves

$$\text{Regret}(T) \lesssim \tilde{O}(\sqrt{H^2 S A T})$$

Regret-optimal algorithm w/o burn-in cost



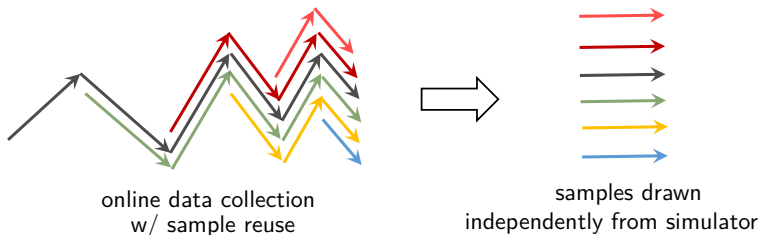
Theorem (Zhang, Chen, Lee, Du '24)

The model-based algorithm Monotonic Value Propagation achieves

$$\text{Regret}(T) \lesssim \tilde{O}(\sqrt{H^2 S A T})$$

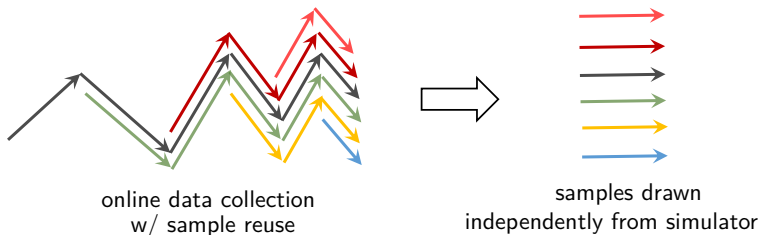
- the only algorithm so far that is regret-optimal w/o burn-ins

Key technical innovation



Decoupling complicated statistical dependency during online learning

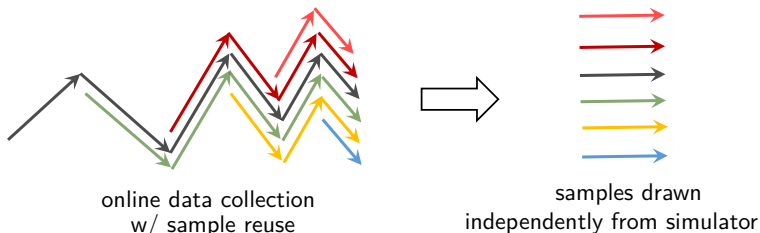
Key technical innovation



Decoupling complicated statistical dependency during online learning

- couples online data collection with i.i.d. sampling

Key technical innovation



Decoupling complicated statistical dependency during online learning

- couples online data collection with i.i.d. sampling
- exploit *compressibility* of visitation counts
 - w/ the aid of doubling algorithmic trick

How about memory complexity?

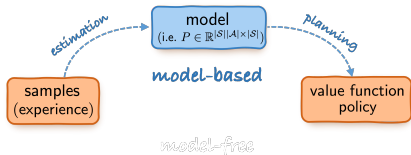
Algorithm	Regret upper bound	Range of K that attains optimal regret	Memory complexity
UCBVI (Azar et al, 2017)	$\sqrt{SAH^2T} + S^2AH^3$	$[S^3AH^3, \infty)$	S^2AH
UCB-Adv (Zhang et al, 2020)	$\sqrt{SAH^2T} + S^2A^{3/2}H^{33/4}K^{1/4}$	$[S^6A^4H^{27}, \infty)$	SAH
MVP (Zhang et al, 2020)	$\sqrt{SAH^2T} + S^2AH^2$	$[S^3AH, \infty)$	S^2AH
UCB-M-Q (Menard et al. '21)	$\sqrt{SAH^2T} + SAH^4$	$[SAH^5, \infty)$	S^2AH
MVP (Zhang et al, 2024)	$\sqrt{SAH^2T}$	$[1, \infty)$	S^2AH

How about memory complexity?

Algorithm	Regret upper bound	Range of K that attains optimal regret	Memory complexity
UCBVI (Azar et al, 2017)	$\sqrt{SAH^2T} + S^2AH^3$	$[S^3AH^3, \infty)$	S^2AH
UCB-Adv (Zhang et al, 2020)	$\sqrt{SAH^2T} + S^2A^{3/2}H^{33/4}K^{1/4}$	$[S^6A^4H^{27}, \infty)$	SAH
MVP (Zhang et al, 2020)	$\sqrt{SAH^2T} + S^2AH^2$	$[S^3AH, \infty)$	S^2AH
UCB-M-Q (Menard et al. '21)	$\sqrt{SAH^2T} + SAH^4$	$[SAH^5, \infty)$	S^2AH
MVP (Zhang et al, 2024)	$\sqrt{SAH^2T}$	$[1, \infty)$	S^2AH

Can we find a regret-optimal algorithm with
(1) low burn-in cost and (2) low memory complexity?

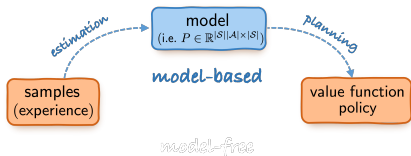
Model-free RL is often more memory-efficient



store transition kernel estimates

→ $O(S^2AH)$ memory

Model-free RL is often more memory-efficient

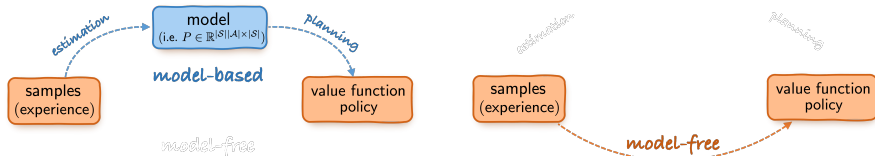


store transition kernel estimates
→ $O(S^2AH)$ memory



maintain Q -estimates
→ $O(SAH)$ memory

Model-free RL is often more memory-efficient



store transition kernel estimates
 $\rightarrow O(S^2AH)$ memory

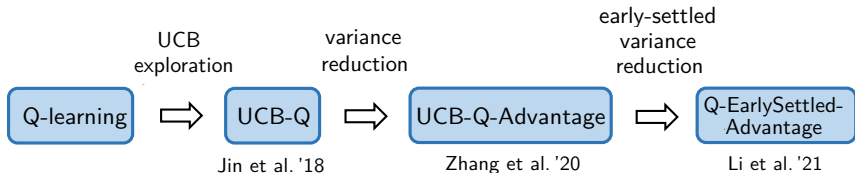
maintain Q -estimates
 $\rightarrow O(SAH)$ memory

Definition (Jin et al. '18)

An RL algorithm is **model-free** if its space complexity is $o(S^2AH)$

Which model-free algorithms are sample-efficient for online RL?

Which model-free algorithms are sample-efficient for online RL?



Q-learning: a classical model-free algorithm



Chris Watkins



Peter Dayan

Stochastic approximation for solving Bellman equation

$$Q_h(s_h, a_h) \longleftarrow (1 - \eta_k)Q_h(s_h, a_h) + \eta_k \mathcal{T}_k(Q_{h+1})(s_h, a_h)$$

Q-learning: a classical model-free algorithm



Chris Watkins



Peter Dayan

Stochastic approximation for solving Bellman equation

$$Q_h(s_h, a_h) \leftarrow (1 - \eta_k)Q_h(s_h, a_h) + \eta_k \mathcal{T}_k(Q_{h+1})(s_h, a_h)$$

$$\mathcal{T}_k(Q_h)(s_h, a_h) = r(s_h, a_h) + \max_{a'} Q(s_{h+1}, a')$$

using sample in k -th episode

Q-learning with UCB exploration (Jin et al., 2018)

$$Q_h(s_h, a_h) \leftarrow \underbrace{(1 - \eta_k)Q_h(s_h, a_h) + \eta_k \mathcal{T}_k(Q_{h+1})(s_h, a_h)}_{\text{classical Q-learning}} + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{exploration bonus}}$$

Q-learning with UCB exploration (Jin et al., 2018)

$$Q_h(s_h, a_h) \leftarrow \underbrace{(1 - \eta_k)Q_h(s_h, a_h) + \eta_k \mathcal{T}_k(Q_{h+1})(s_h, a_h)}_{\text{classical Q-learning}} + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{exploration bonus}}$$

- $b_h(s, a)$: upper confidence bound
— *optimism in the face of uncertainty*
- inspired by UCB bandit algorithm (Lai, Robbins '85)

Q-learning with UCB exploration (Jin et al., 2018)

$$Q_h(s_h, a_h) \leftarrow \underbrace{(1 - \eta_k)Q_h(s_h, a_h) + \eta_k \mathcal{T}_k(Q_{h+1})(s_h, a_h)}_{\text{classical Q-learning}} + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{exploration bonus}}$$

- $b_h(s, a)$: upper confidence bound
— *optimism in the face of uncertainty*
- inspired by UCB bandit algorithm (Lai, Robbins '85)

$$\text{Regret}(T) \lesssim \sqrt{H^3 SAT} \implies \text{sub-optimal by a factor of } \sqrt{H}$$

Q-learning with UCB exploration (Jin et al., 2018)

$$Q_h(s_h, a_h) \leftarrow \underbrace{(1 - \eta_k)Q_h(s_h, a_h) + \eta_k \mathcal{T}_k(Q_{h+1})(s_h, a_h)}_{\text{classical Q-learning}} + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{exploration bonus}}$$

- $b_h(s, a)$: upper confidence bound
— *optimism in the face of uncertainty*
- inspired by UCB bandit algorithm (Lai, Robbins '85)

$\text{Regret}(T) \lesssim \sqrt{H^3 SAT} \implies$ sub-optimal by a factor of $\sqrt{H}$

Issue: large variability in stochastic update rules

Q-learning with UCB and variance reduction

— *Zhang et al. '20*

Incorporates **variance reduction** into UCB-Q:

Q-learning with UCB and variance reduction

— *Zhang et al. '20*

Incorporates **reference-advantage decomposition** into UCB-Q:

Q-learning with UCB and variance reduction

— Zhang et al. '20

Incorporates **reference-advantage decomposition** into UCB-Q:

$$Q_h(s_h, a_h) \leftarrow (1 - \eta_k)Q_h(s_h, a_h) + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{UCB bonus}} + \eta_k \left(\underbrace{\mathcal{T}_k(Q_{h+1}) - \mathcal{T}_k(\bar{Q}_{h+1})}_{\text{advantage}} + \underbrace{\hat{\mathcal{T}}(\bar{Q}_{h+1})}_{\text{reference}} \right)(s_h, a_h)$$

- Reference $\bar{Q}_{h+1}$, batch estimate $\hat{\mathcal{T}}$: help reduce variability

Q-learning with UCB and variance reduction

— Zhang et al. '20

Incorporates **reference-advantage decomposition** into UCB-Q:

$$Q_h(s_h, a_h) \leftarrow (1 - \eta_k)Q_h(s_h, a_h) + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{UCB bonus}} + \eta_k \left(\underbrace{\mathcal{T}_k(Q_{h+1}) - \mathcal{T}_k(\bar{Q}_{h+1})}_{\text{advantage}} + \underbrace{\hat{\mathcal{T}}(\bar{Q}_{h+1})}_{\text{reference}} \right)(s_h, a_h)$$

- Reference $\bar{Q}_{h+1}$, batch estimate $\hat{\mathcal{T}}$: help reduce variability

UCB-Q-Advantage is asymptotically regret-optimal

Q-learning with UCB and variance reduction

— Zhang et al. '20

Incorporates **reference-advantage decomposition** into UCB-Q:

$$Q_h(s_h, a_h) \leftarrow (1 - \eta_k)Q_h(s_h, a_h) + \underbrace{\eta_k b_h(s_h, a_h)}_{\text{UCB bonus}} + \eta_k \left(\underbrace{\mathcal{T}_k(Q_{h+1}) - \mathcal{T}_k(\bar{Q}_{h+1})}_{\text{advantage}} + \underbrace{\hat{\mathcal{T}}(\bar{Q}_{h+1})}_{\text{reference}} \right)(s_h, a_h)$$

- Reference $\bar{Q}_{h+1}$, batch estimate $\hat{\mathcal{T}}$: help reduce variability

UCB-Q-Advantage is asymptotically regret-optimal

Issue: high burn-in cost $O(S^6 A^4 H^{28})$

Part 2

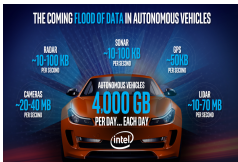
1. Online RL
2. Offline RL
3. Multi-agent RL
4. Robust RL

Offline/batch RL

- Collecting new data might be costly, unsafe, unethical, or time-consuming



medical records



data of self-driving



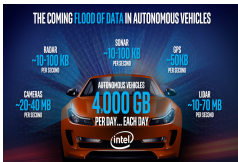
clicking times of ads

Offline/batch RL

- Collecting new data might be costly, unsafe, unethical, or time-consuming
- But we have already stored tons of historical data



medical records



data of self-driving



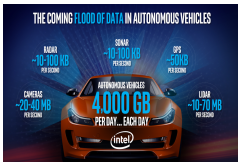
clicking times of ads

Offline/batch RL

- Collecting new data might be costly, unsafe, unethical, or time-consuming
- But we have already stored tons of historical data



medical records



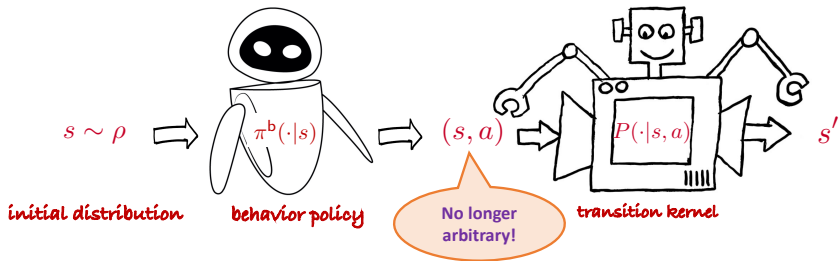
data of self-driving



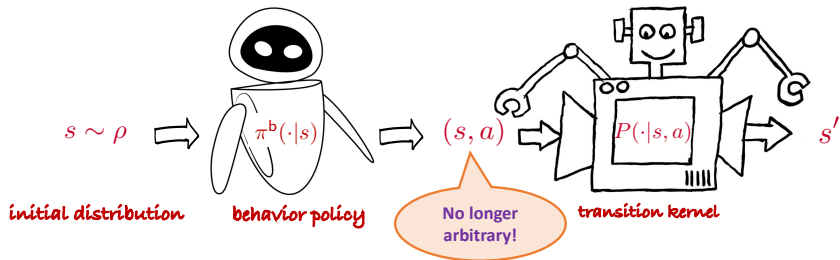
clicking times of ads

Question: can we learn based solely on historical data w/o active exploration?

A mathematical model of offline data



A mathematical model of offline data

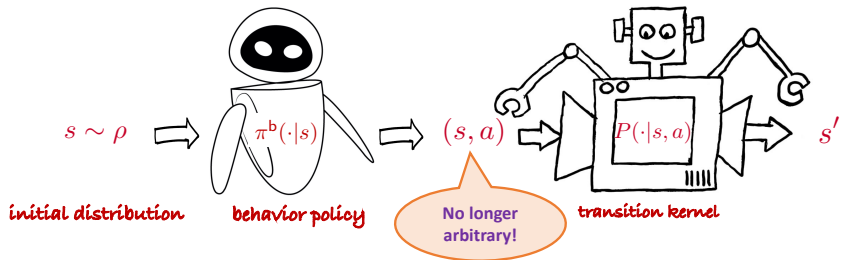


historical dataset $\mathcal{D} = \{(s^{(i)}, a^{(i)}, s'^{(i)})\}$: N independent copies of

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

- ρ : initial state distribution; π^b : behavior policy

A mathematical model of offline data



Goal: given a target accuracy level $\varepsilon \in (0, H]$, find $\hat{\pi}$ s.t.

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

— in a sample-efficient manner

Challenges of offline RL

- **Distribution shift:**

distribution($\mathcal{D}$) $\neq$ target distribution under optimal π^*

Challenges of offline RL

- **Distribution shift:**

distribution($\mathcal{D}$) $\neq$ target distribution under optimal π^*

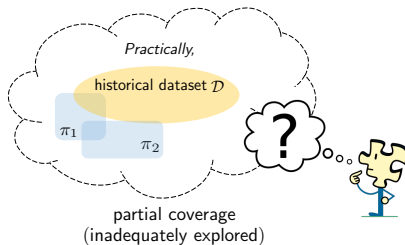
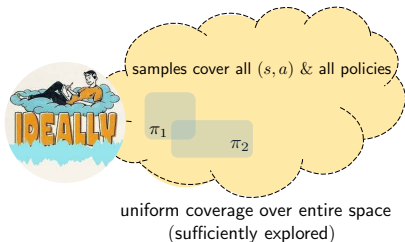


Challenges of offline RL

- **Distribution shift:**

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

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



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 (Rashidineiad et al. '21)

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

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

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

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

- captures distributional shift

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

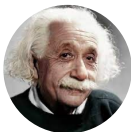
Single-policy concentrability coefficient (Rashidineiad et al. '21)

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

- captures distributional shift

$C^* = O(1)$

large C^*



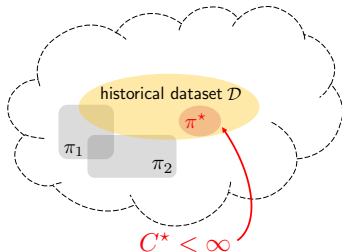
expert data

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

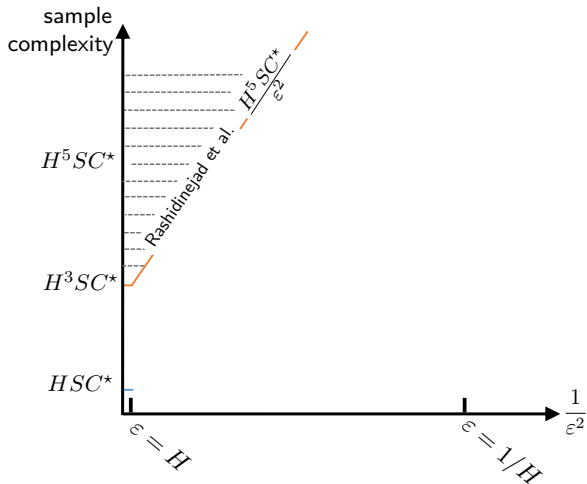
Single-policy concentrability coefficient (Rashidineiad et al. '21)

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

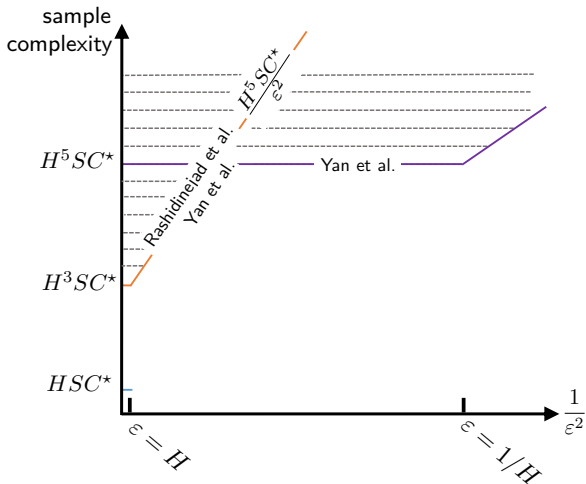
- captures distributional shift
- allows for partial coverage
 - as long as it covers the part reachable by $\pi^\star$



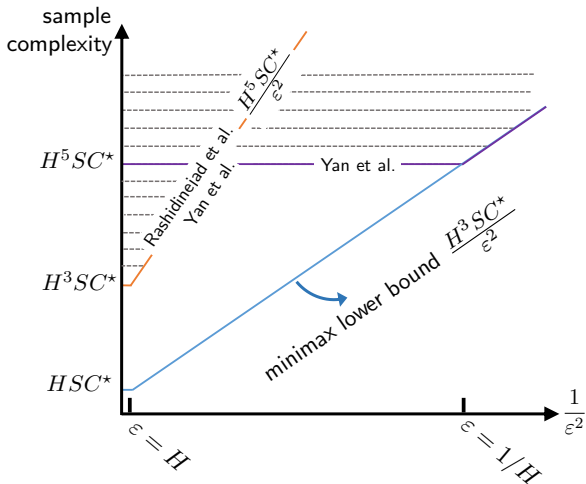
Prior art: sample complexity bounds



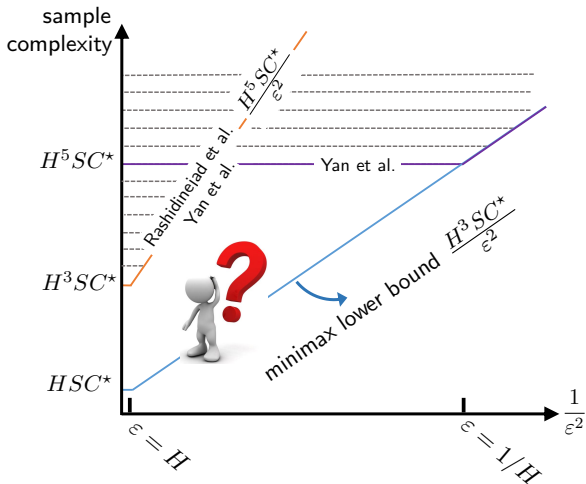
Prior art: sample complexity bounds



Prior art: sample complexity bounds

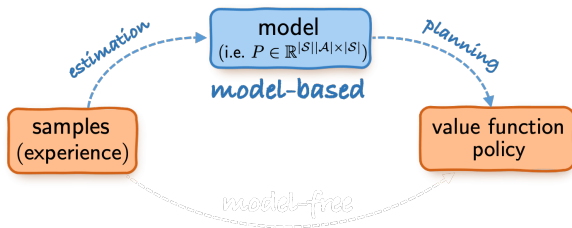


Prior art: sample complexity bounds

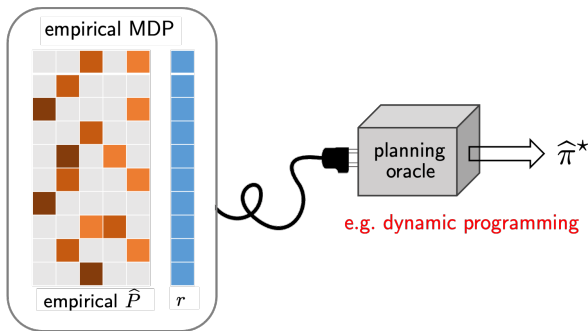


Can we close the gap between upper & lower bounds?

Model-based (“plug-in”) approach?



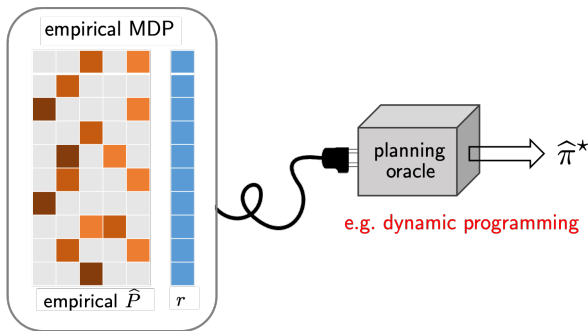
Model-based (“plug-in”) approach?



1. construct empirical model $\hat{P}$:

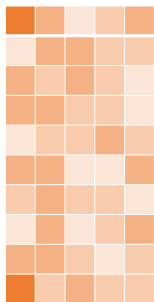
$$\hat{P}(s' | s, a) = \underbrace{\frac{1}{N} \sum_{i=1}^N \mathbb{1}\{s'^{(i)} = s'\}}_{\text{empirical frequency}}$$

Model-based (“plug-in”) approach?

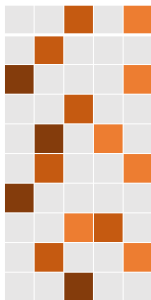


1. construct empirical model $\hat{P}$
2. planning (e.g. value iteration) based on empirical MDP

Issues & challenges in the sample-starved regime



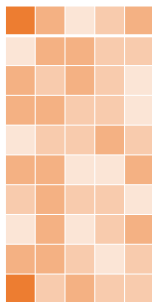
truth: $P \in \mathbb{R}^{SA \times S}$



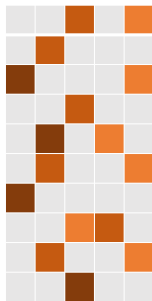
empirical $\hat{P}$ (simulator)

- can't recover P faithfully if sample size $\ll S^2 A$

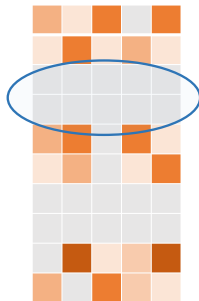
Issues & challenges in the sample-starved regime



truth: $P \in \mathbb{R}^{SA \times S}$



empirical $\hat{P}$ (simulator)



empirical $\hat{P}$ (offline)

- can't recover P faithfully if sample size $\ll S^2 A$
- (possibly) insufficient coverage under offline data

Key idea: pessimism in the face of uncertainty

— *Jin et al, 2020, Rashidinejad et al, 2021, Xie et al, 2021*



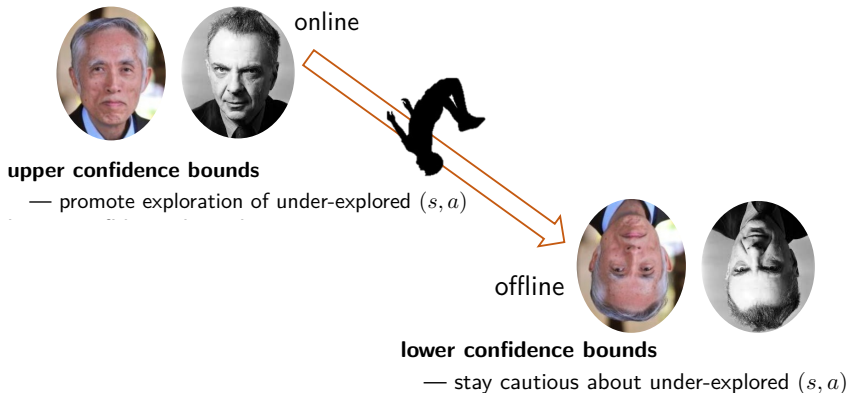
online

upper confidence bounds

— promote exploration of under-explored (s, a)

Key idea: pessimism in the face of uncertainty

— Jin et al, 2020, Rashidinejad et al, 2021, Xie et al, 2021



Key idea: pessimism in the face of uncertainty

— Jin et al, 2020, Rashidinejad et al, 2021, Xie et al, 2021

1. build empirical model $\hat{P}$
2. **(value iteration)** repeat: for all (s, a)

$$\hat{Q}(s, a) \leftarrow \max \left\{ r(s, a) + \gamma \langle \hat{P}(\cdot | s, a), \hat{V} \rangle, 0 \right\}$$

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

Key idea: pessimism in the face of uncertainty

— Jin et al, 2020, Rashidinejad et al, 2021, Xie et al, 2021

Penalize those poorly visited (s, a) ...

1. build empirical model $\hat{P}$
2. **(pessimistic value iteration)** repeat: for all (s, a)

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

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

Key idea: pessimism in the face of uncertainty

— Jin et al, 2020, Rashidinejad et al, 2021, Xie et al, 2021

Penalize those poorly visited $(s, a) \dots$

1. build empirical model $\hat{P}$
2. **(pessimistic value iteration)** repeat: for all (s, a)

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

compared w/ Rashidinejad et al, 2021

- sample-reuse across iterations
- Bernstein-style penalty

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '24)

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

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

with high prob., with sample complexity at most

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

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '24)

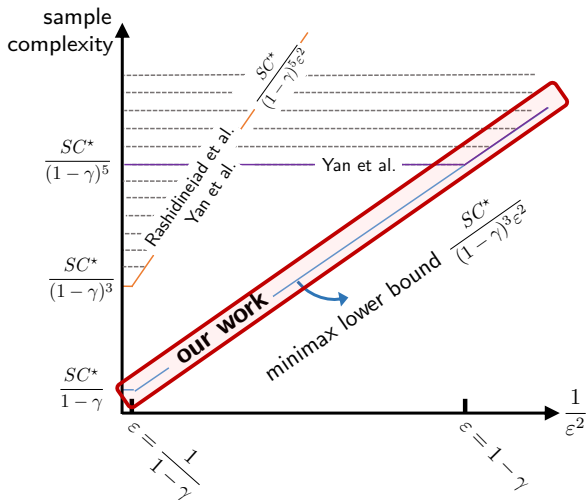
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, the policy $\hat{\pi}$ returned by VI-LCB using a Bernstein-style penalty term achieves

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

with high prob., with sample complexity at most

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

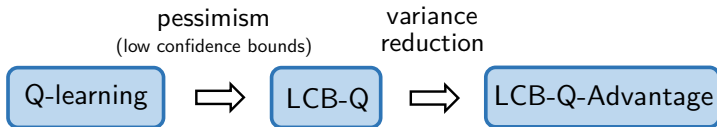
- depends on distribution shift (as reflected by C^*)
- achieves minimax optimality
- full ε -range (no burn-in cost)



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

*Is it possible to design offline model-free algorithms
with optimal sample efficiency?*

*Is it possible to design offline model-free algorithms
with optimal sample efficiency?*



LCB-Q: Q-learning with LCB penalty

— Shi et al, 2022, Yan et al, 2023

$$Q_{t+1}(s_t, a_t) \leftarrow \underbrace{(1 - \eta_t)Q_t(s_t, a_t) + \eta_t \mathcal{T}_t(Q_t)(s_t, a_t)}_{\text{classical Q-learning}} - \underbrace{\eta_t b_t(s_t, a_t)}_{\text{LCB penalty}}$$

LCB-Q: Q-learning with LCB penalty

— Shi et al, 2022, Yan et al, 2023

$$Q_{t+1}(s_t, a_t) \leftarrow \underbrace{(1 - \eta_t)Q_t(s_t, a_t) + \eta_t \mathcal{T}_t(Q_t)(s_t, a_t)}_{\text{classical Q-learning}} - \underbrace{\eta_t b_t(s_t, a_t)}_{\text{LCB penalty}}$$

- $b_t(s, a)$: Hoeffding-style confidence bound
- pessimism in the face of uncertainty

LCB-Q: Q-learning with LCB penalty

— Shi et al, 2022, Yan et al, 2023

$$Q_{t+1}(s_t, a_t) \leftarrow \underbrace{(1 - \eta_t)Q_t(s_t, a_t) + \eta_t \mathcal{T}_t(Q_t)(s_t, a_t)}_{\text{classical Q-learning}} - \underbrace{\eta_t b_t(s_t, a_t)}_{\text{LCB penalty}}$$

- $b_t(s, a)$: Hoeffding-style confidence bound
- pessimism in the face of uncertainty

sample size: $\tilde{O}\left(\frac{SC^*}{(1-\gamma)^5 \varepsilon^2}\right) \implies$ sub-optimal by a factor of $\frac{1}{(1-\gamma)^2}$

Issue: large variability in stochastic update rules

Q-learning with LCB and variance reduction

— Shi et al, 2022, Yan et al, 2023

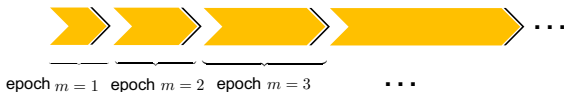
$$Q_{t+1}(s_t, a_t) \leftarrow (1 - \eta_t)Q_t(s_t, a_t) - \underbrace{\eta_t b_t(s_t, a_t)}_{\text{LCB penalty}} \\ + \eta_t \left(\underbrace{\mathcal{T}_t(Q_t) - \mathcal{T}_t(\bar{Q})}_{\text{advantage}} + \underbrace{\hat{\mathcal{T}}(\bar{Q})}_{\text{reference}} \right)(s_t, a_t)$$

Q-learning with LCB and variance reduction

— Shi et al, 2022, Yan et al, 2023

$$Q_{t+1}(s_t, a_t) \leftarrow (1 - \eta_t)Q_t(s_t, a_t) - \underbrace{\eta_t b_t(s_t, a_t)}_{\text{LCB penalty}} + \eta_t \left(\underbrace{\mathcal{T}_t(Q_t) - \mathcal{T}_t(\bar{Q})}_{\text{advantage}} + \underbrace{\hat{\mathcal{T}}(\bar{Q})}_{\text{reference}} \right)(s_t, a_t)$$

- incorporates **variance reduction** into LCB-Q

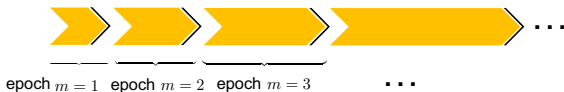


Q-learning with LCB and variance reduction

— Shi et al, 2022, Yan et al, 2023

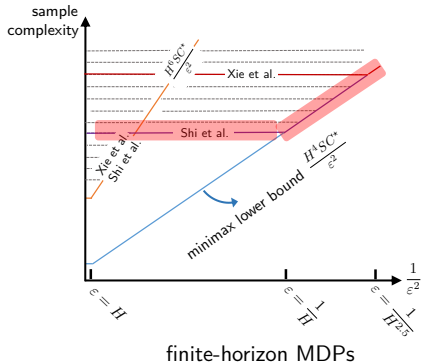
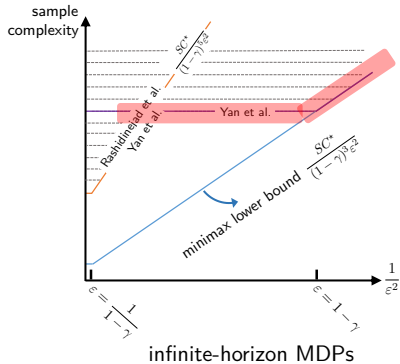
$$Q_{t+1}(s_t, a_t) \leftarrow (1 - \eta_t)Q_t(s_t, a_t) - \underbrace{\eta_t b_t(s_t, a_t)}_{\text{LCB penalty}} + \eta_t \left(\underbrace{\mathcal{T}_t(Q_t) - \mathcal{T}_t(\bar{Q})}_{\text{advantage}} + \underbrace{\hat{\mathcal{T}}(\bar{Q})}_{\text{reference}} \right)(s_t, a_t)$$

- incorporates **variance reduction** into LCB-Q



Theorem (Yan, Li, Chen, Fan '23, Shi, Li, Wei, Chen, Chi '22)

For $\varepsilon \in (0, 1 - \gamma]$, LCB-Q-Advantage achieves $V^*(\rho) - V^{\hat{\pi}}(\rho) \leq \varepsilon$ with optimal sample complexity $\tilde{O}\left(\frac{SC^*}{(1-\gamma)^3 \varepsilon^2}\right)$



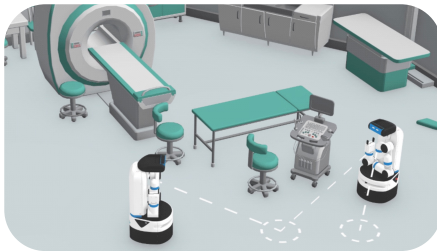
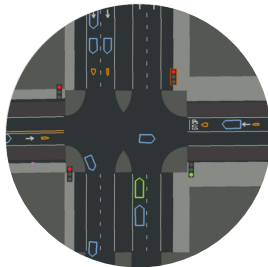
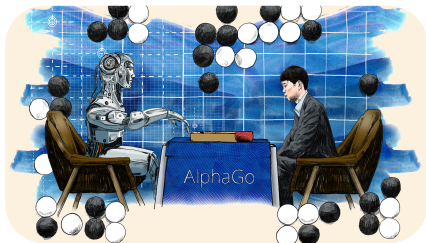
Model-free offline RL attains sample optimality too!

— with some burn-in cost though ...

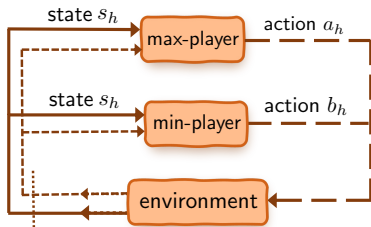
Part 2

1. Online RL
2. Offline RL
3. Multi-agent RL
4. Robust RL

Multi-agent reinforcement learning (MARL)

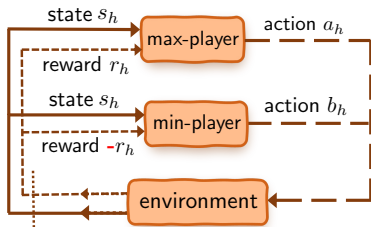


Two-player zero-sum Markov games (finite-horizon)



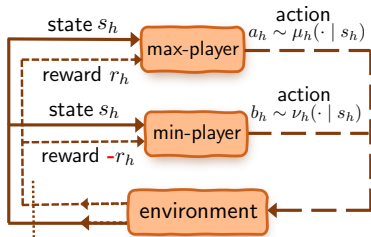
- $\mathcal{S} = [S]$: state space
- $\mathcal{A} = [A]$: action space of max-player
- H : horizon
- $\mathcal{B} = [B]$: action space of min-player

Two-player zero-sum Markov games (finite-horizon)



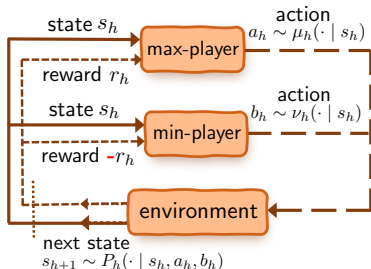
- $\mathcal{S} = [S]$: state space
- $\mathcal{A} = [A]$: action space of max-player
- H : horizon
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s, a, b) \in [0, 1]$
min-player $-r(s, a, b)$

Two-player zero-sum Markov games (finite-horizon)



- $\mathcal{S} = [S]$: state space
- $\mathcal{A} = [A]$: action space of max-player
- H : horizon
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s, a, b) \in [0, 1]$
min-player $-r(s, a, b)$
- $\mu : \mathcal{S} \times [H] \rightarrow \Delta(\mathcal{A})$: policy of max-player
 $\nu : \mathcal{S} \times [H] \rightarrow \Delta(\mathcal{B})$: policy of min-player

Two-player zero-sum Markov games (finite-horizon)



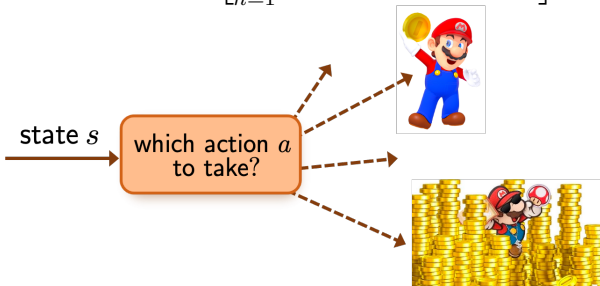
- $\mathcal{S} = [S]$: state space
- $\mathcal{A} = [A]$: action space of max-player
- H : horizon
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s, a, b) \in [0, 1]$
min-player $-r(s, a, b)$
- $\mu : \mathcal{S} \times [H] \rightarrow \Delta(\mathcal{A})$: policy of max-player
 $\nu : \mathcal{S} \times [H] \rightarrow \Delta(\mathcal{B})$: policy of min-player
- $P_h(\cdot | s, a, b)$: **unknown** transition probabilities

Value function under *independent* policies (μ, ν) (no coordination)

$$V^{\mu, \nu}(s) := \mathbb{E} \left[\sum_{h=1}^H r_h(s_h, a_h, b_h) \mid s_1 = s \right]$$

Value function under *independent* policies (μ, ν) (no coordination)

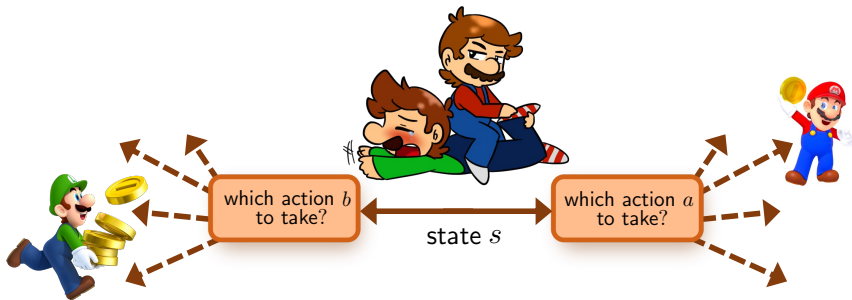
$$V^{\mu, \nu}(s) := \mathbb{E} \left[\sum_{h=1}^H r_h(s_h, a_h, b_h) \mid s_1 = s \right]$$



- Each agent seeks **optimal policy** maximizing her own value

Value function under *independent* policies (μ, ν) (no coordination)

$$V^{\mu, \nu}(s) := \mathbb{E} \left[\sum_{h=1}^H r_h(s_h, a_h, b_h) \mid s_1 = s \right]$$



- Each agent seeks **optimal policy** maximizing her own value
- But two agents have conflicting goals ...

Compromise: Nash equilibrium (NE)



John von Neumann



John Nash

An NE policy pair (μ^*, ν^*) obeys

$$\max_{\mu} V^{\mu, \nu^*} = V^{\mu^*, \nu^*} = \min_{\nu} V^{\mu^*, \nu}$$

Compromise: Nash equilibrium (NE)



John von Neumann



John Nash

An NE policy pair (μ^*, ν^*) obeys

$$\max_{\mu} V^{\mu, \nu^*} = V^{\mu^*, \nu^*} = \min_{\nu} V^{\mu^*, \nu}$$

- no unilateral deviation is beneficial

Compromise: Nash equilibrium (NE)



John von Neumann



John Nash

An NE policy pair (μ^*, ν^*) obeys

$$\max_{\mu} V^{\mu, \nu^*} = V^{\mu^*, \nu^*} = \min_{\nu} V^{\mu^*, \nu}$$

- no unilateral deviation is beneficial
- no coordination between two agents (they act *independently*)

Compromise: Nash equilibrium (NE)



John von Neumann



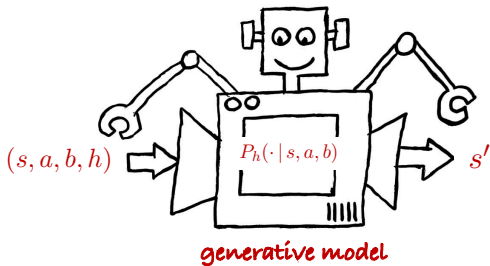
John Nash

An ε -NE policy pair $(\hat{\mu}, \hat{\nu})$ obeys

$$\max_{\mu} V^{\mu, \hat{\nu}} - \varepsilon \leq V^{\hat{\mu}, \hat{\nu}} \leq \min_{\nu} V^{\hat{\mu}, \nu} + \varepsilon$$

- no unilateral deviation is beneficial
- no coordination between two agents (they act *independently*)

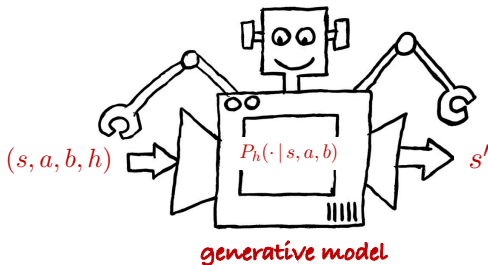
Learning NEs with a simulator



input: any (s, a, b, h)

output: an independent sample $s' \sim P_h(\cdot | s, a, b)$

Learning NEs with a simulator



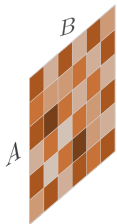
input: any (s, a, b, h)

output: an independent sample $s' \sim P_h(\cdot | s, a, b)$

Question: how many samples are sufficient to learn an ε -Nash policy pair?

Model-based approach (non-adaptive sampling)

— Zhang, Kakade, Başar, Yang '20

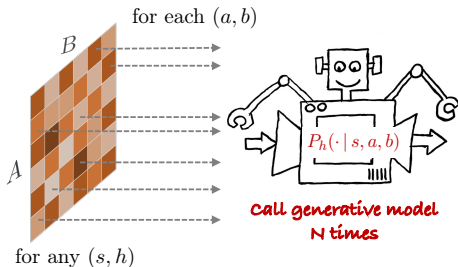


for any (s, h)

1. for each (s, a, b, h) , call simulator N times

Model-based approach (non-adaptive sampling)

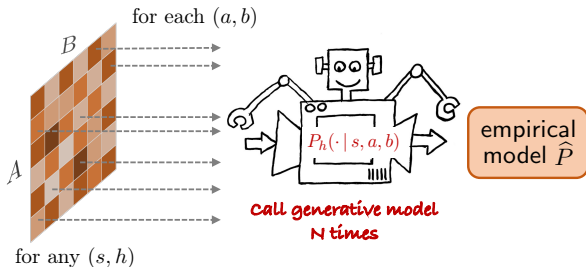
— Zhang, Kakade, Başar, Yang '20



1. for each (s, a, b, h) , call simulator N times

Model-based approach (non-adaptive sampling)

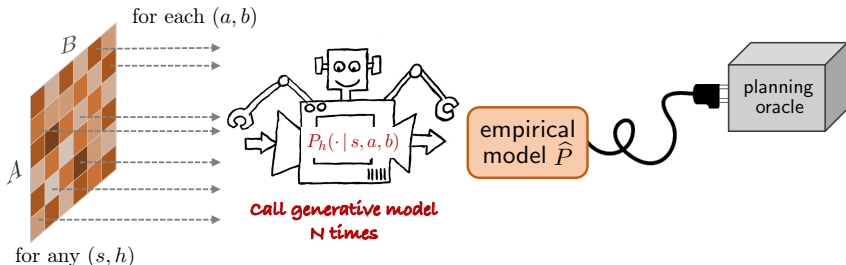
— Zhang, Kakade, Başar, Yang '20



1. for each (s, a, b, h) , call simulator N times
2. build empirical model $\hat{P}$

Model-based approach (non-adaptive sampling)

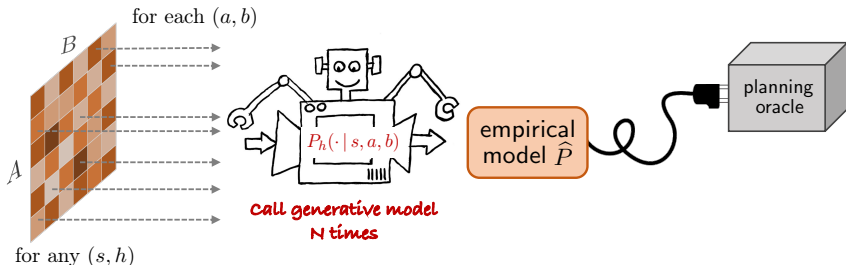
— Zhang, Kakade, Başar, Yang '20



1. for each (s, a, b, h) , call simulator N times
2. build empirical model $\hat{P}$, and run "plug-in" methods

Model-based approach (non-adaptive sampling)

— Zhang, Kakade, Başar, Yang '20



1. for each (s, a, b, h) , call simulator N times
2. build empirical model $\hat{P}$, and run “plug-in” methods

sample complexity: $\frac{H^4 S_{AB}}{\epsilon^2}$

Curse of multiple agents



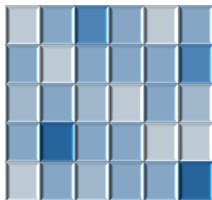
1 player: A

Let's look at the **size** of joint action space ...

Curse of multiple agents



1 player: A



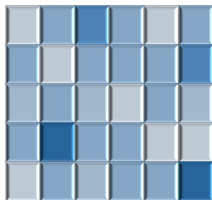
2 players: AB

Let's look at the **size** of joint action space ...

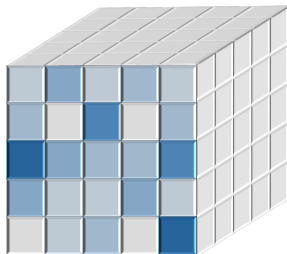
Curse of multiple agents



1 player: A



2 players: AB



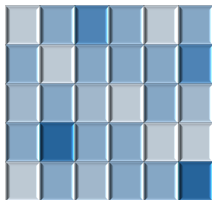
m players: $A_1 A_2 \cdots A_m$

Let's look at the **size** of joint action space ...

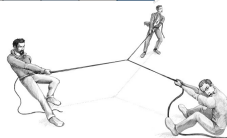
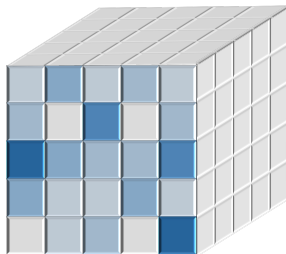
Curse of multiple agents



1 player: A



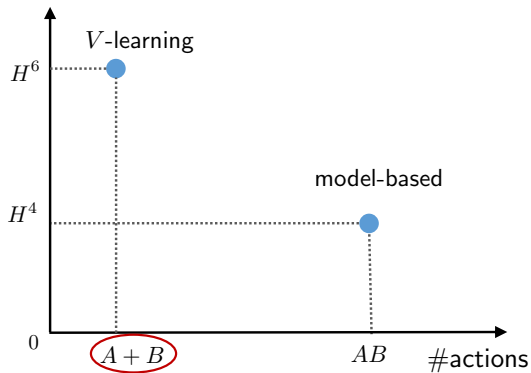
2 players: AB



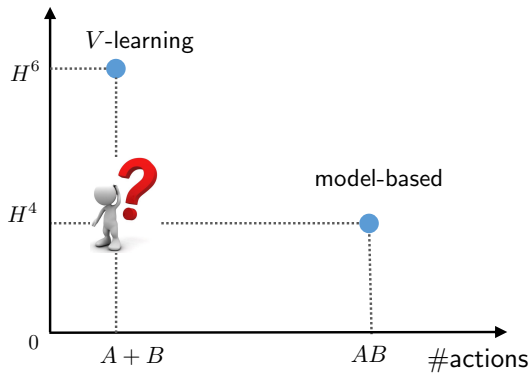
m players: $A_1 A_2 \cdots A_m$

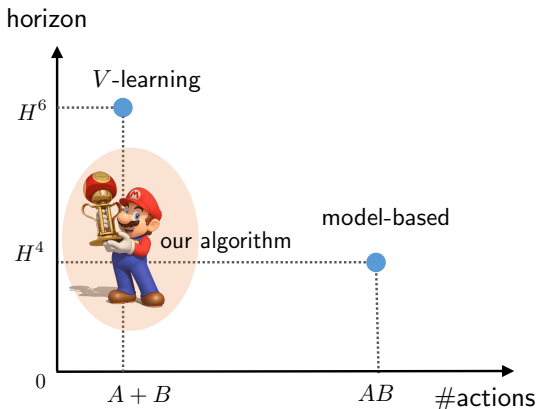
joint actions **blows up geometrically** in # players!

horizon



horizon





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

For any $0 < \varepsilon \leq H$, one can design an algorithm that finds an ε -Nash policy pair $(\hat{\mu}, \hat{\nu})$ with high prob., with sample complexity at most

$$\tilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right) \quad (\text{minimax-optimal } \forall \varepsilon)$$

Part 2

1. Online RL
2. Offline RL
3. Multi-agent RL
4. Robust RL

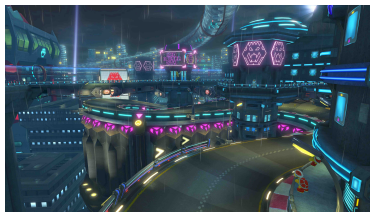
Safety and robustness in RL

(Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022;)



Training environment

≠



Test environment

Safety and robustness in RL

(Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022;)



Training environment

≠



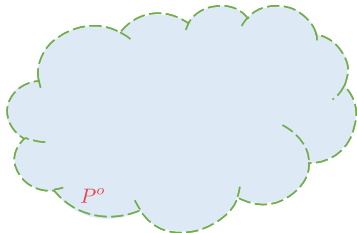
Test environment

Sim2Real Gap: Can we learn optimal policies that are robust to model perturbations?

Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o :

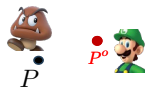
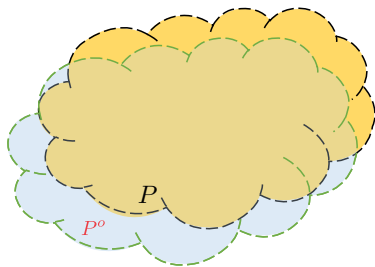
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$



Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o :

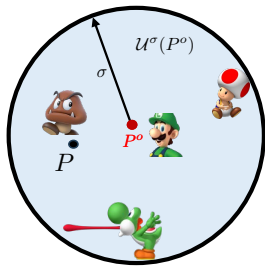
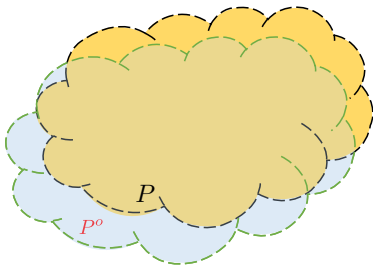
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$



Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o :

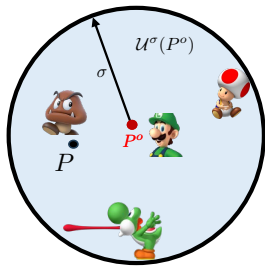
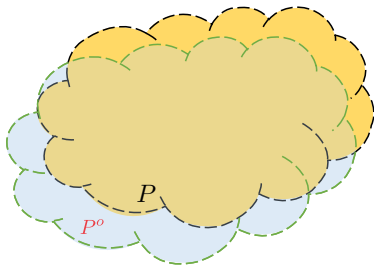
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$



Modeling environment uncertainty

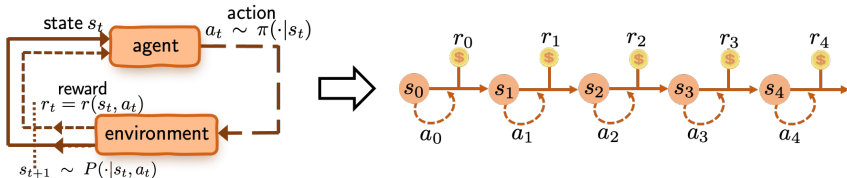
Uncertainty set of the nominal transition kernel P^o :

$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$



- Examples of ρ : f-divergence (TV, χ^2 , KL...)

Robust value/Q function



Robust value/Q function of policy π :

$$\forall s \in \mathcal{S} : \quad V^{\pi, \sigma}(s) := \inf_{P \in \mathcal{U}^{\sigma}(P^o)} \mathbb{E}_{\pi, P} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right]$$

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

Measures the **worst-case** performance of the policy in the uncertainty set.

Distributionally robust MDP

Robust MDP

Find the policy π^ that maximizes $V^{\pi, \sigma}$*

(Iyengar. '05, Nilim and El Ghaoui. '05)

Distributionally robust MDP

Robust MDP

Find the policy π^ that maximizes $V^{\pi,\sigma}$*

(Iyengar. '05, Nilim and El Ghaoui. '05)

Robust Bellman's optimality equation: the optimal robust policy π^* and optimal robust value $V^{*,\sigma} := V^{\pi^*,\sigma}$ satisfy

$$Q^{*,\sigma}(s, a) = r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(P_{s,a}^o)} \langle P_{s,a}, V^{*,\sigma} \rangle,$$
$$V^{*,\sigma}(s) = \max_a Q^{*,\sigma}(s, a)$$

Distributionally robust MDP

Robust MDP

Find the policy π^ that maximizes $V^{\pi,\sigma}$*

(Iyengar. '05, Nilim and El Ghaoui. '05)

Robust Bellman's optimality equation: the optimal robust policy π^* and optimal robust value $V^{*,\sigma} := V^{\pi^*,\sigma}$ satisfy

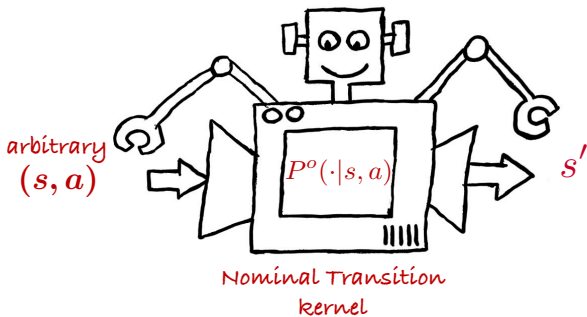
$$Q^{*,\sigma}(s, a) = r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(P_{s,a}^o)} \langle P_{s,a}, V^{*,\sigma} \rangle,$$
$$V^{*,\sigma}(s) = \max_a Q^{*,\sigma}(s, a)$$

Distributionally robust value iteration (DRVI):

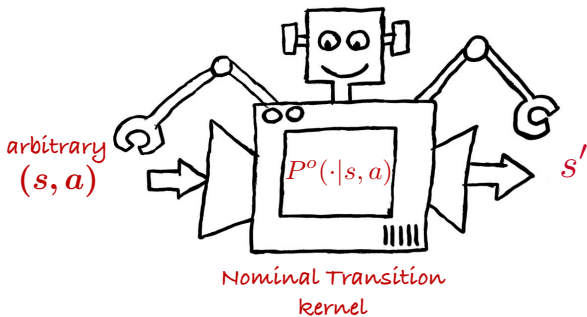
$$Q(s, a) \leftarrow r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(P_{s,a}^o)} \langle P_{s,a}, V \rangle,$$

where $V(s) = \max_a Q(s, a)$.

Learning distributionally robust MDPs



Learning distributionally robust MDPs

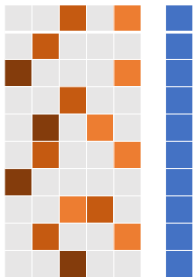


Goal of robust RL: given $\mathcal{D} := \{(s_i, a_i, s'_i)\}_{i=1}^N$ from the *nominal* environment P^0 , find an ε -optimal robust policy $\hat{\pi}$ obeying

$$V^{*,\sigma} - V^{\hat{\pi},\sigma} \leq \varepsilon$$

— in a sample-efficient manner

A curious question



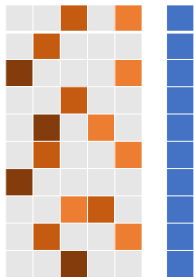
empirical MDP

Learn the optimal policy of the nominal MDP?

Learn the **robust** policy around the nominal MDP?



A curious question



empirical MDP

Learn the optimal policy of the nominal MDP?

Learn the **robust** policy around the nominal MDP?



Robustness-statistical trade-off? Is there a statistical premium that one needs to pay in quest of additional robustness?

Sample complexity under TV uncertainty

Theorem (Shi et al., 2023)

Assume the uncertainty set is measured via the TV distance with radius $\sigma \in [0, 1]$. For sufficiently small $\varepsilon > 0$, DRVI outputs a policy $\hat{\pi}$ that satisfies $V^{*,\sigma} - V^{\hat{\pi},\sigma} \leq \varepsilon$ with sample complexity at most

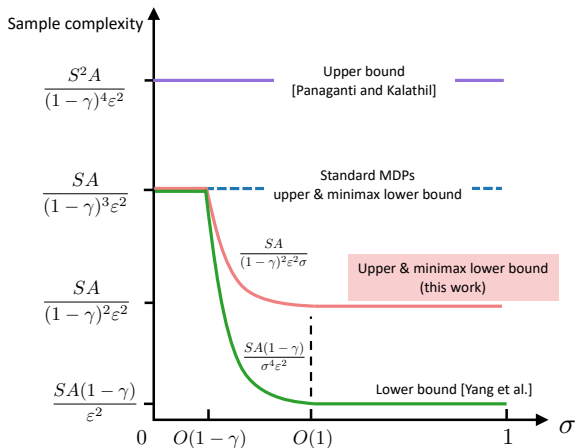
$$\tilde{O}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\} \varepsilon^2}\right)$$

ignoring logarithmic factors. In addition, no algorithm can succeed if the sample size is below

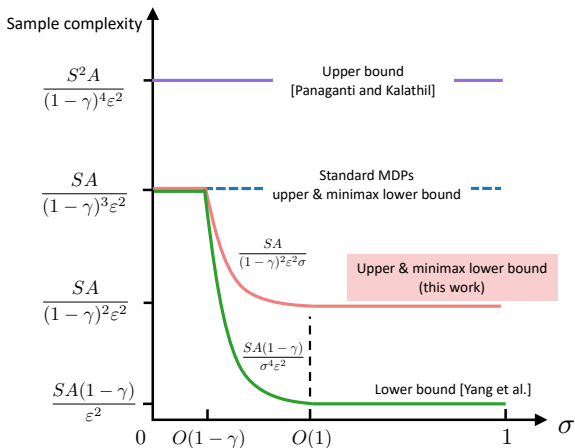
$$\tilde{\Omega}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\} \varepsilon^2}\right).$$

- Establish the minimax optimality of DRVI for RMDP under the TV uncertainty set over the full range of σ .

When the uncertainty set is TV



When the uncertainty set is TV



RMDPs are **easier** to learn than standard MDPs.

Sample complexity theorem under χ^2 uncertainty

Theorem (Shi et al., 2025)

Assume the uncertainty set is measured via the χ^2 divergence with radius $\sigma \in [0, \infty)$. For sufficiently small $\varepsilon > 0$, DRVI outputs a policy $\hat{\pi}$ that satisfies $V^{*,\sigma} - V^{\hat{\pi},\sigma} \leq \varepsilon$ with sample complexity at most

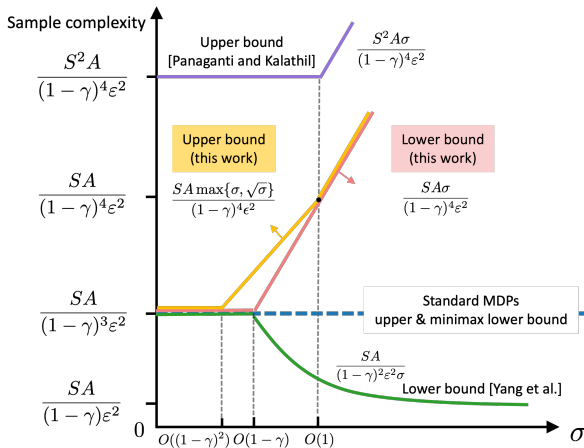
$$\tilde{O} \left(\frac{SA}{(1-\gamma)^3 \varepsilon^2} \left(1 + \frac{\sqrt{\sigma} + \sigma}{1-\gamma} \right) \right)$$

ignoring logarithmic factors. In addition, no algorithm can succeed if the sample size is below

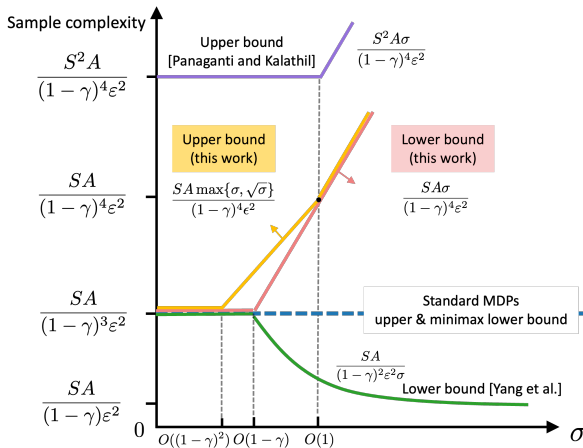
$$\tilde{\Omega} \left(\frac{SA}{(1-\gamma)^3 \varepsilon^2} \left(1 + \frac{\sigma}{1-\gamma} \right) \right)$$

- Establish the minimax optimality of DRVI for RMDP under the TV uncertainty set over the full range of σ .

When the uncertainty set is χ^2 divergence



When the uncertainty set is χ^2 divergence



RMDPs are **harder** to learn than standard MDPs.

Reference: online RL I

- “*Asymptotically efficient adaptive allocation rules*,” T. L. Lai, H. Robbins, *Advances in applied mathematics*, vol. 6, no. 1, 1985
- “*Finite-time analysis of the multiarmed bandit problem*,” P. Auer, N. Cesa-Bianchi, P. Fischer, *Machine learning*, vol. 47, pp. 235-256, 2002
- “*Minimax regret bounds for reinforcement learning*,” M. G. Azar, I. Osband, R. Munos, *ICML*, 2017
- “*Is Q-learning provably efficient?*” C. Jin, Z. Allen-Zhu, S. Bubeck, and M. Jordan, *NeurIPS*, 2018
- “*Provably efficient Q-learning with low switching cost*,” Y. Bai, T. Xie, N. Jiang, Y. X. Wang, *NeurIPS*, 2019
- “*Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited*” O. D. Domingues, P. Menard, E. Kaufmann, M. Valko, *Algorithmic Learning Theory*, 2021
- “*Almost optimal model-free reinforcement learning via reference-advantage decomposition*,” Z. Zhang, Y. Zhou, X. Ji, *NeurIPS*, 2020

Reference: online RL II

- *"Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon,"* Z. Zhang, X. Ji, and S. Du, *COLT*, 2021
- *"Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning,"* G. Li, L. Shi, Y. Chen, Y. Gu, Y. Chi, *NeurIPS*, 2021
- *"Regret-optimal model-free reinforcement learning for discounted MDPs with short burn-in time,"* X. Ji, G. Li, *NeurIPS*, 2023
- *"Reward-free exploration for reinforcement learning,"* C. Jin, A. Krishnamurthy, M. Simchowitz, T. Yu, *ICML*, 2020
- *"Minimax-optimal reward-agnostic exploration in reinforcement learning,"* G. Li, Y. Yan, Y. Chen, J. Fan, *COLT*, 2024
- *"Settling the sample complexity of online reinforcement learning,"* Z. Zhang, Y. Chen, J. D. Lee, S. S. Du, *COLT*, 2024

Reference: offline RL I

- “*Bridging offline reinforcement learning and imitation learning: A tale of pessimism*,” P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, S. Russell, *NeurIPS*, 2021
- “*Is pessimism provably efficient for offline RL?*” Y. Jin, Z. Yang, Z. Wang, *ICML*, 2021
- “*Settling the sample complexity of model-based offline reinforcement learning*,” G. Li, L. Shi, Y. Chen, Y. Chi, Y. Wei, *Annals of Statistics*, vol. 52, no. 1, pp. 233-260, 2024
- “*Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity*,” L. Shi, G. Li, Y. Wei, Y. Chen, Y. Chi, *ICML*, 2022
- “*The efficacy of pessimism in asynchronous Q-learning*,” Y. Yan, G. Li, Y. Chen, J. Fan, *IEEE Transactions on Information Theory*, 2023
- “*Policy finetuning: Bridging sample-efficient offline and online reinforcement learning*” T. Xie, N. Jiang, H. Wang, C. Xiong, Y. Bai, *NeurIPS*, 2021

Reference: multi-agent RL I

- “*Stochastic games*,” L. S. Shapley, *Proceedings of the national academy of sciences*, 1953
- “*Twenty lectures on algorithmic game theory*,” T. Roughgarden, 2016
- “*Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity*,” K. Zhang, S. Kakade, T. Basar, L. Yang, *NeurIPS*, 2020
- “*When can we learn general-sum Markov games with a large number of players sample-efficiently?*” Z. Song, S. Mei, Y. Bai, *ICLR*, 2021
- “*V-learning—A simple, efficient, decentralized algorithm for multiagent RL*,” C. Jin, Q. Liu, Y. Wang, T. Yu, 2021
- “*Minimax-optimal multi-agent RL in Markov games with a generative model*,” G. Li, Y. Chi, Y. Wei, Y. Chen, *NeurIPS*, 2022
- “*When are offline two-player zero-sum Markov games solvable?*” Q. Cui, S. S. Du, *NeurIPS*, 2022
- “*Model-based reinforcement learning for offline zero-sum Markov games*,” Y. Yan, G. Li, Y. Chen, J. Fan, *Operations Research*, 2024

Reference: robust RL I

- “*Robust dynamic programming*,” G. Iyengar, *Mathematics of Operations Research*, 2005
- “*The curious price of distributional robustness in reinforcement learning with a generative model.*,” L. Shi, G. Li, Y. Wei, Y. Chen, M. Geist, Y. Chi, *Operations Research*, 2026
- “*Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity*,” L. Shi, Y. Chi, 2022
- “*On the foundation of distributionally robust reinforcement learning*,” S. Wang, N. Si, J. Blanchet, and Z. Zhou, 2023
- “*Sample complexity of robust reinforcement learning with a generative model*,” K. Panaganti, D. Kalathil, *AISTATS*, 2022
- “*Sample-Efficient Robust Multi-Agent Reinforcement Learning in the Face of Environmental Uncertainty*,” L. Shi, E. Mazumdar, Y. Chi, and A. Wierman, *ICML*, 2024