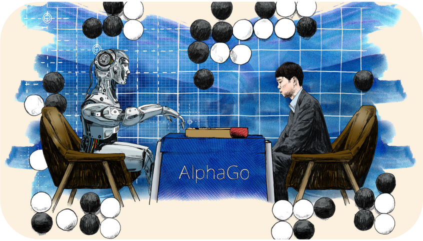
Non-Asymptotic Analysis for Reinforcement Learning (Part 2)

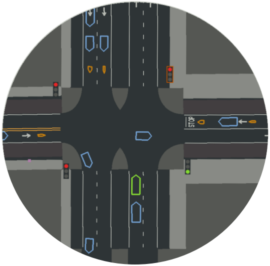
Yuxin Chen

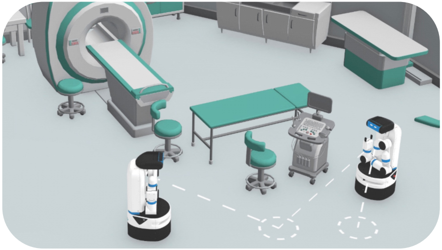
Wharton Statistics & Data Science, SIGMETRICS 2023

Multi-agent RL with a generative model

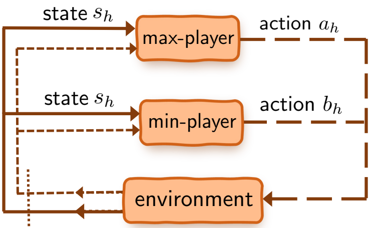
Multi-agent reinforcement learning (MARL)





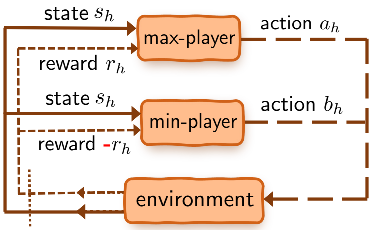


Two-player zero-sum Markov games (finite-horizon)



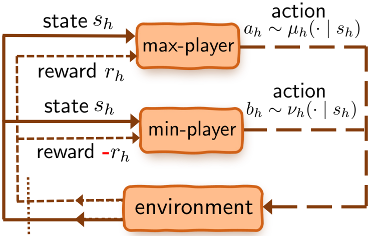
- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player

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



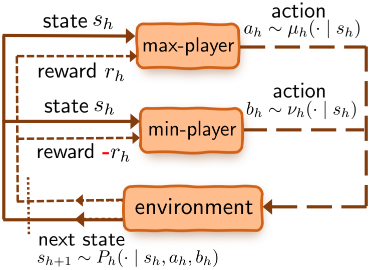
- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s,a,b) \in [0,1]$
  min-player $-r(s,a,b)$

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 $-r(s, a, b)$
- $\mu : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$: policy of max-player
  $\nu : \mathcal{S} \times [H] \to \Delta(\mathcal{B})$: policy of min-player

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



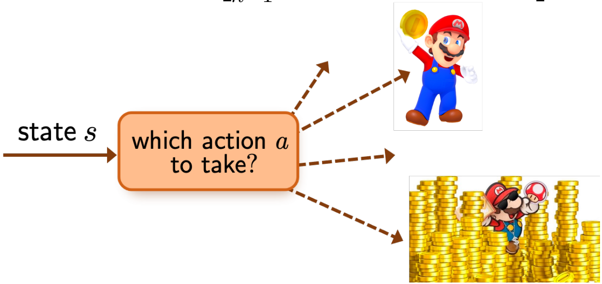
- $\mathcal{S} = [S]$: state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r(s,a,b) \in [0,1]$
  min-player $-r(s,a,b)$
- $\mu : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$: policy of max-player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

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

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

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

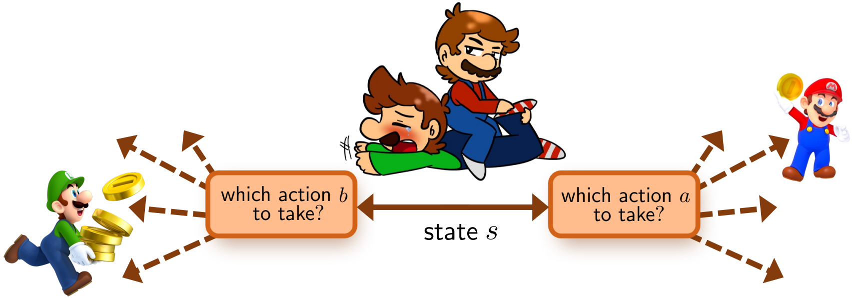
$$V^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{h=1}^{H} r_h(s_h, a_h, b_h) \,\middle|\, s_1 = s\right]$$



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

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

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



- 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 $(\mu^\star, \nu^\star)$ obeys

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

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

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

- no unilateral deviation is beneficial

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

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

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

Compromise: Nash equilibrium (NE)

John von Neumann

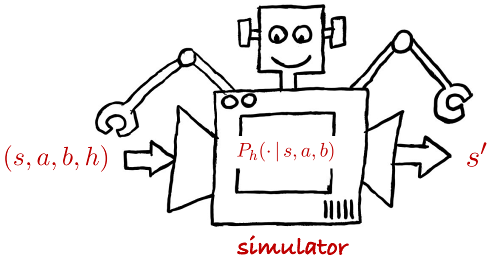
John Nash

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

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

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

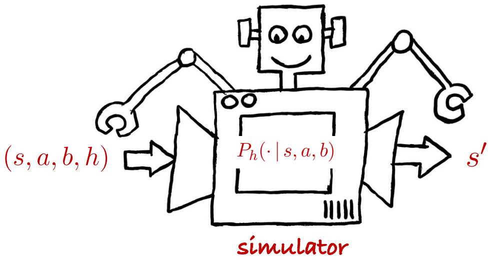
Learning NEs with a simulator



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

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

Learning NEs with a simulator



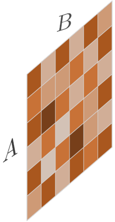
input: any (s, a, b, h)

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*

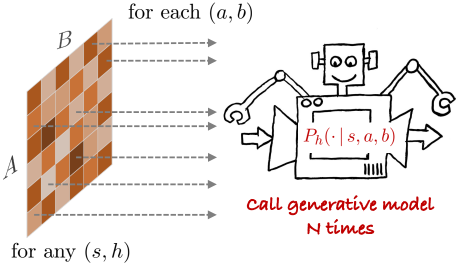


for any (s, h)

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

Model-based approach (non-adaptive sampling)

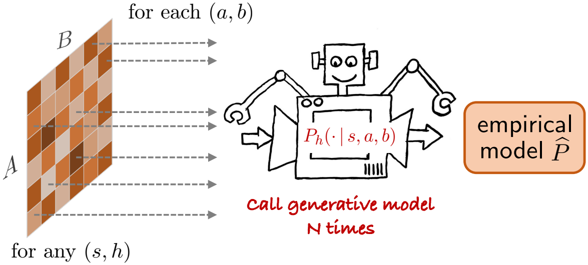
— *Zhang, Kakade, Başar, Yang '20*



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

Model-based approach (non-adaptive sampling)

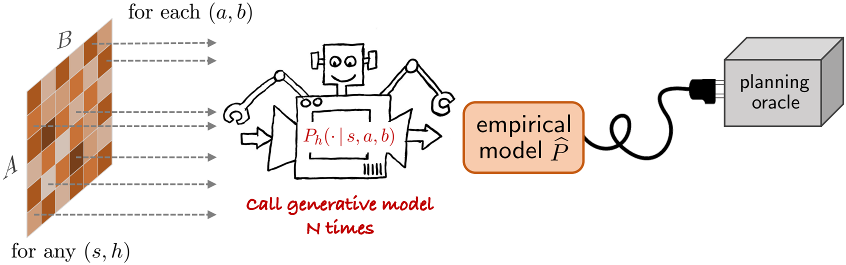
— Zhang, Kakade, Başar, Yang '20



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

Model-based approach (non-adaptive sampling)

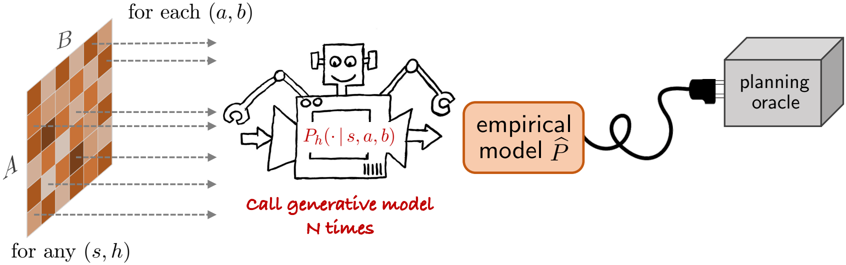
— *Zhang, Kakade, Başar, Yang '20*



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

Model-based approach (non-adaptive sampling)

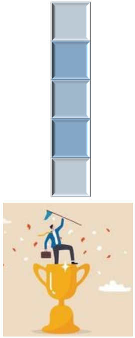
— *Zhang, Kakade, Başar, Yang '20*



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

sample complexity: $\frac{H^4 S A B}{\varepsilon^2}$

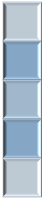
Curse of multiple agents



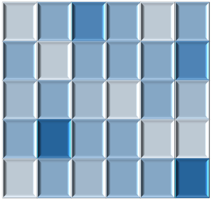
1 player: A

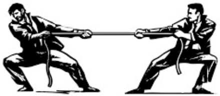
Let's look at the size of joint action space . . .

Curse of multiple agents



1 player: A

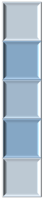




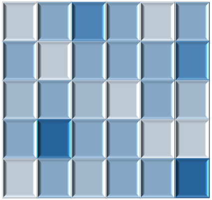
2 players: AB

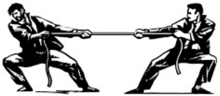
Let's look at the size of joint action space ...

Curse of multiple agents

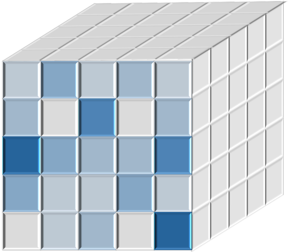


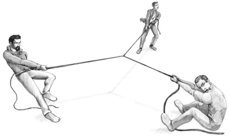
1 player: A





2 players: AB

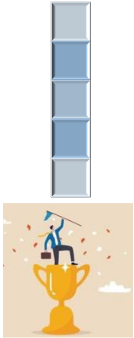




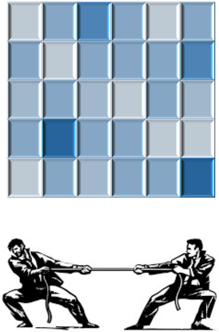
m players: $A_1 A_2 \cdots A_m$

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

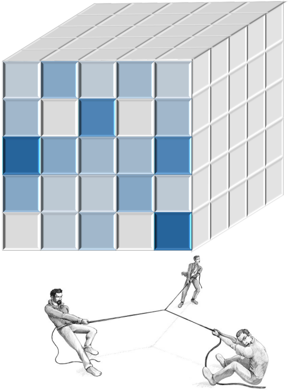
Curse of multiple agents



1 player: A

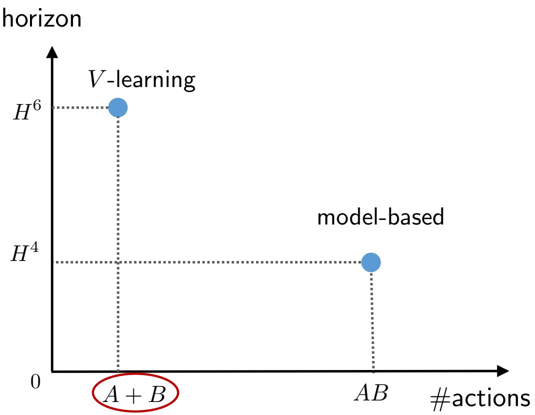


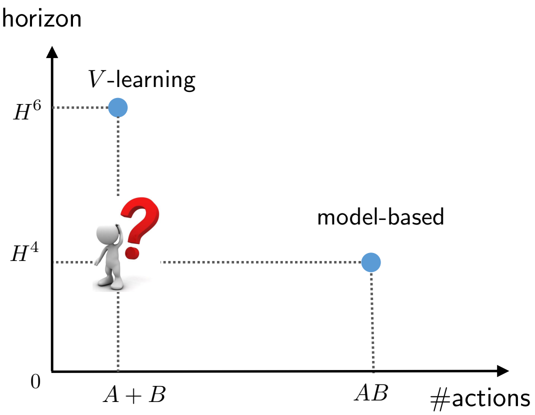
2 players: AB

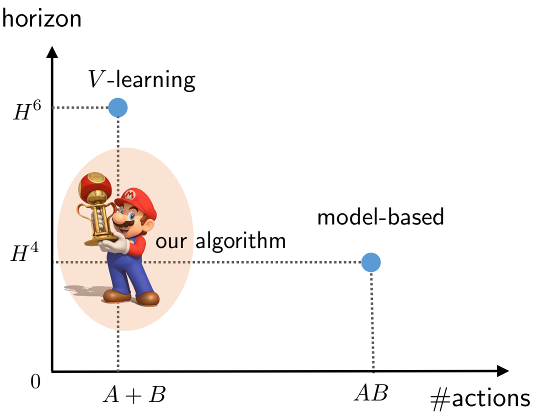


m players: $A_1 A_2 \cdots A_m$

joint actions blows up geometrically in # players!


horizon
V-learning
H^6
model-based
H^4
0
$A + B$
AB
#actions


horizon
V-learning
H^6
model-based
H^4
0
$A + B$
AB
#actions



Theorem 1 (Li, Chi, Wei, Chen '22)

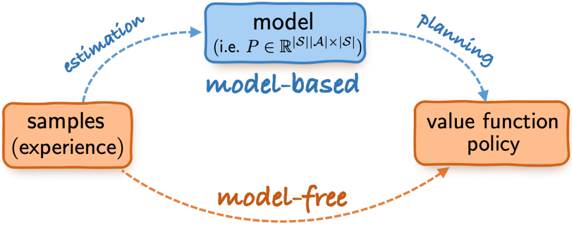
For any $0 < \varepsilon \le H$, one can design an algorithm that finds an ε-Nash policy pair $(\widehat{\mu}, \widehat{\nu})$ with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right) \qquad \textit{(minimax-optimal } \forall\varepsilon\textit{)}$$

Model-free / value-based RL

1. Basics of Q-learning
2. Synchronous Q-learning and variance reduction (simulator)
3. Asynchronous Q-learning (Markovian data)
4. Q-learning with lower confidence bounds (offline RL)
5. Q-learning with upper confidence bounds (online RL)

Model-based vs. model-free RL



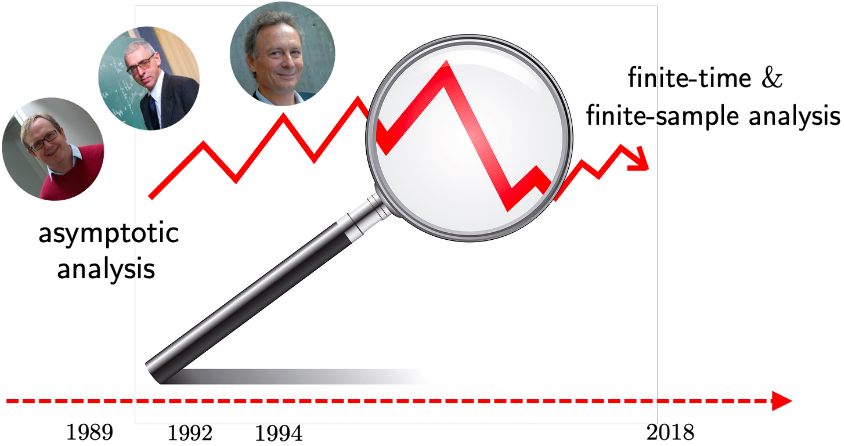
Model-based approach ("plug-in")

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

Model-free / value-based approach

— learning w/o modeling & estimating environment explicitly

— memory-efficient, online, . . .



Focus of this part: classical **Q-learning** algorithm and its variants

A starting point: Bellman optimality principle

Bellman operator

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

- one-step look-ahead

A starting point: Bellman optimality principle

Bellman operator

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

- one-step look-ahead

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

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

A starting point: Bellman optimality principle

Bellman operator

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

- one-step look-ahead

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

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

- **takeaway message:** it suffices to solve the Bellman equation
- **challenge:** how to solve it using stochastic samples?

Richard Bellman

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation (Robbins & Monro, 1951) for solving the **Bellman equation**

$$\mathcal{T}(Q) - Q = 0$$

where

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

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $\mathcal{T}(Q) - Q = 0$

$$\underbrace{Q_{t+1}(s,a) = Q_t(s,a) + \eta_t\big(\mathcal{T}_t(Q_t)(s,a) - Q_t(s,a)\big)}_{\text{sample transition } (s,a,s')}, \quad t \geq 0$$

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $\mathcal{T}(Q) - Q = 0$

$$\underbrace{Q_{t+1}(s,a) = Q_t(s,a) + \eta_t\big(\color{blue}{\mathcal{T}_t}(Q_t)(s,a) - Q_t(s,a)\big)}_{\text{sample transition } (s,a,s')}, \quad t \geq 0$$

$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

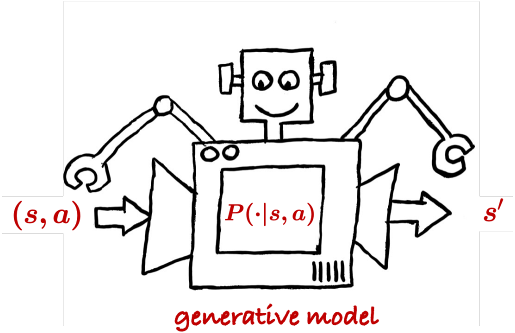
$$\mathcal{T}(Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \Big[\max_{a'} Q(s',a')\Big]$$

Model-free RL

1. Basics of Q-learning
2. Synchronous Q-learning and variance reduction (simulator)
3. Asynchronous Q-learning (Markovian data)
4. Q-learning with lower confidence bounds (offline RL)
5. Q-learning with upper confidence bounds (online RL)

A generative model / simulator

— *Kearns, Singh '99*



Each iteration, draw an independent sample (s, a, s') for given (s, a)

Synchronous Q-learning

Chris Watkins

Peter Dayan

for $t = 0, 1, \ldots, T$

for each $(s, a) \in \mathcal{S} \times \mathcal{A}$

draw a sample (s, a, s'), run

$$Q_{t+1}(s, a) = (1 - \eta_t) Q_t(s, a) + \eta_t \Big\{ r(s, a) + \gamma \max_{a'} Q_t(s', a') \Big\}$$

synchronous: all state-action pairs are updated simultaneously

- total sample size: $T|\mathcal{S}||\mathcal{A}|$

Sample complexity of synchronous Q-learning

Theorem 2 (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \le 1$, *synchronous Q-learning yields* $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ *with high prob. and* $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \le \varepsilon$, *with sample size at most*

$$\begin{cases} \widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) & \text{if } |\mathcal{A}| \ge 2 \\ \widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big) & \text{if } |\mathcal{A}| = 1 \qquad (\textit{TD learning}) \end{cases}$$

Sample complexity of synchronous Q-learning

Theorem 2 (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$, with sample size at most

$$\begin{cases} \widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\Big) & \text{if } |\mathcal{A}| \geq 2 \\ \widetilde{O}\Big(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\Big) & \text{if } |\mathcal{A}| = 1 \qquad (\textit{TD learning}) \end{cases}$$

- Covers both *constant* and *rescaled linear* learning rates:

$$\eta_t \equiv \frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \quad \text{or} \quad \eta_t = \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}}$$

Sample complexity of synchronous Q-learning

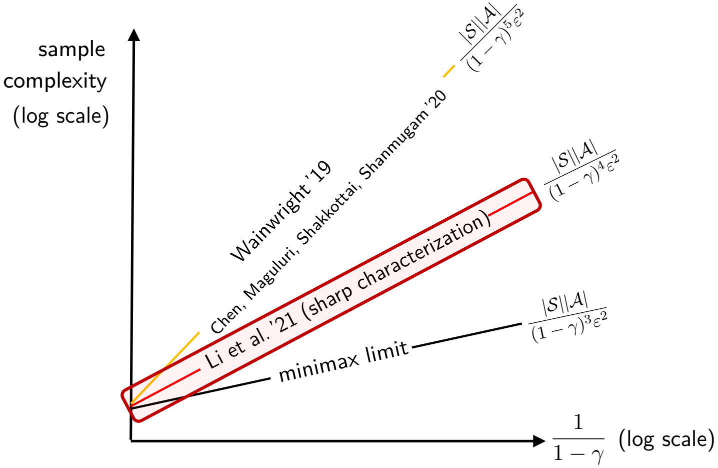
Theorem 2 (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$, with sample size at most

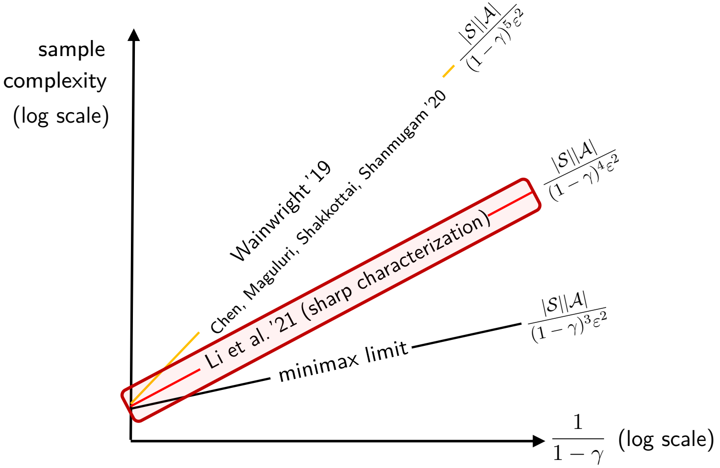
$$\begin{cases} \widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) & \text{if } |\mathcal{A}| \geq 2 \qquad \text{(?)} \\ \widetilde{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)^3\varepsilon^2}\right) & \text{if } |\mathcal{A}| = 1 \qquad \text{(minimax optimal)} \end{cases}$$

other papers	sample complexity
Even-Dar & Mansour '03	$2^{\frac{1}{1-\gamma}} \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$
Beck & Srikant '12	$\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^5\varepsilon^2}$
Wainwright '19	$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$
Chen, Maguluri, Shakkottai, Shanmugam '20	$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$

All this requires sample size at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ $(|\mathcal{A}| \geq 2)$ . . .



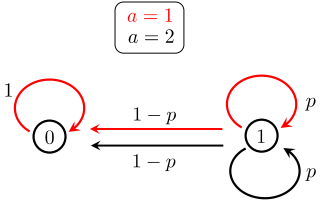
All this requires sample size at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ $(|\mathcal{A}| \geq 2)$ ...



Question: *Is Q-learning sub-optimal, or is it an analysis artifact?*

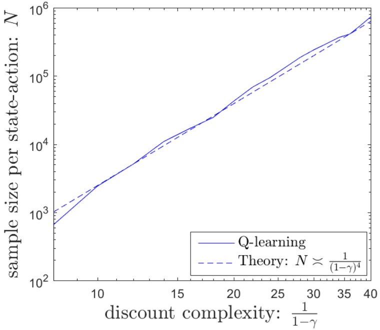
A numerical example: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ samples seem necessary ...

— *observed in Wainwright '19*



$$p = \frac{4\gamma - 1}{3\gamma}$$

$$r(0,1) = 0, \quad r(1,1) = r(1,2) = 1$$



Q-learning is NOT minimax optimal

Theorem 3 (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, *there exists an MDP with* $|\mathcal{A}| \geq 2$ *such that to achieve* $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$, *synchronous Q-learning needs at least*

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$

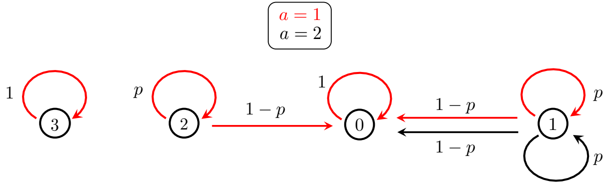
Q-learning is NOT minimax optimal

Theorem 3 (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, *there exists an MDP with* $|\mathcal{A}| \geq 2$ *such that to achieve* $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, *synchronous Q-learning needs at least*

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$

- Tight **algorithm-dependent** lower bound
- Holds for both constant and rescaled linear learning rates

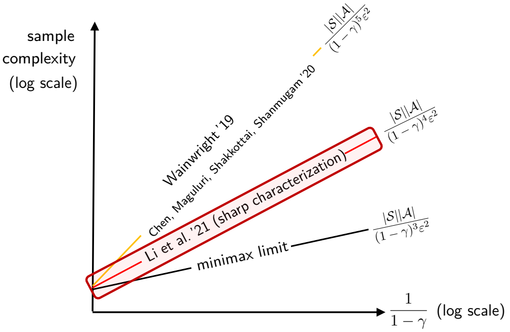


Q-learning is NOT minimax optimal

Theorem 3 (Li, Cai, Chen, Wei, Chi, 2021)

For any $0 < \varepsilon \le 1$, there exists an MDP with $|\mathcal{A}| \ge 2$ such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$



Improving sample complexity via ***variance reduction***

— *a powerful idea from finite-sum stochastic optimization*

Variance-reduced Q-learning updates (Wainwright '19)

— *inspired by SVRG (Johnson & Zhang '13)*

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

Variance-reduced Q-learning updates (Wainwright '19)

— *inspired by SVRG (Johnson & Zhang '13)*

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

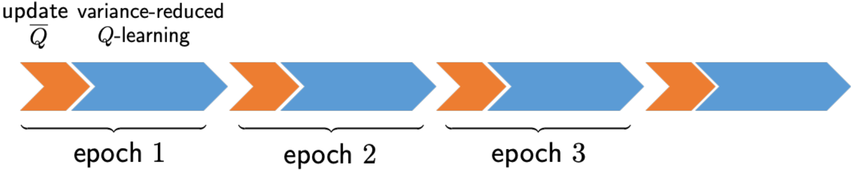
- $\overline{Q}$: some reference Q-estimate
- $\widetilde{\mathcal{T}}$: empirical Bellman operator (using a *batch* of samples)

$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

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

An epoch-based stochastic algorithm

— *inspired by Johnson & Zhang '13*



for each epoch

1. update $\overline{Q}$ and $\widetilde{\mathcal{T}}(\overline{Q})$ (which stay fixed in the rest of the epoch)
2. run variance-reduced Q-learning updates iteratively

Sample complexity of variance-reduced Q-learning

Theorem 4 (Wainwright '19)

For any $0 < \varepsilon \leq 1$, sample complexity for ***variance-reduced synchronous Q-learning*** *to yield $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$ is at most*

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

- allows for more aggressive learning rates

Sample complexity of variance-reduced Q-learning

Theorem 4 (Wainwright '19)

For any $0 < \varepsilon \leq 1$, sample complexity for ***variance-reduced synchronous Q-learning*** *to yield $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ is at most*

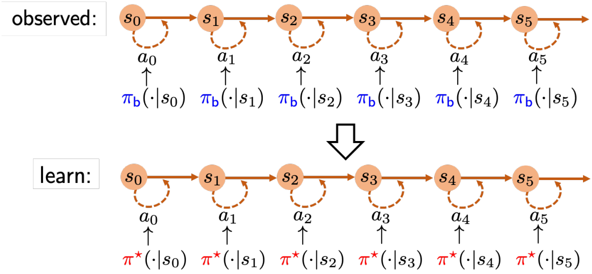
$$\widetilde{O}\Big(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\Big)$$

- allows for more aggressive learning rates
- minimax-optimal for $0 < \varepsilon \leq 1$
  - remains suboptimal if $1 < \varepsilon < \frac{1}{1-\gamma}$

Model-free RL

1. Basics of Q-learning
2. Synchronous Q-learning and variance reduction (simulator)
3. Asynchronous Q-learning (Markovian data)
4. Q-learning with lower confidence bounds (offline RL)
5. Q-learning with upper confidence bounds (online RL)

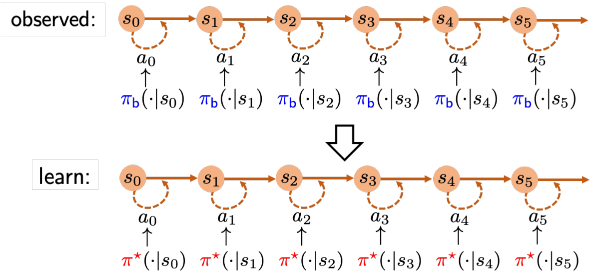
Markovian samples and behavior policy



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

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

Markovian samples and behavior policy



Key quantities of sample trajectory

- minimum state-action occupancy probability (uniform coverage)

$$\mu_{\min} := \min \underbrace{\mu_{\pi_b}(s,a)}_{\text{stationary distribution}} \quad \in \left[0, \frac{1}{|\mathcal{S}||\mathcal{A}|}\right]$$

- mixing time: t_{mix}

Q-learning on Markovian samples

Chris Watkins

Peter Dayan

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

Q-learning on Markovian samples

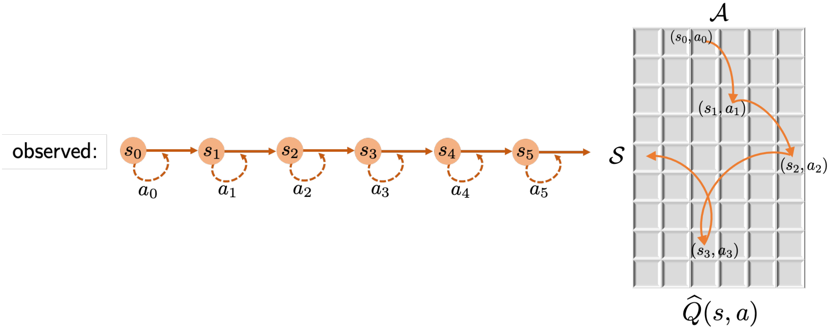
Chris Watkins

Peter Dayan

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

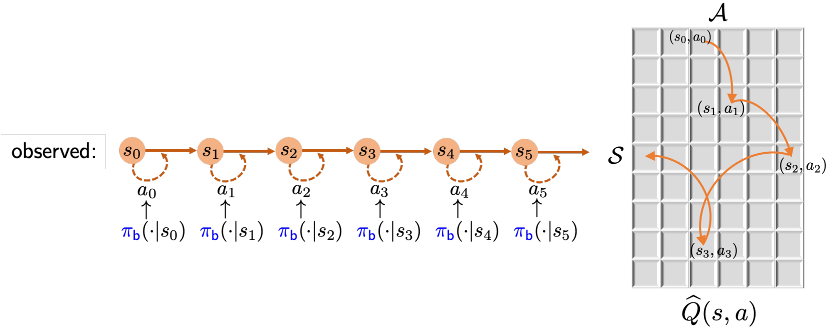
$$\mathcal{T}_t(Q)(s_t, a_t) = r(s_t, a_t) + \gamma \max_{a'} Q(s_{t+1}, a')$$

Q-learning on Markovian samples



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

Q-learning on Markovian samples



- **asynchronous:** only a single entry is updated each iteration
- **off-policy:** target policy $\pi^\star \neq$ behavior policy π_b

Sample complexity of asynchronous Q-learning

Theorem 5 (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$*, sample complexity of async Q-learning to yield* $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$ *with high prob. (or* $\mathbb{E}[\|\widehat{Q} - Q^{\star}\|_{\infty}] \leq \varepsilon$*) is at most*

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)} \qquad \textit{(up to log factor)}$$

Sample complexity of asynchronous Q-learning

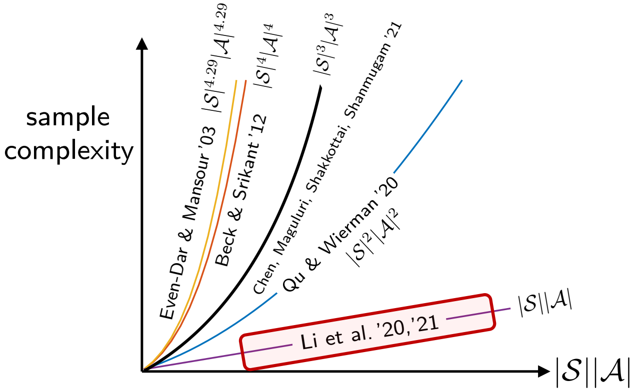
Theorem 5 (Li, Cai, Chen, Wei, Chi '21)

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$*, sample complexity of async Q-learning to yield* $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ *with high prob. (or* $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$*) is at most*

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)} \qquad \textit{(up to log factor)}$$

other papers	sample complexity
Even-Dar, Mansour '03	$\frac{(t_{\mathsf{cover}})^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\varepsilon^2}$
Even-Dar, Mansour '03	$\big(\frac{t_{\mathsf{cover}}^{1+3\omega}}{(1-\gamma)^4\varepsilon^2}\big)^{\frac{1}{\omega}} + \big(\frac{t_{\mathsf{cover}}}{1-\gamma}\big)^{\frac{1}{1-\omega}}$, $\omega \in (\frac{1}{2}, 1)$
Beck & Srikant '12	$\frac{t_{\mathsf{cover}}^3 \vert\mathcal{S}\vert\vert\mathcal{A}\vert}{(1-\gamma)^5\varepsilon^2}$
Qu & Wierman '20	$\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}$
Li, Wei, Chi, Gu, Chen '20	$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$
Chen, Maguluri, Shakkottai, Shanmugam '21	$\frac{1}{\mu_{\mathsf{min}}^3(1-\gamma)^5\varepsilon^2}$ + other-term(t_{mix})

Linear dependency on $1/\mu_{\text{min}}$

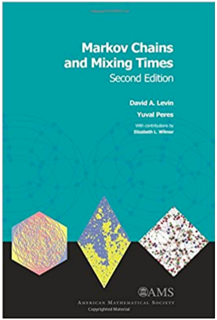


if we take $\mu_{\text{min}} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}$, $t_{\text{cover}} \asymp \frac{t_{\text{mix}}}{\mu_{\text{min}}}$

Effect of mixing time on sample complexity

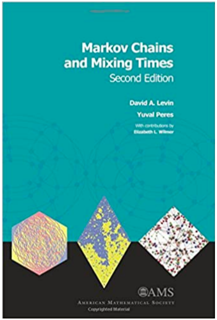
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

- reflects cost taken to reach steady state



Effect of mixing time on sample complexity

$$\frac{1}{\mu_{\min}(1-\gamma)^4\varepsilon^2} + \frac{t_{\text{mix}}}{\mu_{\min}(1-\gamma)}$$



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

  — it becomes amortized as algorithm runs

— *prior art:* $\frac{t_{\text{mix}}}{\mu_{\min}^2(1-\gamma)^5\varepsilon^2}$ (Qu & Wierman '20)

Model-free RL

1. Basics of Q-learning
2. Synchronous Q-learning and variance reduction (simulator)
3. Asynchronous Q-learning (Markovian data)
4. Q-learning with lower confidence bounds (offline RL)
5. Q-learning with upper confidence bounds (online RL)

Recap: 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}

Recap: offline RL / batch RL

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

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

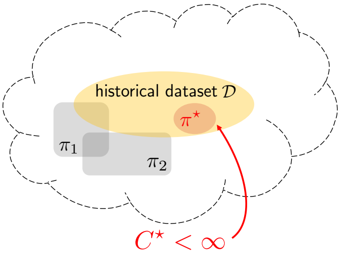
for some state distribution ρ^{b} and behavior policy π^{b}

Single-policy concentrability

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1$$

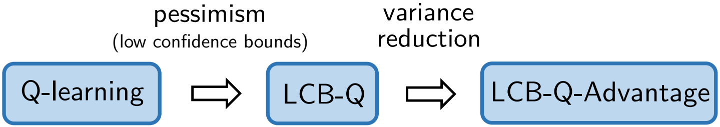
where d^π: occupancy distribution under π

- captures distributional shift
- allows for partial coverage



How to design offline model-free algorithms with optimal sample efficiency?

How to design offline model-free algorithms with optimal sample efficiency?



LCB-Q: Q-learning with LCB penalty

— *Shi et al. '22, Yan et al. '22*

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

LCB-Q: Q-learning with LCB penalty

— *Shi et al. '22, Yan et al. '22*

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

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

LCB-Q: Q-learning with LCB penalty

— *Shi et al. '22, Yan et al. '22*

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

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

sample size: $\widetilde{O}\left(\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}\right) \quad \Longrightarrow \quad$ 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. '22, Yan et al. '22*

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

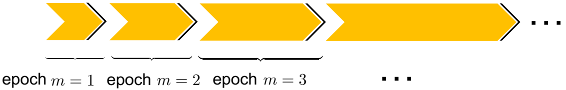
Q-learning with LCB and variance reduction

— *Shi et al. '22, Yan et al. '22*

$$Q_{t+1}(s_t, a_t) \leftarrow (1-\eta_t)Q_t(s_t, a_t) - \eta_t \underbrace{b_t(s_t, a_t)}_{\text{LCB penalty}}$$

$$+ \eta_t \Big(\underbrace{\mathcal{T}_t(Q_t) - \mathcal{T}_t(\overline{Q})}_{\text{advantage}} + \underbrace{\widehat{\mathcal{T}}(\overline{Q})}_{\text{reference}} \Big)(s_t, a_t)$$

- incorporates variance reduction into LCB-Q



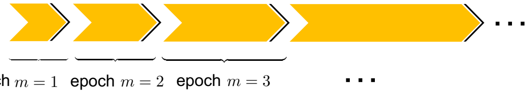
Q-learning with LCB and variance reduction

— *Shi et al. '22, Yan et al. '22*

$$Q_{t+1}(s_t,a_t) \leftarrow (1-\eta_t)Q_t(s_t,a_t) - \eta_t \underbrace{b_t(s_t,a_t)}_{\text{LCB penalty}}$$

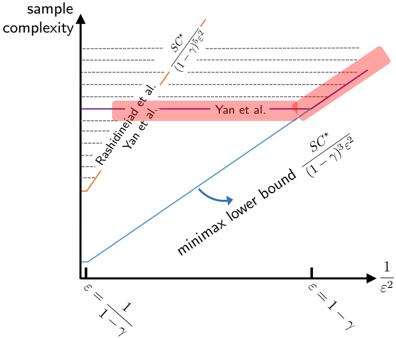
$$+ \eta_t\Big(\underbrace{\mathcal{T}_t(Q_t) - \mathcal{T}_t(\overline{Q})}_{\text{advantage}} + \underbrace{\widehat{\mathcal{T}}(\overline{Q})}_{\text{reference}}\Big)(s_t,a_t)$$

- incorporates variance reduction into LCB-Q

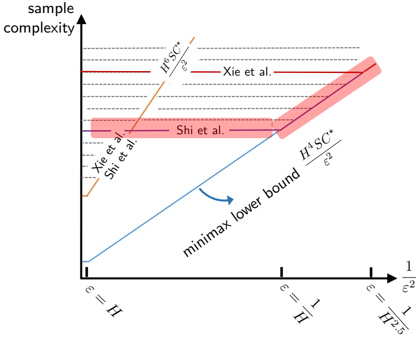


Theorem 6 (Yan, Li, Chen, Fan '22, Shi, Li, Wei, Chen, Chi '22)

For $\varepsilon \in (0, 1-\gamma]$, LCB-Q-Advantage *achieves* $V^\star(\rho) - V^{\widehat{\pi}}(\rho) \le \varepsilon$ *with optimal sample complexity* $\widetilde{O}\big(\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}\big)$



infinite-horizon MDPs



finite-horizon MDPs

Model-free offline RL attains sample optimality too!

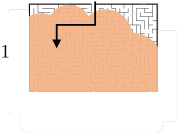
— *with some burn-in cost though . . .*

Model-free RL

1. Basics of Q-learning
2. Synchronous Q-learning and variance reduction (simulator)
3. Asynchronous Q-learning (Markovian data)
4. Q-learning with lower confidence bounds (offline RL)
5. Q-learning with upper confidence bounds (online RL)

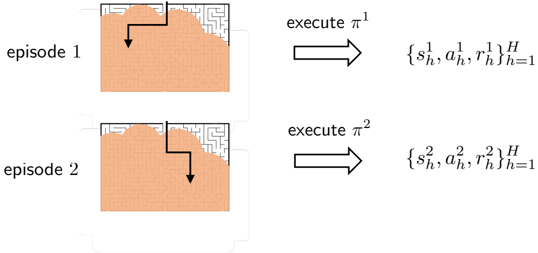
Online RL: interacting with real environments

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

episode 1  execute π^1 $\Longrightarrow$ $\{s_h^1, a_h^1, r_h^1\}_{h=1}^H$

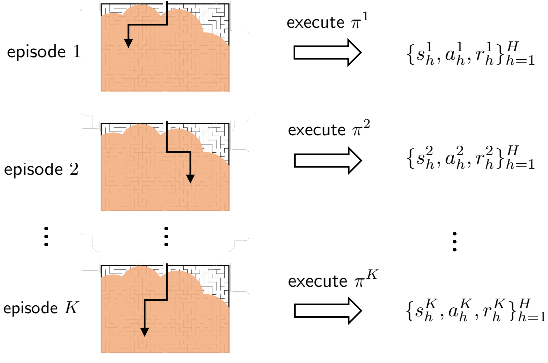
Online RL: interacting with real environments

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



Online RL: interacting with real environments

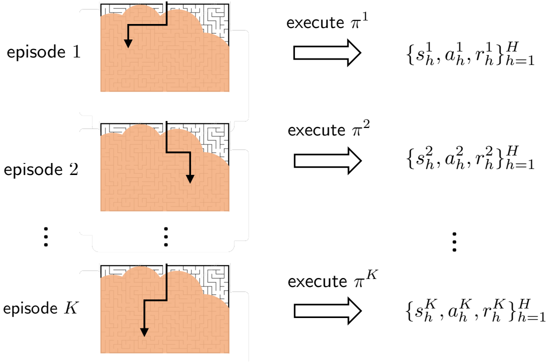
Sequentially execute MDP for K episodes, each consisting of H steps



Online RL: interacting with real environments

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

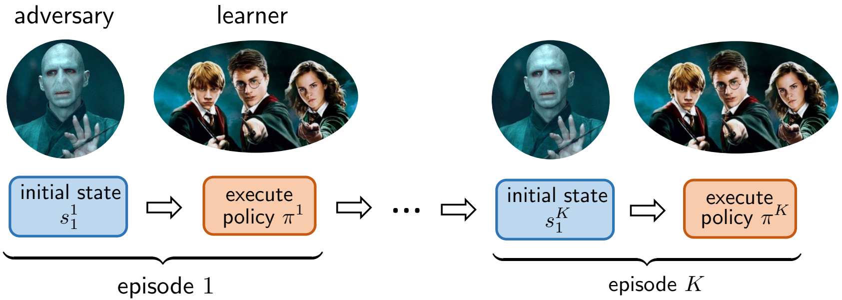
adversary

learner

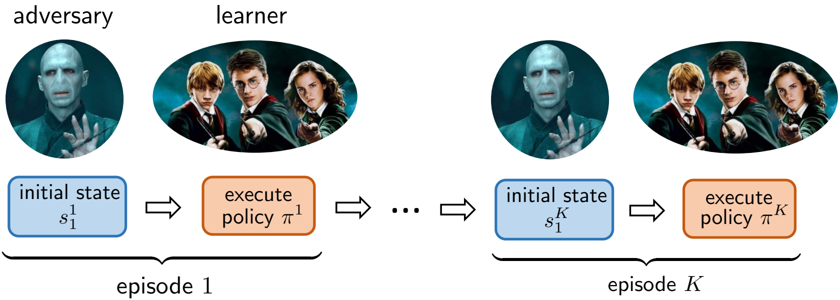
initial state s_1^1 ⇨ execute policy π^1

episode 1

Regret: gap between learned policy & optimal policy



Regret: gap between learned policy & optimal policy



Performance metric: given $\underbrace{\text{initial states } \{s_1^k\}_{k=1}^K}_{\text{chosen by nature/adversary}}$, define

$$\mathsf{Regret}(T) := \sum_{k=1}^{K} \Big(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) \Big)$$

Lower bound

(Domingues et al. '21)

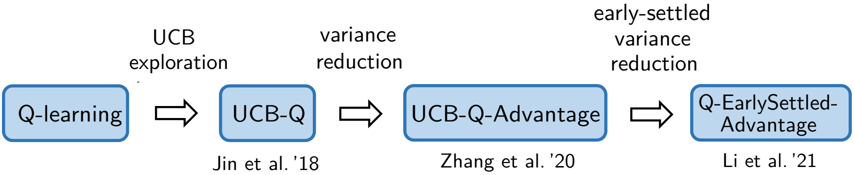
$$\text{Regret}(T) \gtrsim \sqrt{H^2SAT}$$

Existing algorithms

- UCB-VI: Azar et al. '17
- UBEV: Dann et al. '17
- UCB-Q-Hoeffding: Jin et al. '18
- **UCB-Q-Bernstein: Jin et al. '18**
- UCB2-Q-Bernstein: Bai et al. '19
- EULER: Zanette et al. '19
- **UCB-Q-Advantage: Zhang et al. '20**
- UCB-M-Q: Menard et al. '21
- **Q-EarlySettled-Advantage: Li et al. '21**

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

Issue: *large variability in stochastic update rules*

UCB Q-learning with UCB and variance reduction

Incorporates variance reduction into UCB-Q: — *Zhang, Zhou, Ji '20*

- asymptotically regret-optimal

UCB Q-learning with UCB and variance reduction

Incorporates variance reduction into UCB-Q: — *Zhang, Zhou, Ji '20*

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

UCB Q-learning with UCB and variance reduction

Incorporates variance reduction into UCB-Q: — *Zhang, Zhou, Ji '20*

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

One additional idea: early settlement of reference updates — *Li, Shi, Chen, Chi '23*

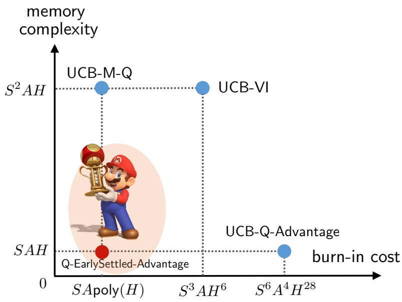
UCB Q-learning with UCB and variance reduction

Incorporates variance reduction into UCB-Q: — *Zhang, Zhou, Ji '20*

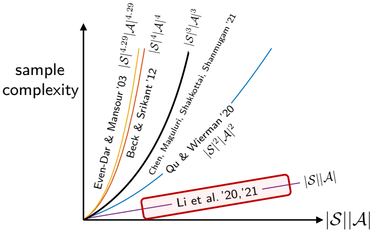
- asymptotically regret-optimal
- ***Issue:*** *high burn-in cost* $O(S^6A^4H^{28})$

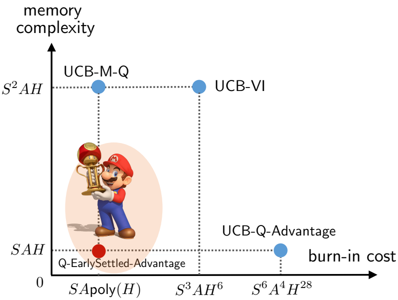
One additional idea: early settlement of reference updates — *Li, Shi, Chen, Chi '23*

- regret-optimal w/ near-minimal burn-in cost in S and A
- memory-efficient $O(SAH)$
- computationally efficient: runtime $O(T)$



Summary of this part





Model-free RL can achieve memory efficiency, computational efficiency, and sample efficiency at once!

— *with some burn-in cost though*

Reference I

- "*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* 2022
- "*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
- "*The complexity of Markov equilibrium in stochastic games,*" C. Daskalakis, N. Golowich, K. Zhang, *COLT*, 2023
- "*A stochastic approximation method,*" H. Robbins, S. Monro, *Annals of mathematical statistics*, 1951

Reference II

- "*Robust stochastic approximation approach to stochastic programming,*" A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, *SIAM Journal on optimization*, 2009
- "*Learning from delayed rewards,*" C. Watkins, 1989
- "*Q-learning,*" C. Watkins, P. Dayan, *Machine learning*, 1992
- "*Learning to predict by the methods of temporal differences,*" R. Sutton, *Machine learning*, 1988
- "*Analysis of temporal-diffference learning with function approximation,*" B. van Roy, J. Tsitsiklis, *IEEE transactions on automatic control*, 1997
- "*Learning Rates for Q-learning,*" E. Even-Dar, Y. Mansour, *Journal of machine learning Research*, 2003
- "*The asymptotic convergence-rate of Q-learning,*" C. Szepesvari, *NeurIPS*, 1998

Reference III

- "*Stochastic approximation with cone-contractive operators: Sharp ℓ_∞ bounds for Q-learning,*" M. Wainwright, arXiv:1905.06265, 2019
- "*Is Q-Learning minimax optimal? A tight sample complexity analysis,*" G. Li, Y. Wei, Y. Chi, Y. Chen, accepted to *Operations Research*, 2023
- "*Accelerating stochastic gradient descent using predictive variance reduction*," R. Johnson, T. Zhang, *NeurIPS*, 2013
- "*Variance-reduced Q-learning is minimax optimal,*" M. Wainwright, arXiv:1906.04697, 2019
- "*Asynchronous stochastic approximation and Q-learning,*" J. Tsitsiklis, *Machine learning*, 1994
- "*On the convergence of stochastic iterative dynamic programming algorithms,*" T. Jaakkola, M. Jordan, S. Singh, *Neural computation*, 1994

Reference IV

- "*Error bounds for constant step-size Q-learning,*" C. Beck, R. Srikant, *Systems and control letters*, 2012
- "*Sample complexity of asynchronous Q-learning: sharper analysis and variance reduction,*" G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, *NeurIPS* 2020
- "*Finite-time analysis of asynchronous stochastic approximation and Q-learning,*" G. Qu, A. Wierman, *COLT* 2020.
- "*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, arXiv:2203.07368, 2022.
- "*Asymptotically efficient adaptive allocation rules*," T. L. Lai, H. Robbins, *Advances in applied mathematics*, vol. 6, no. 1, 1985.

Reference V

- "*Is Q-learning provably efficient?*" C. Jin, Z. Allen-Zhu, S. Bubeck, and M. Jordan, *NeurIPS* 2018.
- "*Almost optimal model-free reinforcement learning via reference-advantage decomposition,*" Z. Zhang, Y. Zhou, X. Ji, *NeurIPS* 2020.
- "*Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning,*" G. Li, L. Shi, Y. Chen, Y. Chi, *Information and Inference: A Journal of the IMA*, 2023.