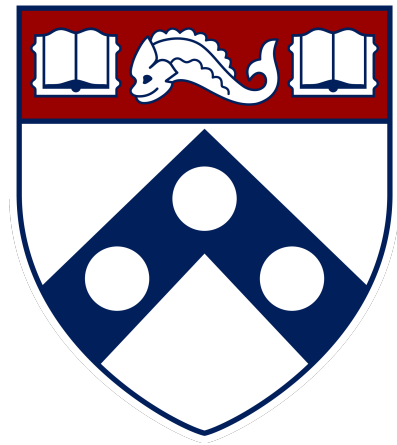
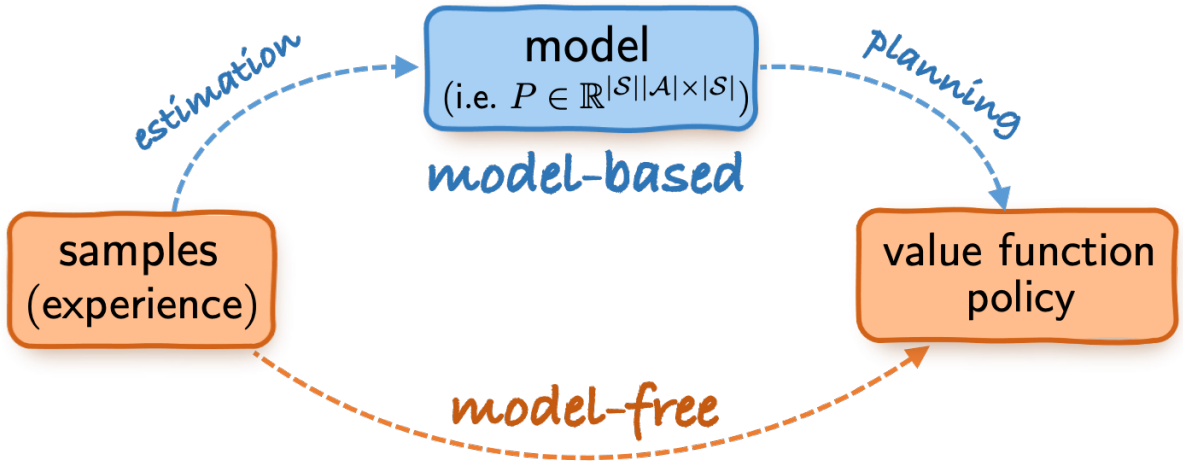
Reinforcement learning (Part 2): Model-free RL



Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

Model-based vs. model-free RL



Model-based approach ("plug-in")

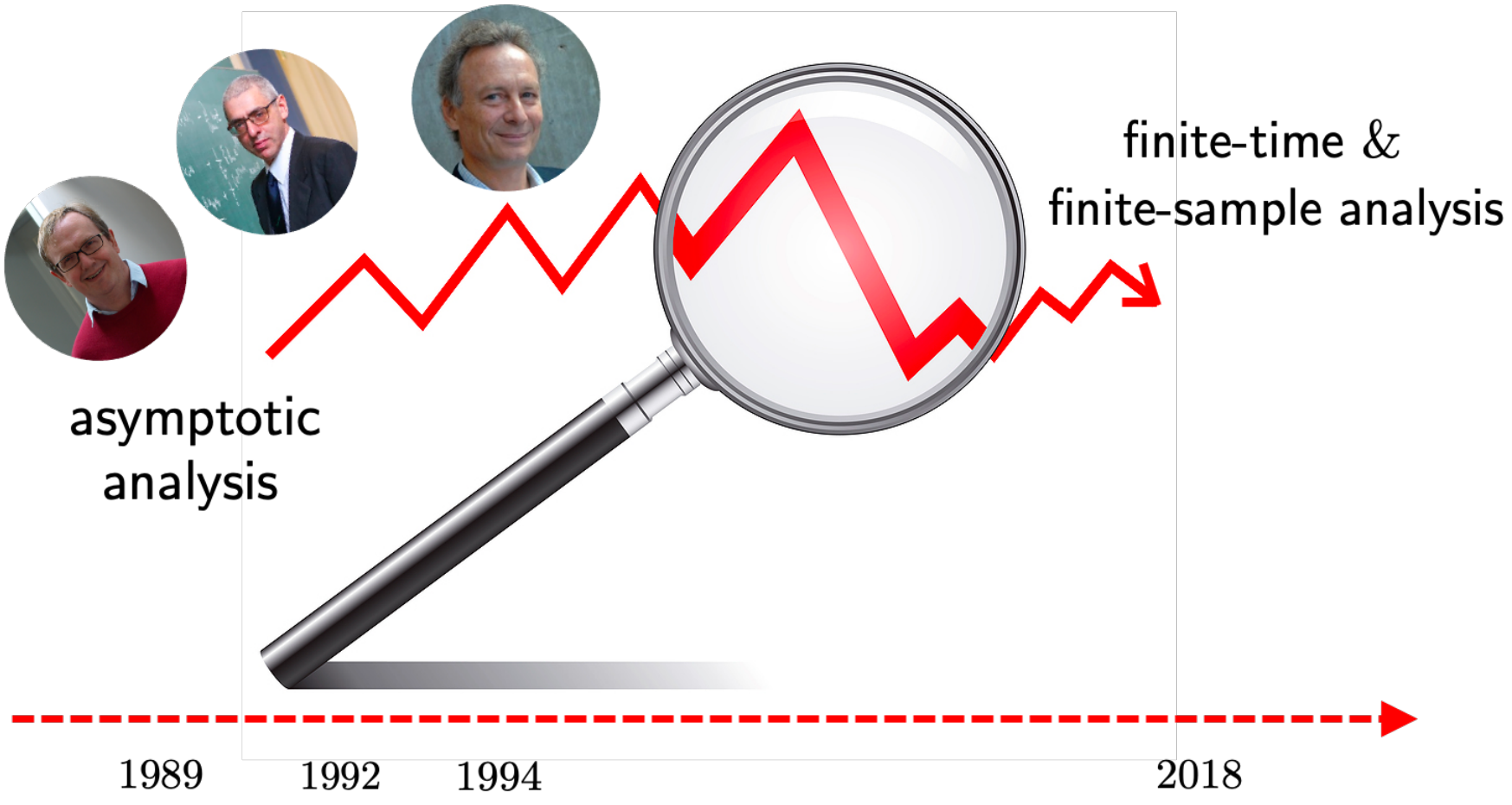
1. build empirical estimate $\widehat{P}$ for P
2. planning based on empirical $\widehat{P}$

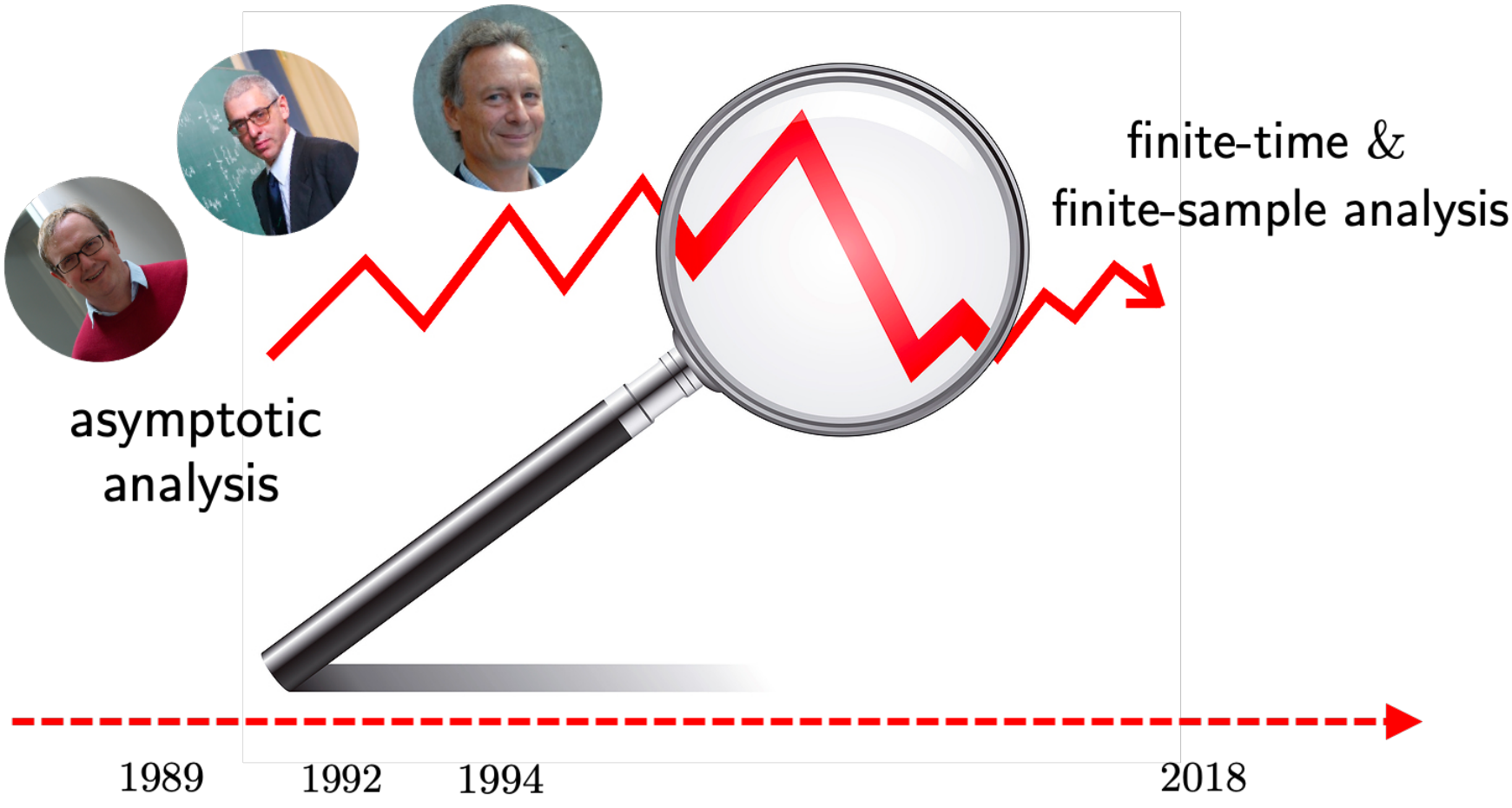
Model-free approach

— learning w/o modeling & estimating environment explicitly

— memory-efficient, online, ...

Is model-free RL minimax optimal?


asymptotic
analysis
finite-time &
finite-sample analysis
1989
1992
1994
2018



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

A starting point: Bellman optimality principle

Bellman operator

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

- one-step look-ahead

A starting point: Bellman optimality principle

Bellman operator

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

- 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

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

- 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

A detour: stochastic approximation

- **Goal:** solve
$$\boldsymbol{G}\left(\boldsymbol{x}\right) = \mathbb{E}\big[\boldsymbol{g}\left(\boldsymbol{x};\boldsymbol{\xi}\right)\big] = \mathbf{0}$$
  - $\boldsymbol{\xi}$: randomness in problem

- **What we can query:** for any given input $\boldsymbol{x}$, we receive a *random* sample $\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\xi})$ obeying $\mathbb{E}\big[\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\xi})\big] = \boldsymbol{G}(\boldsymbol{x})$

Stochastic approximation (Robbins, Monro '51)

Herbert Robbins

Sutton Monro

stochastic approximation

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \, \boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t) \tag{1}$$

where $\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t)$ is *unbiased* estimate of $\boldsymbol{G}(\boldsymbol{x}^t)$, i.e.

$$\mathbb{E}\big[\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t)\big] = \boldsymbol{G}(\boldsymbol{x}^t)$$

Stochastic approximation (Robbins, Monro '51)

Herbert Robbins

Sutton Monro

stochastic approximation

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \, \boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t) \tag{1}$$

a stochastic algorithm for finding roots of $\boldsymbol{G}(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\xi})]$

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving the **Bellman equation**

Robbins & Monro, 1951

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

where

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

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

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

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

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

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

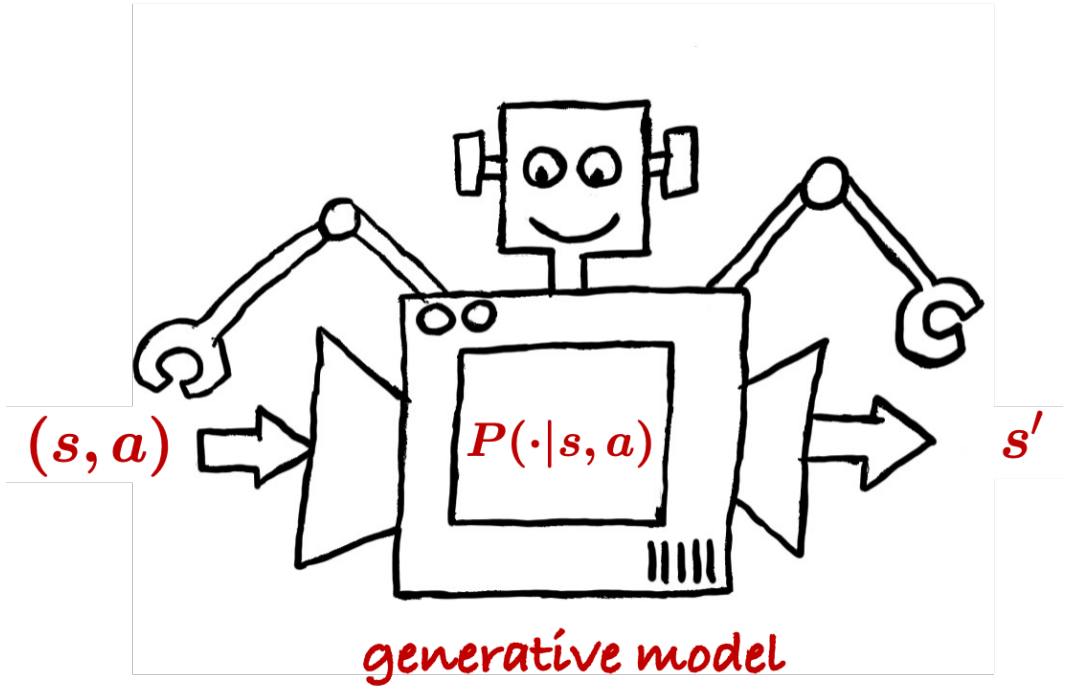
$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

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

Synchronous Q-learning

Sampling from a generative model

— *Kearns, Singh '99*



In each iteration, collect an independent sample (s, a, s') for each (s, a)

Synchronous Q-learning

Chris Watkins

Peter Dayan

for $t = 0, 1, \ldots$

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

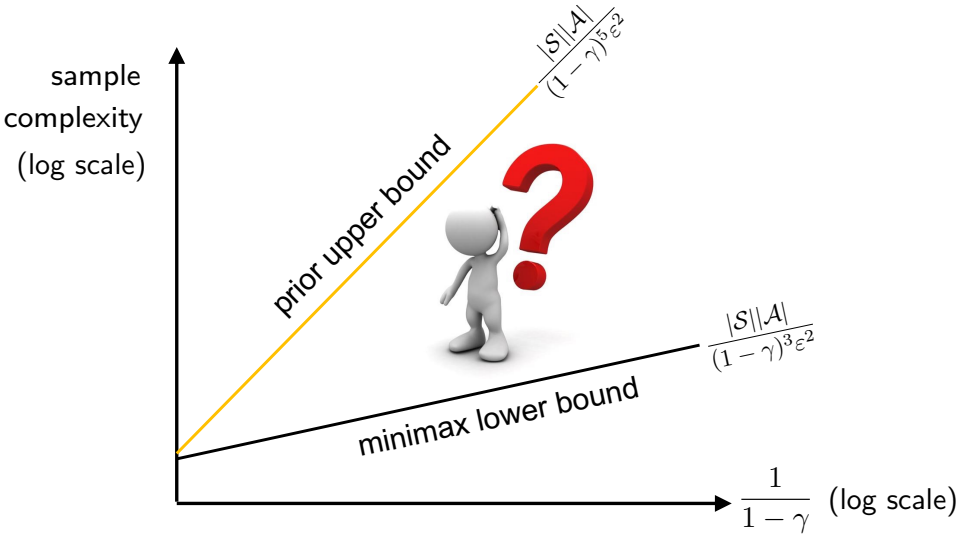
Prior art: achievability

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

Prior art: achievability

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

paper	sample complexity
Even-Dar & Mansour '03	$2^{\frac{1}{1-\gamma}} \frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
Beck & Srikant '12	$\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^5\varepsilon^2}$
Wainwright '19	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
Chen et al. '20	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$

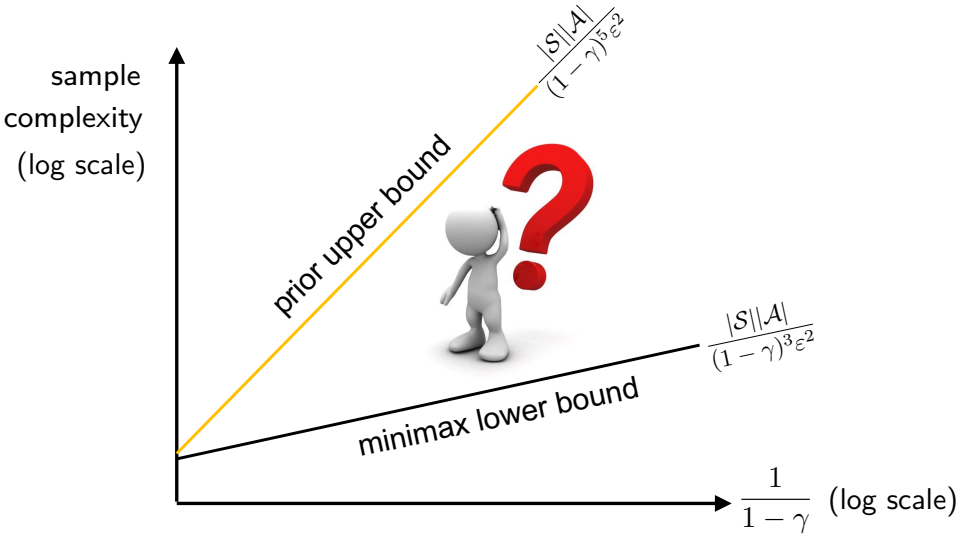


All prior results require sample size of at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$!

Prior art: achievability

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

paper	sample complexity
Even-Dar & Mansour '03	$2^{\frac{1}{1-\gamma}} \frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$
Beck & Srikant '12	$\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^5\varepsilon^2}$
Wainwright '19	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$
Chen et al. '20	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$



All prior results require sample size of at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$!

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

Sample complexity of Q-learning

Theorem 1 (Li, Cai, Chen, Gu, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, *Q-learning yields*

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

with sample complexity at most

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

- Improves dependency on effective horizon $\frac{1}{1-\gamma}$

Sample complexity of Q-learning

Theorem 1 (Li, Cai, Chen, Gu, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, *Q-learning yields*

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

with sample complexity at most

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

- Improves dependency on effective horizon $\frac{1}{1-\gamma}$
- Covers both constant and rescaled linear learning rates:

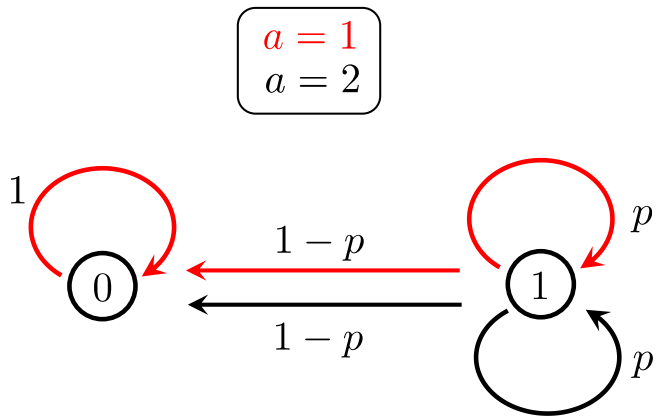
$$\eta_t \equiv \frac{1}{1+\frac{c_1(1-\gamma)T}{\log^2 T}} \qquad \text{or} \qquad \eta_t = \frac{1}{1+\frac{c_2(1-\gamma)t}{\log^2 T}}$$

How sharp is sample complexity bound $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}\right)$*?*

A curious numerical example

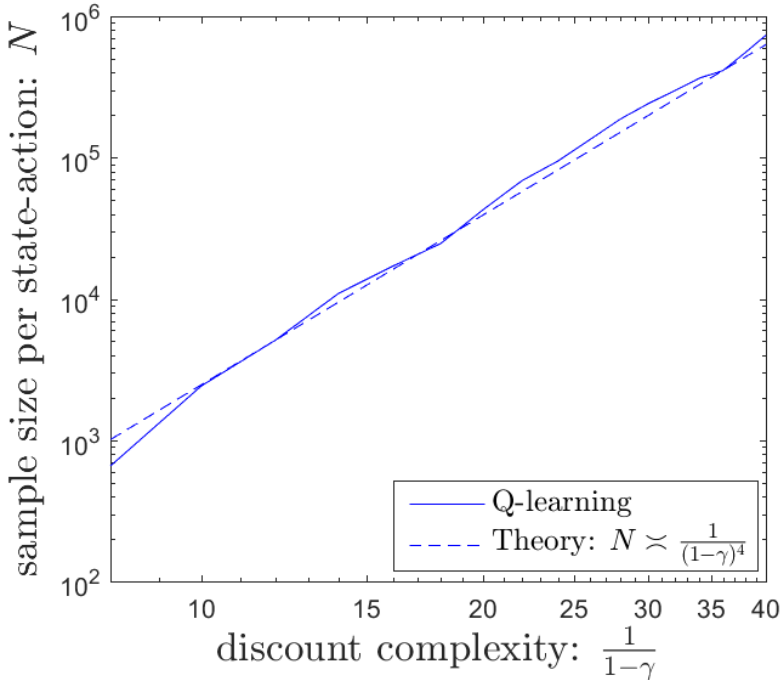
Numerical evidence: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ samples seem necessary ...

— 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 2 (Li, Cai, Chen, Gu, Wei, Chi, 2021)

For any $0 < \varepsilon \leq 1$, there exist an MDP such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, Q-learning needs at least a sample complexity of

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

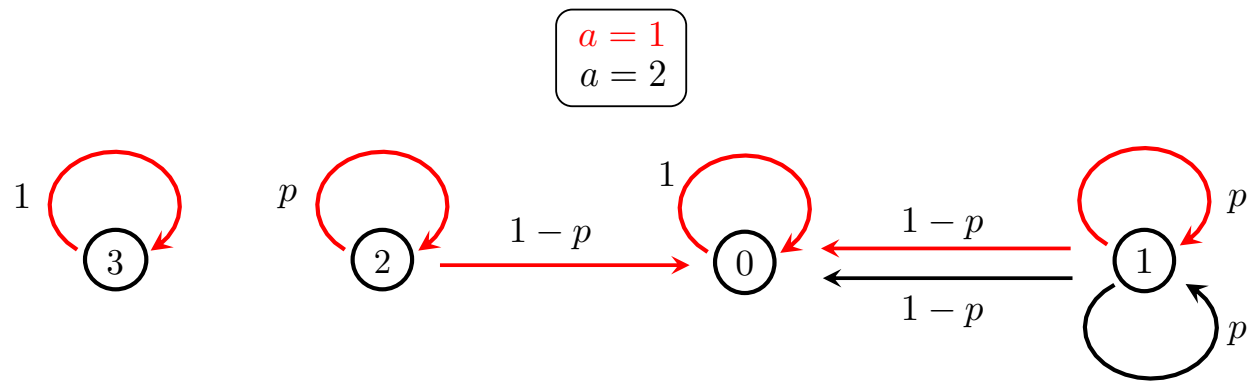
Q-learning is NOT minimax optimal

Theorem 2 (Li, Cai, Chen, Gu, Wei, Chi, 2021)

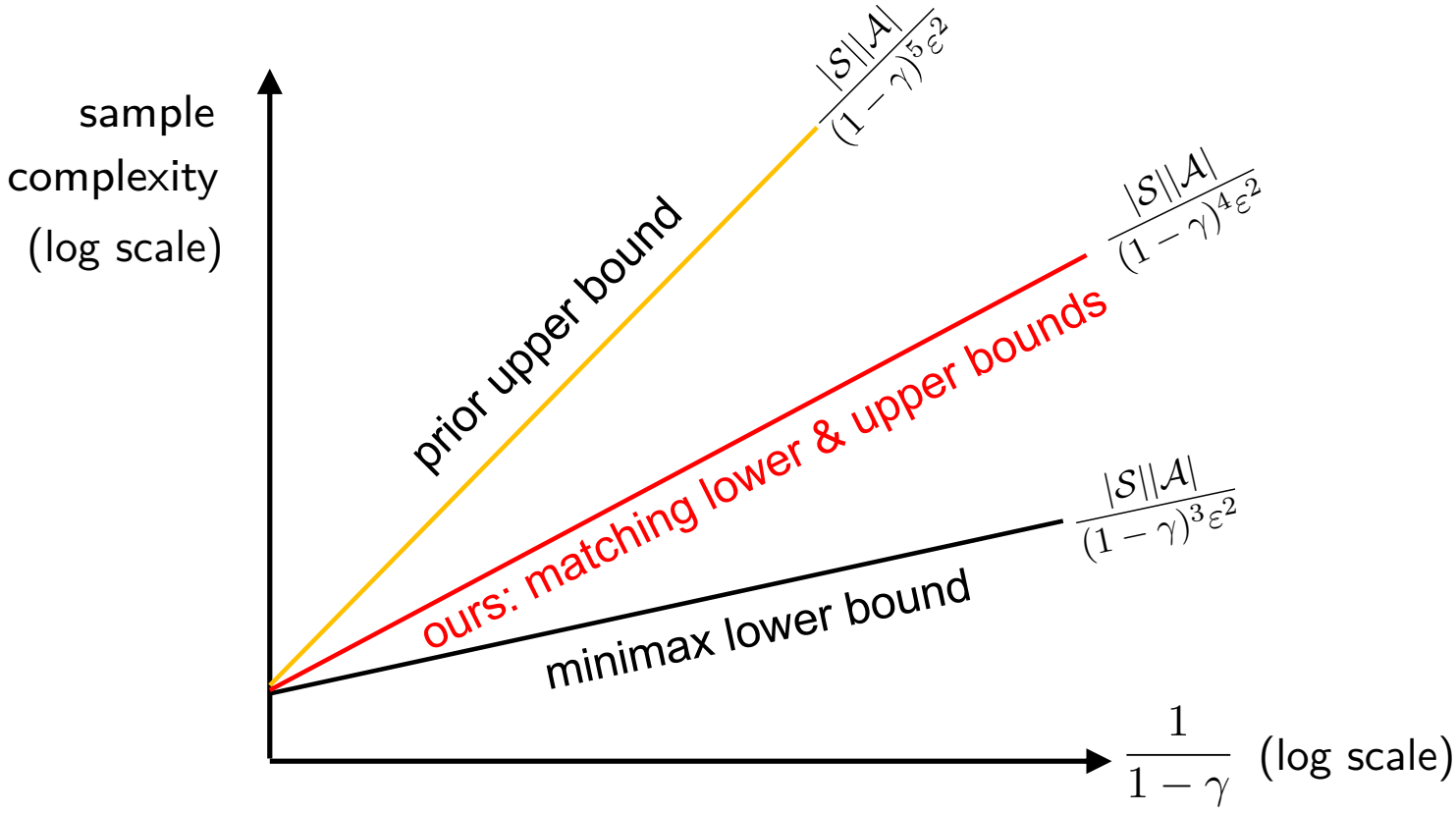
For any $0 < \varepsilon \leq 1$, there exist an MDP such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, Q-learning needs at least a sample complexity of

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

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



Where we stand now



Q-learning requires a sample size of $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$.

Why is Q-learning sub-optimal?

Over-estimation of Q-functions (Thrun and Schwartz, 1993; Hasselt, 2010):

- $\max_{a \in \mathcal{A}} \mathbb{E}[X(a)]$ tends to be over-estimated (high positive bias) when $\mathbb{E}[X(a)]$ is replaced by its empirical estimates using a small sample size;
- often gets worse with a large number of actions (Hasselt, Guez, Silver, 2015).

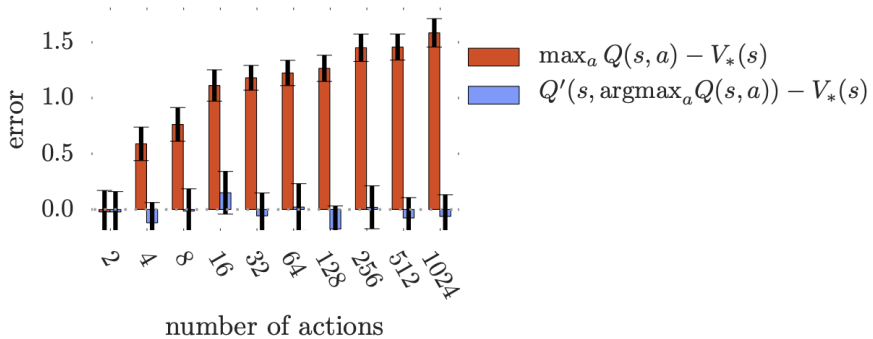


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically and independently. All bars are the average of 100 repetitions.

Why is Q-learning sub-optimal?

Over-estimation of Q-functions (Thrun and Schwartz, 1993; Hasselt, 2010):

- $\max_{a \in \mathcal{A}} \mathbb{E}[X(a)]$ tends to be over-estimated (high positive bias) when $\mathbb{E}[X(a)]$ is replaced by its empirical estimates using a small sample size;
- often gets worse with a large number of actions (Hasselt, Guez, Silver, 2015).

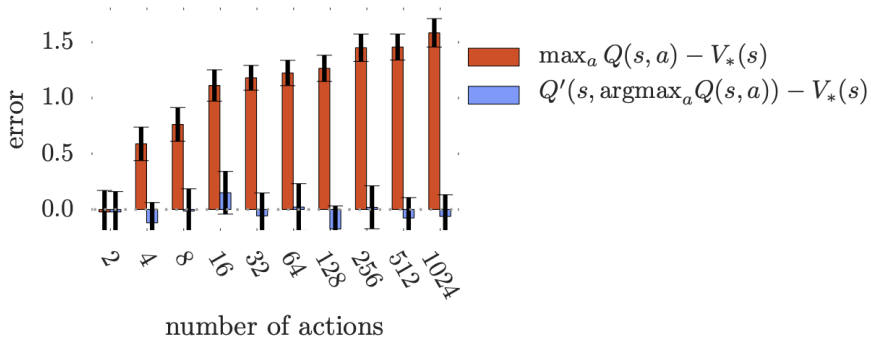


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically and independently. All bars are the average of 100 repetitions.

A provable fix: Q-learning with variance reduction (Wainwright 2019) is *provably* minimax optimal.

Variance-reduced Q-learning

Back to Q-learning ...

— inspired by Johnson & Zhang '13

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

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

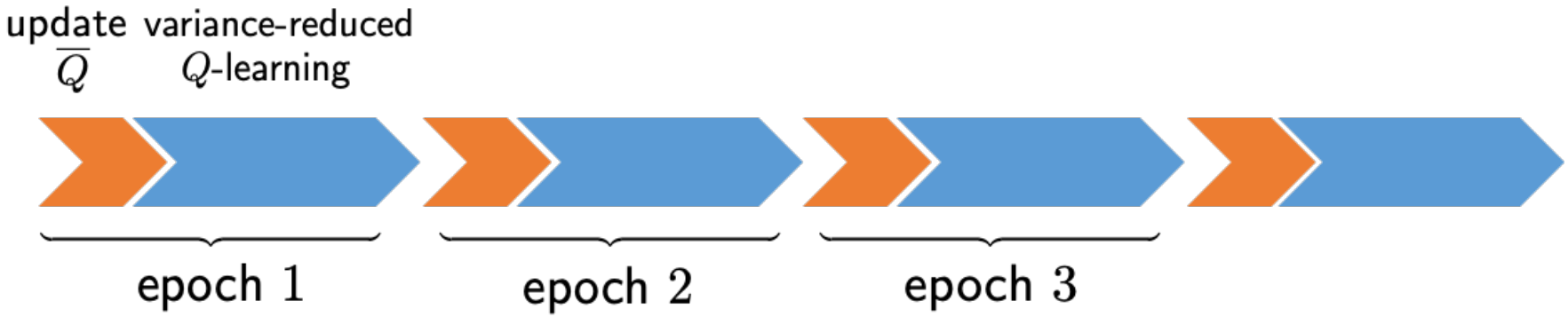
- $\overline{Q}$: some reference Q-estimate
- $\widetilde{\mathcal{T}}$: empirical Bellman operator (using a batch of samples)

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

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

Variance-reduced Q-learning

— *inspired by Johnson & Zhang '13*



for each epoch

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

Main result: ℓ_∞-based sample complexity

Theorem 3 (Wainwright '19)

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

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

- more aggressive learning rates: $\eta_t \equiv \frac{\cancel{(1-\gamma)^4}(1-\gamma)^2}{\gamma^2}$

Main result: ℓ_∞-based sample complexity

Theorem 3 (Wainwright '19)

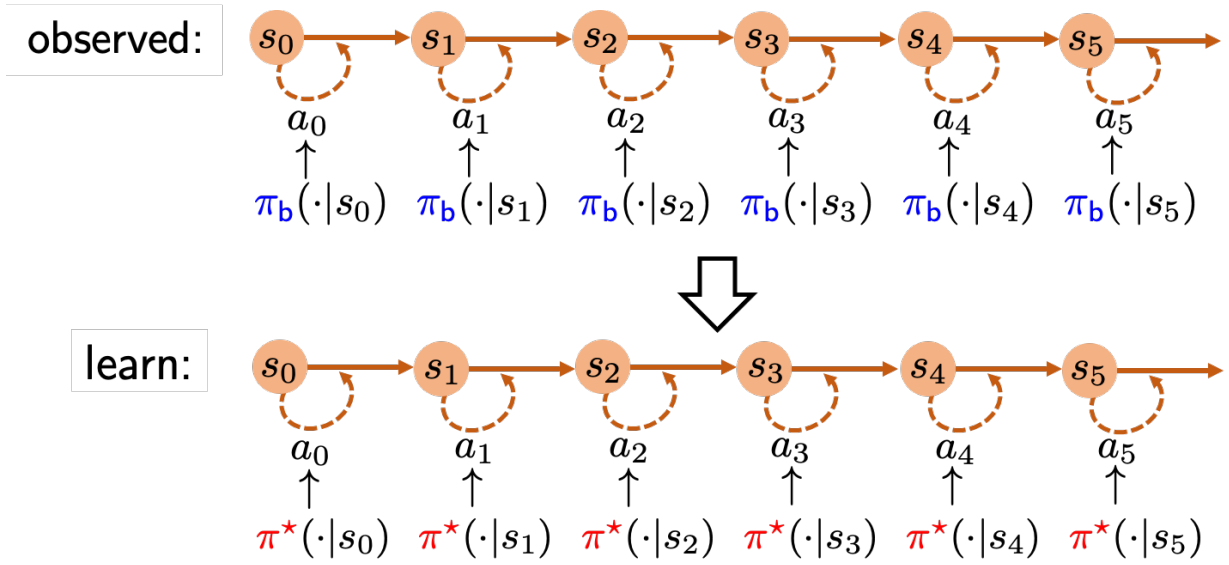
*For any $0 < \varepsilon \leq 1$, sample complexity for **variance-reduced Q-learning** to yield $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ is at most on the order of*

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

- more aggressive learning rates: $\eta_t \equiv \frac{\cancel{(1-\gamma)^4}(1-\gamma)^2}{\gamma^2}$
- minimax-optimal for $0 < \varepsilon \leq 1$
  - suboptimal if $1 < \varepsilon < \frac{1}{1-\gamma}$

Asynchronous Q-learning (on Markovian samples)

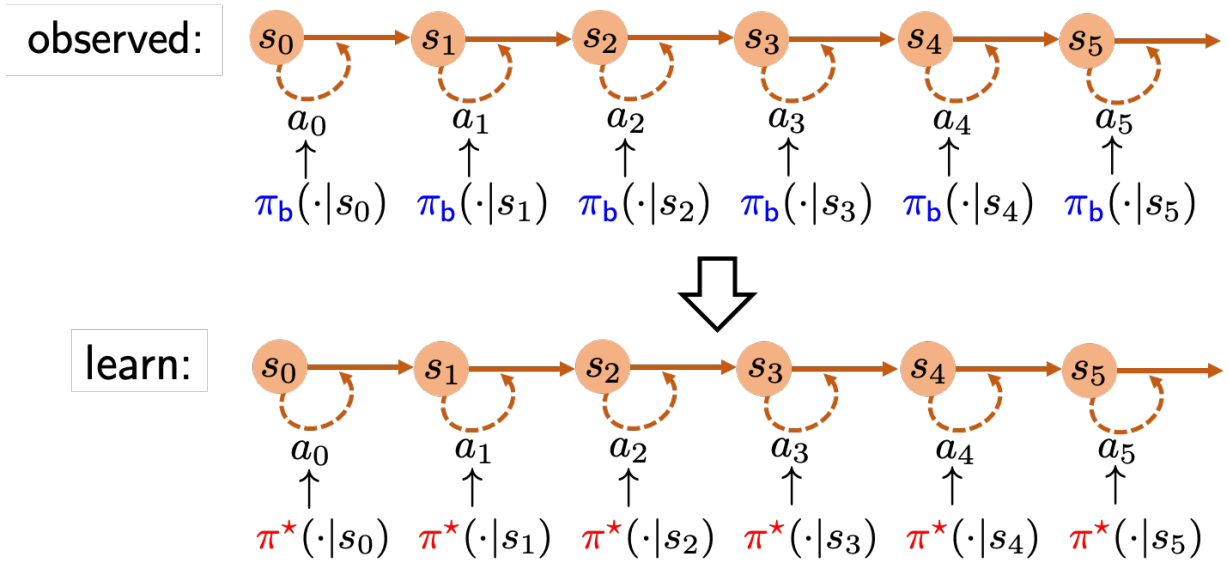
Markovian samples and behavior policy



Observed: $\underbrace{\{s_t, a_t, r_t\}_{t\geq 0}}_{\text{Markovian trajectory}}$ 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

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

- mixing time: t_{mix}

Q-learning on Markovian samples

Chris Watkins

Peter Dayan

$$\underbrace{Q_{t+1}(s_t,a_t) = (1-\eta_t)Q_t(s_t,a_t) + \eta_t \mathcal{T}_t(Q_t)(s_t,a_t)}, \quad t \geq 0$$

only update (s_t,a_t)-th entry

Q-learning on Markovian samples

Chris Watkins

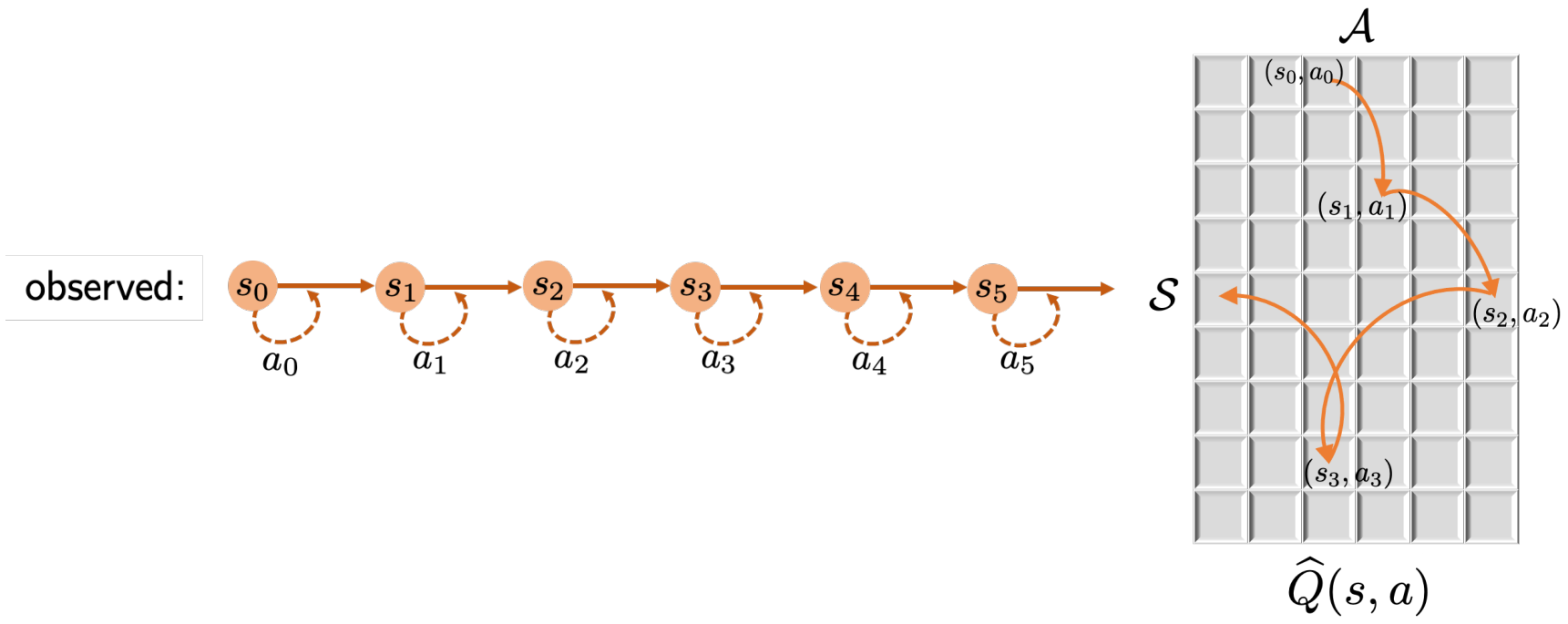
Peter Dayan

$$\underbrace{Q_{t+1}(s_t,a_t) = (1-\eta_t)Q_t(s_t,a_t) + \eta_t \mathcal{T}_t(Q_t)(s_t,a_t)}, \quad t \geq 0$$

only update (s_t,a_t)-th entry

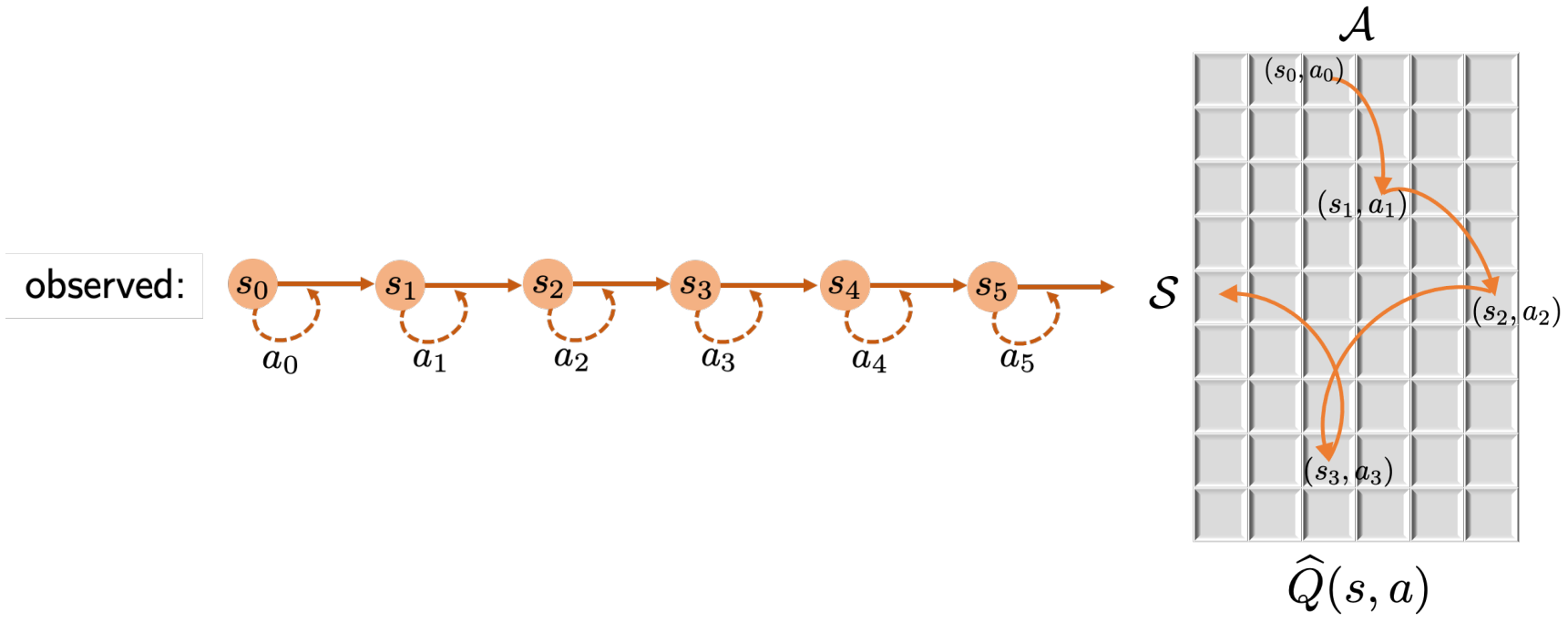
$$\mathcal{T}_t(Q)(s_t,a_t) = r(s_t,a_t) + \gamma \max_{a'} Q(s_{t+1},a')$$

Q-learning on Markovian samples



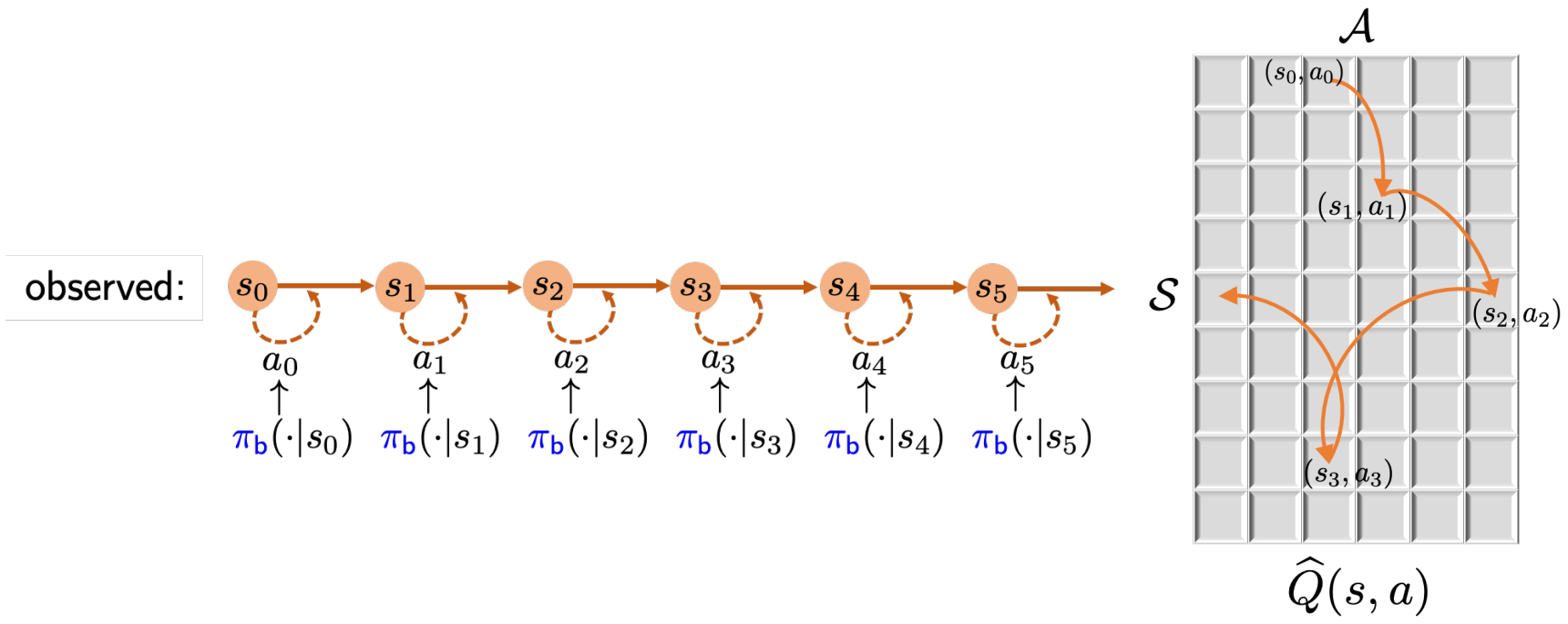
- **asynchronous:** only a single entry is updated each iteration

Q-learning on Markovian samples



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

Q-learning on Markovian samples



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

A highly incomplete list of prior work

- Watkins, Dayan '92
- Tsitsiklis '94
- Jaakkola, Jordan, Singh '94
- Szepesvári '98
- Kearns, Singh '99
- Borkar, Meyn '00
- Even-Dar, Mansour '03
- Beck, Srikant '12
- Chi, Zhu, Bubeck, Jordan '18
- Shah, Xie '18
- Lee, He '18
- Wainwright '19
- Chen, Zhang, Doan, Maguluri, Clarke '19
- Yang, Wang '19
- Du, Lee, Mahajan, Wang '20
- Chen, Maguluri, Shakkottai, Shanmugam '20
- Qu, Wierman '20
- Devraj, Meyn '20
- Weng, Gupta, He, Ying, Srikant '20
- ...

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

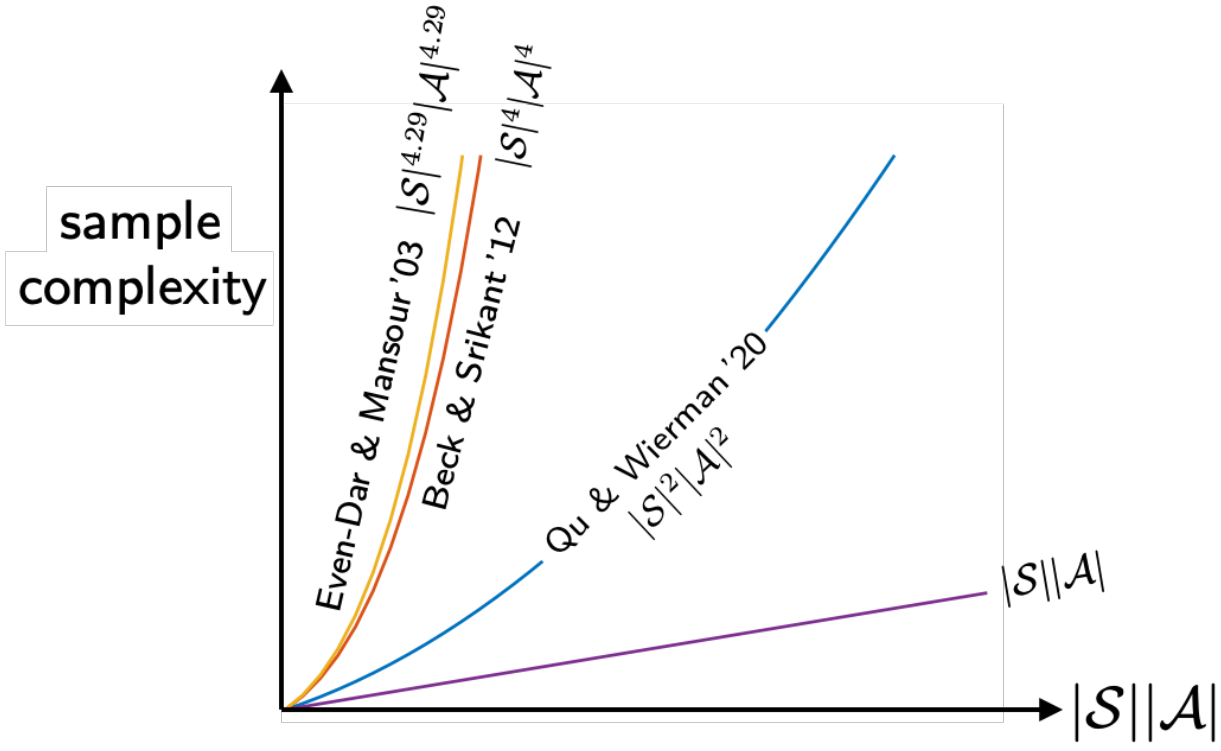
Prior art: async Q-learning

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

paper	sample complexity	learning rate
Even-Dar & Mansour '03	$\frac{(t_{\mathsf{cover}})^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\varepsilon^2}$	linear: $\frac{1}{t}$
Even-Dar & Mansour '03	$\left(\frac{t_{\mathsf{cover}}^{1+3\omega}}{(1-\gamma)^4\varepsilon^2}\right)^{\frac{1}{\omega}} + \left(\frac{t_{\mathsf{cover}}}{1-\gamma}\right)^{\frac{1}{1-\omega}}$	poly: $\frac{1}{t^\omega}$, $\omega \in (\frac{1}{2}, 1)$
Beck & Srikant '12	$\frac{t_{\mathsf{cover}}^3 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\varepsilon^2}$	constant
Qu & Wierman '20	$\frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}^2(1-\gamma)^5\varepsilon^2}$	rescaled linear

Prior art: async Q-learning

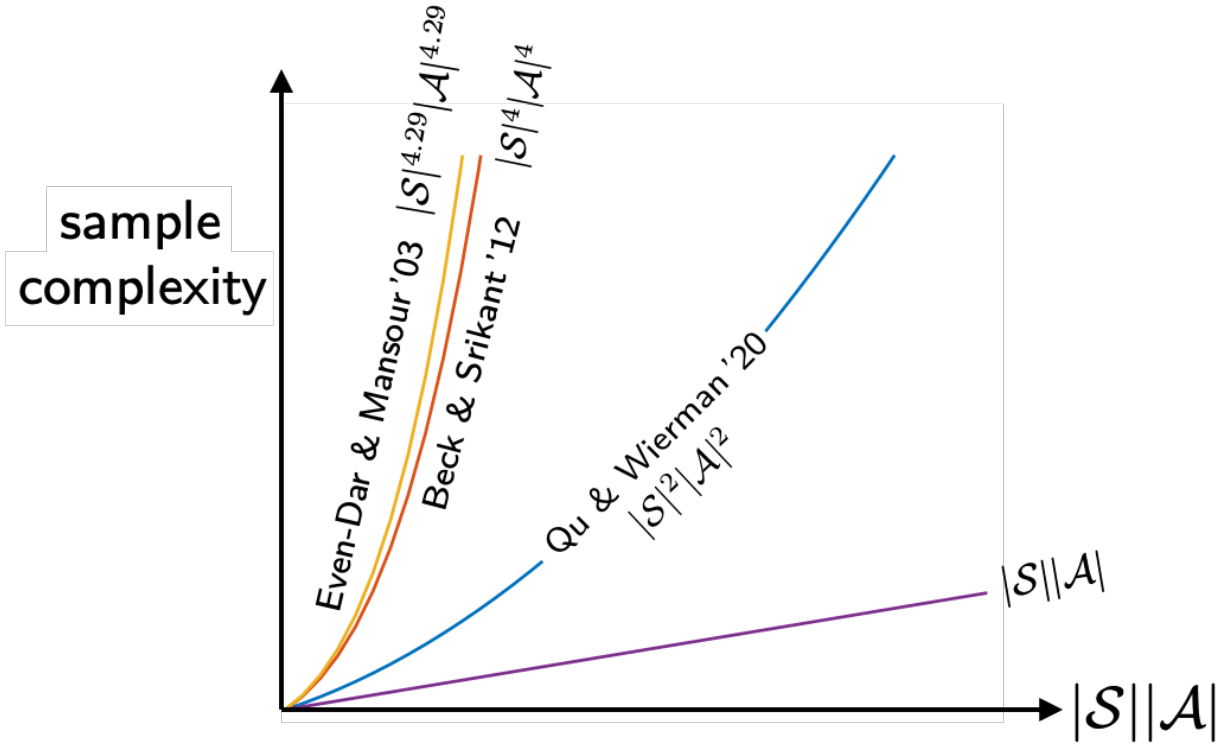
Question: how many samples are needed to ensure $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$?



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

Prior art: async Q-learning

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

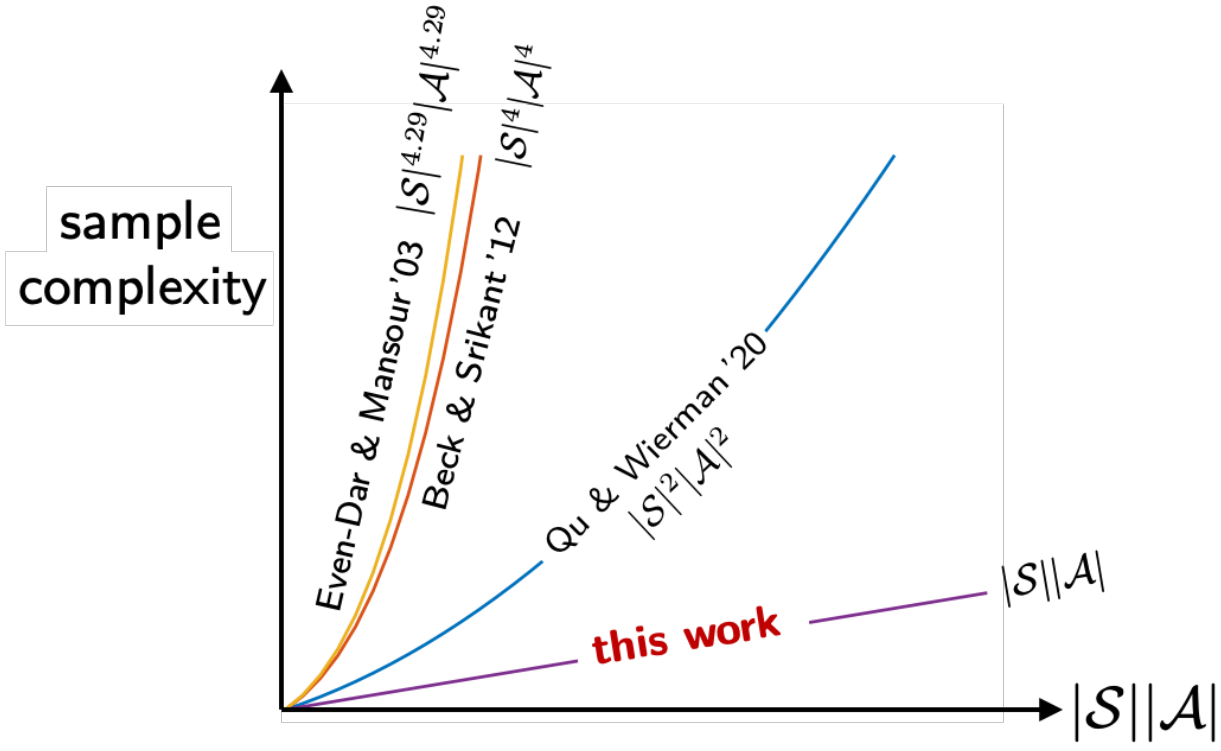


if we take $\mu_{\min} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}$, $t_{\text{cover}} \asymp \frac{t_{\text{mix}}}{\mu_{\min}}$

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

Prior art: async Q-learning

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



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

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

Main result: ℓ_∞-based sample complexity

Theorem 4 (Li, Cai, Chen, Gu, Wei, Chi '21)

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

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

Main result: ℓ_∞-based sample complexity

Theorem 4 (Li, Cai, Chen, Gu, Wei, Chi '21)

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

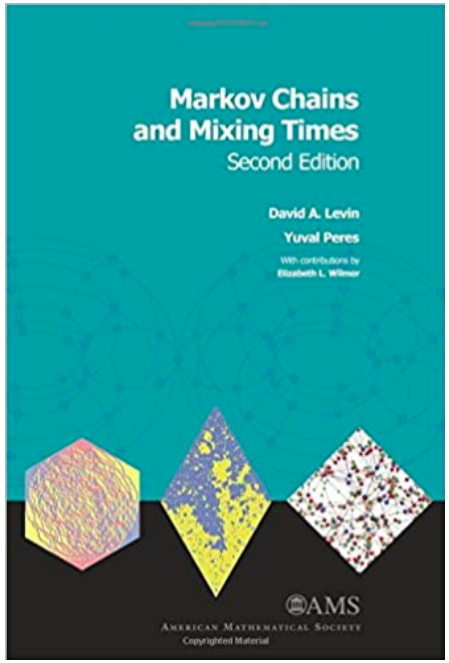
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

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

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

Effect of mixing time on sample complexity

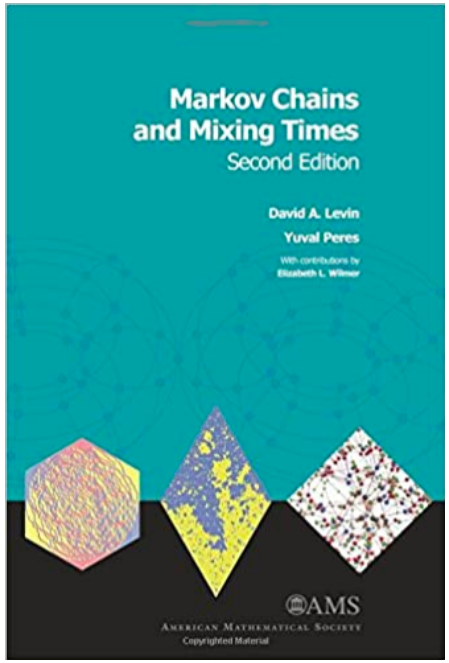
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^4\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$



- reflects cost taken to reach steady state
- one-time expense (almost independent of ε)
  — it becomes amortized as algorithm runs
- can be improved with the aid of variance reduction (Li et al. '20)

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
- one-time expense (almost independent of ε)
  — it becomes amortized as algorithm runs
- can be improved with the aid of variance reduction (Li et al. '20)

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

Reference I

- "*A stochastic approximation method,*" H. Robbins, S. Monro, *Annals of mathematical statistics*, 1951
- "*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

Reference II

- "*The asymptotic convergence-rate of Q-learning,*" C. Szepesvari, *NeurIPS*, 1998
- "*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. Gu, Y. Chen, arXiv:2102.06548, 2021
- "*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

Reference III

- "*On the convergence of stochastic iterative dynamic programming algorithms,*" T. Jaakkola, M. Jordan, S. Singh, *Neural computation*, 1994
- "*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