Understanding the Efficacy of Reinforcement Learning Through a Non-asymptotic Lens

Yuejie Chi

Carnegie Mellon University

IEEE Data Science and Learning Workshop
May 2022

My wonderful students and collaborators

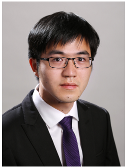
Shicong Cen
CMU

Chen Cheng
Stanford

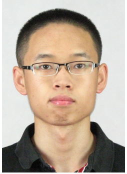
Yuxin Chen
UPenn

Yuting Wei
UPenn

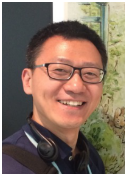
Laixi Shi
CMU



Changxiao Cai
UPenn

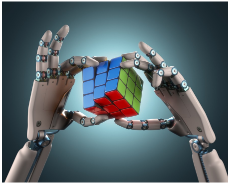


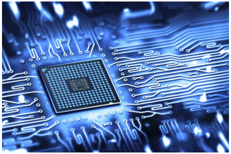
Gen Li
UPenn

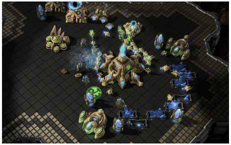


Yuantao Gu
Tsinghua

In RL, an agent learns by interacting with an environment.



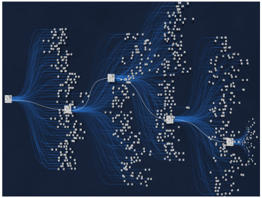




RL holds great promise in the next era of artificial intelligence.

Challenges of RL

- explore or exploit: unknown or changing environments
- credit assignment problem: delayed rewards or feedback
- enormous state and action space
- nonconcavity in value maximization



Sample efficiency

Collecting data samples might be expensive or time-consuming

clinical trials

autonomous driving

online ads

Sample efficiency

Collecting data samples might be expensive or time-consuming

clinical trials

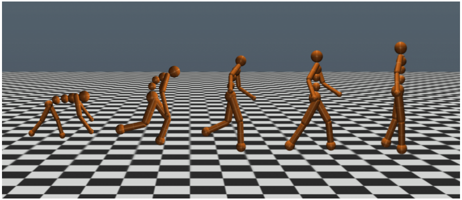
autonomous driving

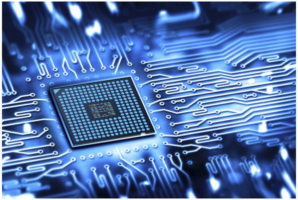
online ads

Calls for design of sample-efficient RL algorithms!

Computational efficiency

Running RL algorithms might take a long time and space

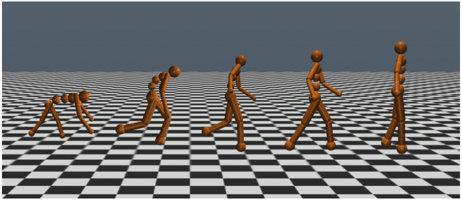


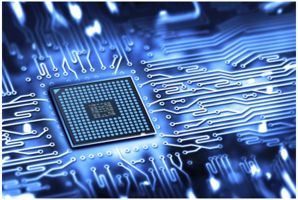


many CPUs / GPUs / TPUs + computing hours

Computational efficiency

Running RL algorithms might take a long time and space





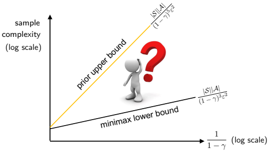
many CPUs / GPUs / TPUs + computing hours

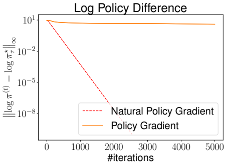
Calls for computationally efficient RL algorithms!

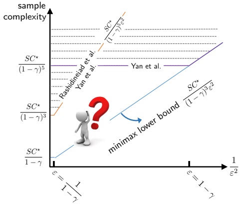
From asymptotic to non-asymptotic analyses

Non-asymptotic analyses are key to understand sample and computational efficiency in modern RL.

This talk: non-asymptotic analysis of RL





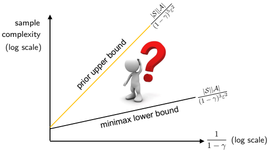


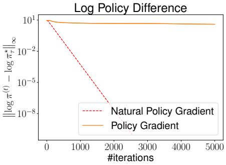
Value-based approach:
Q-learning

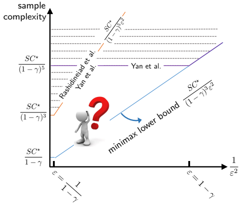
Policy-based approach:
Policy Optimization

Model-based approach:
Offline RL

This talk: non-asymptotic analysis of RL







Value-based approach:
Q-learning

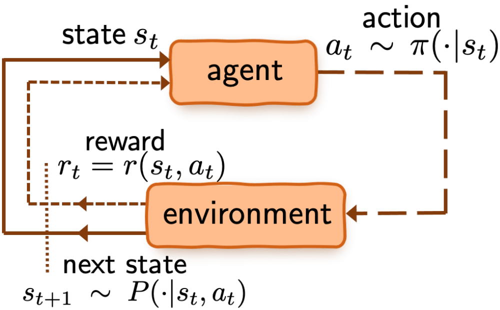
Policy-based approach:
Policy Optimization

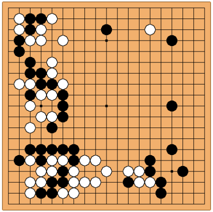
Model-based approach:
Offline RL

Does reinforcement learning learn the optimal policy, optimally?

Backgrounds: Markov decision processes

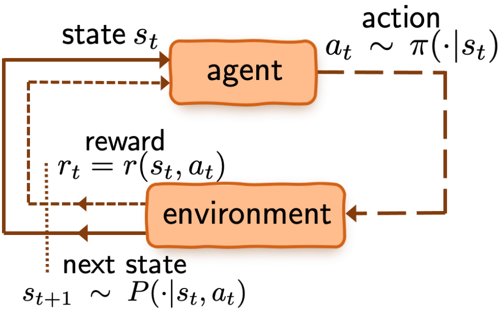
Markov decision process (MDP)

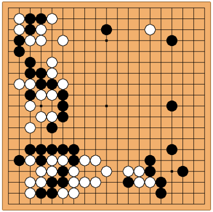




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

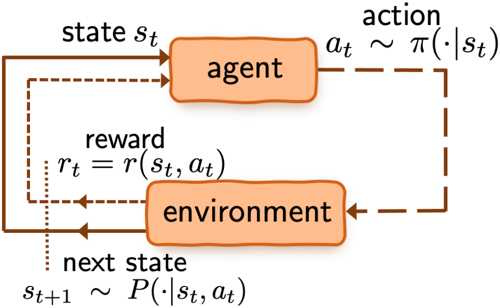
Markov decision process (MDP)

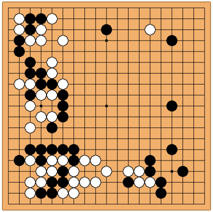




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward

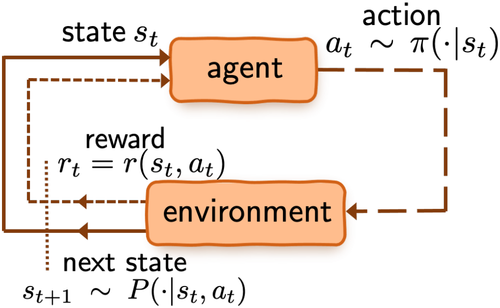
Markov decision process (MDP)

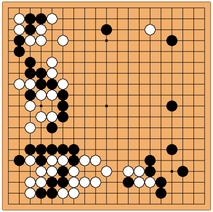




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

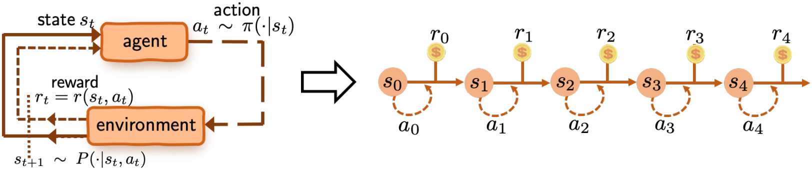
Markov decision process (MDP)





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

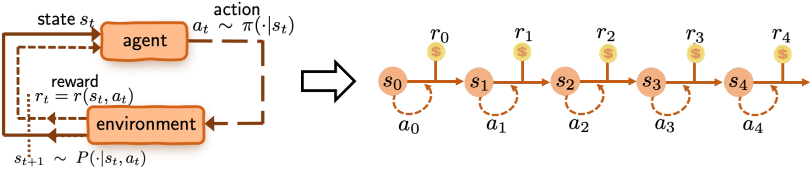
Value function



Value function of policy π:

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

Value function

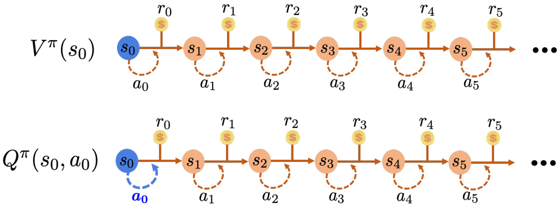


Value function of policy π:

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

- $\gamma \in [0,1)$ is the discount factor; $\frac{1}{1-\gamma}$ is effective horizon
- Expectation is w.r.t. the sampled trajectory under π

Q-function

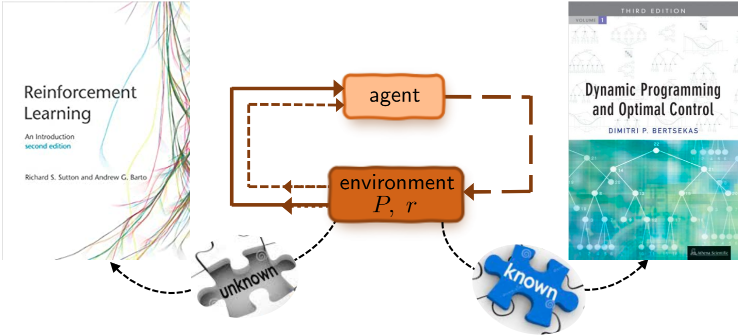


Q-function of policy π:

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

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

Searching for the optimal policy



Goal: find the optimal policy $\pi^\star$ that maximize $V^\pi(s)$

- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$
- optimal policy $\pi^\star(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, Q^\star(s,a)$

Bellman's optimality principle

Bellman operator

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

- one-step look-ahead

Bellman's optimality principle

Bellman operator

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

- one-step look-ahead

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

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

γ**-contraction of Bellman operator:**

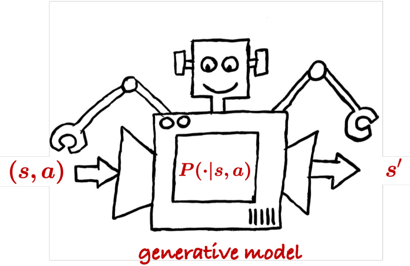
$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$$

Richard Bellman

Is Q-learning minimax-optimal?

RL with a generative model / simulator

— *Kearns and Singh, 1999*

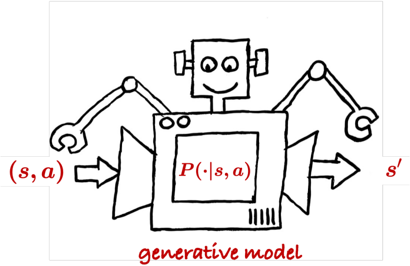


Query *any* state-action pair (s, a), collect sample transition

$$(s, a, s')$$

RL with a generative model / simulator

— *Kearns and Singh, 1999*



Query *any* state-action pair (s, a), collect sample transition

$$(s, a, s')$$

Question: How many samples are necessary and sufficient to solve the RL problem without worrying about exploration?

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

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

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

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 classical model-free algorithm

Chris Watkins

Peter Dayan

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

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

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

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

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

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

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

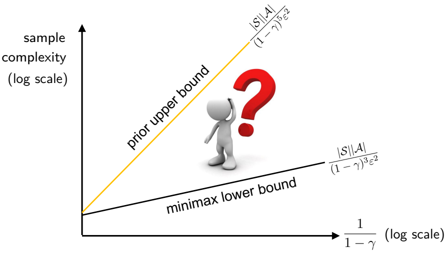
Prior art: achievability

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

Prior art: achievability

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

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

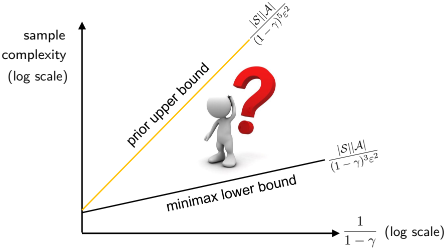


All prior results require sample size of at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\epsilon^2}$!

Prior art: achievability

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

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



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

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

A sharpened sample complexity of Q-learning

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

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

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

with sample complexity at most

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

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

A sharpened sample complexity of Q-learning

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

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

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

with sample complexity at most

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

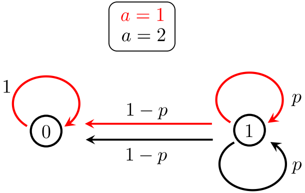
- Improves dependency on effective horizon $\frac{1}{1-\gamma}$
- Allows both constant and rescaled linear learning rate:

$$\frac{1}{1+\frac{c_1(1-\gamma)T}{\log^2 T}} \leq \eta_t \leq \frac{1}{1+\frac{c_2(1-\gamma)t}{\log^2 T}}$$

A curious numerical example

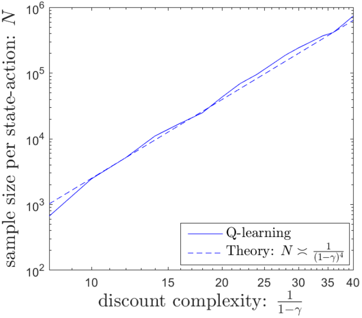
Numerical evidence: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}$ samples seem necessary . . .

— *observed in Wainwright '19*



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

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



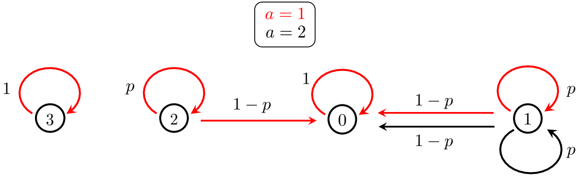
Q-learning is not minimax optimal

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

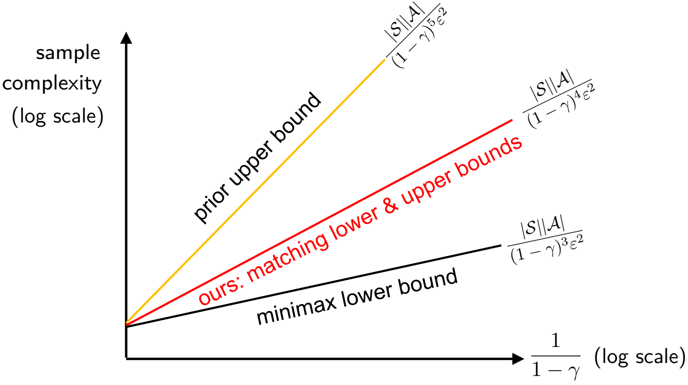
For any $0 < \epsilon \leq 1$*, there exists an MDP such that to achieve* $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \epsilon$*, Q-learning needs at least a sample complexity of*

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

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

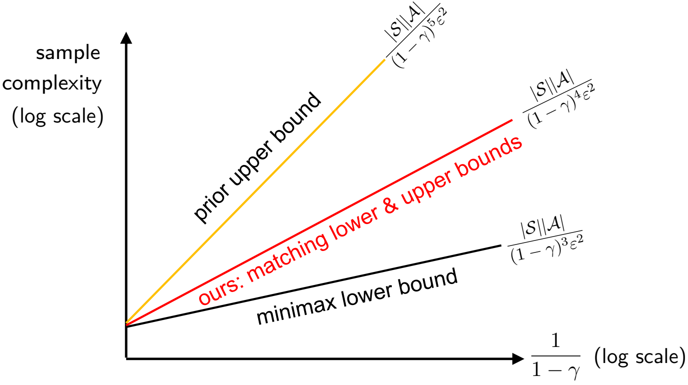


Where we stand now



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

Where we stand now



Q-learning is not minimax optimal!

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).

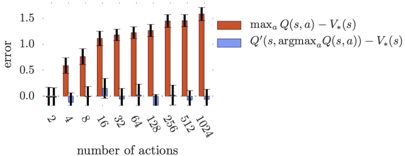


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically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

TD-learning: when the action space is a singleton

Richard Sutton

Stochastic approximation for solving Bellman equation $V = \mathcal{T}(V)$

$$V_{t+1}(s) = (1-\eta_t)V_t(s) + \eta_t \mathcal{T}_t(V_t)(s)$$
$$= V_t(s) + \eta_t \underbrace{\Big[r(s) + \gamma V_t(s') - V_t(s)\Big]}_{\text{temporal difference}}, \quad t \geq 0$$

$$\mathcal{T}_t(V)(s) = r(s) + \gamma V(s')$$
$$\mathcal{T}(V)(s) = r(s) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s)} V(s')$$

A sharpened sample complexity of TD-learning

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

For any $0 < \epsilon \leq 1$*, TD-learning yields*

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

with sample complexity at most

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

- Near minimax-optimal without the need of averaging or variance reduction.

A sharpened sample complexity of TD-learning

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

For any $0 < \epsilon \leq 1$, *TD-learning yields*

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

with sample complexity at most

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

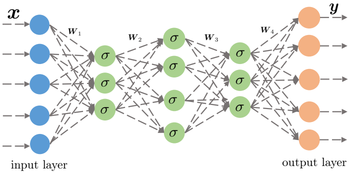
- Near minimax-optimal without the need of averaging or variance reduction.
- Allows both constant and rescaled linear learning rate.

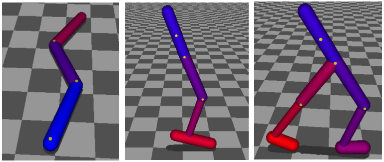
How to accelerate the convergence of policy gradient methods?

Policy optimization

$$\text{maximize}_\theta \quad \text{value}(\text{policy}(\theta))$$

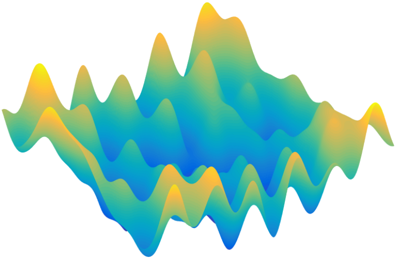
- directly optimize the policy, which is the quantity of interest;
- allow flexible differentiable parameterizations of the policy;
- work with both continuous and discrete problems.





Theoretical challenges: non-concavity

Little understanding on the global convergence of policy gradient methods until very recently, e.g. (Fazel et al., 2018; Bhandari and Russo, 2019; Agarwal et al., 2019; Mei et al. 2020), and many many more.



Can we understand and accelerate the global convergence of policy gradient methods?

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_{\pi} \quad V^{\pi}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

$$\Downarrow$$

softmax parameterization:
$$\pi_{\theta}(a|s) \propto \exp(\theta(s,a))$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_{\pi} \quad V^{\pi}(\rho) := \mathbb{E}_{s \sim \rho}\left[V^{\pi}(s)\right]$$

$$\Downarrow$$

softmax parameterization:

$$\pi_\theta(a|s) \propto \exp(\theta(s,a))$$

$$\text{maximize}_{\theta} \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s \sim \rho}\left[V^{\pi_\theta}(s)\right]$$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

$$\text{maximize}_\pi \quad V^\pi(\rho) := \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]$$

$$\Downarrow$$

softmax parameterization:

$$\pi_\theta(a|s) \propto \exp(\theta(s,a))$$

$$\text{maximize}_\theta \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

Policy gradient method (Sutton et al., 2000)

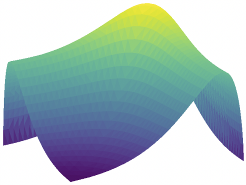
For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate.*

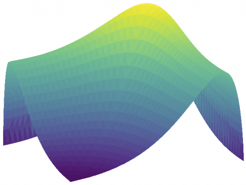
— *we'll assume exact gradient evaluation*

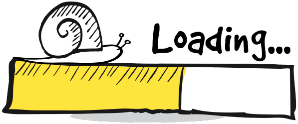
Global convergence of the PG method?



- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.

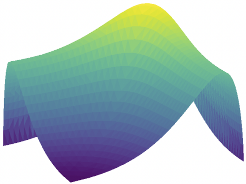
Global convergence of the PG method?

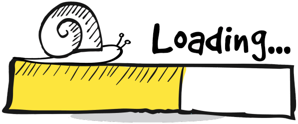




- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.

Global convergence of the PG method?

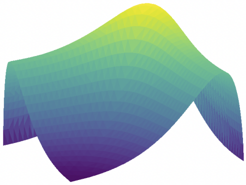


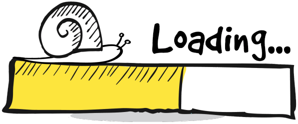


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

  $O\left(\frac{1}{\epsilon}\right)$ iterations.

Global convergence of the PG method?

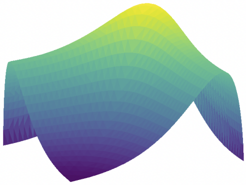


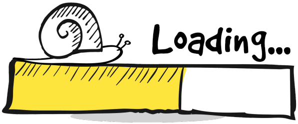


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

$$c\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \cdots\right) O\left(\frac{1}{\epsilon}\right) \text{ iterations.}$$

Global convergence of the PG method?





- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

$$c\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \cdots\right) O\left(\frac{1}{\epsilon}\right) \text{ iterations.}$$

Is the rate of PG good, bad or ugly?

A negative message

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

Starting from a uniform initial state distribution, there exists an MDP s.t. it takes softmax PG at least

$$\frac{1}{\eta}\,|\mathcal{S}|^{2^{\Theta(\frac{1}{1-\gamma})}}$$

iterations to achieve $\|V^{(t)} - V^{\star}\|_{\infty} \leq 0.15$.

A negative message

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

Starting from a uniform initial state distribution, there exists an MDP s.t. it takes softmax PG at least

$$\frac{1}{\eta}|\mathcal{S}|^{2^{\Theta(\frac{1}{1-\gamma})}}$$

iterations to achieve $\|V^{(t)} - V^{\star}\|_{\infty} \leq 0.15$.

- Softmax PG can take (super)-exponential time to converge (in problems w/ large state space & long effective horizon)!

A negative message

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

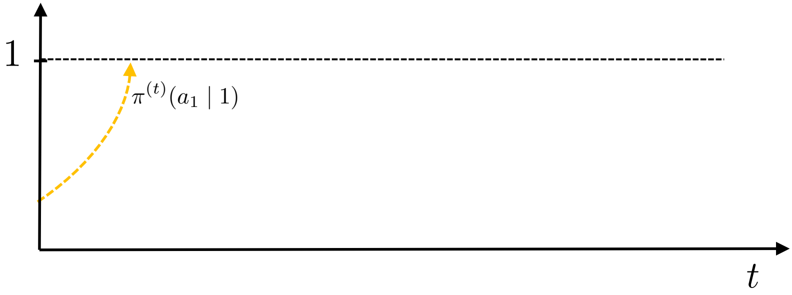
Starting from a uniform initial state distribution, there exists an MDP s.t. it takes softmax PG at least

$$\frac{1}{\eta} |\mathcal{S}|^{2^{\Theta(\frac{1}{1-\gamma})}}$$

iterations to achieve $\|V^{(t)} - V^{\star}\|_{\infty} \leq 0.15$.

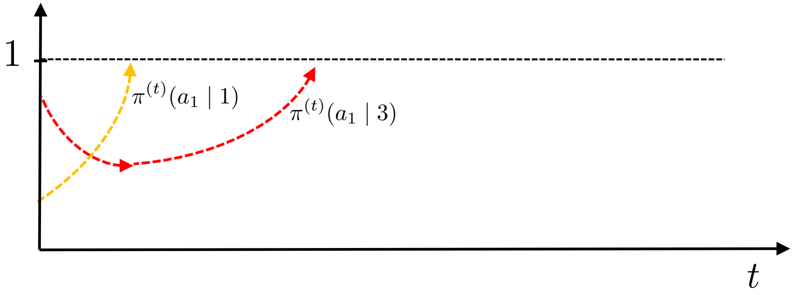
- Softmax PG can take (super)-exponential time to converge (in problems w/ large state space & long effective horizon)!
- Also hold for average sub-opt gap $\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left[V^{(t)}(s) - V^{\star}(s)\right]$.

What is happening in our constructed MDP?



We constructed a chain-structured MDP where the convergence time for state s grows geometrically as s increases

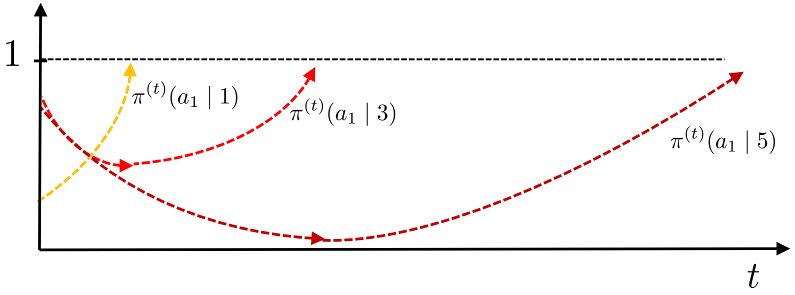
What is happening in our constructed MDP?



We constructed a chain-structured MDP where the convergence time for state s grows geometrically as s increases

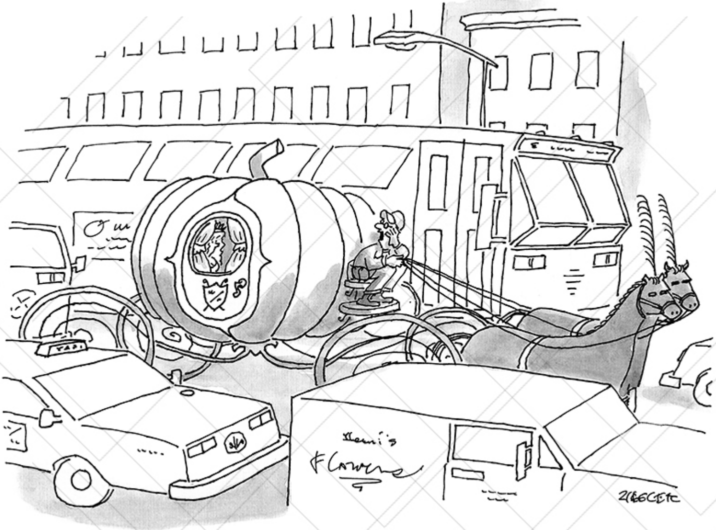
$$\text{convergence-time}(s) \gtrsim \big(\text{convergence-time}(s-2)\big)^{1.5}$$

What is happening in our constructed MDP?



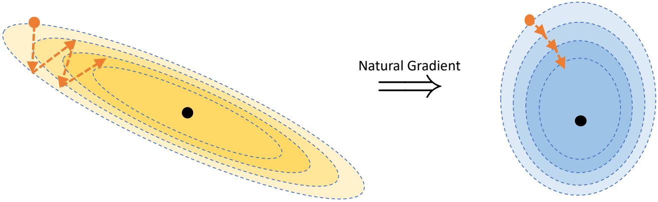
We constructed a chain-structured MDP where the convergence time for state s grows geometrically as s increases

$$\text{convergence-time}(s) \gtrsim \big(\text{convergence-time}(s-2)\big)^{1.5}$$



"Seriously, lady, at this hour you'd make a lot better time taking the subway."

Booster #1: natural policy gradient



Natural policy gradient (NPG) method (Kakade, 2002)

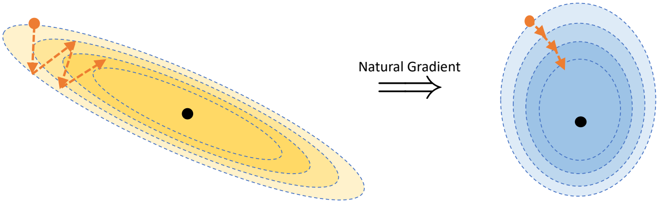
For $t = 0, 1, \cdots$

$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate and* $\mathcal{F}_\rho^\theta$ *is the Fisher information matrix:*

$$\mathcal{F}_\rho^\theta := \mathbb{E}\left[\left(\nabla_\theta \log \pi_\theta(a|s)\right)\left(\nabla_\theta \log \pi_\theta(a|s)\right)^\top\right].$$

Booster #1: natural policy gradient



Natural policy gradient (NPG) method (Kakade, 2002)

For $t = 0, 1, \cdots$

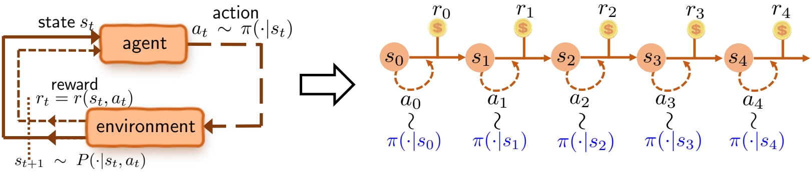
$$\theta^{(t+1)} = \theta^{(t)} + \eta(\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

where η *is the learning rate and* $\mathcal{F}_\rho^\theta$ *is the Fisher information matrix:*

$$\mathcal{F}_\rho^\theta := \mathbb{E}\left[\left(\nabla_\theta \log \pi_\theta(a|s)\right)\left(\nabla_\theta \log \pi_\theta(a|s)\right)^\top\right].$$

In fact, popular heuristic TRPO (Schulman et al., 2015) = NPG + line search.

Booster #2: entropy regularization

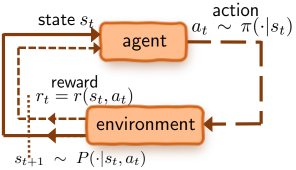


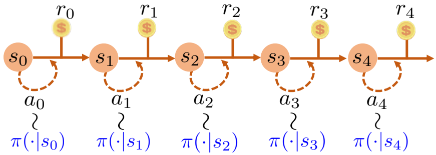
To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t))\Big) \,\Big|\, s_0 = s\right]$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

Booster #2: entropy regularization





To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

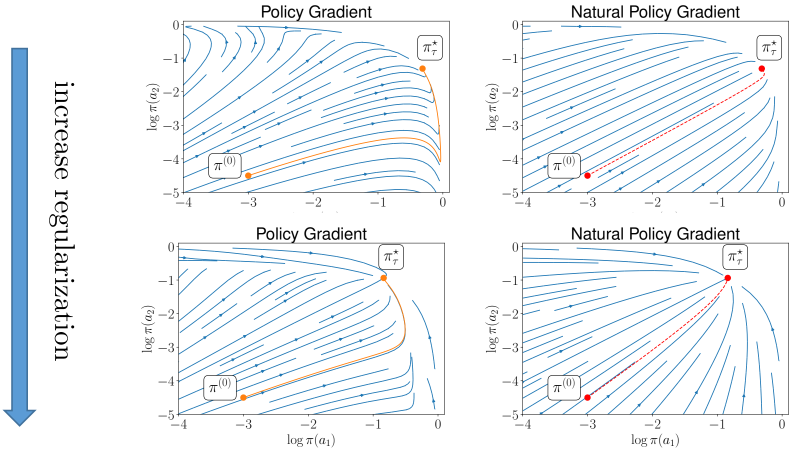
$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t))\Big) \,\Big|\, s_0 = s\right]$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

$$\text{maximize}_\theta \quad V_\tau^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V_\tau^{\pi_\theta}(s)\right]$$

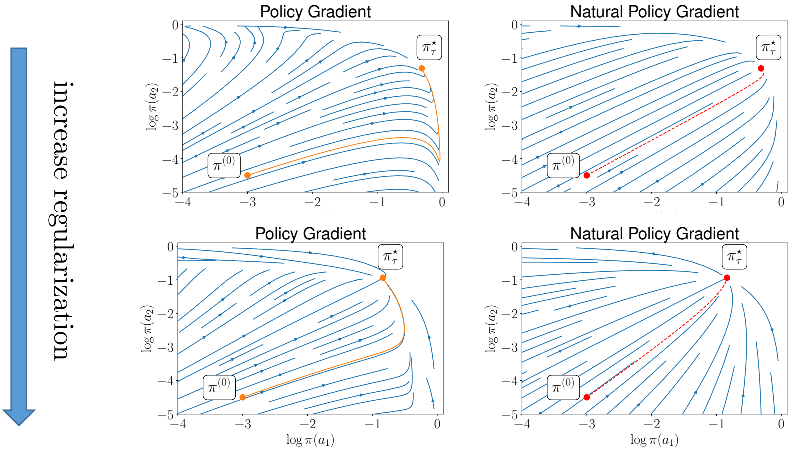
Entropy-regularized natural gradient helps!

Toy example: a bandit with 3 arms of rewards 1, 0.9 and 0.1.



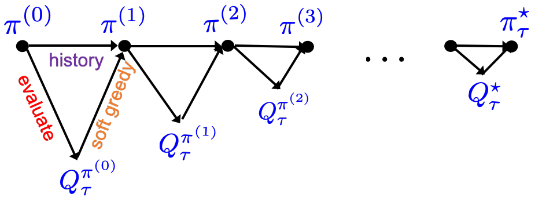
Entropy-regularized natural gradient helps!

Toy example: a bandit with 3 arms of rewards 1, 0.9 and 0.1.



Can we justify the efficacy of entropy-regularized NPG?

Entropy-regularized NPG in the tabular setting



Entropy-regularized NPG (Tabular setting)

For $t = 0, 1, \cdots$, *the policy is updated via*

$$\pi^{(t+1)}(\cdot|s) \propto \underbrace{\pi^{(t)}(\cdot|s)}_{\text{current policy}}{}^{1-\frac{\eta\tau}{1-\gamma}} \underbrace{\exp(Q_\tau^{(t)}(s,\cdot)/\tau)}_{\text{soft greedy}}{}^{\frac{\eta\tau}{1-\gamma}}$$

where $Q_\tau^{(t)} := Q_\tau^{\pi^{(t)}}$ *is the soft Q-function of* $\pi^{(t)}$, *and* $0 < \eta \le \frac{1-\gamma}{\tau}$.

- invariant with the choice of ρ
- Reduces to soft policy iteration (SPI) when $\eta = \frac{1-\gamma}{\tau}$.

Linear convergence with exact gradient

Theorem (Cen, Cheng, Chen, Wei, Chi, 2020)

For any learning rate $0 < \eta \leq (1-\gamma)/\tau$, the entropy-regularized NPG needs no more than

$$\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$$

iterations to reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \epsilon$.

- Soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$): $\frac{1}{1-\gamma} \log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\epsilon}\right)$.

Linear convergence with exact gradient

Theorem (Cen, Cheng, Chen, Wei, Chi, 2020)

For any learning rate $0 < \eta \leq (1-\gamma)/\tau$, the entropy-regularized NPG needs no more than

$$\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$$

iterations to reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \epsilon$.

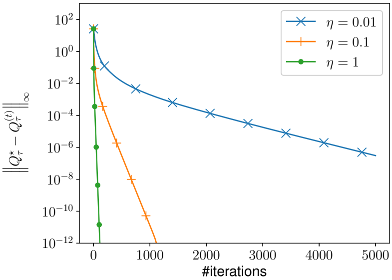
- Soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$): $\frac{1}{1-\gamma} \log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\epsilon}\right)$.

Global linear convergence of entropy-regularized NPG
at a rate independent of $|\mathcal{S}|$, $|\mathcal{A}|$!

Entropy helps

Regularized NPG

$\tau = 0.001$

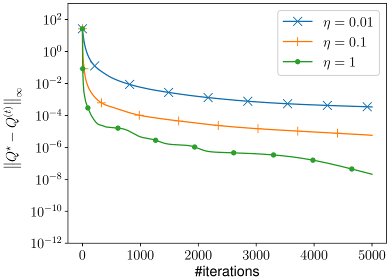


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Vanilla NPG

$\tau = 0$



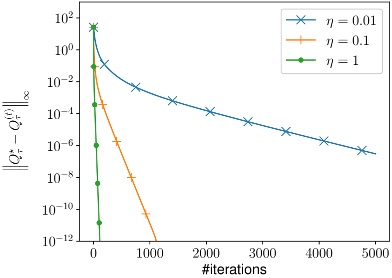
Sublinear rate: $\frac{1}{\min\{\eta,(1-\gamma)^2\}\epsilon}$

(Agarwal et al. 2019)

Entropy helps

Regularized NPG

$\tau = 0.001$

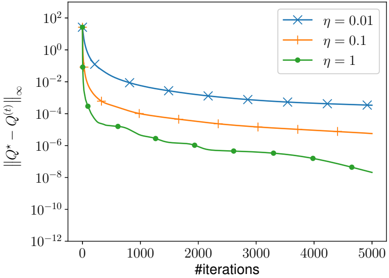


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Vanilla NPG

$\tau = 0$



Sublinear rate: $\frac{1}{\min\{\eta,(1-\gamma)^2\}\epsilon}$

(Agarwal et al. 2019)

Entropy regularization enables fast convergence!

A key operator: soft Bellman operator

Soft Bellman operator

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

A key operator: soft Bellman operator

Soft Bellman operator

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

Soft Bellman equation: $Q_\tau^\star$ is *unique* solution to

$$\mathcal{T}_\tau(Q_\tau^\star) = Q_\tau^\star$$

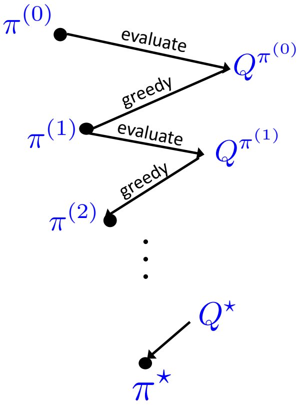
γ-contraction of soft Bellman operator:

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

Richard Bellman

Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

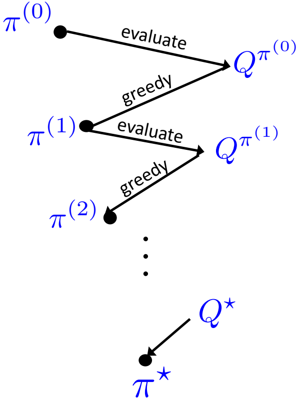
Policy iteration



Bellman operator

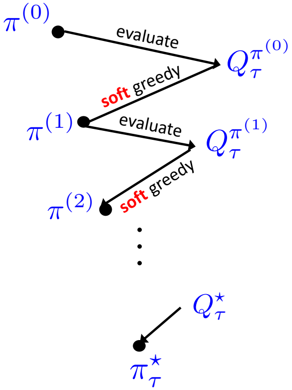
Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

Policy iteration



Bellman operator

Soft policy iteration



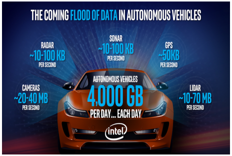
Soft Bellman operator

Offline RL: learning without exploration

Offline RL / Batch RL

- Sometimes we can not explore or generate new data
- But we have already stored tons of historical data

medical records

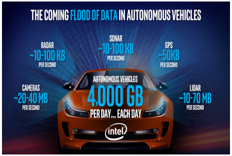


data of self-driving

clicking times of ads

- Sometimes we can not explore or generate new data
- But we have already stored tons of historical data

medical records

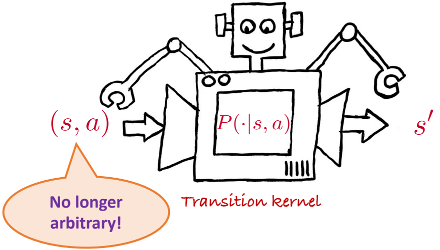


data of self-driving

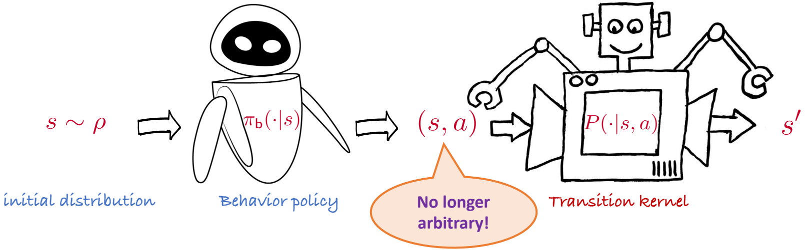
clicking times of ads

Can we learn a good policy based solely on historical data without active exploration?

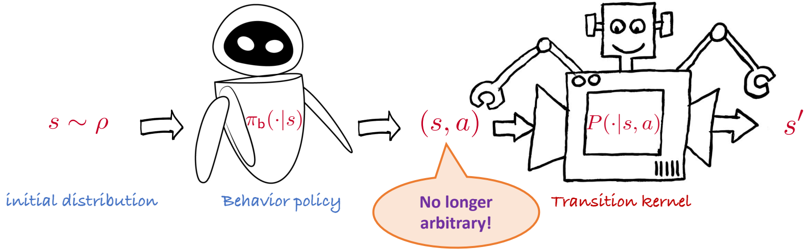
A simplified model of history data from behavior policy



A simplified model of history data from behavior policy



A simplified model of history data from behavior policy



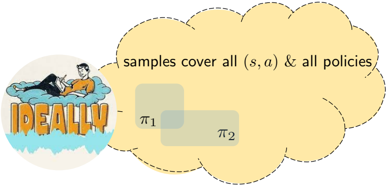
Goal of offline RL: given history data $\mathcal{D} := \{(s_i, a_i, s'_i)\}_{i=1}^N$, find an ϵ-optimal policy $\widehat{\pi}$ obeying

$$V^\star(\rho) - V^{\widehat{\pi}}(\rho) \le \epsilon$$

— *in a sample-efficient manner*

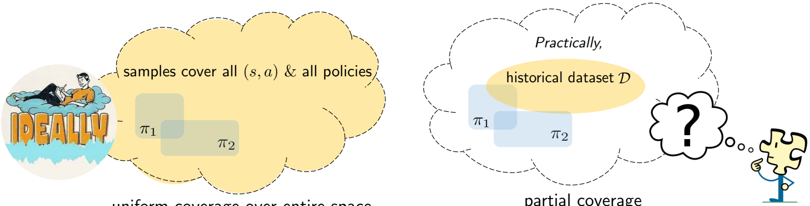
Challenges of offline RL

Partial coverage of state-action space:



uniform coverage over entire space
(sufficiently explored)

Partial coverage of state-action space:

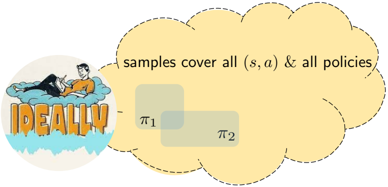


uniform coverage over entire space
(sufficiently explored)

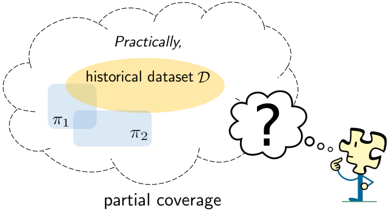
partial coverage
(inadequately explored)

Challenges of offline RL

Partial coverage of state-action space:



uniform coverage over entire space
(sufficiently explored)



partial coverage
(inadequately explored)

Distribution shift:

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

How to quantify the distribution shift?

Single-policy concentrability coefficient (Rashidineiad et al.)

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

where $d^\pi(s,a)$ is the state-action occupation density of policy π.

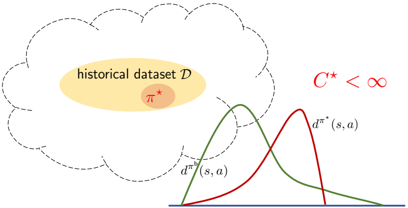
How to quantify the distribution shift?

Single-policy concentrability coefficient (Rashidineiad et al.)

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

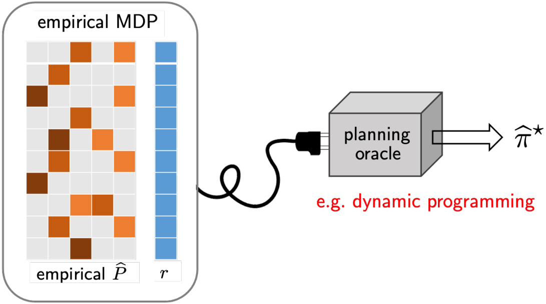
where $d^{\pi}(s,a)$ is the state-action occupation density of policy π.

- captures distribution shift
- allows for partial coverage
- Behavior cloning $C^{\star} = 1$



A "plug-in" model-based approach

— (Azar et al. '13, Agarwal et al. '19, Li et al. '20)



Planning (e.g., value iteration) based on the the empirical MDP $\widehat{P}$:

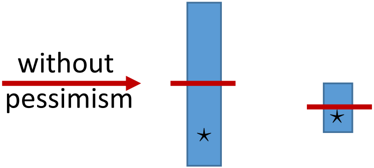
$$\widehat{Q}(s,a) \ \leftarrow \ r(s,a) + \gamma \langle \widehat{P}(\cdot \,|\, s,a), \widehat{V} \rangle, \quad \widehat{V}(s) = \max_a \widehat{Q}(s,a).$$

Issue: poor value estimates under partial and poor coverage.

Pessimism in the face of uncertainty

Penalize value estimate of (s, a) pairs that were poorly visited

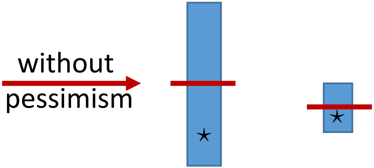
— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)



Pessimism in the face of uncertainty

Penalize value estimate of (s, a) pairs that were poorly visited

— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)



Value iteration with lower confidence bound (VI-LCB):

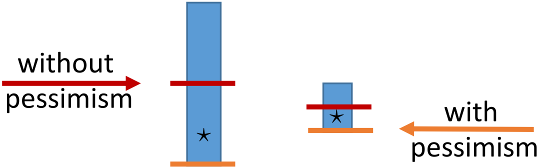
$$\widehat{Q}(s,a) \leftarrow \max\big\{r(s,a) + \gamma\big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}\big\rangle - \underbrace{b(s,a;\widehat{V})}_{\text{uncertainty penalty}}, 0\big\},$$

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

Pessimism in the face of uncertainty

Penalize value estimate of (s, a) pairs that were poorly visited

— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)

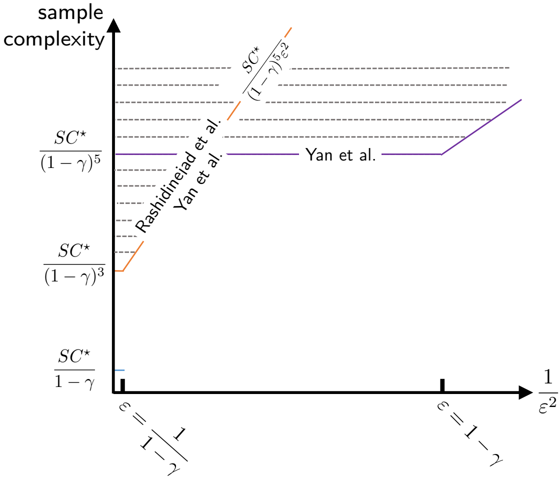


Value iteration with lower confidence bound (VI-LCB):

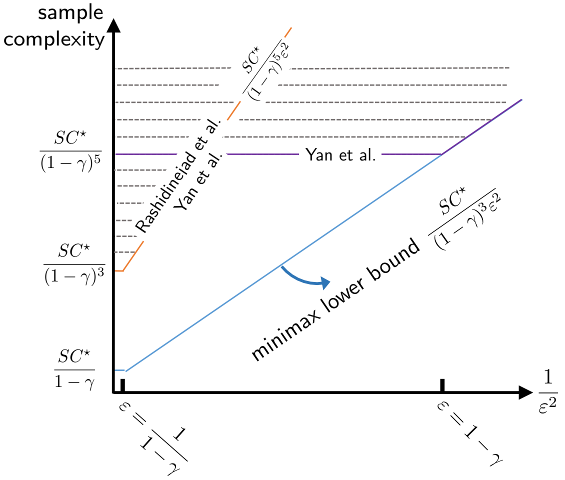
$$\widehat{Q}(s,a) \leftarrow \max\Big\{r(s,a) + \gamma\big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}\big\rangle - \underbrace{b(s,a;\widehat{V})}_{\text{uncertainty penalty}}, 0\Big\},$$

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

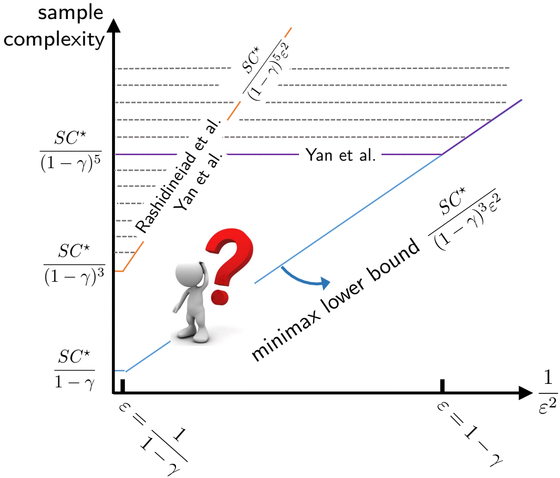
A benchmark of prior arts



A benchmark of prior arts



A benchmark of prior arts



Can we close the gap with the minimax lower bound?

Minimax optimality of model-based offline RL

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

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

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

with high prob., with sample complexity at most

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

Minimax optimality of model-based offline RL

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

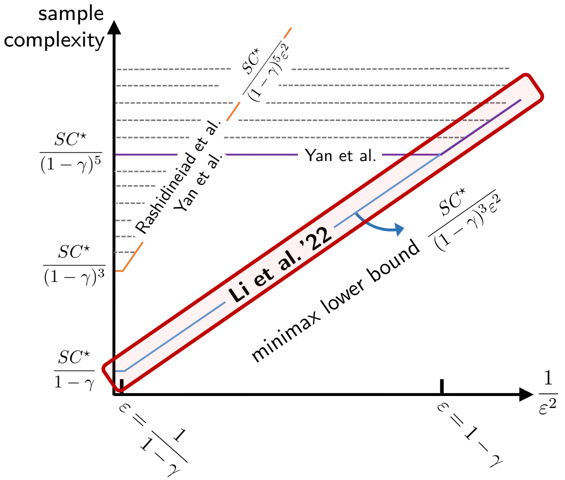
For any $0 < \epsilon \leq \frac{1}{1-\gamma}$*, the policy* $\widehat{\pi}$ *returned by VI-LCB using a Bernstein-style penalty term achieves*

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

with high prob., with sample complexity at most

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

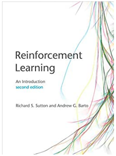
- matches minimax lower bound: $\widetilde{\Omega}\left(\frac{SC^{\star}}{(1-\gamma)^{3}\epsilon^{2}}\right)$
- depends on distribution shift (as reflected by $C^{\star}$)
- full ϵ-range (no burn-in cost)

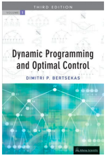


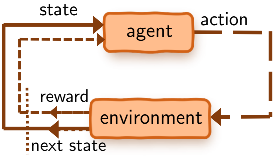
Model-based RL is minimax optimal with no burn-in cost!

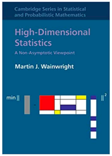
Concluding remarks

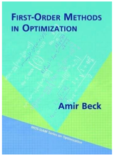
Concluding remarks











Understanding non-asymptotic performances of RL algorithms sheds light to their empirical successes (and failures)!

Future directions:

- function approximation
- multi-agent RL
- robust RL
- many more...

References

Q-learning and variants:

- Is Q-learning minimax optimal? a tight sample complexity analysis, arXiv:2102.06548, short version at ICML 2021.
- Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction, *IEEE Trans. on Information Theory*, short version at NeurIPS 2020.

Policy optimization:

- Fast global convergence of natural policy gradient methods with entropy regularization, *Operations Research*, in press.
- Softmax policy gradient methods can take exponential time to converge, arXiv:2102.11270, short version at COLT 2021.
- Fast policy extragradient methods for competitive games with entropy regularization, arXiv:2105.15186, short version at NeurIPS 2021.

Offline RL:

- Settling the sample complexity of model-based offline reinforcement learning, arXiv:2204.05275.

Thank you!

https://users.ece.cmu.edu/~yuejiec/