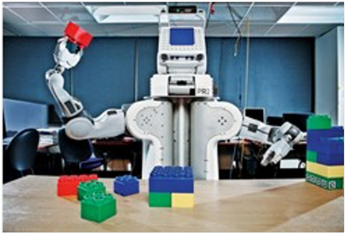
Demystifying the efficiency of reinforcement learning: A few recent stories

Yuxin Chen

EE, Princeton University

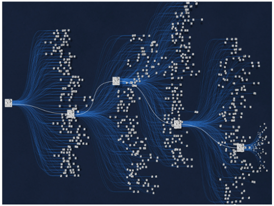
Reinforcement learning (RL)



RL challenges

In RL, an agent learns by interacting with an environment

- unknown or changing environments
- delayed rewards or feedback
- enormous state and action space
- nonconvexity



Sample efficiency

Collecting data samples might be expensive or time-consuming

clinical trials

online ads

Sample efficiency

Collecting data samples might be expensive or time-consuming

clinical trials

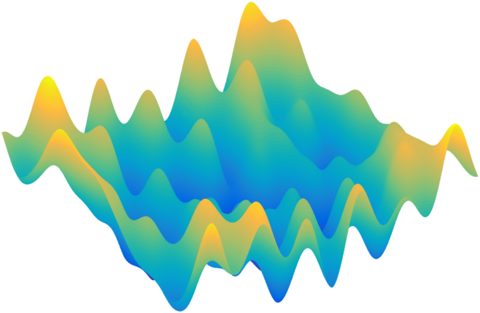
online ads

Calls for design of sample-efficient RL algorithms!

Computational efficiency

Running RL algorithms might take a long time . . .

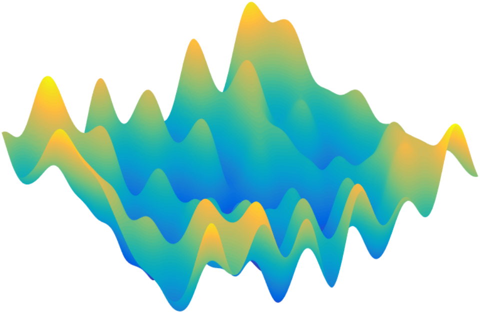
- enormous state-action space
- nonconvexity



Computational efficiency

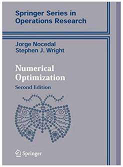
Running RL algorithms might take a long time ...

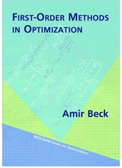
- enormous state-action space
- nonconvexity

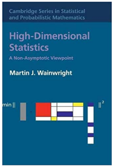


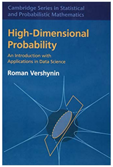
Calls for computationally efficient RL algorithms!

This talk: three recent stories





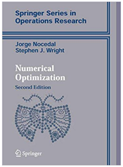


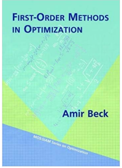
(large-scale) optimization

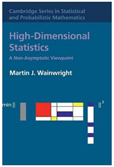
(high-dimensional) statistics

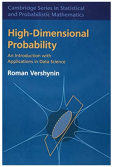
Demystify sample- and computational efficiency of RL algorithms

This talk: three recent stories





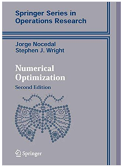


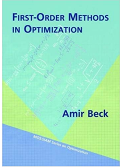
(large-scale) optimization (high-dimensional) statistics

Demystify sample- and computational efficiency of RL algorithms

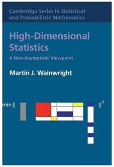
1. **model-based RL**
2. **policy-based RL**
3. **value-based RL**

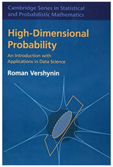
This talk: three recent stories





(large-scale) optimization





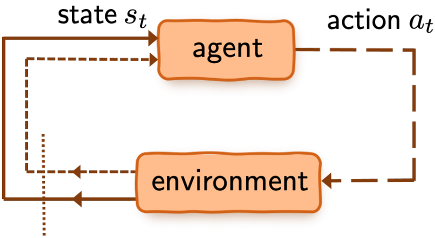
(high-dimensional) statistics

Demystify sample- and computational efficiency of RL algorithms

1. **model-based RL**: breaking a sample size barrier
2. **policy-based RL**: natural policy gradient (NPG) methods
3. **value-based RL**: Q-learning over Markovian samples

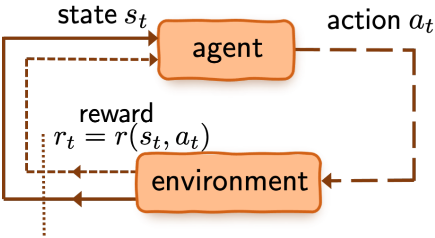
Background: Markov decision processes

Markov decision process (MDP)



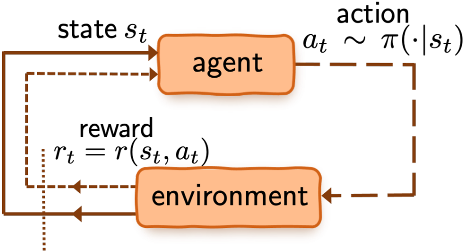
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

Markov decision process (MDP)



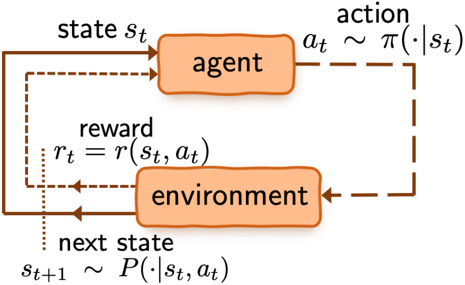
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

Markov decision process (MDP)



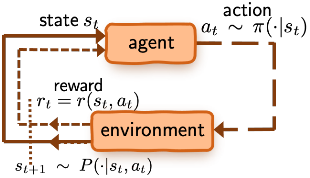
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

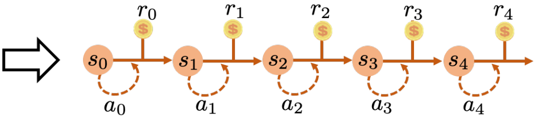
Markov decision process (MDP)



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

Value function

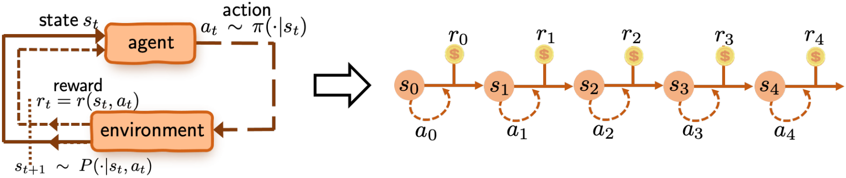




Value of policy π: long-term *discounted* reward

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

Value function

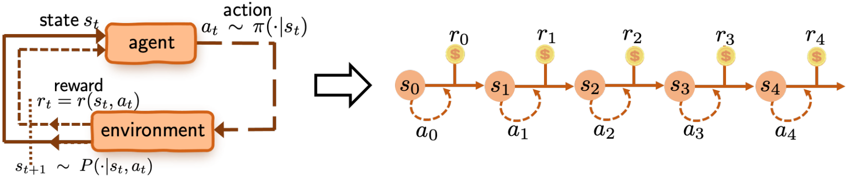


Value of policy π: long-term *discounted* reward

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

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

Value function

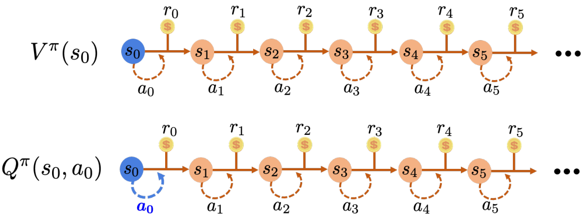


Value of policy π: long-term *discounted* reward

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

- $(a_0, s_1, a_1, s_2, a_2, \cdots)$: generated under policy π
- $\gamma \in [0, 1)$: discount factor
  - take $\gamma \to 1$ to approximate *long-horizon* MDPs

Q-function

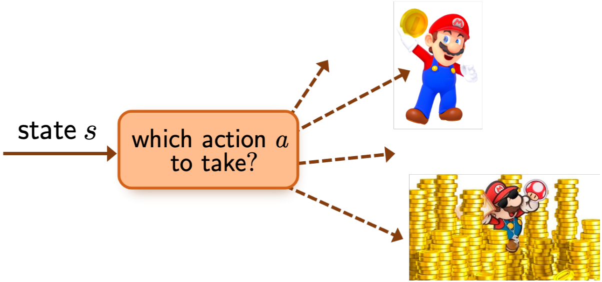


Q-function of policy π

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

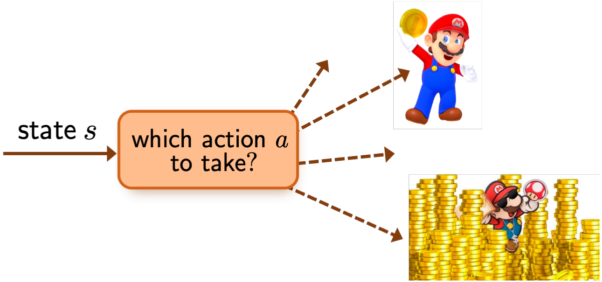
- $(\cancel{a_0,}\ s_1, a_1, s_2, a_2, \cdots)$: generated under policy π

Optimal policy and optimal values



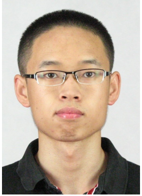
- **Optimal policy** $\pi^\star$: maximizing the value function

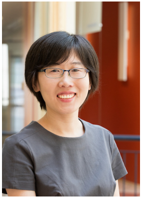
Optimal policy and optimal values

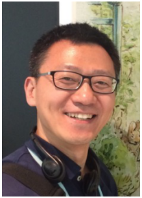


- **Optimal policy** $\pi^\star$: maximizing the value function
- Optimal values: $V^\star := V^{\pi^\star}$

Story 1: breaking the sample size barrier
*via **model-based RL** under a generative model*







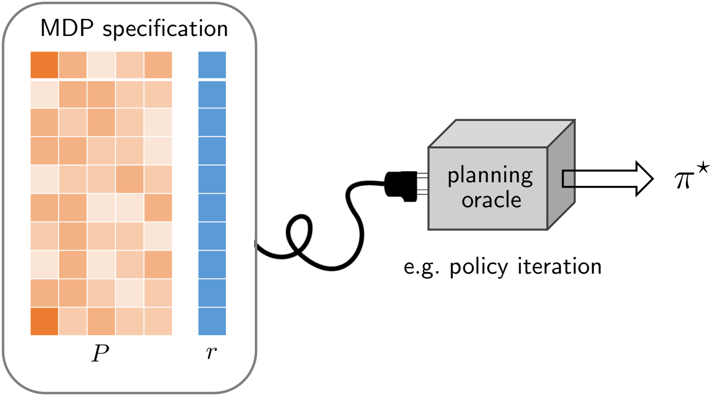
Gen Li
Tsinghua EE

Yuting Wei
CMU Stats

Yuejie Chi
CMU ECE

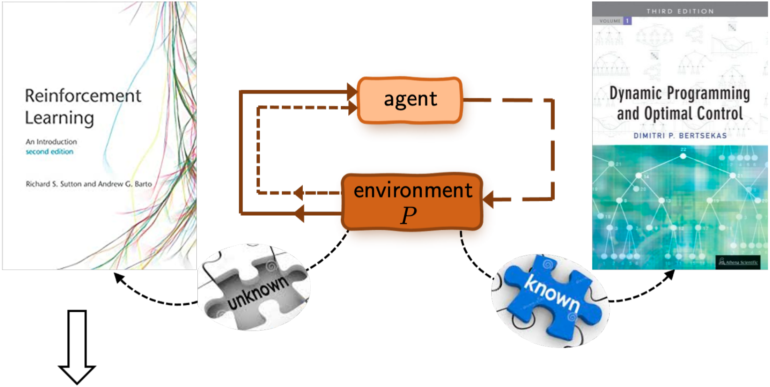
Yuantao Gu
Tsinghua EE

When the model is known ...



Planning: computing the optimal policy $\pi^\star$ given MDP specification

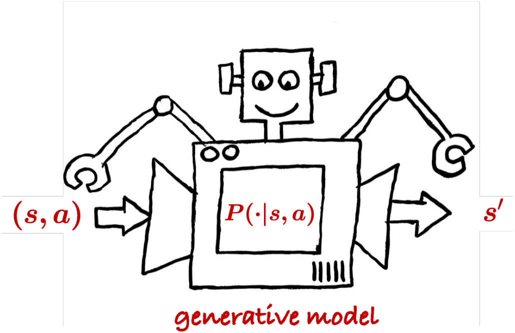
When the model is unknown . . .



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

This work: RL with a generative model / simulator

— *Kearns, Singh '99*



For each state-action pair (s,a), collect N samples $\{(s,a,s'_{(i)})\}_{1\le i\le N}$

Question: *how many samples are sufficient to learn an* $\underbrace{\varepsilon\text{-optimal policy}}$ *?*

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

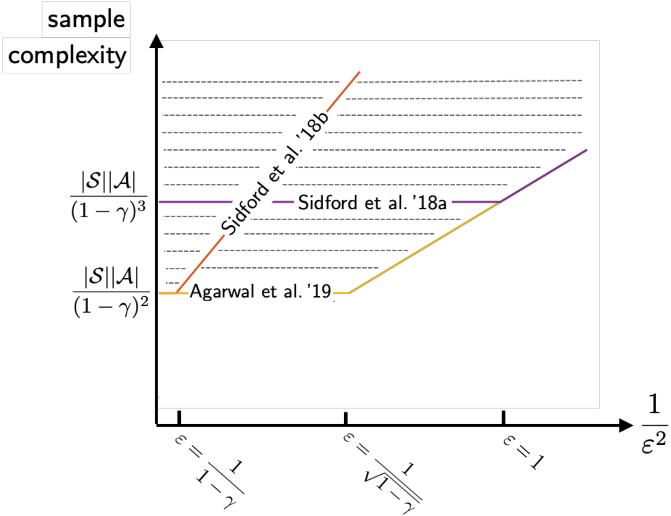
An incomplete list of prior art

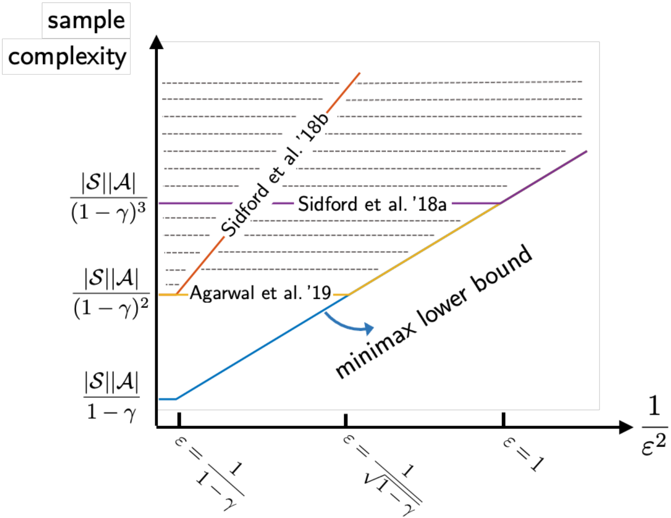
- Kearns & Singh '99
- Kakade '03
- Kearns, Mansour & Ng '02
- Azar, Munos & Kappen '12
- Azar, Munos, Ghavamzadeh & Kappen '13
- Sidford, Wang, Wu, Yang & Ye '18
- Sidford, Wang, Wu & Ye '18
- Wang '17
- Agarwal, Kakade & Yang '19
- Wainwright '19a
- Wainwright '19b
- Pananjady & Wainwright '20
- Yang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

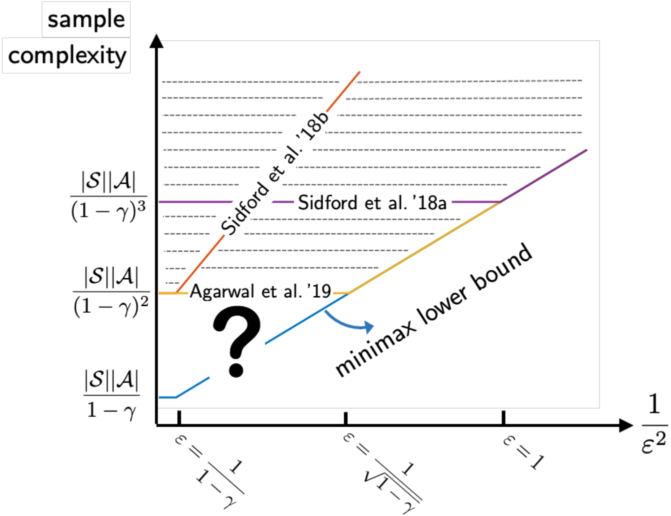
An even shorter list of prior art

algorithm	sample size range	sample complexity	ε-range
empirical QVI Azar et al. '13	$\left[\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}}]$
sublinear randomized VI Sidford et al. '18a	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\varepsilon^2}$	$\left(0, \frac{1}{1-\gamma}\right]$
variance-reduced QVI Sidford et al. '18b	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, 1]$
empirical MDP + planning Agarwal et al. '19	$\left[\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^2}, \infty\right)$	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\varepsilon^2}$	$(0, \frac{1}{\sqrt{1-\gamma}}]$

— *see also Wainwright '19 (for estimating optimal values)*


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

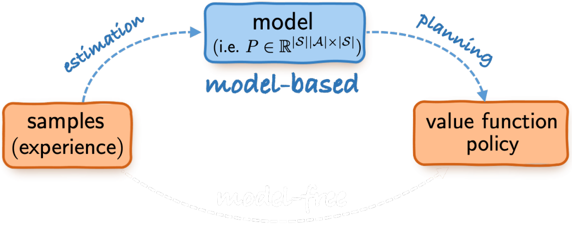

sample
complexity
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3}$
$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$
$\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}$
Sidford et al. '18b
Sidford et al. '18a
Agarwal et al. '19
minimax lower bound
$\varepsilon = \frac{1}{1-\gamma}$
$\varepsilon = \frac{1}{\sqrt{1-\gamma}}$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



All prior theory requires $\underbrace{\text{sample size} > \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}}_{\text{sample size barrier}}$

Is it possible to close the gap?

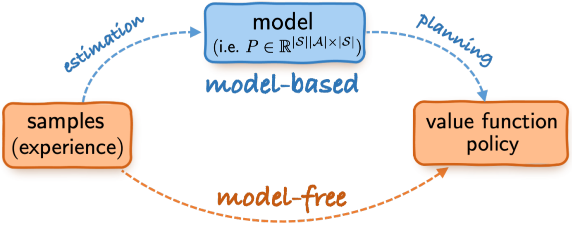
Two approaches



Model-based approach ("plug-in")

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

Two approaches



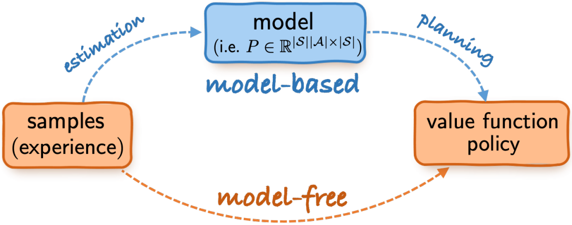
Model-based approach ("plug-in")

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

Model-free approach

— learning w/o constructing model explicitly

Two approaches



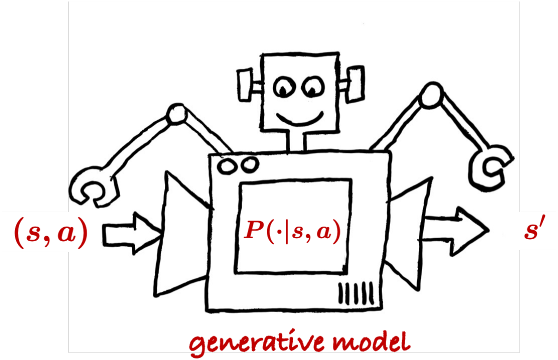
Model-based approach ("plug-in")

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

Model-free approach

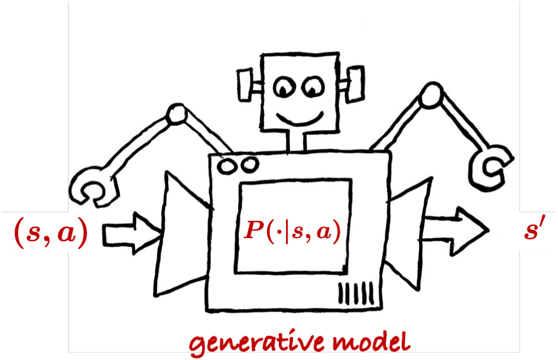
— learning w/o constructing model explicitly

Model estimation



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

Model estimation

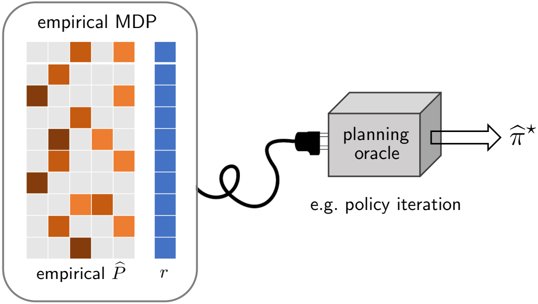


Sampling: for each (s,a), collect N ind. samples $\{(s,a,s'_{(i)})\}_{1\le i\le N}$

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

Model-based (plug-in) estimator

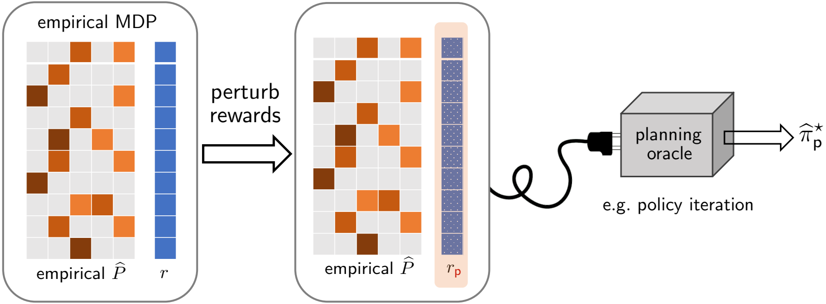
— *Azar et al. '13, Agarwal et al. '19, Pananjady et al. '20*



Planning based on the *empirical* MDP with *slightly perturbed rewards*

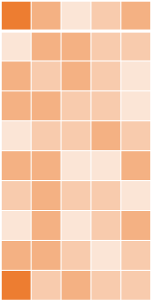
Our method: plug-in estimator + perturbation

— *Li, Wei, Chi, Gu, Chen '20*

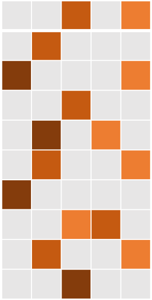


Run planning algorithms based on the *empirical* MDP

Challenges in the sample-starved regime



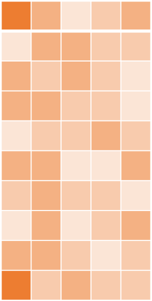
truth:
$P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



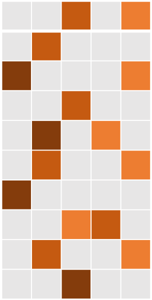
empirical estimate:
$\widehat{P}$

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

Challenges in the sample-starved regime



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



empirical estimate:
$\widehat{P}$

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

Main result

Theorem 1 (Li, Wei, Chi, Gu, Chen '20)

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

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

with sample complexity at most

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

Main result

Theorem 1 (Li, Wei, Chi, Gu, Chen '20)

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

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

with sample complexity at most

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

- $\widehat{\pi}_{\mathrm{p}}^{\star}$: obtained by empirical QVI or PI within $\widetilde{O}(\frac{1}{1-\gamma})$ iterations

Main result

Theorem 1 (Li, Wei, Chi, Gu, Chen '20)

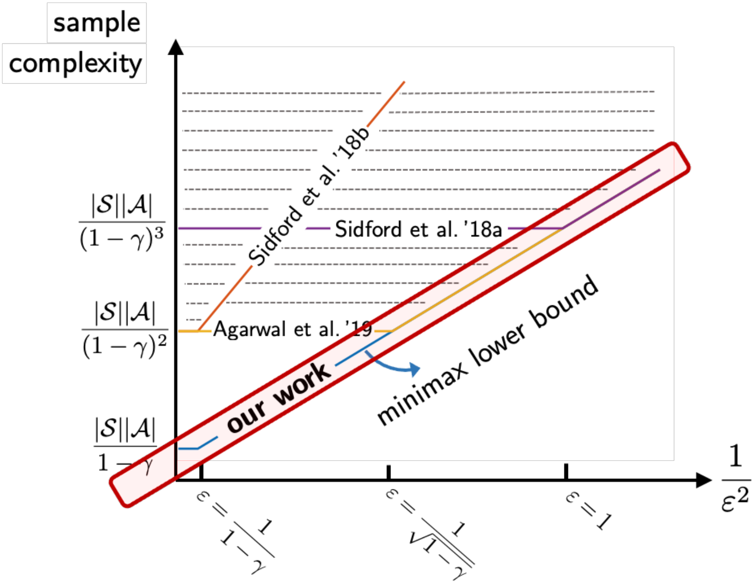
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *the optimal policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of the perturbed empirical MDP achieves*

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

with sample complexity at most

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

- $\widehat{\pi}_{\mathrm{p}}^{\star}$: obtained by empirical QVI or PI within $\widetilde{O}\left(\frac{1}{1-\gamma}\right)$ iterations
- **Minimax lower bound:** $\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}\right)$ (Azar et al. '13)


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

Analysis

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V^\pi = (I - P_\pi)^{-1} r$

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V^\pi = (I - P_\pi)^{-1} r$
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \widehat{P}_\pi)^{-1} r$

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V^\pi = (I - P_\pi)^{-1} r$
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \widehat{P}_\pi)^{-1} r$
- $\pi^\star$: optimal policy w.r.t. true value function
- $\widehat{\pi}^\star$: optimal policy w.r.t. empirical value function

Notation and Bellman equation

- V^π: true value function under policy π
  - Bellman equation: $V^\pi = (I - P_\pi)^{-1} r$
- $\widehat{V}^\pi$: estimate of value function under policy π
  - Bellman equation: $\widehat{V}^\pi = (I - \widehat{P}_\pi)^{-1} r$
- $\pi^\star$: optimal policy w.r.t. true value function
- $\widehat{\pi}^\star$: optimal policy w.r.t. empirical value function
- $V^\star := V^{\pi^\star}$: optimal values under true models
- $\widehat{V}^\star := \widehat{V}^{\widehat{\pi}^\star}$: optimal values under empirical models

Proof ideas

Elementary decomposition:

$$V^\star - V^{\widehat{\pi}^\star} = \big(V^\star - \widehat{V}^{\pi^\star}\big) + \big(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\big) + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big)$$

Proof ideas

Elementary decomposition:

$$\begin{aligned} V^\star - V^{\widehat{\pi}^\star} &= \big(V^\star - \widehat{V}^{\pi^\star}\big) + \big(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\big) + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big) \\ &\le \big(V^{\pi^\star} - \widehat{V}^{\pi^\star}\big) + 0 + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big) \end{aligned}$$

- **Step 1:** control $V^\pi - \widehat{V}^\pi$ for a fixed π
  (**Bernstein inequality** + **high-order decomposition**)

Proof ideas

Elementary decomposition:

$$\begin{aligned} V^{\star} - V^{\widehat{\pi}^{\star}} &= \left(V^{\star} - \widehat{V}^{\pi^{\star}}\right) + \left(\widehat{V}^{\pi^{\star}} - \widehat{V}^{\widehat{\pi}^{\star}}\right) + \left(\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}\right) \\ &\leq \left(V^{\pi^{\star}} - \widehat{V}^{\pi^{\star}}\right) + 0 + \left(\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}\right) \end{aligned}$$

- **Step 1:** control $V^{\pi} - \widehat{V}^{\pi}$ for a fixed π
  (**Bernstein inequality** + **high-order decomposition**)
- **Step 2:** extend it to control $\widehat{V}^{\widehat{\pi}^{\star}} - V^{\widehat{\pi}^{\star}}$ ($\widehat{\pi}^{\star}$ depends on samples)
  (**decouple statistical dependency**)

Step 1: improved theory for policy evaluation

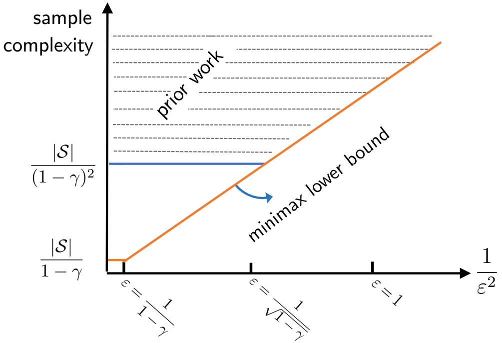
Model-based policy evaluation:

— given a fixed policy π, estimate V^{π} via the plug-in estimate $\widehat{V}^{\pi}$

Step 1: improved theory for policy evaluation

Model-based policy evaluation:

— given a fixed policy π, estimate V^{π} via the plug-in estimate $\widehat{V}^{\pi}$



- A sample size barrier $\frac{|\mathcal{S}|}{(1-\gamma)^2}$ already appeared in prior work (Agarwal et al. '19, Pananjady & Wainwright '19, Khamaru et al. '20)

Step 1: improved theory for policy evaluation

Model-based policy evaluation:

— given a fixed policy π, estimate V^π via the plug-in estimate $\widehat{V}^\pi$

Theorem 2 (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. For $0 < \varepsilon \le \frac{1}{1-\gamma}$, the plug-in estimator $\widehat{V}^\pi$ obeys

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

with sample complexity at most

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

Step 1: improved theory for policy evaluation

Model-based policy evaluation:

— given a fixed policy π, estimate V^π via the plug-in estimate $\widehat{V}^\pi$

Theorem 2 (Li, Wei, Chi, Gu, Chen'20)

Fix any policy π. For $0 < \varepsilon \leq \frac{1}{1-\gamma}$, the plug-in estimator $\widehat{V}^\pi$ obeys

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

with sample complexity at most

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

- Minimax optimal for all ε (Azar et al. '13, Pananjady & Wainwright '19)

Key idea 1: a peeling argument

Agarwal, Kakade, Yang 19: first-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)\widehat{V}^{\pi} \qquad (\star)$$

Ours: higher-order expansion $\longrightarrow$ tighter control

$$\widehat{V}^{\pi} - V^{\pi} = \gamma\big(I - \gamma P_{\pi}\big)^{-1}\big(\widehat{P}_{\pi} - P_{\pi}\big)V^{\pi} +$$

Key idea 1: a peeling argument

Agarwal, Kakade, Yang 19: first-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma(I - \gamma P_{\pi})^{-1}(\widehat{P}_{\pi} - P_{\pi})\widehat{V}^{\pi} \qquad (\star)$$

Ours: higher-order expansion $\longrightarrow$ tighter control

$$\begin{aligned}\widehat{V}^{\pi} - V^{\pi} = \gamma(I - \gamma P_{\pi})^{-1}(\widehat{P}_{\pi} - P_{\pi}){\color{red}V^{\pi}} + \\ + \gamma(I - \gamma P_{\pi})^{-1}(\widehat{P}_{\pi} - P_{\pi})\,{\color{red}\left(\widehat{V}^{\pi} - V^{\pi}\right)}\end{aligned}$$

Key idea 1: a peeling argument

Agarwal, Kakade, Yang 19: first-order expansion

$$\widehat{V}^{\pi} - V^{\pi} = \gamma(I - \gamma P_{\pi})^{-1}(\widehat{P}_{\pi} - P_{\pi})\widehat{V}^{\pi} \qquad (\star)$$

Ours: higher-order expansion $\longrightarrow$ tighter control

$$\begin{aligned}\widehat{V}^{\pi} - V^{\pi} = \gamma(I - \gamma P_{\pi})^{-1}(\widehat{P}_{\pi} - P_{\pi})V^{\pi} + \\ + \gamma^2\Big((I - \gamma P_{\pi})^{-1}(\widehat{P}_{\pi} - P_{\pi})\Big)^2 V^{\pi} \\ + \gamma^3\Big((I - \gamma P_{\pi})^{-1}(\widehat{P}_{\pi} - P_{\pi})\Big)^3 V^{\pi} \\ + \dots\end{aligned}$$

Step 2: controlling $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

A natural idea: apply our policy evaluation theory + union bound

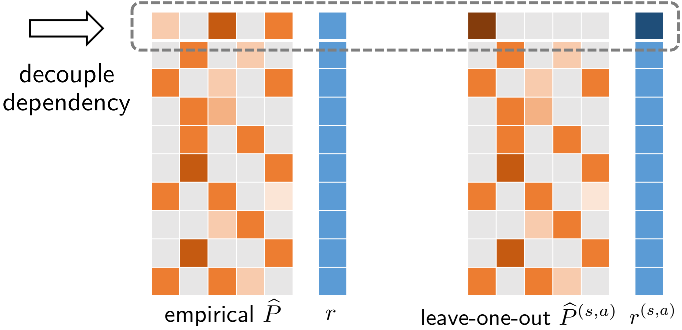
Step 2: controlling $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

A natural idea: apply our policy evaluation theory + union bound

- highly suboptimal! (there are exponentially many policies)

Key idea 2: leave-one-out analysis

Decouple dependency by introducing auxiliary state-action absorbing MDPs by dropping randomness for each (s, a)

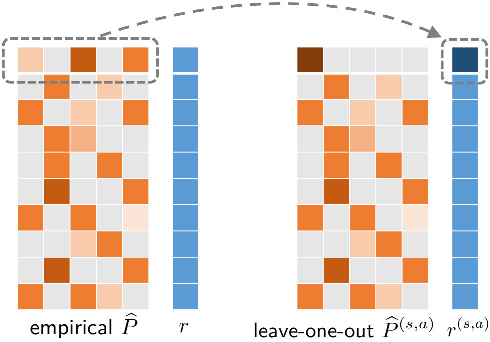


— *inspired by Agarwal et al. '19 but quite different . . .*

Key idea 2: leave-one-out analysis

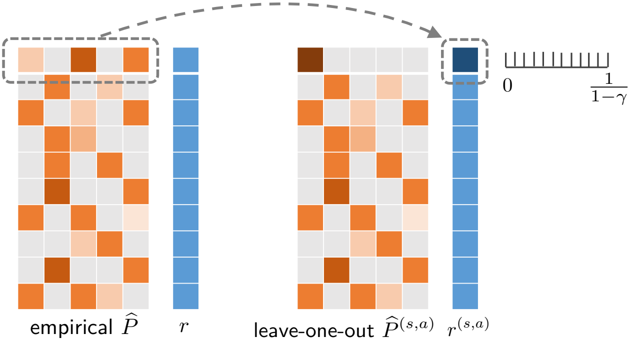
- Stein '72
- El Karoui, Bean, Bickel, Lim, Yu '13
- El Karoui '15
- Javanmard, Montanari '15
- Zhong, Boumal '17
- Lei, Bickel, El Karoui '17
- Sur, Chen, Candès '17
- Abbe, Fan, Wang, Zhong '17
- Chen, Fan, Ma, Wang '17
- Ma, Wang, Chi, Chen '17
- Chen, Chi, Fan, Ma '18
- Ding, Chen '18
- Dong, Shi '18
- Chen, Chi, Fan, Ma, Yan '19
- Chen, Fan, Ma, Yan '19
- Cai, Li, Poor, Chen '19
- Agarwal, Kakade, Yang '19
- Pananjady, Wainwright '19
- Ling '20

Key idea 2: leave-one-out analysis



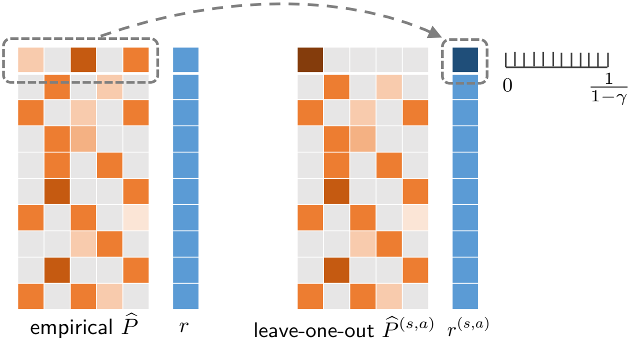
1. embed all randomness from $\widehat{P}_{s,a}$ into a single scalar (i.e. $r_{s,a}^{(s,a)}$)

Key idea 2: leave-one-out analysis



1. embed all randomness from $\widehat{P}_{s,a}$ into a single scalar (i.e. $r_{s,a}^{(s,a)}$)
2. build an ϵ-net for this scalar

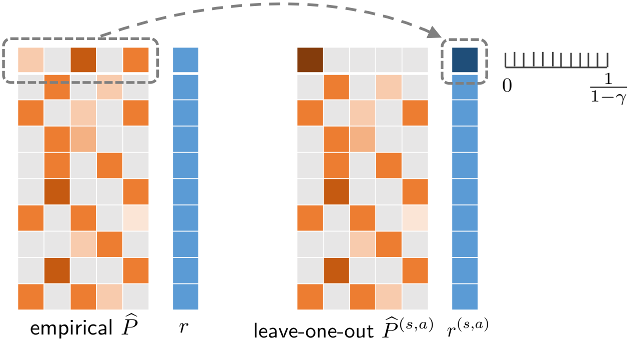
Key idea 2: leave-one-out analysis



1. embed all randomness from $\widehat{P}_{s,a}$ into a single scalar (i.e. $r_{s,a}^{(s,a)}$)
2. build an ϵ-net for this scalar
3. $\widehat{\pi}^\star$ can be determined by this ϵ-net under separation condition

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^\star(s, \widehat{\pi}^\star(s)) - \max_{a:\, a \neq \widehat{\pi}^\star(s)} \widehat{Q}^\star(s,a) > 0$$

Key idea 2: leave-one-out analysis



Our decoupling argument vs. Agarwal, Kakade, Yang '19

- *Agarwal et al. '19:* dependency btw value $\widehat{V}$ & samples
- *Ours:* dependency btw policy $\widehat{\pi}$ & samples

Key idea 3: tie-breaking via perturbation

- How to ensure separation between the optimal policy and others?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a:\, a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s, a) > 0$$

Key idea 3: tie-breaking via perturbation

- How to ensure separation between the optimal policy and others?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a:\, a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s, a) > 0$$

- **Solution:** *slightly perturb rewards* $r \implies \widehat{\pi}^{\star}_{\mathrm{p}}$
  - *ensures* $\widehat{\pi}^{\star}_{\mathrm{p}}$ *can be differentiated from others*
  - $V^{\widehat{\pi}^{\star}_{\mathrm{p}}} \approx V^{\widehat{\pi}^{\star}}$

Key idea 3: tie-breaking via perturbation

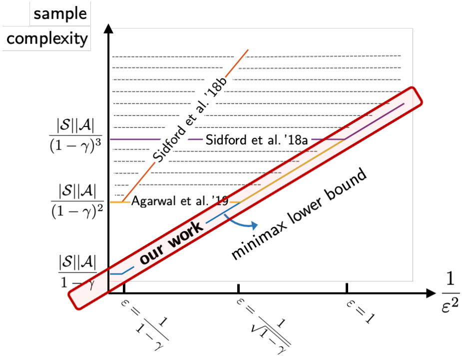
- How to ensure separation between the optimal policy and others?

$$\forall s \in \mathcal{S}, \quad \widehat{Q}^{\star}(s, \widehat{\pi}^{\star}(s)) - \max_{a:\, a \neq \widehat{\pi}^{\star}(s)} \widehat{Q}^{\star}(s,a) > \frac{(1-\gamma)\varepsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$$

- **Solution:** *slightly perturb rewards* $r \implies \widehat{\pi}^{\star}_{\mathrm{p}}$
  - *ensures* $\widehat{\pi}^{\star}_{\mathrm{p}}$ *can be differentiated from others*
  - $V^{\widehat{\pi}^{\star}_{\mathrm{p}}} \approx V^{\widehat{\pi}^{\star}}$

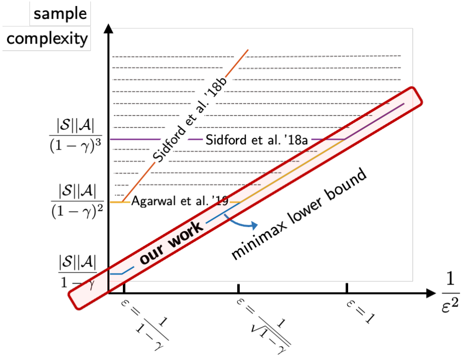
Summary

Model-based RL is minimax optimal and does not suffer from a sample size barrier!



Summary

Model-based RL is minimax optimal and does not suffer from a sample size barrier!



future directions

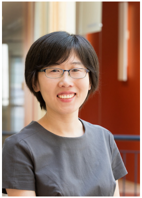
- finite-horizon episodic MDPs
- Markov games

Story 2: fast global convergence of entropy-regularized
natural policy gradient (NPG) methods

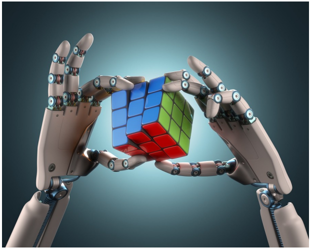
Shicong Cen
CMU ECE

Chen Cheng
Stanford Stats

Yuting Wei
CMU Stats



Yuejie Chi
CMU ECE



Policy optimization: a major contributor to these successes

Policy gradient (PG) methods

Given initial state distribution $s \sim \rho$:

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

Policy gradient (PG) methods

Given initial state distribution $s \sim \rho$:

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

$$\Downarrow$$

softmax parameterization:

$$\pi_\theta(a|s) = \frac{\exp(\theta(s,a))}{\sum_a \exp(\theta(s,a))}$$

Policy gradient (PG) methods

Given initial state distribution $s \sim \rho$:

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

$$\Downarrow$$

softmax parameterization:

$$\pi_\theta(a|s) = \frac{\exp(\theta(s,a))}{\sum_a \exp(\theta(s,a))}$$

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

Policy gradient (PG) methods

Given initial state distribution $s \sim \rho$:

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

$$\Downarrow$$

softmax parameterization:

$$\pi_\theta(a|s) = \frac{\exp(\theta(s,a))}{\sum_a \exp(\theta(s,a))}$$

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

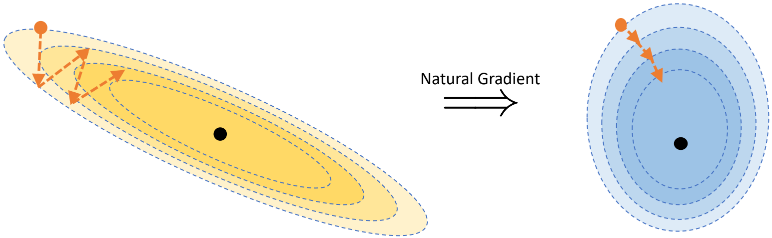
PG method (Sutton et al. '00)

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta V^{\pi_\theta^{(t)}}(\rho), \qquad t = 0, 1, \cdots$$

- η: learning rate

Booster 1: natural policy gradient (NPG)

precondition gradients to improve search directions ...



NPG method (Kakade '02)

$$\theta^{(t+1)} = \theta^{(t)} + \eta(\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho), \qquad t = 0, 1, \cdots$$

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

Booster 2: entropy regularization

accelerate convergence by regularizing objective function

$$
\begin{aligned}
V_\tau^\pi(s_0) &:= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r_t - \tau \log \pi(a_t|s_t)\big) \,\Big|\, s_0\right] \\
&= V^\pi(s) + \frac{\tau}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_s^\pi} \Big[\underbrace{-\sum_a \pi(a|s)\log \pi(a|s)}_{\text{entropy}} \,\Big|\, s_0\Big]
\end{aligned}
$$

- τ: regularization parameter
- d_s^π: discounted state visitation distribution

Booster 2: entropy regularization

accelerate convergence by regularizing objective function

$$V_\tau^\pi(s_0) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r_t - \tau \log \pi(a_t|s_t)\big) \,\Big|\, s_0\right]$$
$$= V^\pi(s) + \frac{\tau}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_s^\pi} \Big[\underbrace{-\sum_a \pi(a|s) \log \pi(a|s)}_{\text{entropy}} \,\Big|\, s_0\Big]$$

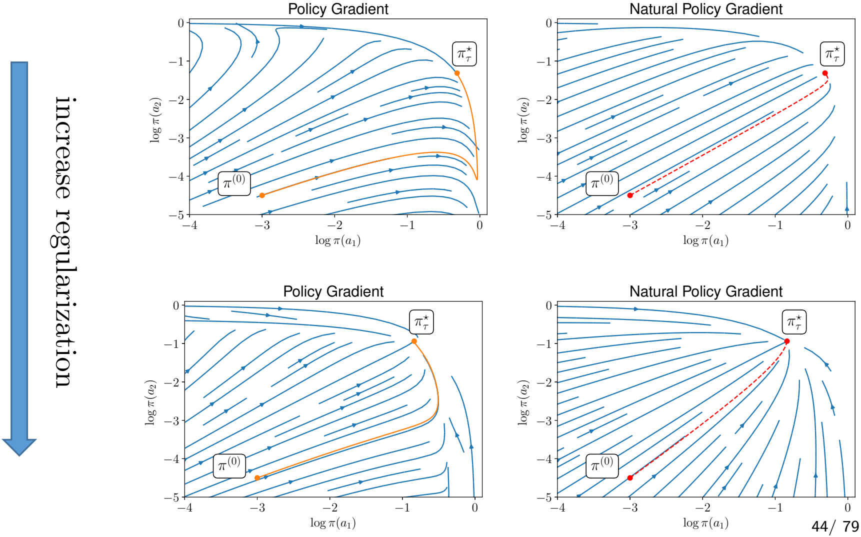
- τ: regularization parameter
- d_s^π: discounted state visitation distribution

entropy-regularized value maximization

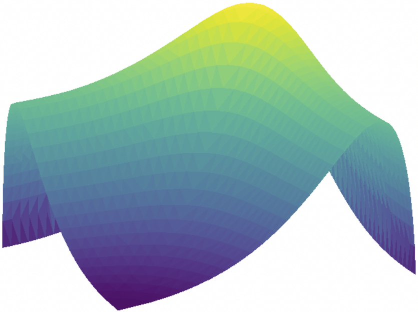
$$\text{maximize}_\theta \quad V_\tau^{\pi_\theta}(\rho) := \mathbb{E}_{s\sim\rho}\left[V_\tau^{\pi_\theta}(s)\right] \qquad \text{(“soft” value function)}$$

Entropy-regularized natural gradient helps!

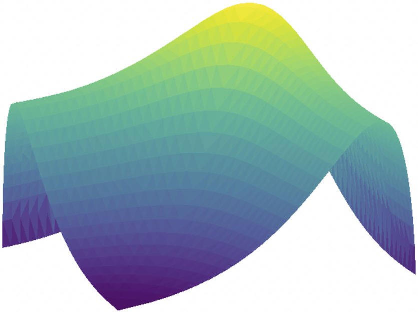
A toy bandit example: 3 arms with rewards 1, 0.9 and 0.1



Challenge: non-concavity



Challenge: non-concavity

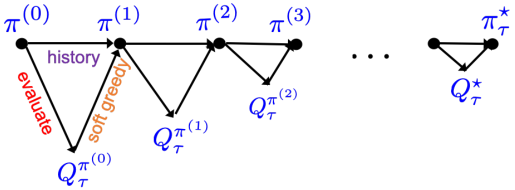


Recent advances

- PG for control (Fazel et al., 2018; Bhandari and Russo, 2019)
- PG for tabular MDPs (Agarwal et al. 19, Bhandari and Russo '19, Mei et al '20)
- unregularized NPG for tabular MDPs (Agarwal et al. '19, Bhandari and Russo '20)
- ...

This work: understanding entropy-regularized NPG methods in tabular settings

Entropy-regularized NPG in tabular settings



An alternative expression in policy space (tabular setting)

$$\pi^{(t+1)}(a|s) \propto \pi^{(t)}(a|s)^{1-\frac{\eta\tau}{1-\gamma}} \exp\left(\frac{\eta Q_\tau^{(t)}(s,a)}{1-\gamma}\right), \qquad t = 0, 1, \cdots$$

- $Q_\tau^{(t)}$: soft Q-function of $\pi^{(t)}$; $\quad 0 < \eta \le \frac{1-\gamma}{\tau}$: learning rate

- invariant to the choice of initial state distribution ρ

Linear convergence with exact gradients

optimal policy: $\pi_\tau^\star$*; optimal "soft" Q function:* $Q_\tau^\star := Q_\tau^{\pi_\tau^\star}$

Exact oracle: perfect gradient evaluation

Theorem 3 (Cen, Cheng, Chen, Wei, Chi '20)

For any $0 < \eta \leq (1-\gamma)/\tau$, *entropy-regularized NPG achieves*

$$\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq C_1 \gamma (1-\eta\tau)^t, \quad t = 0, 1, \cdots$$

- $C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\tau \left(1 - \frac{\eta\tau}{1-\gamma}\right) \| \log \pi_\tau^\star - \log \pi^{(0)}\|_\infty$

Implications: iteration complexity

number of iterations needed to reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le \varepsilon$ is at most

- **General learning rates ($0 < \eta < \frac{1-\gamma}{\tau}$):**

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

- **Soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$):**

$$\frac{1}{1-\gamma} \log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\varepsilon}\right)$$

Implications: iteration complexity

number of iterations needed to reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \varepsilon$ is at most

- **General learning rates ($0 < \eta < \frac{1-\gamma}{\tau}$):**

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

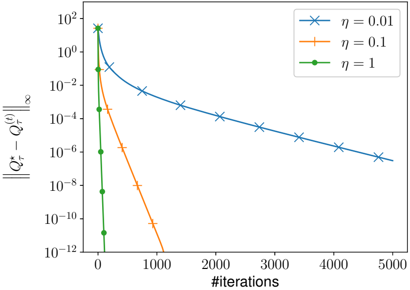
- **Soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$):**

$$\frac{1}{1-\gamma} \log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\varepsilon}\right)$$

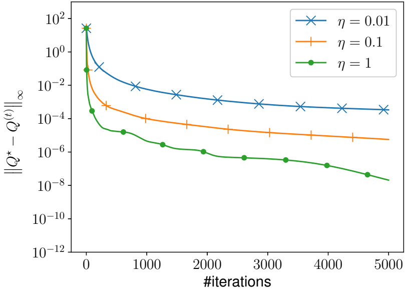
Nearly dimension-free global linear convergence!

Regularized NPG vs. unregularized NPG

regularized NPG
$\tau = 0.001$



unregularized NPG
$\tau = 0$



linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\varepsilon}\right)$

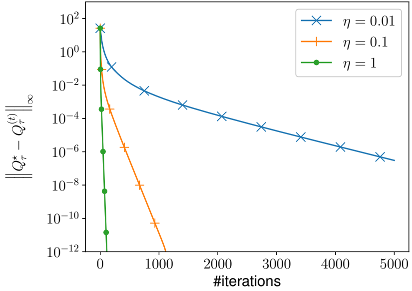
Ours

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

(Agarwal et al. '19)

Regularized NPG vs. unregularized NPG

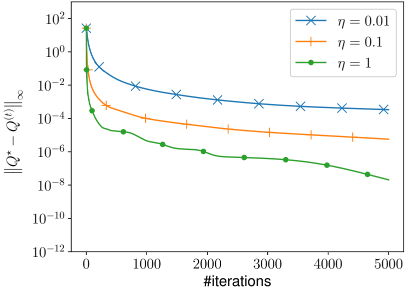
regularized NPG
$\tau = 0.001$



linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\varepsilon}\right)$

Ours

unregularized NPG
$\tau = 0$



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

(Agarwal et al. '19)

Entropy regularization enables faster convergence!

Returning to the original MDP?

How to employ entropy-regularized NPG to find an ε-optimal policy for the original (unregularized) MDP?

- suffices to find an $\frac{\varepsilon}{2}$-optimal policy of regularized MDP w/ regularization parameter $\tau = \frac{(1-\gamma)\varepsilon}{4\log|\mathcal{A}|}$
- iteration complexity is the same as before (up to log factor)

Entropy-regularized NPG with inexact gradients

Inexact oracle: inexact evaluation of $Q_\tau^{(t)}$, which returns $\widehat{Q}_\tau^{(t)}$ s.t.

$$\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \leq \delta,$$

e.g. using sample-based estimators

Entropy-regularized NPG with inexact gradients

Inexact oracle: inexact evaluation of $Q_\tau^{(t)}$, which returns $\widehat{Q}_\tau^{(t)}$ s.t.

$$\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \le \delta,$$

e.g. using sample-based estimators

Inexact entropy-regularized NPG:

$$\pi^{(t+1)}(a|s) \;\propto\; \big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}} \exp\Big(\frac{\eta \widehat{Q}_\tau^{(t)}(s,a)}{1-\gamma}\Big)$$

Entropy-regularized NPG with inexact gradients

Inexact oracle: inexact evaluation of $Q_\tau^{(t)}$, which returns $\widehat{Q}_\tau^{(t)}$ s.t.

$$\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \leq \delta,$$

e.g. using sample-based estimators

Inexact entropy-regularized NPG:

$$\pi^{(t+1)}(a|s) \;\propto\; \big(\pi^{(t)}(a|s)\big)^{1-\frac{\eta\tau}{1-\gamma}} \exp\Big(\frac{\eta \widehat{Q}_\tau^{(t)}(s,a)}{1-\gamma}\Big)$$

Question: stability vis-à-vis inexact gradient evaluation?

Linear convergence with inexact gradients

$$\big\|\widehat{Q}_\tau^{(t)} - Q_\tau^{(t)}\big\|_\infty \le \delta$$

Theorem 4 (Cen, Cheng, Chen, Wei, Chi '20)

For any stepsize $0 < \eta \le (1-\gamma)/\tau$*, entropy-regularized NPG attains*

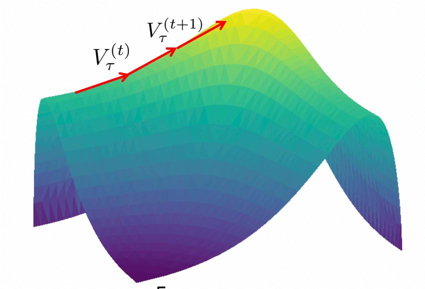
$$\big\|Q_\tau^\star - Q_\tau^{(t+1)}\big\|_\infty \le \gamma(1-\eta\tau)^t C_1 + C_2$$

- $C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\tau\left(1 - \frac{\eta\tau}{1-\gamma}\right)\|\log \pi_\tau^\star - \log \pi^{(0)}\|_\infty$
- $C_2 = \frac{2\gamma\left(1 + \frac{\gamma}{\eta\tau}\right)}{1-\gamma}\delta$: error floor

- converges linearly at the same rate until an error floor is hit

A little analysis when $\eta = \frac{1-\gamma}{\tau}$

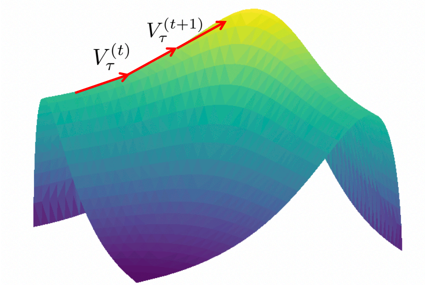
A key lemma: monotonic performance improvement



$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathop{\mathbb{E}}_{s \sim d_\rho^{(t+1)}} \Bigg[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right) \underbrace{\mathsf{KL}\Big(\pi^{(t+1)}(\cdot|s) \,\Big\|\, \pi^{(t)}(\cdot|s)\Big)}_{\text{KL divergence}}$$

$$+ \frac{1}{\eta} \underbrace{\mathsf{KL}\left(\pi^{(t)}(\cdot|s) \,\Big\|\, \pi^{(t+1)}(\cdot|s)\right)}_{\text{KL divergence}} \Bigg]$$

A key lemma: monotonic performance improvement



$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}} \Bigg[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right) \underbrace{\mathsf{KL}\Big(\pi^{(t+1)}(\cdot|s) \,\big\|\, \pi^{(t)}(\cdot|s)\Big)}_{\text{KL divergence}}$$

$$+ \frac{1}{\eta} \underbrace{\mathsf{KL}\left(\pi^{(t)}(\cdot|s) \,\big\|\, \pi^{(t+1)}(\cdot|s)\right)}_{\text{KL divergence}} \Bigg]$$

$$\geq 0 \qquad \left(\text{if } 0 < \eta \leq \frac{1-\gamma}{\tau}\right)$$

“Soft” Bellman operator

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

“Soft” Bellman operator

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

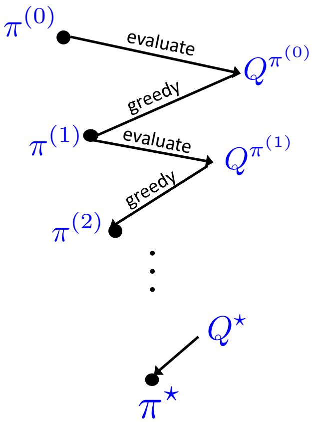
Soft Bellman equation: $Q_\tau^\star$ is the *unique* solution to

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

γ**-contraction of soft Bellman operator:**

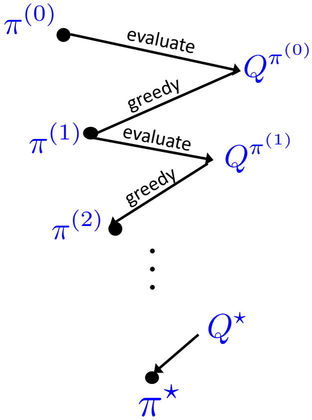
$$\|\mathcal{T}_\tau(Q_1) - \mathcal{T}_\tau(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$$

policy iteration



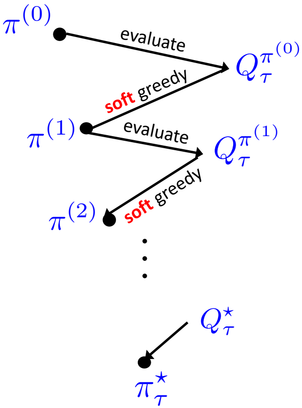
Bellman operator

policy iteration



Bellman operator

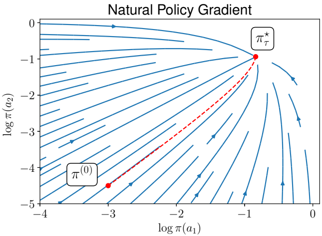
soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

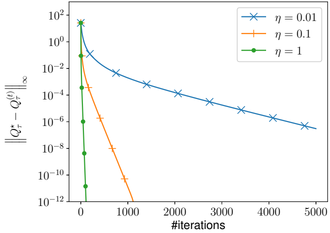


soft Bellman operator

Summary

Global linear convergence of entropy-regularized NPG methods for tabular discounted MDPs

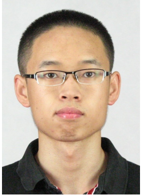




future directions:

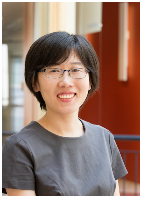
- function approximation
- sample complexities
- soft actor-critic algorithms

Story 3: sample complexity of *(**asynchronous) Q-learning** on Markovian samples*

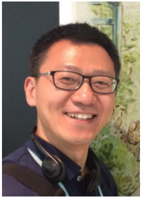


Gen Li
Tsinghua EE

Yuting Wei
CMU Stats

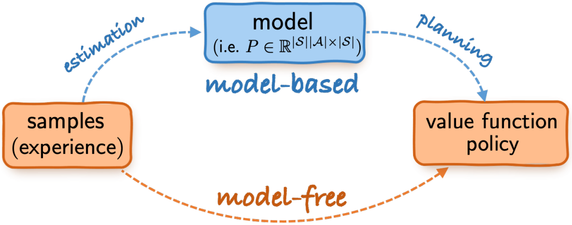


Yuejie Chi
CMU ECE



Yuantao Gu
Tsinghua EE

Model-based vs. model-free RL

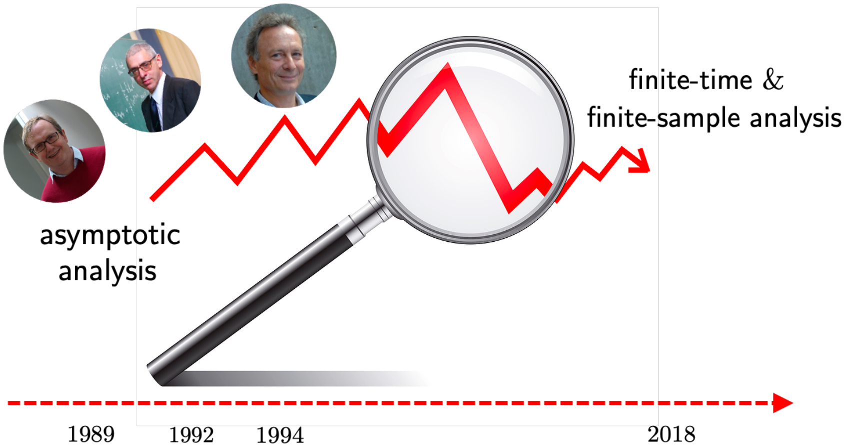


Model-based approach ("plug-in")

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

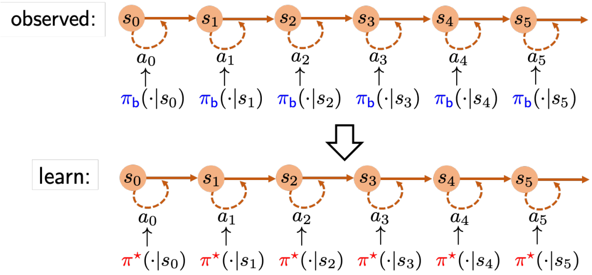
Model-free approach

— learning w/o modeling & estimating environment explicitly



A classical example: **Q-learning** on Markovian samples

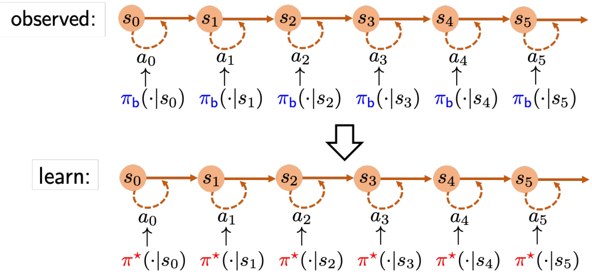
Markovian samples and behavior policy



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

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

- mixing time: t_{mix}

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

$\underbrace{\text{Stochastic approximation}}_{\text{Robbins \& Monro '51}}$ for solving **Bellman equation** $Q = \mathcal{T}(Q)$

Aside: 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

Aside: 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

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

Richard Bellman

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

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

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

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

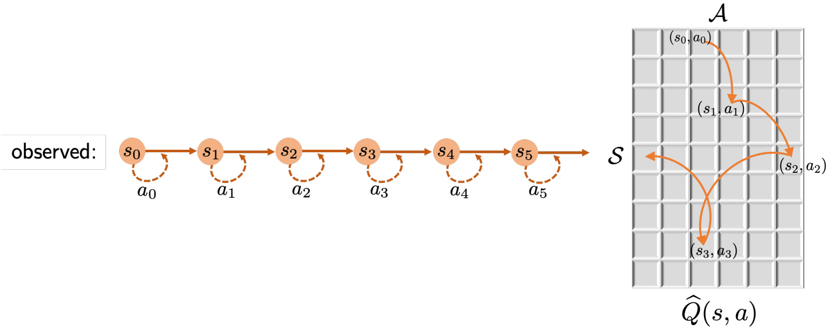
Stochastic approximation for solving Bellman equation $Q = \mathcal{T}(Q)$

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

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

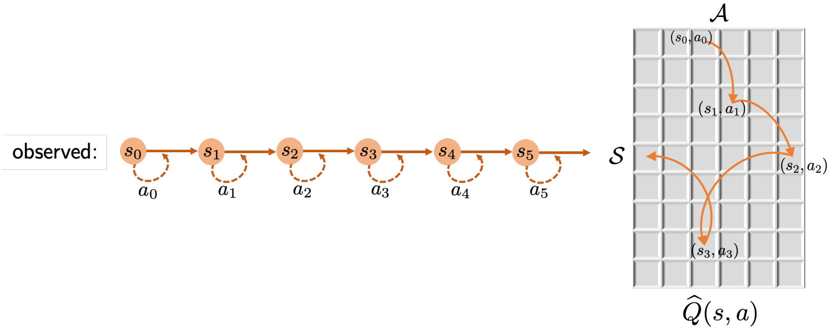
$$\mathcal{T}(Q)(s, a) = r(s, a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s', a') \right]$$

Q-learning on Markovian samples



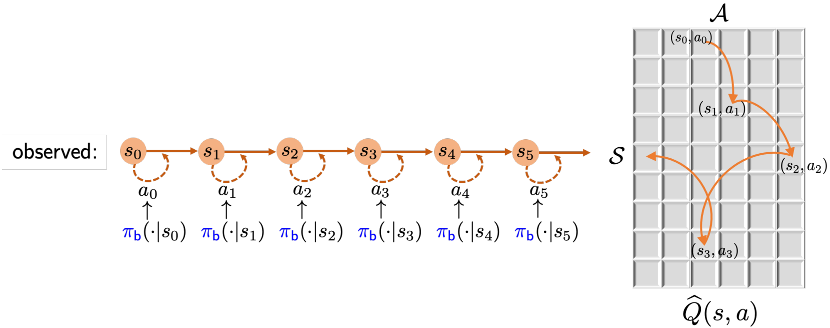
- **asynchronous:** only a single entry is updated each iteration

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?

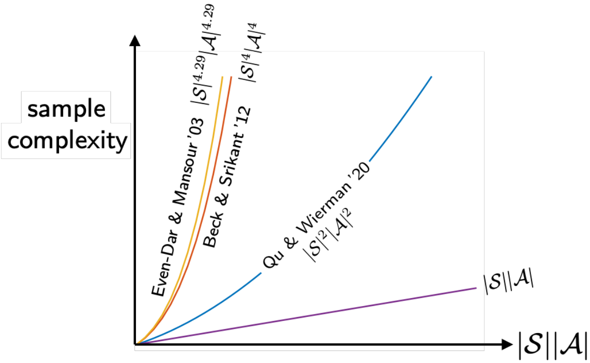
Prior art: async Q-learning

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	$\Big(\frac{t_{\mathsf{cover}}^{1+3\omega}}{(1-\gamma)^4\varepsilon^2}\Big)^{\frac{1}{\omega}} + \Big(\frac{t_{\mathsf{cover}}}{1-\gamma}\Big)^{\frac{1}{1-\omega}}$	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

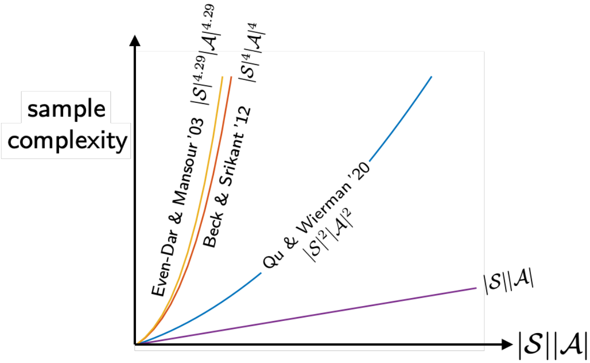
Question: how many samples are needed to ensure $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$?



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

Prior art: async Q-learning

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

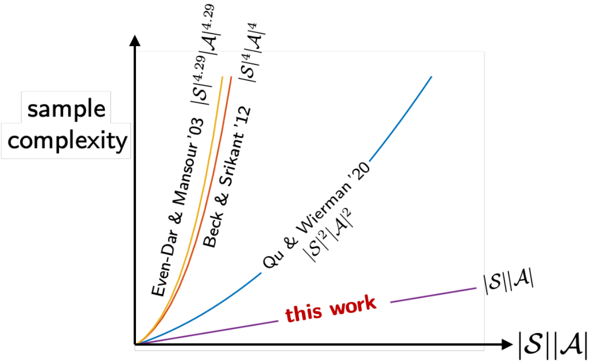


if we take $\mu_{\mathsf{min}} \asymp \frac{1}{|\mathcal{S}||\mathcal{A}|}$, $t_{\mathsf{cover}} \asymp \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}}$

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

Prior art: async Q-learning

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



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

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 5 (Li, Wei, Chi, Gu, Chen '20)

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)

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

Main result: ℓ_∞-based sample complexity

Theorem 5 (Li, Wei, Chi, Gu, Chen '20)

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)*

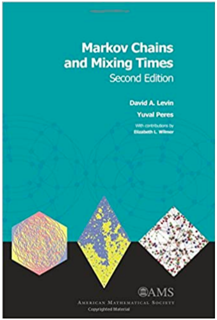
$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

- 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

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

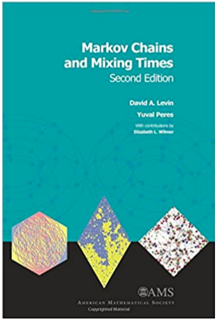


- reflects cost taken to reach steady state
- one-time expense (almost independent of ε)

  — it becomes amortized as algorithm runs

Effect of mixing time on sample complexity

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$



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

  — it becomes amortized as algorithm runs

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

Learning rates

Our choice: constant stepsize $\eta_t \equiv \min\left\{\frac{(1-\gamma)^4\varepsilon^2}{\gamma^2}, \frac{1}{t_{\mathsf{mix}}}\right\}$

- Qu & Wierman '20: rescaled linear $\eta_t = \frac{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)}}{t+\max\{\frac{1}{\mu_{\mathsf{min}}(1-\gamma)}, t_{\mathsf{mix}}\}}$
- Beck & Srikant '12: constant $\underbrace{\eta_t \equiv \frac{(1-\gamma)^4\varepsilon^2}{|\mathcal{S}||\mathcal{A}|t_{\mathsf{cover}}^2}}_{\text{too conservative}}$
- Even-Dar & Mansour '03: polynomial $\eta_t = t^{-\omega}$ $(\omega \in (\frac{1}{2}, 1])$

Minimax lower bound

minimax lower bound
(Azar et al. '13)

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2}$$

asyn Q-learning
(ignoring dependency on t_{mix})

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}$$

Minimax lower bound

minimax lower bound
(Azar et al. '13)

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2}$$

asyn Q-learning
(ignoring dependency on t_{mix})

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^5\varepsilon^2}$$

Can we improve dependency on **discount complexity** $\frac{1}{1-\gamma}$?

One strategy: variance reduction

— *inspired by Johnson & Zhang '13, Wainwright '19*

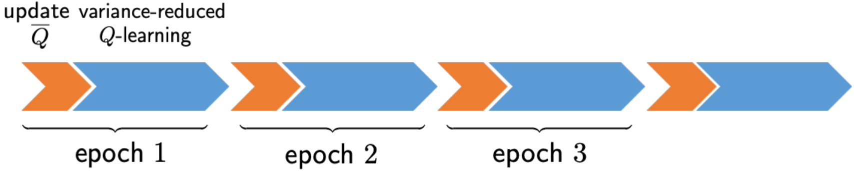
Variance-reduced Q-learning updates

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

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

Variance-reduced Q-learning

— *inspired by Johnson & Zhang '13, Sidford et al. '18, Wainwright '19*



for each epoch

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

Main result: ℓ_∞-based sample complexity

Theorem 6 (Li, Wei, Chi, Gu, Chen '20)

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

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

- more aggressive learning rates: $\eta_t \equiv \min\left\{\frac{\cancel{(1-\gamma)^4}(1-\gamma)^2}{\gamma^2}, \frac{1}{t_{\mathsf{mix}}}\right\}$

Main result: ℓ_∞-based sample complexity

Theorem 6 (Li, Wei, Chi, Gu, Chen '20)

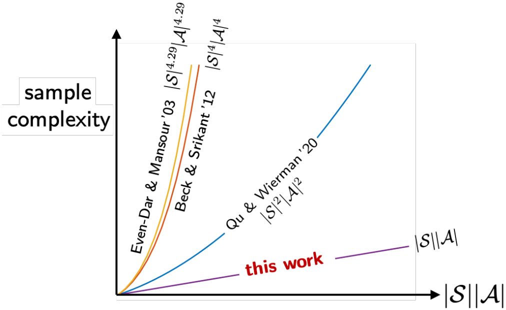
For any $0 < \varepsilon \le 1$, sample complexity for ***(async) variance-reduced Q-learning*** *to yield $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ is at most on the order of*

$$\frac{1}{\mu_{\mathsf{min}}(1-\gamma)^3\varepsilon^2} + \frac{t_{\mathsf{mix}}}{\mu_{\mathsf{min}}(1-\gamma)}$$

- more aggressive learning rates: $\eta_t \equiv \min\left\{\frac{\cancel{(1-\gamma)^4}(1-\gamma)^2}{\gamma^2}, \frac{1}{t_{\mathsf{mix}}}\right\}$
- minimax-optimal for $0 < \varepsilon \le 1$

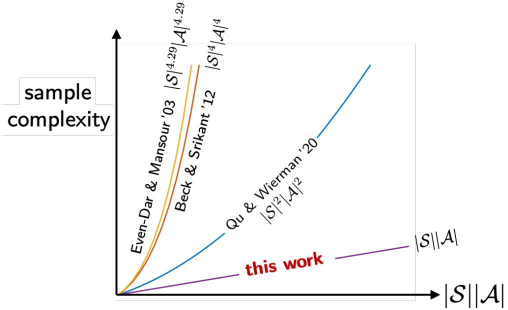
Summary

Sharpens finite-sample understanding of Q-learning on Markovian data



Summary

Sharpens finite-sample understanding of Q-learning on Markovian data

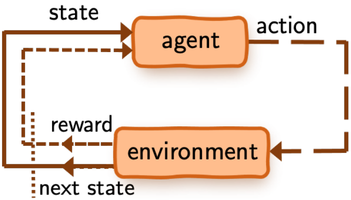


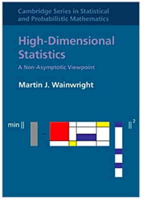
future directions

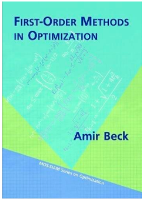
- function approximation
- on-policy algorithms like SARSA
- general Markov-chain-based optimization algorithms

Concluding remarks

Understanding RL requires modern statistics and optimization

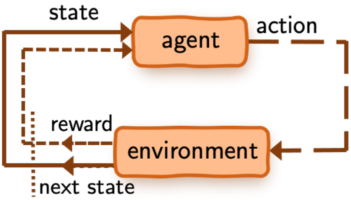


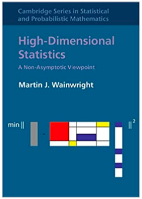


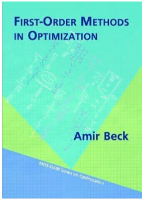


Concluding remarks

Understanding RL requires modern statistics and optimization







future directions

- beyond tabular settings
- finite-horizon episodic MDPs
- multi-agent RL (e.g. Markov games)
- ...

Papers:

"Breaking the sample size barrier in model-based reinforcement learning with a generative model," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, NeurIPS, 2020

"Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, NeurIPS 2020

"Fast global convergence of natural policy gradient methods with entropy regularization," S. Cen, C. Cheng, Y. Chen, Y. Wei, Y. Chi, arxiv:2007.06558, 2020