Non-asymptotic Statistical and Computational Guarantees of Reinforcement Learning Algorithms

Yuejie Chi

Carnegie Mellon University

2021 Goldsmith Lecture

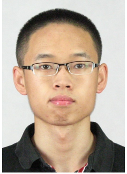
Special thanks to...

Dean Andrea Goldsmith
Princeton University

My wonderful collaborators

Shicong Cen
CMU

Chen Cheng
Stanford

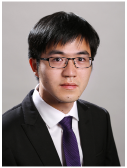


Gen Li
Princeton

Yuxin Chen
Princeton

Yuting Wei
UPenn

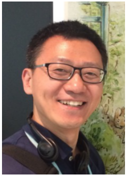
Laixi Shi
CMU



Changxiao Cai
UPenn

Wenhao Zhan
Princeton

Jason Lee
Princeton



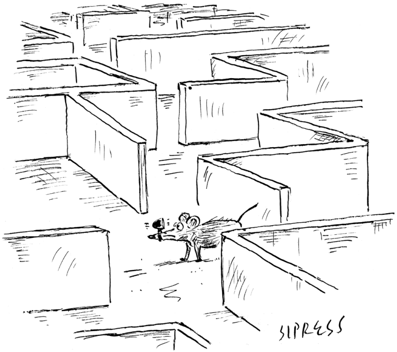
Yuantao Gu
Tsinghua

Many thanks to Y. Wei and Y. Chen for significant contributions to the slides.

Reinforcement learning (RL)

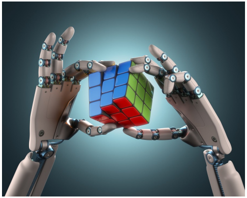
In RL, an agent learns by interacting with an environment.

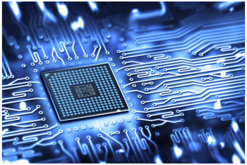
- unknown environments
- maximize total rewards
- trial-and-error
- sequential and online

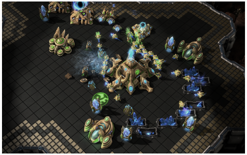


"Recalculating ... recalculating ..."

Recent successes in RL



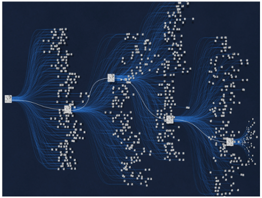




RL holds great promise in the next era of artificial intelligence.

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

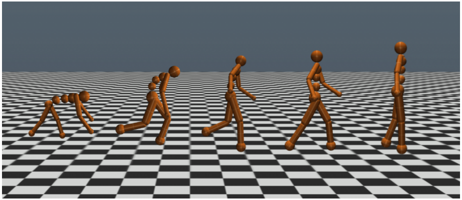
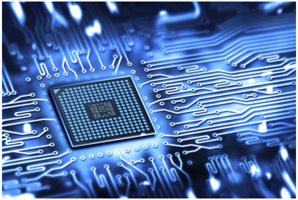
autonomous driving

online ads

Calls for design of sample-efficient RL algorithms!

Computational efficiency

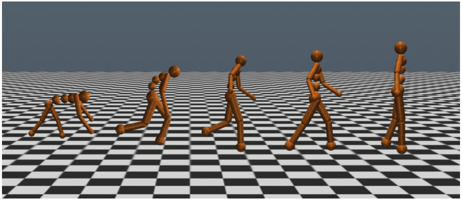
Running RL algorithms might take a long time and space

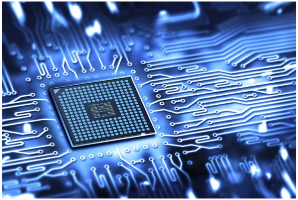
 

many CPUs / GPUs / TPUs + computing hours

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

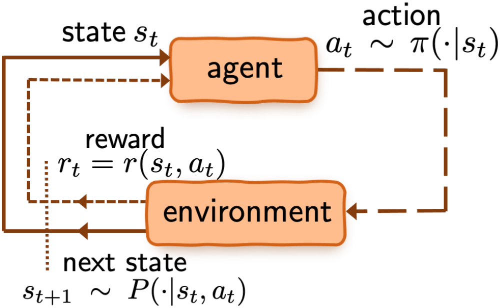
Non-asymptotic analyses are key to understand sample and computational efficiency in modern RL.

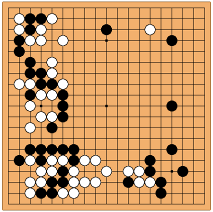
This tutorial

- Part I: backgrounds and basics
  - Markov decision processes
  - Planning
- Part II: statistical guarantees under the generative model
  - minimax lower bound
  - Is model-based RL minimax optimal?
  - Is Q-learning minimax optimal?
- Part III: computational guarantees of policy optimization
  - (natural) policy gradient methods
  - finite-time rate of global convergence
  - entropy regularization and beyond
- Part IV: concluding remarks and further pointers

Part I: backgrounds and basics

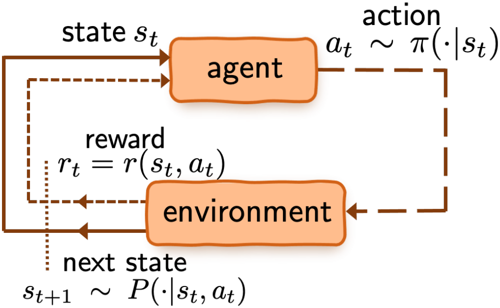
Markov decision process (MDP)

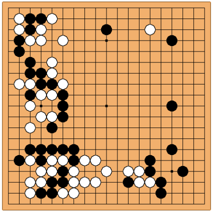




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

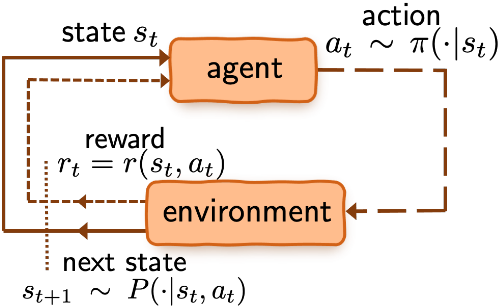
Markov decision process (MDP)

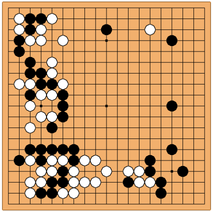




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

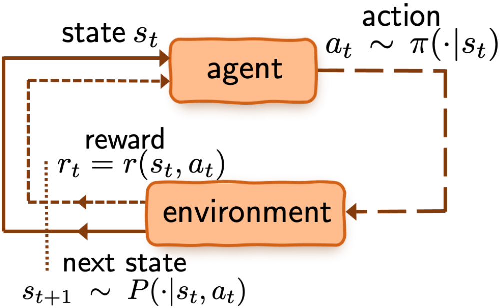
Markov decision process (MDP)

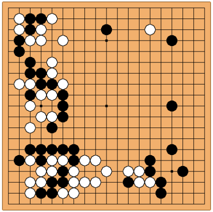




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

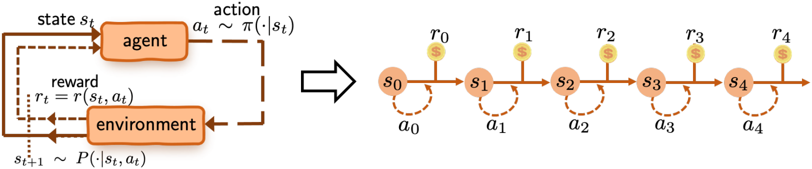
Markov decision process (MDP)





- $\mathcal{S}$: 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

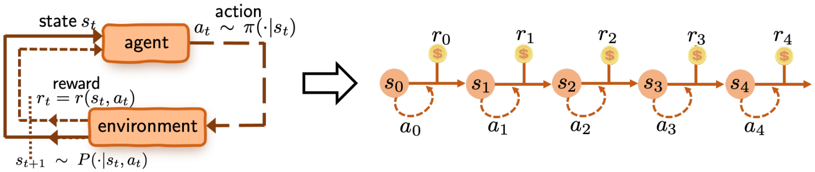
Value function



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

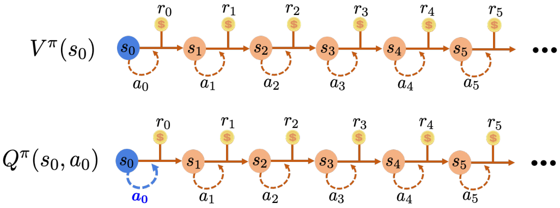


Value function of policy π:

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

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

Q-function

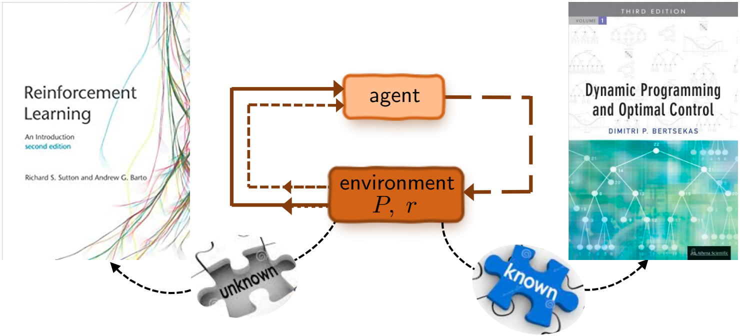


Q-function of policy π:

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

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

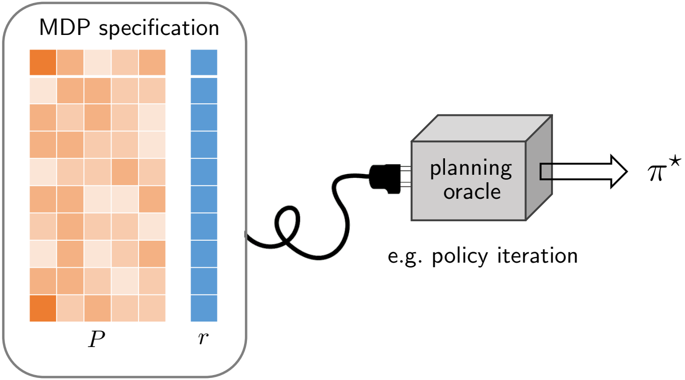
Searching for the optimal policy



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 $\pi^\star(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, Q^\star(s,a)$

Planning: when the model is known



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

Policy evaluation: Bellman's consistency equation

- V^π / Q^π: value / action-value function under policy π

Bellman's consistency equation

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right]$$

$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s'\sim P(\cdot|s,a)}\Big[\underbrace{V^\pi(s')}_{\text{next state's value}}\Big]$$

- one-step look-ahead
- Let P^π be the state-action transition matrix induced by π:

$$Q^\pi = r + \gamma P^\pi Q^\pi \quad \Longrightarrow \quad Q^\pi = (I - \gamma P^\pi)^{-1} r$$

Richard Bellman

Bellman's optimality principle

Bellman operator

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

- 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

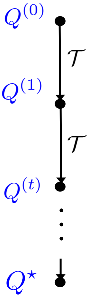
Bellman equation: $Q^\star$ is *unique* solution to

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

γ-contraction of Bellman operator:

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

Richard Bellman

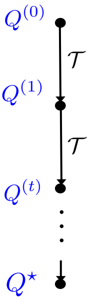


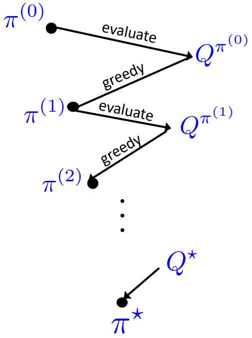
Value iteration (VI)

For $t = 0, 1, \ldots,$

$$Q^{(t+1)} = \mathcal{T}(Q^{(t)})$$

Value iteration and policy iteration





Value iteration (VI)

For $t = 0, 1, \ldots,$

$$Q^{(t+1)} = \mathcal{T}(Q^{(t)})$$

Policy iteration (PI)

For $t = 0, 1, \ldots,$

$$\pi^{(t)} = \mathsf{Greedy}(Q^{(t-1)})$$

$$Q^{(t)} = Q^{\pi^{(t)}}$$

Iteration complexity

Proposition (Linear convergence of policy/value iteration)

$$\|Q^{(t)} - Q^{\star}\|_{\infty} \le \gamma^{t} \|Q^{(0)} - Q^{\star}\|_{\infty}$$

Iteration complexity

Proposition (Linear convergence of policy/value iteration)

$$\|Q^{(t)} - Q^{\star}\|_{\infty} \le \gamma^{t} \|Q^{(0)} - Q^{\star}\|_{\infty}$$

Implications: to achieve $\|Q^{(t)} - Q^{\star}\|_{\infty} \le \epsilon$, it takes no more than

$$\frac{1}{1-\gamma} \log\left(\frac{\|Q^{(0)} - Q^{\star}\|_{\infty}}{\epsilon}\right)$$

iterations.

Iteration complexity

Proposition (Linear convergence of policy/value iteration)

$$\|Q^{(t)} - Q^\star\|_\infty \le \gamma^t \|Q^{(0)} - Q^\star\|_\infty$$

Implications: to achieve $\|Q^{(t)} - Q^\star\|_\infty \le \epsilon$, it takes no more than

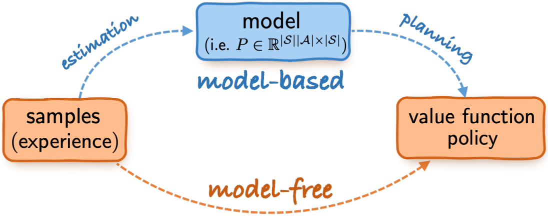
$$\frac{1}{1-\gamma} \log\left(\frac{\|Q^{(0)} - Q^\star\|_\infty}{\epsilon}\right)$$

iterations.

Linear convergence at a **dimension-free** rate!

Part II: statistical guarantees under the generative model

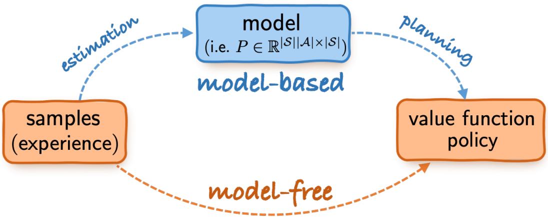
Two approaches to RL



Model-based approach ("plug-in")

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

Two approaches to RL



Model-based approach ("plug-in")

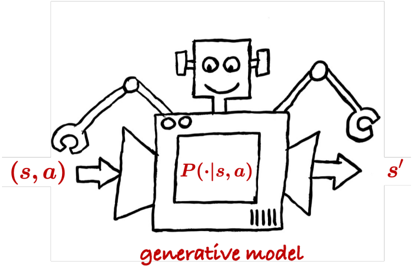
1. build an empirical estimate $\widehat{P}$ for P
2. planning based on empirical $\widehat{P}$

Model-free approach

— learning w/o constructing model explicitly

RL with a generative model / simulator

— *Kearns and Singh, 1999*

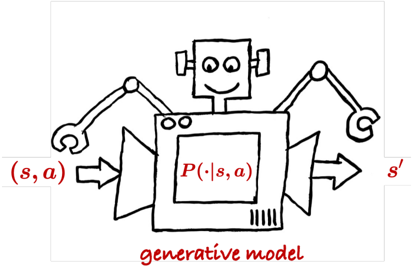


For each state-action pair (s,a), collect N samples

$$\{(s,a,s'_{(i)})\}_{1\le i\le N}$$

RL with a generative model / simulator

— *Kearns and Singh, 1999*



For each state-action pair (s, a), collect N samples

$$\{(s, a, s'_{(i)})\}_{1 \le i \le N}$$

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

Minimax lower bound

Theorem (minimax lower bound; Azar et al., 2013)

For all $\epsilon \in [0, \frac{1}{1-\gamma})$, there exists some MDP such that the total number of samples need to be at least

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

to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \epsilon$, where $\widehat{Q}$ is the output of any RL algorithm.

Minimax lower bound

Theorem (minimax lower bound; Azar et al., 2013)

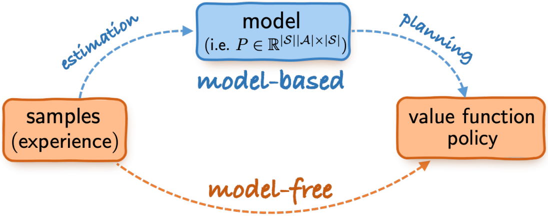
For all $\epsilon \in [0, \frac{1}{1-\gamma})$, there exists some MDP such that the total number of samples need to be at least

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

to achieve $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \epsilon$, where $\widehat{Q}$ is the output of any RL algorithm.

- holds for both finding the optimal Q-function and the optimal policy over the entire range of ϵ
- much smaller than the model dimension $|\mathcal{S}|^2|\mathcal{A}|$

Is model-based RL minimax optimal?



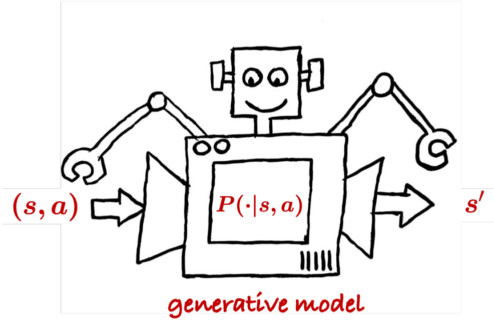
Model-based approach ("plug-in")

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

Model-free approach

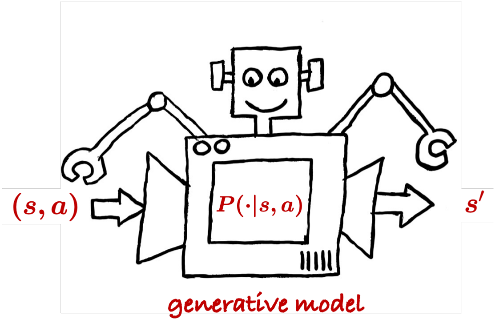
— learning w/o constructing model explicitly

Model estimation under the generative model



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

Model estimation under the generative model

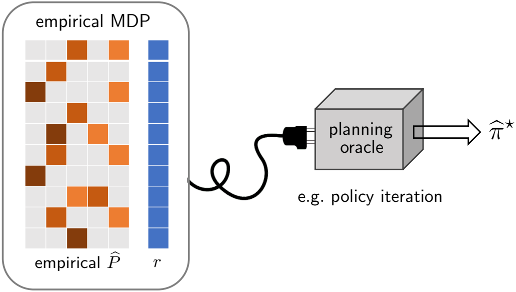


For each (s, a), collect N ind. samples $\{(s, a, s'_{(i)})\}_{1 \le i \le N}$

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

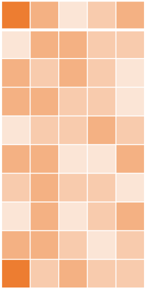
Model-based (plug-in) estimator

— *Azar et al., 2013; Agarwal et al., 2019*

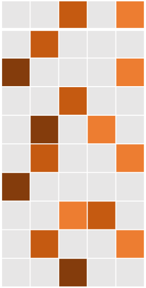


Run planning algorithms based on the *empirical* MDP

Challenges in the sample-starved regime



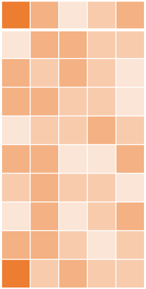
truth:
$P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$



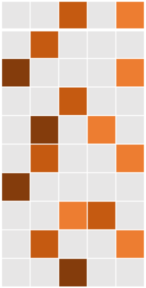
empirical estimate:
$\widehat{P}$

- 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?

Sample complexity of the plug-in estimator

Theorem (Azar et al., 2013)

For any $0 < \epsilon \leq 1$, *the optimal Q-function* $\widehat{Q}$ *of the empirical MDP achieves*

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

with sample complexity at most $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$.

- matches with the minimax lower bound whenever $\epsilon \in (0, 1]$.

Sample complexity of the plug-in estimator

Theorem (Azar et al., 2013)

For any $0 < \epsilon \leq 1$, the optimal Q-function $\widehat{Q}$ of the empirical MDP achieves

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

with sample complexity at most $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$.

- matches with the minimax lower bound whenever $\epsilon \in (0, 1]$.
- **Question:** Does it imply a near minimax-optimal policy $\widehat{\pi}$?

From Q-function to policy

Proposition (Singh and Yee, 1994)

Let the greedy policy w.r.t. $\widehat{Q}$ *be* $\widehat{\pi}$*, then*

$$V^\star - V^{\widehat{\pi}} \leq \frac{2}{1-\gamma} \|Q^\star - \widehat{Q}\|_\infty.$$

From Q-function to policy

Proposition (Singh and Yee, 1994)

Let the greedy policy w.r.t. $\widehat{Q}$ *be* $\widehat{\pi}$*, then*

$$V^\star - V^{\widehat{\pi}} \le \frac{2}{1-\gamma}\|Q^\star - \widehat{Q}\|_\infty.$$

$$\boxed{\|\widehat{Q} - Q^\star\|_\infty \le \epsilon} \xrightarrow{\widehat{\pi} = \mathrm{Greedy}(\widehat{Q})} \boxed{V^\star - V^{\widehat{\pi}} \le \frac{\epsilon}{1-\gamma}}$$

From Q-function to policy

Proposition (Singh and Yee, 1994)

Let the greedy policy w.r.t. $\widehat{Q}$ *be* $\widehat{\pi}$*, then*

$$V^\star - V^{\widehat{\pi}} \leq \frac{2}{1-\gamma}\|Q^\star - \widehat{Q}\|_\infty.$$

$$\boxed{\|\widehat{Q} - Q^\star\|_\infty \leq \epsilon} \xrightarrow{\widehat{\pi} = \text{Greedy}(\widehat{Q})} \boxed{V^\star - V^{\widehat{\pi}} \leq \frac{\epsilon}{1-\gamma}}$$

This **error amplification** has consequences in sample complexities.

- To reach ϵ-optimality, the greedy policy of a minimax-optimal Q-function estimator needs

$$\widetilde{\Omega}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\epsilon^2}\right)$$

samples invoking the above naive argument.

Sample complexity of the plug-in estimator

Theorem (Agarwal et al., 2019)

For any $0 < \epsilon \leq \frac{1}{\sqrt{1-\gamma}}$, *the optimal policy* $\widehat{\pi}^{\star}$ *of the empirical MDP achieves*

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

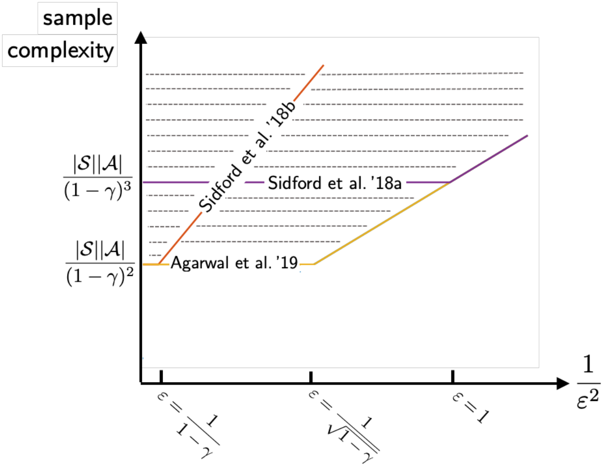
with sample complexity at most $\widetilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$.

- matches with the minimax lower bound whenever

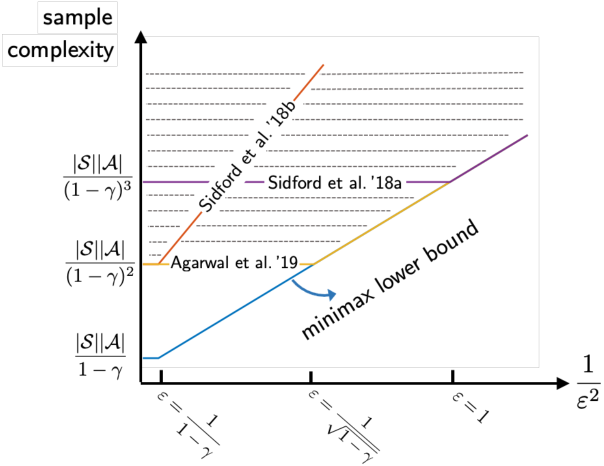
$$\epsilon \in (0, \frac{1}{\sqrt{1-\gamma}}].$$

- requires a sample size of at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$.

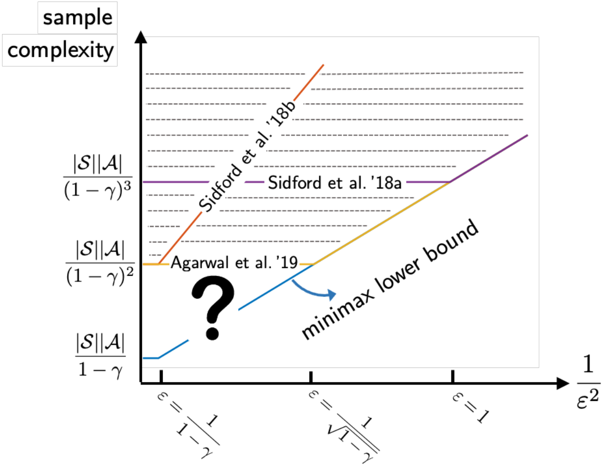
A benchmark of the prior art



A benchmark of the prior art

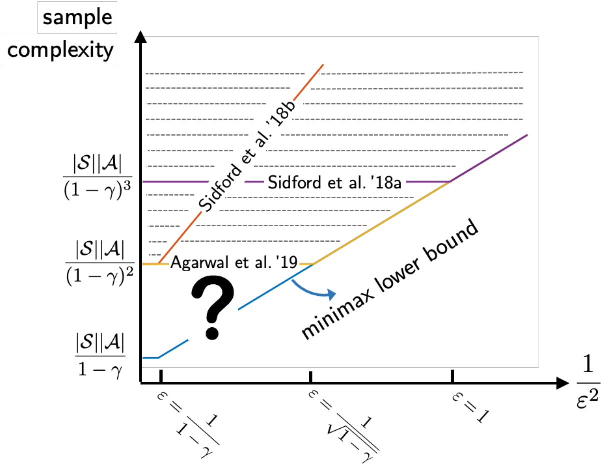


A benchmark of the prior art



All prior theory requires sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

A benchmark of the prior art

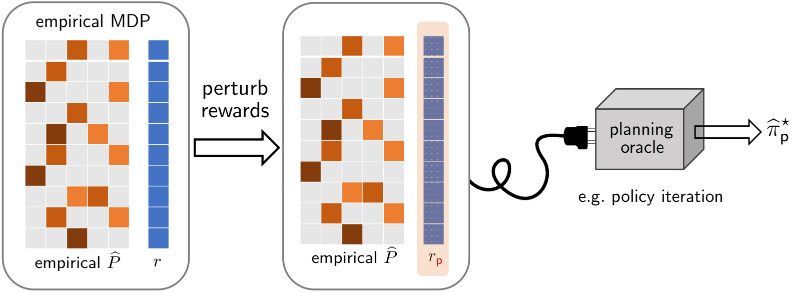


All prior theory requires sample size $\gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}$

Is it possible to close the gap?

Our method: a perturbed plug-in estimator

— *Li, Wei, Chi, Gu, Chen, 2020*



Run planning algorithms based on the *empirical* MDP with *slightly perturbed rewards*

$$r_{\mathsf{p}}(s,a) = r(s,a) + \zeta(s,a), \qquad \zeta(s,a) \sim \mathsf{Unif}(0,\xi).$$

Sample complexity of a perturbed plug-in estimator

Theorem (Li, Wei, Chi, Gu, Chen, 2020)

For any $0 < \epsilon \le \frac{1}{1-\gamma}$, *the optimal policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of the perturbed empirical MDP with* $\xi \asymp \frac{(1-\gamma)\epsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$ *achieves*

$$V^{\star} - V^{\widehat{\pi}_{\mathrm{p}}^{\star}} \le \epsilon$$

with sample complexity at most

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

Sample complexity of a perturbed plug-in estimator

Theorem (Li, Wei, Chi, Gu, Chen, 2020)

For any $0 < \epsilon \leq \frac{1}{1-\gamma}$*, the optimal policy* $\widehat{\pi}_{\mathrm{p}}^{\star}$ *of the perturbed empirical MDP with* $\xi \asymp \frac{(1-\gamma)\epsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$ *achieves*

$$V^{\star} - V^{\widehat{\pi}_{\mathrm{p}}^{\star}} \leq \epsilon$$

with sample complexity at most

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

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

Sample complexity of a perturbed plug-in estimator

Theorem (Li, Wei, Chi, Gu, Chen, 2020)

For any $0 < \epsilon \leq \frac{1}{1-\gamma}$, *the optimal policy* $\widehat{\pi}^\star_{\mathrm{p}}$ *of the perturbed empirical MDP with* $\xi \asymp \frac{(1-\gamma)\epsilon}{|\mathcal{S}|^5|\mathcal{A}|^5}$ *achieves*

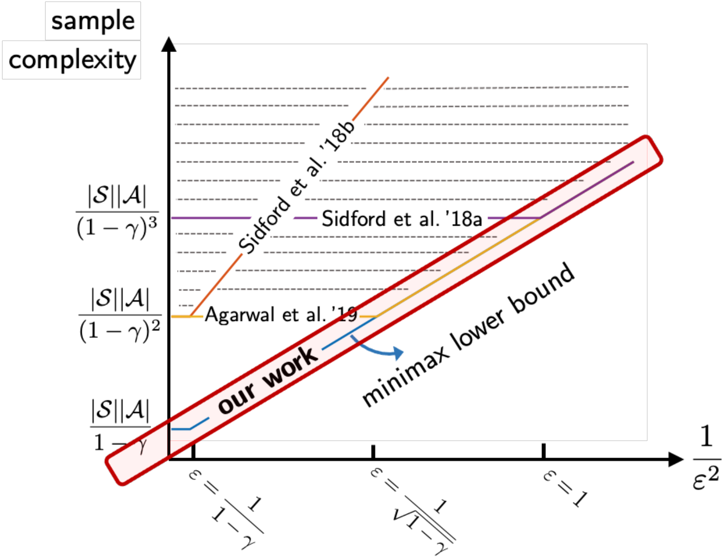
$$V^\star - V^{\widehat{\pi}^\star_{\mathrm{p}}} \leq \epsilon$$

with sample complexity at most

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

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

Close the gap



A glimpse of the analysis: notation

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

A glimpse of the analysis: notation

- 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: $\widehat{V}^{\pi} = (I - \widehat{P}_{\pi})^{-1} r$

A glimpse of the analysis: notation

- 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

A glimpse of the analysis: notation

- 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:

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

Proof ideas

Elementary decomposition:

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

Proof ideas

Elementary decomposition:

$$V^\star - V^{\widehat{\pi}^\star} = \big(V^\star - \widehat{V}^{\pi^\star}\big) + \big(\widehat{V}^{\pi^\star} - \widehat{V}^{\widehat{\pi}^\star}\big) + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big)$$
$$\leq \big(V^{\pi^\star} - \widehat{V}^{\pi^\star}\big) + 0 + \big(\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}\big)$$

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

Proof ideas

Elementary decomposition:

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

- **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**)

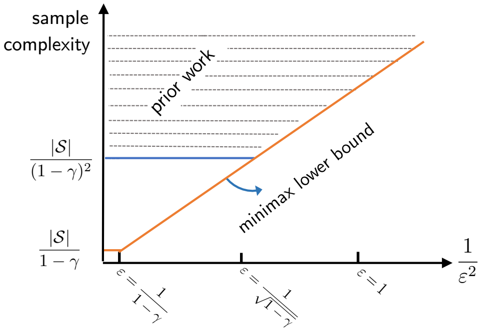
Step 1: improved theory for policy evaluation

Model-based policy evaluation:

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

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 (Li, Wei, Chi, Gu, Chen, 2020)

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

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

with sample complexity at most

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

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 (Li, Wei, Chi, Gu, Chen, 2020)

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

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

with sample complexity at most

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

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

Key idea 1: a peeling argument

First-order expansion:

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

Higher-order expansion $\longrightarrow$ tighter control:

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

Key idea 1: a peeling argument

First-order expansion:

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

Higher-order expansion $\longrightarrow$ tighter control:

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

Key idea 1: a peeling argument

First-order expansion:

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

Higher-order expansion $\longrightarrow$ tighter control:

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

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

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

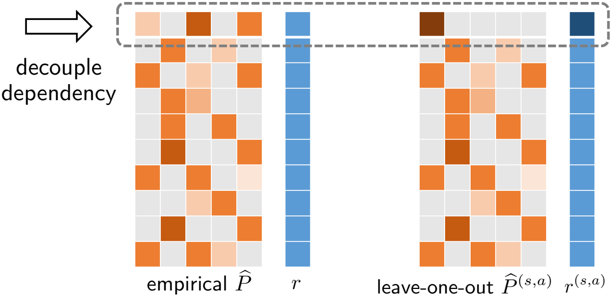
Step 2: controlling $\widehat{V}^{\widehat{\pi}^\star} - V^{\widehat{\pi}^\star}$

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

- highly suboptimal! (there are exponentially many policies)

Key idea 2: leave-one-out analysis

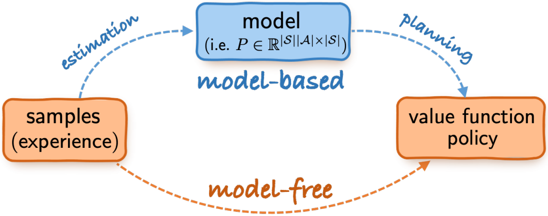
Decouple dependency by introducing auxiliary state-action absorbing MDPs by dropping randomness for each (s, a)



— *inspired by (Agarwal et al. 2019) but quite different ...*

Other leave-one-out analysis: (El Karoui, 2015; Javanmard, Montanari, 2015; Abbe et al., 2017; Zhong, Boumal, 2017; Ma et al., 2017; Pananjady, Wainwright, 2019)

Is model-free RL minimax optimal?



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

Q-learning: a classical model-free algorithm

Chris Watkins

Peper Dayan

$\underbrace{\text{Stochastic approximation}}_{\text{Robbins \& Monro, 1951}}$ for solving the **Bellman equation**

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

where

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

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

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

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

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

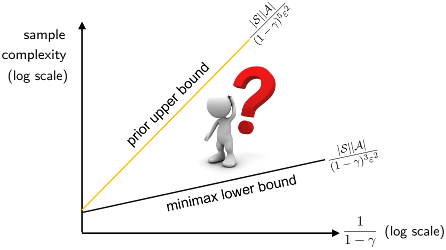
Prior art: achievability

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

Prior art: achievability

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

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	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\epsilon^2}$

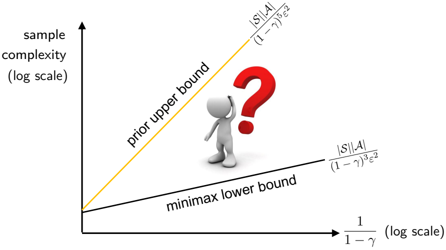


All prior results require sample size of at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\epsilon^2}$!

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

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

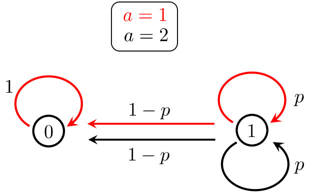
- Improves dependency on effective horizon $\frac{1}{1-\gamma}$
- Allows both constant and rescaled linear learning rate:

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

A curious numerical example

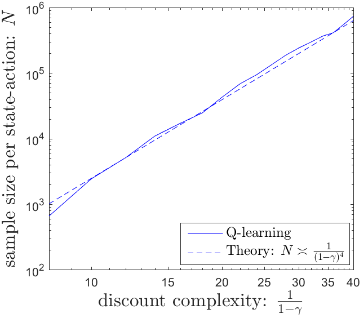
Numerical evidence: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}$ samples seem necessary . . .

— *observed in Wainwright '19*



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

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



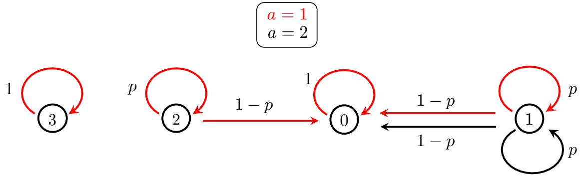
Q-learning is not minimax optimal

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

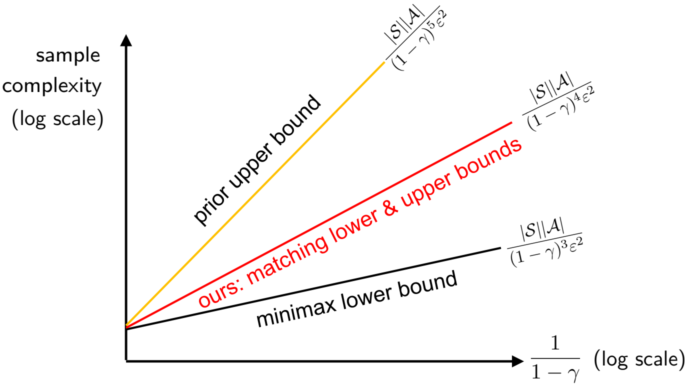
For any $0 < \epsilon \le 1$, there exists an MDP such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \le \epsilon$, Q-learning needs at least a sample complexity of

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

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

Why is Q-learning sub-optimal?

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

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

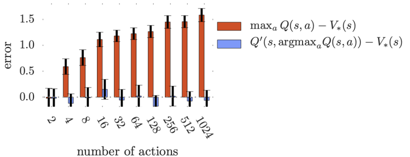


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically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

Why is Q-learning sub-optimal?

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

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

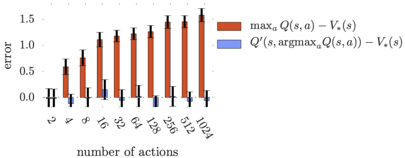


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically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

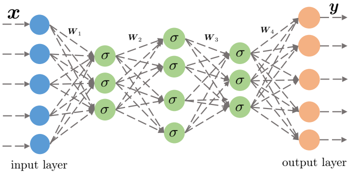
A provable fix: Q-learning with variance reduction (Wainwright 2019) is *provably* minimax optimal.

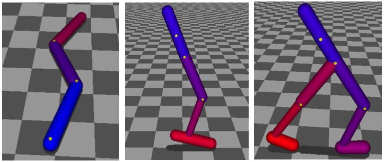
Part III: policy optimization

Policy optimization

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

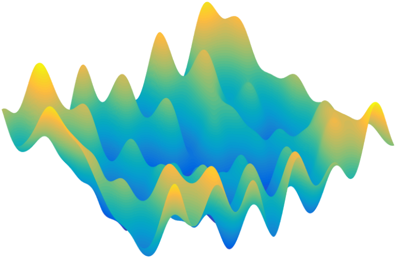
- directly optimize the policy, which is the quantity of interest;
- allow flexible differentiable parameterizations of the policy;
- work with both continuous and discrete problems.





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 more.



Our goal:

- understand finite-time convergence rates of popular heuristics;
- design fast-convergent algorithms that scale for finding policies with desirable properties.

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:

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

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

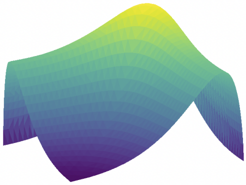
Policy gradient method (Sutton et al., 2000)

For $t = 0, 1, \cdots$

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

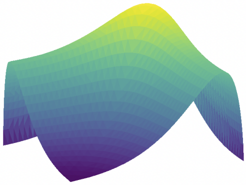
where η *is the learning rate.*

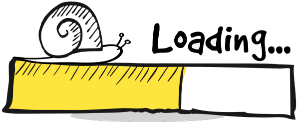
Global convergence of the PG method?



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

Global convergence of the PG method?

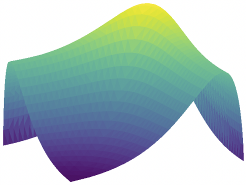


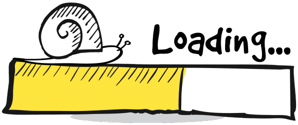


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

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

Global convergence of the PG method?

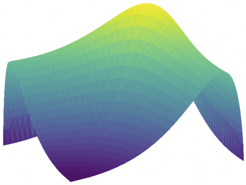


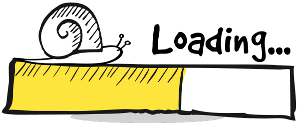


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

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

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)

There exists an MDP s.t. it takes softmax PG at least

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

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

A negative message

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

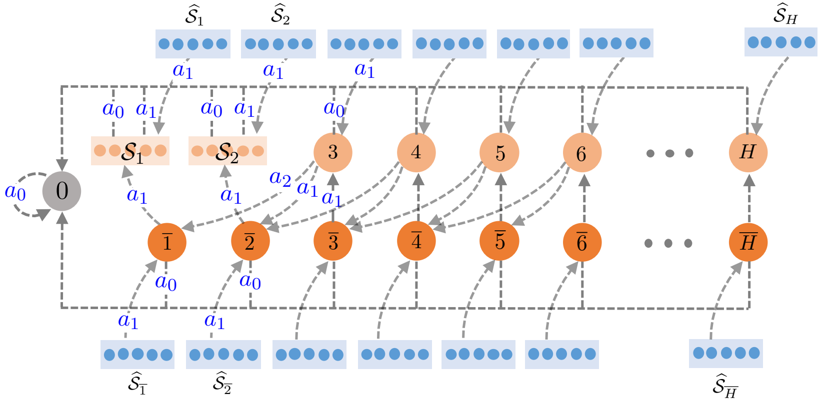
There exists an MDP s.t. it takes softmax PG at least

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

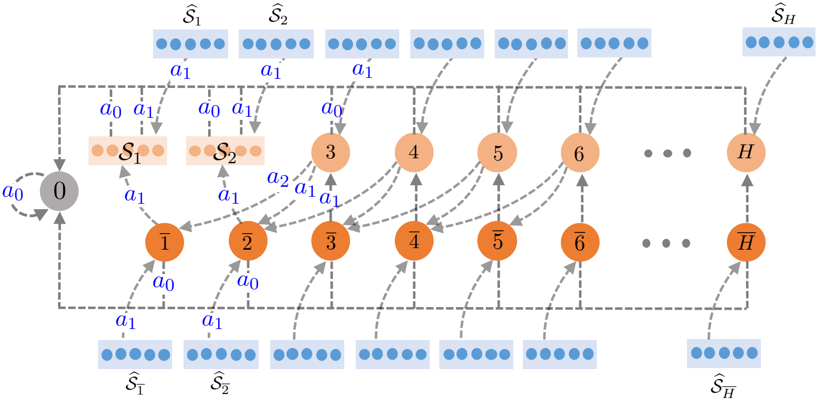
to achieve $\|V^{(t)} - V^\star\|_\infty \leq 0.15$.

- 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 $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\big[V^{(t)}(s) - V^\star(s)\big]$.

MDP construction for our lower bound

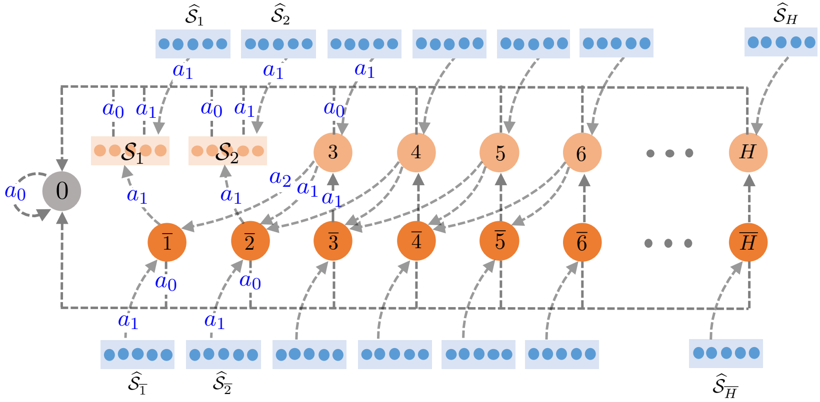


MDP construction for our lower bound



Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$,

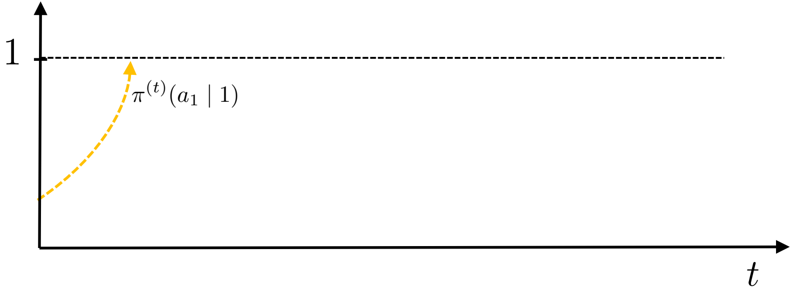
MDP construction for our lower bound



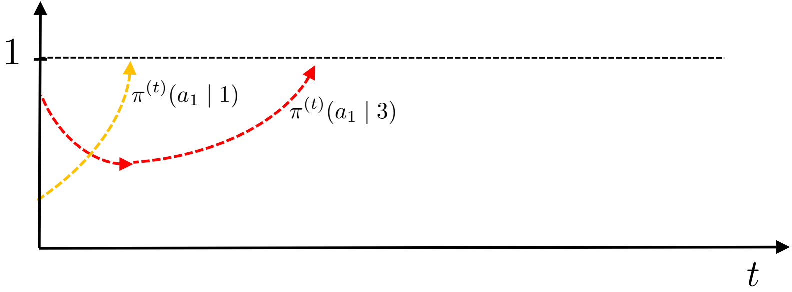
Key ingredients: for $3 \le s \le H \asymp \frac{1}{1-\gamma}$,

- $\pi^{(t)}(a_{\text{opt}} \mid s)$ keeps decreasing until $\pi^{(t)}(a_{\text{opt}} \mid s-2) \approx 1$

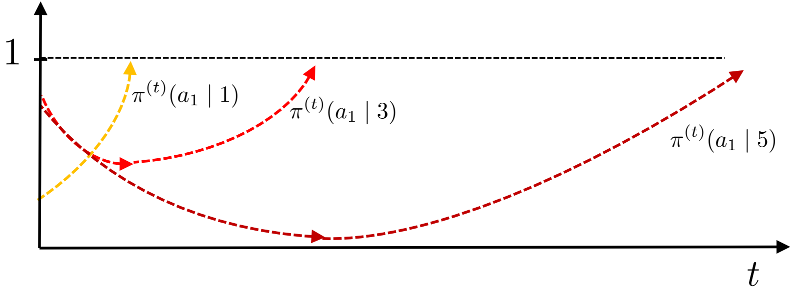
What is happening in our constructed MDP?



What is happening in our constructed MDP?

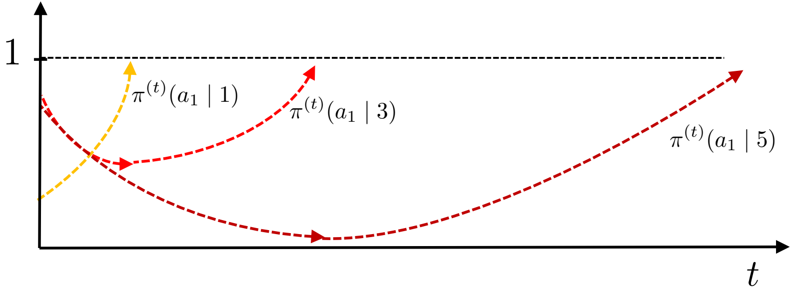


What is happening in our constructed MDP?



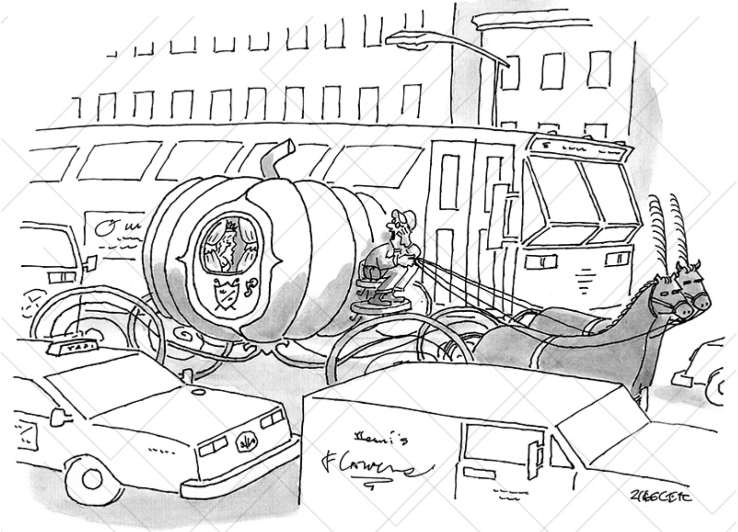
Convergence time for state s grows geometrically as s increases

What is happening in our constructed MDP?



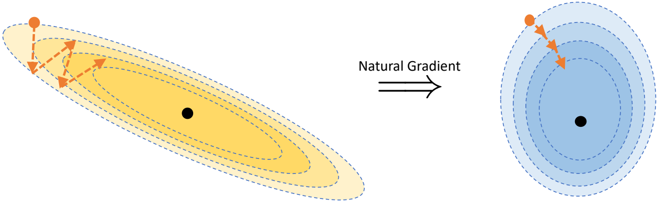
Convergence time for state s grows geometrically as s increases

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



"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)

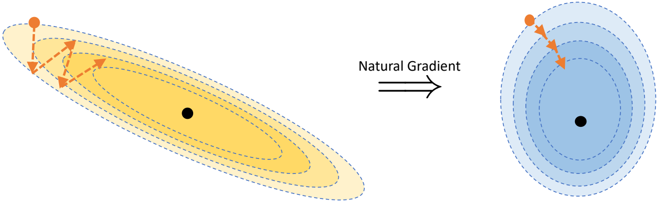
For $t = 0, 1, \cdots$

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

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:*

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

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

NPG in the tabular setting

Natural policy gradient (NPG) method (Tabular setting)

For $t = 0, 1, \cdots$, NPG updates the policy via

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

where $Q^{(t)} := Q^{\pi^{(t)}}$ is the Q-function of $\pi^{(t)}$, and $\eta > 0$.

- invariant with the choice of ρ
- Reduces to policy iteration (PI) when $\eta = \infty$.

Global convergence of NPG

Theorem (Agarwal et al., 2019)

Set $\pi^{(0)}$ *as a uniform policy. For all* $t \geq 0$*, we have*

$$V^{(t)}(\rho) \geq V^{\star}(\rho) - \left(\frac{\log|\mathcal{A}|}{\eta} + \frac{1}{(1-\gamma)^2}\right)\frac{1}{t}.$$

Global convergence of NPG

Theorem (Agarwal et al., 2019)

Set $\pi^{(0)}$ *as a uniform policy. For all* $t \geq 0$*, we have*

$$V^{(t)}(\rho) \geq V^{\star}(\rho) - \left(\frac{\log |\mathcal{A}|}{\eta} + \frac{1}{(1-\gamma)^2} \right) \frac{1}{t}.$$

Implication: set $\eta \geq (1-\gamma)^2 \log |\mathcal{A}|$, we find an ϵ-optimal policy within at most

$$\frac{2}{(1-\gamma)^2 \epsilon} \text{ iterations.}$$

Global convergence of NPG

Theorem (Agarwal et al., 2019)

Set $\pi^{(0)}$ *as a uniform policy. For all* $t \geq 0$*, we have*

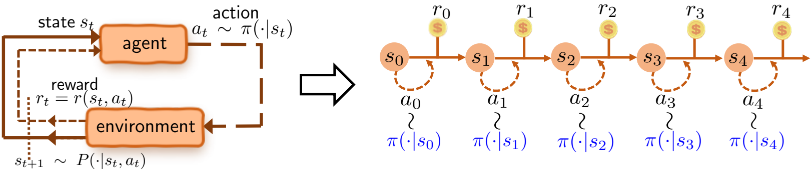
$$V^{(t)}(\rho) \geq V^{\star}(\rho) - \left(\frac{\log |\mathcal{A}|}{\eta} + \frac{1}{(1-\gamma)^2} \right) \frac{1}{t}.$$

Implication: set $\eta \geq (1-\gamma)^2 \log |\mathcal{A}|$, we find an ϵ-optimal policy within at most

$$\frac{2}{(1-\gamma)^2 \epsilon} \text{ iterations.}$$

Global convergence at a sublinear rate independent of $|\mathcal{S}|$, $|\mathcal{A}|$!

Booster #2: entropy regularization

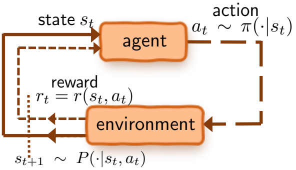


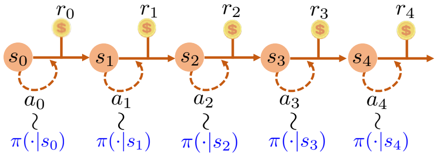
To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

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

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):

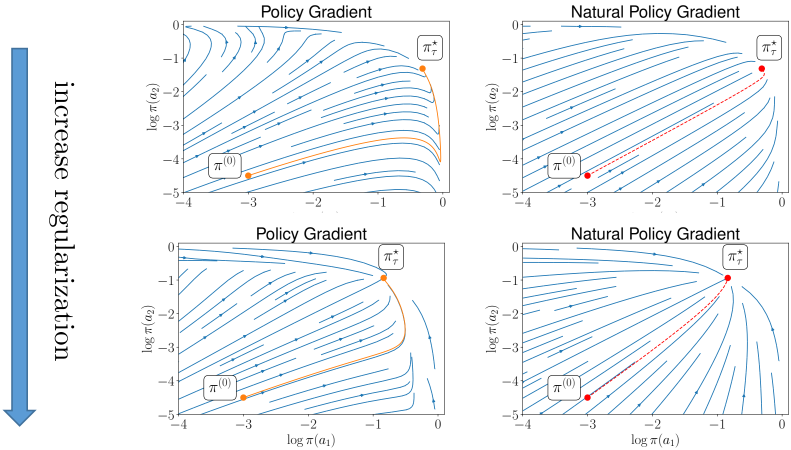
$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t))\Big) \,\Big|\, s_0 = s\right]$$

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

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

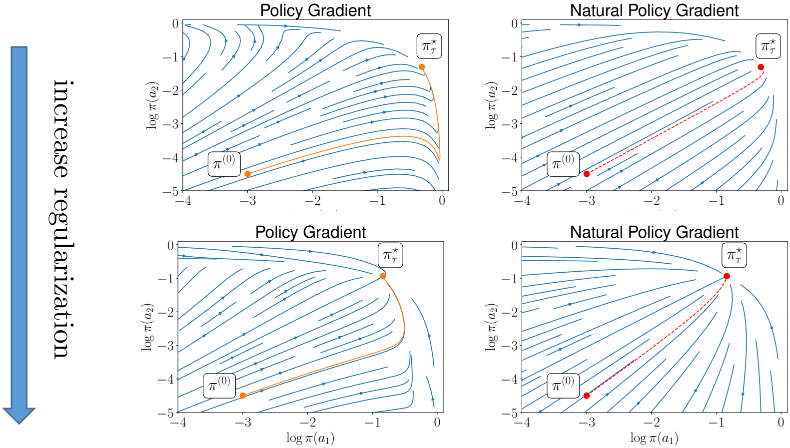
Entropy-regularized natural gradient helps!

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

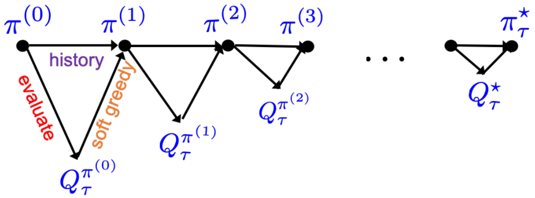


Entropy-regularized natural gradient helps!

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 (Tabular setting)

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

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

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

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$;

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

For any learning rate $0 < \eta \leq (1-\gamma)/\tau$, the entropy-regularized NPG updates satisfy

- **Linear convergence of soft Q-functions:**

$$\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq C_1 \gamma \left(1 - \eta\tau\right)^t$$

for all $t \geq 0$, where $Q_\tau^\star$ is the optimal soft Q-function, and

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

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le \epsilon$, the iteration complexity is at most

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

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

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

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

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le \epsilon$, the iteration complexity is at most

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

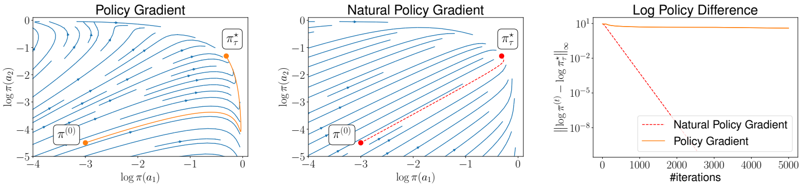
$$\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$$

- **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 $|\mathcal{S}|$, $|\mathcal{A}|$!

Comparisons with entropy-regularized PG



(Mei et al., 2020) showed entropy-regularized PG achieves

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le \left(V_\tau^\star(\rho) - V_\tau^{(0)}(\rho)\right)$$

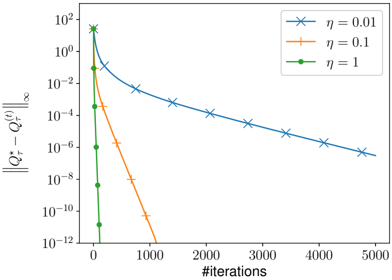
$$\cdot \exp\left(-\frac{(1-\gamma)^4 t}{(8/\tau + 4 + 8\log|\mathcal{A}|)|\mathcal{S}|}\left\|\frac{d_\rho^{\pi_\tau^\star}}{\rho}\right\|_\infty^{-1} \min_s \rho(s) \underbrace{\left(\inf_{0\le k\le t-1}\min_{s,a}\pi^{(k)}(a|s)\right)^2}_{\text{can be exponential in } |\mathcal{S}| \text{ and } \frac{1}{1-\gamma}}\right)$$

Much faster convergence of entropy-regularized NPG
at a **dimension-free** rate!

Comparison with unregularized NPG

Regularized NPG

$\tau = 0.001$

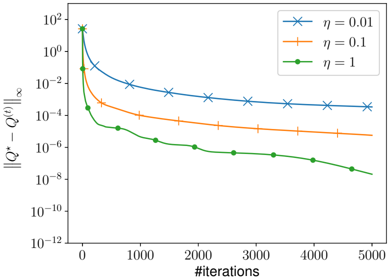


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Vanilla NPG

$\tau = 0$



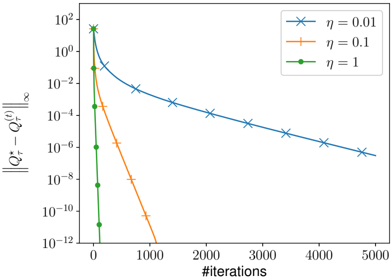
Sublinear rate: $\frac{1}{\min\{\eta,(1-\gamma)^2\}\epsilon}$

(Agarwal et al. 2019)

Comparison with unregularized NPG

Regularized NPG

$\tau = 0.001$

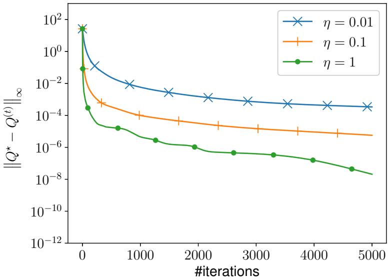


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Vanilla NPG

$\tau = 0$



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

(Agarwal et al. 2019)

Entropy regularization enables fast convergence!

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

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

e.g., using sample-based estimators (Williams, 1992).

Entropy-regularized NPG with inexact gradients

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

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

e.g., using sample-based estimators (Williams, 1992).

Inexact entropy-regularized NPG:

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

Entropy-regularized NPG with inexact gradients

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

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

e.g., using sample-based estimators (Williams, 1992).

Inexact entropy-regularized NPG:

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

Question: Robustness of entropy-regularized NPG?

Linear convergence with inexact gradients

Theorem (Cen, Cheng, Chen, Wei, Chi '20; improved)

For any learning rate $0 < \eta \leq (1-\gamma)/\tau$*, the entropy-regularized NPG updates achieve the same iteration complexity as the exact case, as long as*

$$\delta \leq \frac{1-\gamma}{\gamma} \cdot \min\left\{\frac{\epsilon}{4}, \sqrt{\frac{\epsilon\tau}{2}}\right\}$$

Linear convergence with inexact gradients

Theorem (Cen, Cheng, Chen, Wei, Chi '20; improved)

For any learning rate $0 < \eta \le (1-\gamma)/\tau$, *the entropy-regularized NPG updates achieve the same iteration complexity as the exact case, as long as*

$$\delta \le \frac{1-\gamma}{\gamma} \cdot \min\left\{\frac{\epsilon}{4}, \sqrt{\frac{\epsilon\tau}{2}}\right\}$$

- **Intuition:** assume $\tau = O(\epsilon)$, the per-iteration policy evaluation error is no larger than

$$\frac{\text{final error}}{\text{iteration complexity}} = \frac{\epsilon}{\widetilde{O}((1-\gamma)^{-1})} \approx (1-\gamma)\epsilon.$$

Aside: statistical implication

Question: how many samples are sufficient to find an ϵ-optimal policy of the unregularized MDP?

Aside: statistical implication

Question: how many samples are sufficient to find an ϵ-optimal policy of the unregularized MDP?

Recipe:

- set $\tau = \frac{(1-\gamma)\epsilon}{\log|\mathcal{A}|}$;
- use fresh samples for policy evaluation with a targeted accuracy $\delta \asymp \frac{(1-\gamma)^{1.5}\epsilon}{\gamma\sqrt{\log|\mathcal{A}|}}$, e.g. using model-based plug-in estimators (Li et al., 2020).

Aside: statistical implication

Question: how many samples are sufficient to find an ϵ-optimal policy of the unregularized MDP?

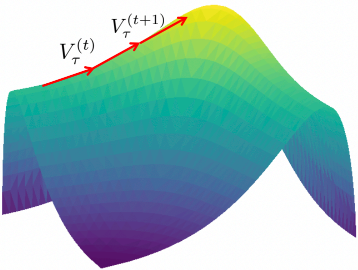
Recipe:

- set $\tau = \frac{(1-\gamma)\epsilon}{\log|\mathcal{A}|}$;
- use fresh samples for policy evaluation with a targeted accuracy $\delta \asymp \frac{(1-\gamma)^{1.5}\epsilon}{\gamma\sqrt{\log|\mathcal{A}|}}$, e.g. using model-based plug-in estimators (Li et al., 2020).

A crude answer:

$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^7\epsilon^2}\right) \text{ samples}$$

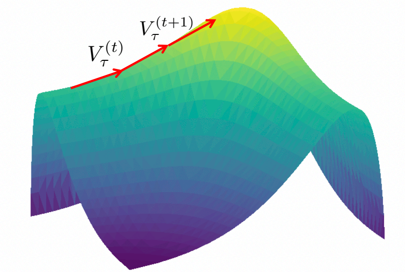
A key lemma: monotonic performance improvement



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

discounted state visitation distribution

A key lemma: monotonic performance improvement



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

discounted state visitation distribution

Implication: monotonic improvement of $V_\tau(s)$ and $Q_\tau(s,a)$.

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

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

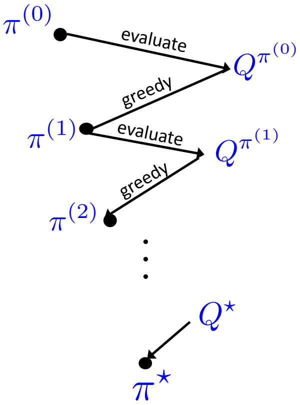
γ**-contraction of soft Bellman operator:**

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

Richard Bellman

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

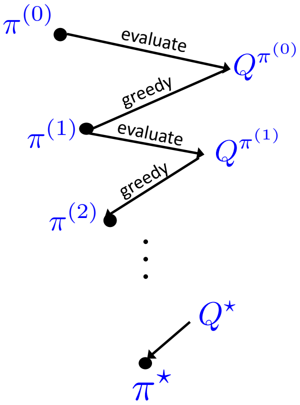
Policy iteration



Bellman operator

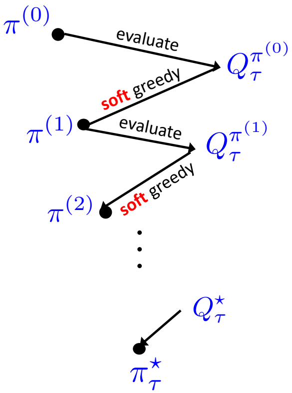
Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

Policy iteration



Bellman operator

Soft policy iteration



Soft Bellman operator

A key linear system: general learning rates

Let $x_t := \begin{bmatrix} \big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty \\ \big\|Q_\tau^\star - \tau \log \xi^{(t)}\big\|_\infty \end{bmatrix}$ and $y := \begin{bmatrix} \big\|Q_\tau^{(0)} - \tau \log \xi^{(0)}\big\|_\infty \\ 0 \end{bmatrix}$,

where $\xi^{(t)} \propto \pi^{(t)}$ is an auxiliary sequence, then

A key linear system: general learning rates

Let $x_t := \begin{bmatrix} \big\|Q_\tau^\star - Q_\tau^{(t)}\big\|_\infty \\ \big\|Q_\tau^\star - \tau \log \xi^{(t)}\big\|_\infty \end{bmatrix}$ and $y := \begin{bmatrix} \big\|Q_\tau^{(0)} - \tau \log \xi^{(0)}\big\|_\infty \\ 0 \end{bmatrix}$,

where $\xi^{(t)} \propto \pi^{(t)}$ is an auxiliary sequence, then

$$x_{t+1} \leq A x_t + \gamma \left(1 - \frac{\eta\tau}{1-\gamma}\right)^{t+1} y,$$

where

$$A := \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{\eta\tau}{1-\gamma} & 1 - \frac{\eta\tau}{1-\gamma} \end{bmatrix}$$

is a rank-1 matrix with a non-zero eigenvalue $\underbrace{1 - \eta\tau}_{\text{contraction rate!}}$.

Beyond entropy regularization

Leverage regularization to promote structural properties of the learned policy.

cost-sensitive RL

weighted 1-norm

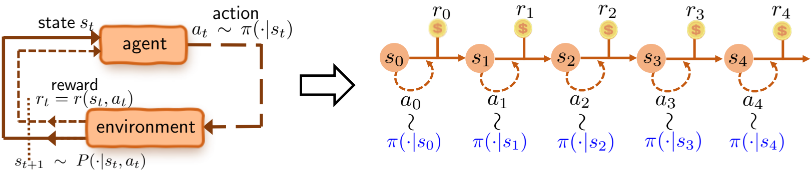
sparse exploration

Tsallis entropy

constrained and safe RL

log-barrier

Regularized RL in general form

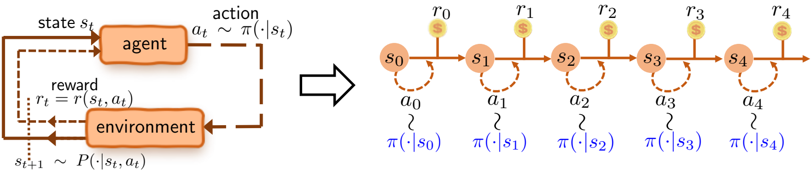


The regularized value function is defined as

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

where h_s is convex (and possibly nonsmooth) w.r.t. $\pi(\cdot|s)$.

Regularized RL in general form



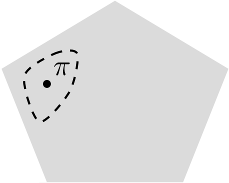
The regularized value function is defined as

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

where h_s is convex (and possibly nonsmooth) w.r.t. $\pi(\cdot|s)$.

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

Detour: a mirror descent view of entropy-regularized NPG



Entropy-reg. NPG = mirror descent with KL divergence:

(Lan, 2021; Shani et al., 2020)

$$\pi^{(t+1)}(\cdot|s) = \operatorname*{argmin}_{p\in\Delta(\mathcal{A})} \big\langle -Q_\tau^{(t)}(s,\cdot), p\big\rangle - \tau\mathcal{H}(p) + \frac{1}{\eta}\mathsf{KL}\big(p\|\pi^{(t)}(\cdot|s)\big)$$

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

for all $s \in \mathcal{S}$.

Definition (Generalized Bregman divergence, Kiwiel 1997)

The generalized Bregman divergence w.r.t. to a convex $h : \Delta(\mathcal{A}) \mapsto \mathbb{R}$ is defined as:

$$\begin{aligned} D_h(p, q; g) &= h(p) - h(q) - \langle g, p - q \rangle \\ &= h(p) - h(q) - \langle g - c \cdot \mathbf{1}, p - q \rangle, \end{aligned}$$

for $p, q \in \Delta(\mathcal{A})$, where $g \in \partial h(q)$ and $c \in \mathbb{R}$.

Generalized policy mirror descent (GPMD)

Definition (Generalized Bregman divergence, Kiwiel 1997)

The generalized Bregman divergence w.r.t. to a convex $h : \Delta(\mathcal{A}) \mapsto \mathbb{R}$ is defined as:

$$\begin{aligned} D_h(p, q; g) &= h(p) - h(q) - \langle g, p - q \rangle \\ &= h(p) - h(q) - \langle g - c \cdot \mathbf{1}, p - q \rangle, \end{aligned}$$

for $p, q \in \Delta(\mathcal{A})$, where $g \in \partial h(q)$ and $c \in \mathbb{R}$.

A natural idea

For $t = 0, 1, \cdots,$

$$\pi^{(t+1)}(\cdot|s) = \operatorname*{argmin}_{p \in \Delta(\mathcal{A})} \langle -Q_\tau(s, \cdot), p \rangle + \tau h_s(p) + \frac{1}{\eta} D_{h_s}(p, \pi^{(t)}(\cdot|s); \partial h_s(\pi^{(t)}(\cdot|s)))$$

PMD with Generalized Bregman Divergence (**GPMD**)

Plugging in a recursive surrogate $\{\xi^{(t)}\}$ of $\partial h_s(\pi^{(t)}(\cdot|s))$, we obtain the formal algorithm.

Generalized policy mirror descent (GPMD) method

For $t = 0, 1, \cdots$, *update*

$$\pi^{(t+1)}(\cdot|s) = \operatorname*{argmin}_{p \in \Delta(\mathcal{A})} \langle -Q_\tau(s,\cdot), p \rangle + \tau h_s(p) + \frac{1}{\eta} D_{h_s}(p, \pi^{(t)}(\cdot|s); \xi^{(t)}(s,\cdot)),$$

and

$$\xi^{(t+1)}(s,\cdot) = \frac{1}{1+\eta\tau}\xi^{(t)}(s,\cdot) + \frac{\eta}{1+\eta\tau} Q_\tau^{(t)}(s,\cdot).$$

The subproblem does not admit closed-form solution in general.

Linear convergence with exact gradient

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$; exact solution to subproblems.

— *Read our paper for the inexact case!*

Linear convergence with exact gradient

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$; exact solution to subproblems.

— *Read our paper for the inexact case!*

Theorem (Zhan*, Cen*, Huang, Chen, Lee, Chi '21)

For any learning rate $\eta > 0$, the GPMD updates satisfy

- **Linear convergence of soft Q-functions:**

$$\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq C_1 \gamma \left(1 - \frac{\eta\tau(1-\gamma)}{1+\eta\tau}\right)^t$$

where $C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + \frac{2}{1+\eta\tau}\|Q_\tau^\star - \tau\xi^{(0)}\|_\infty$.

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \epsilon$, the iteration complexity is at most

- **General learning rates ($\eta > 0$):**

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

- **Regularized policy iteration ($\eta = \infty$):**

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

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \epsilon$, the iteration complexity is at most

- **General learning rates ($\eta > 0$):**

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

- **Regularized policy iteration ($\eta = \infty$):**

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

Global linear convergence of GPMD at a **dimension-free** rate!

Comparison with PMD (Lan, 2021)

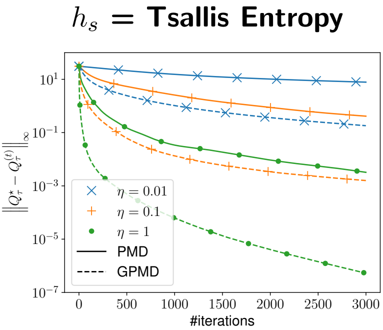
Policy mirror descent (PMD) method (Lan, 2021)

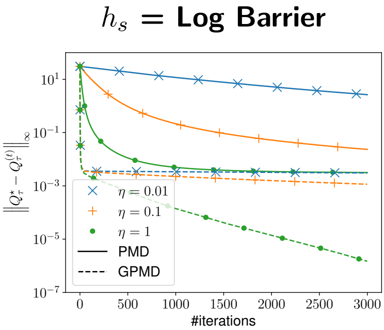
For $t = 0, 1, \cdots,$

$$\pi^{(t+1)}(\cdot|s) = \operatorname*{argmin}_{p \in \Delta(\mathcal{A})} \langle -Q_\tau(s,\cdot), p\rangle + \tau h_s(p) + \frac{1}{\eta}\mathsf{KL}(p||\pi^{(t)}(\cdot|s))$$

- Linear convergence is established only when h_s is stronger than entropy regularization ($h_s + \mathcal{H}$ is convex).
- In contrast, GPMD converges linearly for general convex and nonsmooth h_s!

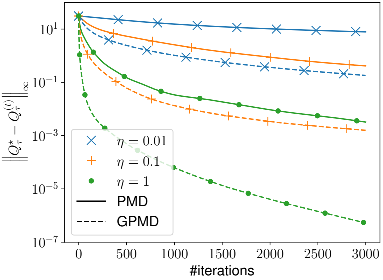
Numerical examples



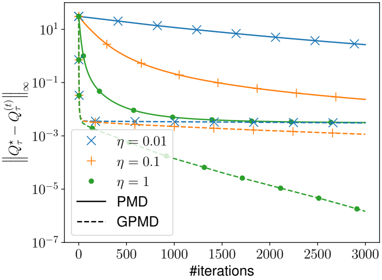


Numerical examples

h_s = **Tsallis Entropy**



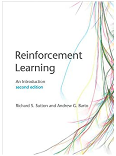
h_s = **Log Barrier**

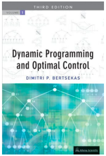


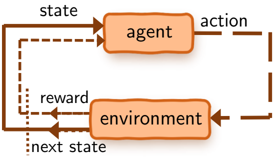
GPMD achieves faster convergence than PMD!

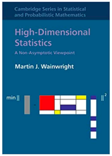
Part IV: concluding remarks and further pointers

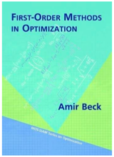
Concluding remarks









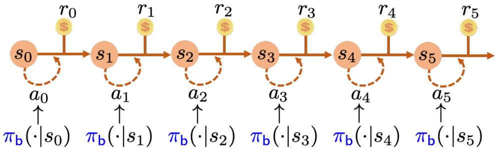


Understanding non-asymptotic performances of RL algorithms is a fruitful playground!

Future directions:

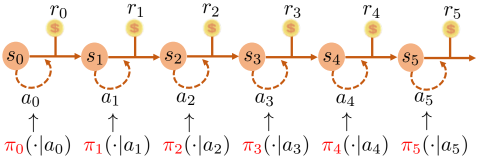
- function approximation
- multi-agent RL
- offline RL
- many more...

Sampling under a behavior policy: asynchronous/offline RL



(Bhandari et al, 2018; Srikant and Ying, 2019; Qu and Wierman, 2020; Li et al., 2020)

Exploration under an adaptive policy: minimize the regret against the optimal policy



(Azar et al., 2017; Jin et al., 2018; Li et al., 2021)

Beyond the tabular setting

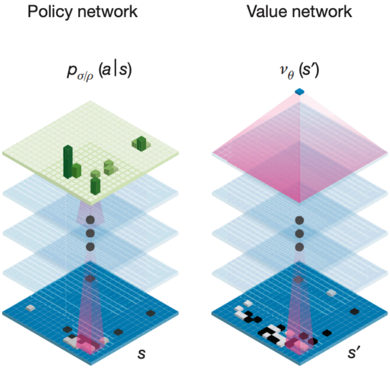
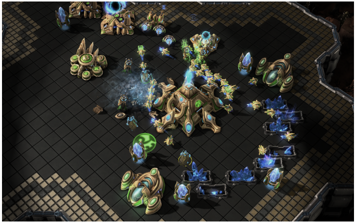


Figure credit: (Silver et al., 2016)

- function approximation for dimensionality reduction
- Provably efficient RL algorithms under minimal assumptions

(Osband and Van Roy, 2014; Dai et al., 2018; Du et al., 2019; Jin et al., 2020)

Multi-agent RL



- **Competitive setting:** finding Nash equilibria for Markov games
- **Collaborative setting:** multiple agents jointly optimize the policy to maximize the total reward

(Zhang, Yang, and Basar, 2021; Cen, Wei, and Chi, 2021)

Can we design RL algorithms based on history data?

(Rashidinejad, Zhu, Ma, Jiao, and Russell, 2021)

Bibliography I

Disclaimer: this straw-man list is by no means exhaustive (in fact, it is quite the opposite given the fast pace of the field), and biased towards materials most related to this tutorial; readers are invited to further delve into the references therein to gain a more complete picture.

Books and monographs:

- Sutton and Barto. *Reinforcement learning: An introduction, 2nd edition*. MIT press, 2018.
- Agarwal, Jiang, Kakade, and Sun. *Reinforcement learning: Theory and algorithms*, monograph, 2021+.
- Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific, 2019.
- Szepesvári. *Algorithms for reinforcement learning*. Synthesis lectures on artificial intelligence and machine learning, 2010.
- Bertsekas and Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

Bibliography II

Model-based RL:

- Singh and Yee. "*An upper bound on the loss from approximate optimal-value functions.*" Machine Learning, 1994.
- Azar, Munos, and Kappen. "*Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model.*" Machine Learning, 2013.
- Kearns, Mansour, and Ng. "*A sparse sampling algorithm for near-optimal planning in large Markov decision processes.*" Machine Learning, 2002.
- Agarwal, Kakade, and Yang, "*Model-based reinforcement learning with a generative model is minimax optimal.*" COLT 2020.
- Dann and Brunskill. "*Sample complexity of episodic fixed-horizon reinforcement learning.*" NeurIPS 2015.
- Li, Wei, Chi, Gu, and Chen. "*Breaking the sample size barrier in model-based reinforcement learning with a generative model.*" NeurIPS 2020.

Bibliography III

- Sidford, Wang, Wu, Yang, and Ye. "*Near-optimal time and sample complexities for solving Markov decision processes with a generative model.*" NeurIPS 2018.
- Sidford, Wang, Wu, and Ye. "*Variance reduced value iteration and faster algorithms for solving Markov decision processes.*" SODA 2018.
- Pananjady and Wainwright. "*Instance-Dependent ℓ_∞-Bounds for Policy Evaluation in Tabular Reinforcement Learning.*" IEEE Transactions on Information Theory, 2020.
- Osband and Van Roy. "*Model-based reinforcement learning and the Eluder dimension.*" NeurIPS 2014.
- Azar, Osband, and Munos. "*Minimax regret bounds for reinforcement learning.*" ICML 2017.

Value-based RL:

- Sutton. "*Learning to predict by the methods of temporal differences.*" Machine Learning, 1988.
- Watkins and Dayan. "*Q-learning.*" Machine Learning, 1992.

Bibliography IV

- Tsitsiklis. "*Asynchronous stochastic approximation and Q-learning.*" Machine Learning, 1994.
- Borkar and Meyn. "*The ODE method for convergence of stochastic approximation and reinforcement learning.*" SIAM Journal on Control and Optimization, 2000.
- Tsitsiklis and Van Roy. "*An analysis of temporal-difference learning with function approximation.*" IEEE Transactions on Automatic Control, 1997.
- Kearns and Singh. "*Finite-sample convergence rates for Q-learning and indirect algorithms.*" NeurIPS 1999.
- Jaakkola, Jordan, and Singh. "*On the convergence of stochastic iterative dynamic programming algorithms.*" Neural Computation, 1994.
- Singh, Jaakkola, Littman, and Szepesvári. "*Convergence results for single-step on-policy reinforcement-learning algorithms.*" Machine Learning, 2000.
- Even-Dar, Mansour, and Bartlett. "*Learning rates for Q-learning.*" Journal of Machine Learning Research, 2003.

Bibliography V

- Beck and Srikant. "*Error bounds for constant step-size Q-learning.*" Systems & Control Letters, 2012.
- Bhandari, Russo, and Singal. "*A finite time analysis of temporal difference learning with linear function approximation.*" COLT 2018.
- Jin, Allen-Zhu, Bubeck, and Jordan. "*Is Q-learning provably efficient?*" NeurIPS 2018.
- Srikant and Ying. "*Finite-time error bounds for linear stochastic approximation and TD learning.*" COLT 2019.
- Wainwright. "*Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning.*" arXiv preprint arXiv:1905.06265, 2019.
- Wainwright. "*Variance-reduced Q-learning is minimax optimal.*" arXiv preprint arXiv:1906.04697, 2019.
- Zou, Xu, and Liang. "*Finite-sample analysis for SARSA with linear function approximation.*" NeurIPS 2019.
- Li, Cai, Chen, Gu, Wei, and Chi. "*Is Q-learning minimax optimal? a tight sample complexity analysis.*" arXiv preprint arXiv:2102.06548, 2021.

- Qu and Wierman. "*Finite-time analysis of asynchronous stochastic approximation and Q-learning.*" COLT 2020.
- Li, Wei, Chi, Gu, and Chen. "*Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction.*" NeurIPS 2020.
- Chen, Maguluri, Shakkottai, and Shanmugam. "*Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes.*" NeurIPS 2020.
- Li, Shi, Chen, Gu, and Chi, "*Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning.*" Preprint, 2021+.

Policy optimization:

- Williams. "*Simple statistical gradient-following algorithms for connectionist reinforcement learning.*" Machine Learning, 1992.
- Sutton, McAllester, Singh, and Mansour. "*Policy gradient methods for reinforcement learning with function approximation.*" NeurIPS 1999.
- Kakade. "*A natural policy gradient.*" NeurIPS 2001.

Bibliography VII

- Fazel, Ge, Kakade, and Mesbahi. "*Global convergence of policy gradient methods for the linear quadratic regulator.*" ICML 2018.
- Agarwal, Kakade, Lee, and Mahajan. "*On the theory of policy gradient methods: Optimality, approximation, and distribution shift.*" Journal of Machine Learning Research, 2021.
- Mei, Xiao, Szepesvári, and Schuurmans. "*On the global convergence rates of softmax policy gradient methods.*" ICML 2020.
- Bhandari and Russo. "*Global optimality guarantees for policy gradient methods.*" arXiv preprint arXiv:1906.01786, 2019.
- Cai, Yang, Jin, and Wang. "*Provably efficient exploration in policy optimization.*" ICML 2020.
- Shani, Efroni, and Mannor."*Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs.*" AAAI 2020.
- Li, Gen, Wei, Chi, Gu, and Chen. "*Softmax policy gradient methods can take exponential time to converge.*" arXiv preprint arXiv:2102.11270, 2021.

Bibliography VIII

- Cen, Cheng, Chen, Wei, and Chi. "*Fast global convergence of natural policy gradient methods with entropy regularization.*" Operations Research, 2021+.
- Zhan, Cen, Huang, Chen, Lee, and Chi."*Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence.*" arXiv preprint arXiv:2105.11066, 2021.
- Lan. "*Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes.*" arXiv preprint arXiv:2102.00135, 2021.
- Liu, Zhang, Basar, and Yin. "*An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods.*" NeurIPS 2020.
- Zhang, Koppel, Bedi, Szepesvári, and Wang. "*Variational policy gradient method for reinforcement learning with general utilities.*" NeurIPS 2020.
- Cen, Wei, and Chi. "*Fast policy extragradient methods for competitive games with entropy regularization.*" arXiv preprint arXiv:2105.15186, 2021.

Additional ad-hoc pointers:

- Neu, Jonsson, and Gómez. "*A unified view of entropy-regularized Markov Decision Processes.*" arXiv preprint arXiv:1705.07798, 2017.
- Dai, Shaw, Li, Xiao, He, Liu, Chen, and Song. "*SBEED: Convergent reinforcement learning with nonlinear function approximation.*" ICML 2018.
- Geist, Scherrer, and Pietquin. "*A theory of regularized Markov Decision Processes.*" ICML 2019.
- Du, Kakade, Wang, and Yang. "*Is a good representation sufficient for sample efficient reinforcement learning?*" ICLR 2019.
- Jin, Yang, Wang, and Jordan. "*Provably efficient reinforcement learning with linear function approximation.*" COLT 2020.
- Zhang, Yang, and Basar. "*Multi-agent reinforcement learning: A selective overview of theories and algorithms.*" Handbook of Reinforcement Learning and Control, 2021.
- Rashidinejad, Zhu, Ma, Jiao, and Russell. "*Bridging offline reinforcement learning and imitation learning: A tale of pessimism.*" arXiv preprint arXiv:2103.12021, 2021.

Thanks!

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