Model-Free RL: Non-asymptotic Statistical and Computational Guarantees

Yuejie Chi

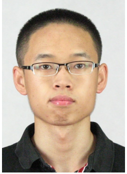
Carnegie Mellon University

2022 MIT LIDS Student Conference

My wonderful collaborators

Shicong Cen
CMU

Chen Cheng
Stanford

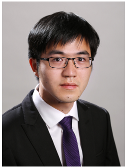


Gen Li
Princeton

Yuxin Chen
UPenn

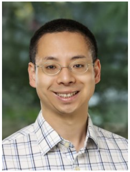
Yuting Wei
UPenn

Laixi Shi
CMU

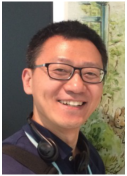


Changxiao Cai
UPenn

Wenhao Zhan
Princeton



Jason Lee
Princeton

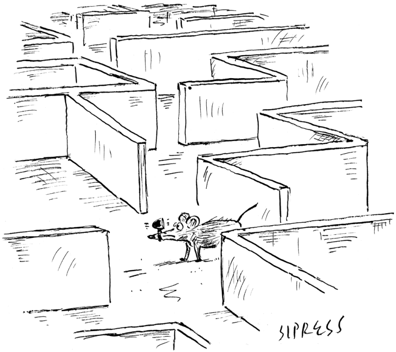


Yuantao Gu
Tsinghua

Reinforcement learning (RL)

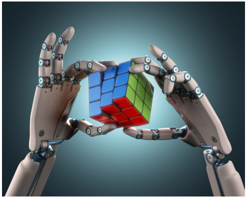
In RL, an agent learns by interacting with an environment.

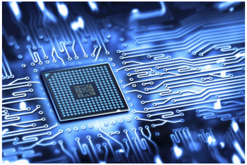
- unknown environments
- maximize total rewards
- trial-and-error
- sequential and online

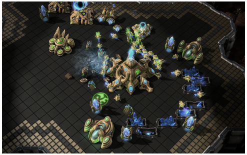


"Recalculating ... recalculating ..."

Recent successes in RL



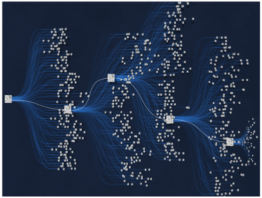




RL holds great promise in the next era of artificial intelligence.

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

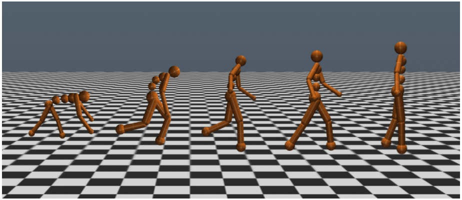
autonomous driving

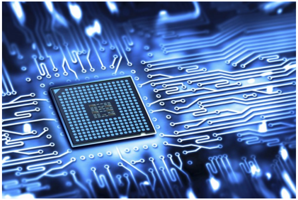
online ads

Calls for design of sample-efficient RL algorithms!

Computational efficiency

Running RL algorithms might take a long time and space

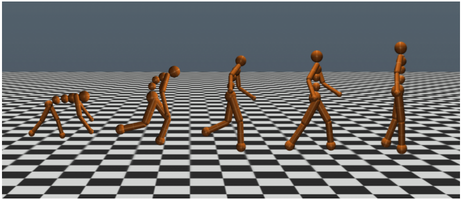


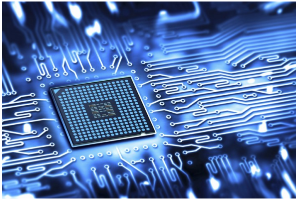


many CPUs / GPUs / TPUs + computing hours

Computational efficiency

Running RL algorithms might take a long time and space





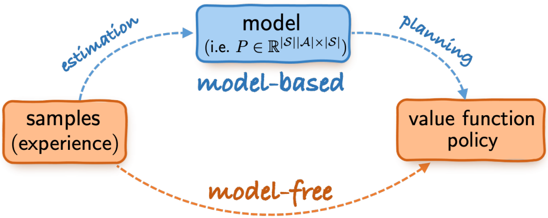
many CPUs / GPUs / TPUs + computing hours

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.

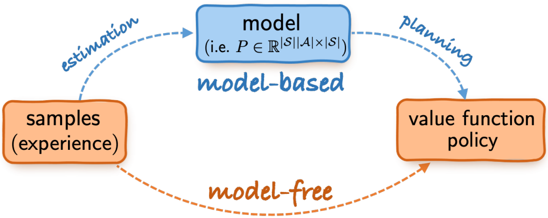
Two approaches to RL



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



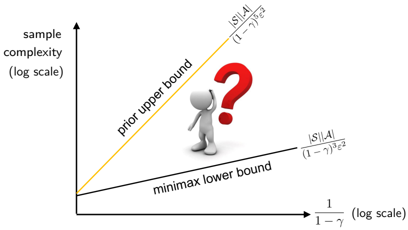
Model-based approach ("plug-in")

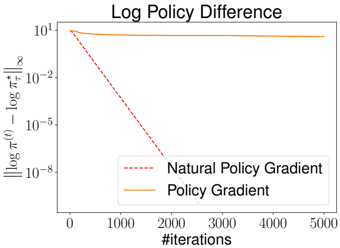
1. build an empirical estimate $\widehat{P}$ for P
2. planning based on empirical $\widehat{P}$

Model-free approach

1. learning w/o constructing model explicitly
2. widely popular and successful in practice

This talk: model-free approach



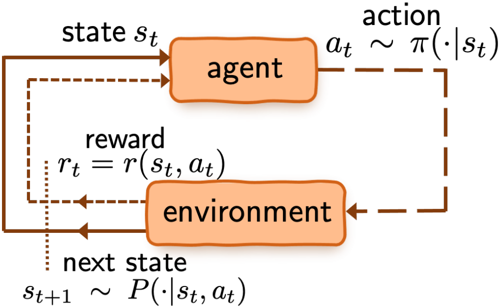


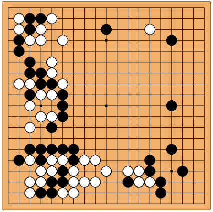
Value-based approach:
Finite-sample complexity of Q-learning

Policy-based approach:
Finite-time convergence of policy optimization

Backgrounds: Markov decision processes

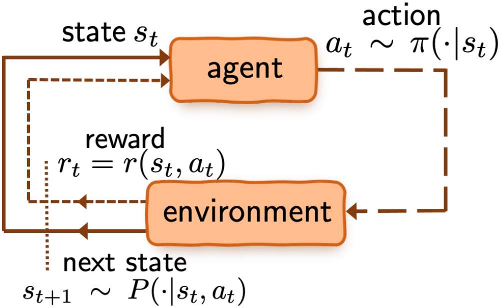
Markov decision process (MDP)

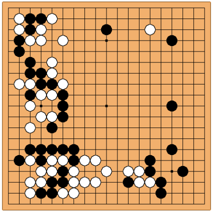




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

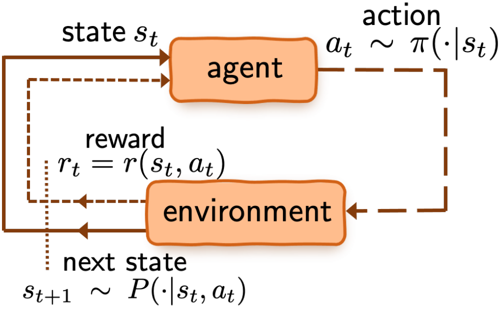
Markov decision process (MDP)

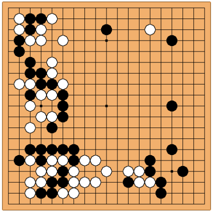




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

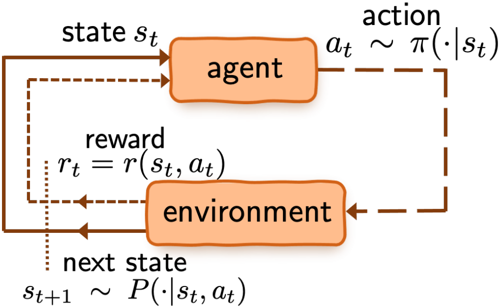
Markov decision process (MDP)

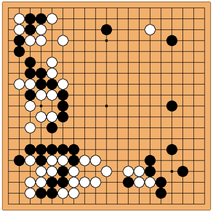




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

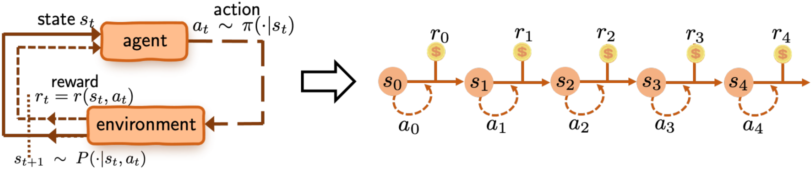
Markov decision process (MDP)





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

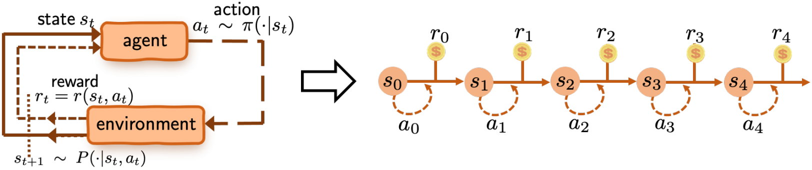
Value function



Value function of policy π:

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

Value function

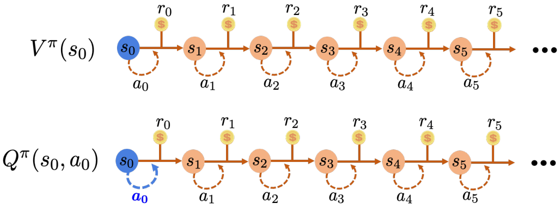


Value function of policy π:

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

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

Q-function

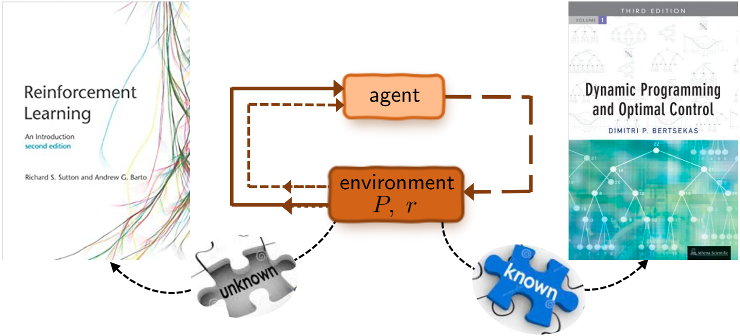


Q-function of policy π:

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

- $(\cancel{a_0,}\, s_1, a_1, s_2, a_2, \cdots)$: 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 $\pi^\star(s) = \text{argmax}_{a \in \mathcal{A}} Q^\star(s, a)$

Bellman's optimality principle

Bellman operator

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

- one-step look-ahead

Bellman's optimality principle

Bellman operator

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

- one-step look-ahead

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

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

γ**-contraction of Bellman operator:**

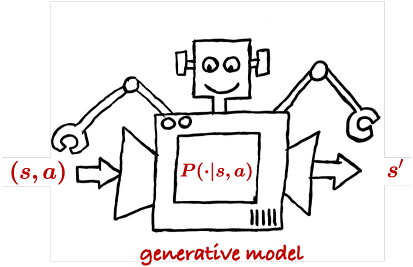
$$\|\mathcal{T}(Q_1) - \mathcal{T}(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$$

Richard Bellman

Is Q-learning minimax-optimal?

RL with a generative model / simulator

— *Kearns and Singh, 1999*

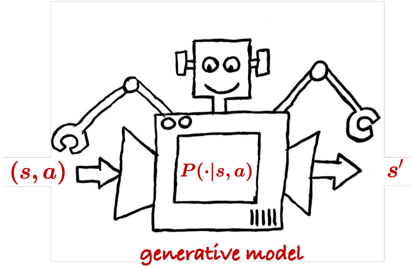


For each state-action pair (s, a), collect N samples

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

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 $\|\widehat{Q} - Q^\star\|_\infty \le \epsilon$, 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 $|\mathcal{S}|^2|\mathcal{A}|$

Q-learning: a classical model-free algorithm

Chris Watkins

Peter Dayan

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

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

where

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

Q-learning: a classical model-free algorithm


Chris Watkins


Peter Dayan

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

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

Q-learning: a classical model-free algorithm


Chris Watkins


Peter Dayan

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

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

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

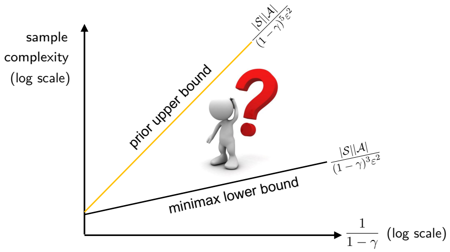
Prior art: achievability

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

Prior art: achievability

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

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

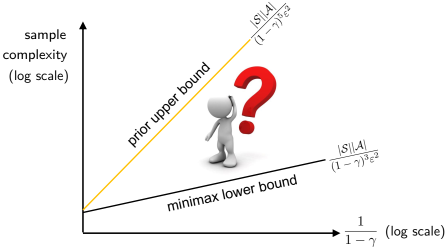


All prior results require sample size of at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\epsilon^2}$!

Prior art: achievability

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

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



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 $0 < \epsilon \le 1$, Q-learning yields

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

with sample complexity at most

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

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

A sharpened sample complexity of Q-learning

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

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

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

with sample complexity at most

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

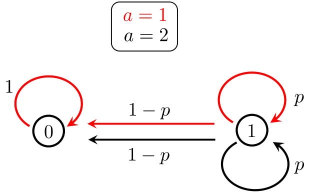
- Improves dependency on effective horizon $\frac{1}{1-\gamma}$
- Allows both constant and rescaled linear learning rate:

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

A curious numerical example

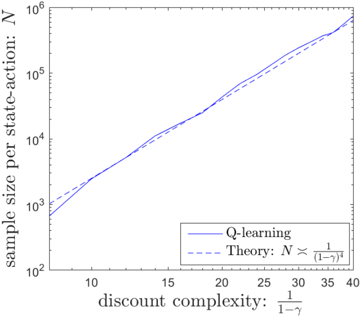
Numerical evidence: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}$ samples seem necessary ...

— *observed in Wainwright '19*



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

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



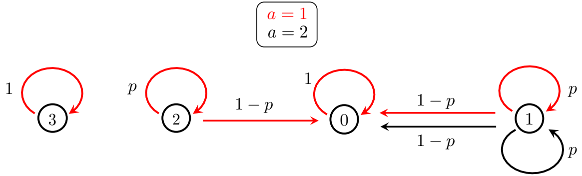
Q-learning is not minimax optimal

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

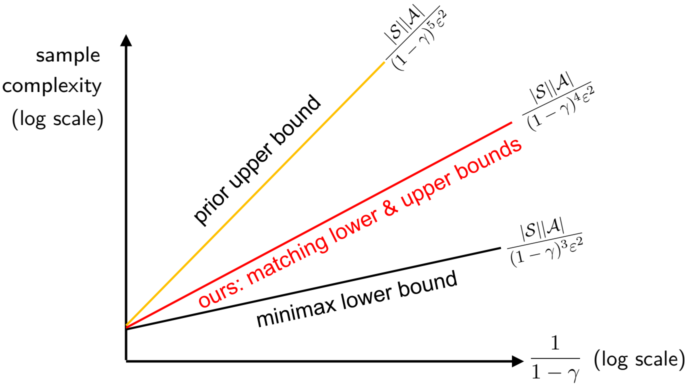
For any $0 < \epsilon \leq 1$, *there exists an MDP such that to achieve* $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \epsilon$, *Q-learning needs at least a sample complexity of*

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

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



Where we stand now



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

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).

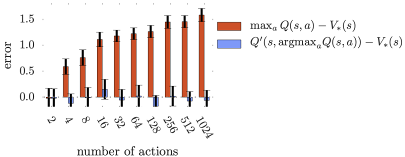


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically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

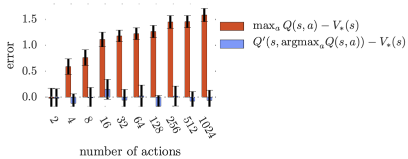


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically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

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

TD-learning: when the action space is a singleton

Richard Sutton

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

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

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

A sharpened sample complexity of TD-learning

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

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

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

with sample complexity at most

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

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

A sharpened sample complexity of TD-learning

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

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

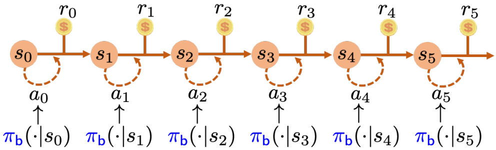
$$\|\widehat{V} - V^{\star}\|_{\infty} \leq \epsilon$$

with sample complexity at most

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

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

Sampling under a behavior policy: asynchronous Q-Learning



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

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

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

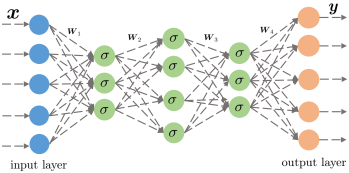
where μ_{min} is the smallest entry in the stationary distribution, and t_{mix} is the mixing time of the Markov chain.

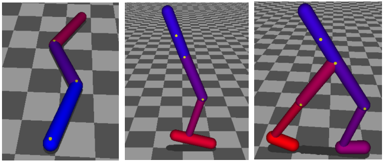
Understanding finite-time convergence of policy optimization, and how to accelerate it

Policy optimization

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

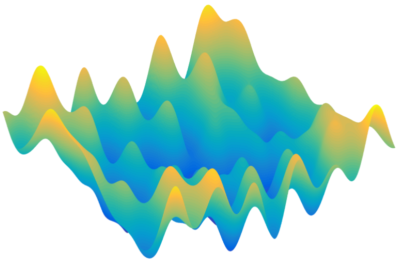
- directly optimize the policy, which is the quantity of interest;
- allow flexible differentiable parameterizations of the policy;
- work with both continuous and discrete problems.





Theoretical challenges: non-concavity

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



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:

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

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

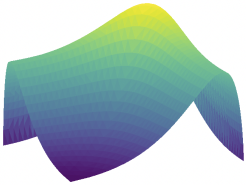
Policy gradient method (Sutton et al., 2000)

For $t = 0, 1, \cdots$

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

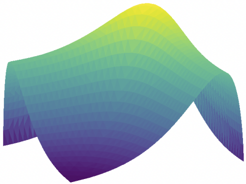
where η *is the learning rate.*

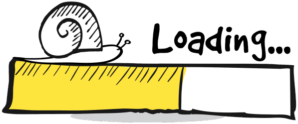
Global convergence of the PG method?



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

Global convergence of the PG method?

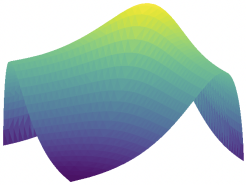


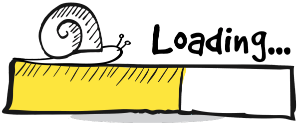


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

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

Global convergence of the PG method?

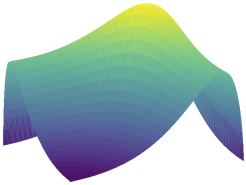


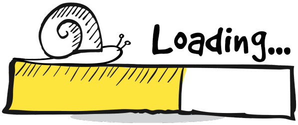


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

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

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)

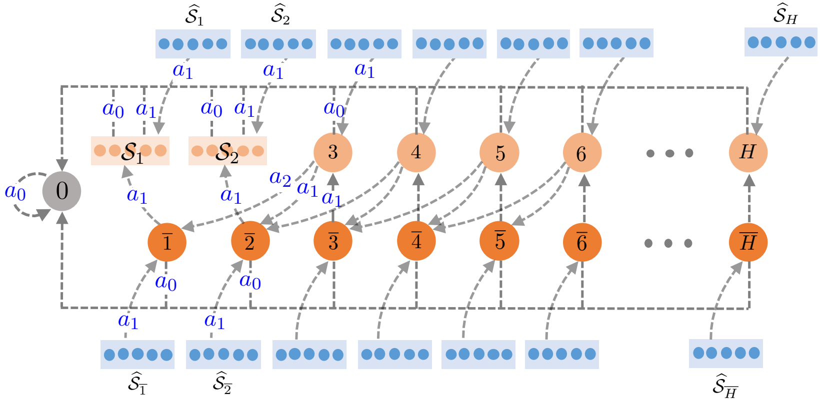
There exists an MDP s.t. it takes softmax PG at least

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

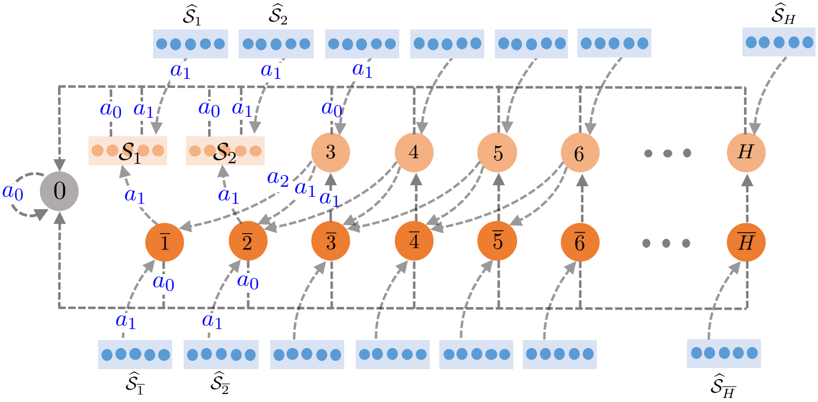
to achieve $\|V^{(t)} - V^{\star}\|_{\infty} \leq 0.15$.

- Softmax PG can take (super)-exponential time to converge (in problems w/ large state space & long effective horizon)!
- Even when starting from a uniform initial state distribution!
- 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

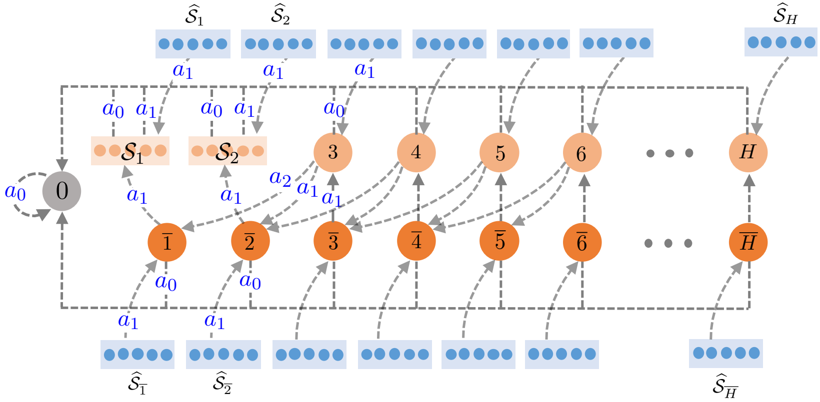


MDP construction for our lower bound



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

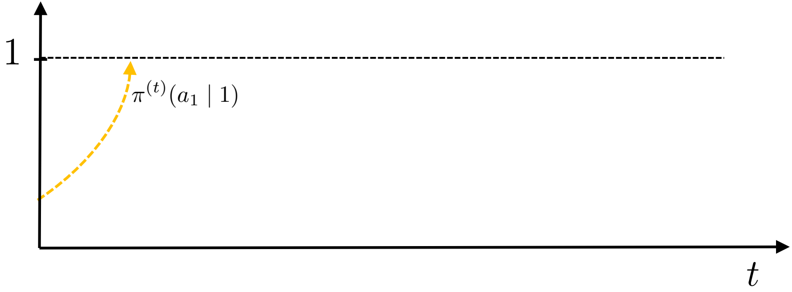
MDP construction for our lower bound



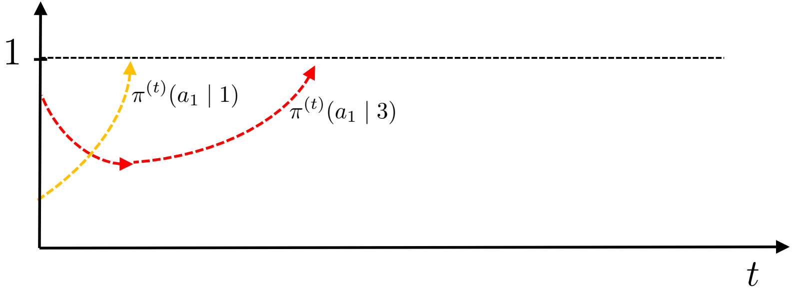
Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$,

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

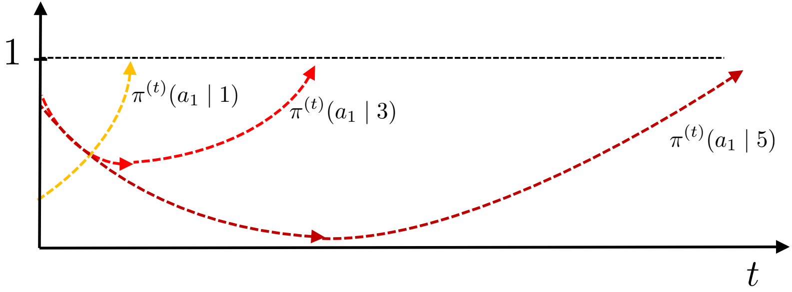
What is happening in our constructed MDP?



What is happening in our constructed MDP?

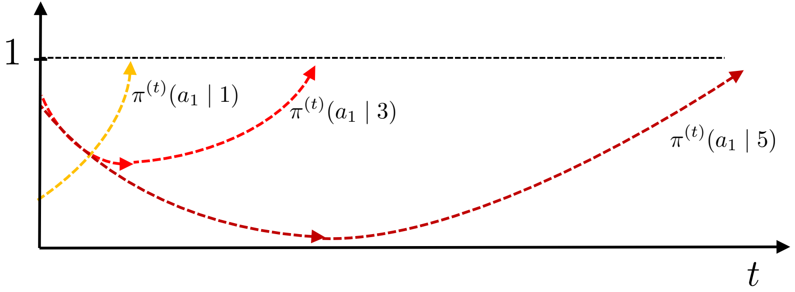


What is happening in our constructed MDP?



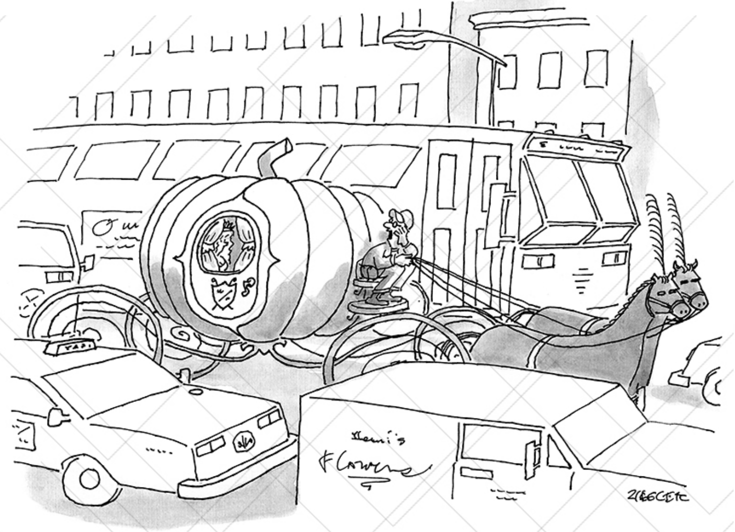
Convergence time for state s grows geometrically as s increases

What is happening in our constructed MDP?



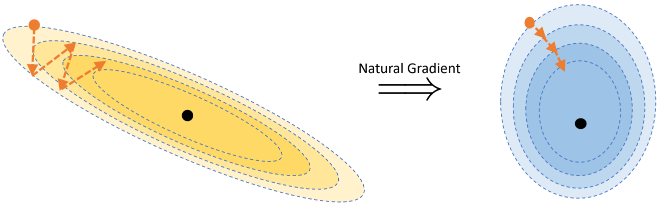
Convergence time for state s grows geometrically as s increases

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



"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)

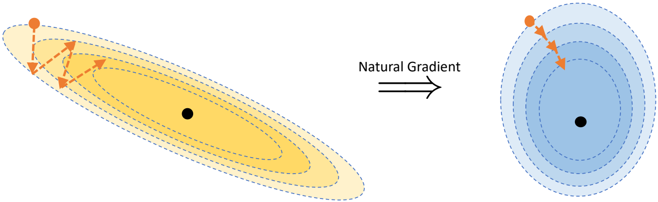
For $t = 0, 1, \cdots$

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

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 $t = 0, 1, \cdots$

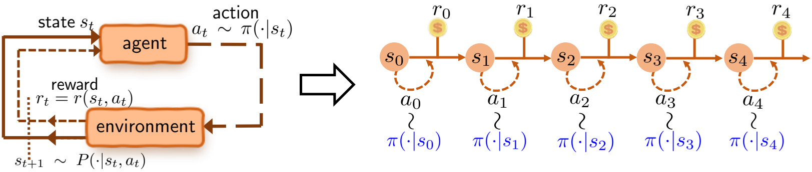
$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

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.

Booster #2: entropy regularization

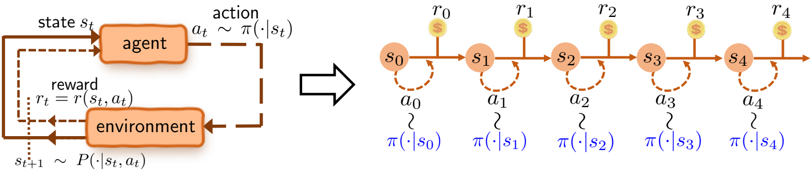


To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

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

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

Booster #2: entropy regularization



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

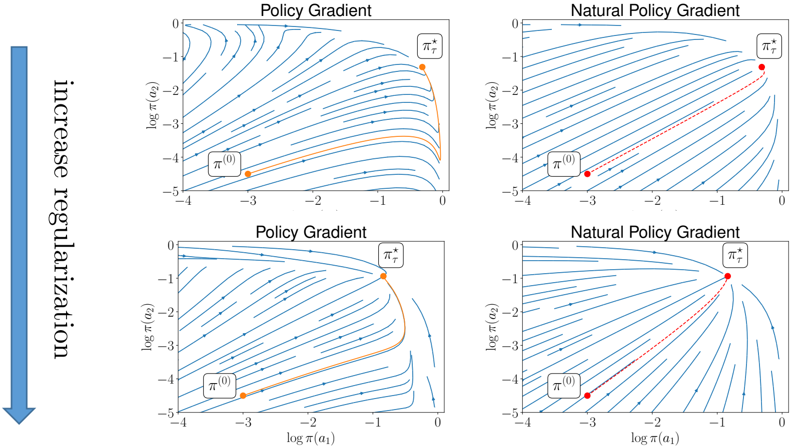
$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t))\Big) \,\Big|\, s_0 = s\right]$$

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

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

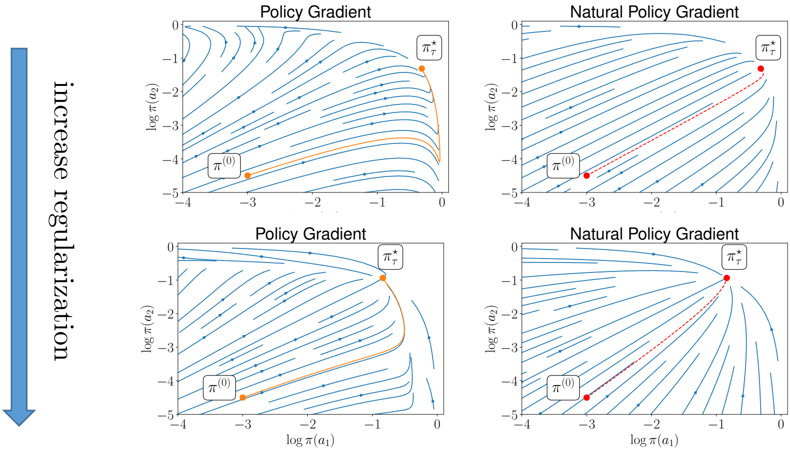
Entropy-regularized natural gradient helps!

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



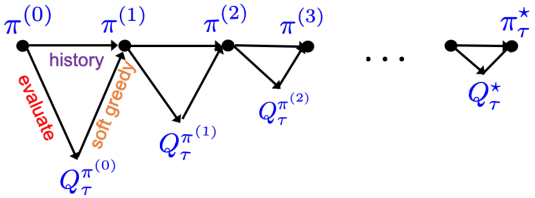
Entropy-regularized natural gradient helps!

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



Can we justify the efficacy of entropy-regularized NPG?

Entropy-regularized NPG in the tabular setting



Entropy-regularized NPG (Tabular setting)

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

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

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

- 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 $0 < \eta \le (1-\gamma)/\tau$, the entropy-regularized NPG updates satisfy

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

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

for all $t \ge 0$, 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 $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \epsilon$, the iteration complexity is at most

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

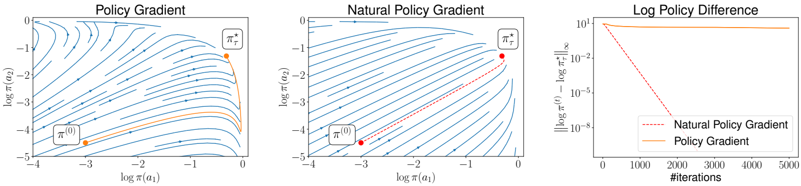
$$\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$$

- **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

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \leq \Big(V_\tau^\star(\rho) - V_\tau^{(0)}(\rho)\Big)$$

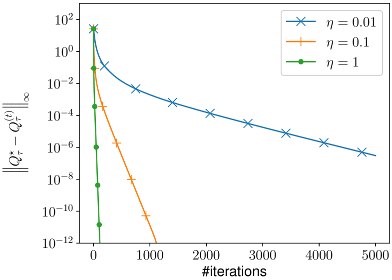
$$\cdot \exp\left(- \frac{(1-\gamma)^4 t}{(8/\tau + 4 + 8\log|\mathcal{A}|)|\mathcal{S}|} \left\| \frac{d_\rho^{\pi_\tau^\star}}{\rho} \right\|_\infty^{-1} \min_s \rho(s) \underbrace{\left(\inf_{0 \leq k \leq t-1} \min_{s,a} \pi^{(k)}(a|s) \right)^2}_{\text{can be exponential in } |\mathcal{S}| \text{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

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

Comparison with unregularized NPG

Regularized NPG

$\tau = 0.001$

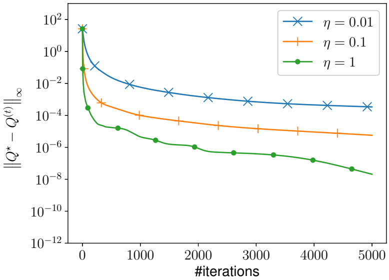


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Vanilla NPG

$\tau = 0$



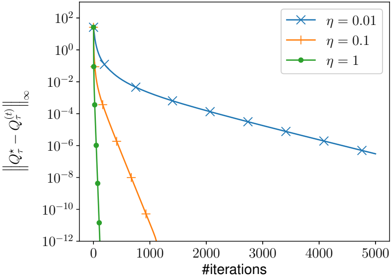
Sublinear rate: $\frac{1}{\min\{\eta,(1-\gamma)^2\}\epsilon}$

(Agarwal et al. 2019)

Comparison with unregularized NPG

Regularized NPG

$\tau = 0.001$

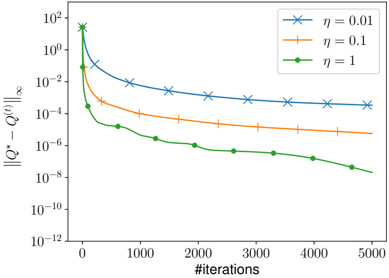


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Vanilla NPG

$\tau = 0$



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

(Agarwal et al. 2019)

Entropy regularization enables fast convergence!

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

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

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

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

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

Inexact entropy-regularized NPG:

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

Entropy-regularized NPG with inexact gradients

Inexact oracle: inexact evaluation of $Q_\tau^{\pi^{(t)}}$ 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).

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: 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

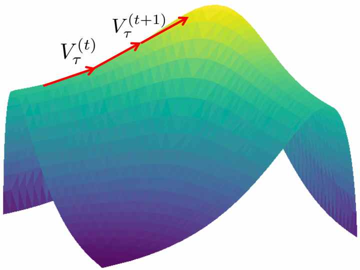
Theorem (Cen, Cheng, Chen, Wei, Chi '20; improved)

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

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

- **Sample complexity for the original MDP:** set $\tau = \frac{(1-\gamma)\epsilon}{\log|\mathcal{A}|}$; using fresh samples for policy evaluation at every iteration requires

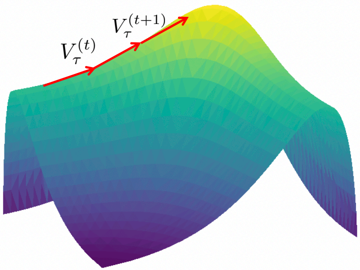
$$\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^7\epsilon^2}\right) \text{ samples.}$$



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

discounted state visitation distribution

A key lemma: monotonic performance improvement



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

discounted state visitation distribution

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: $Q^\star_\tau$ is *unique* solution to

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

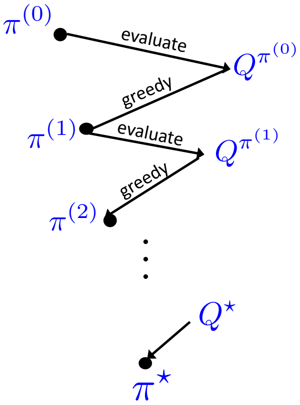
γ**-contraction of soft Bellman operator:**

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

Richard Bellman

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

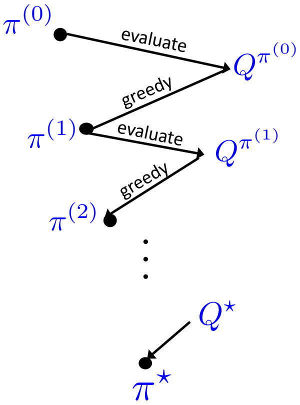
Policy iteration



Bellman operator

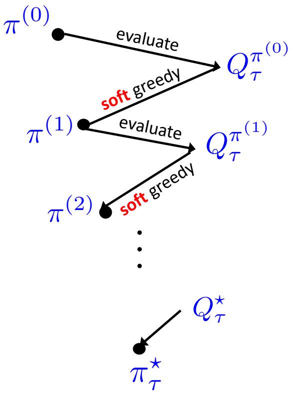
Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

Policy iteration



Bellman operator

Soft policy iteration



Soft Bellman operator

A key linear system: general learning rates

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

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

A key linear system: general learning rates

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

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

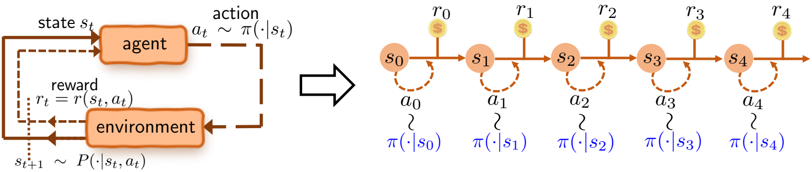
sparse exploration

Tsallis entropy

constrained and safe RL

log-barrier

Regularized RL in general form

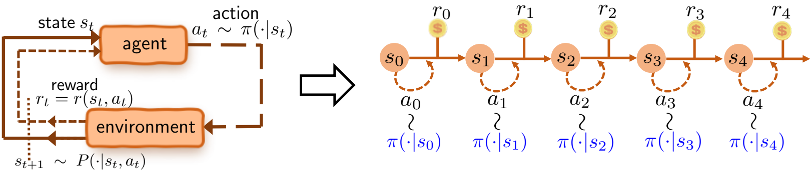


The regularized value function is defined as

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

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

Regularized RL in general form



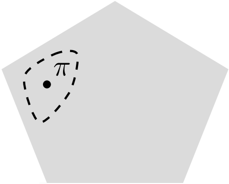
The regularized value function is defined as

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

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)

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

for all $s\in\mathcal{S}$, where the KL divergence is the Bregman divergence w.r.t. the negative Shannon entropy.

Generalized Policy Mirror Descent (GPMD)

Generalized policy mirror descent (GPMD) method

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

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

where a surrogate of $\partial h_s(\pi^{(t)}(\cdot|s))$ *is updated recursively.*

Generalized Policy Mirror Descent (GPMD)

Generalized policy mirror descent (GPMD) method

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

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

where a surrogate of $\partial h_s(\pi^{(t)}(\cdot|s))$ *is updated recursively.*

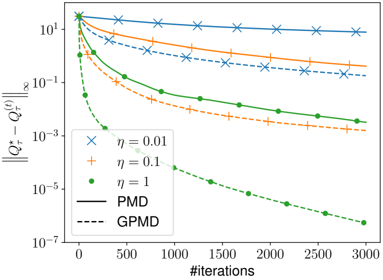
- Compare with PMD (Lan, 2021):

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

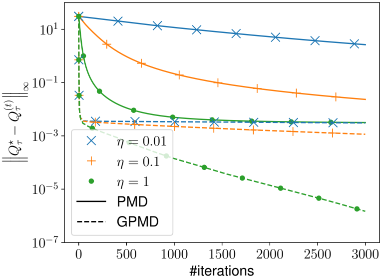
  GPMD achieves linear convergence for general convex and nonsmooth h_s! In contrast, PMD requires $h_s + \mathcal{H}$ is convex.

Numerical examples

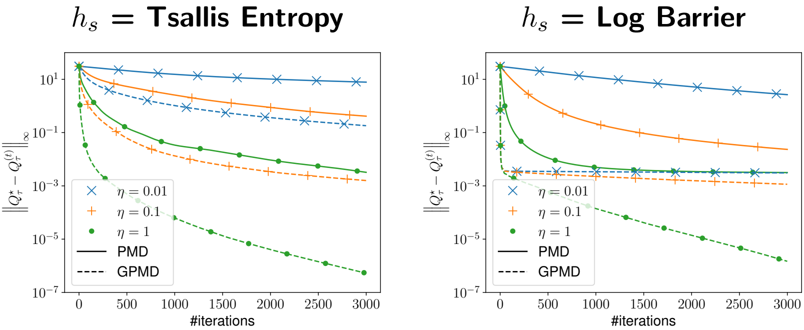
h_s = **Tsallis Entropy**



h_s = **Log Barrier**



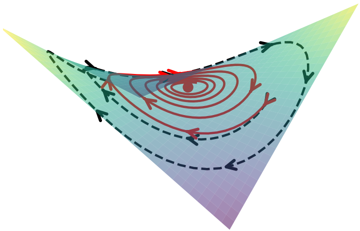
Numerical examples

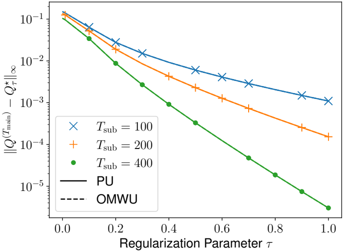


GPMD achieves faster convergence than PMD!

Entropy-regularized zero-sum two-player Markov game

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho)$$

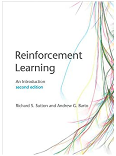


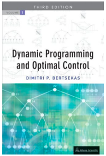


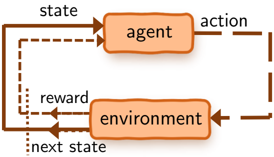
(Cen et. al., NeurIPS 2021): OMWU with value iteration = dimension-free rate, last-iterate convergence, symmetric updates

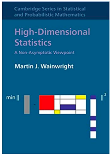
Concluding remarks

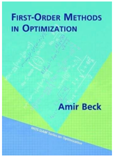
Concluding remarks











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

Future directions:

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

References

Q-learning and variants:

- Is Q-learning minimax optimal? a tight sample complexity analysis, arXiv preprint arXiv:2102.06548, short version at ICML 2021.
- Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction, *IEEE Trans. on Information Theory*, short version at NeurIPS 2020.
- Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning, arXiv:2110.04645, short version at NeurIPS 2021.

Policy optimization:

- Fast global convergence of natural policy gradient methods with entropy regularization, *Operations Research*, in press.
- Softmax policy gradient methods can take exponential time to converge, arXiv:2102.11270, short version at COLT 2021.
- Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence, arXiv:2105.11066.
- Fast policy extragradient methods for competitive games with entropy regularization, arXiv:2105.15186, short version at NeurIPS 2021.

Thank you!

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