Taming nonconvexity in policy optimization

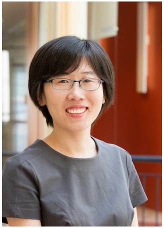
Yuxin Chen

ECE, Princeton University

Shicong Cen
CMU

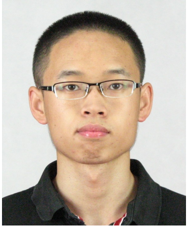
Chen Cheng
Stanford

Yuting Wei
CMU



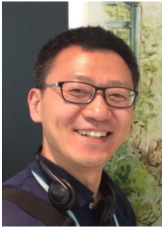
Yuejie Chi
CMU

"Fast global convergence of natural policy gradient methods with entropy regularization," S. Cen, C. Cheng, Y. Chen, Y. Wei, Y. Chi, under revision, *Operations Research*, 2020

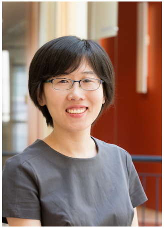


Gen Li
Tsinghua

Yuting Wei
CMU



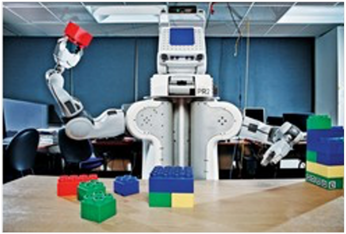
Yuantao Gu
Tsinghua



Yuejie Chi
CMU

"Softmax policy gradient methods can take exponential time to converge,"
G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, arxiv:2102.11270, 2021

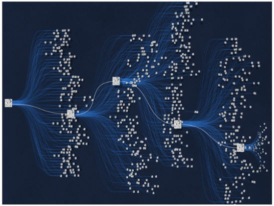
Reinforcement learning (RL)



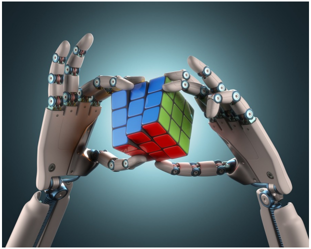
RL challenges

In RL, an agent learns by interacting with an environment

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



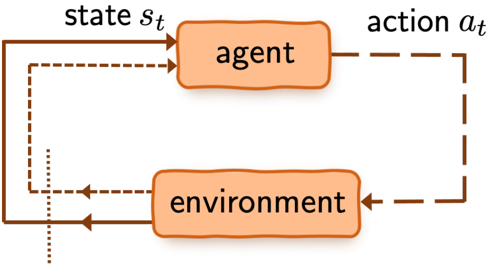
Recent successes in RL

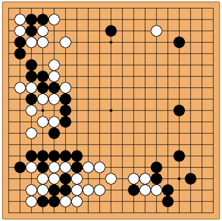


Policy optimization: a major contributor to these successes

Backgrounds: policy optimization for MDPs

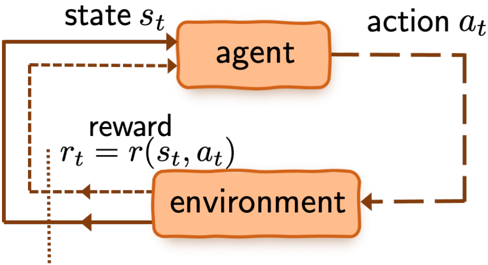
Markov decision process (MDP)

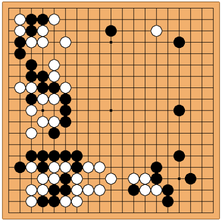




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

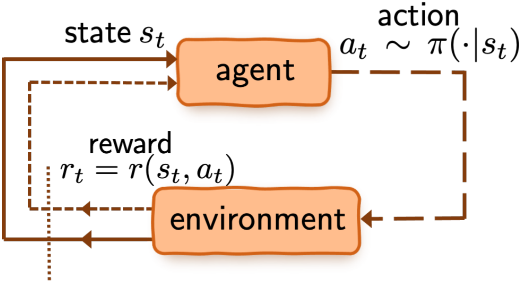
Markov decision process (MDP)

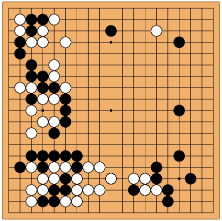




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

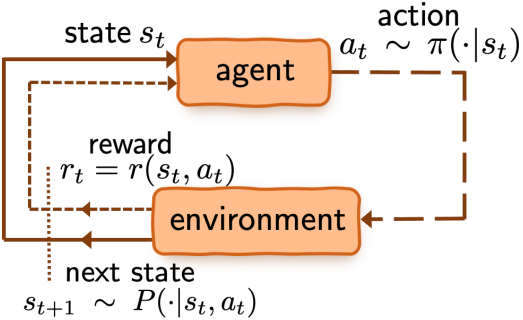
Markov decision process (MDP)

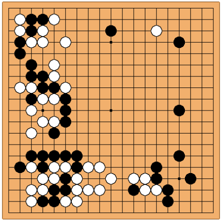




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

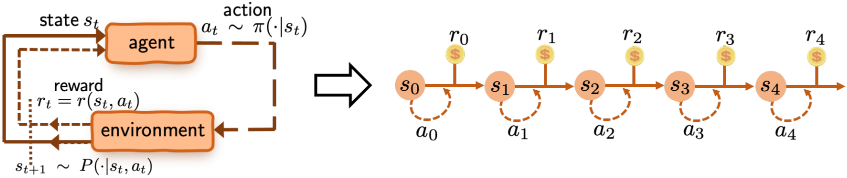
Markov decision process (MDP)





- $\mathcal{S}$: 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 and Q-function of policy π

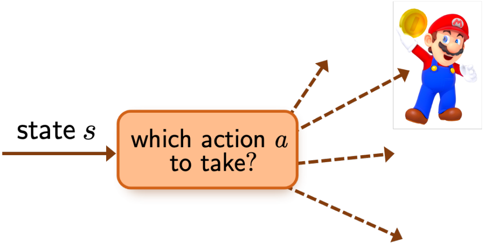


$$\forall s \in \mathcal{S} : \qquad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right]$$

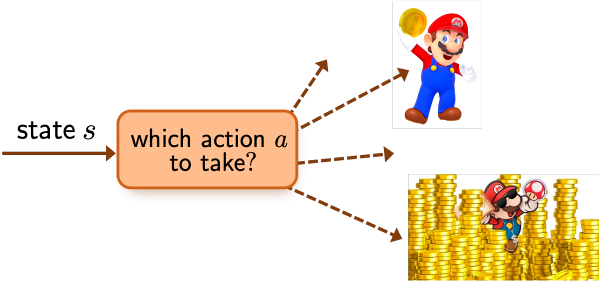
$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

- cumulative *discounted* reward; $\gamma \in [0,1)$: discount factor
  - **effective horizon**: $\frac{1}{1-\gamma}$
- sampled trajectory is generated under π

Optimal policy and optimal value



Optimal policy and optimal value



- **goal:** find optimal policy $\pi^\star$ that maximizes values
- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$

Policy optimization

Given state distribution $s \sim \rho$ (e.g. uniform)

$$\boxed{\max_{\pi} \; V^{\pi}(\rho) \coloneqq \mathbb{E}_{s\sim\rho}\left[V^{\pi}(s)\right]}$$

Policy optimization

Given state distribution $s \sim \rho$
(e.g. uniform)

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

Policy optimization

Given state distribution $s \sim \rho$
(e.g. uniform)

$$\max_{\pi} \; V^{\pi}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi}(s)\right] \xrightarrow{\text{parameterize}} \max_{\theta} \; V^{\pi_\theta}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

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

softmax parameterization

Policy optimization

Given state distribution $s \sim \rho$
(e.g. uniform)

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

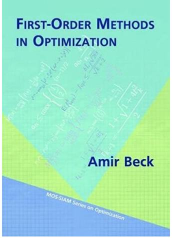
parameterize $\Longrightarrow$

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

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

softmax parameterization

$\Downarrow$



Policy optimization

Given state distribution $s \sim \rho$
(e.g. uniform)

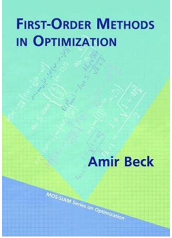
$$\max_{\pi} \; V^{\pi}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

parameterize

$$\max_{\theta} \; V^{\pi_\theta}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$$

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

softmax parameterization

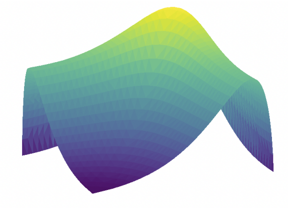


Policy gradient method (Sutton et al. '00)

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

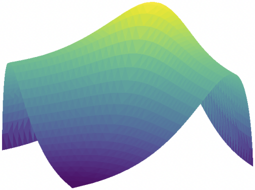
- η: learning rate

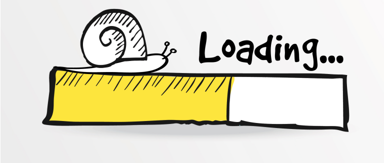
Does policy gradient (PG) method converge?



- (Agarwal et al. '19) Softmax PG converges to global opt as $t \to \infty$

Does policy gradient (PG) method converge?

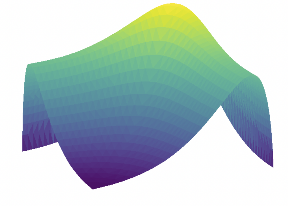


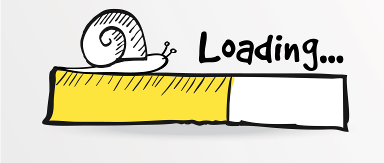


- (Agarwal et al. '19) Softmax PG converges to global opt as $t \to \infty$

However, "asymptotic convergence" might mean "taking forever"

Does policy gradient (PG) method converge?



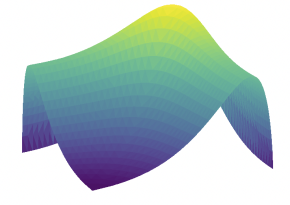


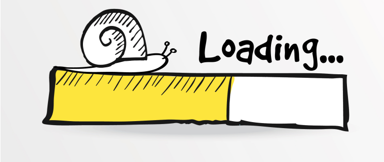
- (Agarwal et al. '19) Softmax PG converges to global opt as $t \to \infty$
- (Mei et al. '20) Softmax PG converges to global opt in

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

However, "asymptotic convergence" might mean "taking forever"

Does policy gradient (PG) method converge?





- (Agarwal et al. '19) Softmax PG converges to global opt as $t \to \infty$
- (Mei et al. '20) Softmax PG converges to global opt in

$$\color{red}{c\left(|\mathcal{S}|, |\mathcal{A}|, \tfrac{1}{1-\gamma}, \cdots\right)} \color{blue}{O\left(\tfrac{1}{\varepsilon}\right)} \text{ iterations}$$

However, "asymptotic convergence" might mean "taking forever"

A negative message

Theorem 1 (Li, Wei, Chi, Gu, Chen '21)

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

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

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

A negative message

Theorem 1 (Li, Wei, Chi, Gu, Chen '21)

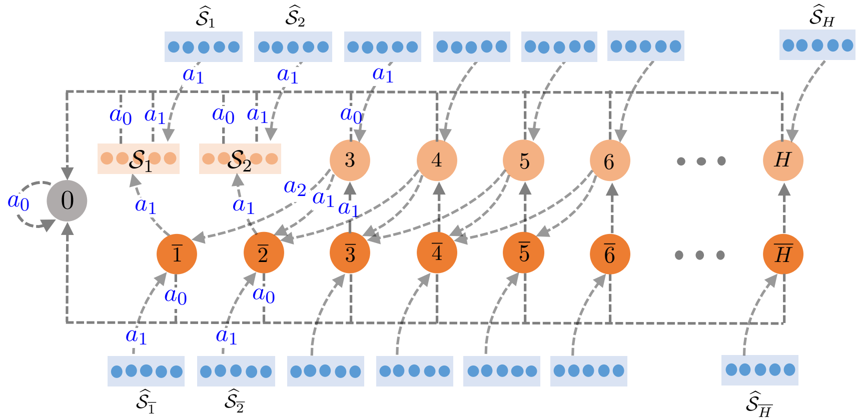
There exists an MDP s.t. it takes softmax PG at least

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

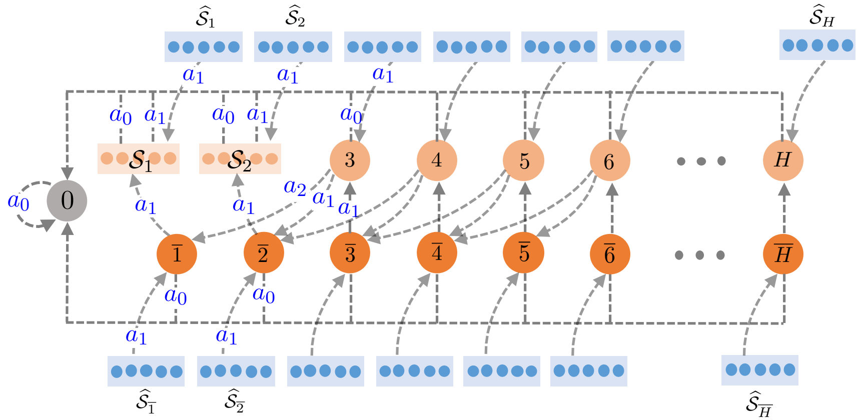
to achieve $\|V^{(t)} - V^{\star}\|_{\infty} \leq 0.15$

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

MDP construction for our lower bound

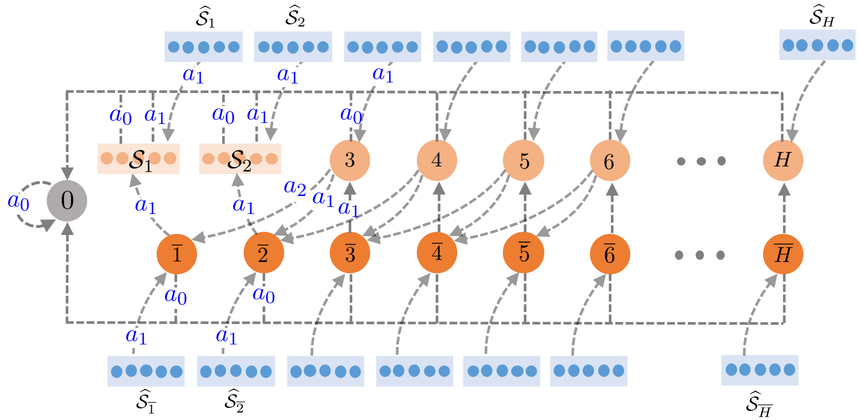


MDP construction for our lower bound



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

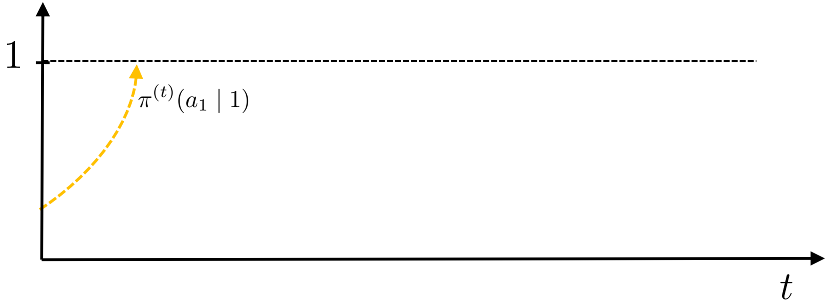
MDP construction for our lower bound



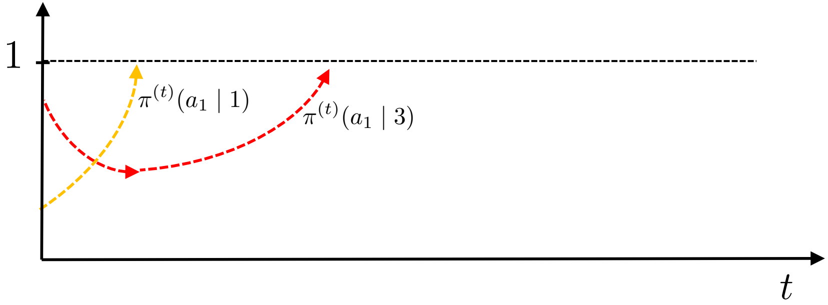
Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$,

- $\pi^{(t)}(a_{\mathsf{opt}} \,|\, s)$ keeps decreasing until $\pi^{(t)}(a_{\mathsf{opt}} \,|\, s-2) \approx 1$

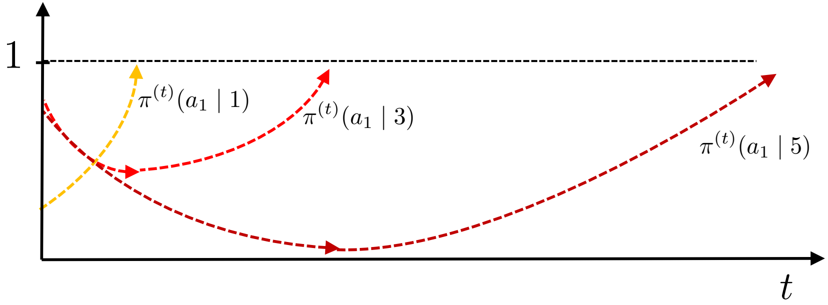
What is happening in our constructed MDP?



What is happening in our constructed MDP?

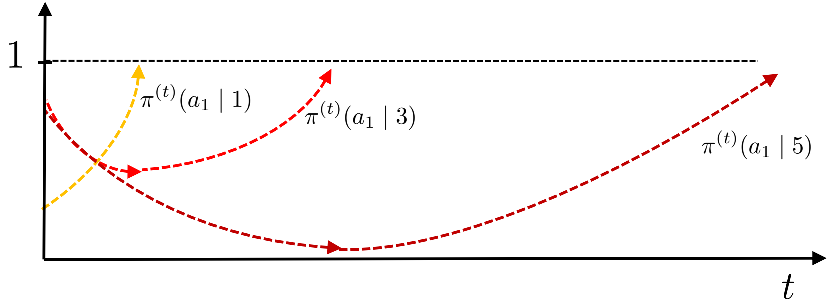


What is happening in our constructed MDP?



observation: convergence time for state s grows geometrically as $s \uparrow$

What is happening in our constructed MDP?



observation: convergence time for state s grows geometrically as $s \uparrow$

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

Booster 1: entropy regularization

accelerate convergence by regularizing value function

$$V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r_t - \tau \log \pi(a_t \mid s_t)\big) \,\Big|\, s_0 = s\right]$$

Booster 1: entropy regularization

accelerate convergence by regularizing value function

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

- τ: regularization parameter
- d_s^π: certain marginal distribution

Booster 1: entropy regularization

accelerate convergence by regularizing value function

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

- τ: regularization parameter
- d_s^π: certain marginal distribution

entropy-regularized value maximization

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

Entropy-regularized PG remains slow ...

Theorem 2 (Li, Wei, Chi, Gu, Chen '21)

There is an MDP s.t. it takes entropy-regularized softmax PG at least

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

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

- Softmax PG method with entropy regularization can still take exponential time to converge!

Entropy-regularized PG remains slow . . .

Theorem 2 (Li, Wei, Chi, Gu, Chen '21)

There is an MDP s.t. it takes entropy-regularized softmax PG at least

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

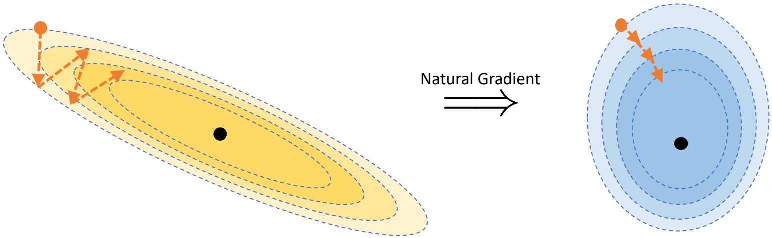
to achieve $\|V^{(t)} - V^{\star}\|_{\infty} \leq \varepsilon$

- Softmax PG method with entropy regularization can still take exponential time to converge!
- (Mei et al. '20) entropy-regularized softmax PG converges in

$$c\big(|\mathcal{S}|, |\mathcal{A}|, \tfrac{1}{1-\gamma}, \cdots\big)\, O\big(\tfrac{1}{\varepsilon}\big) \text{ iterations}$$

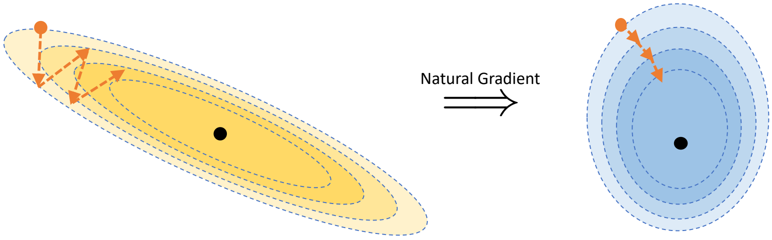
Booster 2: natural policy gradient (NPG)

precondition gradients to improve search directions ...



Booster 2: natural policy gradient (NPG)

precondition gradients to improve search directions ...



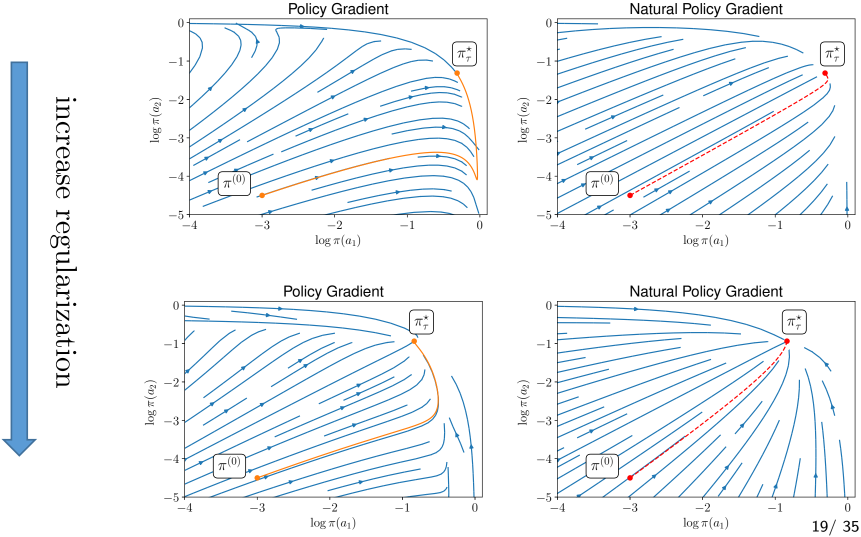
NPG method (Kakade '02)

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

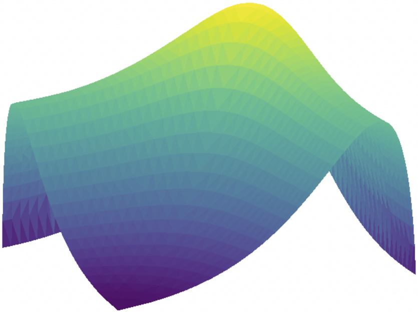
- $\mathcal{F}_\rho^\theta := \mathbb{E}\Big[\big(\nabla_\theta \log \pi_\theta(a \,|\, s)\big)\big(\nabla_\theta \log \pi_\theta(a \,|\, s)\big)^\top\Big]$: Fisher info

Entropy-regularized natural gradient helps!

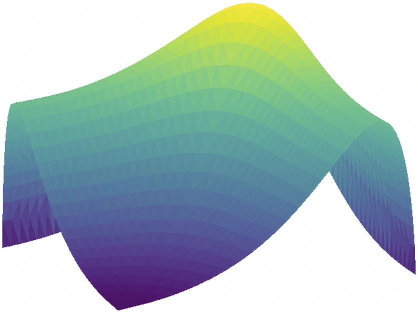
A toy bandit example: 3 arms with rewards 1, 0.9 and 0.1



Challenge: non-concavity

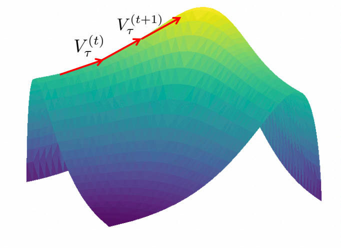


Challenge: non-concavity



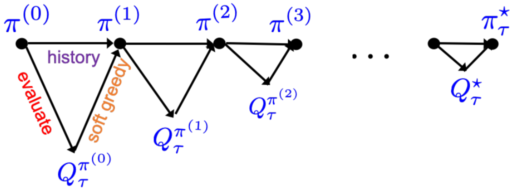
Recent advances

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



How to characterize the efficiency of entropy-regularized NPG in tabular settings?

Entropy-regularized NPG in tabular settings



An alternative expression in policy space (tabular setting)

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

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

- invariant to the choice of initial state distribution ρ

Linear convergence with exact gradients

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

Exact oracle: perfect gradient evaluation

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

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

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

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

Implications: iteration complexity

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

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

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

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

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

Implications: iteration complexity

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

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

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

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

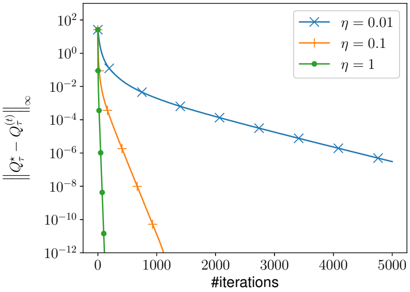
$$\frac{1}{1-\gamma} \log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\varepsilon}\right)$$

Nearly dimension-free global linear convergence!

Regularized NPG vs. unregularized NPG

regularized NPG

$\tau = 0.001$

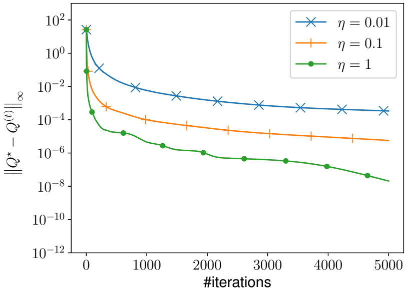


linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\varepsilon}\right)$

Ours

unregularized NPG

$\tau = 0$

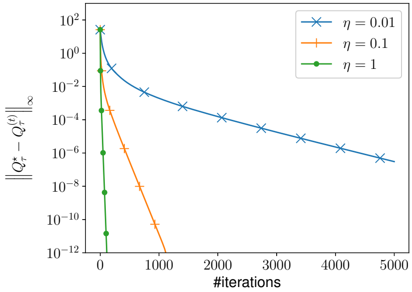


sublinear rate: $\frac{1}{\min\{\eta,(1-\gamma)^2\}\varepsilon}$

(Agarwal et al. '19)

Regularized NPG vs. unregularized NPG

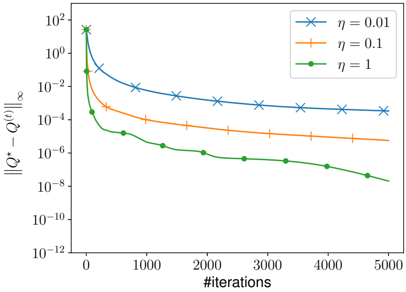
regularized NPG
$\tau = 0.001$



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

Ours

unregularized NPG
$\tau = 0$



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

(Agarwal et al. '19)

Entropy regularization enables faster convergence!

Entropy-regularized NPG with inexact gradients

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

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

e.g. using sample-based estimators

Entropy-regularized NPG with inexact gradients

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

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

e.g. using sample-based estimators

Inexact entropy-regularized NPG:

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

Entropy-regularized NPG with inexact gradients

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

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

e.g. using sample-based estimators

Inexact entropy-regularized NPG:

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

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

Linear convergence with inexact gradients

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

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

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

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

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

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

Returning to the original MDP?

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

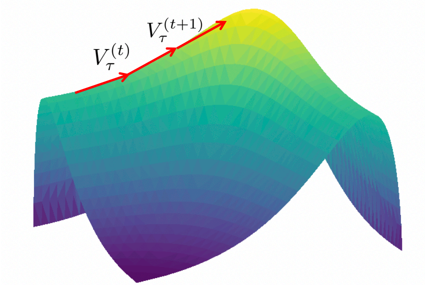
Returning to the original MDP?

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

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

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

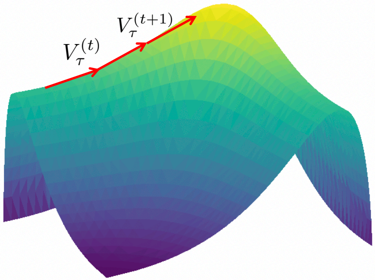
A key lemma: monotonic performance improvement



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

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

A key lemma: monotonic performance improvement



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

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

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

Recall: 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

Recall: 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

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

γ**-contraction of Bellman operator:**

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

Richard Bellman

Soft Bellman operator

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

Soft Bellman operator

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

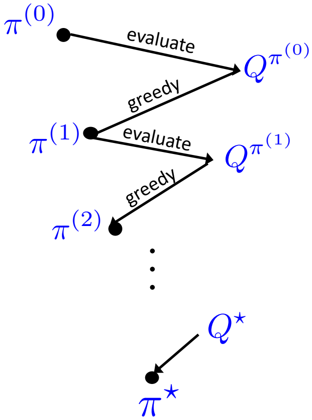
Soft Bellman equation: $Q^\star_\tau$ is *unique* solution to

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

γ-contraction of soft Bellman operator:

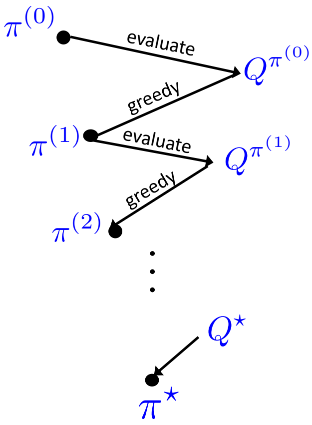
$$\|\mathcal{T}_\tau(Q_1) - \mathcal{T}_\tau(Q_2)\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$$

policy iteration



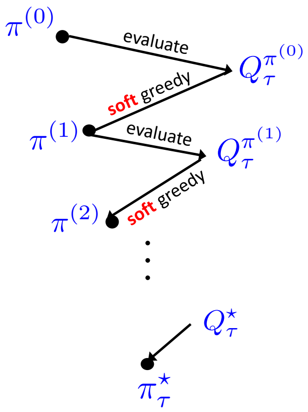
Bellman operator

policy iteration



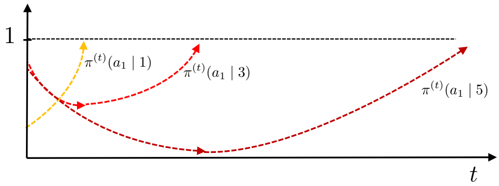
Bellman operator

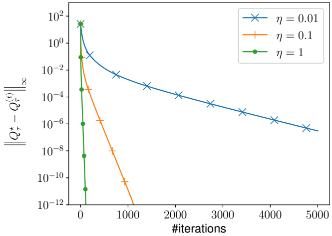
soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)



soft Bellman operator

Concluding remarks





- Softmax policy gradient can take exponential time to converge
- Entropy regularization & natural gradients help!

Papers:

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

"Softmax policy gradient methods can take exponential time to converge," G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, arxiv:2102.11270, 2021