

On the Effectiveness of Nonconvex Policy Optimization



Yuxin Chen

ECE, Princeton University



Shicong Cen
CMU



Chen Cheng
Stanford

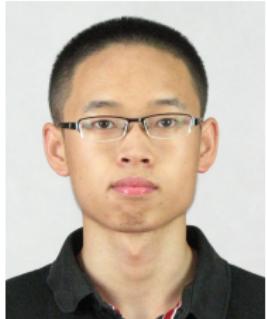


Yuting Wei
Wharton



Yuejie Chi
CMU

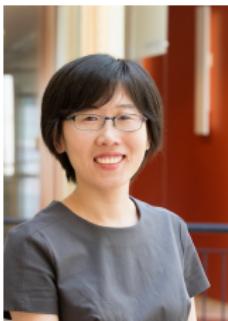
“Fast global convergence of natural policy gradient methods with entropy regularization,” S. Cen, C. Cheng, Y. Chen, Y. Wei, Y. Chi, accepted to *Operations Research*, 2020



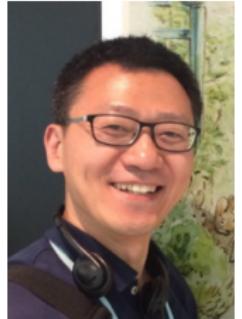
Gen Li
Tsinghua



Yuting Wei
Wharton



Yuejie Chi
CMU



Yuantao Gu
Tsinghua

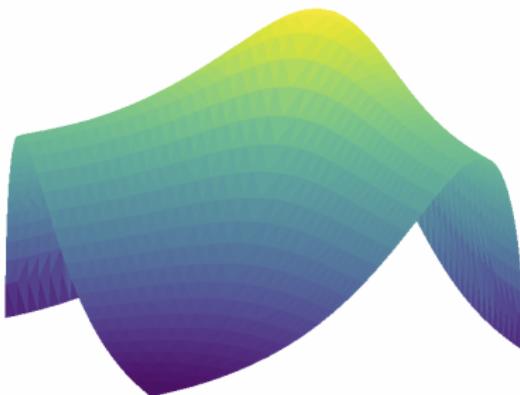
“Softmax policy gradient methods can take exponential time to converge,”
G. Li, Y. Wei, Y. Chi, Y. Gu, Y. Chen, arxiv:2102.11270, accepted to COLT
2021

Recent successes in RL



Policy optimization: a major contributor to these successes

Challenges: large dimensionality and non-concavity



Recent advances towards understanding policy optimization

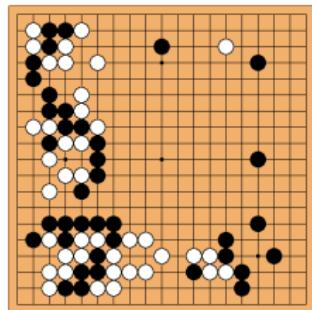
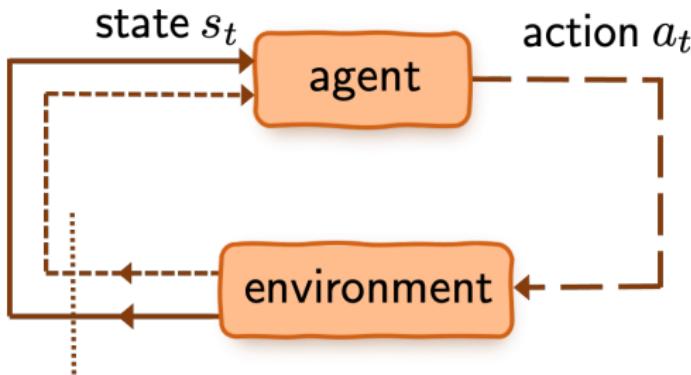
- finite-state MDPs ([Agarwal et al. 19](#), [Bhandari and Russo '19](#), [Shani et al. '19](#), [Mei et al. '20](#), [Cen et al. '20](#), [Zhang et al. '20](#), [Lan '21](#), [Zhan et al. '21](#), [Cen et al. '21](#), [Cayci et al. '21](#), [Khodadadian et al. '21](#), ...)
- control ([Fazel et al., 2018](#); [Bhandari and Russo, 2019](#), ...)

This talk

1. *A negative message*: exponential lower bound for softmax PG
2. *A positive message*: fast convergence of entropy-regularized NPG

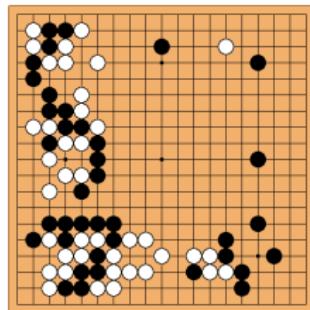
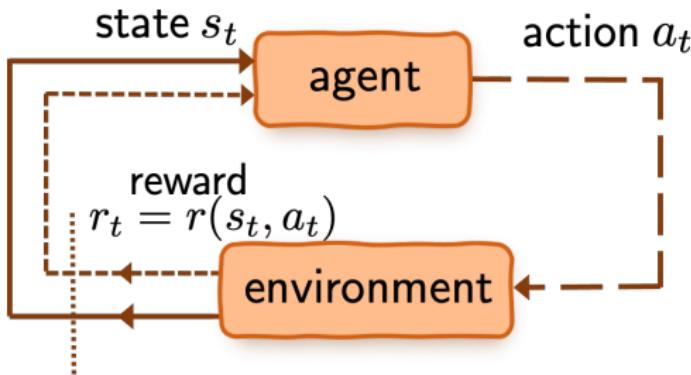
Backgrounds: policy optimization for MDPs

Markov decision process (MDP)



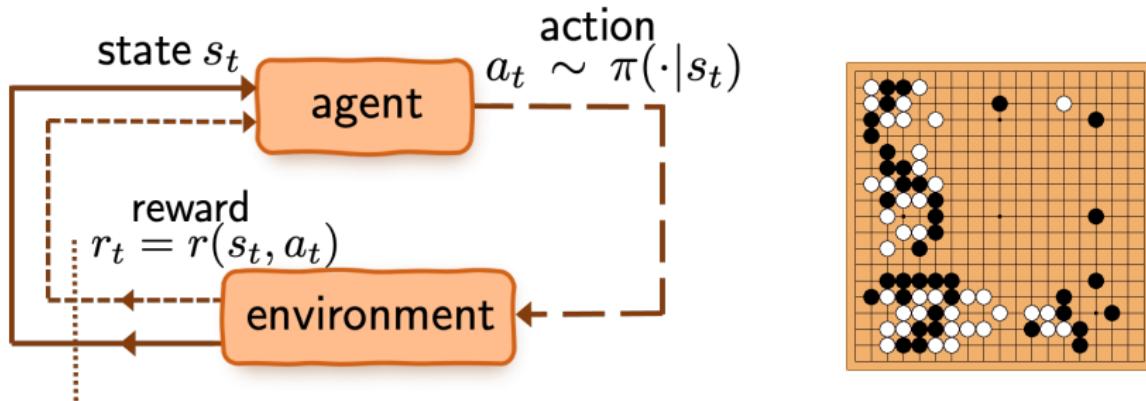
- \mathcal{S} : state space
- \mathcal{A} : action space

Markov decision process (MDP)



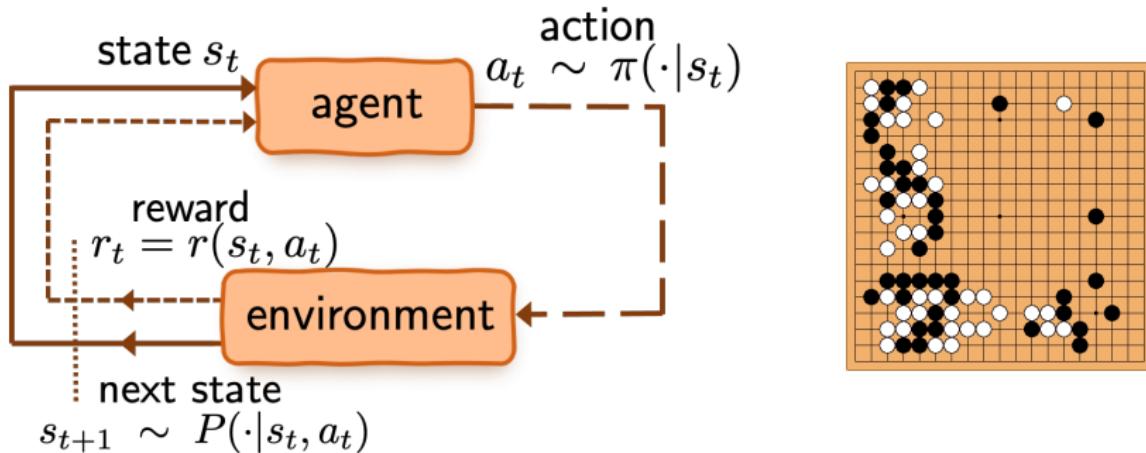
- \mathcal{S} : state space
- $r(s, a) \in [0, 1]$: immediate reward
- \mathcal{A} : action space

Markov decision process (MDP)



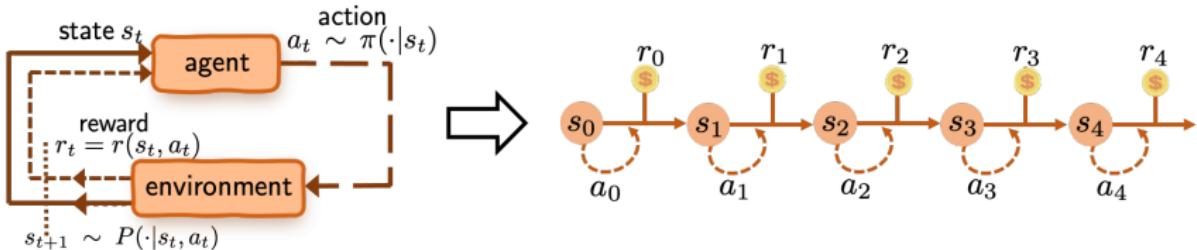
- \mathcal{S} : state space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot | s)$: policy (or action selection rule)
- \mathcal{A} : action space

Markov decision process (MDP)



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

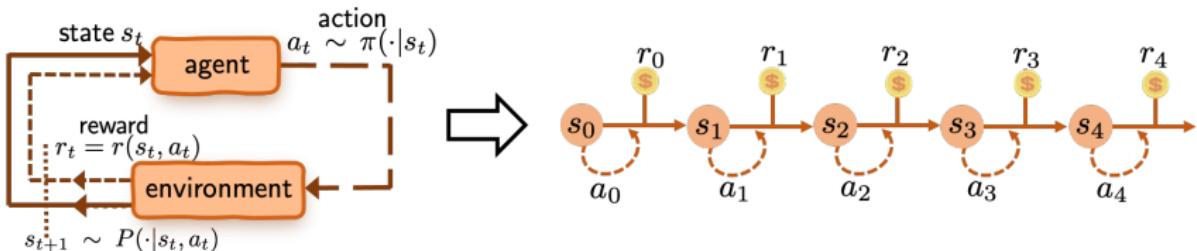
Value function of policy π



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

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

Value function of policy π



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

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

Our focus: large state/action space + long effective horizon $\frac{1}{1-\gamma}$

Optimal policy and optimal value



- **goal:** find optimal policy π^* that maximizes values
- optimal value: $V^* := V^{\pi^*}$

Policy optimization

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

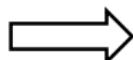
$$\max_{\pi} V^{\pi}(\rho) := \mathbb{E}_{s \sim \rho} [V^{\pi}(s)]$$

Policy optimization

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

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

parameterize



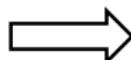
$$\max_{\theta} V^{\pi_{\theta}}(\rho) := \mathbb{E}_{s \sim \rho} [V^{\pi_{\theta}}(s)]$$

Policy optimization

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

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

parameterize



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

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

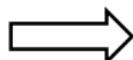
softmax parameterization

Policy optimization

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

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

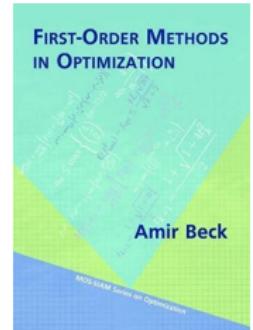
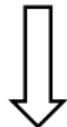
parameterize



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

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

softmax parameterization

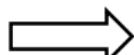


Softmax policy gradient methods

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

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

parameterize



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

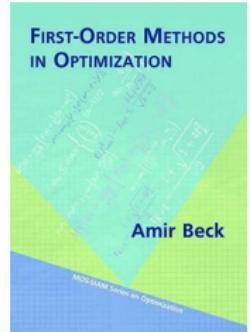
$$\pi_{\theta}(a|s) = \frac{\exp(\theta(s, a))}{\sum_a \exp(\theta(s, a))}$$

softmax parameterization

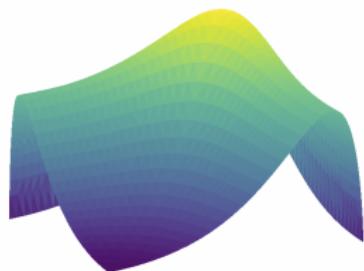
Policy gradient method (Sutton et al. '00)

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

- η : learning rate

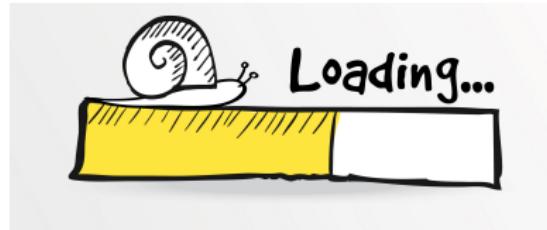
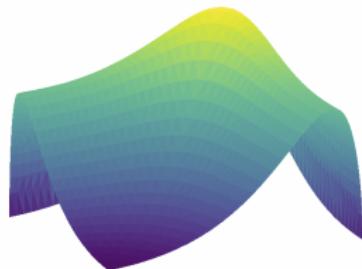


Does policy gradient (PG) method converge?



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

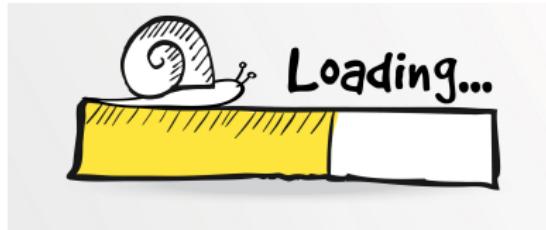
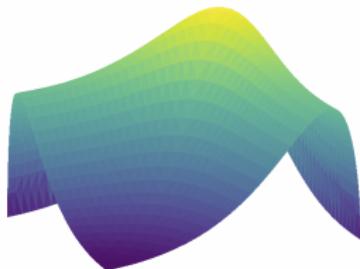
Does policy gradient (PG) method converge?



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

However, “asymptotic convergence” might mean “taking forever”

Does policy gradient (PG) method converge?

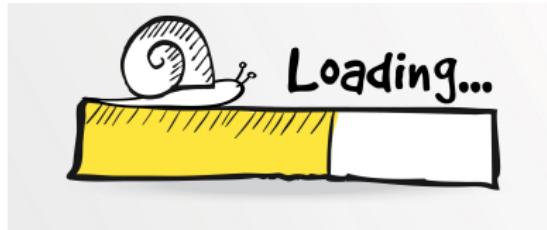
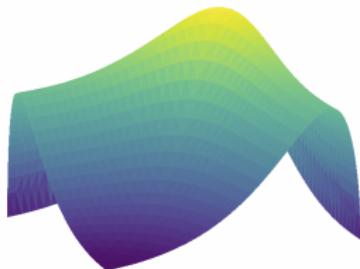


- (Agarwal et al. '19) Softmax PG converges to global opt as $t \rightarrow \infty$
- (Mei et al. '20) Softmax PG converges to global opt in

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

However, “asymptotic convergence” might mean “taking forever”

Does policy gradient (PG) method converge?



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

$$c(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \dots) O(\frac{1}{\varepsilon}) \text{ iterations}$$

However, “asymptotic convergence” might mean “taking forever”

A negative message: exponential lower bound

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 $\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} |V^{(t)}(s) - V^*(s)| \leq 0.07$

A negative message: exponential lower bound

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 $\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} |V^{(t)}(s) - V^*(s)| \leq 0.07$

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

A negative message: exponential lower bound

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

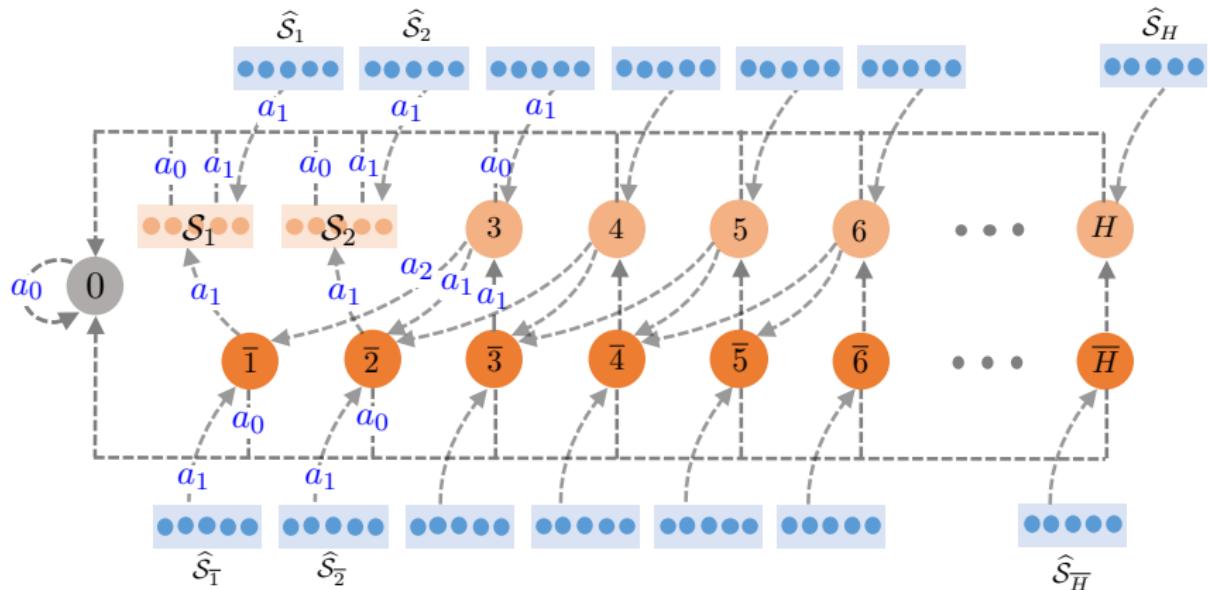
There exists an MDP s.t. it takes softmax PG at least

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

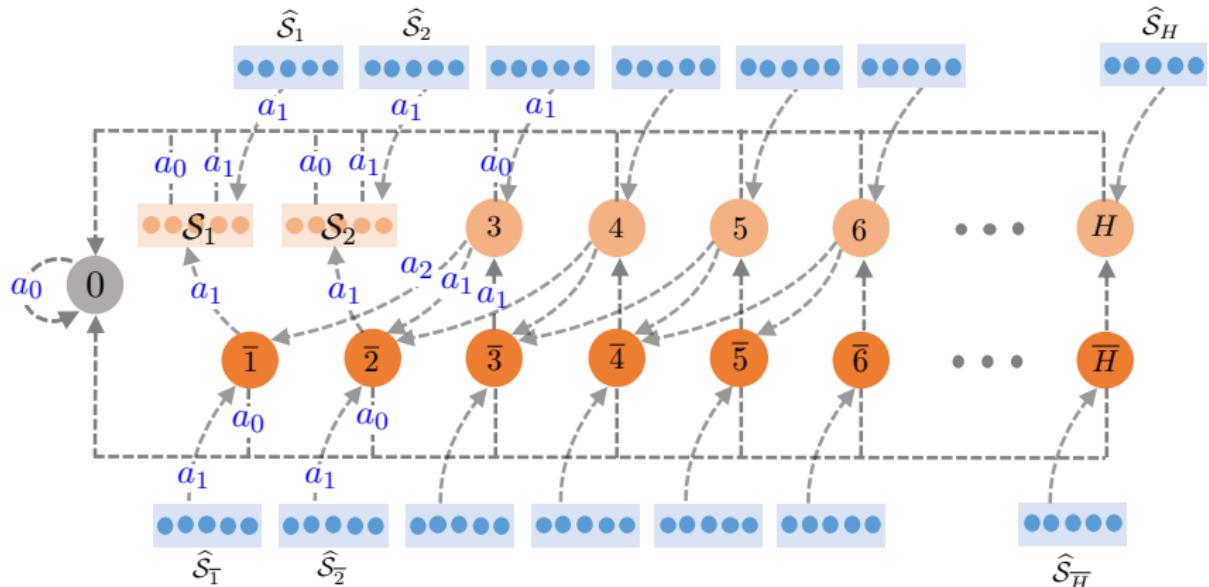
to achieve $\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} |V^{(t)}(s) - V^*(s)| \leq 0.07$

- This (super)-exponential lower bound arises even with
 - uniform initial state distribution
→ benign distribution mismatch $\left\| \frac{d_\rho^\pi}{\rho} \right\|_\infty \leq |\mathcal{S}|$
 - uniform policy initialization

MDP construction for our lower bound



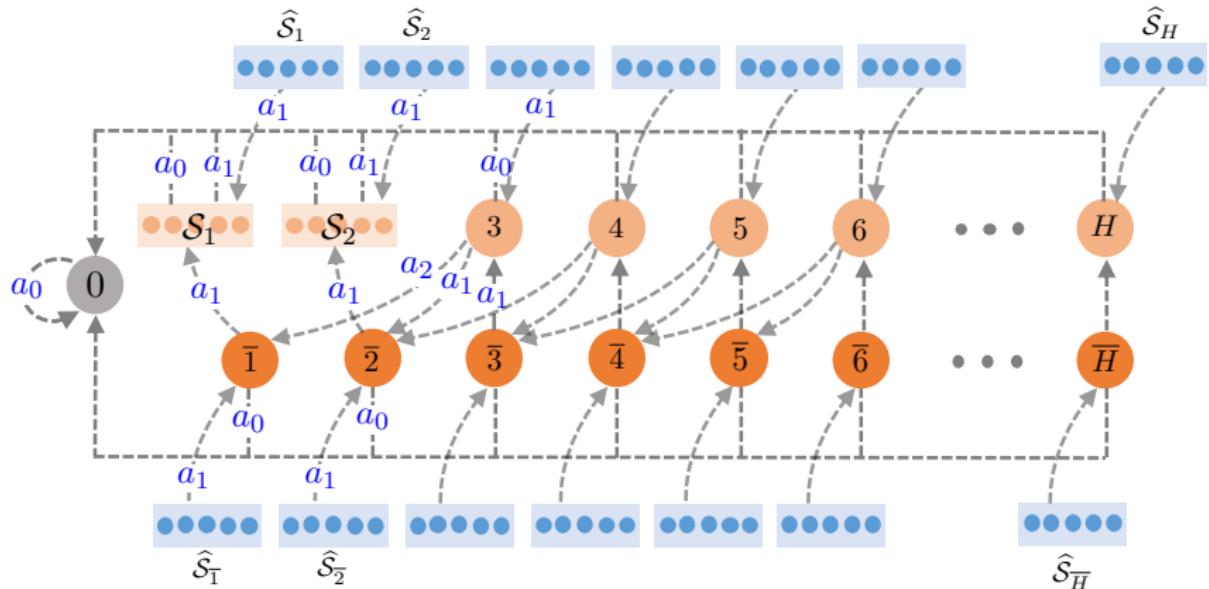
MDP construction for our lower bound



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

- augmented chain-like structure

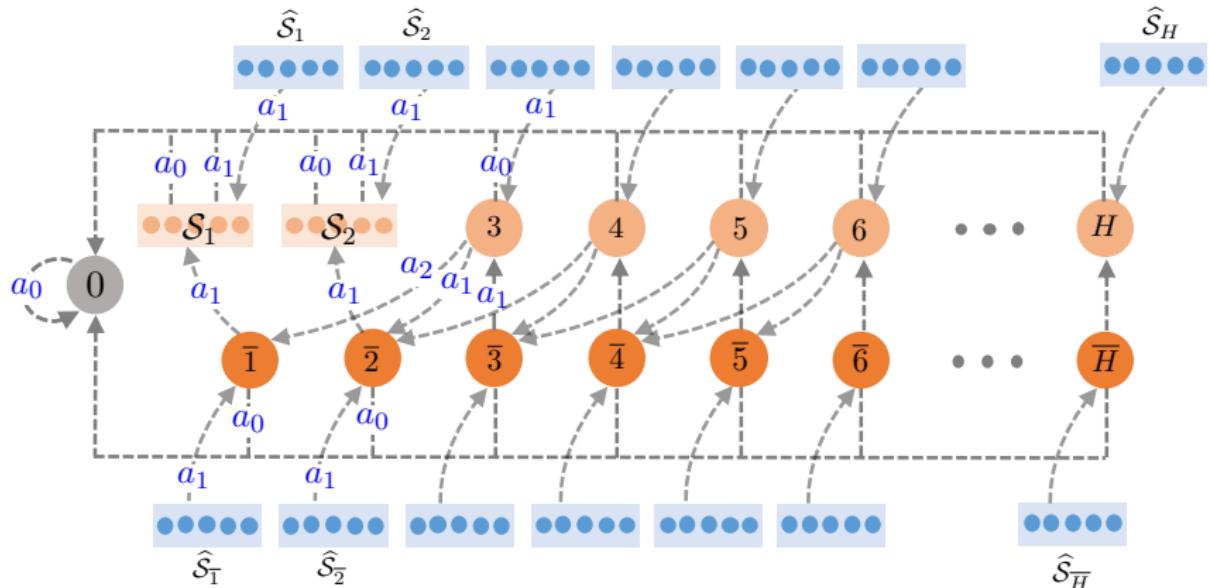
MDP construction for our lower bound



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

- augmented chain-like structure
- $V^{(t)}(s)$ relies on $V^{(t)}(s-1), V^{(t)}(s-2), \dots$ (delayed impacts)

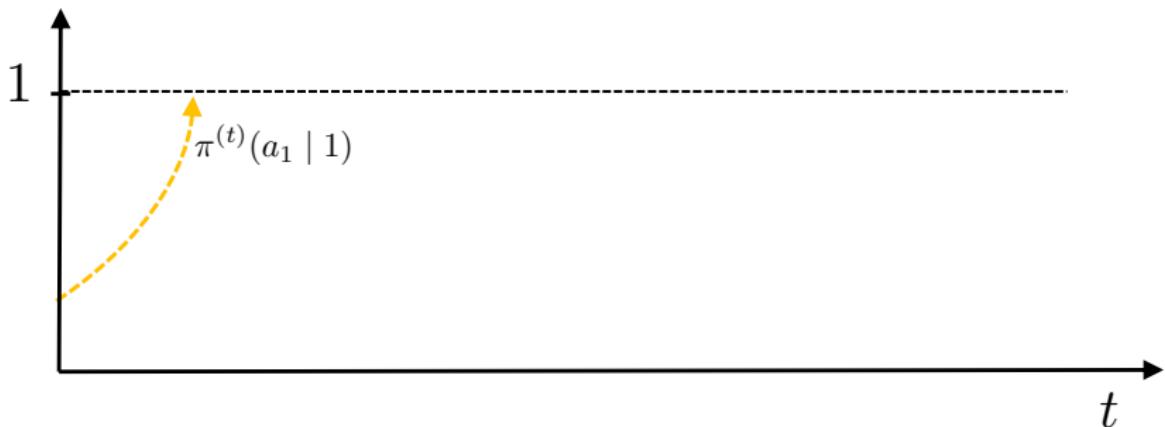
MDP construction for our lower bound



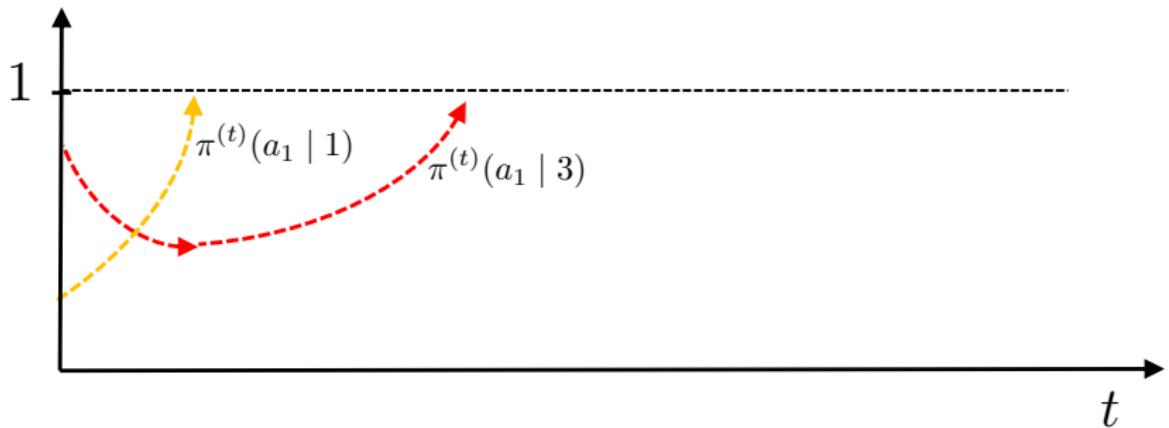
Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$

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

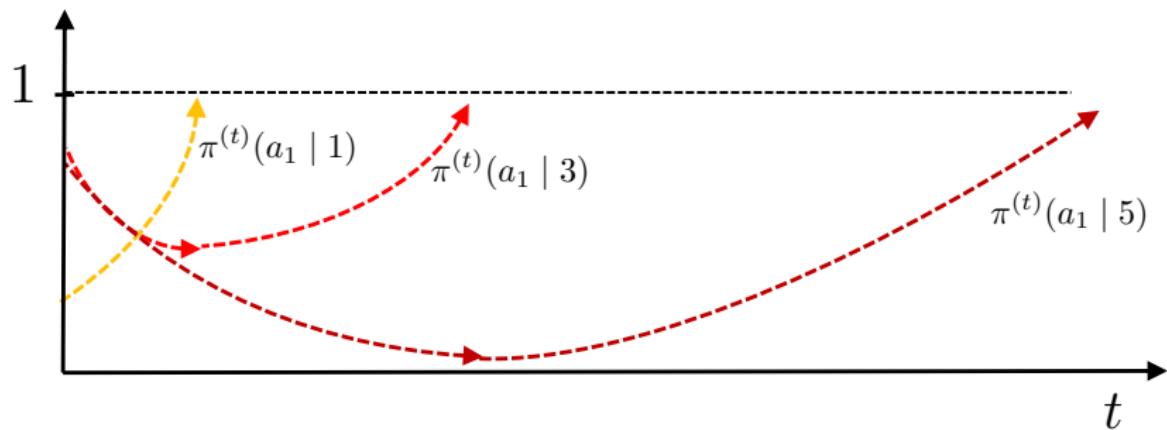
What is happening in our constructed MDP?



What is happening in our constructed MDP?

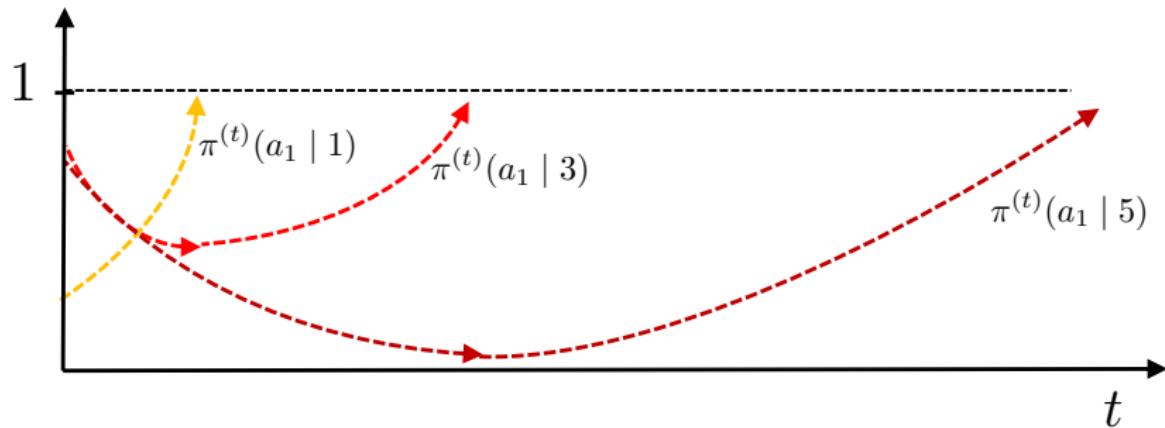


What is happening in our constructed MDP?



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 $s \uparrow$

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

Booster 1: entropy regularization

accelerate convergence by regularizing value function

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

Booster 1: entropy regularization

accelerate convergence by regularizing value function

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

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

Booster 1: entropy regularization

accelerate convergence by regularizing value function

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

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

entropy-regularized value maximization

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

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

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

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

- 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

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

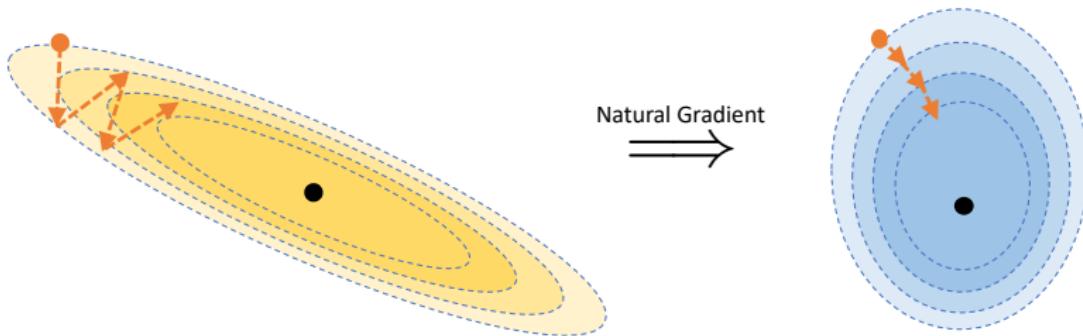
to achieve $\|V^{(t)} - V^*\|_\infty \leq \varepsilon$

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

$$c(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \dots) O(\frac{1}{\varepsilon}) \text{ iterations}$$

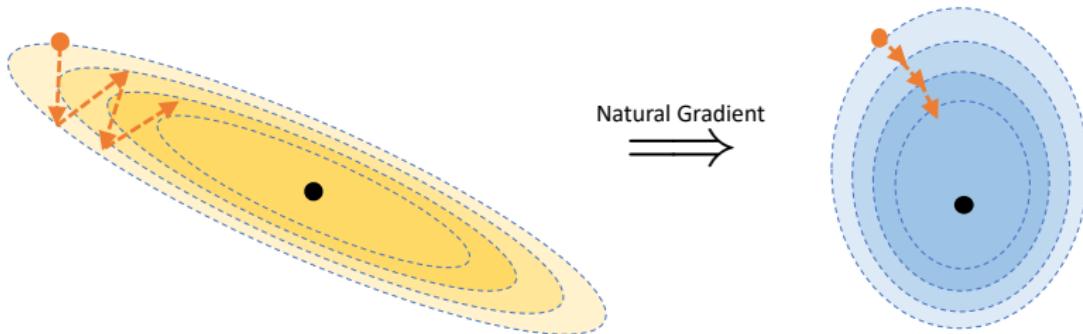
Booster 2: natural policy gradient (NPG)

precondition gradients to improve search directions ...



Booster 2: natural policy gradient (NPG)

precondition gradients to improve search directions ...



NPG method (Kakade '02)

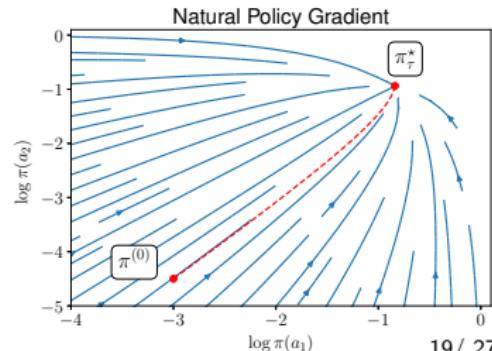
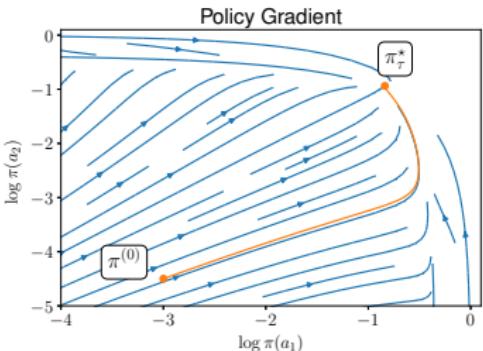
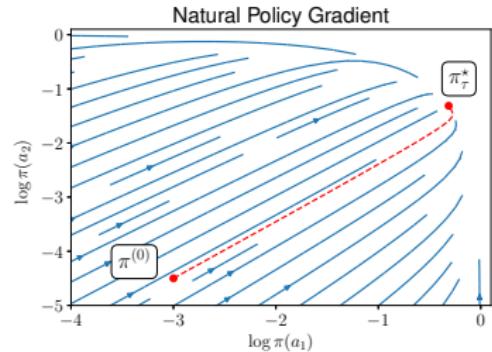
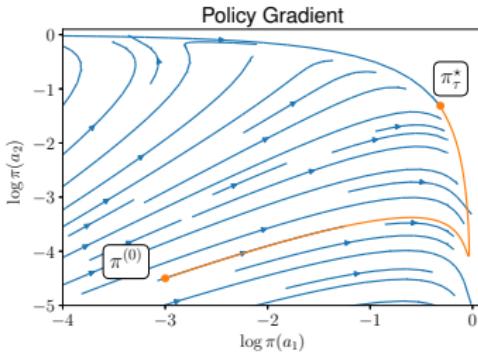
$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V_\tau^{(t)}(\rho), \quad t = 0, 1, \dots$$

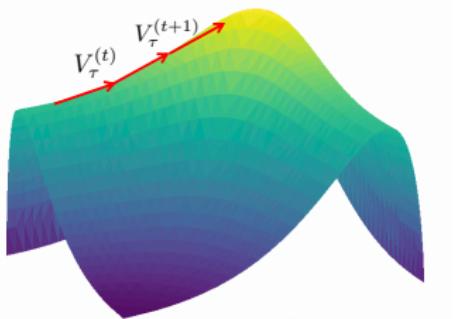
- $\mathcal{F}_\rho^\theta := \mathbb{E} \left[(\nabla_\theta \log \pi_\theta(a | s)) (\nabla_\theta \log \pi_\theta(a | s))^\top \right]$: Fisher info

Entropy-regularized natural gradient helps!

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

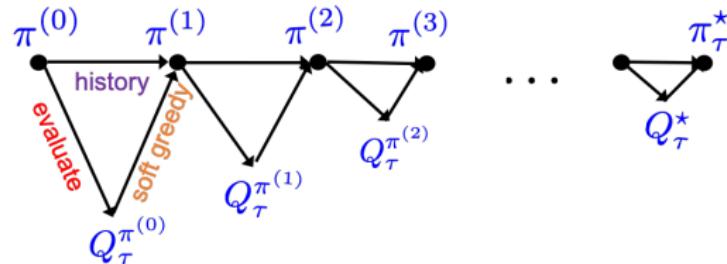
increase regularization





*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)

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

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

- invariant to the choice of initial state distribution ρ

Linear convergence with exact gradients

optimal policy: π_τ^* ; *optimal “soft” Q function:* $Q_\tau^* := Q_\tau^{\pi_\tau^*}$

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

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

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

Implications: iteration complexity

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

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

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

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

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

Implications: iteration complexity

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

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

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

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

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

Nearly dimension-free global linear convergence!

Entropy-regularized NPG with inexact gradients

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

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

e.g. using sample-based estimators

Entropy-regularized NPG with inexact gradients

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

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

e.g. using sample-based estimators

Inexact entropy-regularized NPG:

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

Entropy-regularized NPG with inexact gradients

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

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

e.g. using sample-based estimators

Inexact entropy-regularized NPG:

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

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

Linear convergence with inexact gradients

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

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

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

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

- $C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + 2\tau\left(1 - \frac{\eta\tau}{1 - \gamma}\right)\|\log \pi_\tau^\star - \log \pi^{(0)}\|_\infty$
- $C_2 = \frac{2\gamma\left(1 + \frac{\gamma}{\eta\tau}\right)}{1 - \gamma} \delta$: error floor
- converges linearly at the same rate until an error floor is hit

Returning to the original MDP?

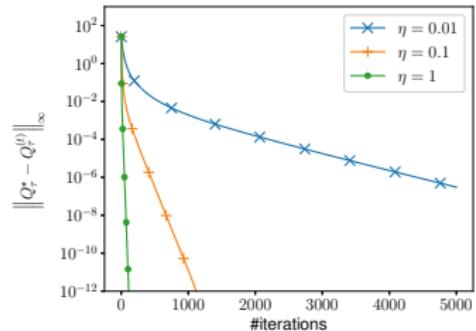
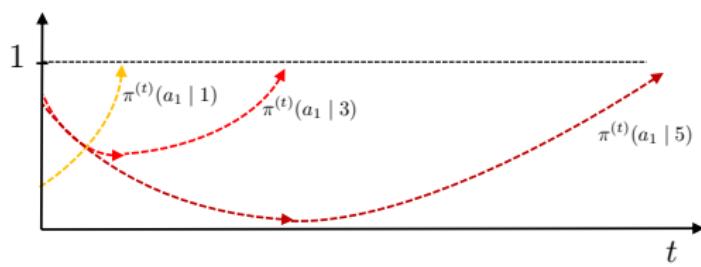
How to employ entropy-regularized NPG to find an ε -optimal policy for the original (unregularized) MDP?

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)

Concluding remarks



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

"Fast global convergence of natural policy gradient methods with entropy regularization,"
S. Cen, C. Cheng, Y. Chen, Y. Wei, Y. Chi, accepted to Operations Research, 2021+

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