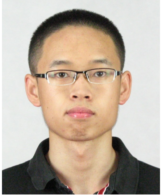
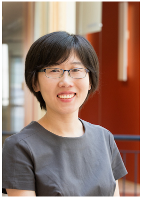
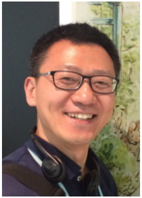
Softmax policy gradient methods can take exponential time to converge







Gen Li
Tsinghua

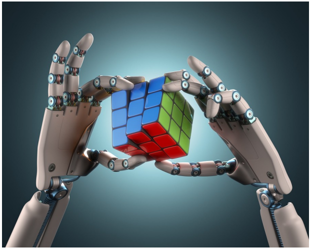
Yuting Wei
Wharton

Yuejie Chi
CMU

Yuantao Gu
Tsinghua

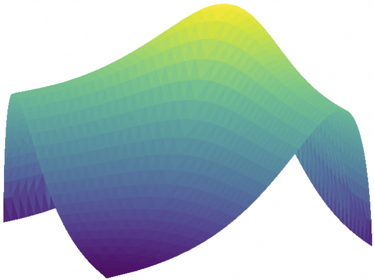
Yuxin Chen
Princeton

Recent successes in RL



Policy optimization: a major contributor to these successes

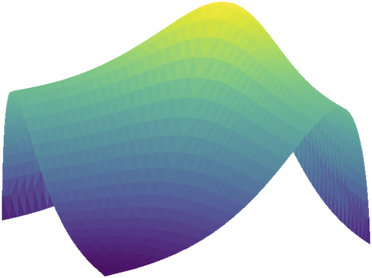
Challenges: large dimensionality and non-concavity



Recent advances towards understanding policy optimization

- tabular 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 et al '21, Zhan et al '21, Cen et al '21, . . .)
- control (Fazel et al., 2018; Bhandari and Russo, 2019, . . .)

Challenges: large dimensionality and non-concavity



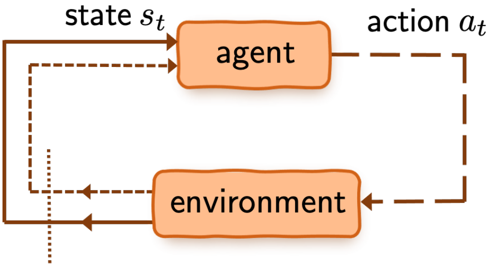
Recent advances towards understanding policy optimization

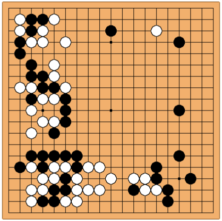
- tabular MDPs (Agarwal et al 19, Bhandari and Russo '19, Shani et al '19, Mei et al '20, Cen et al '20, Zhang et al '20, Lan et al '21, Zhan et al '21, Cen et al '21, ...)
- control (Fazel et al., 2018; Bhandari and Russo, 2019, ...)

This talk: a (super)-exponential lower bound on a popular variant of policy gradient methods

Backgrounds: policy optimization for tabular MDPs

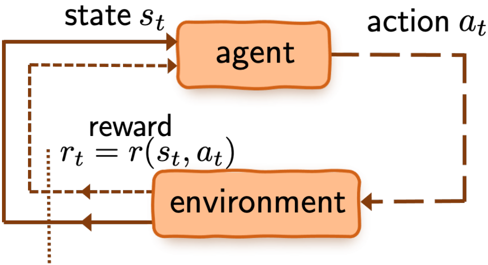
Markov decision process (MDP)

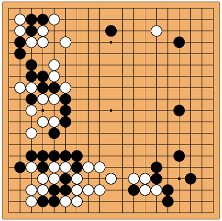




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

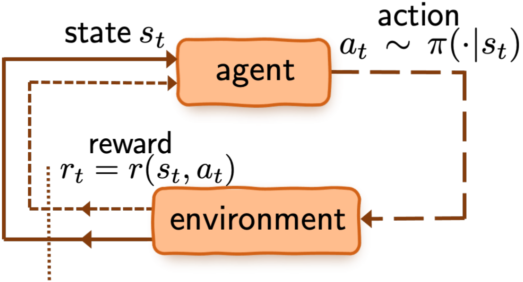
Markov decision process (MDP)

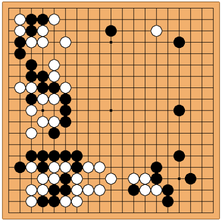




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

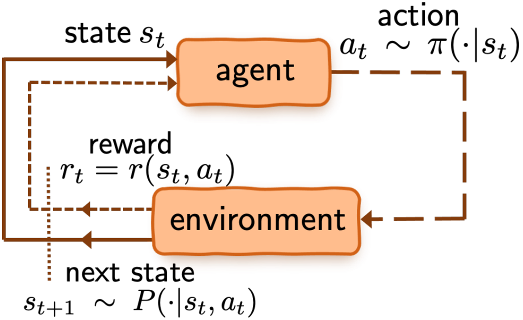
Markov decision process (MDP)

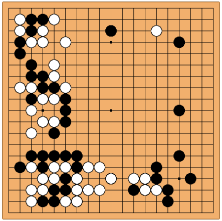




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

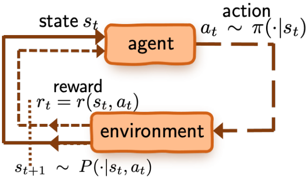
Markov decision process (MDP)

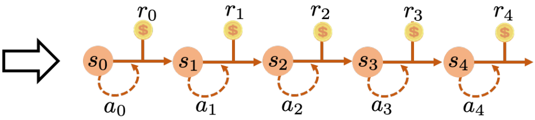




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)
- $P(\cdot|s, a)$: transition probabilities

Value function of policy π

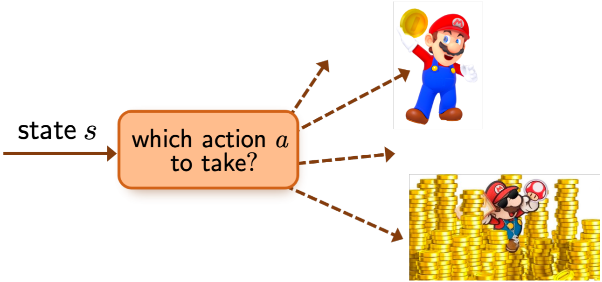




$$\forall s \in \mathcal{S}: \qquad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right]$$

- 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



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

Policy optimization

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

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

Policy optimization

Given state distribution $s \sim \rho$

(e.g. uniform)

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

parameterize $\Longrightarrow$

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

Softmax policy gradient (PG) methods

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

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

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

softmax parameterization

Softmax policy gradient (PG) methods

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

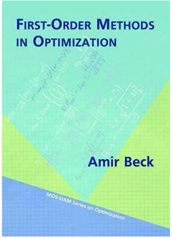
$$\max_{\pi} \; V^{\pi}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

parameterize $\Longrightarrow$

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

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

softmax parameterization



Softmax policy gradient (PG) methods

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

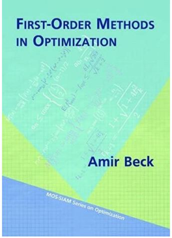
$$\max_{\pi} \; V^{\pi}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\pi}(s)\right]$$

parameterize $\Longrightarrow$

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

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

softmax parameterization

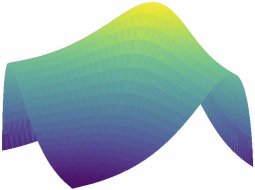


Policy gradient method (Sutton et al. '00)

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

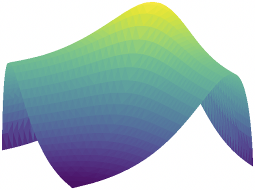
- η: learning rate

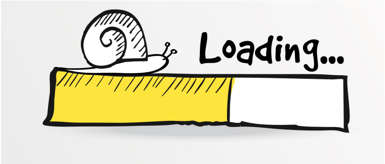
Does policy gradient (PG) method converge?



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

Does policy gradient (PG) method converge?

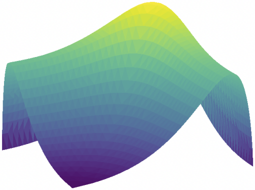


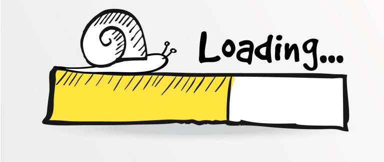


- (Agarwal et al. '19) Softmax PG converges to global opt as $t \to \infty$

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

Does policy gradient (PG) method converge?



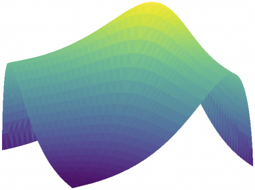


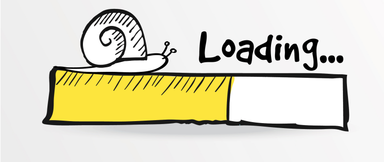
- (Agarwal et al. '19) Softmax PG converges to global opt as $t \to \infty$
- (Mei et al. '20) Softmax PG converges to global opt in

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

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

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

A negative message

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

Suppose the learning rate obeys $0 < \eta < (1-\gamma)^2/5$. *There exists an MDP with* $|\mathcal{S}|$ *states and 3 actions s.t. it takes softmax PG at least*

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

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

A negative message

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

Suppose the learning rate obeys $0 < \eta < (1-\gamma)^2/5$. *There exists an MDP with* $|\mathcal{S}|$ *states and 3 actions s.t. it takes softmax PG at least*

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

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

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

A negative message

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

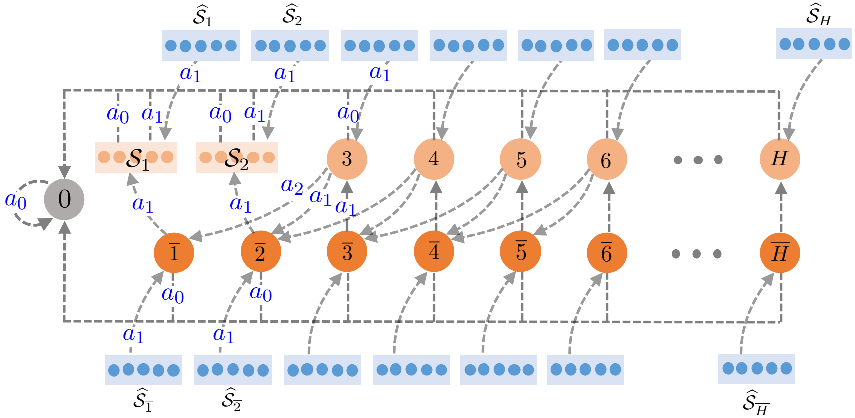
Suppose the learning rate obeys $0 < \eta < (1-\gamma)^2/5$. *There exists an MDP with* $|\mathcal{S}|$ *states and 3 actions s.t. it takes softmax PG at least*

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

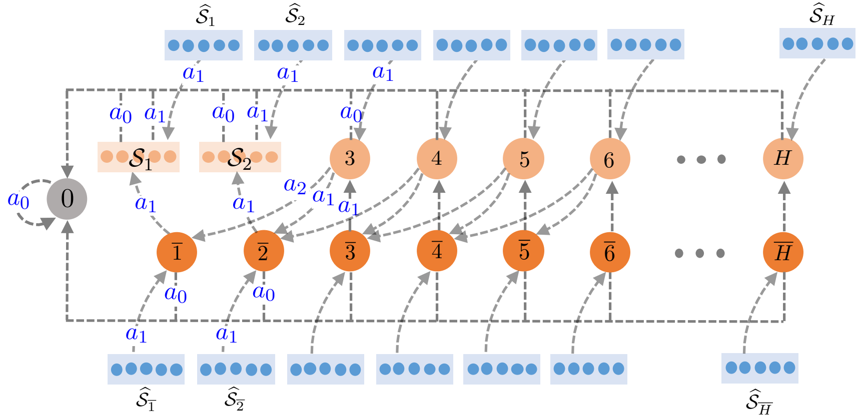
to achieve $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} |V^{(t)}(s) - V^{\star}(s)| \le 0.07$.

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

MDP construction for our lower bound



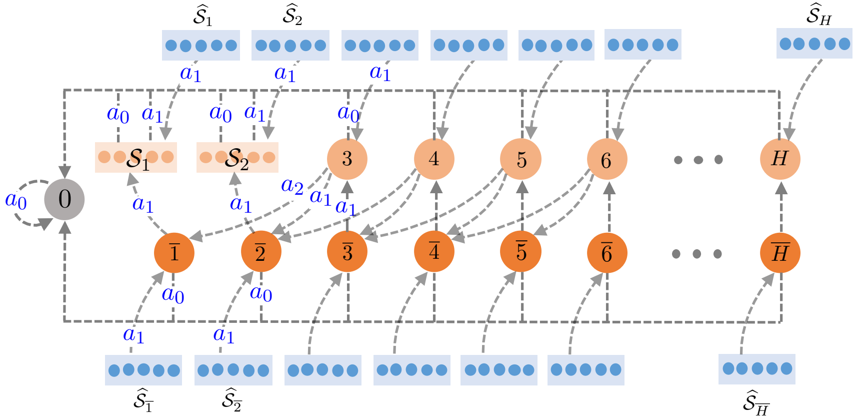
MDP construction for our lower bound



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

- augmented chain-like structure

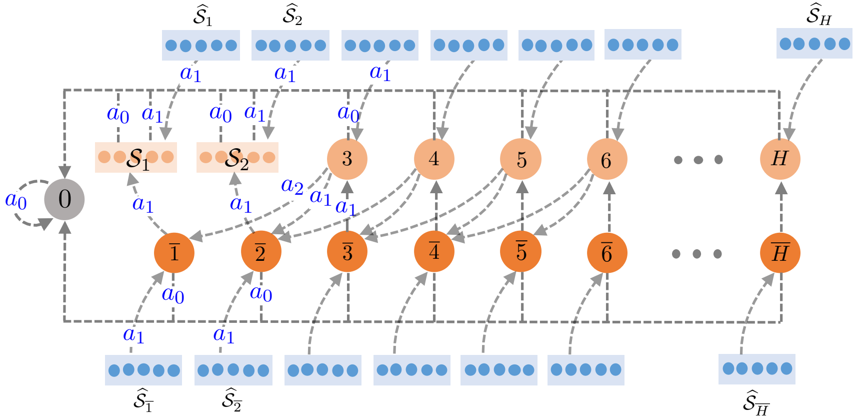
MDP construction for our lower bound



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

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

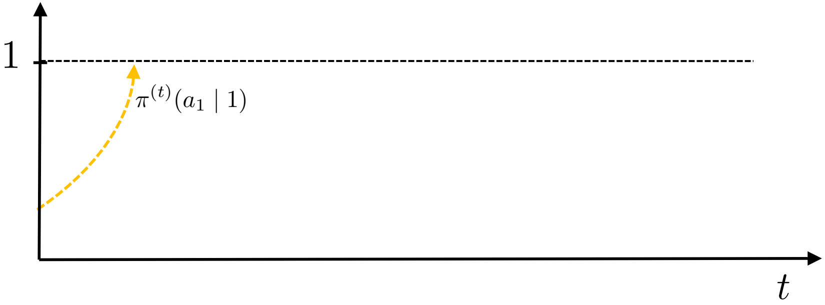
MDP construction for our lower bound



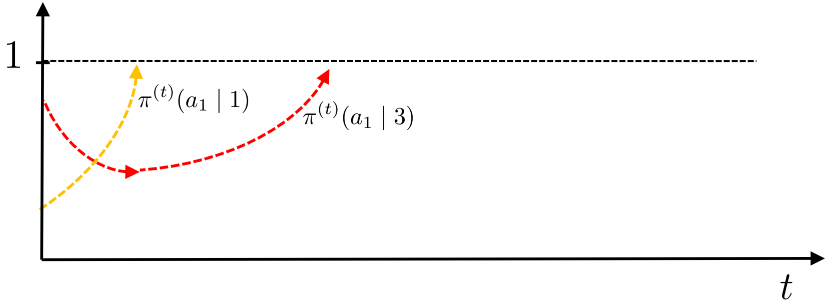
Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$

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

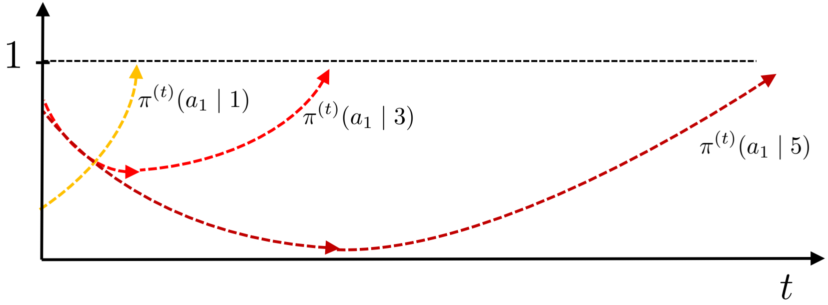
What is happening in our constructed MDP?



What is happening in our constructed MDP?

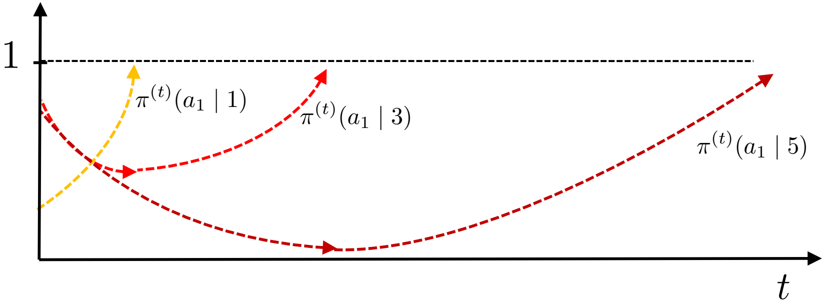


What is happening in our constructed MDP?



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 $s \uparrow$

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

Concluding remarks

- Popular policy optimization algorithms like softmax PG can be extremely slow for large-dimensional & long-horizon problems

Concluding remarks

- Popular policy optimization algorithms like softmax PG can be extremely slow for large-dimensional & long-horizon problems

- Proper modifications of update rules are needed for acceleration
  - e.g. KL-type regularization (Agarwal et al '19)
    — iteration complexity: $\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\varepsilon^2}$

Concluding remarks

- Popular policy optimization algorithms like softmax PG can be extremely slow for large-dimensional & long-horizon problems

- Proper modifications of update rules are needed for acceleration
  - e.g. KL-type regularization (Agarwal et al '19)
    — iteration complexity: $\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\varepsilon^2}$
  - e.g. preconditioning / natural policy gradient (Agarwal et al '19)
    — iteration complexity: $\frac{1}{(1-\gamma)^2\varepsilon}$

Concluding remarks

- Popular policy optimization algorithms like softmax PG can be extremely slow for large-dimensional & long-horizon problems

- Proper modifications of update rules are needed for acceleration
  - e.g. KL-type regularization (Agarwal et al '19)
    — iteration complexity: $\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\varepsilon^2}$
  - e.g. preconditioning / natural policy gradient (Agarwal et al '19)
    — iteration complexity: $\frac{1}{(1-\gamma)^2\varepsilon}$
  - e.g. NPG with entropy regularization (Cen et al '20)
    — iteration complexity: $\frac{1}{1-\gamma}\log\frac{1}{\varepsilon}$

Concluding remarks

- Popular policy optimization algorithms like softmax PG can be extremely slow for large-dimensional & long-horizon problems

- Proper modifications of update rules are needed for acceleration
  - e.g. KL-type regularization (Agarwal et al '19)
    — iteration complexity: $\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\varepsilon^2}$
  - e.g. preconditioning / natural policy gradient (Agarwal et al '19)
    — iteration complexity: $\frac{1}{(1-\gamma)^2\varepsilon}$
  - e.g. NPG with entropy regularization (Cen et al '20)
    — iteration complexity: $\frac{1}{1-\gamma}\log\frac{1}{\varepsilon}$

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