Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization

Yuejie Chi

Carnegie Mellon University

Bath Symposium on the Mathematics of Machine Learning
August 2020

Shicong Cen
CMU ECE

Chen Cheng
Stanford Stat

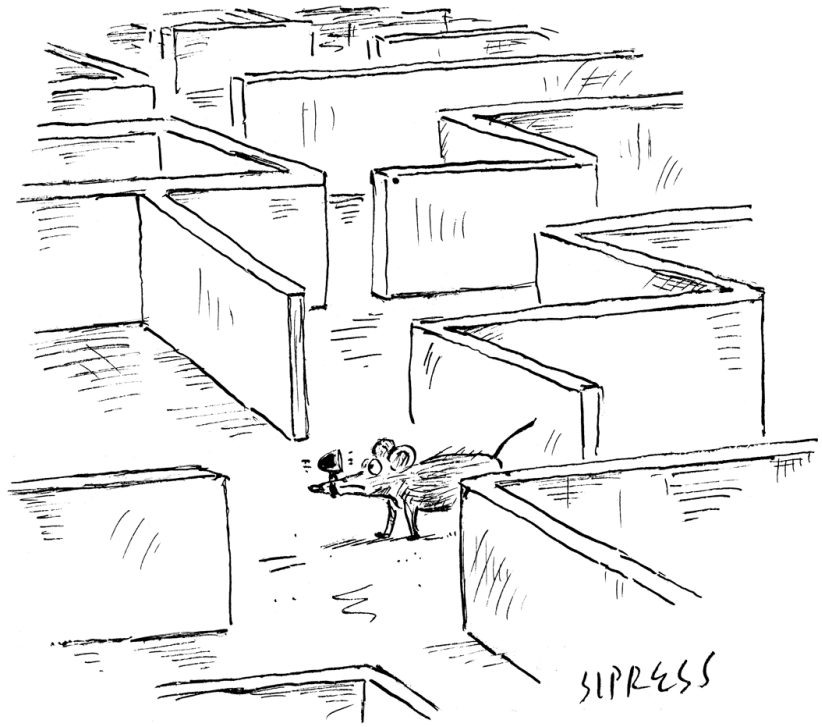
Yuxin Chen
Princeton EE

Yuting Wei
CMU Stat

Reinforcement learning (RL)

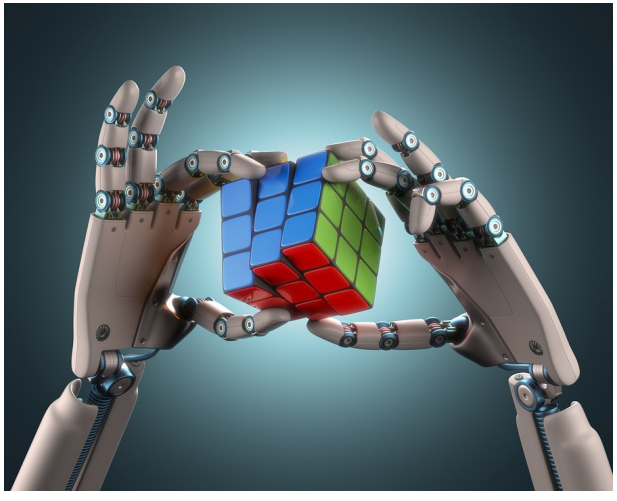
In RL, an agent learns by interacting with an environment.

- unknown environments
- non-stationarity
- delayed feedback or rewards
- trial-and-error
- sequential and online



"Recalculating ... recalculating ..."

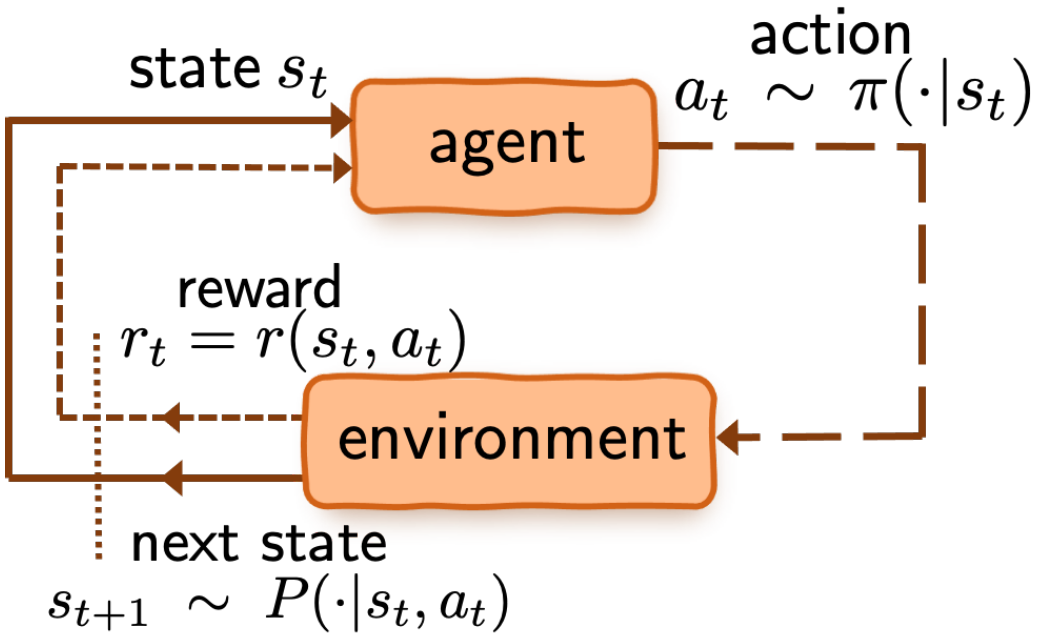
Recent successes in RL

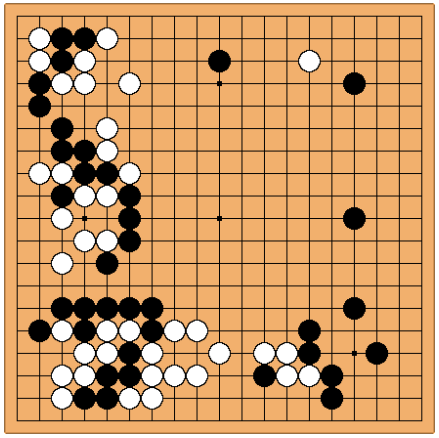


Policy optimization is a major driver to these successes.

Backgrounds: policy optimization for MDPs

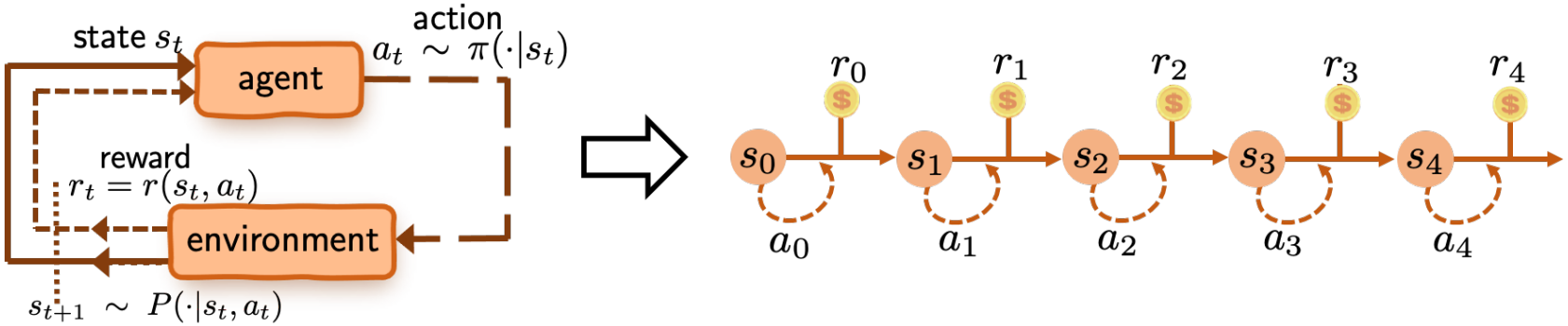
Markov decision process (MDP)





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

Value function and Q-function



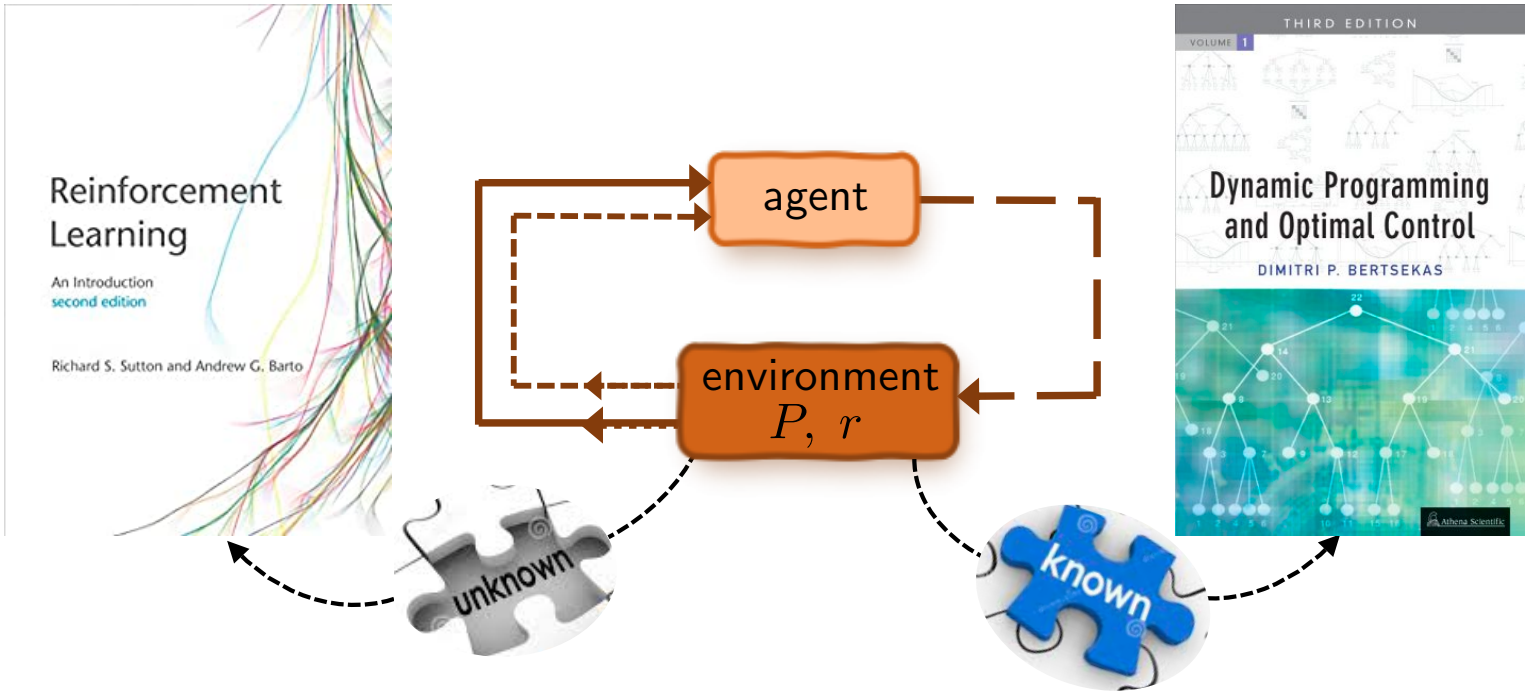
Value function and **state-action (Q) function** of policy π:

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

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

- Long-term *discounted* reward: $\gamma \in [0, 1)$ is the discount factor
- Expectation is w.r.t. the sampled trajectory under π

Searching for the optimal policy



Goal: find the optimal policy $\pi^\star$ that maximizes $V^\pi(s)$

- optimal value / Q function: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$

Policy gradient methods

Given an initial state distribution $s \sim \rho$, find policy π such that

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

softmax parameterization:

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

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

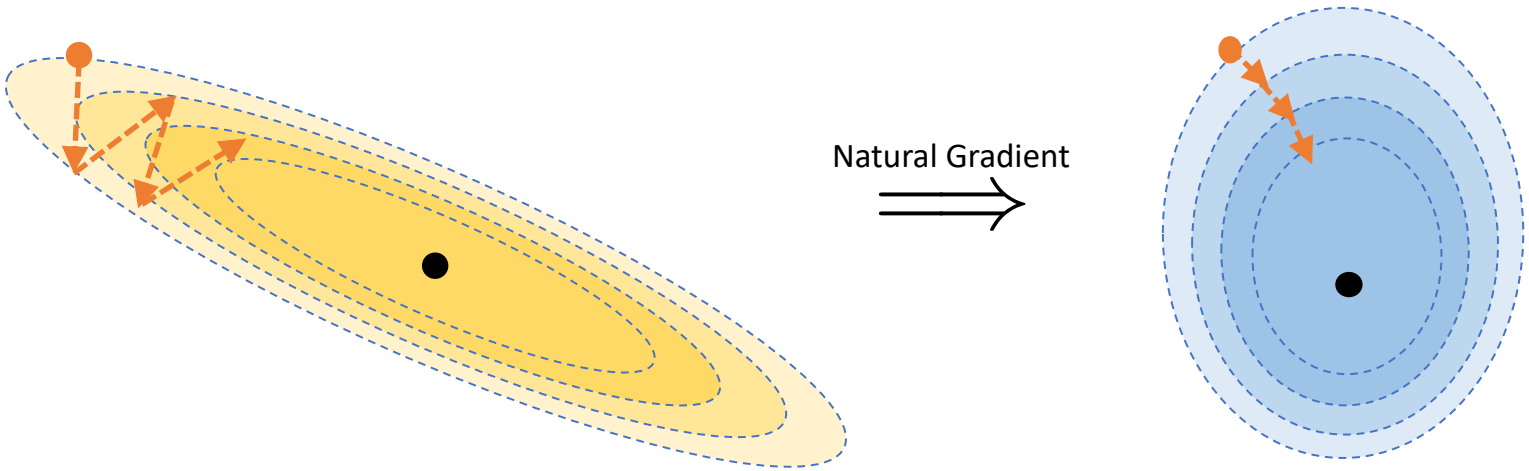
Policy gradient method (Sutton et al., 2000)

For $t = 0, 1, \cdots$

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

where η *is the learning rate.*

Booster #1: natural policy gradient



Natural policy gradient method (Kakade, 2002)

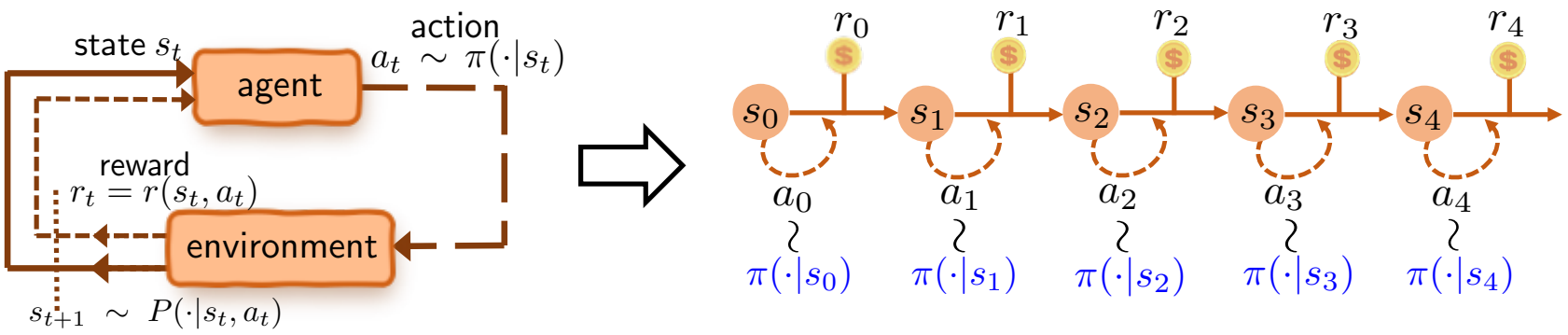
For $t = 0, 1, \cdots$

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

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 #2: entropy regularization



To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function:

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

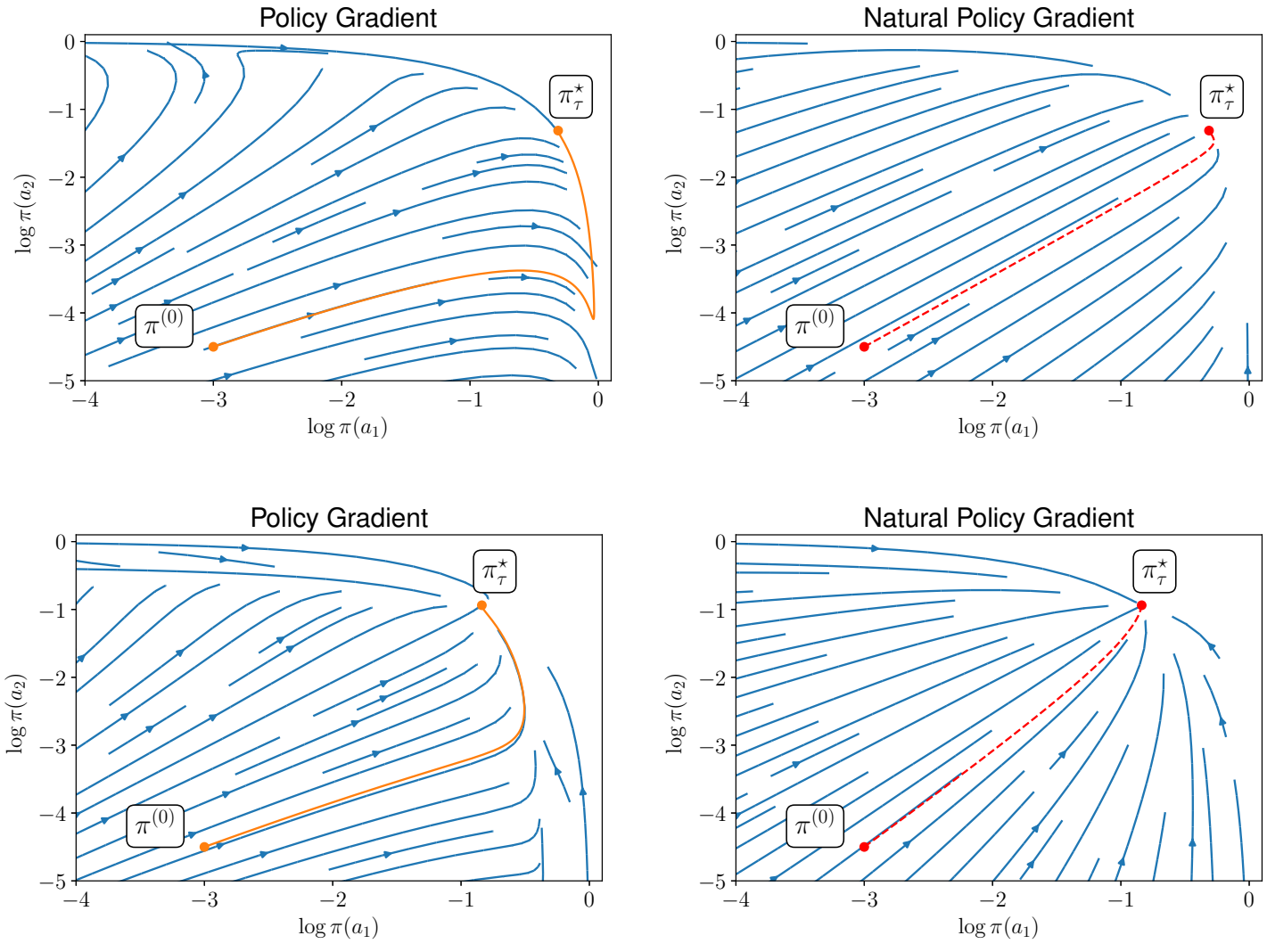
where τ is the entropy regularization parameter.

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

Entropy-regularized natural gradient helps!

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

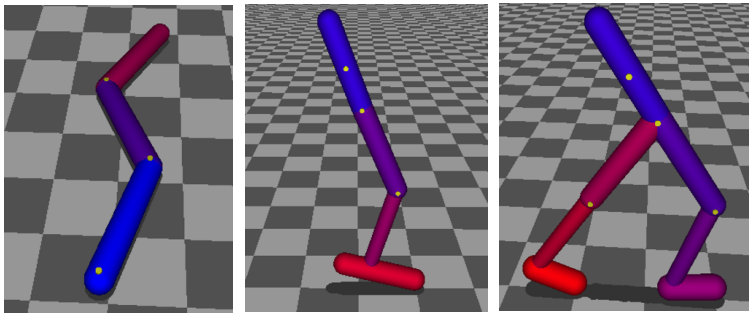
increase regularization



Unreasonable effectiveness in practice

Advantages of policy gradient methods:

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

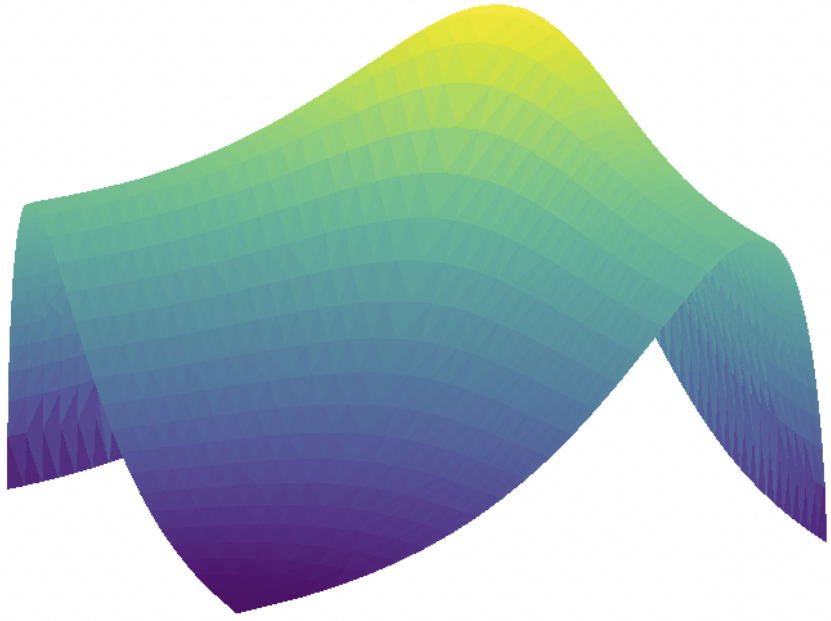


TRPO = NPG + line search
(Schulman et al., 2015)

We also found that adding the entropy of the policy π to the objective function improved exploration by discouraging premature convergence to suboptimal deterministic policies. This technique was originally proposed by (Williams & Peng, 1991), who found that it was particularly helpful on tasks requiring hierarchical behavior. The gradi-

A3C (Mnih et al., 2016)
SAC (Haarnoja et al., 2018)

Theoretical challenges: non-concavity



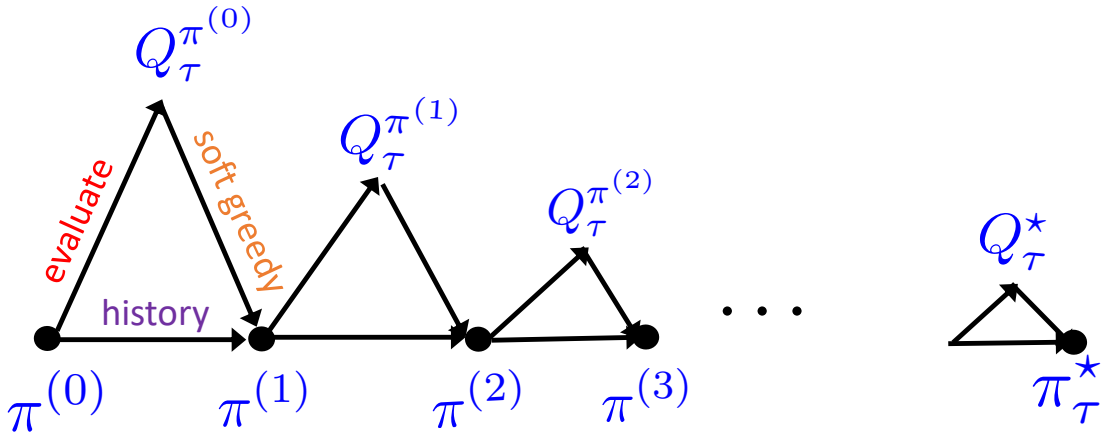
Recent breakthroughs on understanding global convergence of

- policy gradient methods for control (Fazel et al., 2018; Bhandari and Russo, 2019);
- (un)regularized policy gradients for tabular MDPs (Agarwal et al., 2019, Bhandari and Russo, 2019; Mei et al. 2020);
- unregularized NPG for tabular MDPs (Agarwal et al., 2019);

and many others.

This talk: understanding 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

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

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

- invariant with the choice of ρ
- optimal policy: $\pi_\tau^\star$
- optimal "soft" value / Q function: $V_\tau^\star := V_\tau^{\pi_\tau^\star}$, $Q_\tau^\star := Q_\tau^{\pi_\tau^\star}$

Linear convergence with exact gradient

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

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

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

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

$$\forall t \ge 0: \quad \|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le C_1 \gamma \, (1-\eta\tau)^t$$

- **Linear convergence of log policies:**

$$\forall t \ge 0: \quad \|\log \pi_\tau^\star - \log \pi^{(t+1)}\|_\infty \le 2C_1 \tau^{-1} (1-\eta\tau)^t$$

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

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

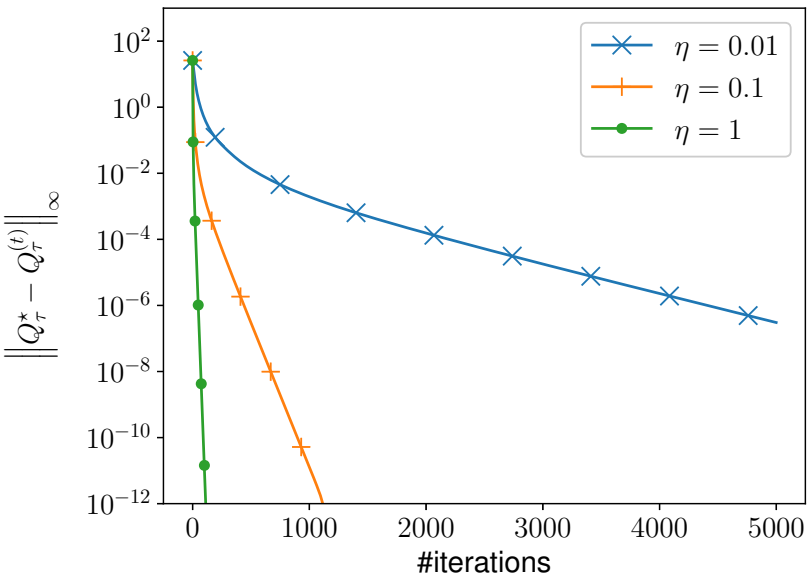
$$\frac{1}{1-\gamma} \log\left(\frac{\|Q_\tau^\star - Q_\tau^{(0)}\|_\infty \gamma}{\epsilon}\right)$$

Global linear convergence of entropy-regularized NPG
at a **dimension-free** rate!

Comparison with unregularized NPG

Regularized NPG

$\tau = 0.001$

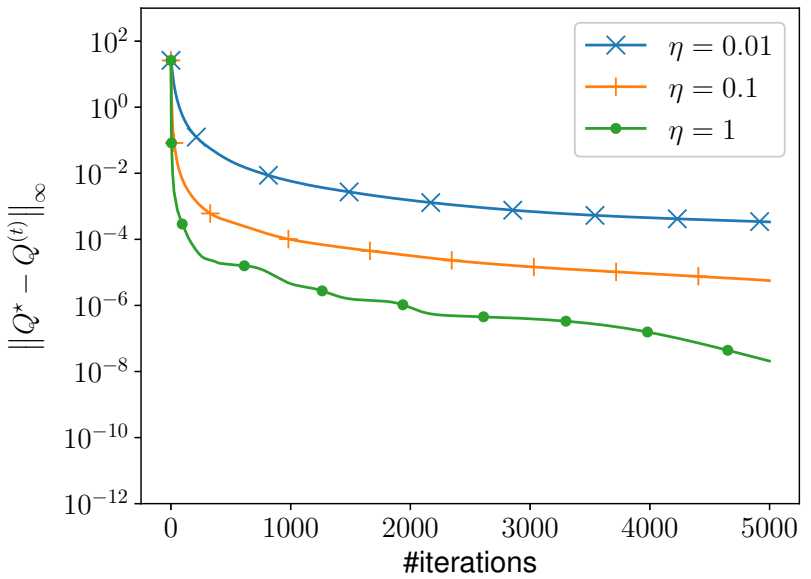


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

Ours

Vanilla NPG

$\tau = 0$



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

(Agarwal et al. 2019)

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

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

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

Inexact entropy-regularized NPG:

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

Question: Robustness of entropy-regularized NPG?

Linear convergence with inexact gradients

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

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

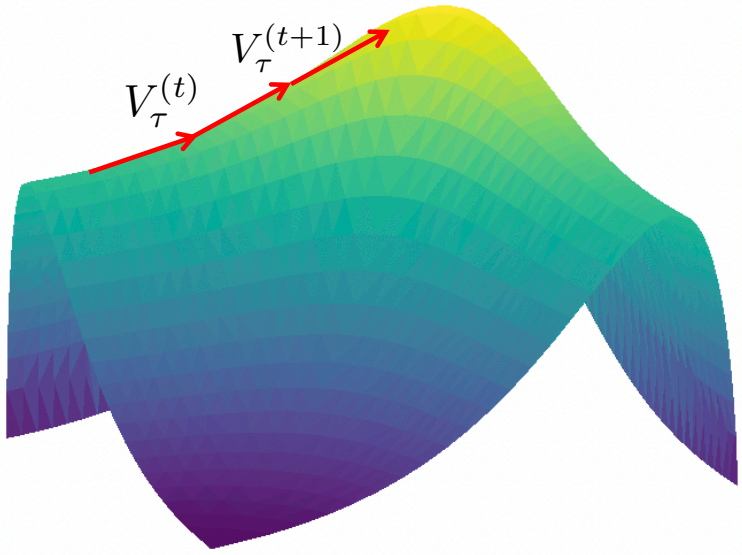
$$\delta \le \frac{(1-\gamma)^2 \epsilon}{2\gamma\Big[1+\gamma\Big(\underbrace{\frac{1-\gamma}{\eta\tau}}_{\text{impact of learning rate}} - 1\Big)\Big]}.$$

- The tolerence level δ is maximized when $\eta = \frac{1-\gamma}{\tau}$:

$$\delta \le \frac{(1-\gamma)^2 \epsilon}{2\gamma}.$$

A glimpse of the analysis:
a soft version of Bellman's equation

A key lemma: monotonic performance improvement



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

discounted state visitation distribution

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

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

Recall: Bellman's optimality principle

Bellman operator

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

- 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

Soft Bellman operator

Soft Bellman operator

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

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

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

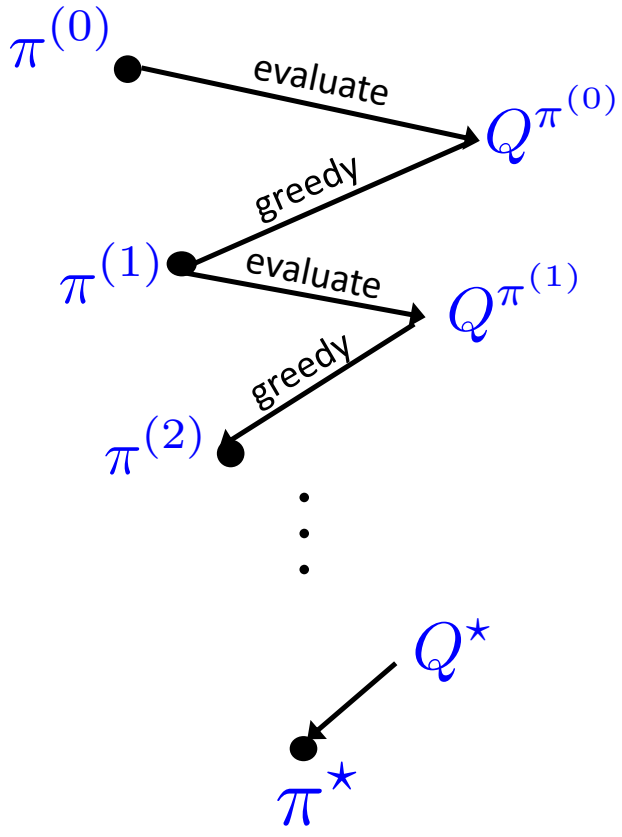
γ-contraction of soft Bellman operator:

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

Richard Bellman

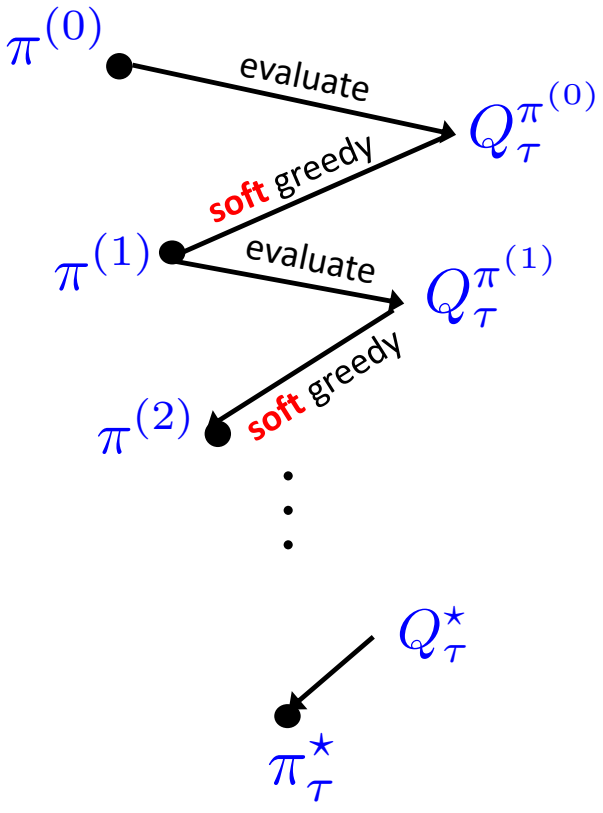
Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

Policy iteration



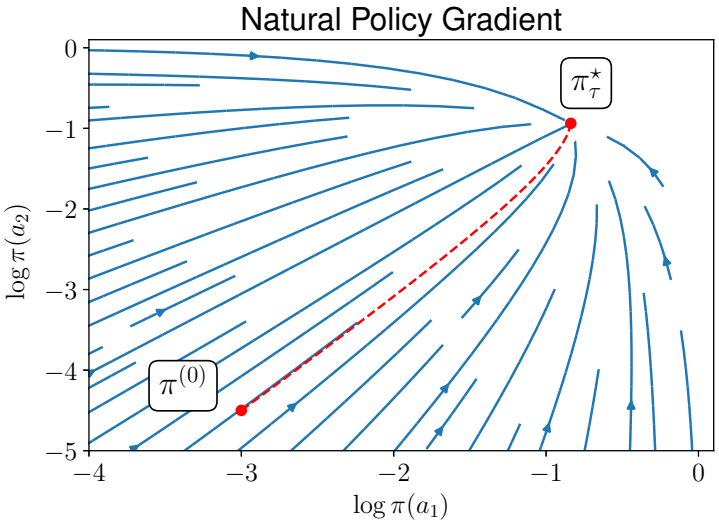
Bellman operator

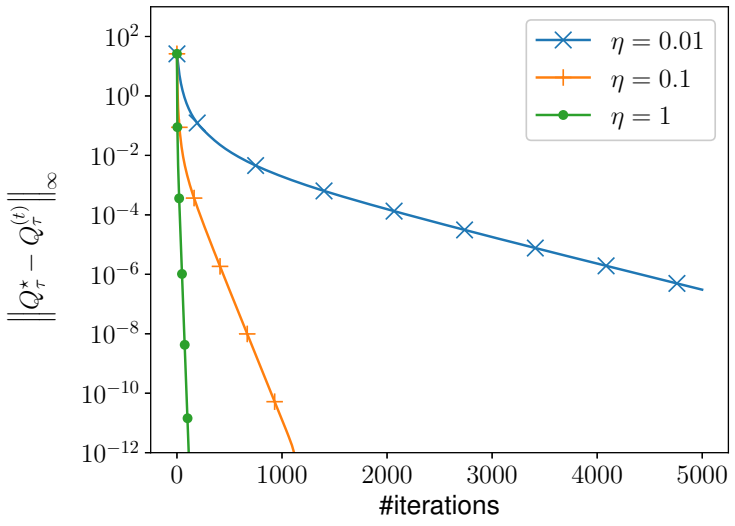
Soft policy iteration



Soft Bellman operator

Concluding remarks





Global linear convergence of entropy-regularized NPG for tabular discounted MDP

Future directions:

- function approximation
- sample complexities
- analysis of soft actor-critic algorithms

Thanks!

Our research is supported by NSF (to Chen, Wei, Chi), ONR (to Chen, Chi), ARO (to Chen, Chi), and AFOSR (to Chen).

arXiv:2007.06558