Policy Optimization in Reinforcement Learning: A Tale of Preconditioning and Regularization

Yuejie Chi

Carnegie Mellon University

Instituto Superior Técnico, June 2021

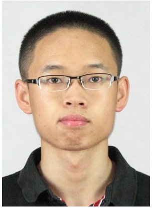
My wonderful collaborators

Shicong Cen
CMU

Chen Cheng
Stanford

Yuxin Chen
Princeton

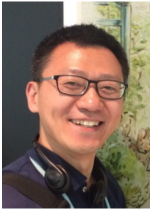
Yuting Wei
CMU



Gen Li
Princeton

Wenhao Zhan
Princeton

Jason Lee
Princeton

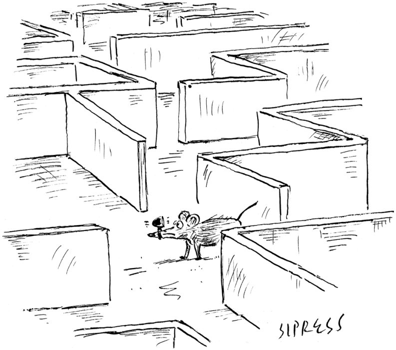


Yuantao Gu
Tsinghua

Reinforcement learning (RL)

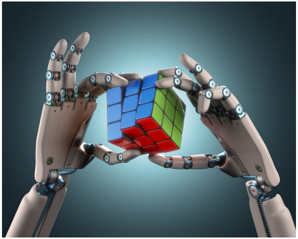
In RL, an agent learns by interacting with an environment.

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



"Recalculating ... recalculating ..."

Recent successes in RL

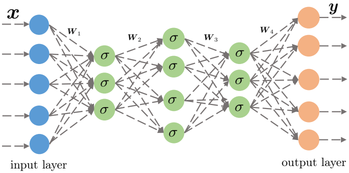


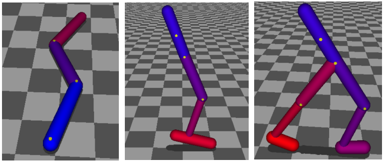
Policy optimization is a major driver to these successes.

Policy optimization in practice

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

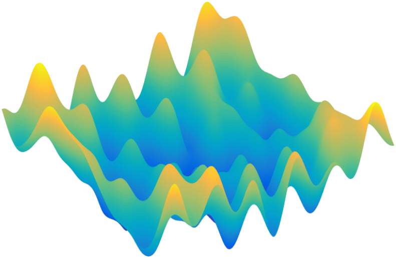
- directly optimize the policy, which is the quantity of interest;
- allow flexible differentiable parameterizations of the policy;
- work with both continuous and discrete problems.





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 more.

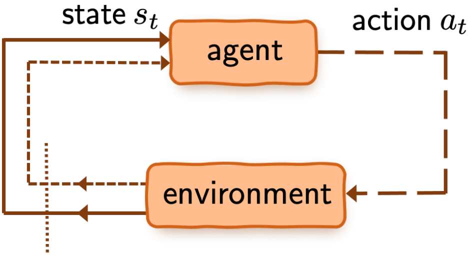


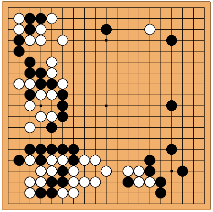
Our goal:

- understand finite-time convergence rates of popular heuristics;
- design fast-convergent algorithms that scale for finding policies with desirable properties.

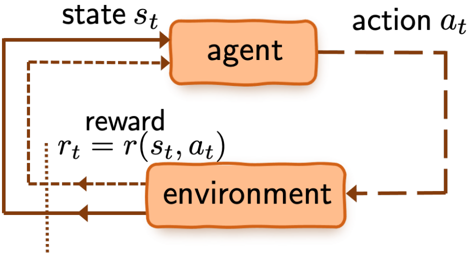
Backgrounds: policy optimization in tabular Markov decision processes

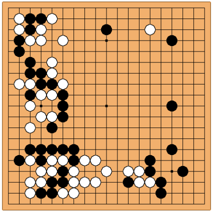
Markov decision process (MDP)





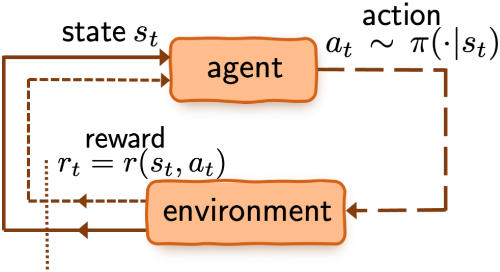
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

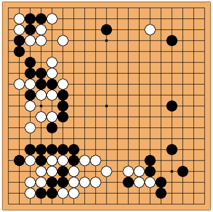




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

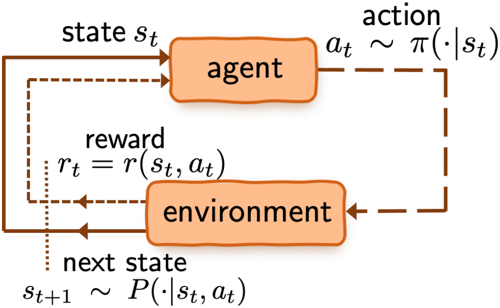
Markov decision process (MDP)

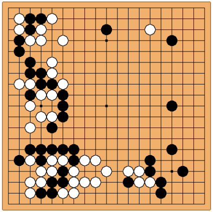




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

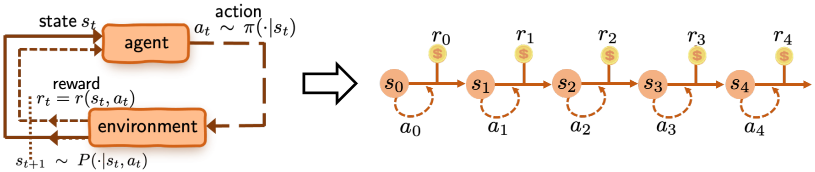
Markov decision process (MDP)





- $\mathcal{S}$: 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

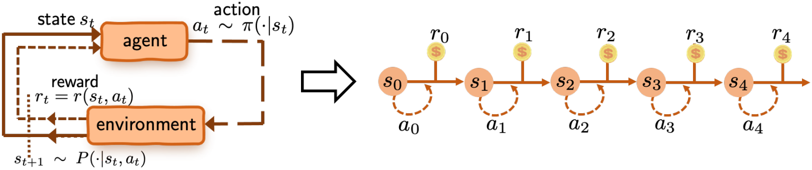


Value function and **Q function** of policy π:

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

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

Value function and Q-function



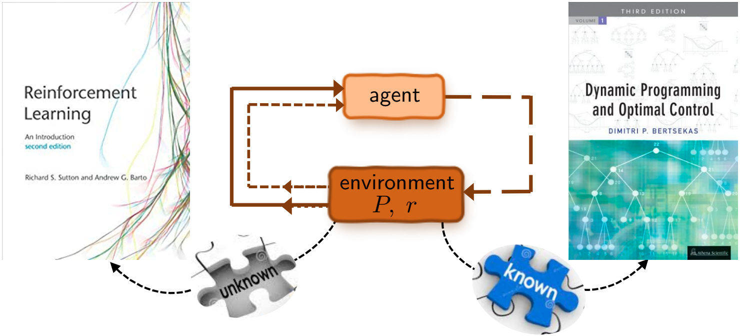
Value function and **Q function** of policy π:

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

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

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

Searching for the optimal policy



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

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

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

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

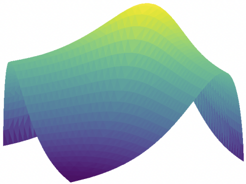
Policy gradient method (Sutton et al., 2000)

For $t = 0, 1, \cdots$

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

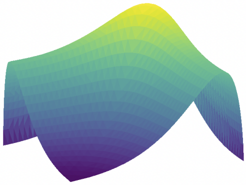
where η *is the learning rate.*

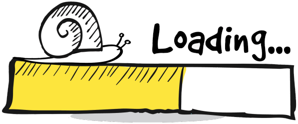
Global convergence of the PG method?



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

Global convergence of the PG method?

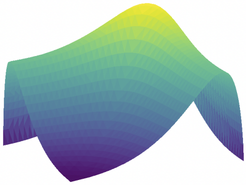


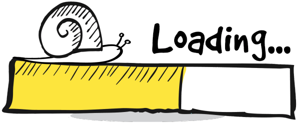


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

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

Global convergence of the PG method?

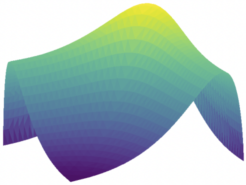


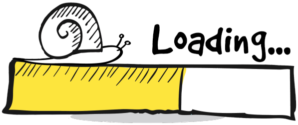


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

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

Global convergence of the PG method?



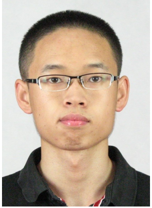


- (Agarwal et al., 2019) showed that softmax PG converges asymptotically to the global optimal policy.
- (Mei et al., 2020) Softmax PG converges to global opt in

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

Is the rate of PG good, bad or ugly?

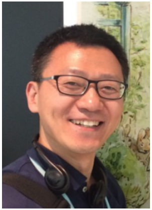
Softmax PG can take exponential time to converge



Gen Li
Princeton

Yuting Wei
CMU

Yuxin Chen
Princeton



Yuantao Gu
Tsinghua

A negative message

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

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

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

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

A negative message

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

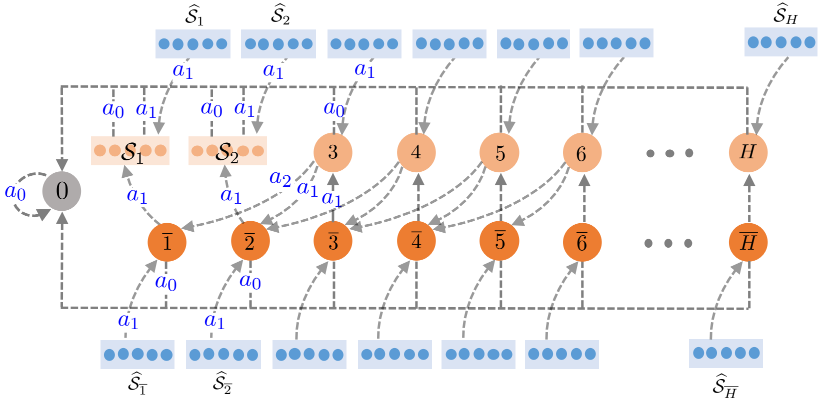
There exists an MDP s.t. it takes softmax PG at least

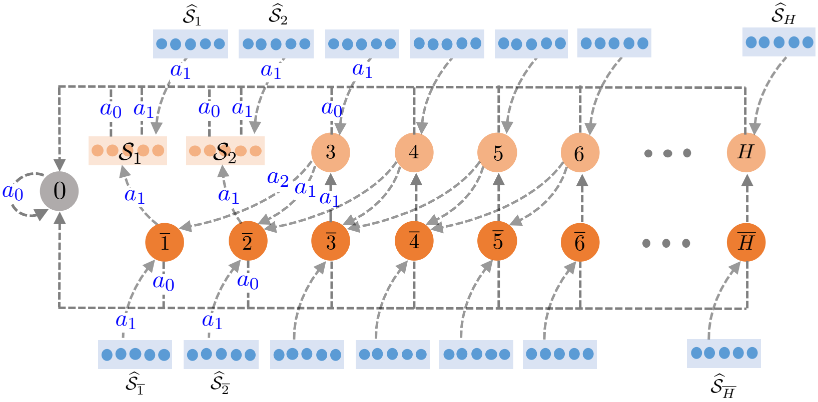
$$\frac{1}{\eta}\,|\mathcal{S}|^{2^{\Theta(\frac{1}{1-\gamma})}} \ \textit{iterations}$$

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)!
- Also hold for average sub-opt gap $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\big[V^{(t)}(s) - V^{\star}(s)\big]$.

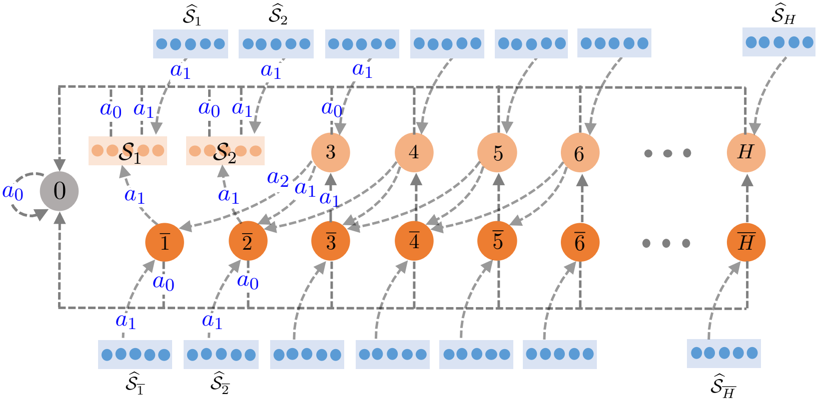
MDP construction for our lower bound





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

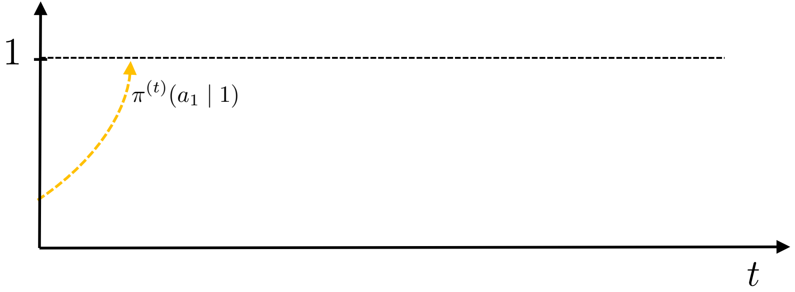
MDP construction for our lower bound



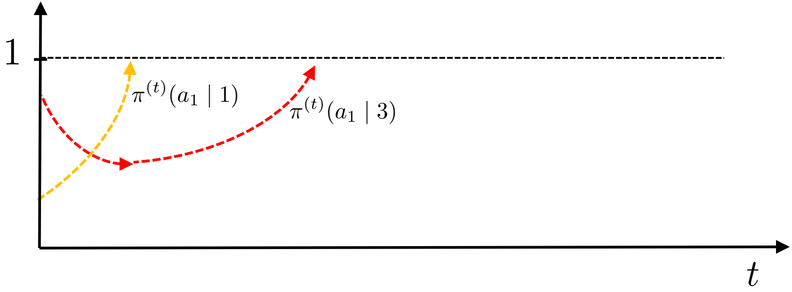
Key ingredients: for $3 \leq s \leq H \asymp \frac{1}{1-\gamma}$,

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

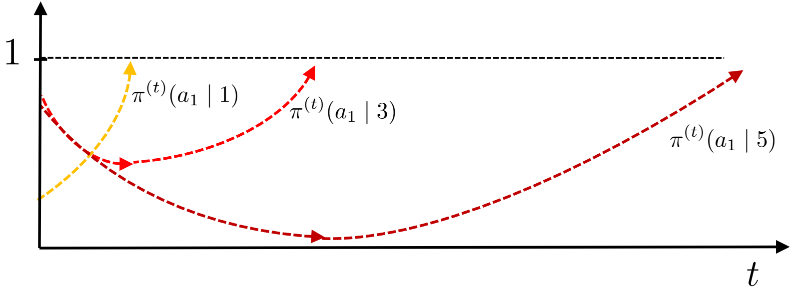
What is happening in our constructed MDP?



What is happening in our constructed MDP?

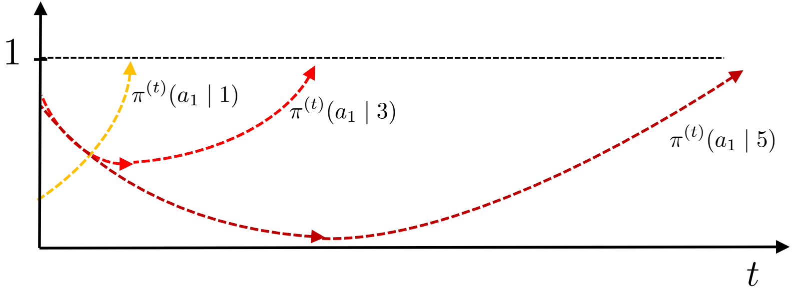


What is happening in our constructed MDP?



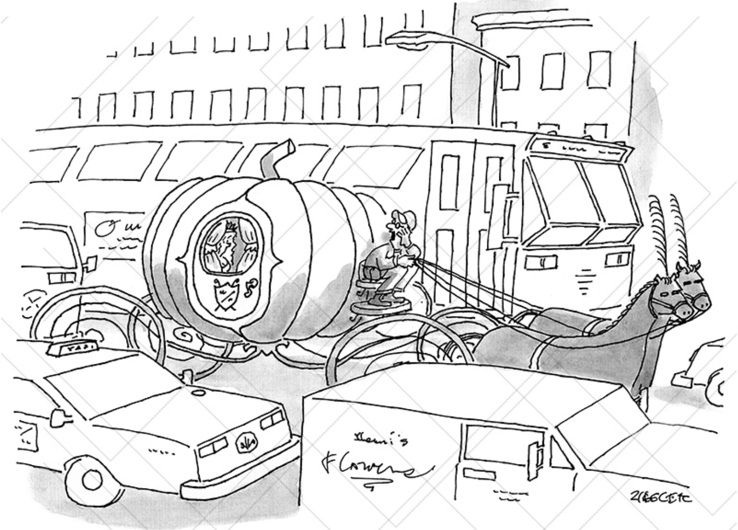
Convergence time for state s grows geometrically as s increases

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."

Accelerating the convergence via preconditioning and regularization

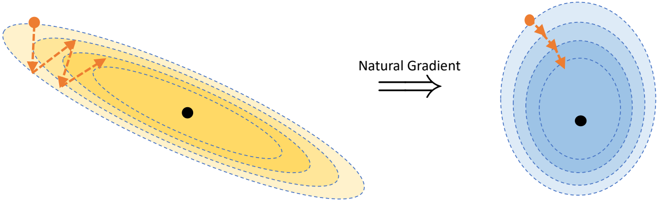
Shicong Cen
CMU

Chen Cheng
Stanford

Yuxin Chen
Princeton

Yuting Wei
CMU

Booster #1: natural policy gradient



Natural policy gradient (NPG) method (Kakade, 2002)

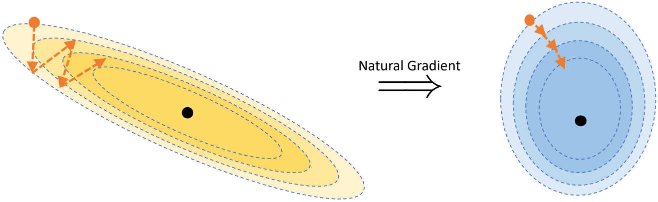
For $t = 0, 1, \cdots$

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

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 $t = 0, 1, \cdots$

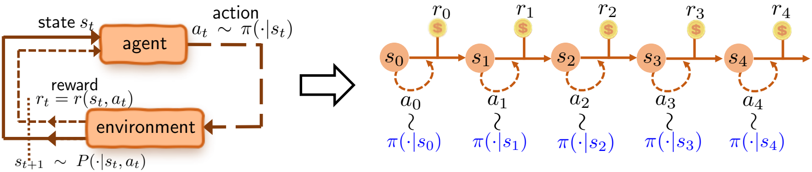
$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V^{\pi_\theta^{(t)}}(\rho)$$

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

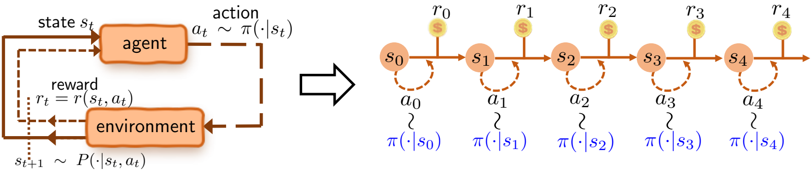


To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

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

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):

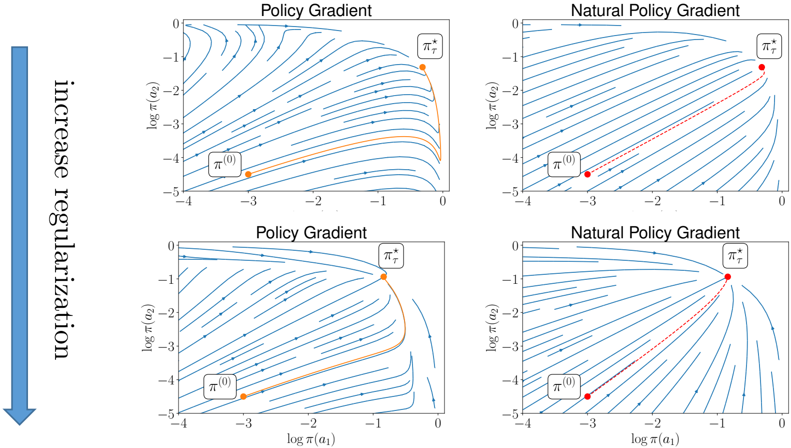
$$\forall s \in \mathcal{S}: \qquad V_\tau^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \Big(r_t + \tau \mathcal{H}(\pi(\cdot|s_t))\Big) \,\Big|\, s_0 = s\right]$$

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

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

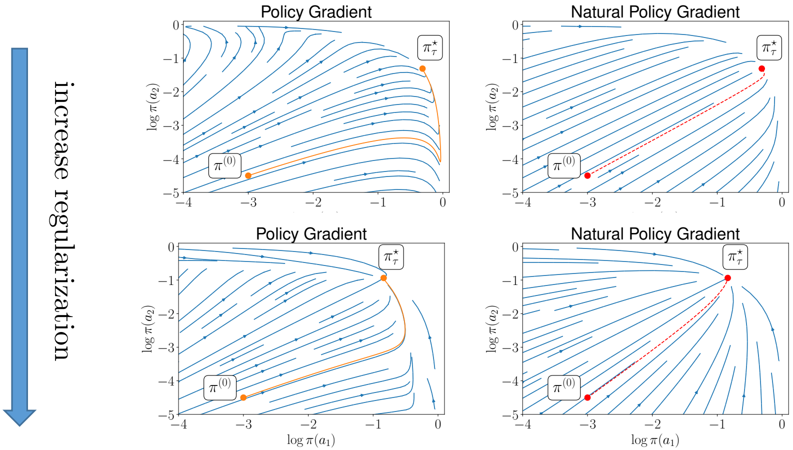
Entropy-regularized natural gradient helps!

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



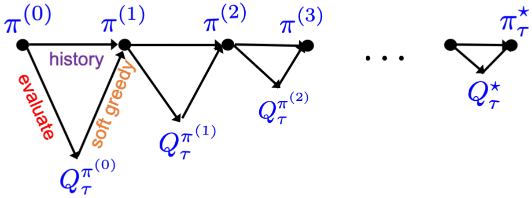
Entropy-regularized natural gradient helps!

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

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

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 '20)

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

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

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

for all $t \geq 0$, 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 $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \leq \epsilon$, 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 ($\eta = \frac{1-\gamma}{\tau}$):**

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

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le \epsilon$, the iteration complexity is at most

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

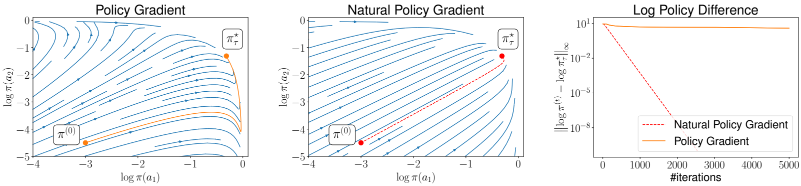
$$\frac{1}{\eta\tau} \log\left(\frac{C_1 \gamma}{\epsilon}\right)$$

- **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 rate independent of $|\mathcal{S}|$, $|\mathcal{A}|$!

Comparisons with entropy-regularized PG



(Mei et al., 2020) showed entropy-regularized PG achieves

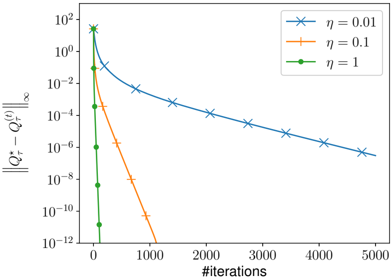
$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \le \left(V_\tau^\star(\rho) - V_\tau^{(0)}(\rho)\right)$$

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

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

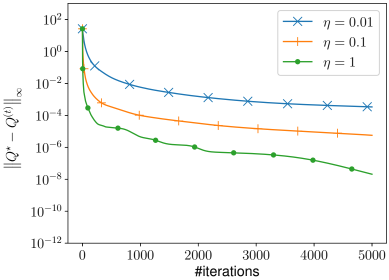
Comparison with unregularized NPG

Regularized NPG
$\tau = 0.001$



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

Vanilla NPG
$\tau = 0$

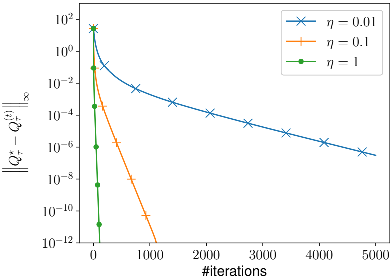


Sublinear rate: $\frac{1}{\min\{\eta,(1-\gamma)^2\}\epsilon}$
(Agarwal et al. 2019)

Comparison with unregularized NPG

Regularized NPG

$\tau = 0.001$

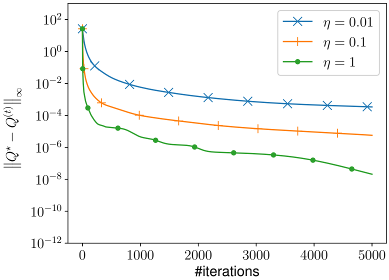


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

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

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

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

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

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; 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

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

Statistical implication: how many samples are sufficient to find an ϵ-optimal policy of the unregularized MDP?

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

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

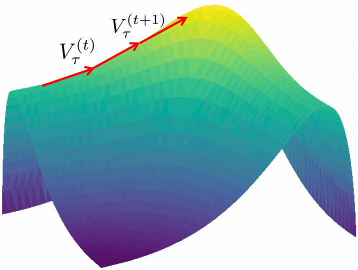
Statistical implication: how many samples are sufficient to find an ϵ-optimal policy of the unregularized MDP?

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

Recipe: set $\tau = \frac{(1-\gamma)\epsilon}{\log|\mathcal{A}|}$; use fresh samples for policy evaluation (Li et al., 2020).

A glimpse of the analysis

A key lemma: monotonic performance improvement

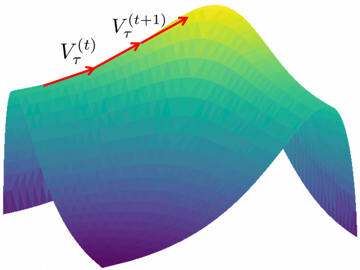


$$V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) = \mathbb{E}_{s\sim d_\rho^{(t+1)}}\Bigg[\left(\frac{1}{\eta} - \frac{\tau}{1-\gamma}\right)\underbrace{\mathsf{KL}\Big(\pi^{(t+1)}(\cdot|s)\,\big\|\,\pi^{(t)}(\cdot|s)\Big)}_{\text{KL divergence}}$$

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

discounted state visitation distribution

A key lemma: monotonic performance improvement



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

discounted state visitation distribution

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

Recall: Bellman's optimality principle

Bellman operator

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

- 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

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

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

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

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: $Q_\tau^\star$ is *unique* solution to

$$\mathcal{T}_\tau(Q_\tau^\star) = Q_\tau^\star$$

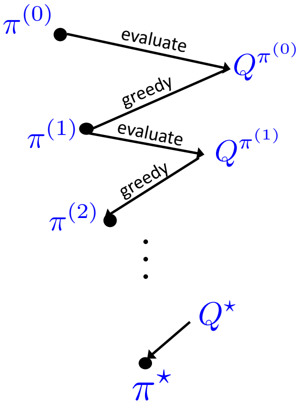
γ**-contraction of soft Bellman operator:**

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

Richard Bellman

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

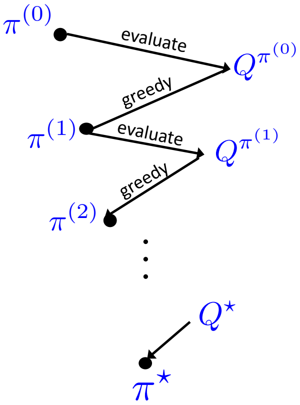
Policy iteration



Bellman operator

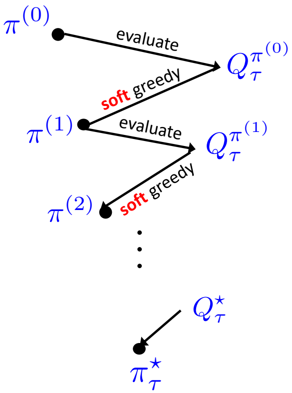
Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

Policy iteration



Bellman operator

Soft policy iteration



Soft Bellman operator

A key linear system: general learning rates

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

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

A key linear system: general learning rates

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

$$x_{t+1} \leq A x_t + \gamma \left(1 - \frac{\eta\tau}{1-\gamma}\right)^{t+1} y,$$

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

Shicong Cen
CMU

Wenhao Zhan
Princeton

Yuxin Chen
Princeton

Jason Lee
Princeton

The ever-important role of regularization

Leverage regularization to promote structural properties of the learned policy.

cost-sensitive RL

weighted 1-norm

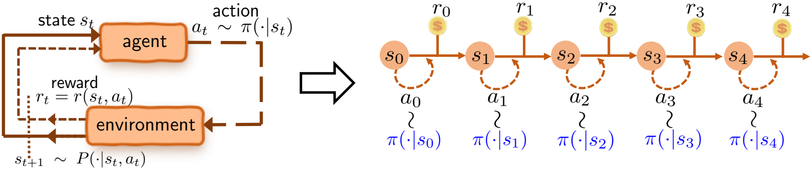
sparse exploration

Tsallis entropy

constrained and safe RL

log-barrier

Regularized RL in general form

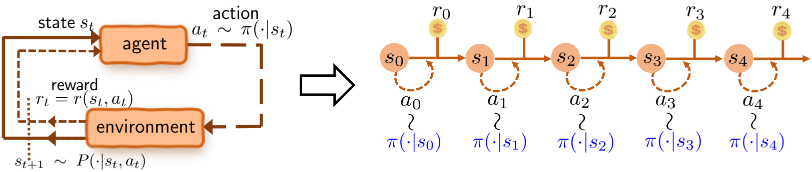


The regularized value function is defined as

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

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

Regularized RL in general form



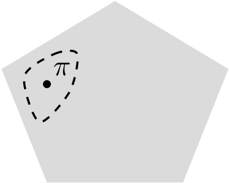
The regularized value function is defined as

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

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-regularized NPG = mirror descent with KL divergence (Lan, 2021; Shani et al., 2020):

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

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

for all $s \in \mathcal{S}$.

Generalized policy mirror descent (GPMD)

Definition (Generalized Bregman divergence, Kiwiel 1997)

The generalized Bregman divergence w.r.t. to a convex $h : \Delta(\mathcal{A}) \mapsto \mathbb{R}$ is defined as:

$$\begin{aligned} D_h(p, q; g) &= h(p) - h(q) - \langle g, p - q \rangle \\ &= h(p) - h(q) - \langle g - c \cdot \mathbf{1}, p - q \rangle, \end{aligned}$$

for $p, q \in \Delta(\mathcal{A})$, where $g \in \partial h(q)$ and $c \in \mathbb{R}$.

Generalized policy mirror descent (GPMD)

Definition (Generalized Bregman divergence, Kiwiel 1997)

The generalized Bregman divergence w.r.t. to a convex $h : \Delta(\mathcal{A}) \mapsto \mathbb{R}$ is defined as:

$$
\begin{aligned}
D_h(p, q; g) &= h(p) - h(q) - \langle g, p - q \rangle \\
&= h(p) - h(q) - \langle g - c \cdot \mathbf{1}, p - q \rangle,
\end{aligned}
$$

for $p, q \in \Delta(\mathcal{A})$, where $g \in \partial h(q)$ and $c \in \mathbb{R}$.

A natural idea

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

$$
\begin{aligned}
\pi^{(t+1)}(\cdot|s) = \underset{p \in \Delta(\mathcal{A})}{\operatorname{argmin}} \, &\langle -Q_\tau(s, \cdot), p \rangle + \tau h_s(p) \\
&+ \frac{1}{\eta} D_{h_s}(p, \pi^{(t)}(\cdot|s); \partial h_s(\pi^{(t)}(\cdot|s)))
\end{aligned}
$$

PMD with Generalized Bregman Divergence (**GPMD**)

Plugging in a recursive surrogate $\{\xi^{(t)}\}$ of $\partial h_s(\pi^{(t)}(\cdot|s))$, we obtain the formal algorithm.

Generalized policy mirror descent (GPMD) method

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

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

and

$$\xi^{(t+1)}(s,\cdot) = \frac{1}{1+\eta\tau}\xi^{(t)}(s,\cdot) + \frac{\eta}{1+\eta\tau} Q_\tau^{(t)}(s,\cdot).$$

The subproblem does not admit closed-form solution in general.

Linear convergence with exact gradient

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$; exact solution to subproblems.

— *Read our paper for the inexact case!*

Linear convergence with exact gradient

Exact oracle: perfect evaluation of $Q_\tau^{\pi^{(t)}}$ given $\pi^{(t)}$; exact solution to subproblems.

— *Read our paper for the inexact case!*

Theorem (Zhan*, Cen*, Huang, Chen, Lee, Chi '21)

For any learning rate $\eta > 0$, the GPMD updates satisfy

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

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

where $C_1 = \|Q_\tau^\star - Q_\tau^{(0)}\|_\infty + \frac{2}{1+\eta\tau}\|Q_\tau^\star - \tau\xi^{(0)}\|_\infty$.

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le \epsilon$, the iteration complexity is at most

- **General learning rates ($\eta > 0$):**

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

- **Regularized policy iteration ($\eta = \infty$):**

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

Implications

To reach $\|Q_\tau^\star - Q_\tau^{(t+1)}\|_\infty \le \epsilon$, the iteration complexity is at most

- **General learning rates ($\eta > 0$):**

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

- **Regularized policy iteration ($\eta = \infty$):**

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

Global linear convergence of GPMD at a **dimension-free** rate!

Comparison with PMD (Lan, 2021)

Policy mirror descent (PMD) method (Lan, 2021)

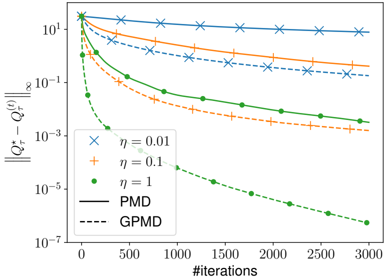
For $t = 0, 1, \cdots,$

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

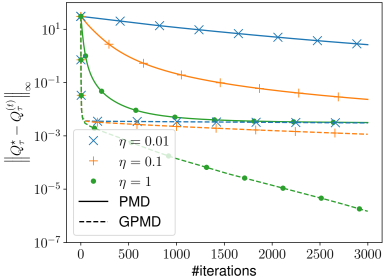
- Linear convergence is established only when h_s is stronger than entropy regularization ($h_s + \mathcal{H}$ is convex).
- In contrast, GPMD converges linearly for general convex and nonsmooth h_s!

Numerical examples

h_s = **Tsallis Entropy**

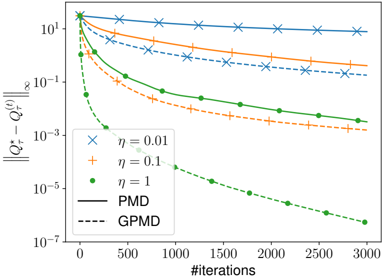


h_s = **Log Barrier**

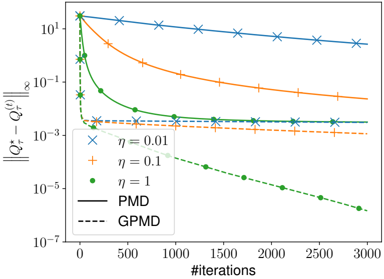


Numerical examples

h_s = **Tsallis Entropy**



h_s = **Log Barrier**



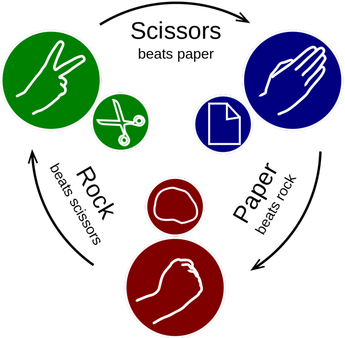
GPMD achieves faster convergence than PMD!

Bonus track: entropy-regularized games

Shicong Cen
CMU

Yuting Wei
CMU

Zero-sum entropy-regularized two-player matrix game



			
	0	-1	1
	1	0	-1
	-1	1	0

Quantal response equilibrium (McKelvey and Palfrey, 1995)

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top A \nu + \tau \mathcal{H}(\mu) - \tau \mathcal{H}(\nu)$$

- Basic building block for solving value iteration in zero-sum two-player Markov games.

Motivation: an implicit update method

Implicit update (IU) method

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

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\nu^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \mu^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

Motivation: an implicit update method

Implicit update (IU) method

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

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\nu^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \mu^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

Theorem (Cen, Wei, Chi, 2021)

Suppose that $0 < \eta \le 1/\tau$*, then for all* $t \ge 0$*,*

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \le (1 - \eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big),$$

where $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) = \mathsf{KL}\big(\mu_\tau^\star \| \mu^{(t)}\big) + \mathsf{KL}\big(\nu_\tau^\star \| \nu^{(t)}\big)$.

Motivation: an implicit update method

Implicit update (IU) method

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

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\nu^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top\mu^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

Theorem (Cen, Wei, Chi, 2021)

Suppose that $0 < \eta \le 1/\tau$, *then for all* $t \ge 0$,

$$\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) \le (1-\eta\tau)^t \mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(0)}\big),$$

where $\mathsf{KL}\big(\zeta_\tau^\star \,\|\, \zeta^{(t)}\big) = \mathsf{KL}\big(\mu_\tau^\star \| \mu^{(t)}\big) + \mathsf{KL}\big(\nu_\tau^\star \| \nu^{(t)}\big)$.

Can we make this practical?

From implicit updates to policy extragradient methods

Predictive update (PU) method

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

2. *update:*

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top\bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

From implicit updates to policy extragradient methods

Predictive update (PU) method

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

1. *extrapolate/predict:*

$$\begin{cases} \bar{\mu}^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\nu^{(t)}]/\tau\right)^{\eta\tau} \\ \bar{\nu}^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \mu^{(t)}]/\tau\right)^{\eta\tau} \end{cases}$$

2. *update:*

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

From implicit updates to policy extragradient methods

Optimistic multiplicative weights update (OMWU) method

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

1. *extrapolate/predict:*

$$\begin{cases} \bar{\mu}^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t)}]/\tau\right)^{\eta\tau} \\ \bar{\nu}^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t)}]/\tau\right)^{\eta\tau} \end{cases}$$

2. *update:*

$$\begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

Last-iterate convergence

- **Entropy-regularized matrix game:** To get an ϵ-optimal solution to the regularized problem (ϵ-**QRE**), the iteration complexity is at most

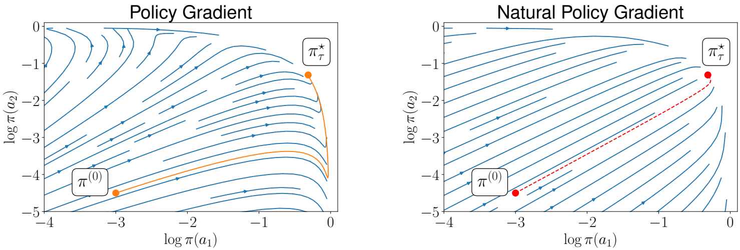
$$\widetilde{O}\left(\left(1+\frac{\|A\|_\infty}{\tau}\right)\log\frac{1}{\epsilon}\right).$$

- **Unregularized matrix game:** To get an ϵ-optimal solution to the unregularized problem (ϵ-**NE**), the iteration complexity is at most

$$\widetilde{O}\left(\frac{\|A\|_\infty}{\epsilon}\right).$$

  No need to assume unique Nash equilibrium!

Concluding remarks



fast global linear convergence of RL enabled by
regularization + preconditioning

Future directions:

- function approximation
- sample complexities
- Markov games
- multi-agent RL

Thanks!

- Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization, *Operations Research*; arXiv: 2007.06558.
- Softmax Policy Gradient Methods Can Take Exponential Time to Converge, COLT 2021; arXiv: 2102.11270.
- 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.

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