Foundations of Reinforcement Learning

Policy optimization: the role of regularization

Shicong Cen and Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2023

Outline

Global convergence of entropy-regularized NPG

A mirror descent perspective and alternative analysis

Beyond entropy regularization

Softmax 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:

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

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

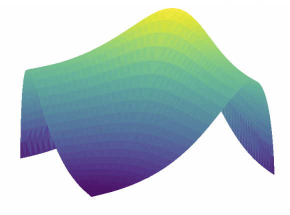
Policy gradient method

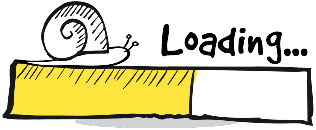
For $t = 0, 1, \cdots$

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

where η *is the learning rate.*

How fast does softmax PG converge?

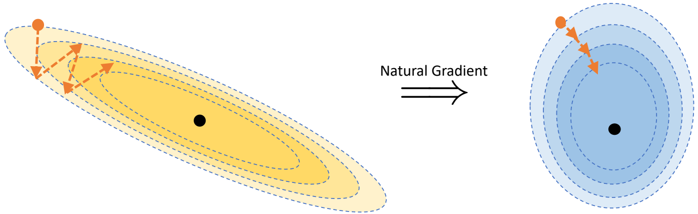




- [Agarwal et al., 2021] showed that softmax PG converges asymptotically to the global optimal policy.
- [Li et al., 2023] showed that softmax PG may take $|\mathcal{S}|^{2^{\Theta\left(\frac{1}{1-\gamma}\right)}}$ iterations to converge!

Can we accelerate the convergence using algorithmic tricks?

Natural policy gradient



Natural policy gradient (NPG) method [Kakade, 2001]

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:

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

Global convergence of NPG

Theorem 1 ([Agarwal et al., 2021])

Set $\pi^{(0)}$ as a uniform policy. For all $t \geq 0$, we have

$$V^{(t)}(\rho) \geq V^\star(\rho) - \left(\frac{\log|\mathcal{A}|}{\eta} + \frac{1}{(1-\gamma)^2}\right)\frac{1}{t}.$$

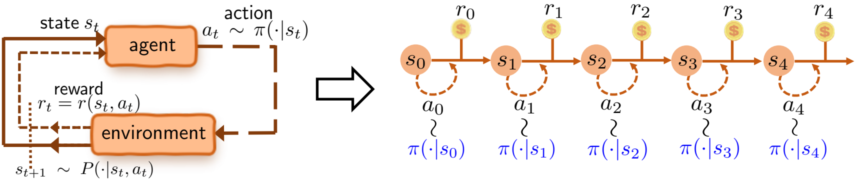
Implication: set $\eta \geq (1-\gamma)^2 \log|\mathcal{A}|$, we find an ϵ-optimal policy within at most

$$\frac{2}{(1-\gamma)^2\epsilon} \text{ iterations.}$$

Global convergence at a sublinear rate independent of $|\mathcal{S}|$, $|\mathcal{A}|$!

Global convergence of entropy-regularized NPG

Entropy regularization



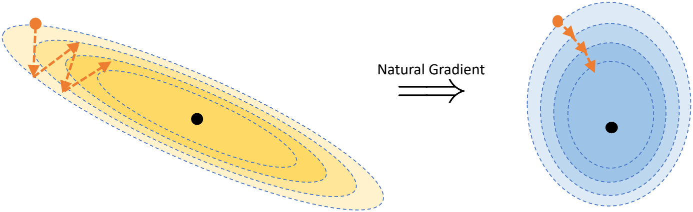
To encourage exploration, promote the stochasticity of the policy using the **"soft"** value function (Williams and Peng, 1991):

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

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 NPG



Entropy-regularized NPG

For $t = 0, 1, \cdots$

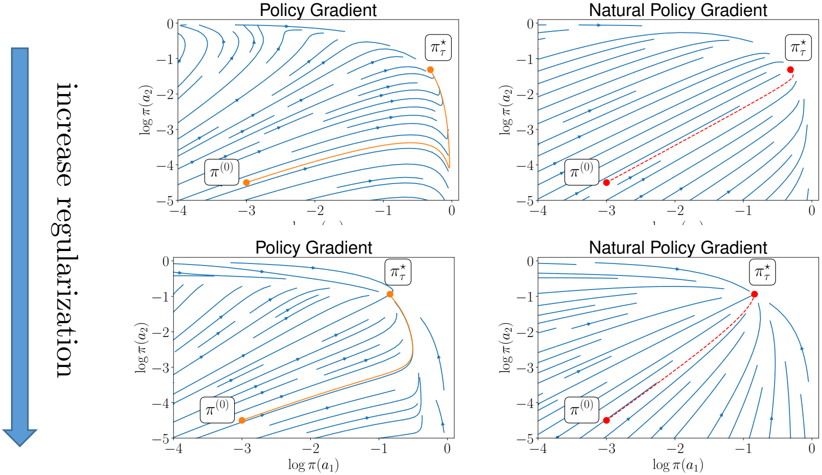
$$\theta^{(t+1)} = \theta^{(t)} + \eta (\mathcal{F}_\rho^\theta)^\dagger \nabla_\theta V_\tau^{\pi_\theta^{(t)}}(\rho)$$

where η is the learning rate and $\mathcal{F}_\rho^\theta$ is the Fisher information matrix:

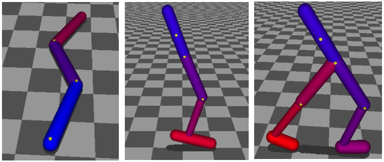
$$\mathcal{F}_\rho^\theta := \mathbb{E}\left[\big(\nabla_\theta \log \pi_\theta(a|s)\big) \big(\nabla_\theta \log \pi_\theta(a|s)\big)^\top \right].$$

Entropy-regularized natural gradient helps!

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



Unreasonable effectiveness in practice



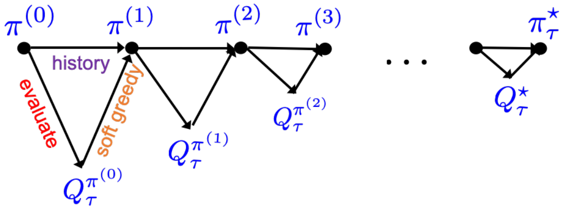
TRPO = NPG + line search
(Schulman et al., 2015)

> 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)

Can we justify the efficacy of entropy-regularized NPG?

Entropy-regularized NPG in the tabular setting



Entropy-regularized NPG

For $t = 0, 1, \cdots$, the policy is updated via

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

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 2 ([Cen et al., 2022])

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

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

$$\|V_\tau^\star - V_\tau^{(t+1)}\|_\infty \leq 3C_1 \left(1 - \eta\tau\right)^t,$$

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

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

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

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

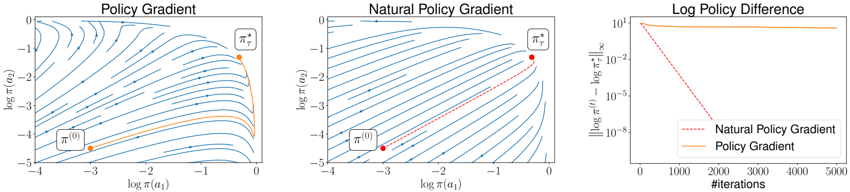
$$\frac{1}{\eta\tau} \log\left(\frac{C_1\gamma}{\epsilon}\right)$$

- **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

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

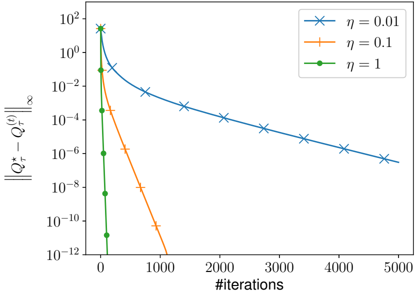
$$\cdot \exp\left(-\frac{(1-\gamma)^4 t}{(8/\tau + 4 + 8\log|\mathcal{A}|)|\mathcal{S}|} \left\| \frac{d_\rho^{\pi_\tau^\star}}{\rho} \right\|_\infty^{-1} \min_s \rho(s) \underbrace{\left(\inf_{0\le k\le t-1} \min_{s,a} \pi^{(k)}(a|s) \right)^2}_{\text{can be exponential in } |\mathcal{S}| \text{ and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

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

Comparison with unregularized NPG

Regularized NPG

$\tau = 0.001$

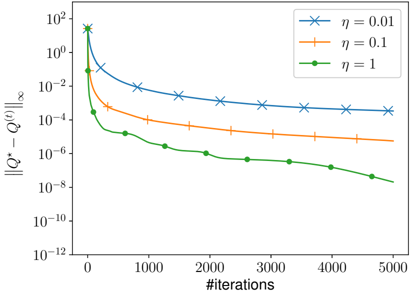


Linear rate: $\frac{1}{\eta\tau}\log\left(\frac{1}{\epsilon}\right)$

[Cen et al., 2022]

Vanilla NPG

$\tau = 0$



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

[Agarwal et al., 2020]

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 such as REINFORCE (Williams, 1992).

Inexact entropy-regularized NPG:

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

Question: Robustness of entropy-regularized NPG?

Linear convergence with inexact gradients

Theorem 3 ([Cen et al., 2022]; 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

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

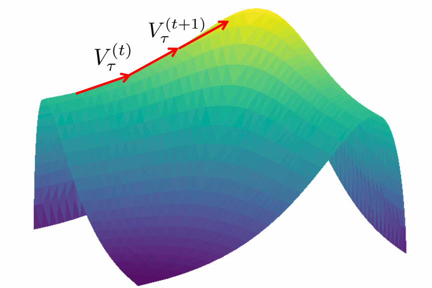
- **Crude sample complexity for finding an ϵ-optimal policy in the original MDP using a generative model:**

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

  - set $\tau = (1-\gamma)\epsilon/\log|\mathcal{A}|$;
  - in a generative model takes no larger than $\widetilde{O}(|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-3}\delta^{-2})$ samples to achieve δ-accurate estimate of $Q_\tau^{(t)}$ per iteration;

A glimpse of the analysis

A key lemma: monotonic performance improvement



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

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

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

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

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

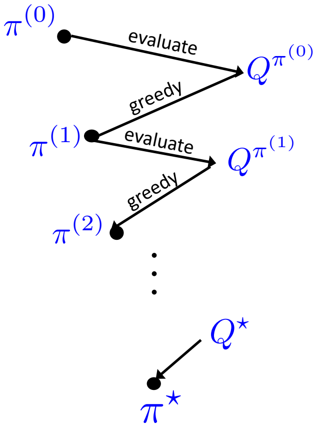
γ-contraction of soft Bellman operator:

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

Richard Bellman

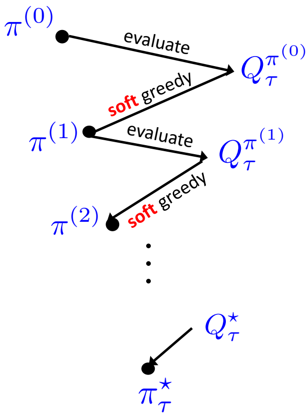
Analysis of soft policy iteration ($\eta = \frac{1-\gamma}{\tau}$)

Policy iteration



Bellman operator

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

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

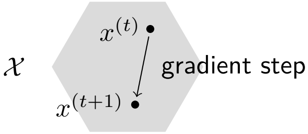
where

$$A := \begin{bmatrix} \gamma \\ 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{\eta\tau}{1-\gamma} & 1 - \frac{\eta\tau}{1-\gamma} \end{bmatrix}$$

is a rank-1 matrix with a non-zero eigenvalue $\underbrace{1-\eta\tau}_{\text{contraction rate!}}$.

A mirror descent perspective and alternative analysis

Detour: mirror descent



- The gradient descent update rule

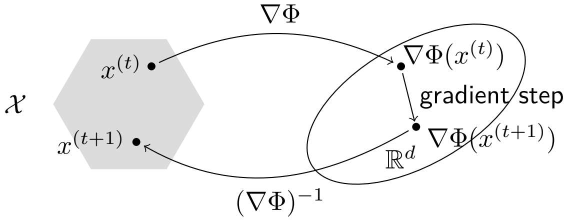
$$x^{(t+1)} = P_{\mathcal{X}}\left(x^{(t)} - \eta_{\mathsf{GD}} \nabla f(x^{(t)})\right)$$

  is equivalent to minimizing local quadratic approximation of f:

$$x^{(t+1)} = \arg\min_{x \in \mathcal{X}} \left\langle \nabla f(x^{(t)}), x - x^{(t)} \right\rangle + \frac{1}{2\eta_{\mathsf{GD}}} \|x - x^{(t)}\|_2^2.$$

- $\eta_{\mathsf{GD}} > 0$ is the step size and $P_{\mathcal{X}}$ is the projection operator to $\mathcal{X}$.

Detour: mirror descent



- The mirror descent update rule

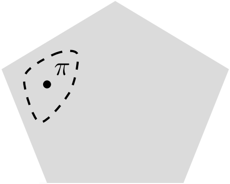
$$x^{(t+1)} = \arg\min_{x\in\mathcal{X}} \left\langle \nabla f(x^{(t)}), x - x^{(t)} \right\rangle + \frac{1}{\eta_{\mathsf{MD}}} D_\Phi(x, x^{(t)})$$

is obtained by replacing $\frac{1}{2}\|x - x^{(t)}\|_2^2$ with Bregman divergence

$$D_\Phi(x, x^{(t)}) = \Phi(x) - \Phi(x^{(t)}) - \left\langle x - x^{(t)}, \nabla\Phi(x^{(t)}) \right\rangle.$$

- $\eta_{\mathsf{MD}} > 0$ is the step size.

A mirror descent view of entropy-regularized NPG



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

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

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

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

A mirror descent view of entropy-regularized NPG

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

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

for all $s\in\mathcal{S}$, with

$$\eta_{\mathsf{MD}} = \frac{\eta}{1-\gamma-\eta\tau}.$$

Redux: Linear convergence with exact gradient

[Lan, 2022] provided an alternative framework for analyzing regularized natural policy gradient (called policy mirror descent - PMD).

Theorem 4 ([Lan, 2022])

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

$$V_\tau^\star(\rho) - V_\tau^{(t+1)}(\rho) \leq C_2 \left\| \frac{\rho}{\nu_\tau^\star} \right\|_\infty \max\left\{\gamma, 1 - \frac{\eta\tau}{1-\gamma}\right\}^{t+1}$$

for all $t \geq 0$, *where* $\nu_\tau^\star$ *is the stationary distribution of* $\pi_\tau^\star$,

$$\left\| \frac{\rho}{\nu_\tau^\star} \right\|_\infty = \max_{s \in \mathcal{S}} \frac{\rho(s)}{\nu_\tau^\star(s)},$$

and $\qquad C_2 = V_\tau^\star(\nu_\tau^\star) - V_\tau^{(0)}(\nu_\tau^\star) + \frac{1-\gamma}{\eta} \mathop{\mathbb{E}}_{s \sim \nu_\tau^\star}\left[\mathsf{KL}(\pi_\tau^\star \,\|\, \pi^{(0)}(s))\right].$

Take-away message

With a *fixed* learning rate $0 < \eta \leq (1-\gamma)/\tau$, the iteration complexity for entropy-regularized NPG to reach

$$V_\tau^\star(\rho) - V_\tau^{(t)}(\rho) \leq \epsilon$$

is no larger than the **minimum** of

$$\widetilde{O}\Big(\frac{1}{\eta\tau}\log\Big(\frac{\text{init. error}}{\epsilon}\Big)\Big) \qquad \text{[Cen et al., 2022]}$$

and

$$\widetilde{O}\Big(\max\Big\{\frac{1}{1-\gamma}, \frac{1-\gamma}{\eta\tau}\Big\}\log\Big\|\frac{\rho}{\nu_\tau^\star}\Big\|_\infty \log\Big(\frac{\text{init. error}}{\epsilon}\Big)\Big). \qquad \text{[Lan, 2022]}$$

Key lemmas

Regularized performance difference lemma: for any two policies π and π',

$$
\begin{aligned}
& V_\tau^{\pi}(\rho) - V_\tau^{\pi'}(\rho) \\
& = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim d_\rho^\pi} \left[\left\langle Q_\tau^{\pi'}(s), \pi^{(t+1)}(s) - \pi'(s) \right\rangle + \tau \mathcal{H}(\pi^{(t+1)}(s)) - \tau \mathcal{H}(\pi'(s)) \right].
\end{aligned}
$$

Regularized three-point identity: for any policy π,

$$
\begin{aligned}
& \frac{\eta}{1-\gamma-\eta\tau} \left[\left\langle Q_\tau^{(t)}(s), \pi^{(t+1)}(s) - \pi(s) \right\rangle + \tau \mathcal{H}(\pi^{(t+1)}(s)) - \tau \mathcal{H}(\pi(s)) \right] \\
& = \frac{1-\gamma}{1-\gamma-\eta\tau} \mathsf{KL}(\pi \,\|\, \pi^{(t+1)}(s)) + \mathsf{KL}(\pi^{(t+1)}(s) \,\|\, \pi^{(t)}(s)) - \mathsf{KL}(\pi \,\|\, \pi^{(t)}(s)).
\end{aligned}
$$

Proof sketch

Applying regularized performance difference lemma gives:

$$
\begin{aligned}
&V_\tau^{(t+1)}(\rho) - V_\tau^{(t)}(\rho) \\
&= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_\rho^{(t+1)}} \left[\left\langle Q_\tau^{(t)}(s), \pi^{(t+1)}(s) - \pi^{(t)}(s) \right\rangle + \tau\mathcal{H}(\pi^{(t+1)}(s)) - \tau\mathcal{H}(\pi^{(t)}(s)) \right] \\
&\geq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{(t+1)}}{d_\rho^{\pi_\tau^\star}} \right\|_\infty \mathop{\mathbb{E}}_{s\sim d_\rho^{\pi_\tau^\star}} \left[\left\langle Q_\tau^{(t)}(s), \pi^{(t+1)}(s) - \pi^{(t)}(s) \right\rangle + \tau\mathcal{H}(\pi^{(t+1)}(s)) - \tau\mathcal{H}(\pi^{(t)}(s)) \right]
\end{aligned}
$$

With ρ set to stationary state distribution $\nu_\tau^\star$ of $\pi_\tau^\star$, we have

$$
\frac{1}{1-\gamma} \left\| \frac{d_\rho^{(t+1)}}{d_\rho^{\pi_\tau^\star}} \right\| = \frac{1}{1-\gamma} \left\| \frac{d_{\nu_\tau^\star}^{(t+1)}}{\nu_\tau^\star} \right\| \geq 1.
$$

Proof sketch

We end up with:

$$
\begin{aligned}
&V_\tau^{(t+1)}(\nu_\tau^\star) - V_\tau^{(t)}(\nu_\tau^\star) \\
&\geq \mathop{\mathbb{E}}_{s \sim d_{\nu_\tau^\star}^{\pi_\tau^\star}} \left[\left\langle Q_\tau^{(t)}(s), \pi^{(t+1)}(s) - \pi^{(t)}(s) \right\rangle + \tau \mathcal{H}(\pi^{(t+1)}(s)) - \tau \mathcal{H}(\pi^{(t)}(s)) \right].
\end{aligned}
$$

Adding and subtracting terms,

$$
\begin{aligned}
&V_\tau^{(t+1)}(\nu_\tau^\star) - V_\tau^{(t)}(\nu_\tau^\star) \\
&= \mathop{\mathbb{E}}_{s \sim d_{\nu_\tau^\star}^{\pi_\tau^\star}} \left[\left\langle Q_\tau^{(t)}(s), \pi_\tau^\star(s) - \pi^{(t)}(s) \right\rangle + \tau \mathcal{H}(\pi_\tau^\star(s)) - \tau \mathcal{H}(\pi^{(t)}(s)) \right] \\
&\qquad + \mathop{\mathbb{E}}_{s \sim \nu_\tau^\star} \left[\left\langle Q_\tau^{(t)}(s), \pi^{(t+1)}(s) - \pi_\tau^\star(s) \right\rangle + \tau \mathcal{H}(\pi^{(t+1)}(s)) - \tau \mathcal{H}(\pi_\tau^\star(s)) \right]
\end{aligned}
$$

Proof sketch

Applying the two key lemmas gives

$$
\begin{aligned}
&V_\tau^{(t+1)}(\nu_\tau^\star) - V_\tau^{(t)}(\nu_\tau^\star) \\
&\geq (1-\gamma)(V_\tau^\star(\nu_\tau^\star) - V_\tau^{(t)}(\nu_\tau^\star)) \\
&\qquad + \frac{1}{\eta} \mathop{\mathbb{E}}_{s\sim\nu_\tau^\star} \Big[(1-\gamma)\mathsf{KL}(\pi_\tau^\star \,\|\, \pi^{(t+1)}(s)) - (1-\gamma-\eta\tau)\mathsf{KL}(\pi_\tau^\star \,\|\, \pi^{(t)}(s))\Big].
\end{aligned}
$$

Rearranging the terms,

$$
\begin{aligned}
&V_\tau^\star(\nu_\tau^\star) - V_\tau^{(t+1)}(\nu_\tau^\star) + \frac{1-\gamma}{\eta} \mathop{\mathbb{E}}_{s\sim\nu_\tau^\star} \Big[\mathsf{KL}(\pi_\tau^\star \,\|\, \pi^{(t+1)}(s))\Big] \\
&\leq \gamma(V_\tau^\star(\nu_\tau^\star) - V_\tau^{(t)}(\nu_\tau^\star)) + \frac{1-\gamma-\eta\tau}{\eta} \mathop{\mathbb{E}}_{s\sim\nu_\tau^\star} \Big[\mathsf{KL}(\pi_\tau^\star \,\|\, \pi^{(t)}(s))\Big] \\
&\leq \max\Big\{\gamma, 1-\frac{\eta\tau}{1-\gamma}\Big\}\Big\{V_\tau^\star(\nu_\tau^\star) - V_\tau^{(t)}(\nu_\tau^\star) + \frac{1-\gamma}{\eta} \mathop{\mathbb{E}}_{s\sim\nu_\tau^\star} \Big[\mathsf{KL}(\pi_\tau^\star \,\|\, \pi^{(t)}(s))\Big]\Big\}.
\end{aligned}
$$

Proof sketch

Finally, we have

$$
\begin{aligned}
&V_\tau^\star(\nu_\tau^\star) - V_\tau^{(t+1)}(\nu_\tau^\star) + \frac{1-\gamma}{\eta} \mathop{\mathbb{E}}_{s\sim\nu_\tau^\star}\Big[\mathsf{KL}\big(\pi_\tau^\star \,\|\, \pi^{(t+1)}(s)\big)\Big] \\
&\le \max\Big\{\gamma, 1-\frac{\eta\tau}{1-\gamma}\Big\}^{t+1}\Big\{V_\tau^\star(\nu_\tau^\star) - V_\tau^{(0)}(\nu_\tau^\star) + \frac{1-\gamma}{\eta} \mathop{\mathbb{E}}_{s\sim\nu_\tau^\star}\Big[\mathsf{KL}\big(\pi_\tau^\star \,\|\, \pi^{(0)}(s)\big)\Big]\Big\}
\end{aligned}
$$

Applying the bound

$$
V_\tau^\star(\rho) - V_\tau^{(t+1)}(\rho) \le \Big\|\frac{\rho}{\nu_\tau^\star}\Big\|_\infty \big(V_\tau^\star(\nu_\tau^\star) - V_\tau^{(t+1)}(\nu_\tau^\star)\big)
$$

finishes the proof.

Beyond entropy regularization

Beyond entropy regularization

Leverage regularization to promote structural properties of the learned policy.

cost-sensitive RL

weighted 1-norm

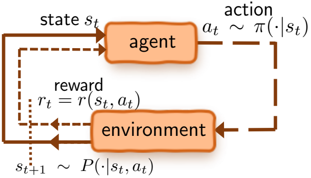
sparse exploration

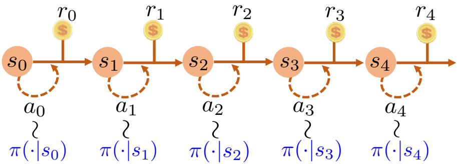
Tsallis entropy

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. $\pi(\cdot|s)$.

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

Generalized Policy Mirror Descent (GPMD)

Generalized Policy Mirror Descent (GPMD) [Zhan et al., 2023]

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

$$\pi^{(t+1)}(\cdot|s) = \operatorname*{argmin}_{p \in \Delta(\mathcal{A})} \langle -Q_\tau(s, \cdot), p \rangle + \tau h_s(p) + \frac{1}{\eta_{\text{MD}}} \underbrace{D_{h_s}(p, \pi^{(t)}(\cdot|s); \partial h_s(\pi^{(t)}(\cdot|s)))}_{\text{Generalized Bregman divergence w.r.t. } h_s},$$

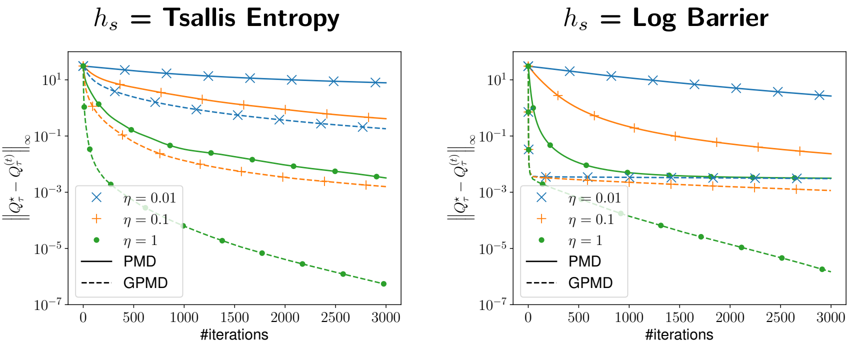
where a surrogate of $\partial h_s(\pi^{(t)}(\cdot|s))$ is updated recursively.

Compare with PMD [Lan, 2022]:

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

- GPMD achieves linear convergence for general convex and nonsmooth h_s! In contrast, PMD requires $h_s + \mathcal{H}$ is convex.

Numerical examples



GPMD achieves faster convergence than PMD!

References I

Agarwal, A., Kakade, S., and Yang, L. F. (2020).
Model-based reinforcement learning with a generative model is minimax optimal.
Conference on Learning Theory, pages 67–83.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021).
On the theory of policy gradient methods: Optimality, approximation, and distribution shift.
The Journal of Machine Learning Research, 22(1):4431–4506.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2022).
Fast global convergence of natural policy gradient methods with entropy regularization.
Operations Research, 70(4):2563–2578.

Kakade, S. M. (2001).
A natural policy gradient.
Advances in Neural Information Processing Systems, 14.

Lan, G. (2022).
Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes.
Mathematical programming, pages 1–48.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2023).
Softmax policy gradient methods can take exponential time to converge.
Mathematical Programming.

References II

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020).
On the global convergence rates of softmax policy gradient methods.
In *International Conference on Machine Learning*, pages 6820–6829. PMLR.

Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. (2023).
Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence.
SIAM Journal on Optimization.