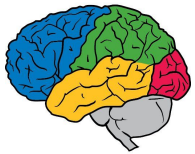
On the Global Convergence Rates of Softmax Policy Gradient Methods

Jincheng Mei[1,2*]

Joint work w/ Chenjun Xiao[1], Csaba Szepesvari[1,3], Dale Schuurmans[1,2]

[1]University of Alberta, [2]Google Brain, [3]DeepMind

* Work done as an intern in Google Research, Brain team



Main contributions

With true gradient:

1. Softmax policy gradient converges to optimal policy in a rate $O(1/t)$.
2. Entropy regularized softmax policy gradient converges to softmax optimal policy in a rate $O(1/e^{-t})$.
3. Softmax policy gradient follows a rate lower bound $\Omega(1/t)$.
4. Non-uniform Łojasiewicz degree as a deeper reason for rate separation.

Reinforcement Learning (RL)

Finite Markov Decision Processes (MDPs) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$

- state space $\mathcal{S}$; action space $\mathcal{A}$
- reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$; transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
- discount factor $\gamma \in [0, 1)$

Bounded reward assumption

Assumption 1 (Bounded reward). $r(s, a) \in [0, 1], \forall (s, a)$.

Notations

State value: $V^{\pi}(s) \coloneqq \mathop{\mathbb{E}}_{\substack{s_0=s, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \qquad V^{\pi}(\rho) \coloneqq \mathbb{E}_{s \sim \rho} \left[V^{\pi}(s) \right]$

State-action value: $Q^{\pi}(s, a) \coloneqq r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) V^{\pi}(s')$

Advantage function: $A^{\pi}(s, a) \coloneqq Q^{\pi}(s, a) - V^{\pi}(s)$

(discounted) state distribution: $d^{\pi}_{s_0}(s) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | s_0, \pi, \mathcal{P}) \qquad d^{\pi}_{\rho}(s) \coloneqq \mathbb{E}_{s_0 \sim \rho} \left[d^{\pi}_{s_0}(s) \right]$

Policy gradient (PG)

Policy gradient: foundational concept of policy search and actor-critic

Theorem 1 (Policy gradient theorem (Sutton et al., 2000)). *Suppose* $\theta \mapsto \pi_\theta(a|s)$ *is differentiable w.r.t.* θ, $\forall(s,a)$,

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim d_\mu^{\pi_\theta}} \left[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} Q^{\pi_\theta}(s,a) \right], \quad (8)$$

where $\mu \in \Delta(\mathcal{S})$ *is an initial state distribution.*

Algorithm 1 Policy Gradient Method

Input: Learning rate $\eta > 0$.
Initialize logit $\theta_1(s,a)$ for all (s,a).
for $t = 1$ **to** T **do**
  $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}$.
end for

Settings

Tabular cases: $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

Softmax parametrized policies: $\pi_\theta(a|s) = \frac{\exp\{\theta(s,a)\}}{\sum_{a'} \exp\{\theta(s,a')\}}$

True gradients:

Lemma 1. *Softmax policy gradient w.r.t. θ is*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a).$$

Open problems

1. Convergence rate of softmax policy gradient was unknown.

   The best results was asymptotic convergence in Agarwal et al. $V^{\pi_{\theta_t}}(\rho) \to V^*(\rho)$ as $t \to \infty$.

2. Convergence rate of entropy regularized softmax policy gradient was unknown.

   Stated as an open question in Agarwal et al.

3. No theoretical understanding why entropy helps policy optimization.

   There were some empirical suggestive observations in Ahmed et al.

   Results for mirror descent in Shani et al. and Vieillard et al., but lower bounds have been missing to make conclusions.

General MDPs

Non-concavity: $\max_{\theta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} V^{\pi_\theta}(\rho)$ is non-concave maximization problem.

Main results:

Theorem 4. *Let Assumption 2 hold and let* $\{\theta_t\}_{t\geq 1}$ *be generated using Algorithm 1 with* $\eta = (1-\gamma)^3/8$, c *the positive constant from Lemma 9. Then, for all* $t \geq 1$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{16S}{c^2(1-\gamma)^6 t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty .$$

Assumption 2 (Sufficient exploration). *The initial state distribution satisfies* $\min_s \mu(s) > 0$.

First convergence-rate result for softmax policy gradient.

General MDPs

1. Smoothness: **Lemma 7** (Smoothness). *$V^{\pi_\theta}(\rho)$ is $8/(1-\gamma)^3$-smooth.*

2. Non-uniform Lojasiewicz inequality:

Lemma 8 (Non-uniform Łojasiewicz). *Suppose $\mu(s) > 0$ for all state s. Then,*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S} \cdot \left\| d_\rho^{\pi^*} / d_\mu^{\pi_\theta} \right\|_\infty} \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right],$$

where $a^*(s) := \arg\max_a \pi^*(a|s)$, $s \in \mathcal{S}$.

3. Minimum probability of optimal action:

Lemma 9. *Let Assumption 2 hold. Then, $c := \inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) > 0$.*

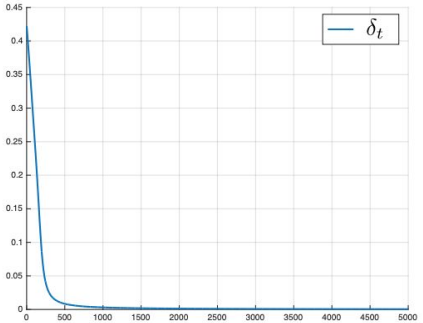
Sketch

Ascent lemma for smooth function: guaranteed progress

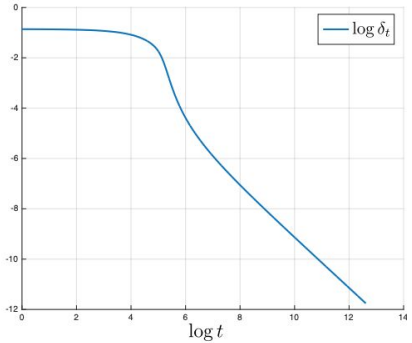
$$
\begin{aligned}
V^{\pi_{\theta_t}}(\mu) - V^{\pi_{\theta_{t+1}}}(\mu) &\leq -\frac{(1-\gamma)^3}{16} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 \\
&\leq -\frac{(1-\gamma)^3}{16S} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\theta_t}}} \right\|_\infty^{-2} \cdot \left[\min_s \pi_{\theta_t}(a^*(s)|s) \right]^2 \cdot \left[V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \right]^2 \\
&\leq -\frac{(1-\gamma)^5}{16S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot \left[\inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) \right]^2 \cdot \left[V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \right]^2
\end{aligned}
$$

Verifications

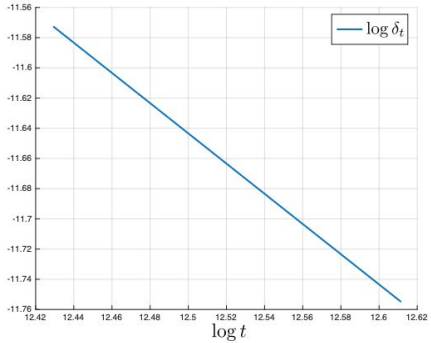
Problem: One-state MDPs, with $K = 20$ actions, with randomly generated reward $r \in [0, 1]^K$, and randomly initialized policy π_{θ_1}.



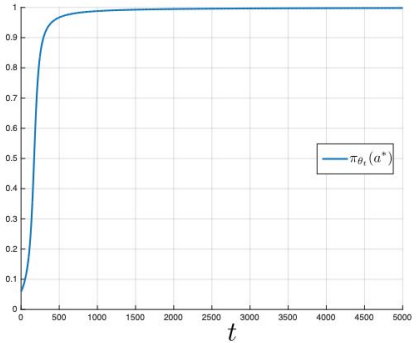
(a) $\delta_t = (\pi^* - \pi_{\theta_t})^\top r$



(b)



(c) slope ≈ -1.0005



(d)

$\log \delta_t = -\log t + C$, which is equivalent to $\delta_t = C'/t$.

Entropy regularized softmax policy gradient

Problem: $\tilde{V}^{\pi}(\rho) \coloneqq V^{\pi}(\rho) + \tau \cdot \mathbb{H}(\rho, \pi)\,,$

where $\mathbb{H}(\rho, \pi)$ is the "discounted entropy", defined as

$$\mathbb{H}(\rho, \pi) \coloneqq \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \right]. \quad (16)$$

Regularized policy gradient:

Lemma 10. *It holds that*

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_{\mu}^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \tilde{A}^{\pi_\theta}(s,a), \quad (22)$$

where $\tilde{A}^{\pi_\theta}(s,a)$ is the "soft" advantage function defined as

$$\tilde{A}^{\pi_\theta}(s,a) \coloneqq \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s)\,, \quad (23)$$

$$\tilde{Q}^{\pi_\theta}(s,a) \coloneqq r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \tilde{V}^{\pi_\theta}(s')\,. \quad (24)$$

Softmax optimal policy

Path consistency conditions:

For general MDPs, the problem is to maximize $\tilde{V}^{\pi_\theta}(\rho)$ in Eq. (20). The softmax optimal policy π_τ^* is known to satisfy the following consistency conditions (Nachum et al., 2017):

$$\pi_\tau^*(a|s) = \exp\left\{(\tilde{Q}^{\pi_\tau^*}(s,a) - \tilde{V}^{\pi_\tau^*}(s))/\tau\right\}, \tag{30}$$

$$\tilde{V}^{\pi_\tau^*}(s) = \tau \log \sum_a \exp\left\{\tilde{Q}^{\pi_\tau^*}(s,a)/\tau\right\}. \tag{31}$$

General MDPs

Main results:

Theorem 6. *Suppose $\mu(s) > 0$ for all state s. Using Algorithm 1 with the entropy regularized objective and softmax parametrization and $\eta = (1-\gamma)^3/(8+\tau(4+8\log A))$, there exists a constant $C > 0$ such that for all $t \geq 1$,*

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \left\| \frac{1}{\mu} \right\|_\infty \cdot \frac{1+\tau \log A}{(1-\gamma)^2} \cdot e^{-Ct}.$$

$$C = \frac{(1-\gamma)^4}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \min_s \mu(s) \cdot c \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} > 0$$

First convergence-rate result for entropy regularized softmax policy gradient.

General MDPs

1. Smoothness: **Lemma 14** (Smoothness). *$\mathbb{H}(\rho, \pi_\theta)$ is $(4 + 8 \log A)/(1 - \gamma)^3$-smooth, where $A := |\mathcal{A}|$ is the total number of actions.*

2. Non-uniform **Lemma 15** (Non-uniform Łojasiewicz). Suppose $\mu(s) > 0$ for all states $s \in \mathcal{S}$ and $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s, \cdot))$. Then,

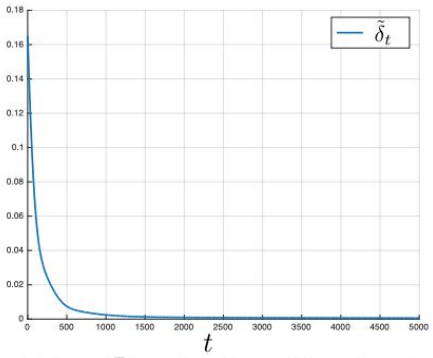
Lojasiewicz inequality

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}}.$$

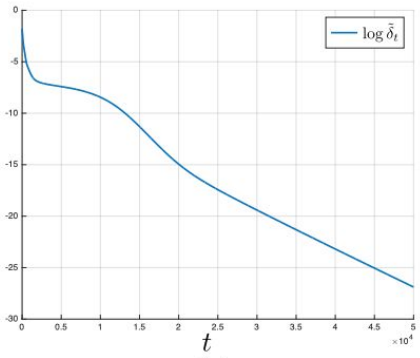
3. Minimum probability: **Lemma 16.** *Using Algorithm 1 with the entropy regularized objective, we have $c := \inf_{t \geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$.*

Verifications

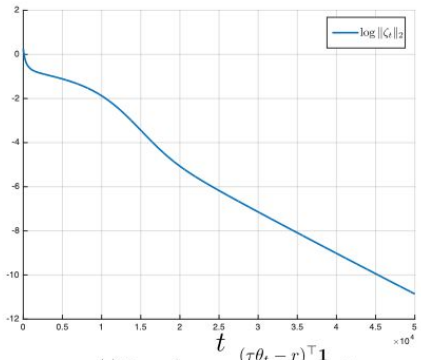
Problem: One-state MDPs, with $K = 20$ actions, with randomly generated reward $r \in [0, 1]^K$, and randomly initialized policy π_{θ_1}.



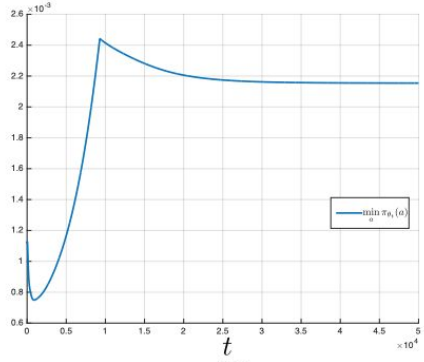
(a) $\tilde{\delta}_t = \pi_\tau^{*\top}(r - \tau \log \pi_\tau^*) - \pi_{\theta_t}{}^\top(r - \tau \log \pi_{\theta_t})$



(b)



(c) $\zeta_t = \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}$



(d)

$\log \tilde{\delta}_t = -C_1 \cdot t + C_2$, which is equivalent to $\tilde{\delta}_t = C_2' / \exp\{C_1' \cdot t\}$.

Separation of rates

Lower bound:

Theorem 8 (Lower bound). *Take any MDP. For large enough $t \geq 1$, using softmax policy gradient Algorithm 1 with $\eta_t \in (0, 1]$,*

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \geq \frac{(1-\gamma)^5 \cdot (\Delta^*)^2}{12 \cdot t}, \qquad (31)$$

where $\Delta^* := \min_{s \in \mathcal{S}, a \neq a^*(s)} \{Q^*(s, a^*(s)) - Q^*(s, a)\} > 0$ is the optimal value gap of the MDP.

First convergence-rate low bound result for softmax policy gradient.

General MDPs

Smoothness + Reversed Lojasiewicz

Lemma 28 (Reversed Łojasiewicz). *Denote* $\Delta^*(s) = Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a) > 0$ *as the optimal value gap of state* s*, where* $a^*(s)$ *is the action that the optimal policy selects under state* s*, and* $\Delta^* = \min_{s \in \mathcal{S}} \Delta^*(s) > 0$ *as the optimal value gap of the MDP. Then we have,*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \leq \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \left[V^*(\mu) - V^{\pi_\theta}(\mu) \right].$$

The bounds are matching up to constants, $O(1/t)$ and $\Omega(1/t)$.

Even with access to the true gradient, entropy helps policy gradient converge faster than any achievable rate of softmax policy gradient ascent without regularization.

Deeper reason

Non-uniform Lojasiewicz degree:

Definition 1 (Non-uniform Łojasiewicz degree). *A function* $f : \mathcal{X} \to \mathbb{R}$ *has Łojasiewicz degree* $\xi \in [0, 1]$ *if*[3]

$$\|\nabla_x f(x)\|_2 \geq C(x) \cdot |f(x) - f(x^*)|^{1-\xi}, \tag{32}$$

$\forall x \in \mathcal{X}$, *where* $C(x) > 0$ *holds for all* $x \in \mathcal{X}$.

Without regularization:

Proposition 4. *Let* $r \in [0, 1]^K$ *be arbitrary and consider* $\theta \mapsto \mathbb{E}_{a\sim\pi_\theta}[r(a)]$. *The non-uniform Łojasiewicz degree of this map with constant* $C(\theta) = \pi_\theta(a^*)$ *is zero.*

With regularization:

Proposition 5. *Fix* $\tau > 0$. *With* $C(\theta) = \sqrt{2\tau} \cdot \min_a \pi_\theta(a)$, *the Łojasiewicz degree of* $\theta \mapsto \mathbb{E}_{a\sim\pi_\theta}[r(a) - \tau \log \pi_\theta(a)]$ *is at least* $1/2$.

Summary

With true gradient:

1. Softmax policy gradient converges to optimal policy in a rate $O(1/t)$.
2. Entropy regularized softmax policy gradient converges to softmax optimal policy in a rate $O(1/e^{-t})$.
3. Softmax policy gradient follows a rate lower bound $\Omega(1/t)$.
4. Non-uniform Łojasiewicz degree as a deeper reason for rate separation.

Future Work

1. Stochastic policy gradient: similar separation of rates.

2. Function approximations, e.g., linear, over-parameterized NNs.

3. More efficient policy gradient based methods.