

# On the Global Convergence Rates of Softmax Policy Gradient Methods

Authors: Jincheng Mei ,Chenjun Xiao,Csaba Szepesv ´ari, Dale Schuurmans □

Presented by: Sadaf Raoufi

Fall2023

# Introduction:

- Analysis of the convergence rate of an iterative policy gradient methods
- In this method, Convergence towards the optimal policy improves (or the distance to the optimal policy decreases) as the number of iterations increases.

# Motivation:

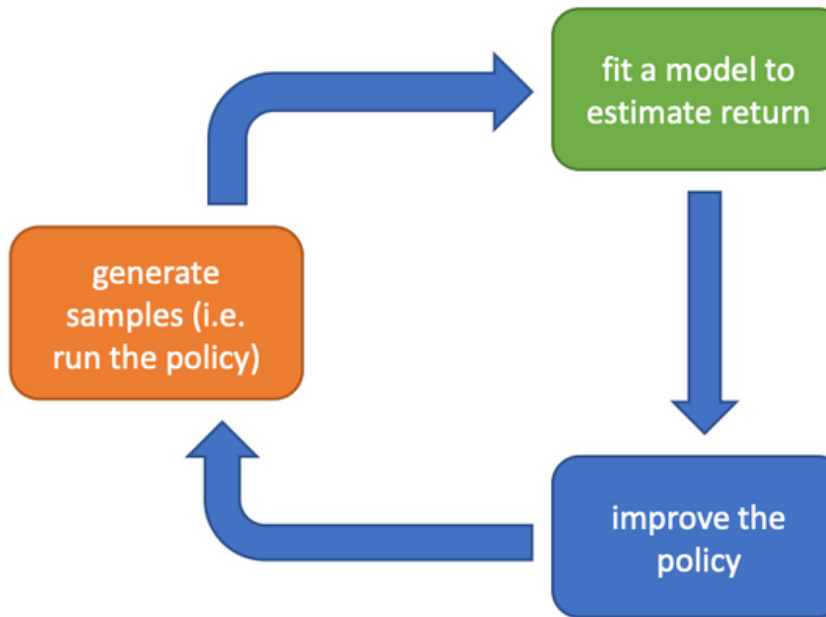
Why policy gradient methods?

- Conceptually straightforward
- Under some regularity conditions they guarantee monotonic improvement of the value
- Achieve effective empirical performance

# Basic concepts

## Policy gradient:

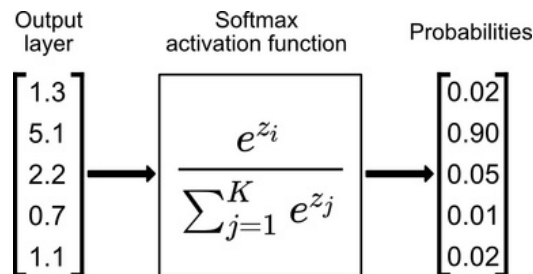
- Generate Samples
- Estimate Return
- Improve Policy
- Iterative Process



# Basic concepts:

## Softmax parameterization:

- Convert a set of numerical values (logits) into a normalized probability distribution over available actions.
- Ensures that the probabilities of all possible actions sum to one
- This allows the agent to stochastically select actions in proportion to their estimated value



# 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  $O(e^{-t})$ .
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  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ ; transition function  $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$

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:  $V^\pi(s) := \mathbb{E}_{\substack{s_0=s, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$  discount factor  $\gamma \in [0, 1)$ .

$V^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V^\pi(s)]$  where  $\rho \in \Delta(S)$  is an initial state distribution.

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

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

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



# Settings

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

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.  $\theta$  is*

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

# 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.  $\theta$ ,  $\forall(s, a)$ ,

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

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**

---

# Open problems

1. Convergence rate of softmax policy gradient was unknown.

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

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

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  $\min_s \mu(s) > 0$ .*

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

First convergence-rate result for softmax policy gradient.

**Lemma 3** (Non-uniform Łojasiewicz). *Assume  $r$  has one unique maximizing action  $a^*$ . Let  $\pi^* = \arg \max_{\pi \in \Delta} \pi^\top r$ . Then,*

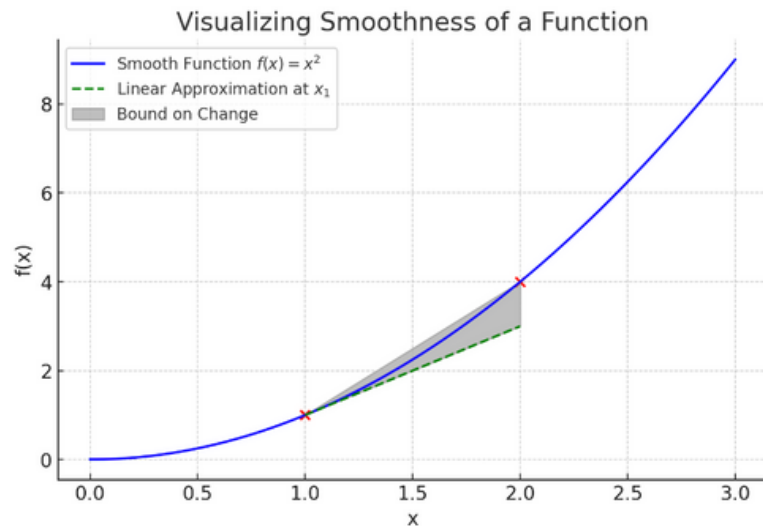
$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r. \quad (11)$$

**Smoothness.** A function  $f : \Theta \rightarrow \mathbb{R}$  with  $\Theta \subset \mathbb{R}^d$  is  $\beta$ -smooth (w.r.t.  $\ell_2$  norm,  $\beta > 0$ ) if for all  $\theta, \theta' \in \Theta$ ,

$$\left| f(\theta') - f(\theta) - \left\langle \frac{df(\theta)}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2. \quad (7)$$

# Smoothness

- a smaller  $\beta$  restricts the function to change more gradually, resulting in a tighter bound and smoother curve.
- a larger  $\beta$  permits more rapid changes, allowing the function to deviate more from the tangent line.



# Proof Idea

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

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

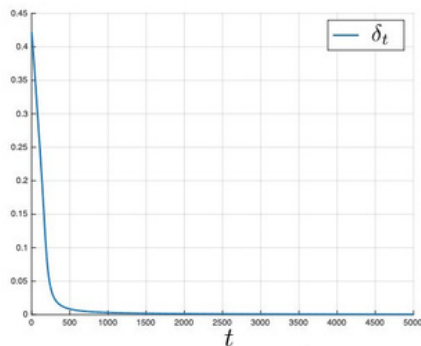
Łojasiewicz

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

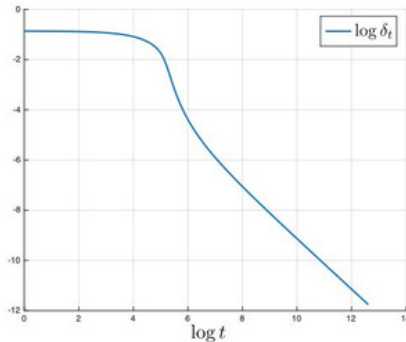
inequality: where  $a^*(s) := \arg \max_a \pi^*(a|s)$ ,  $s \in \mathcal{S}$ .

# Verifications

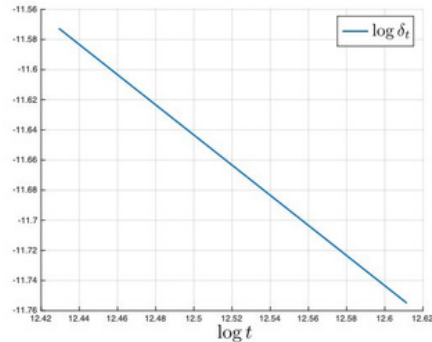
Problem: One-state MDPs, with  $K = 20$  actions, with randomly generated reward  $r \in [0, 1]^K$ , and randomly initialized policy  $\pi_{\theta_1}$ .



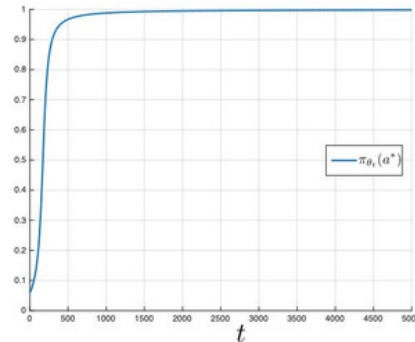
(a)  $\delta_t = (\pi^* - \pi_{\theta_t})^\top r$



(b)



(c) slope  $\approx -1.0005$



(d)

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



# Entropy regularized softmax policy gradient

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

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

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

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

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

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

# 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  $\pi_\tau^*$  is known to satisfy the following consistency conditions (Nachum et al., 2017):

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

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

# General MDPs

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

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.

# Proof Idea

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.

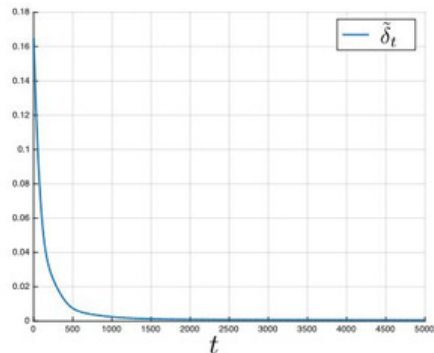
2. Non-uniform **Lemma 15** (Non-uniform Łojasiewicz). Suppose  $\mu(s) > 0$  for all states  $s \in \mathcal{S}$  and  $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s, \cdot))$ . Then,

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

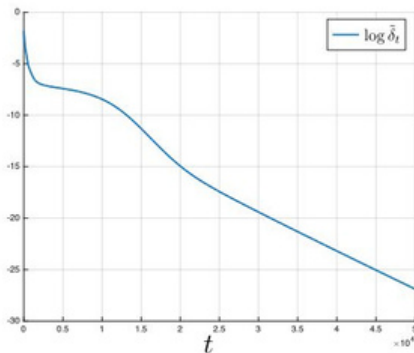
3. Minimum probability: **Lemma 16.** Using Algorithm 1 with the entropy regularized objective, we have  $c := \inf_{t \geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$ .

# Verifications

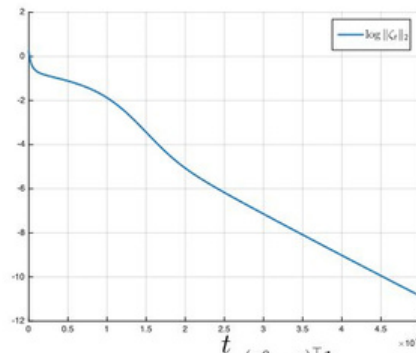
Problem: One-state MDPs, with  $K = 20$  actions, with randomly generated reward  $r \in [0, 1]^K$ , and randomly initialized policy  $\pi_{\theta_1}$ .



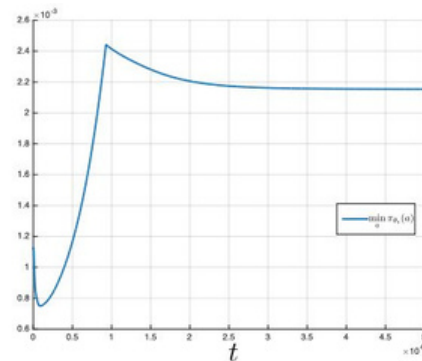
(a)  $\tilde{\delta}_t = \pi_\tau^{*\top} (r - \tau \log \pi_\tau^*) - \pi_{\theta_t}^\top (r - \tau \log \pi_{\theta_t})$



(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  $\eta_t \in (0, 1]$ ,*

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

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.

# Proof Idea

## Smoothness + Reversed Łojasiewicz

**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,*

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

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 Łojasiewicz degree: **Definition 1** (Non-uniform Łojasiewicz degree). *A function  $f : \mathcal{X} \rightarrow \mathbb{R}$  has Łojasiewicz degree  $\xi \in [0, 1]$  if<sup>8</sup>*

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

$\forall x \in \mathcal{X}$ , 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  $O(e^{-t})$ .
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.