Linear Convergence of Natural Policy Gradient Methods with Log-Linear Policies

Rui Yuan [1,4,5] Simon S. Du [1,3] Robert M. Gower [2] Alessandro Lazaric [1] Lin Xiao [1]

[1]FAIR, Meta AI [2]CCM, Flatiron Institute [3]University of Washington [4]LTCI, Télécom Paris [5]Institut Polytechnique de Paris

Overview

Log-linear policy parametrization:

$$\pi_{s,a}(\theta) \;=\; \frac{\exp(\phi_{s,a}^\top \theta)}{\sum_{a' \in \mathcal{A}} \exp(\phi_{s,a'}^\top \theta)}.$$

Objective:

$$\min_{\theta \in \mathbb{R}^m} V_\rho(\theta) \;=\; \mathbb{E}_{\substack{s_0 \sim \rho, a_t \sim \pi_{s_t}(\theta) \\ s_{t+1} \sim \mathcal{P}(\cdot | s_t, a_t)}} \left[\sum_{t=0}^\infty \gamma^t c(s_t, a_t) \right].$$

Define the Q-function $Q(\theta)$ and the advantage function $A(\theta)$.

Natural Policy Gradient (NPG) Method:

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k F_\rho\big(\theta^{(k)}\big)^\dagger \nabla_\theta V_\rho\big(\theta^{(k)}\big), \tag{1}$$

where $\nabla_\theta V_\rho\big(\theta^{(k)}\big)$ is the policy gradient, $F_\rho(\theta)$ is the Fisher information matrix and d^θ is the state visitation distribution:

$$\nabla_\theta V_\rho(\theta) \;=\; \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^\theta, a \sim \pi_s(\theta)} \left[A_{s,a}(\theta) \nabla_\theta \log \pi_{s,a}(\theta) \right],$$

$$F_\rho(\theta) \;\stackrel{\text{def}}{=}\; \mathbb{E}_{s \sim d^\theta, a \sim \pi_s(\theta)} \left[\nabla_\theta \log \pi_{s,a}(\theta) \big(\nabla_\theta \log \pi_{s,a}(\theta) \big)^\top \right],$$

$$d_s^\theta \;\stackrel{\text{def}}{=}\; (1-\gamma)\, \mathbb{E}_{s_0 \sim \rho} \left[\sum_{t=0}^\infty \gamma^t \Pr^{\pi(\theta)}(s_t = s \mid s_0) \right].$$

NPG with Compatible Function Approximation

We define the compatible function approximation error as

$$L_A(w, \theta, \zeta) \;\stackrel{\text{def}}{=}\; \mathbb{E}_{(s,a) \sim \zeta} \left[\big(w^\top \nabla_\theta \log \pi_{s,a}(\theta) - A_{s,a}(\theta) \big)^2 \right].$$

The NPG update (1) is equivalent to

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w_\star^{(k)}, \qquad w_\star^{(k)} \in \arg\min_{w \in \mathbb{R}^m} L_A\big(w, \theta^{(k)}, d^{(k)}\big),$$

where $d^{(k)}$ is the state-action visitation distribution

$$d_{s,a}^{(k)} \;\stackrel{\text{def}}{=}\; d_s^\theta \pi_{s,a}^{(k)} \;=\; (1-\gamma)\, \mathbb{E}_{s_0 \sim \rho} \left[\sum_{t=0}^\infty \gamma^t \Pr^{\pi^{(k)}}(s_t = s, a_t = a \mid s_0) \right].$$

Consider a more general state-action visitation distribution

$$\tilde{d}_{s,a}^{(k)} \;\stackrel{\text{def}}{=}\; (1-\gamma)\, \mathbb{E}_{(s_0,a_0) \sim \nu} \left[\sum_{t=0}^\infty \gamma^t \Pr^{\pi^{(k)}}(s_t = s, a_t = a \mid s_0, a_0) \right],$$

we proceed the following approximate NPG update rule:

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \arg\min_w L_A\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big).$$

We can define the compatible function approximation error as

$$L_Q(w, \theta, \zeta) \;\stackrel{\text{def}}{=}\; \mathbb{E}_{(s,a) \sim \zeta} \left[\big(w^\top \phi_{s,a} - Q_{s,a}(\theta) \big)^2 \right]$$

and derive a variant of the approximate NPG update called Q-NPG:

$$\theta^{(k+1)} \;=\; \theta^{(k)} - \eta_k w^{(k)}, \qquad w^{(k)} \approx \arg\min_w L_Q\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big).$$

NPG as Policy Mirror Descent

Given $w^{(k)}$ an approximate solution for minimizing $L_A\big(w, \theta^{(k)}, \tilde{d}^{(k)}\big)$, we can write the approximate NPG as a mirror descent update:

$$\pi_s^{(k+1)} = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ \eta_k \left\langle \bar{\Phi}_s^{(k)} w^{(k)}, p \right\rangle + \mathrm{KL}(p, \pi_s^{(k)}) \right\}, \quad \forall s \in \mathcal{S}, \tag{2}$$

where $\bar{\Phi}_s^{(k)} \in \mathbb{R}^{|\mathcal{A}| \times m}$ is a matrix whose rows consist of the *centered feature maps* $\bar{\phi}_{s,a}(\theta^{(k)})$ defined as

$$\nabla_\theta \log \pi_{s,a}(\theta) \;=\; \bar{\phi}_{s,a}(\theta) \;\stackrel{\text{def}}{=}\; \phi_{s,a} - \mathbb{E}_{a' \sim \pi_s(\theta)} \left[\phi_{s,a'} \right].$$

NPG Algorithm

Algorithm 1 NPG: Natural Policy Gradient

Input: initial state-action distribution ν, policy $\pi^{(0)}$, step size $\eta_0 > 0$, `NPG-SGD` for minimizing $L_A(w, \theta, \tilde{d}^\theta)$

For $k = 0, \ldots, K-1$ do:

  Call `NPG-SGD` to obtain $w^{(k)}$

  Update $\theta^{(k+1)} = \theta^{(k)} - \eta_k w^{(k)}$ and η_k

Output: last iterate $\pi^{(K)}$

Convergence analysis (1/2)

Two key ingredients:

▶ Performance difference lemma [2] : For any policy $\pi(\theta), \pi(\theta')$,

$$V_\rho(\theta) - V_\rho(\theta') = \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim \bar{d}^\theta} \left[A_{s,a}(\theta') \right].$$

▶ Three-point descent lemma [1, 3] : Suppose that $\mathcal{C} \subset \mathbb{R}^m$ is a closed convex set, $f: \mathcal{C} \to \mathbb{R}$ is a proper, closed convex function, $D_h(\cdot,\cdot)$ is the Bregman divergence generated by a function h of Lengendre type and $\mathrm{rint}\,\mathrm{dom}\,h \cap \mathcal{C} \neq \emptyset$. For any $x \in \mathrm{rint}\,\mathrm{dom}\,h$, let

$$x^+ \in \arg\min_{u \,\in\, \mathrm{dom}\,h \,\cap\, \mathcal{C}} \{ f(u) + D_h(u, x) \}.$$

Then $x^+ \in \mathrm{rint}\,\mathrm{dom}\,h \cap \mathcal{C}$ and for any $u \in \mathrm{dom}\,h \cap \mathcal{C}$,

$$f(x^+) + D_h(x^+, x) \le f(u) + D_h(u, x) - D_h(u, x^+).$$

Decompose the compatible function approximation error as

$$L_A\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) \;=\; \underbrace{L_A\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) - L_A\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)}_{\text{Statistical error (excess risk)}} + \underbrace{L_A\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big)}_{\text{Approximation error}}.$$

Convergence analysis (2/2)

Define the *distribution mismatch coefficient* of p relative to q as

$$\left\| \frac{p}{q} \right\|_\infty \;\stackrel{\text{def}}{=}\; \max_{s \in \mathcal{S}} \frac{p_s}{q_s}.$$

Let π^* be an arbitrary *comparator policy*. We define

$$\vartheta_\rho \;\stackrel{\text{def}}{=}\; \frac{1}{1-\gamma} \left\| \frac{d^{\pi^*}}{\rho} \right\|_\infty \;\ge\; \frac{1}{1-\gamma}.$$

Assumption 1: There exists $\epsilon_{\text{stat}}, \epsilon_{\text{approx}} > 0$ such that for all iterations $k \ge 0$ of the NPG method (2), we have

$$\mathbb{E}\left[L_A\big(w^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) - L_A\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) \right] \;\le\; \epsilon_{\text{stat}},$$

$$\mathbb{E}\left[L_A\big(w_\star^{(k)}, \theta^{(k)}, \tilde{d}^{(k)}\big) \right] \;\le\; \epsilon_{\text{approx}}.$$

Assumption 2: There exists $C_\nu < \infty$ such that for all iterations $k \ge 0$ of the NPG method (2), we have

$$\mathbb{E}_{(s,a) \sim \tilde{d}^{(k)}} \left[\left(\frac{d_{s,a}^{(k+1)}}{\tilde{d}_{s,a}^{(k)}} \right)^2 \right] \;\le\; C_\nu \quad \text{and} \quad \mathbb{E}_{(s,a) \sim \tilde{d}^{(k)}} \left[\left(\frac{\bar{d}_{s,a}^{\pi^*}}{\tilde{d}_{s,a}^{(k)}} \right)^2 \right] \;\le\; C_\nu.$$

Theorem: Fix a state distribution ρ, a state-action distribution ν, and a comparator policy π^*. We consider the NPG method (2) with the step sizes satisfying $\eta_{k+1} \ge \frac{1}{\gamma} \eta_k$. Then for all $k \ge 0$,

$$\mathbb{E}\left[V_\rho(\pi^{(k)}) \right] - V_\rho(\pi^*) \le \left(1 - \frac{1}{\vartheta_\rho} \right)^k \frac{2}{1-\gamma} + \frac{\sqrt{C_\nu}(\vartheta_\rho + 1)}{1-\gamma} \left(\sqrt{\epsilon_{\text{stat}}} + \sqrt{\epsilon_{\text{approx}}} \right).$$

Corollary: By further assuming that the feature maps are bounded and the Fisher information matrix is non-degenerate, we obtain an $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity for NPG.

Remark: Similar linear convergence and $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity results can be established for Q-NPG.

Take-away

We show that both NPG and Q-NPG with log-linear policies enjoy **linear convergence rates** and $\boldsymbol{\mathcal{O}(1/\epsilon^2)}$ sample complexities using a simple, **non-adaptive geometrically increasing** step size.

References

[1] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[2] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of 19th International Conference on Machine Learning*, pages 267–274, 2002.

[3] Lin Xiao. On the convergence rates of policy gradient methods, 2022.