Linear convergence for natural policy gradient with general policy parametrizations

Carlo Alfano

UNIVERSITY OF OXFORD

4th IMA Conference on The Mathematical Challenges of Big Data
21/09/2022

Joint work with Patrick Rebeschini

Setting

A (finite) Markov Decision Process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \nu)$ is specified by:

- a finite state space $\mathcal{S}$;
- a finite action space $\mathcal{A}$;
- a transition model $P(s'|s, a)$;
- a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$;
- a discount factor $\gamma \in (0, 1)$;
- a starting state distribution ν over $\mathcal{S}$
- a policy π.

Objective of Reinforcement Learning: find $\pi^\star \in \mathrm{argmax}_\pi V^\pi(\nu)$, where V^π is the value function of policy π.

Policies

Objective: find $(\pi_t)_{t\geqslant 1}$ such that $\exists \alpha > 0, \beta \geqslant 0$

$$V^{\pi^\star}(\nu) - V^{\pi_t}(\nu) \leqslant e^{-\alpha t} + \beta.$$

Stochastic policy: $\pi : \mathcal{S} \to \Delta(\mathcal{A})$.
General function approximation:

$$\pi_\theta(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(f_\theta(s,a'))}$$

$\to$ e.g. Tabular policies: $f_\theta(s,a) = \theta_{s,a}$.

$\to$ e.g. Log-linear policies: $f_\theta(s,a) = \langle \theta, \phi(s,a) \rangle$ for $\theta, \phi \in \mathbb{R}^d$.

Policy Gradient

Value functions:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\Big|\pi, s_0=s, a_0=a\right]$$

$$V^\pi(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}Q^\pi(s,a)$$

$$V^\pi(\nu) := \mathbb{E}_{s\sim\nu}\left[V^\pi(s)\right]$$

Policy gradient:

$$\nabla_\pi V^\pi(\nu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\nu^\pi, a\sim\pi_\theta(\cdot|s)}\left[Q^{\pi_\theta}(s,a)\right]$$

where

$$d_\nu^\pi(s) := (1-\gamma)\mathbb{E}_{s\sim\nu}\sum_{t=0}^{\infty}\gamma^t Pr^\pi(s_t=s|s_0=s)$$

is the discounted visitation distribution of policy π.

Natural Policy Gradient

Mirror Descent update:

$$\pi_{\theta^{t+1}}$$

$$= \operatorname{argmax}_{\pi_\theta} \left\{ \sum_{s \in \mathcal{S}} d_\nu^t(s) \left(\langle Q^t(s, \cdot), \pi_\theta(\cdot|s) \rangle - D_h(\pi_\theta(\cdot|s), \pi_{\theta^t}(\cdot|s)) \right) \right\}$$

If $h = \sum_a \pi(a|s) \log \pi(a|s)$, then $\pi_{\theta^{t+1}} \propto \pi_{\theta^t} e^{\eta_t Q^t(s,a)}$.

On the Theory of Policy Gradient Methods: Optimality, Approximation and Distribution Shift. A. Agarwal, S. M. Kakade, J. D. Lee, G. Mahajan (2021). J. Mach. Learn. Res.

Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization. S. Cen, C. Cheng, Y. Chen, Y. Wei, Y. Chi (2021). Operation Research

Latest result

On the convergence rates of policy gradient methods. L. Xiao (2022). arXiv:2201.07443

Theorem

Assume $\left\| Q^\pi - \widehat{Q}^\pi \right\|_\infty \leqslant \tau$.

Let π *be a tabular policy.*

Let $\eta_{t+1} \geqslant \eta_t/\gamma$ *and* $\delta_\nu = \frac{1}{1-\gamma} \left\| \frac{d_\nu^{\pi^\star}}{\nu} \right\|_\infty$, *then*

$$V^\star(\nu) - V^{\pi_{\theta^T}}(\nu) \leqslant \left(1 - \frac{1}{\delta_\nu}\right)^T \left(\frac{1}{1-\gamma} + \frac{\mathtt{KL}_0^\star}{(1-\gamma)\eta_0(\delta_\nu - 1)}\right) + \frac{4\delta_\nu}{1-\gamma}\tau$$

Estimating Q^π

Find $\widehat{Q}^\pi$ close to Q^π. Given a feature function $\phi(s,a) \in \mathbb{R}^d$, assume that $\exists\, w^t$ and that we can find a $\widehat{w}^t$ such that

$$Q^\pi - \widehat{Q}^\pi = \big(Q^\pi - \langle w^t, \phi(s,a)\rangle\big) + \big(\langle w^t, \phi(s,a)\rangle - \langle \widehat{w}^t, \phi(s,a)\rangle\big)$$

is bounded. In particular, we assume:

- $\mathbb{E}_{s\sim d_\nu^\star, a\sim \mathsf{Unif}_\mathcal{A}}\left[\big(Q^t(s,a) - \langle w^t, \phi(s,a)\rangle\big)^2\right] \leqslant \varepsilon_{\mathsf{bias}}$
- $\mathbb{E}_{s\sim d_\nu^t, a\sim \mathsf{Unif}_\mathcal{A}}\left[\big(\langle w^t - \widehat{w}^t, \phi(s,a)\rangle\big)^2\right] \leqslant \varepsilon_{\mathsf{stat}}$.

Once we find $\widehat{w}^t$, the update for the log-linear policy class becomes $\theta^{t+1} = \theta^t + \eta_t \widehat{w}^t$.

Main result

Theorem

Under the assumptions on the previous slide, let $\eta_{t+1} \geqslant \eta_t/\gamma$ *and* $\delta_\nu = \frac{1}{1-\gamma} \left\| \frac{d_\nu^{\pi^\star}}{\nu} \right\|_\infty$, *then*

$$V^\star(\nu) - V^{\pi_{\theta^T}}(\nu) \leqslant \left(1 - \frac{1}{\delta_\nu}\right)^T \left(\frac{1}{1-\gamma} + \frac{\mathtt{KL}_0^\star}{(1-\gamma)\eta_0(\delta_\nu - 1)} \right) + 4\delta_\nu \sqrt{\frac{|\mathcal{A}|\kappa}{(1-\gamma)^3}(\varepsilon_{\mathit{stat}} + \varepsilon_{\mathit{bias}})}$$

- First result on linear convergence of unregularized natural policy gradient for general policy parametrization.
- Improved sample complexity with respect to the tabular case.