Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes

Alekh Agarwal* Sham M. Kakade† Jason D. Lee‡ Gaurav Mahajan§

【强化学习 98】PG Theory Summary

张楚珩
清华大学 交叉信息院博士在读

这篇文章太经典了，今天在想某个问题的时候需要用到，由推导了一遍，做了个总结。

原文传送门

Agarwal, Alekh, et al. "Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes." arXiv preprint arXiv:1908.00261 (2019).

正文

Algorithm	Error	#Iterations
Direct parametrization w/ projected gradient ascent $\pi^{(t+1)} = P_{\Delta(A)^{|S|}}(\pi^{(t)} + \eta\nabla_\pi V^{(t)}(\mu))$	$\min_{t<T}\{V^*(\rho) - V^{(t)}(\rho)\} \le \epsilon$ (Theorem 4.2)	$\frac{64\gamma\|S\|\|A\|}{(1-\gamma)^6\epsilon^2}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty^2$
Softmax parametrization (w/o regularization) $\theta^{(t+1)} = \theta^{(t)} + \eta\nabla_\theta V^{(t)}(\mu)$	Only asymptotic convergence (Theorem 5.1)	May need exponential number of iterations
Softmax parametrization w/ relative entropy regularization $L_\lambda(\theta) = V^{\pi_\theta}(\mu) + \frac{\lambda}{\|S\|\|A\|}\sum_{s,a}\log\pi_\theta(a\|s)$ $\theta^{(t+1)} = \theta^{(t)} + \eta\nabla_\theta L_\lambda(\theta^{(t)})$	$\min_{t<T}\{V^*(\rho) - V^{(t)}(\rho)\} \le \epsilon$ (Corollary 5.4)	$\frac{320\|S\|^2\|A\|^2}{(1-\gamma)^6\epsilon^2}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty^2$
NPG with softmax parametrization $\theta^{(t+1)} = \theta^{(t)} + \eta F_\rho(\theta^{(t)})^{-1}\nabla_\theta V^{(t)}(\rho)$ $\Leftrightarrow \theta_{s,a}^{(t+1)} = \theta_{s,a}^{(t)} + \frac{\eta}{1-\gamma}A^{(t)}(s,a)$	$V^*(\rho) - V^{(T)}(\rho) \le \epsilon$ (Theorem 5.7)	$\frac{2}{(1-\gamma)^2\epsilon}$
Unconstrained approximate NPG $\theta^{(t+1)} = \theta^{(t)} + \eta\omega^{(t)}$ ($\omega^{(t)}$ is sample based NPG) $\Leftrightarrow \omega^{(t)} \in \arg\min_\omega L_\nu(\omega;\theta)$ $L_\nu(\omega;\theta) = \mathbb{E}_{s,a\sim\nu}[(A^{\pi_\theta}(s,a) - \omega\cdot\nabla_\theta\log\pi_\theta(a\|s))^2]$ $\nu(s,a) = d_\rho^{\pi_\theta}(s)\pi_\theta(a\|s)$	$\min_{t<T}\{V^{\pi}(\rho) - V^{(t)}(\rho)\} \le \epsilon + \frac{\sqrt{\bar\epsilon_{approx}}}{1-\gamma}$ where $\overline{L_\nu(\omega^{(t)},\theta^{(t)})} \le \bar\epsilon_{approx}$ (Theorem 6.4)	$\frac{2\beta W^2\log\|A\|}{(1-\gamma)^2\epsilon^2}$ where β smooth, $\|\omega^{(t)}\|_2 \le W$
Unconstrained approximate NPG $\nu(s,a) = d_{\nu_0}^{\pi_\theta}(s)\pi_\theta(a\|s)$ $\omega^{(t)}$ can achieve $\arg\min_\omega L_\nu(\omega;\theta)$	$\min_{t<T}\{V^*(\rho) - V^{(t)}(\rho)\} \le \epsilon + \sqrt{\frac{1}{(1-\gamma)^3}\left\|\frac{\nu^*}{\nu_0}\right\|_\infty\epsilon_{approx}}$ where $L_\nu^*(\theta^{(t)}) = \min_\omega L_\nu(\omega;\theta^{(t)}) \le \epsilon_{approx}$ (Corollary 6.5)	$\frac{2\beta W^2\log\|A\|}{(1-\gamma)^2\epsilon^2}$
Unconstrained approximate NPG $\nu(s,a) = d_{\nu_0}^{\pi_\theta}(s)\pi_\theta(a\|s)$ Consider estimation error for $\min_\omega L_\nu(\omega;\theta)$	$\min_{t<T}\{V^*(\rho) - V^{(t)}(\rho)\} \le \epsilon + \sqrt{\frac{1}{(1-\gamma)^3}\left\|\frac{\nu^*}{\nu_0}\right\|_\infty}(\sqrt{\epsilon_{approx}} + \frac{\epsilon_1}{\sqrt{N}})$ (Corollary 6.5)	$\frac{2\beta\tilde W^2\log\|A\|}{(1-\gamma)^2\epsilon^2}$ where $\|\tilde\omega^{(t)}\|_2 \le \tilde W$
Projected Policy Gradient for constrained policy class $\theta^{(t+1)} = P_\Theta(\theta^{(t)} + \eta\nabla_\theta V^{(t)}(\nu))$	$\min_{t<T}\{V^*(\rho) - V^{(t)}(\rho)\} \le (W^* + 1)\epsilon + \frac{1}{(1-\gamma)^3}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty\epsilon_{approx}$ where $L_{BPE}^*(\theta^{(t)}) =\le \epsilon_{approx}$ and $\|\omega^{(t)}\|_2 \le W^*$ (Corollary 6.14)	$\frac{8\beta}{(1-\gamma)^3\epsilon^2}\left\|\frac{d_\rho^{\pi^*}}{\mu}\right\|_\infty$

Direct parametrization

Direct parametrization with projected gradient ascent

$$\pi^{(t+1)} = P_{\Delta(A)^{|S|}}(\pi^{(t)} + \eta \nabla_\pi V^{(t)}(\mu))$$

Theorem 4.2. *The projected gradient ascent algorithm (9) on $V^\pi(\mu)$ with stepsize $\eta = \frac{(1-\gamma)^3}{2\gamma|\mathcal{A}|}$ satisfies for all distributions $\rho \in \Delta(\mathcal{S})$,*

$$\min_{t<T} \left\{ V^\star(\rho) - V^{(t)}(\rho) \right\} \le \epsilon \quad \textit{whenever} \quad T > \frac{64\gamma|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^6\epsilon^2} \left\| \frac{d_\rho^{\pi^\star}}{\mu} \right\|_\infty^2 .$$

Theorem 4.2 proof sketch

1. Smoothness of $V^\pi(\mu)$ with $\beta = \frac{2\gamma|A|}{(1-\gamma)^3}$ (Lemma E3)
2. #samples to magnitude of the update (projected gradient ascent, Bech, 2017)
3. Small magnitude of the update to small gradient (projection, Ghadimi and Lan, 2016)
4. Small gradient to optimality (Lemma 4.1)

Softmax parametrization w/ relative entropy regularization

$$L_\lambda(\theta) = V^{\pi_\theta}(\mu) + \frac{\lambda}{|S||A|}\sum_{s,a} \log \pi_\theta(a|s)\,, \qquad \theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta L_\lambda(\theta^{(t)})$$

Corollary 5.4. *(Iteration complexity with relative entropy regularization) Let $\beta_\lambda := \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|\mathcal{S}|}$. Starting from any initial $\theta^{(0)}$, consider the updates (13) with $\lambda = \frac{\epsilon(1-\gamma)}{2\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty}$ and $\eta = 1/\beta_\lambda$. Then for all starting state distributions ρ, we have*

$$\min_{t<T}\left\{V^\star(\rho) - V^{(t)}(\rho)\right\} \le \epsilon \quad \textit{whenever} \quad T \ge \frac{320|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^6\,\epsilon^2}\left\|\frac{d_\rho^{\pi^\star}}{\mu}\right\|_\infty^2 .$$

Corollary 5.4 proof sketch

1. Smoothness of $L_\lambda(\theta)$ with $\beta_\lambda = \frac{8\gamma}{(1-\gamma)^3} + \frac{2\lambda}{|S|}$(Lemma E4)
2. #samples to small gradient (unconstrained gradient ascent, Ghadimi and Lan, 2013)
3. Small gradient to optimality (Theorem 5.3)

Natural policy gradient with softmax parametrization and exact gradient

$$F_\rho(\theta) = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[(\nabla_\theta \log \pi_\theta(a|s))\,(\nabla_\theta \log \pi_\theta(a|s))^T\right] \quad \theta^{(t+1)} = \theta^{(t)} + \eta F_\rho\left(\theta^{(t)}\right)^{-1}\nabla_\theta V^{(t)}(\rho)$$

Theorem 5.7 (Global convergence for Natural Policy Gradient Ascent). *Suppose we run the NPG updates (15) using $\rho \in \Delta(\mathcal{S})$ and with $\theta^{(0)} = 0$. Fix $\eta > 0$. For all $T > 0$, we have:*

$$V^{(T)}(\rho) \ge V^\star(\rho) - \frac{\log|\mathcal{A}|}{\eta T} - \frac{1}{(1-\gamma)^2 T}.$$

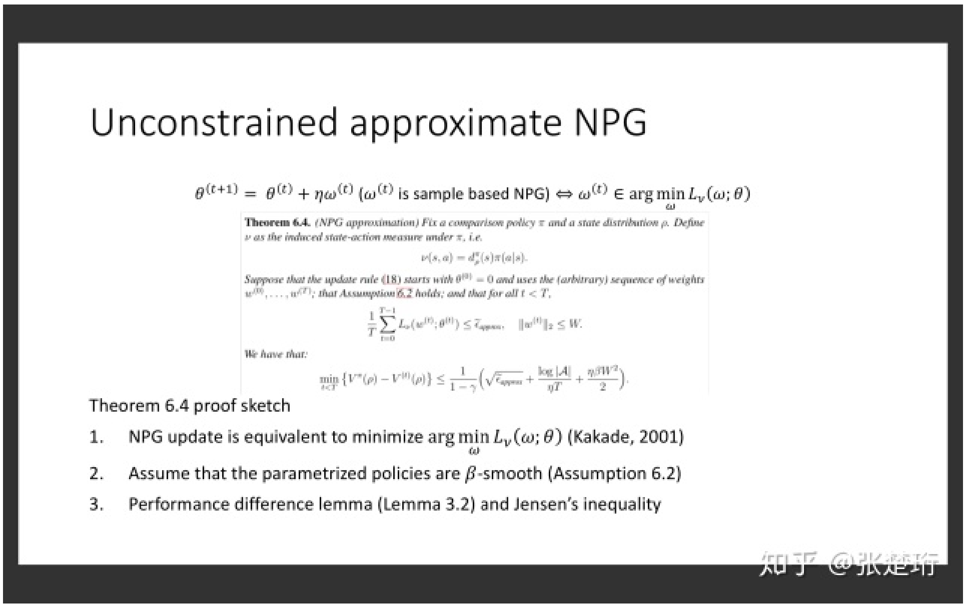
In particular, setting $\eta \ge (1-\gamma)^2\log|\mathcal{A}|$, we see that NPG finds an ϵ-optimal policy in a number of iterations that is at most:

$$T \le \frac{2}{(1-\gamma)^2\epsilon},$$

Theorem 5.7 proof sketch

1. NPG update has a simple form under softmax parametrization (Lemma 5.6)
2. Improvement lower bound exists in terms of the partition function (Lemma 5.8)

NPG gives large weights to rare situations, and therefore the algorithm does not need μ and the sample complexity does not involve distribution mismatch coefficient using exact gradients.


Unconstrained approximate NPG
$\theta^{(t+1)} = \theta^{(t)} + \eta\omega^{(t)}$ ($\omega^{(t)}$ is sample based NPG) $\Leftrightarrow \omega^{(t)} \in \arg\min_{\omega} L_{\nu}(\omega;\theta)$
Theorem 6.4. *(NPG approximation) Fix a comparison policy π and a state distribution ρ. Define ν as the induced state-action measure under π, i.e.*
$\nu(s,a) = d_{\rho}^{\pi}(s)\pi(a|s).$
Suppose that the update rule (18) starts with $\theta^{(0)} = 0$ and uses the (arbitrary) sequence of weights $w^{(0)}, \dots, w^{(T)}$; that Assumption 6.2 holds; and that for all $t < T$,
$\frac{1}{T}\sum_{t=0}^{T-1} L_{\nu}(w^{(t)};\theta^{(t)}) \le \bar{\epsilon}_{\text{approx}}, \quad \|w^{(t)}\|_2 \le W.$
We have that:
$\min_{t<T}\{V^{\pi}(\rho) - V^{(t)}(\rho)\} \le \frac{1}{1-\gamma}\left(\sqrt{\bar{\epsilon}_{\text{approx}}} + \frac{\log|\mathcal{A}|}{\eta T} + \frac{\eta\beta W^2}{2}\right).$
Theorem 6.4 proof sketch
1. NPG update is equivalent to minimize $\arg\min_{\omega} L_{\nu}(\omega;\theta)$ (Kakade, 2001)
2. Assume that the parametrized policies are β-smooth (Assumption 6.2)
3. Performance difference lemma (Lemma 3.2) and Jensen's inequality
知乎 @张楚珩