A general sample complexity analysis of vanilla policy gradient

Rui Yuan[1,2], Robert M. Gower[2], Alessandro Lazaric[1] [1] Facebook AI Research [2] LTCI, Télécom Paris, Institut Polytechnique de Paris

Summary

Objective: $\max_{\theta\in\mathbb{R}^d} J(\theta) = \mathbb{E}_{\tau\sim p(\cdot|\pi_\theta,\mathcal{M})}[\mathcal{R}(\tau)] = \mathbb{E}_{\tau\sim p(\cdot|\theta)}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\right]$

Empirical estimated policy gradient by sampling m truncated trajectories τ_i:

REINFORCE $\quad \widehat{\nabla}_m J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\gamma^t\mathcal{R}(s_t^i,a_t^i)\cdot\sum_{t'=0}^{H-1}\nabla_\theta\log\pi_\theta(a_{t'}^i\mid s_{t'}^i)$

GPOMDP $\quad \widehat{\nabla}_m J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H-1}\left(\sum_{k=0}^{t}\nabla_\theta\log\pi_\theta(a_k^i\mid s_k^i)\right)\gamma^t\mathcal{R}(s_t^i,a_t^i)$

Vanilla policy gradient (**REINFORCE**, **GPOMDP**): $\theta_{k+1}=\theta_k+\eta\widehat{\nabla}_m J(\theta_k)$

▶ *Question:* What is the sample complexity (i.e. number of single-step interactions with the environment) of vanilla policy gradient ?

▶ *Contribution:* Recover existing $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity guarantees for vanilla policy gradient (**REINFORCE** and **GPOMDP**) with *weaker* assumptions for *wider* ranges of parameters (e.g. mini-batch $m=1$).

ABC assumption

▶ Assumption 1 (smoothness)

$$|J(\theta')-J(\theta)-\langle\nabla J(\theta),\theta'-\theta\rangle| \le \frac{L}{2}\|\theta'-\theta\|^2$$

▶ Assumption 2 (ABC, [Khaled and Richtarik, 2020])

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \le 2A(J^*-J(\theta))+B\|\nabla J_H(\theta)\|^2+C,$$

$J_H(\theta)=\mathbb{E}_\tau\left[\sum_{t=0}^{H-1}\gamma^t\mathcal{R}(s_t,a_t)\right]$ is the expected truncated total reward.

▶ Assumption 3

$$|\langle\nabla J_H(\theta),\nabla J_H(\theta)-\nabla J(\theta)\rangle| \le D\gamma^H$$
$$\|\nabla J_H(\theta)-\nabla J(\theta)\| \le D'\gamma^H$$

Existing assumptions are special cases

▶ *Bounded variance* [Ghadimi and Lan, 2013] $(A=0,B=1)$:

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \le \|\nabla J(\theta)\|^2+C$$

▶ *Expected strong growth* [Vaswani et al., 2019] $(A=C=0)$:

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \le B\|\nabla J(\theta)\|^2$$

▶ *Relaxed growth condition* [Bottou et al., 2018] $(A=0)$:

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \le B\|\nabla J(\theta)\|^2+C$$

▶ *Sure-smoothness* [Lei et al., 2020] , *Gradient confusion* [Sankararaman et al., 2020] , etc.

General convergence analysis of policy gradient under ABC assumption

Proposition 1. *Suppose that Assumption 1, 2 and 3 are satisfied. We choose a constant stepsize η such that $\eta\in\left(0,\frac{2}{LB}\right)$ where B can be zero. Let $\delta_0\overset{def}{=}J^*-J(\theta_0)$. If $A>0$, then policy gradient satisfies*

$$\min_{0\le t\le T-1}\mathbb{E}\left[\|\nabla J(\theta_t)\|^2\right] \le \frac{2\delta_0(1+L\eta^2A)^T}{\eta T(2-LB\eta)}+\frac{LC\eta}{2-LB\eta}+\left(\frac{2D(3-LB\eta)}{2-LB\eta}+D'^2\gamma^H\right)\gamma^H.$$

If $A=0$, we have

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \le \frac{2\delta_0}{\eta T(2-LB\eta)}+\frac{LC\eta}{2-LB\eta}+\left(\frac{2D(3-LB\eta)}{2-LB\eta}+D'^2\gamma^H\right)\gamma^H,$$

where θ_U is uniformly sampled from $\{\theta_0,\theta_1,\cdots,\theta_{T-1}\}$.

☞ Mini-batch $m=1$, step size $\eta=\min\left\{\frac{1}{\sqrt{LAT}},\frac{1}{LB},\frac{\epsilon}{2LC}\right\}$, number of iterations $T\ge\frac{12\delta_0 L}{\epsilon^2}\max\left\{B,\frac{12\delta_0 A}{\epsilon^2},\frac{2C}{\epsilon^2}\right\}$, horizon $H=\mathcal{O}(\log\epsilon^{-1})$, sample complexity: $TH=\widetilde{\mathcal{O}}(\epsilon^{-4})$ [Liu et al., 2020]

☞ For the full exact gradient $(A=C=D=D'=0,B=1)$: $T=\mathcal{O}(\epsilon^{-2})$ [Agarwal et al., 2021]

Lipschitz and smooth policy assumptions

▶ Assumption 4 (Lipschitz and smooth policy, [Xu et al., 2020])

$$\|\nabla_\theta\log\pi_\theta(a\mid s)\| \le G, \qquad \left\|\nabla^2_\theta\log\pi_\theta(a\mid s)\right\| \le F$$

Under Asssumption 4, we have Asssumption 3 holds and :

▷ Assumption 1 holds with

$$L = \frac{2G^2\mathcal{R}_{\max}}{(1-\gamma)^3}+\frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}.$$

⚠ The smoothness constant L is different to the one in [Xu et al., 2020] which is $\frac{F\mathcal{R}_{\max}}{(1-\gamma)^2}$.

▷ Assumption 2 holds with

$$\mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right] \le \underbrace{\left(1-\frac{1}{m}\right)}_{=B}\|\nabla J_H(\theta)\|^2+\underbrace{\frac{\Gamma_g^2}{m}}_{=C},$$

where $\Gamma_g=\frac{HG\mathcal{R}_{\max}}{1-\gamma}$ when using **REINFORCE** or $\Gamma_g=\frac{G\mathcal{R}_{\max}}{(1-\gamma)^2}$ when using **GPOMDP** gradient estimator.

☞ Bounded variance of the gradient estimator

$$\mathrm{Var}\left[\widehat{\nabla}_m J(\theta)\right] = \mathbb{E}\left[\left\|\widehat{\nabla}_m J(\theta)\right\|^2\right]-\|\nabla J_H(\theta)\|^2 \le \frac{\Gamma_g^2-\|\nabla J_H(\theta)\|^2}{m} \le \frac{\Gamma_g^2}{m}$$

⚠ This was used as an assumption in [Xu et al., 2020], while it can be directly deduced from Assumption 4.

Convergence under the Lipschitz and smooth policy assumptions

Corollary 1. *Under Assumption 4, let $\delta_0\overset{def}{=}J^*-J(\theta_0)$. Any vanilla policy gradient method with a mini-batch sampling of size m and stepsize $\eta\in\left(0,\frac{2}{L(1-1/m)}\right)$, we have*

$$\mathbb{E}\left[\|\nabla J(\theta_U)\|^2\right] \le \frac{2\delta_0}{\eta T\left(2-L\eta\left(1-\frac{1}{m}\right)\right)}+\frac{L\Gamma_g^2\eta}{m\left(2-L\eta\left(1-\frac{1}{m}\right)\right)}+\mathcal{O}(\gamma^H).$$

☞ Mini-batch $m\in\left[1,\frac{2\Gamma_g^2}{\epsilon^2}\right]$, number of iterations T s.t. $Tm\ge\frac{8\delta_0 L\Gamma_g^2}{\epsilon^4}$, step size $\eta=\frac{\epsilon^2 m}{2L\Gamma_g^2}$, horizon $H=\mathcal{O}(\log\epsilon^{-1})$, sample complexity: $TmH=\widetilde{\mathcal{O}}(\epsilon^{-4})$ [Zhang et al., 2020]

Main references

- Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world, 2020.
- Liu, Y., Zhang, K., Basar, T., and Yin, W. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), Advances in Neural Information Processing Systems, volume 33, pp. 7624–7636. Curran Associates, Inc., 2020.
- Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. Journal of Machine Learning Research, 22(98): 1–76, 2021.
- Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In International Conference on Learning Representations, 2020.
- Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. Sample efficient reinforcement learning with reinforce, 2020.