Sample Efficient Reinforcement Learning with REINFORCE

Junzi Zhang[1], Jongho Kim[2], Brendan O'Donoghue[3] and Stephen Boyd[2]

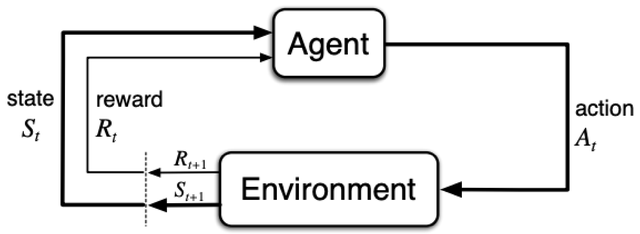
[1]ICME, Stanford University, [2]EE, Stanford University, [3]Google DeepMind

Abstract

We study the convergence theory of practical policy gradient methods and try to address the limitation of prior works on applicability in practical scenarios. Specifically, we provide the first set of global convergence and sample efficiency results for the well-known REINFORCE algorithm and contribute to a better understanding of its performance in practice.

Success of Reinforcement Learning

Reinforcement Learning (RL): algorithms for solving MDPs with incomplete information of transitions and rewards.



Heroes behind the success: combination of RL algorithms

- Value function learning (Q-learning): global convergence ✓
- Monte Carlo Tree Search (UCT): global convergence ✓
- Policy optimization (REINFORCE): global convergence ✓✗

Set-up: Episodic Online Model-free RL

Goal: maximize $\mathbf{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$, where $s_0\sim\rho$, $a_t\sim\pi(s_t,\cdot)$, $s_{t+1}\sim p(\cdot|s_t,a_t)$, and $\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})$ is called policy.

- Episodic (allow restart in the trajectory);
- Fully online (no simulator);
- Model-free (no storage of transition & reward models).

Policy Optimization & Policy Gradient

MDP/RL as policy optimization:

$$\text{maximize}_{\theta\in\Theta}\, F(\pi_\theta),$$
$$F(\pi)=\mathbf{E}\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t),$$

$s_0\sim\rho$, $a_t\sim\pi(s_t,\cdot)$, $s_{t+1}\sim p(\cdot|s_t,a_t)$, $\forall t\geq 0$, $\pi_\theta:\Theta\to\Pi$, and

$$\Pi=\left\{\pi\in\mathbf{R}^{SA}\,\middle|\,\textstyle\sum_{a=1}^{A}\pi_{s,a}=1\,(\forall s\in\mathcal{S}),\ \pi_{s,a}\geq 0\,(\forall s\in\mathcal{S},\ a\in\mathcal{A})\right\}.$$

Assume: $\rho>0$, $|\mathcal{S}|=S<\infty$, $|\mathcal{A}|=A<\infty$.

(Vanilla) policy gradient update: $n=1,\ldots,N$ (N episodes)

$$\theta^{n+1}=\theta^n+\alpha^n\nabla_\theta L_{\lambda^n}(\theta^n),$$

with $L_\lambda=F(\pi_\theta)+\lambda R(\theta)$: *e.g.*, entropy regularization.

Implementable (vanilla) policy gradient update: $n=1,\ldots,N$ (N episodes)

$$\theta^{n+1}=\theta^n+\alpha^n\frac{1}{M}\sum_{i=1}^{M}\tilde{\nabla}_\theta^{(i)}L_{\lambda^n}(\theta^n).$$

- Monte-Carlo policy gradient estimators (PGE): $\tilde{\nabla}_\theta^{(i)}$, $i=1,\ldots,M$.
- M independent mini-batch (MB) trajectories (variance reduction).

Theory vs. Practice: What was Missing?

	Global?	Practical PGE?	Finite MB?	High-Prob Rate?
Long Ago	No	Yes	Yes	No (a.s. Asymp)
$\sim$ 10 years	No	Yes	Yes	No (Rate in Expect.)
$\sim$ 2 years	Yes	No	No: $\Omega(1/M^p)$	No (Rate in Expect.)
Ours	**Yes**	**Yes**	**Yes**	**Yes**

Table: PGE: policy gradient estimators; MB: mini-batch; $p=1/2$ typically; a.s.: almost sure

Good news: recent advances in global convergence;
Bad news: limitations in characterizing practical performance

- Impractical PGE $\tilde{\nabla}_\theta^{(i)}$, $i=1,\ldots,M$;
- With finite MB M, only convergence to $\Omega(1/M^p)$ (*i.e.*, linear regret)
- Weak convergence guarantees: convergence rate in expectation.

Our Contribution:

- Practical PGE: *e.g.*, REINFORCE estimators;
- Finite MB M (including $M=1$), convergence to 0 (*i.e.*, sub-linear regret)
- Strong convergence guarantees: with probability at least $1-\delta$, regret sub-linear in N ($\forall\ N\geq 0$, *i.e.*, anytime) with poly-log$(1/\delta)$ terms + almost sure convergence.

Algorithm Specifications

We make the following choices:

- **Regularization**: $R(\theta)=\frac{1}{SA}\sum_{s\in\mathcal{S},a\in\mathcal{A}}\log\pi_\theta(s,a)$.
- **Phased schedule**:
  - updates divided into phases of lengths 2^l ($l\geq 0$);
  - λ^k constant in each phase and decreasing over phases: $\lambda^k=2^{-l/6-1}(1-\gamma)$ in phase l.
- **Post-processing**: add a simple truncation after each phase to keep $\pi_\theta(s,a)\geq 1/(2A)$ for all $s\in\mathcal{S}$, $a\in\mathcal{A}$.

Main Results (Informal)

Theorem [ZKOB20] For general PGEs, with probability at least $1-\delta$, we have a regret bound of

$$O((M^{\frac{1}{6}}+M^{-\frac{5}{6}})(N+M)^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}}+M(\log N)^2)=\tilde{O}(N^{\frac{5}{6}}).$$

In addition, we also have almost sure convergence of the average regret with an asymptotic rate of

$$O\left((M^{\frac{1}{6}}+M^{-\frac{5}{6}})N^{-\frac{1}{6}}\left(1+\frac{M}{N}\right)^{\frac{5}{6}}(\log N)^{\frac{5}{2}}+\frac{M(\log N)^2}{N}\right)=\tilde{O}(N^{-\frac{1}{6}}).$$

Remark: trade-off between lower variances with larger batch sizes ($M^{-\frac{5}{6}}$) and more frequent updates with smaller batch sizes ($M^{\frac{1}{6}}$).

Corollary [ZKOB20] For REINFORCE PGE with $M=1$, with probability at least $1-\delta$, we have a regret bound of

$$O(((1-\gamma)^{-7}S^2A^2+\|d_\rho^{\pi^\star}/\rho\|_\infty)N^{\frac{5}{6}}(\log(N/\delta))^{\frac{5}{2}}).$$

In addition, we also have almost sure convergence of the average regret with an asymptotic rate of

$$O(((1-\gamma)^{-7}S^2A^2+\|d_\rho^{\pi^\star}/\rho\|_\infty)N^{-1/6}(\log N)^{5/2}).$$

Here $d_\rho^\pi=(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbf{Prob}_\pi(s_t=s|s_0\sim\rho)$.

Open Problems

- Entropy regularization used in practice?
- $\rho\not>0$? (may need additional exploration)
- Function approximation?

References

[ZKOB20] J. Zhang, J. Kim, B. O'Donoghue, S. Boyd. *Sample efficient reinforcement learning with REINFORCE*, arXiv preprint arXiv:2010.11364, 2020.

[AKLM19] A. Agarwal, S. Kakade, J. Lee, and G. Mahajan. *On the theory of policy gradient methods: optimality, approximation, and distribution shift*, arXiv preprint arXiv:1908.00261, 2019.