Sample Efficient Reinforcement Learning with REINFORCE

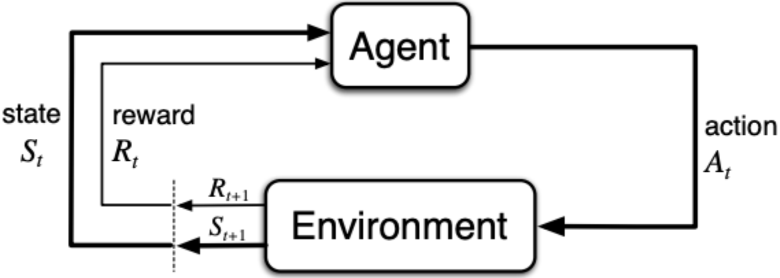
Junzi Zhang, Jongho Kim, Brendan O'Donoghue, Stephen Boyd

EE & ICME Departments, Stanford University
Google DeepMind

Algorithm Analysis for Learning and Games
INFORMS Annual Meeting, 2020

Overview

Markov Decision Process (MDP)



MDP (stationary, discounted): $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma, \rho)$, $\gamma \in [0, 1)$.

- $\rho > 0$, $S = |\mathcal{S}| < \infty$, $A = |\mathcal{A}| < \infty$. W.l.o.g., $r(s, a) \in [0, 1]$.
- 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.

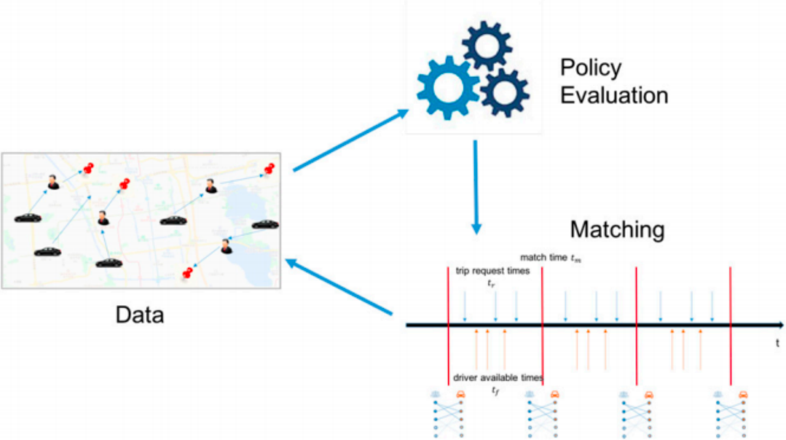
Reinforcement Learning (RL)

- **RL**: algorithms for solving MDPs with incomplete information of $\mathcal{M}$ (*e.g.*, p, r accessible by interacting with the environment) as input.
- **Today**: fully online (no simulator), episodic (allow restart in the trajectory) and model-free (no storage of transition & reward models).

Success of RL

Success of RL

Success of RL



Heroes Behind the Success: RL algorithms

Success in practice is a combination of several major families of RL algorithms:

- Value function learning (relatively well understood)
  - Q-learning, SARSA, Bellman Residue Minimization, etc.
- Monte Carlo Tree Search (relatively well understood):
  - ϵ-greedy tree search, UCT, BRUE, etc.
- Policy optimization (not very well understood apart from first-order local convergence)
  - Policy gradient, random search, actor-critic, etc.

Today: practical versions of policy gradient methods including REINFORCE (one of the least understood).

Policy Optimization

- Policy optimization reformulation:

$$\text{maximize}_{\pi \in \Pi} \, F(\pi),$$

where

$$F(\pi) = \mathbf{E} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t),$$

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

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

Policy Optimization

- Policy optimization reformulation:

$$\text{maximize}_{\pi \in \Pi} F(\pi),$$

- $F(\pi)$ is also written as $V^{\pi}(\rho)$ in the value function learning literature.
- Policy parametrization: $\pi_\theta : \Theta \to \Pi$.
- New problem:

$$\text{maximize}_{\theta \in \Theta} F(\pi_\theta).$$

- **Today**: energy-based policies: $\pi_\theta(s, a) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$, $\Theta = \mathbf{R}^{SA}$.
- Practical choice in reality, common basis for more advanced (*e.g.*, neural) parametrization.

Policy Gradient Existence

- Question: Is $F(\pi_\theta)$ differentiable?
- Answer: yes!
  - Indeed, $F(\pi_\theta)$ is at least C^2 and $\nabla_\theta F(\pi_\theta)$ is $8/(1-\gamma)^3$-Lipschitz.

Policy Gradient Methods

- (Vanilla) policy gradient method:

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

  where $L_\lambda(\theta) = F(\pi_\theta) + \lambda R(\theta)$: *e.g.*, entropy reg R.
- Some other variants: NPG (Fisher information matrix scaling), TRPO and PPO (trust region/KL regularization).
- What does the policy gradient look like?
  - **Policy gradient theorems** (PGT): hold for general C^1-smooth π_θ.
  - **Policy gradient estimators** (PGE): Monte Carlo approx of PGT.
- How to reduce variance/errors caused by Monte Carlo approximation?
  - **Mini-batch updates**.

PGE in Theoretical Analysis

- Visitation-measure based PGT:

$$\nabla_\theta F(\pi_\theta) = \frac{1}{1-\gamma} \mathbf{E}_{s \sim d_\rho^{\pi_\theta}} \mathbf{E}_{a \sim \pi_\theta(s,\cdot)} \left[Q^{\pi_\theta}(s,a) \nabla \log \pi_\theta(s,a) \right].$$

Here $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$ denotes a trajectory, and

$$Q^\pi(s,a) = \mathbf{E}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t) \middle| s_0 = s, a_0 = a, a_t \sim \pi(s_t, \cdot), s_{t+1} \sim p(\cdot | s_t, a_t), \forall t > 0 \right],$$

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

PGE in Theoretical Analysis

- Visitation-measure based PGT:

$$\nabla_\theta F(\pi_\theta) = \frac{1}{1-\gamma}\mathbf{E}_{s\sim d_\rho^{\pi_\theta}}\mathbf{E}_{a\sim\pi_\theta(s,\cdot)}\left[Q^{\pi_\theta}(s,a)\nabla\log\pi_\theta(s,a)\right].$$

- Vistation measure based PGE **(used in theory)**:

$$\bar{\nabla}_\theta F(\pi_{\theta^k}) = \tfrac{1}{1-\gamma}(\bar{Q}^k(s,a) - b(s))\nabla\log\pi_\theta(s,a),$$

where $s\sim d_\rho^{\pi_{\theta^k}}$, $a\sim\pi_{\theta^k}(s,\cdot)$, $\bar{Q}^k(s,a)$ approximates $Q^{\pi_{\theta^k}}(s,a)$, b is baseline: trajectory for sampling s is wasted.

  - Example $\bar{Q}$: $\bar{Q}^k(s,a) = \sum_{t'=t}^{H^k}\gamma^{t'-t}r_{t'}^k$, H^k is a truncation horizon, $\tau^k = (s,a,r_0^k,\dots,s_{H^k}^k,a_{H^k}^k,r_{H^k}^k) \sim \mathbf{Prob}_{s,a}^{\pi_{\theta^k}}$.

PGE in Practice

- Trajectory-based PGT:

$$\nabla_\theta F(\pi_\theta) = \mathbf{E}_{\tau \sim \mathbf{Prob}_\rho^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta}(s_t, a_t) \nabla \log \pi_\theta(s_t, a_t) \right]$$

- REINFORCE PGE **(used in practice)**:

$$\hat{\nabla}_\theta F(\pi_{\theta^k}) = \sum_{t=0}^{\lfloor \beta H^k \rfloor} \gamma^t (\widehat{Q}^k(s_t^k, a_t^k) - b(s_t^k)) \nabla_\theta \log \pi_{\theta^k}(a_t^k | s_t^k),$$

where $\beta \in (0, 1)$, $\widehat{Q}^k(s, a)$ approximates $Q^{\pi_{\theta^k}}(s, a)$, b is baseline, H^k is truncation horizon, $\tau^k = (s_0^k, a_0^k, r_0^k, \ldots, s_{H^k}^k, a_{H^k}^k, r_{H^k}^k) \sim \mathbf{Prob}_\rho^{\pi_{\theta^k}}$.

  - Example $\widehat{Q}$: $\widehat{Q}^k(s_t^k, a_t^k) = \sum_{t'=t}^{H^k} \gamma^{t'-t} r_{t'}^k$.

Additional (Practical) PGE

- Actor-critic PGE: REINFORCE or visitation measure based estimators with Q-functions estimated using TD algorithms.
- Many other versions of policy gradient theorems, which is why you see so many different versions of so-called policy gradient algorithms.
  - Finite horizon cases
    - $\nabla_\theta F(\pi_\theta) = \mathbf{E}_{\tau \sim \mathbf{Prob}_\rho^{\pi_\theta}}[Q^{\pi_\theta}(s_t, a_t) \sum_{t=0}^{H} \nabla \log \pi_\theta(s_t, a_t)]$
  - Zeroth-order approximation
    - a.k.a. random search, corresponding to a random perturbation/smoothing type "policy gradient theorem", widely used in PG + LQR literature.
- Question 1: Can we deal with all kinds of (practical) estimators (including REINFORCE)?

Mini-batch Updates

- Sample M independent trajectories $\tau_1^k, \ldots, \tau_M^k$ from $\mathcal{M}$ following policy π_{θ^k} and then compute an approximate gradient $\tilde{\nabla}_\theta^{(i)} L_{\lambda^k}(\theta^k)$ $(i = 1, \ldots, M)$ using each of these M trajectories.
- Then update as follows:

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

- Question 2: Can we accurately characterize the effect of M?

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: $\Omega(\frac{1}{M^p})$	No (Rate in Expect.)
Our Work	Yes	Yes	Yes	Yes (High-Prob + a.s.)

Table: PGE: policy gradient estimators; MB: mini-batch

- Exceptions:
  - LQR [JSW20] (our work: general MDPs);
  - model-based NPG [CYJW19, ESRM20] (our work: model-free);
  - oracle-based NPG with linear regret term [AYBB+19] (our work: sub-linear regret).

Overview

1) choose regularization $R(\theta) = \frac{1}{SA}\sum_{s\in\mathcal{S},a\in\mathcal{A}} \log \pi_\theta(s,a)$; 2) decrease λ^k in doubling phases; 3) add simple truncation after each phase.

Then we obtain (N is the number of episodes):

- any-time sub-linear high-prob regret bound

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

- a.s. convergence of average regret with asymptotic rate

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

Main Ideas

- **A group of easy-to-verify assumptions for PGE**:
  - *e.g.*, satisfied by REINFORCE with $\Theta(\log k)$ truncated horizon H^k;
- **Phase analysis**: bound regret in each phase (with λ^k fixed)
  - **Control of "bad" episodes**: sub-linear upper bound on # episodes with large gradient norms $\|\nabla_\theta L_\lambda(\theta^k)\|_2$.
  - **Gradient domination condition** [AKLM19]: from gradient norm $\|\nabla_\theta L_\lambda(\theta^k)\|_2$ to sub-optimality gap $F^\star - F(\pi_{\theta^k})$.
- **Doubling trick**:
  - stitch together phase regrets with $\log N$ additional terms.
- **From high prob (with** $\log(1/\delta)$ **dependency) to a.s.**:
  - Borel-Cantelli.

Open Problems

Open problems:

- Practically widely used (relative) entropy regularization, and empirical tests of the log-barrier one adopted in our work and [AKLM19].
- Remove the necessity of $\rho > 0$.
- Function approximation.

Any Questions?

Thank you all for listening! Any questions?

[ZKOB20] *Sample efficient reinforcement learning with REINFORCE*, arXiv preprint arXiv:2010.11364, 2020.