Divergence-augmented Policy Optimization

Qing Wang[4], Yingru Li[1,2], Jiechao Xiong[5], Tong Zhang[2,3]
[1] CUHK(SZ) [2] SRIBD [3] HKUST [4] Huya AI [5] Tencent AI Lab

Previews

We are interested in designing **conservative update** to ensure **stable** policy learning when **off-policy** data are reused.

1. We propose to stabilize policy improvement by constraining the **discounted state-action visitation** induced by consecutive policies to be close to one another.
2. We demonstrate the benefits of our approach through an empirical comparison on Atari games where the reuse of off-policy data is necessary.

Problem and challenges

Reinforcement learning paradigm:

- Learning agent that interacts repeatedly with the environment.
- The environment is formalized by a *Markov decision process* $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma, \rho)$.
- Goal: how to map states to distribution over actions (policy learning) in order to maximize the long term cumulative reward.

Policy optimization

Consider some class of **parametric** stochastic policy $\{\pi_\theta, \theta \in \Theta\}$.
Objective: $J(\theta) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \rho, \pi_\theta\right]$

- **Policy gradient**: optimize directly the cumulative rewards
  - Reinforce (Williams 88′, 92′)
  - Actor-critic methods (Sutton et al. 00′)

$$\widehat{\nabla}_\theta J(\theta) = \sum_{t=0}^{\tau} \gamma^t r_t \sum_{t=0}^{\tau} \nabla \log \pi_\theta\left(a_t \mid s_t\right).$$

$$\theta \leftarrow \theta + \eta \widehat{\nabla}_\theta J(\theta),$$

  - **Large variance issue! ✗**
- **Conservative approaches**: optimize surrogate objective that can provide local improvement.
  - Conservative policy Iteration (**CPI**) (Kakade & Langford 02′).
  - Trust region policy optimization (**TRPO**) (Shulman et al. 15′)
  - Proximal policy optimization (**PPO**) (Shulman et al. 17′)
  - Need to ensure that the new policy stays in the vicinity of the current policy!

Conservative optimization

- Discounted state-visitation:

$$d_\rho(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid \rho, \pi).$$

- Discounted state-action-visitation:

$$\mu(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a \mid \rho, \pi).$$

- **Performance difference lemma**:

$$J\left(\pi'\right) = J(\pi) + \mathbb{E}_{s\sim d_\rho^{\pi}} \mathbb{E}_{a\sim\pi'(.|s)}\left[A^{\pi}(s,a)\right].$$

- Implementation of the **vicinity requirement**:
- **Surrogate objective**: given a current policy π_i,

$$L_{\pi_i}\left(\pi'\right) = J\left(\pi_i\right) + \mathbb{E}_{s\sim d_\rho^{\pi_i}} \mathbb{E}_{a\sim\pi'(\cdot|s)}\left[A^{\pi_i}(s,a)\right].$$

- **CPI** returns a stochstic mixture:

$$\pi_{i+1} = \alpha_i \pi_i^+ + (1-\alpha_i)\,\pi_i,$$

where $\pi_i^+ = \arg\max_{\pi'} L_{\pi_i}\left(\pi'\right)$.

- **TRPO** solves a KL-constrained problem:

$$\pi_{i+1} = \arg\max_{\pi'} L_{\pi_i}\left(\pi'\right)$$

$$\text{s.t}\ \mathbb{E}_{s\sim d_\rho^{\pi_i}}\left[\mathrm{KL}\left(\pi'(s)\|\pi_i(s)\right)\right] \le \delta.$$

- **PPO** optimizes a clipped objective for ensuring target policy stays near to current policy.

Theoretical insights

Lower bound on policy's performance: for any two policies π and π'

Long term effects matters

In terms of action probabilities

$$J\left(\pi'\right) \ge L_\pi\left(\pi'\right) - \frac{2\gamma\epsilon^{\pi}}{1-\gamma}\mathbb{E}_{s\sim d_\pi}\left[D_{\mathrm{TV}}\left(\pi'(s)\|\pi(s)\right)\right].$$

Pay extra term of planning horizon!

- Used in CPI, TRPO and PPO.
- **Looser lower bound!**
- **Instability in long horizon problem!**

In terms of discounted state-action-visitation

$$J\left(\pi'\right) \ge L_\pi\left(\pi'\right) - \epsilon^{\pi} D_{\mathrm{TV}}\left(\mu_{\pi'}\|\mu_\pi\right).$$

- **Do not suffer long-horizon factor. ✓**
- **Requires state distribution! ✗**

Divergence-augmented policy optimization (DAPO)

Mirror descent formulation of policy optimization

We seek to solve policy optimization problem by looking at **state-action joint space**:

$$\max_{\pi} J(\pi) = \mathbb{E}_{(s,a)\sim\mu_\pi}\left[r(s,a)\right] = \langle \mu_\pi, r\rangle$$

Then we try to solve the problem by mirror descent (proximal iterative scheme) to promote interior point to prevent premature convergence and instability

$$\mu_{t+1} = \arg\min_{\mu_\pi \in \Delta_\Pi} D_F(\mu_\pi, \mu_t) + \eta\langle -r, \mu_\pi\rangle \tag{1}$$

Potential function as **Conditional Negentropy** $F(\mu) = \sum_{s,a}\mu(s,a)\log\pi_\mu(a|s)$, we get

$$D_F(\mu', \mu) = D'_{\mathrm{KL}}(\mu', \mu) = \sum_{s,a}\mu'(s,a)\log\frac{\pi'(a|s)}{\pi(a|s)}$$

Divergence-augmented pseudo reward

Optimizing equation 1 with conditional negentropy via policy gradient is equivalent to ordinary policy optimization using **pseudo reward** at each state-action pair: (π_t is the a current data-generating policy)

$$r(s,a) - \frac{1}{\eta}\log\frac{\pi(a|s)}{\pi_t(a|s)}$$

So What?

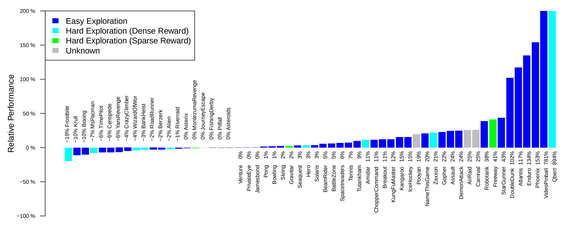

Figure 1: The relative score $\frac{\text{proposed - baseline}}{\max(\text{human, baseline}) - \text{random}}$ of **DAPO (our proposed)** v.s. PPO (baseline) in all 58 atari games.

- In **Figure 1**, our algorithms achieve better performance in a large fraction of game environments.
- In Figure 2, the performance of **PPO**, PPO+DA, PPO+DA (1-step), and PPO+Entropy are plotted in different colors.
- **DAPO** corresponds to multi-step divergence augmentations (**PPO+DA**), which performs best.
- Entropy regularization is not a good choice. PPO performs the worst in the comparison.

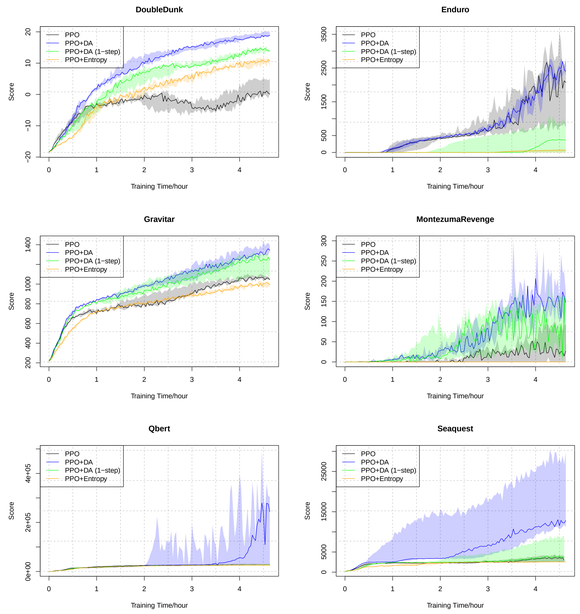

Figure 2: Comparison on selected environments.

More Information

Paper link: tinyurl.com/78wmp42w
Code link: github.com/lns/dapo

WeChat: RichardYRLi
Personal site: richardli.xyz
twitter: @RichardYRLi