Proximal Policy Optimization

ll

Z.Gan
Huazhong University of Science and Technology
School of Artificial Intelligence and Automation
Email: lebmont@hust.edu.cn

Spring 2022

Outline

Problems with Policy Gradient

1. Poor sample efficiency as PG is on-policy learning
2. Large policy update or improper step size destroying the training
    i. This is different from supervised learning where the learning and data are independent
    ii. Step too far $\rightarrow$ bad policy $\rightarrow$ bad data collection
    iii. May not be able to recover from a bad policy, which collapses the overall performance

Standard Definitions

Value function:

$$V_\pi(s_t) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k}]$$

Action-value function:

$$Q_\pi(s_t, a_t) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k}]$$

Advantage function:

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$$

Expected discounted reward:

$$\eta(\pi) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_k]$$

Outline

Rewritting and Local Approximation

$$
\begin{aligned}
\eta(\tilde{\pi}) - \eta(\pi) &= \eta(\tilde{\pi}) - \mathbb{E}[V^\pi(s_0)] \\
&= \eta(\tilde{\pi}) + \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t V^\pi(s_t) - \sum_{t=0}^{\infty} \gamma^t V^\pi(s_t)] \\
&= \eta(\tilde{\pi}) + \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (\gamma V^\pi(s_{t+1}) - V^\pi(s_t))] \\
&= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (r_t + \gamma V^\pi(s_{t+1}) - V^\pi(s_t))] \\
&= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t A^\pi(s_t, a_t)] \\
&= \sum_{t=0}^{\infty} \sum_{s} P(s_t = s|\tilde{\pi}) \sum_{a} \tilde{\pi}(a|s) \gamma^t A^\pi(s_t, a_t) \\
&= \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a|s) A^\pi(s_t, a_t) \approx \sum_{s} \rho_{\pi}(s) \sum_{s} \tilde{\pi}(a|s) A_\pi(s, a)
\end{aligned}
$$

Minorization-Maximization Algorithm

$$L_\pi(\tilde{\pi}) = \eta(\tilde{\pi}) \approx \eta(\pi) + \sum_a \rho_\pi(s) \sum_s \tilde{\pi}(a|s) A_\pi(s,a)$$

Definition

Total variation divergence: $D_{TV}(p,q) = \frac{1}{2}\sum_i |p_i - q_i|$

KL divergence: $D_{KL}(p,q) = \sum_i q_i log(\frac{p_i}{q_i})$

A lower bound(surrogate function): $\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - \frac{4\gamma\epsilon}{(1-\gamma)^2}\alpha^2$

where $\alpha = max_s D_{TV}(\pi, \tilde{\pi}), \epsilon = max_s|\mathbb{E}_{\pi'}[A_\pi(s,a)]|$ and $\pi' = argmax_{\pi'} L_\pi(\pi')$

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - C max_s D_{KL}(\pi, \tilde{\pi})$$

where $C = \frac{4\gamma\epsilon}{(1-\gamma)^2}$ and $D_{TV}(p,q)^2 \leq D_{KL}(p,q)$

Outline

Parameterization and Trust Region Constraint

Consider parameterized policies π_θ

$$maximize_\theta[L_{\tilde{\theta}}(\theta) - Cmax_s D_{KL}(\theta, \tilde{\theta})]$$

To take larger steps

$$maximize_{\tilde{\theta}} L_\theta(\tilde{\theta})$$

$$subject\ to\ max_s D_{KL}(\theta, \tilde{\theta}) \leq \delta$$

Consider the average KL divergence

$$\overline{D}^\rho_{KL}(\theta, \tilde{\theta}) = \mathbb{E}_{s\sim\rho}[D_{KL}(\pi_\theta(\cdot|s), \pi_{\tilde{\theta}}(\cdot|s))]$$

$$maximize_{\tilde{\theta}} L_\theta(\tilde{\theta})$$

$$subject\ to\ \overline{D}^\rho_{KL}(\theta, \tilde{\theta}) \leq \delta$$

Importance Sampling

$$maximize_{\tilde{\theta}} \sum_s \rho_\theta(s) \sum_a \pi_{\tilde{\theta}}(a|s) A_\theta(s,a)$$
$$subject\ to\ \overline{D}^{\rho}_{KL}(\theta, \tilde{\theta}) \leq \delta$$

Use $q = \pi_\theta(a|s)$ to denote the sampling distribution

$$\sum_a \pi_{\tilde{\theta}}(a|s_n) A_\theta(s, a_n) = \mathbb{E}_{a\sim q}[\frac{\pi_{\tilde{\theta}}(a|s_n)}{q(a|s_n)} A_\theta(s_n, a)]$$

Replace advantage value by state-action value and using expectations:

$$maximize_{\tilde{\theta}} \mathbb{E}_{s\sim p_\theta, a\sim q}[\frac{\pi_{\tilde{\theta}}(a|s)}{q(a|s)} Q_\theta(s,a)]$$
$$subject\ to\ \mathbb{E}_{s\sim\rho}[D_{KL}(\pi_\theta(\cdot|s), \pi_{\tilde{\theta}}(\cdot|s))] \leq \delta$$

Linear and Quadratic Approximation

$$maximize_{\tilde{\theta}} \mathbb{E}_{s \sim p_\theta, a \sim q}[\frac{\pi_{\tilde{\theta}}(a|s)}{q(a|s)} Q_\theta(s,a)]$$

$$subject\ to\ \mathbb{E}_{s \sim \rho}[D_{KL}(\pi_\theta(\cdot|s), \pi_{\tilde{\theta}}(\cdot|s))] \leq \delta$$

An analytic estimator has computational benefits in the large-scale setting to approximately solve this constrained optimization problem

$J_{\theta_t}(\theta) \approx \nabla_\theta J_{\theta_t}(\theta)|_{\theta_k}(\theta - \theta_t) = g^T(\theta - \theta_k)$

$\overline{D}_{KL}(\theta, \theta_k) \approx \frac{1}{2}(\theta - \theta_k)^T \nabla_\theta^2 \overline{D}_{KL}(\theta, \theta_k)|_{\theta_k}(\theta - \theta_k) = \frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k)$

$$\boxed{\theta_{t+1} = argmax_\theta g^T(\theta - \theta_t)\ s.t.\ \tfrac{1}{2}(\theta - \theta_t)^T H(\theta - \theta_t) \leq \delta}$$

$$\theta_{t+1} = \theta_t + \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g$$

Algorithm 1 Trust Region Policy Optimization Algorithm

```
Initialize policy parameters θ_0 randomly
for k=0,1,2,... do
    Collect set of trajectories on policy π_k
    Estimate advantages A_t
    Compute PG g_k and KL-divergence Hessian vector product H_k
    Use Conjugate Gradient Algorithm to obtain x_k ≈ H_k^{-1} g_k
    θ_{t+1} = θ_t + α sqrt(2δ / (x_k^T H_k x_k)) x_k
end for
return
```

$$\theta_{t+1} = \theta_t + \alpha\sqrt{\frac{2\delta}{x_k^T H_k x_k}}x_k$$

Outline

Clipped Surrogate Objective

$$maximize_{\tilde{\theta}} \mathbb{E}_t[\frac{\pi_{\tilde{\theta}}(a_t|s_t)}{\pi_{\theta}(a_t|s_t)} A_t]$$

$$subject\ to\ \mathbb{E}_t[D_{KL}(\pi_\theta(\cdot|s_t), \pi_{\tilde{\theta}}(\cdot|s_t))] \leq \delta$$

Definition

$$L^{CPI}(\theta) = \mathbb{E}_t[\frac{\pi_{\tilde{\theta}}(a_t|s_t)}{\pi_{\theta}(a_t|s_t)} A_t] = \mathbb{E}_t[r_t(\theta) A_t]$$

Consider how to modify the objective to penalize changes to the policy

$$L^{CLIP}(\theta) = \mathbb{E}_t[min(r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)]$$

Adaptive KL Penalty Coefficient

$$L^{CLIP}(\theta) = \mathbb{E}_t[min(r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)]$$

Use a penalty on KL divergence to adapt the penalty coefficient(as an alternative to the clipped surrogate objective)

$$L^{KLPEN} = \mathbb{E}_t[\frac{\pi_{\tilde{\theta}}(a_t|s_t)}{\pi_\theta(a_t|s_t)}A_t - \beta D_{KL}[\pi_\theta(\cdot|s_t), \pi_{\tilde{\theta}}(\cdot|s_t)]]$$

Algorithm 2 Proximal Policy Optimization(Actor-Critic Style)

```
for i=1,2,... do
    for actor=0,1,2,... do
        Run policy π_θ for T timesteps
        Compute advantage estimates A_t
    end for
    Optimize surrogate L(θ) with K epochs and minibatch size M ≤ NT
    Update θ
end for
return
```

Reference

[1] Schulman, John , et al. "Trust Region Policy Optimization." ICML 2015.

[2] Schulman, J. , et al. "Proximal Policy Optimization Algorithms." (2017).

Thank you!