Policy Gradient Methods

Sunmook Choi

Dept. of Mathematics
Korea University

Policy Gradient Methods

The goal of solving MDPs is to find an optimal policy $\pi(a|s)$ so that the agent can take an action a at each state s based on the policy.

- DQN uses a neural network to approximate Q-value function. An optimal policy is obtained by

$$\pi_\theta(a|s) = \arg\max_a Q_\theta(s, a).$$

- In Policy Gradient Methods, a policy is directly learned by a parametric distribution $\pi_\theta(a|s)$.
- Let $\tau = (s_0, a_0, r_1, s_1, a_1, r_2, s_2, \ldots, s_{T-1}, a_{T-1}, r_T, s_T)$ be a (random) trajectory by $\pi_\theta(a|s)$.
  The goal of policy gradient method is to maximize the expected total reward $r(\tau) = \sum_{t=0}^{T-1} r_{t+1}$.

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)}[r(\tau)] = \int r(\tau)\, p_\theta(\tau)\, d\tau, \qquad \theta \leftarrow \theta + \alpha \cdot \nabla_\theta J(\theta)$$

- Using the Markov property, the probability of a (sampled) trajectory τ is

$$p_\theta(\tau) = \rho_0 \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t)\, p(s_{t+1}|s_t, a_t).$$

Policy Gradient Theorem

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\tau\sim p_\theta(\tau)}[r(\tau)] = \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[r(\tau)\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

$$\because \nabla_\theta \mathbb{E}_{\tau\sim p_\theta(\tau)}[r(\tau)] = \nabla_\theta \int r(\tau)\, p_\theta(\tau)\, d\tau = \int \nabla_\theta p_\theta(\tau)\, r(\tau)\, d\tau = \int \frac{\nabla_\theta p_\theta(\tau)}{p_\theta(\tau)} p_\theta(\tau)\, r(\tau)\, d\tau = \int p_\theta(\tau)\, r(\tau)\, \nabla_\theta \log p_\theta(\tau)\, d\tau$$

$$= \mathbb{E}_{\tau\sim p_\theta(\tau)}[r(\tau)\, \nabla_\theta \log p_\theta(\tau)] = \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[r(\tau)\ \textstyle\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

* $\nabla_\theta \log p_\theta(\tau) = \nabla_\theta\Big(\log \rho_0(s_0) + \sum_{t=0}^{T-1} \log \pi_\theta(a_t|s_t) + \sum_{t=0}^{T-1} \log p(s_{t+1}|s_t, a_t)\Big) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)$

- The theorem implies that the gradient can be approximated by sampling a large number of trajectories and taking their average (Monte Carlo estimation).
- However, the total reward $r(\tau)$ adds high variance due to erratic trajectories.

Reducing Variance

Causality Problem: policy at time t' cannot affect reward at time t when $t < t'$.

- Total reward: $r(\tau) = \sum_{t=0}^{T-1} r_{t+1}$. $\sum_{t=0}^{k} r_{t+1}$ does not depend on actions $a_{k+1}, a_{k+2}, \ldots$.
- Using this property, we can reduce the variance of gradient estimates.
- Note that it is not related to Markov property.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left(\sum_{t=0}^{T-1} r_t \right) \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \right] = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=0}^{T-1} \left(\sum_{t'=t}^{T-1} r_{t'+1} \right) \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

By introducing discount factor $\gamma \in [0, 1]$, we can further reduce the variance.

- That is, we are slightly changing (or redefining) the goal of policy gradient methods.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=0}^{T-1} \left(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'+1} \right) \nabla_\theta \log \pi_\theta(a_t|s_t) \right] = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=0}^{T-1} G_t \, \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

REINFORCE

Repeat (1) ~ (3)

(1) Execute M trajectories
(each starting in state s and executing (stochastic) policy π_θ)

(2) Approximate the gradient of the objective function $J(\theta)$

$$g_\theta := \frac{1}{M} \sum_{i=1}^{M} \left(\sum_{t=0}^{T-1} G_t^{(i)} \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) \right) \approx \nabla_\theta J(\theta)$$

(3) Update policy (network parameters) to maximize $J(\theta)$

$$\theta := \theta + \alpha g_\theta \approx \theta + \alpha \nabla_\theta J(\theta)$$

REINFORCE with baseline

We try to reduce the variance further. Consider a *baseline* $b(s)$ which does not depend on actions.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=0}^{T-1} \Big(G_t - b(s_t) \Big) \nabla_\theta \log \pi_\theta(a_t|s_t) \right]$$

- Notice that the gradient still remains unchanged because the estimator is unbiased.
  - $\mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=0}^{T-1} b(s_t) \nabla_\theta \log \pi_\theta(a_t|s_t) \right] = 0.$
- A good baseline is $V(s_t)$ because its unbiased estimator is G_t.
- This is a fundamental algorithm for actor-critic methods.
  - 'Actor' : the policy network that defines how to act at each state.
  - 'Critic' : the value network that estimates how good or bad the state is.

Why Unbiased?

$$\because \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} b(s_t)\,\nabla_\theta \log \pi_\theta(a_t|s_t)\right] = \mathbb{E}_{s_0\sim\rho_0(\cdot)}\Big[\mathbb{E}_{a_0\sim\pi_\theta(\cdot|s_0)}\Big[\mathbb{E}_{(s_1,a_1,r_2,\dots)\sim p_\theta(\cdot|s_0,a_0)}\left[\sum_{t=0}^{T-1} b(s_t)\,\nabla_\theta \log \pi_\theta(a_t|s_t)\right]\Big]\Big]$$

$$= \mathbb{E}_{s_0\sim\rho_0(\cdot)}\Big[\mathbb{E}_{a_0\sim\pi_\theta(\cdot|s_0)}\Big[\mathbb{E}_{s_1\sim p(\cdot|s_0,a_0)}\Big[\mathbb{E}_{a_1\sim\pi_\theta(\cdot|s_1)}\cdots\Big[\mathbb{E}_{s_{T-1}\sim p(\cdot|s_{T-2},a_{T-2})}\Big[\mathbb{E}_{a_{T-1}\sim\pi_\theta(\cdot|s_{T-1})}\left[\sum_{t=0}^{T-1} b(s_t)\,\nabla_\theta \log \pi_\theta(a_t|s_t)\right]\Big]\Big]\Big]\Big]\Big]$$

Using the conditional probability, we have

$$\mathbb{E}_{s_{T-1}\sim p(\cdot|s_{T-2},a_{T-2})}\Big[\mathbb{E}_{a_{T-1}\sim\pi_\theta(\cdot|s_{T-1})}\left[\sum_{t=0}^{T-1} b(s_t)\,\nabla_\theta \log \pi_\theta(a_t|s_t)\right]\Big]$$

$$= \mathbb{E}_{s_{T-1}\sim p(\cdot|s_{T-2},a_{T-2})}\Big[\sum_{t=0}^{T-2} b(s_t)\,\nabla_\theta \log \pi_\theta(a_t|s_t) + b(s_{T-1})\cdot\mathbb{E}_{a_{T-1}\sim\pi_\theta(\cdot|s_{T-1})}[\nabla_\theta \log \pi_\theta(a_{T-1}|s_{T-1})]\Big]$$

$$= \mathbb{E}_{s_{T-1}\sim p(\cdot|s_{T-2},a_{T-2})}\Big[\sum_{t=0}^{T-2} b(s_t)\,\nabla_\theta \log \pi_\theta(a_t|s_t)\Big] = \sum_{t=0}^{T-2} b(s_t)\,\nabla_\theta \log \pi_\theta(a_t|s_t).$$

Note that the second equality holds because

$$\begin{aligned}\mathbb{E}_{a_{T-1}\sim\pi_\theta(\cdot|s_{T-1})}[\nabla_\theta \log \pi_\theta(a_{T-1}|s_{T-1})] &= \textstyle\sum_{a_{T-1}} \pi_\theta(a_{T-1}|s_{T-1})\nabla_\theta \log \pi_\theta(a_{T-1}|s_{T-1}) \\ &= \textstyle\sum_{a_{T-1}} \nabla_\theta\, \pi_\theta(a_{T-1}|s_{T-1}) = \nabla_\theta \sum_{a_{T-1}} \pi_\theta(a_{T-1}|s_{T-1}) = \nabla_\theta 1 = 0.\end{aligned}$$

By doing this process iteratively, we obtain that $\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} b(s_t)\,\nabla_\theta \log \pi_\theta(a_t|s_t)\right] = 0.$