Chapter 9 Policy Gradient Methods

Policy Gradient, Monte Carlo Policy Gradient

Policy Gradient

1. What is the idea about Policy Gradient Methods?
2. Explain different forms and distributions of average state value metric and average reward. What is the relationship between these two metrics?
3. Explain the form of derivatives of metrics.

Monte Carlo Policy Gradient (REINFORCE)

4. Explain the process of MCPG. Why is it also called REINFORCE?
5. Why Policy Gradient can keep a balance between exploiation and exploration?
6. Explain the process of MCPG.

1. Policy is used to be represented in tabular form. Policy gradient methods represent policy in function form and use scalar metrics to evaluate and update the policy to find optimal policies. Policy gradient methods are policy-based, instead of value-based in the previous chapters. It is more efficient to handle large state/action space and has stronger generalization ability to make it more efficient in data utilization. The basic method is using the gradient-ascent method to optimize the metric J(θ) to find optimal policy:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta_t)$$

   where J(θ) is the metric, α is the optimizing rate.

2.

   Average state value:

   The basic form is:

$$\bar{v}_\pi = \sum_{s \in \mathcal{S}} d(s) v_\pi(s)$$

   Also can be written as:

$$\bar{v}_\pi = \mathbb{E}_{S \sim d}[v_\pi(S)]$$

   Also:

$$\bar{v}_\pi = d^T v_\pi$$

   Also:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right]$$

   where {R} is gained by the agent following policy π(θ) with suffient episodes and length.

   For the last form, we can prove it is equal to other forms:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R_{t+1}\right] &= \sum_{s\in\mathcal{S}} d(s)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R_{t+1}|S_0=s\right] \\ &= \sum_{s\in\mathcal{S}} d(s)v_\pi(s) \\ &= \bar{v}_\pi.\end{aligned}$$

Usually use $\bar{v}_\pi$ to refer to stationary distribution and use $\bar{v}_\pi^0$ to refer to the case where d(s) is independent of the policy.

Average reward:

basic form:

$$\begin{aligned}\bar{r}_\pi &\doteq \sum_{s\in\mathcal{S}} d_\pi(s) r_\pi(s) \\ &= \mathbb{E}_{S\sim d_\pi}[r_\pi(S)],\end{aligned}$$

where $r_\pi(s)$ is $E[r(s, A)|s] = \Sigma\pi(a|s, \theta) \cdot r(a,s)$, where $r(a,s) = E[R|s,a] = \Sigma_r rp(r|s,a).s$

Also:

$$d_\pi^T r_\pi$$

Also:

$$J(\theta) = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{t=0}^{n-1} R_{t+1}\right]$$

Relationship:

$$\bar{r}_\pi = (1-\gamma)\bar{v}_\pi$$

3.

$$\nabla_\theta J(\theta) = \sum_{s\in\mathcal{S}}\eta(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi(a|s,\theta)q_\pi(s,a)$$

Also:

$$\nabla_\theta J(\theta) = \mathbb{E}_{S\sim\eta, A\sim\pi(S,\theta)}\Big[\nabla_\theta \ln\pi(A|S,\theta)q_\pi(S,A)\Big]$$

To satisfy $\pi(A|S,\theta) > 0$, we use softmax function:

$$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_{a'\in\mathcal{A}} e^{h(s,a',\theta)}}, \quad a\in\mathcal{A}$$

which can ensure $\pi(a|s,\theta) > 0$ and the total is 1. This policy can be realized by neural network.

4. Use gradient-ascent method to maximize J(θ):

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha\nabla_\theta J(\theta_t) \\ &= \theta_t + \alpha\mathbb{E}\Big[\nabla_\theta \ln\pi(A|S,\theta_t)q_\pi(S,A)\Big]\end{aligned}$$

we use samples to approximate the expectation. To be noted, $q_\pi(s_t,a_t)$ is unknown and we use $q_t(s_t,a_t)$ to approximate it. If $q_t(s_t,a_t)$ is obtained from Monte Carlo Estimation, then the method is called REINFORCE or Monte Carlo Policy Gradient. Transform the derivative in the equation:

$$\theta_{t+1} = \theta_t + \alpha \underbrace{\left(\frac{q_t(s_t, a_t)}{\pi(a_t|s_t, \theta_t)} \right)}_{\beta_t} \nabla_\theta \pi(a_t|s_t, \theta_t),$$

$$\theta_{t+1} = \theta_t + \alpha \beta_t \nabla_\theta \pi(a_t|s_t, \theta_t)$$

5. Firstly, if β >= 0, the probability of choosing current (s_t,a_t) increases. The greater β is, the stronger the increase is.

$$\pi(a_t|s_t, \theta_{t+1}) \geq \pi(a_t|s_t, \theta_t)$$

Similarly, if β < 0, the probability of choosing current (s_t,a_t) decreases.

$$\pi(a_t|s_t, \theta_{t+1}) < \pi(a_t|s_t, \theta_t)$$

Then, if q_t is large, then β is large, then the probability increases. This increases the exploitation. Similarly, if π_t is small, then β is large when $q_t > 0$, then the probability increases. This increases the exploration.

6.

Algorithm 9.1: Policy Gradient by Monte Carlo (REINFORCE)

Initialization: Initial parameter θ; $\gamma \in (0, 1)$; $\alpha > 0$.
Goal: Learn an optimal policy for maximizing $J(\theta)$.

For each episode, do
  Generate an episode $\{s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T\}$ following $\pi(\theta)$.
  For $t = 0, 1, \ldots, T-1$:
    Value update: $q_t(s_t, a_t) = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$
    Policy update: $\theta \leftarrow \theta + \alpha \nabla_\theta \ln \pi(a_t|s_t, \theta) q_t(s_t, a_t)$