Chapter 9. Policy Function Approximation.

最流行. (policy gradient)

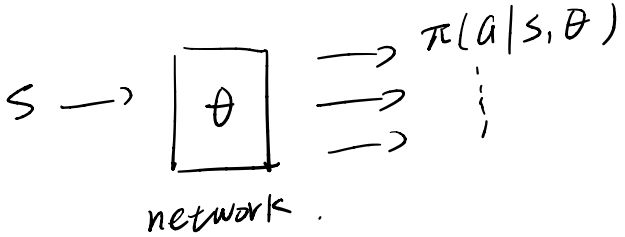
- policy based 直接评价 policy
- value based. 通过 action value 评价 policy.

1. Basic idea of policy gradient ←
2. Metrics to define optimal policies
3. Gradients of the metrics
4. Gradient-ascent algorithm (REINFORCE)
5. Summary

1. previous

	a_1	a_2	a_3	...
s_1	$\pi(a_1 \mid s_1)$	$\pi(a_2 \mid s_1)$	...	
⋮	⋮			
s_9	⋮			

Now: $\pi(a|s,\theta)$ | $\pi_\theta(a|s)$ | $\pi_\theta(a,s)$



Storage / generalization

2. Differences between tabular and function.

① table: π^* $V_{\pi^*}(s) \geq V_\pi(s)$ $\forall \pi$

maximize state value

② function:

maximize certain scalar metrics

3. basic idea of policy gradient:

① metrics (objective function) to define optimal policies $J(\theta)$

② $\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta_t)$

1° how to choose metrics

2° How to calculate the gradients of the metrics

2. Metrics to define optimal policies

① average state value:

(1) weight average. $\bar{V}_\pi = \sum_{s \in S} d(s) V_\pi(s) = d^T V_\pi$

$d(s)$ probability distribution. $\sum d(s) = 1$

written as $\bar{V}_\pi = E[V_\pi(s)] = \sum_s p(s) V_\pi(s)$

$$d^T = [\cdots, d(s), \cdots]^T$$

$$V_\pi = [\cdots, V_\pi(s), \cdots]^T$$

(2) How to select distribution d

1° case 1: d is independent of policy π

1° equally important: $d_0 = \frac{1}{|S|}$

2° a specific state s_0: $d_0(s) = 1$ $\quad d_0(s \neq s_0) = 0$

$\bar{V}_\pi = V_\pi(s_0)$

2° case 2: d depends on policy π

$d_\pi^T P_\pi = d_\pi^T$

P_π: state transition probability matrix

use given π^*

② . average one-step reward or simply average reward.

$$\bar{r}_\pi = \sum_{s\in S} d_\pi(s)\, r_\pi(s) = E[r_\pi(s)]$$

$$r_\pi(s) = \sum_{a\in A} \pi(a|s)\, r(s,a)$$

$$r(s,a) = E[R|s,a] = \sum_r r\, p(r|s,a)$$

An equivalent definition: given policy generate a trajectory $[R_{t+1}, R_{t+2}, \cdots]$

$$\lim_{n\to\infty} \frac{1}{n} E[R_{t+1} + R_{t+2} + \cdots + R_{t+n} \mid S_t = s_0]$$

$$= \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{k=1}^{n} R_{t+k} \mid S_t = s_0\right] = \underbrace{\lim_{n\to\infty} \frac{1}{n} E\left[\sum_{k=1}^{n} R_{t+k}\right]}_{\Delta}$$

$\bar{r}_\pi$ / $\bar{V}_\pi$: $\bar{r}_\pi = (1-\gamma)\bar{V}_\pi$

$$J(\theta) = E\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right] = \bar{V}_\pi$$

$$\bar{r}_\pi = \lim_{n\to\infty} \frac{1}{n} E\left[\sum_{t=1}^{n} R_t\right]$$

3. Gradients of the metrics.

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

$J(\theta)$ can be $\bar{v}_\pi$, $\bar{r}_\pi$, $\bar{v}_\pi^0$

= can be $\propto$ =

η is a distribution or weight of the states.

$$\nabla_\theta J(\theta) = E\left[\nabla_\theta \ln \pi(A|S,\theta)\, q_\pi(S,A)\right]$$

$$\propto \nabla_\theta \ln \pi(a|s,\theta)\, q_\pi(s,a)$$

$$\nabla_\theta \ln \pi(a|s,\theta) = \frac{\nabla_\theta \pi(a|s,\theta)}{\pi(a|s,\theta)}$$

$$\nabla_\theta \pi(a|s,\theta) = \pi(a|s,\theta) \cdot \nabla_\theta \ln \pi(a|s,\theta)$$

代入 $\nabla_\theta J = \sum_s d(s) \sum_a \nabla_\theta \pi(a|s,\theta)\, q_\pi(s,a)$

$$= \sum_s d(s) \sum_a \pi(a|s,\theta) \nabla_\theta \ln \pi(a|s,\theta)\, q_\pi(s,a)$$

$$= E_{S\sim d}\left[\sum_a \pi(a|s,\theta) \nabla_\theta \ln \pi(a|s,\theta)\, q_\pi(s,a)\right]$$

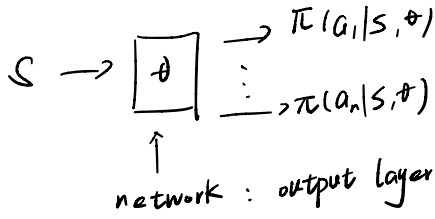
$$= E_{S\sim d, A\sim\pi}\left[\nabla_\theta \ln \pi(A|S,\theta)\, q_\pi(S,A)\right]$$

$$= E\left[\nabla_\theta \ln \pi(A|S,\theta)\, q_\pi(S,A)\right]$$

$\because \ln \pi(a|s,\theta)$ $\underline{\pi(a|s,\theta) > 0}$ $\therefore$ use softmax. from $(-\infty, +\infty)$ to $(0,1)$

$$z_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \qquad \underline{z_i \in (0,1) \quad \sum z_i = 1}$$

$$\therefore \pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_{a' \in \mathcal{A}} e^{h(s,a',\theta)}} \qquad h(s,a,\theta) \text{ is another function}$$



4. Gradient-ascent algorithm (Reinforce)

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

$= \theta_t + \alpha \underbrace{E}_{\triangle} [\nabla_\theta \ln \pi(A|S,\theta_t) q_\pi(S,A)]$

$\Downarrow$

The true gradient can be

$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t,\theta_t) \underbrace{q_\pi(s_t,a_t)}_{\triangle}$

$\Downarrow$

since q_π is unknown.

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

policy gradient method is on policy

How to interpret this algorithm?

$$\nabla_\theta \ln \pi(a_t|s_t,\theta_t) = \frac{\nabla_\theta \pi(a_t|s_t,\theta_t)}{\pi(a_t|s_t,\theta_t)}$$

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

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

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

$$\pi(a_t|s_t,\theta_{t+1}) \approx \pi(a_t|s_t,\theta_t) + (\nabla_\theta \pi(a_t|s_t,\theta_t))^T (\theta_{t+1}-\theta_t)$$

$$= \pi(a_t|s_t,\theta_t) + \alpha\beta_t (\nabla_\theta \pi(a_t|s_t,\theta_t))^T (\nabla_\theta \pi(a_t|s_t,\theta_t))$$

$$= \pi(a_t|s_t,\theta_t) + \alpha\beta_t \|\nabla_\theta \pi(a_t|s_t,\theta_t)\|^2.$$

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

β_t can balance exploration and exploitation

$$\begin{cases} \beta_t \propto q_t(s_t,a_t) & \text{— exploitation} \\ \beta_t \propto \dfrac{1}{\pi(a_t|s_t,\theta_t)} & \text{— exploration.} \end{cases}$$

Policy gradient by Mont. Carlo (Reinforce)

select s_0, $\pi(\theta_k)$ suppose the episode $\{s_0, a_0, r_1, \cdots s_{T-1}, a_{T-1}, r_T\}$

Value update: $q_t(s_t,a_t) = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$

policy update: $\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t,\theta_t) q_t(s_t,a_t)$

$\theta_t = \theta_T$

off-line 离线.