Policy Gradient

1. sampling method

1.1 e-greedy

$$\epsilon = random(0,1)$$

$$\pi(\epsilon) = \begin{cases} net(a|s) & \epsilon < 0.9 \\ rand & \epsilon \geq 0.9 \end{cases}$$

1.2 monte carlo markov chain

- e-greedy
- state transition

trajectory

$$(s_0, a_0, s_1, r_0), (s_1, a_1, s_2, r_1), (s_2, a_2, s_3, r_2), \ldots, (s_t, a_t, s_{t+1}, r_t)$$

2. optimization objective

the goal of policy gradient is trying to find the optimal parameter to maximize the total reward of trajectory. maximizing the total reward will be difficult, but it's possible to maximize the expectation , then we can use MLE and gradient ascent to estimate the optimal parameter.

$$\max_{\theta} E[\sum_{t=1}^{n} R(s_t, a_t); \pi_\theta]$$

$$U(\theta) = E[\sum_{t=1}^{n} R(s_t, a_t); \pi_\theta] = \sum_{\tau} P(\tau; \theta) R(\tau)$$

$$R_t(\tau) = \sum_{t=1}^{n} \gamma^{t-1} r_{n-t+1}$$

$$P(\tau; \theta) = \prod_{i=1}^{n} P(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t)$$

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta U(\theta)$$

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_{\tau} P(\tau; \theta) R(\tau) \\
&= \sum_{\tau} \nabla_\theta P(\tau; \theta) R(\tau) \\
&= \sum_{\tau} \frac{P(\tau; \theta)}{P(\tau; \theta)} \nabla_\theta P(\tau; \theta) R(\tau) \\
&= \sum_{\tau} P(\tau; \theta) \frac{\nabla_\theta P(\tau; \theta) R(\tau)}{P(\tau; \theta)} \\
&= \sum_{\tau} P(\tau; \theta) \nabla_\theta \log\left(P(\tau; \theta)\right) R(\tau) \\
&= E[\nabla_\theta \log\left(P(\tau; \theta)\right) R(\tau)] \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \log\Big(P(\tau^{(i)}; \theta)\Big) R(\tau^{(i)})
\end{aligned}$$

$$\begin{aligned}
\nabla_\theta \log\Big(P(\tau^{(i)}; \theta)\Big) &= \nabla_\theta \log\left(\prod_{t=1}^{m} P(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)})\pi_\theta(a_t^{(i)}|s_t^{(i)})\right) \\
&= \nabla_\theta \{\sum_{t=1}^{m} \log\Big(P(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)})\Big) + \sum_{t=1}^{n} \log\Big(\pi_\theta(a_t^{(i)}|s_t^{(i)})\Big)\} \\
&= \nabla_\theta \sum_{t=1}^{m} \log\Big(\pi_\theta(a_t^{(i)}|s_t^{(i)})\Big) \\
&= \sum_{t=1}^{m} \nabla_\theta \log\Big(\pi_\theta(a_t^{(i)}|s_t^{(i)})\Big)
\end{aligned}$$

$$\nabla_\theta U(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \log\Big(P(\tau^{(i)}; \theta)\Big) R(\tau^{(i)})$$

$$\nabla_\theta U(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{m} \nabla_\theta \log\Big(\pi_\theta(a_t^{(i)}|s_t^{(i)})\Big) R(s_t, a_t)$$

3. policy

3.1 gauss policy

$$\pi_\theta = \mu_\theta + \epsilon$$

$$\pi(a|s) \approx \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(a-\phi(s)^T\theta)^2}{2\sigma^2}\right)$$

$$\nabla_\theta \log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right) \approx \nabla_\theta \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(a^{(i)}-\phi(s^{(i)})^T\theta)^2}{2\sigma^2}\right)\right)$$

$$= \nabla_\theta \log\left(\frac{1}{2\pi\sigma}\right) + \nabla_\theta \log\left(\exp\left(-\frac{(a^{(i)}-\phi(s^{(i)})^T\theta)^2}{2\sigma^2}\right)\right)$$

$$= \nabla_\theta - \frac{(a^{(i)}-\phi(s^{(i)})^T\theta)^2}{2\sigma^2} = \frac{(a^{(i)}-\phi(s^{(i)})^T\theta)\phi(s^{(i)})}{\sigma^2}$$

$$\nabla_\theta U(\theta) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m}\nabla_\theta \log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)R(\tau^{(i)})$$

$$\nabla_\theta \log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right) \approx \frac{(a^{(i)}-\phi(s^{(i)})^T\theta)\phi(s^{(i)})}{\sigma^2}$$

$$\nabla_\theta U(\theta) \approx \frac{1}{n}\sum_{i=1}^{n}\{R(\tau^{(i)})\sum_{t=1}^{m}\frac{(a^{(i)}-\phi(s^{(i)})^T\theta)\phi(s^{(i)})}{\sigma^2}\}$$

$$\theta_{k+1} = \theta_k + \alpha\nabla_\theta U(\theta)$$

3.2 softmax policy

$$\pi(a_t^{(k)}|s_t) \approx softmax(a_t^{(k)}, s_t)$$

$$= \frac{\exp\left(z(s_t, a_t^{(k)})^T\theta\right)}{\sum_{i=1}^{n}\exp\left(z(s_t, a_t^{(i)})^T\theta\right)}$$

$$\log\left(\pi_\theta(a_t^{(k)}|s_t)\right) = z(s_t, a_t^{(k)})^T\theta - \log\left(\sum_{i=1}^{n}\exp\left(z(s_t, a_t^{(i)})^T\theta\right)\right)$$

$$\nabla_\theta \log\left(\pi_\theta(a_t^{(k)}|s_t)\right) = z(s_t, a_t^{(k)}) - \nabla_\theta \log\left(\sum_{i=1}^{n}\exp\left(z(s_t, a_t^{(i)})^T\theta\right)\right)$$

$$\nabla_\theta \log\left(\sum_{i=1}^{n}\exp\left(z(s_t, a_t^{(i)})^T\theta\right)\right) = \frac{\sum_{j=1}^{n} z(s_t, a_t^{(j)})\exp\left(z(s_t, a_t^{(j)})^T\theta\right)}{\sum_{i=1}^{n}\exp\left(z(s_t, a_t^{(i)})^T\theta\right)}$$

$$= \sum_{j=1}^{n} z(s_t, a_t^{(j)})\frac{\exp\left(z(s_t, a_t^{(j)})^T\theta\right)}{\sum_{i=1}^{n}\exp\left(z(s_t, a_t^{(i)})^T\theta\right)} = E_{\pi(a_t,|s_t)}[z(s_t, a_t)]$$

$$\nabla_\theta \log\left(\pi_\theta(a_t^{(k)}|s_t)\right) = z(s_t, a_t^{(k)}) - E_{\pi(a_t,|s_t)}[z(s_t, a_t)]$$

$$\nabla_\theta U(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} R(\tau^{(i)})\{z(s_t, a_t^{(k)}) - E_{\pi(a_t,|s_t)}[z(s_t, a_t)]\}$$

$$\theta_{k+1} = \theta_k + \alpha\nabla_\theta U(\theta)$$

4. baseline

$$\nabla_\theta U(\theta) \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m}\nabla_\theta \log\left(\pi_\theta(a_t^{(i)}|s_t^{(i)})\right)(R(\tau^{(i)})-b)$$

$$b = \frac{1}{n}\sum_{t=1}^{n}R(s_t,a_t)$$

5. transition distribution

if transition distribution is unknown，then we can make an assumption that transition distribution is some kind of distribution, such like gauss distribution and the neural network is powerful which can be used to approximate any distribution. so it will be good to use neural network to approximate transition distribution.

$$P(s_{t+1}|s_t,a_t;\theta) = net(s,a;\theta)$$

6. reinforce

- sample data from environment

$$(s_0,a_0,r_0),(s_1,a_1,r_1),(s_2,a_2,r_2),\ldots,(s_t,a_t,r_t)$$

- calculate discounted reward

$$R_t(s_t,a_t) = \sum_{t=1}^{n}\gamma^{t-1}r_{n-t+1}$$

- update parameter with gradient ascent

$$\theta_{k+1} = \theta_k + \alpha\nabla_\theta U(\theta)$$