Direct Policy Optimization

- lets assume that we have a policy π parameterized by θ. $\pi(a|s) = \theta^T s$.
- we will assume $\gamma = 1$ & terminated episodes.
- $J(\theta) \doteq v_{\pi_\theta}(s_0)$ MDP starts in state s_0.
- Consider $v_\pi(s)$ this is a function of θ.

$\underset{\theta}{\operatorname{argmax}} \; J(\theta)$ →

- Random Search.
- Evolutionary algorithm
- finite differencing gradient for $J(\theta)$

  $\theta_0 \qquad \theta_0 + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \alpha \approx \nabla_\theta J$

- Policy Gradient $\nabla_\theta J$

we want to compute $\nabla_\theta v_\pi(s)$

$$\nabla_\theta v_\pi(s) = \nabla_\theta \left[\sum_a \pi(a|s)\, q_\pi(s,a) \right]$$

$$= \sum_a \left[\nabla_\theta \left(\pi(a|s)\, q_\pi(s,a) \right) \right]$$

$$= \sum_a \left[q_\pi(s,a) \nabla_\theta \pi(a|s) + \pi(a|s) \nabla_\theta q_\pi(s,a) \right]$$

$$\pi(a|s) \nabla_\theta \left[\sum_{s',r} p(s',r|s,a) \left[\cancel{r} + v_\pi(s') \right] \right]$$

$$\pi(a|s) \nabla_\theta \left[\sum_{s'} p(s'|s,a)\, v_\pi(s') \right]$$

$$\pi(a|s) \sum p(s'|s,a)\, \nabla_\theta v_\pi(s')$$

$$= \sum_a \left[q_\pi(s,a) \nabla_\theta \pi(a|s) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla_\theta v_\pi(s') \right]$$

$$\sum_{a'} q_\pi(s',a') \nabla_\theta \pi(a'|s') + \underline{\quad\quad}$$

keep expanding

$$\vdots$$

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

$$+ \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \sum_{a'} q_\pi(s',a') \nabla_\theta \pi(a'|s') -$$

$$+ \sum_a \sum_{s'} \sum_{a'} \sum_{s''} \sum_{a''} q_\pi(s'',a'') \nabla_\theta \pi(a''|s'')$$

$$+$$

$$\sum_x \sum_{i=0}^{\infty} \Pr(s \to x, i, \pi) \sum_a q_\pi(x,a) \nabla_\theta \pi(a|x)$$

$$J(\theta) = v_{\pi_\theta}(s_0)$$

$$\nabla J(\theta) = \nabla v_{\pi_\theta}(s_0).$$

$$= \sum_s \sum_{k=0}^{\infty} \Pr(s_0 \to s, k, \pi) \sum_a q_\pi(s,a) \nabla \pi(a|s).$$

$$= \sum_s \eta_\pi(s) \sum_a q_\pi(s,a) \nabla \pi(a|s) \quad - (1)$$

η how often do you end up in state s when following policy π. & starting from s_0

$$= \mathbb{E}_{\pi}\left[\sum_a q_\pi(s,a)\, \nabla \pi(a|s)\right]$$

$$= \mathbb{E}_{\pi}\left[\sum_a q_\pi(s,a)\, \frac{\nabla_\theta \pi(a|s)}{\pi(a|s)}\, \pi(a|s)\right]$$

$$= \mathbb{E}_{\pi}\left[\sum_a q_\pi(s,a)\, \nabla \log \pi(a|s)\, \pi(a|s)\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi}\left[q_\pi(S,A)\, \nabla \log \pi(A|S)\right]$$

$\pi(a|s)$ function that outputs probability of executing action a when in state s.

REINFORCE [Williams 92]

Loop forever

generate an episode $S_0 A_0 \ldots S_T$

for each step t.

$$G_t \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_t$$

$$\theta \leftarrow \theta + \alpha\, G_t \nabla_\theta \ln \pi(A_t|S_t, \theta)$$

(b) REINFORCE w/ Baseline.

$$\mathbb{E}_\pi\left[q_\pi(s,a)\, \nabla \log \pi_\theta(a|s)\right]$$

$$= \mathbb{E}_\pi\left[(q_\pi(s,a) - b(s))\, \nabla \log \pi_\theta(a|s)\right]$$

$$\mathbb{E}_\pi(b(s)\, \nabla \log \pi_\theta(a|s)) = 0$$

$V(s)$

$$\sum_a b(s)\, \pi(a|s)\, \nabla \log \pi_\theta(a|s)$$

$$= \sum_a b(s)\, \nabla \pi_\theta(a|s)$$

$$= \nabla \sum_a b(s)\, \pi_\theta(a|s)$$

$$= \nabla \cdot (\underline{b(s)}) = 0.$$

$\underline{q_\pi(s,a)} - \underline{V_\pi(s)}$

= Advantage function

$A(s,a)$

$$= E_{\pi_\theta}\left[A(s,a)\, \nabla_\theta \log \pi(a|s) \right]$$

$\nabla_\theta J(\theta) = E_\pi\left(\underline{v_t}\, \nabla \log \pi_\theta(a|s) \right)$ REINFORCE

$= E_{\pi_\theta}\left(\underline{Q(s,a)}\, \nabla \log . \pi_\theta \right)$ Q <u>Actor</u> critic

$= E_{\pi_\theta}\left[\underline{A(s,a)}\, \nabla \log \pi_\theta \right)$ Advantage <u>actor critic</u>

A2C