Actor – Critic Algorithms

$V^*(s) = \max_a \sum_{s'} P(s'|s,a)\big(r + \gamma V^*(s')\big)$

$V^*_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a)[r + \gamma V^*_k(s')]$

$Q^*(s,a) = \sum_{s'} p(s'|s,a)\left[r + \gamma \max_{a'} Q^*(s',a')\right]$

$Q_{i+1}(s) \leftarrow \sum_{s'} P(s'|s,a,)\left[r + \gamma \cdot \max_{a'} Q_i(s',a')\right]$

Policy Evaluation

$V^{\pi_k}_{i+1}(s) = \sum_{s'} P(s'|s,\pi_k(s))\big[r + \gamma V^{\pi_k}_i(s')\big]$

Policy Improvement

$\pi_{k+1}(s) = \operatorname{argmax}_a \sum_{s'} P(s'|s,a)[r + \gamma V^{\pi_k}(s')]$

Q – Learning:

$Q_{i+1}(s,a) \leftarrow Q_i(s,a) + \alpha\Big(r + \gamma \cdot \max_{a'} Q_i(s',a') - Q_i(s,a)\Big)$

DQN Update:

$\theta_{t+1} = \theta_t + \alpha\big(y_t - Q(s_t,a_t;\theta)\big)\nabla_{\theta_t} Q(s_t,a_t;\theta_t)$

DQN Q Update:

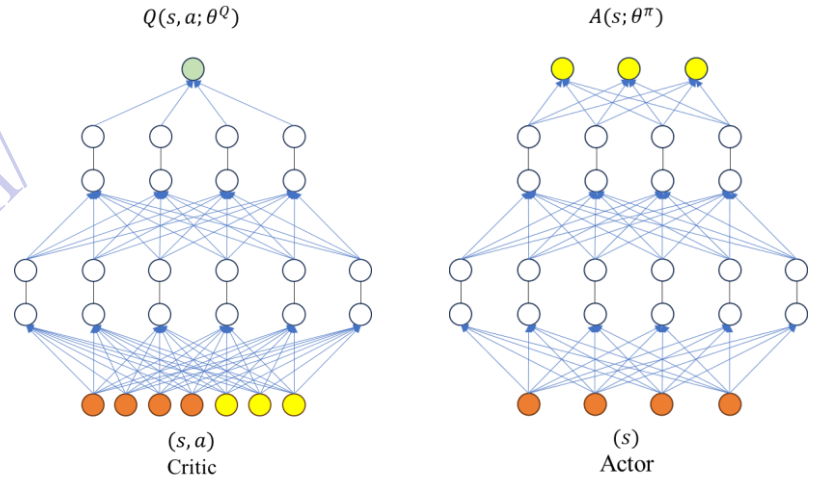
$Q(s_t,a_t;\theta_{t+1}) \approx Q(s_t,a_t;\theta) + \alpha\big(y_t - Q(s_t,a_t;\theta_t)\big)\left|\left|\nabla_{\theta_t} Q(s_t,a_t;\theta_t)\right|\right|^2_2$

n – step return:

$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^n \hat{v}(s_{t+n})$

Steady state distributions: $\mu_\pi(s)$

- Policy Gradient
  - Parameterized policy: $\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,b,\theta)}}$
  - Action preference, $h(s,a,\theta) = \theta^T x(s,a)$ (for discrete action)
- Steady state distribution, μ_π
  - Special distribution under which, if actions are selected according π, state distribution is not changed.
  - $\sum_s \mu_\pi(s) \sum_a \pi(a|s) p(s'|s,a) = \mu_\pi(s')$
- Policy gradient theorem
  - Expectation of sample gradient ∝ Actual gradient of the performance measure
  - $\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla\pi(a|s,\theta)$
  - $\sum_s \mu(s) \sum_a q_\pi(s,a) \nabla\pi(a|s,\theta)$
  - $\mathbb{E}_\pi[\sum_a q_\pi(s,a) \nabla\pi(a|s,\theta)]$
  - $\mathbb{E}_\pi\left[\sum_a \pi(a|s,\theta) \frac{q_\pi(s,a)\big(\nabla\pi(a|s,\theta)\big)}{\pi(a|s,\theta)}\right]$
  - $\mathbb{E}_\pi\left[\frac{q_\pi(s,a)\big(\nabla\pi(a|s,\theta)\big)}{\pi(a|s,\theta)}\right]$
  - $\mathbb{E}_\pi\left[G_t \frac{\nabla\pi(a|s,\theta)}{\pi(a|s,\theta)}\right]$



Algorithm 1: Actor – Critic Algorithms

Input: a differentiable policy parameterization $\pi(a|s,\theta)$
Input: a differentiable state – value function parameterization $\hat{v}(s,w)$
Parameters: step sizes $\alpha^\theta > 0, \alpha^w > 0$
Initialize policy parameter $\theta \in \mathbb{R}^{d'}$ and state – value weights $w \in \mathbb{R}^d$
For episode = 1, M **do**
  Initialize S
  $I \leftarrow 1$
  For t = 1, T **do**
    $a \sim \pi(\cdot|S,\theta)$
    Take action a, observe S', R
    $G_t \leftarrow R + \gamma\hat{v}(S',w)$ (if S' is terminal, then $\hat{v}(S',w) = 0$)
    $w \leftarrow w + \alpha^w G_t \nabla\hat{v}(S,w)$
    $\theta \leftarrow \theta + \alpha^\theta I G_t \nabla \ln\pi(a|S,\theta)$
    $I \leftarrow \lambda I$
    $S \leftarrow S'$
  end For
end For

https://www.profkumar.co.in/

Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.