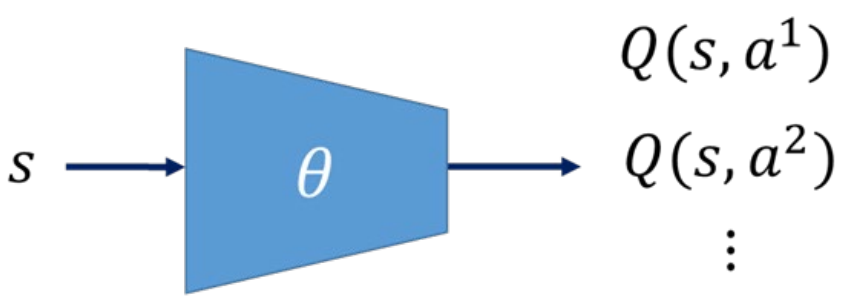
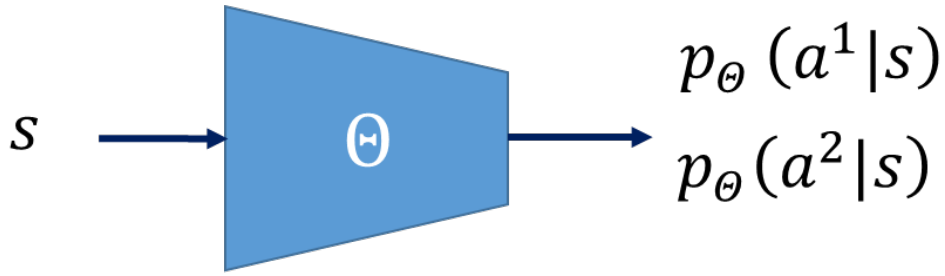
HW4 – Review and comparison of PG with DQN

- HW4 asks you to extend your class notes to review and compare policy-based vs value-based RL algorithms.
  - 1. What to learn?
  - 2. How to learn?
  - 3. How to calculate $\nabla \bar{R}_\Theta$?
  - 4. How to calculate $\nabla \bar{R}_\Theta$ efficiently?
  - 5. How to calculate $Q^*(s,a)$ from $Q^*(s',a')$?
  - 6. Q-learning
  - 7. How to stabilize DQN training?
  - 8. How to stabilize $\nabla \bar{R}_\Theta$ calculation?
- Due: next class meeting
- Upload ppt to Teams

1. What to learn?



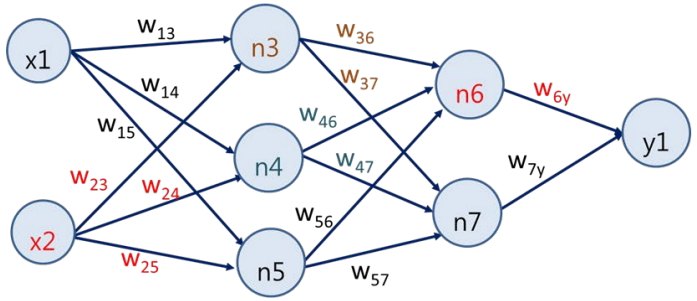
DQN



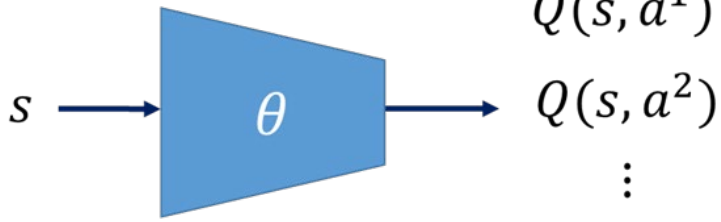
PPO

(week 7, 13 class notes)

2. How to learn?



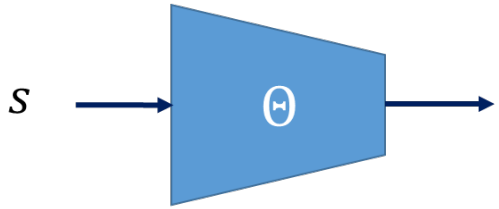
$$L = \frac{1}{N}\sum_{i=1}^{N}(y^i - \hat{y}^i)^2 \qquad \nabla L = [\frac{\partial L}{w_1}, \frac{\partial L}{b_1}, \dots] \qquad w_i \leftarrow w_i - \eta\frac{\partial L}{\partial w_i}$$



$$Q_\pi(s,a) = \mathbb{E}_\pi[R_{t+1} + \gamma Q_\pi(s_{t+1}, a_{t+1}) | s_t = s, a_t = a]$$

$$Loss = \left(R_s^a + \gamma \max_{a'} Q(s', a') - Q(s,a)\right)^2$$

DQN



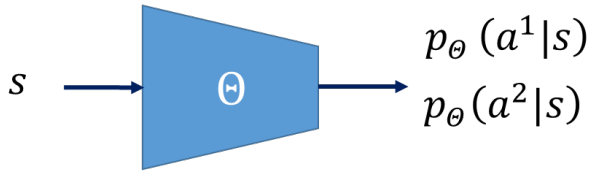
$$\bar{R}_\Theta = E_{\tau \sim p_\Theta(\tau)}[R(\tau)] \qquad R(\tau) = \sum_{t=1}^{T} r_t$$

$$\Theta^{\pi'} \leftarrow \Theta^{\pi} + \eta \nabla \bar{R}_\Theta$$

PPO

(week 5, 7, 13 class notes)

3. How to calculate $\nabla \bar{R}_\Theta$ in PG?



$$\Theta^{\pi'} \leftarrow \Theta^{\pi} + \eta \nabla \bar{R}_\Theta$$

$$\tau = (s_1, a_1, r_1, s_2, a_2, r_2, \cdots s_T, a_T)$$

$$p_\Theta(\tau) = p(s_1) p_\Theta(a_1|s_1) \cdots$$

$$R(\tau) = \sum_{t=1}^{T} r_t$$

$$\bar{R}_\Theta = \sum_{\tau} R(\tau) p_\Theta(\tau) = E_{\tau \sim p_\Theta(\tau)}[R(\tau)]$$

$$\nabla \bar{R}_\theta = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} R(\tau^n) \nabla \log p_\theta(a_t^n | s_t^n)$$

$$\nabla \bar{R}_\theta \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\sum_{t'}^{T_n} \gamma^{t'-t} r_{t'}^n - b \right) \nabla \log p_\theta(a_t^n | s_t^n)$$

$$A^\theta(s_t, a_t) = \left(\sum_{t'}^{T_n} \gamma^{t'-t} r_{t'}^n - b \right)$$

(PG, week 7 class notes)

4. How to calculate $\nabla \bar{R}_{\Theta}$ efficiently in PG?

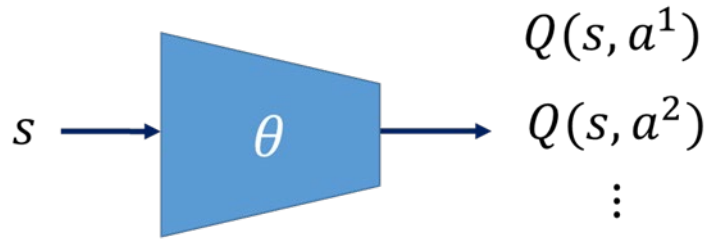
$$\nabla \bar{R}_{\Theta'} \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\sum_{t'}^{T_n} \gamma^{t'-t} r_{t'}^{n} - b \right) \nabla \log p_{\Theta'}(a_t^n | s_t^n)$$

$$\nabla \bar{R}_{\Theta'} = E_{(s_t, a_t) \sim \Theta} \left[\frac{p_{\Theta'}(a_t | s_t)}{p_{\Theta}(a_t | s_t)} A^{\Theta}(s_t, a_t) \nabla \log p_{\Theta}(a_t^n | s_t^n) \right]$$

$$\text{PO2}(\Theta') = \sum_{(s_t, a_t)} min \left(\frac{p_{\Theta'}(a_t | s_t)}{p_{\Theta}(a_t | s_t)} A^{\Theta}(s_t, a_t), clip \left(\frac{p_{\Theta'}(a_t | s_t)}{p_{\Theta}(a_t | s_t)}, 1 - \varepsilon, 1 + \varepsilon \right) A^{\Theta}(s_t, a_t) \right)$$

(PPO, week 8 class notes)

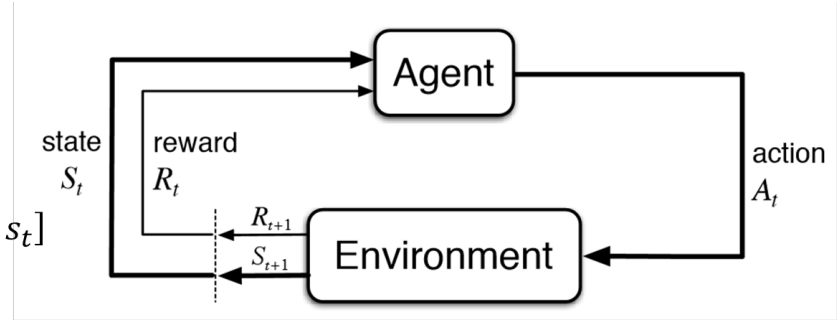
5. How to calculate $Q^*(s,a)$ from $Q^*(s',a')$?



$$Loss = \left(R_s^a + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)^2$$

Markov property

$p[s_{t+1}|s_t] = p[s_{t+1}|s_1, s_2, \dots s_t]$



Value function

$V_\pi(s) = \mathbb{E}_\pi[G_t|s_t = s]$

Markov process (chain)

$s_1, s_2, \dots s_t$

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$$V_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(s_{t+1})|s_t = s]$$

$$Q_\pi(s,a) = \mathbb{E}_\pi[R_{t+1} + \gamma Q_\pi(s_{t+1}, a_{t+1})|s_t = s, a_t = a]$$

$$V^*(s) = \max_a \sum P(s'|s,a)(R_s^a + \gamma V^*(s'))$$

$$Q^*(s,a) = \sum_{s'} P(s'|s,a)\,(R_s^a + \gamma \max_{a'} Q^*(s',a'))$$

Markov decision process

Bellman equation

(week 12 class notes)

6. Q-learning

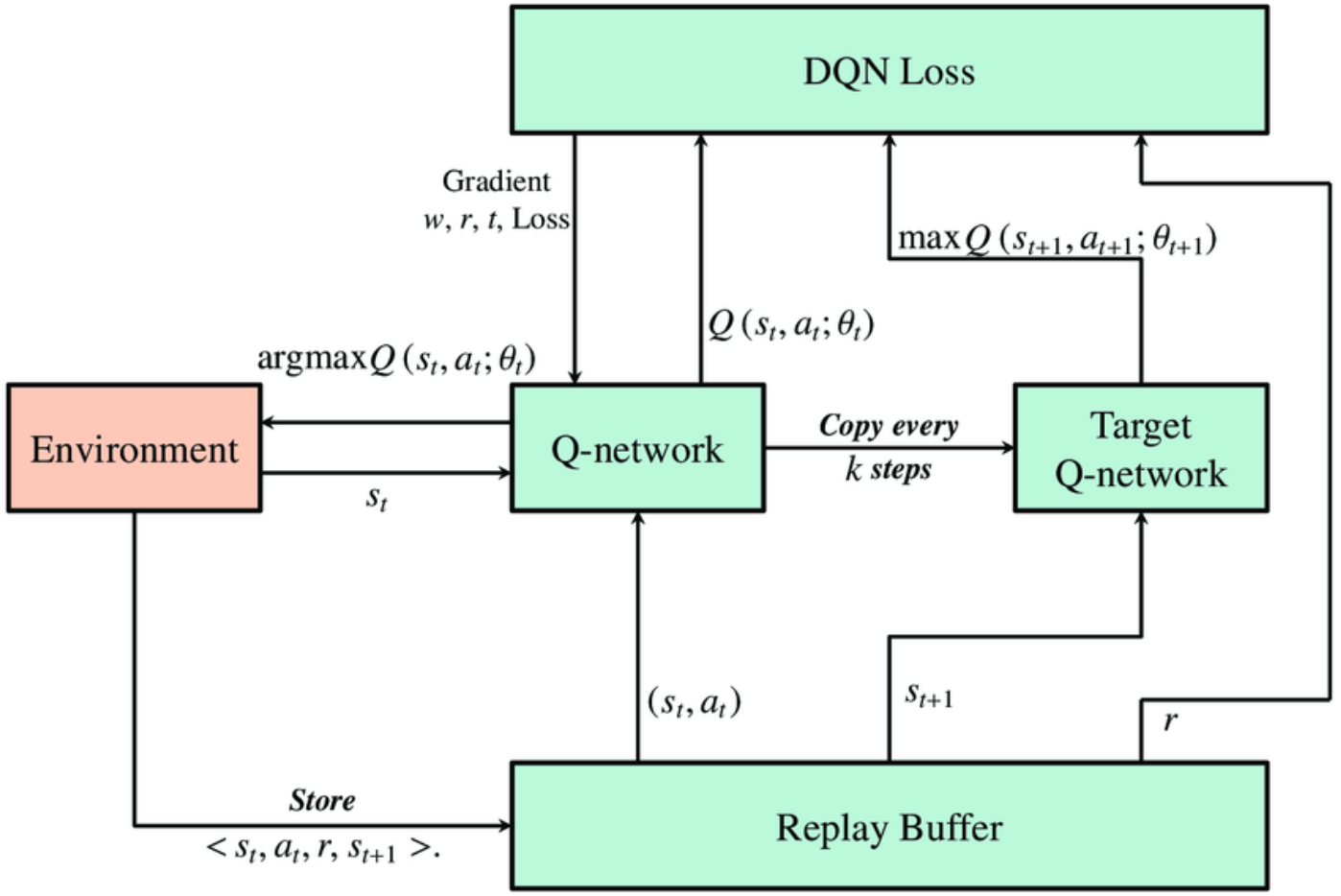
1. Initialize Table Q(S,A) with random values.
2. Take a action (A) with epsilon — greedy policy and move to next state S'
3. Update the Q value of a previous state by following the update equation:

$$Q(s,a) = Q(s,a) + \alpha(R_s^a + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

Q-learning

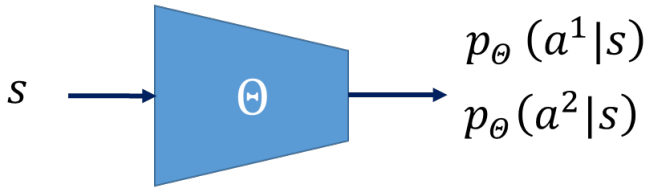
(week 13 class notes)

7. How to stabilize DQN training?



(week 13 class notes)

8. How to stabilize $\nabla \bar{R}_\Theta$ calculation?



$$\Theta^{\pi'} \leftarrow \Theta^{\pi} + \eta \nabla \bar{R}_\Theta$$

$$\nabla \bar{R}_\theta \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \left(\sum_{t'}^{T_n} \gamma^{t'-t} r_{t'}^n - b \right) \nabla \log p_\theta(a_t^n | s_t^n)$$

$V^{\pi_\theta}(s_t^n)$ Expected value of b

$E[G_t^n] = Q^{\pi_\theta}(s_t^n, a_t^n)$ Expected value of G_t^n

$$G_t^n = \sum_{t'}^{T_n} \gamma^{t'-t} r_{t'}^n$$

unstable when sampling amount is not large enough

$$Q^{\pi_\theta}(s_t^n, a_t^n) = \mathrm{E}[r_t^n + V^{\pi_\theta}(s_{t+1}^n)] = r_t^n + V^{\pi_\theta}(s_{t+1}^n)$$

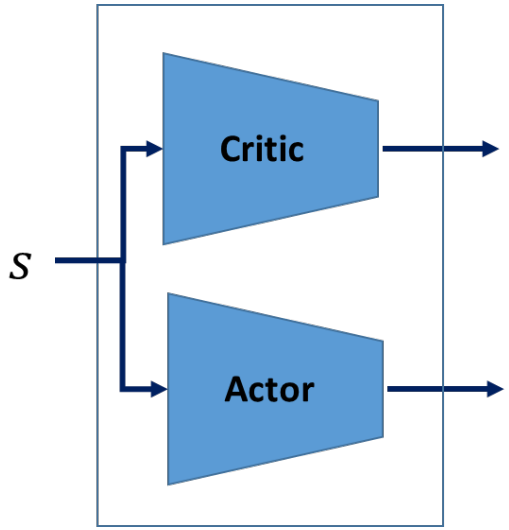
$$Q^{\pi_\theta}(s_t^n, a_t^n) - V^{\pi_\theta}(s_t^n) = r_t^n + V^{\pi_\theta}(s_{t+1}^n) - V^{\pi_\theta}(s_t^n)$$

$$A^\theta(s_t, a_t) = \left(r_t^n + V^{\pi_\theta}(s_{t+1}^n) - V^{\pi_\theta}(s_t^n) \right)$$

PPO-A2C

(week 9 class notes)

8. How to stabilize $\nabla \bar{R}_\Theta$ calculation?



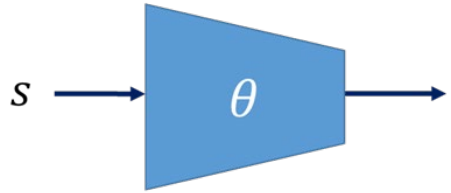
$V(s)$

$p_\Theta(a^1|s)$

$p_\Theta(a^2|s)$

PPO-A2C

$$Loss_v = \left(r_t^n + \gamma V^{\pi_\theta}(s_{t+1}^n) - V^{\pi_\theta}(s_t^n)\right)^2$$



$Q(s, a^1)$

$Q(s, a^2)$

$\vdots$

$$Loss = \left(R_s^a + \gamma \max_{a'} Q(s', a') - Q(s, a)\right)^2$$

(week 9 class notes)