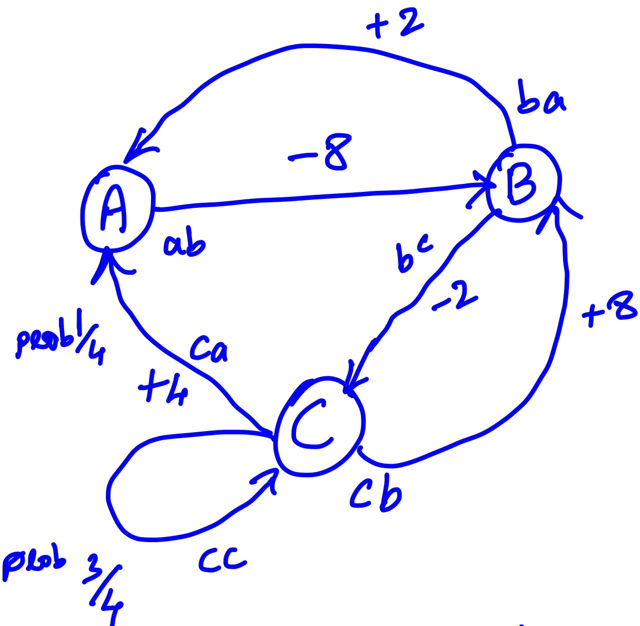

O
G
D
O
O
O
D
O

MDP $\gamma = 0.5$

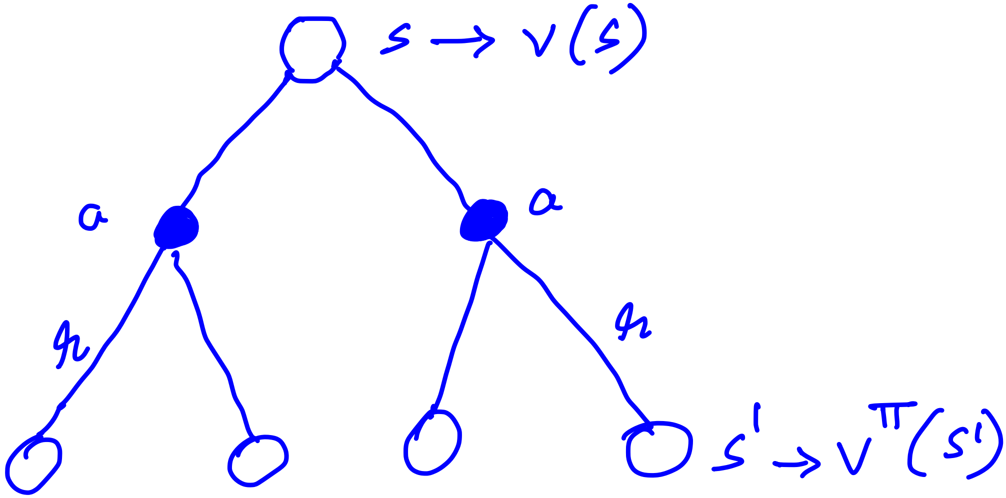


uniform random policy - all actions have equal probability

$V_1(A) = V_1(B) = V_1(C) = 2$

1 iteration of iterative policy evaluation (one back up for each state)
compute new value function $V_2(s)$

$V_2(A)$, $V_2(B)$, $V_2(C) = ?$



$$V_k^{\pi}(s) = E\left[r + \gamma\, V^{\pi}(s')\right]$$

$$= \sum_{a \in A} \pi(a|s)\left(r + \gamma \sum_{s' \in S} P(s'|s,a)\, V_{k-1}^{\pi}(s')\right)$$

policy evaluation

i/p MDP $\langle S, A, P, R, \gamma \rangle$
π policy

o/p $V^{\pi}(s)$

actions → ab, ba, bc, ca, cb

$V_1(s)$

A	B	C
2	2	2

$V_2(s)$

A	B	C
-7	-1	7

$$V_2(A) = -8 + 0.5(2) = -7$$

$$V_2(B) = 0.5\left(-2 + 0.5(2)\right)$$

$$+ 0.5\left(+2 + 0.5(2)\right)$$

$$= -\tfrac{1}{2} + \frac{3}{2} = +\tfrac{2}{2} = +1$$

$$V_2(B) = 0.5\left(8 + 0.5(2)\right)$$

$$+ 0.5\left(+4 + 0.5\left(0.25 * 2 + 0.75 * 2\right)\right)$$

$$= \frac{9}{2} + \frac{5}{2} = \frac{14}{2} = 7$$

Problem 2

Linear function approximation with Q learn using target n/w

$$W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} \in R^3$$

S & a $\in \{-1, 0, 1\}$

feature vector of

$$\phi = \begin{bmatrix} 2.5 \\ a \\ 0.5 \end{bmatrix}$$

$$q(s, a; w) = w^T \phi$$

$$= [w_0 \; w_1 \; w_2] \begin{bmatrix} 2.5 \\ a \\ 0.5 \end{bmatrix}$$

$$= W_0 * 2.5 + w_1 * a + w_2 * 0.5$$

2.

$$Q^{Target} = Q(s, a; w^-)$$

$$w^- = \begin{bmatrix} w_0^- \\ w_1^- \\ w_2^- \end{bmatrix}$$

$$y = r + \gamma \max_{a'} q(s, a; w^-)$$

$$J(w) = MSE\left(q(s, a; w) - y\right)$$

$$J(\omega) = \frac{1}{2}\left(q(s, a; \omega) - (r + \gamma \max_{a'} q(s, a; \bar{\omega}))\right)^2$$

$$J(\omega) = \frac{1}{2}\left(q(s, a; \omega) - y\right)^2$$

minimize this loss function

3. $w = \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix}$ $w^- = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$

Sample (s, a, s', a')

$(1, 0, 2, 2)$

$q(s, a; w) = w^T \phi = \begin{bmatrix} -2 & 1 & -1 \end{bmatrix} \begin{bmatrix} 2.s \\ a \\ 0.5 \end{bmatrix}$

$= -2 . 2.s + 1.a + -1 . 0.5$

$= -2 . 2.1 + 1.0 + -1 . 0.5$

$= -4 + 0 - 0.5 = -4.5$

$q(s', a'; w^-) = w^{-T} \phi = \begin{bmatrix} -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2.s \\ a \\ 0.5 \end{bmatrix}$

next action

$\alpha = 0.2$

assume

$\gamma = 0.9$

$a \in \{-1, 0, 1\}$ $s' = 2$

$a = -1$

$$q(s', a'; \bar{w}) = \bar{w}^T \phi = [-1\ 2\ 1] \begin{bmatrix} 2 \cdot s' \\ -1 \\ 0.5 \end{bmatrix}$$

$= -1 * 2 * 2 + 2 * -1 + 1 * 0.5$

$= -4 - 2 + 0.5 = \boxed{-5.5}$

$a = 0$

$$q(s', a'; \bar{w}) = \bar{w}^T \phi = [-1\ 2\ 1] \begin{bmatrix} 2 \cdot s' \\ 0 \\ 0.5 \end{bmatrix}$$

$= -1 * 2 * 2 + 0 + 0.5$

$= -4 + 0.5 = \boxed{-3.5}$

$a = 1$

$$q(s', a'; \bar{w}) = \bar{w}^T \phi = [-1\ 2\ 1] \begin{bmatrix} 2 \cdot s' \\ 1 \\ 0.5 \end{bmatrix}$$

$= -1 * 2 * 2 + 2 * 1 + 1 * 0.5$

$= -4 + 2 + 0.5 = \boxed{-1.5}$

$$\boxed{\max_{a'} q(s', a'; \bar{w}) = -1.5 \qquad a = 1}$$

$$y = r + \gamma \max_{a'} q(s', a'; w^-)$$

$$= 2 + 0.9 * -1.5$$

$$= \boxed{0.65}$$

$$\delta = -4.5 - 0.65$$

$$= -5.15$$

$$J(w) = \frac{1}{2}\left(q(s,a;w) - y\right)^2$$

$$\nabla_w J(w) = \left(q(s,a;w) - y\right) \nabla_w \left(q(s,a;w)\right)$$

$$= \delta * \nabla_w \left(w^T \phi(s,a)\right)$$

$$\nabla_w J(w) = \delta * \phi(s,a) = -5.15 \begin{bmatrix} 2.5 \\ a \\ 0.5 \end{bmatrix}$$

$$= -5.15 \begin{bmatrix} 2 * 1 \\ 0 \\ 0.5 \end{bmatrix} = \begin{bmatrix} -10.3 \\ 0 \\ -2.575 \end{bmatrix}$$

$$w \leftarrow w - \alpha \nabla_w J(w) * \delta$$

$$W_{new} = \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix} - (0.2) \begin{bmatrix} -10.3 \\ 0 \\ -2.575 \end{bmatrix} * -5.15$$

$$= \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix} + 1.03 \begin{bmatrix} -10.3 \\ 0 \\ -2.575 \end{bmatrix}$$

$$= \begin{bmatrix} -12.609 \\ 1 \\ -3.68727 \end{bmatrix}$$

Problem 3

Actor-critic algorithm

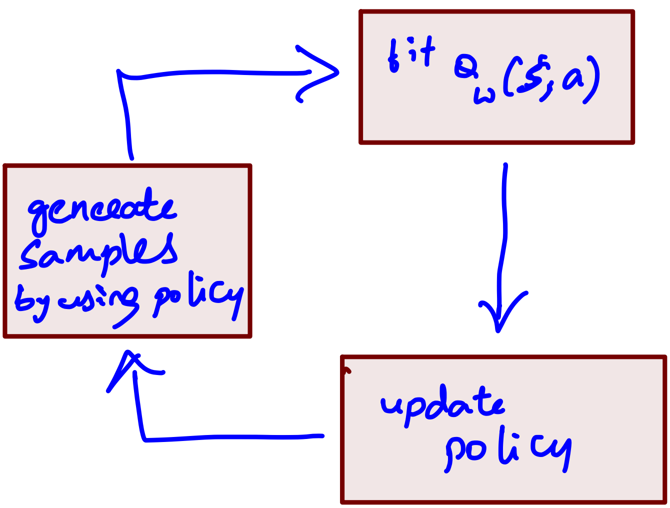
Actor

policy $\pi_\theta(a|s) \rightarrow \theta$ parameters

Critic

state-action value $Q_w(s,a) \rightarrow w$ parameters

online AC $\rightarrow$ update both $\pi_\theta(a|s)$ & $Q_w(s,a)$



$$\theta \Leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

for episodes $1 \to N$

step $i \to$ terminate

- sample $\{s_i, a_i\}$ using $\pi_\theta(a|s)$
- fit $Q_w(s,a)$
- evaluate $\hat{A}^\pi(s_i, a_i) = r(s_i, a_i) + \hat{Q}^\pi_\phi(s_i') - \hat{Q}^\pi_\phi(s_i)$
- $\nabla_\theta J(\theta) \approx \sum_i \nabla_\theta \log \pi_\theta(a_i|s_i) \hat{A}^\pi(s_i, a_i)$
- $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

loss func $= \frac{1}{2} \sum_i \| \hat{Q}^\pi_\phi(s_i, a_i) - y_i \|^2$