22-Sept-2018

Policies & Value functions

Given a policy π,

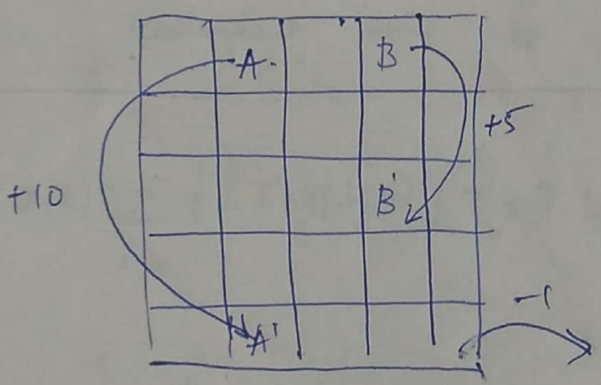
$$V_\pi(s) = E_\pi[G_t | S_t]$$

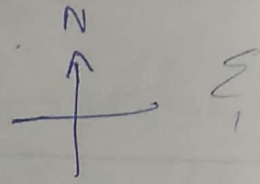
(State-Value f^n)
$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \forall s \in S$$

$$q_\pi(s,a) = E_\pi[G_t | S_t = s, A_t = a]$$

(action-value f^n)
$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

Example: Grid World:





- The cells of the grid correspond to the states of the MDP
- At each cell, four actions are possible N, S, E & W
- An action that takes the state out of the grid results in "state quo" for the state but gives reward of -1
- In states A & B any action takes one to states A' & B' respectively, with a reward of +10 & +5
- All actions in other states result in a reward of "0"

Suppose the agent selects all four actions with equal ~~distribution~~ prob. Suppose discount factor = 0.9

Let $V_\pi(s)$ be the value f^n under policy π

A B

3.3	8.8	4.4	$\boxed{5.3}$	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Value f^n table under policy π

$A \longrightarrow A'$ & $B \longrightarrow B'$ regardless of position of A & B.

$$\frac{1}{4}\left[(0,) + \gamma(v_\pi(s')\right.$$

$$0 + \gamma(v_\pi(s_2')$$

$$0 + \ldots \gamma$$

$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

$$= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s,a)[r + \gamma V_\pi(s')] \quad \forall s \in S$$

Note that, $V_\pi(A) = 8.8$ < immediate reward of 10

Also, $V_\pi(B) = 5.3$ > immediate reward of 5

(i) Suppose we know that $v_\pi(A') = -1.3$; what is $V_\pi(A)$

$$V_\pi(A) = r_\pi(A) + \gamma v_\pi(A') = 10 + 0.9(-1.3) = \approx 8.8$$

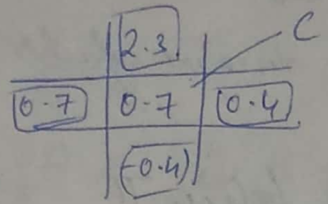
$$\because \sum_a \pi(a|s) \sum_{s'=A', r=R} p(s', r|s,a)[r + \gamma V_\pi(s')]$$

(ii) Suppose wkt, $V_\pi(B') = 0.4$

$$V_\pi(B) = r_\pi(B) + \gamma v_\pi(B')$$

$$= 5 + 0.9 \times 0.4 \approx 5.4$$

Center State



there are equal chance from each state to C

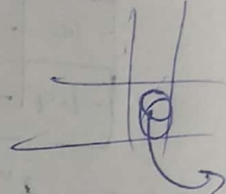
$\frac{1}{4}(0) + \frac{1}{4}(0) + \frac{1}{2}(0)$

$$V_\pi(C) = \underbrace{0}_{\text{Immedi}} + 0.9\left[\frac{1}{4}*2.3 + \frac{1}{4}*0.4 + \frac{1}{4}*(-0.4) + \frac{1}{4}*0.7\right]$$

$$= 0.9 * 0.75$$

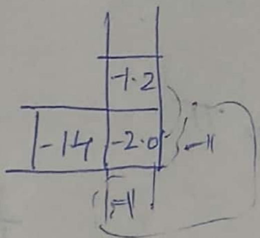
$$\approx 0.7$$

$\frac{1}{4}($ 0



Corner:

$V_\pi(E)$



outside

$$V_\pi(E) = \frac{1}{4}*(-1) + \frac{1}{4}*(-1) + \frac{1}{4}(0) + \frac{1}{4}(0)$$
$$+ 0.9\left[\frac{1}{4}*(-1.4) + \frac{1}{4}(-1.2) + \frac{1}{2}(V_\pi(E))\right]$$

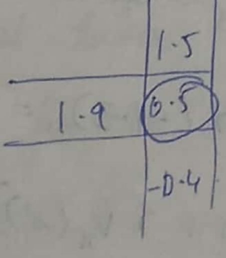
staying in same state if action is outside grid.

$$V_\pi(E) = -0.5 - 0.585 + 0.45*(V_\pi(E))$$

$$0.55\, V_\pi(E) = -1.085$$

$$V_\pi(E) \approx -2.03$$

Ex:



$$V_\pi(D) = \frac{1}{4}(-1) + 0.9\left(\frac{1}{4}(1.5) + \frac{1}{4}(1.9) + \frac{1}{4}(-0.4) + \frac{1}{4}V_\pi(D)\right)$$

$$= -0.25 + 0.9\left(\frac{1}{4}(3) + \frac{1}{4}V_\pi(D)\right)$$

~~$0.75\, V_\pi(D) = -0.25 + 0.675$~~ $\quad (-1 + 2.7)\frac{1}{4} = \frac{3.1}{4}(V_\pi(D)) \Rightarrow 0.566$

Scanned by CamScanner

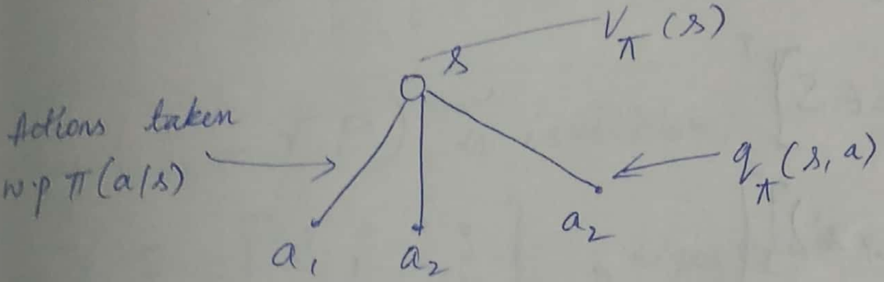
Bellman Eqn for $q_\pi(s,a)$

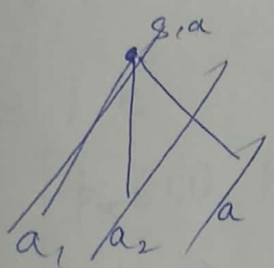
$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = a]$$

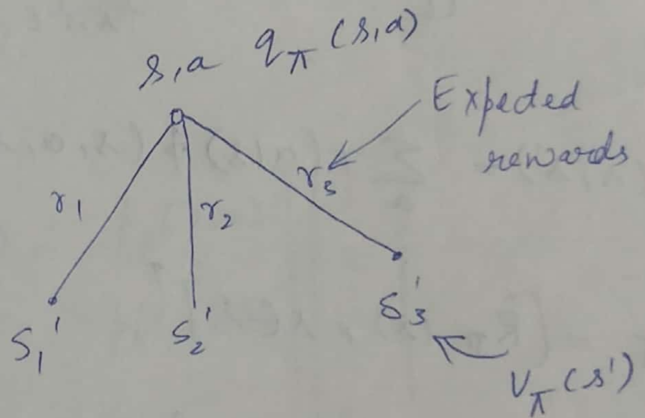
$$= \sum_{s',r} p(s', r \mid s,a)\,(r + \gamma V_\pi(s'))$$

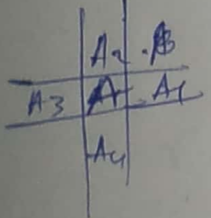
How is state value f^n $V_\pi(s)$ related to action-value f^n $q_\pi(s,a)$?

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



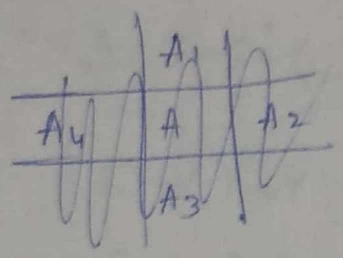






$$V_\pi(A) = \frac{1}{4} * \gamma * V_\pi(A_1) +$$
$$\frac{1}{4} * \gamma * V_\pi(A_2) +$$
$$\frac{1}{4} * \gamma * (V_\pi(A_3) +$$
$$\frac{1}{4} * \gamma\,(V_\pi(A_4))$$

Optimal Policies & Optimal Value functions



Value functions define a partial ordering over policies. A policy π is defined to be better or equal to policy π' if the expectation return under policy π is greater than or equal to the expected return under policy π'

$$\Rightarrow \pi \geq \pi' \quad \text{iff } V_\pi(s) \geq V_{\pi'}(s) \ \forall s \in S$$

$$V_\pi = \left[V_\pi(s), s \in S\right]^T$$

$$P_\pi = \left[\left[p_\pi(s, s')\right]\right]_{s, s' \in S}$$

$$p_\pi(s, s') = \sum_a \pi(a|s)\, p(s, a, s')$$

$$R_\pi = \left[R_\pi(s), s \in S\right]^T$$

$$V_\pi = R_\pi + \gamma P_\pi V_\pi$$

$$\text{or } (I - \gamma P_\pi) V_\pi = R_\pi$$

$$\text{or } V_\pi = (I - \gamma P)^{-1} R_\pi$$

$$R_\pi(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a)\, r$$

Value Iteration:

$$V_\pi(s) = R_\pi(s) + \gamma \sum_{s'} p_\pi(s,s')\, V_\pi(s')$$

Initialize $v_0(s) = 0 \;\forall s$

Iterate:

$$V_{n+1}(s) = R_\pi(s) + \gamma \sum_{s'} p_\pi(s,s')\, v_n(s') \;,\; n \geq 0$$

as $n \longrightarrow \infty$, $v_n(s) \longrightarrow V_\pi(s) \;\forall s \in S$

why $(I - \gamma P)$ is invertible,

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ & 1 & & 0 \\ 0 & & \cdot & 1 \\ 0 & 0 & 0 & \end{bmatrix} \quad \text{eigen values of } I = 1$$

$p_\pi e = 1 \cdot e$

$$\begin{bmatrix} p_\pi(1,1), & p_\pi(1,2) & \dots & p_\pi(1,n) \\ p_\pi(2,1) & \dots & & p_\pi(2,n) \\ & & & p_\pi(n,n) \\ p(n,1) & \dots & & \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \end{bmatrix}$$

$I - \gamma P$ is alway $\neq 0$ since $P = 1$ & $\gamma * P < 1$

$\therefore 1 - \gamma p > 0 \;\therefore\; I - \gamma P$ has no zero E.V hence invertible.

It turns out that there is always at least one policy that is better than or equal to all other policies. That policy will be optimal policy.

We denote all optimal policies by π_*.

The state value f^n for each optimal policy π_* is the same

- We ~~des~~ denote it by v_*

$$v_*(s) = \max_{\pi} v_\pi(s) \quad , \forall s \in S.$$

Optimal policies also share same optimal action value f^n

$q^*(s,a)$

$$q^*(s,a) = \max_{\pi} q_\pi(s,a)$$

$$q^*(s,a) = E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

~~[illegible]~~

$$v^* \equiv \text{optimal value function}$$

Under a given policy π, value function v_π satisfies the bellman equation.

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$

$$= \sum_{a \in A(s)} \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right] \quad s \in S$$

we expect the optimal value function v^* to also satisfy some form of the Bellman equation.

$$V_*(s) = \max_{a \in A(s)} q_{\pi*}(s,a)$$

$$= \max_{a \in A(s)} E_{\pi*} [G_t \mid S_t = s, A_t = a] = \max_{a \in A(s)} E\left[R_{t+1} + \gamma G_{t+1} \middle| S_t = s, A_t = a\right]$$

$$= \max_{a \in A(s)} E_{\pi*} [R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]$$

$$\boxed{V_*(s) = \max_{a \in A(s)} \sum_{s', r} p(s', r \mid s, a) [r + \gamma V_*(s')]}$$ BE for state f^n

The Bellman optimality equation for q^* can similarly written as

$$q^*(s,a) = E\left[R_{t+1} + \gamma \max_{a'} q^*(S_{t+1}, a') \middle| S_t = s, A_t = a\right]$$

This follows from,

$$q^*(s,a) = E\left[R_{t+1} + \gamma v^*(S_{t+1}) \middle| S_t = s, A_t = a\right]$$

Now $v^*(S_{t+1}) = \max_{a'} q^*(S_{t+1}, a')$

$$\boxed{q^*(s,a) = \sum_{s', r} p(s', r \mid s, a) [r + \gamma \max_{a'} q^*(s', a')]}$$ B.E for action value f^n

Note that, $v^*(s) = \max_{a \in A(s)} q^*(s,a)$

$$= \max_{a \in A(s)} \sum_{s', r} p(s', r \mid s, a) (r + \gamma \underbrace{\max_{a'} q^*(s, a')})$$

$$v^*(s) = \max_{a \in A(s)} \sum_{s', r} p(s', r \mid s, a) (r + \gamma v^*(s'))$$

Suppose we let,

$$q^*(s,a) = \sum_{s',r} p(s',r|s,a)\left(r + \gamma\, v^*(s')\right)$$

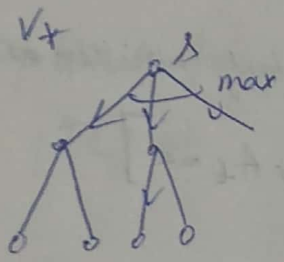
then,

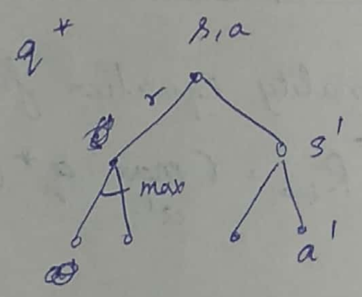
$$v^*(s) = \max_a q^*(s,a)$$

$\Rightarrow$

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

B E for action value function

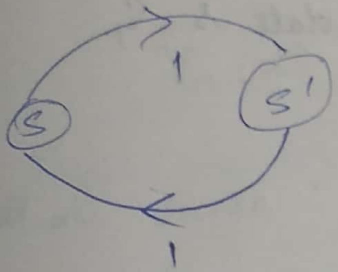




In R.L setup $q^*(s,a)$ is more favorable since the max is within the expectation. Since the system is unknown (Stochastic Approx Rule), E of system is not know Hence ~~E Sys~~ max (E) is possible to calculate. whereas E (max ( ) is possible via q^*

There always exists a policy that is better than or equal to other policies.

Eg: Consider the following MDP with transitions



Under two different policies π & π'

assume π & π' are deterministic policies for simplicity.

The difference in the policies is in terms of rewards obtained.

Note that,

$$V_\pi(s) = r_\pi(s) + \gamma V_\pi(s') \quad \text{—— (1)}$$

$$V_{\pi'}(s) = r_{\pi'}(s) + \gamma V_{\pi'}(s') \quad \text{—— (2)}$$

Suppose that,

$$V_\pi(s) > V_{\pi'}(s) \quad \text{—— (3)}$$

but

$$V_\pi(s') < V_{\pi'}(s') \quad \text{—— (4)}$$

From (1), (2) & (3), we have

$$r_\pi(s) + \gamma V_\pi(s') > r_{\pi'}(s) + \gamma V_{\pi'}(s') \quad \text{from (3)}$$

$$\Rightarrow r_\pi(s) > r_{\pi'}(s) \quad \text{—(5)} \quad \because V_\pi(s') < V_{\pi'}(s') \text{ from 4}$$

Now similar to (1)-(2) we can write,

$$V_\pi(s') = r_\pi(s') + \gamma V_\pi(s) \quad \text{—— (6)}$$

$$V_{\pi'}(s') = r_{\pi'}(s') + \gamma V_{\pi'}(s) \quad \text{—— (7)}$$

$$r_\pi(s') + \gamma V_\pi(s) < r_{\pi'}(s') + \gamma V_{\pi'}(s) \quad \text{—— (3)}$$

$$V_\pi(s) > V_{\pi'}(s)$$

$$\Rightarrow r_\pi(s') < r_{\pi'}(s') \quad \text{—— (4)}$$

It is optimal to pick action $\pi(s)$ in state s & action $\pi'(s')$ in state s'.

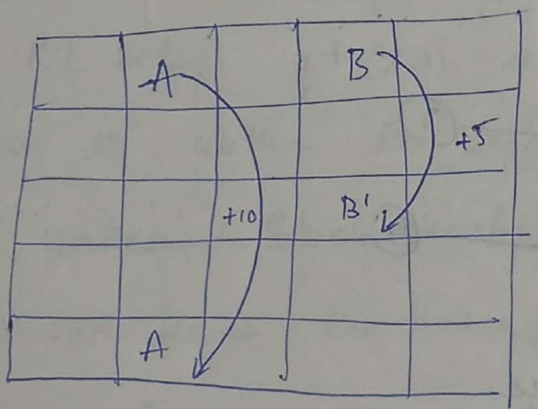
Let π'' be another policy st

$$\pi''(s) = \pi(s)$$

$$\& \quad \pi''(s') = \pi'(s')$$

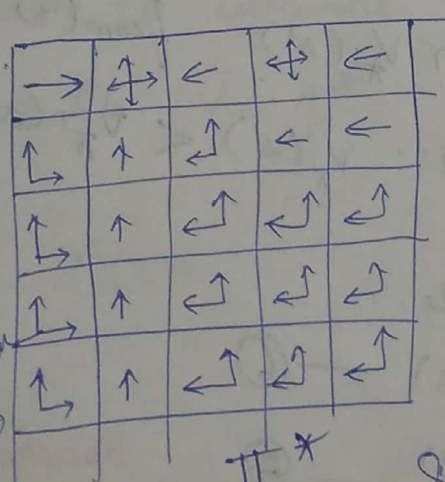
Thus one can construct a policy that is better than or equal to all other policies. In this case, $\underline{\underline{\pi''}}$

Grid World Example:



22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



'a' which max v^* → v_*

$V(s) = r + \gamma V(s')$

$0 + 0.9$

both action '↑' & '→' are optimal

π^*

So action need not be unique.

Init $v_0(s)$ $\forall s \in S$

Iterate

$$V_{n+1}(s) = \max_a \sum_{s',r} p(s',r|s,a)\,(r + \gamma\, v_n(s'))_{n \geq 0}$$

$$V_n(s) \longrightarrow V^*(s) \text{ as } n \longrightarrow \infty$$

Exercise: Implement value iterate for the grid example.

stop when,

$$\max_s |V_n(s) - V_{n-1}(s)| < 0.1$$

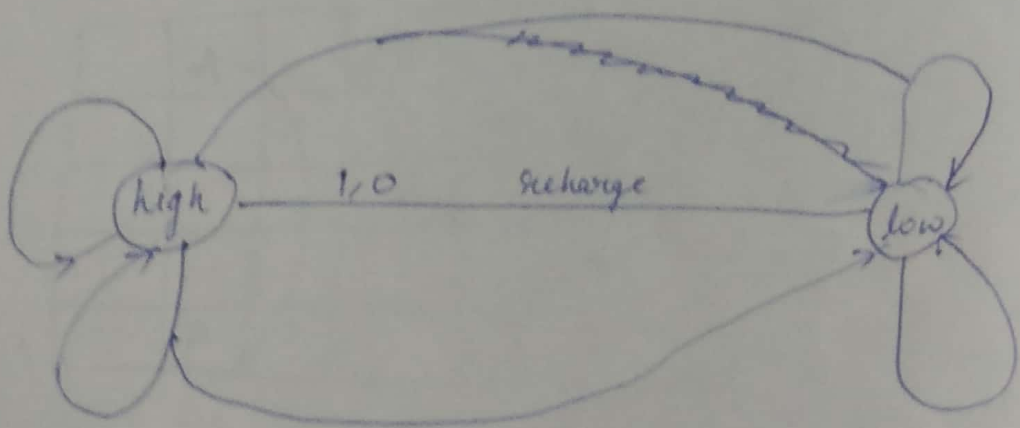
output $V_n(s)$ $\forall s$.

Kavitha

Scanned by CamScanner

Recycling Robot:

s	a	s'	$p(s' \mid s,a)$	$r(s,a,s')$
high	search	high	α	r_{search}
high	search	low	$1-\alpha$	r_{search}
low	search	high	$1-\beta$	-3
low	search	low	β	r_{search}
high	wait	high	1	r_{wait}
high	wait	low	0	r_{wait}
low	wait	high	0	r_{wait}
low	wait	low	1	r_{wait}
low	recharge	high	1	0
low	recharge	low	0	0



Bellman optimality eqn for the robot

high $\equiv h$, low $\equiv l$

search $\equiv s$, wait $\equiv w$, recharge $= r_e$

$$v_*(h) = \max\left\{ p(h|h,s)\,[r(h,s,h) + \gamma v_*(h)] + p(l|h,s)\,[r(h,s,l) + \gamma v_*(l)],\ p(h|h,w)\,[r(h,w,h) + \gamma(v_*(h))] + p(l|h,w)\,[r(h,w,l) + \gamma(v_*(l))] \right\}$$

$$= \max\big(\alpha[r_s + \gamma v_*(h)] + (1-\alpha)[r_s + \gamma v_*(l)],\ 1\,[r_w + \gamma v_*(h)] + 0\,[r_w + \gamma v_*(l)] \big)$$

$$= \max\left\{ \begin{array}{l} r_s + \gamma[\alpha v_*(h) + (1-\alpha) v_*(l)], \\ r_w + \gamma v_*(h) \end{array} \right\}$$

$$v_*(l) = \max\left\{ \begin{array}{ll} \beta r_s - 3(1-\beta) + \gamma[(1-\beta) v_*(h) + \beta(v_*(l)], & (\text{search}) \\ r_w + \gamma v_*(l), & (\text{wait}) \\ \gamma v_*(h) & (\text{recharge}) \end{array} \right\}$$