missed
Dyamic programing 6/oct/2018

$$V_\pi = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1} \mid S_t = s\right]$$

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

$$= \phi c$$

13/oct/2018 (exam)

Policy Improvement

We saw earlier Iterative policy evaluation.

Find value function for a given policy

a) Given value function for a policy, can one find better policy?

policy improvement

$$q_\pi(s,a) \doteq E\left[R_{t+1} + \gamma V_\pi(s,a) \mid S_t = s, A_t = a\right]$$

$$= \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_\pi(s')\right]$$

b) can we use $q_\pi(s,a)$ & $V_\pi(s)$ in order to find a better policy π'?

i.e., $V_{\pi'}(s) \geq V_\pi(s)$

Suppose in state s, we follow the action $\pi'(s)$ s.t

$q_\pi(s, \pi'(s)) \geq V_\pi(s)$ $\Rightarrow$ in state s you use π'

$$V_\pi(s) \leq q_\pi(s, \pi'(s))$$

$$= E_{\pi'}\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= E_{\pi'}\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s\right] \quad \text{--- (3)}$$

(3) action is selected according to π' in the beginning itself here $A_t = a$ absent

by definition

$$\pi'(\bar{s}) = \pi(\bar{s}) \quad \forall \bar{s} \neq s$$

$$\pi'(s) \neq \pi(s) \quad \text{--- (1)}$$

$$q_\pi(s, \pi'(s)) \geq V_\pi(s) \quad \text{--- (2)}$$

So when in s duj s_t we can use $q_\pi(s, \pi'(s))$

$$\leq E_{\pi'}(R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s)$$

$$= E_{\pi'}[R_{t+1} + \gamma E_{\pi'}[R_{t+2} + \gamma V_\pi(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1}), S_t = s]$$

$$\leq E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2}) \mid S_t = s]$$

$$\leq E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 (V_\pi(S_{t+3}) \mid S_t = s]$$

$$\leq E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} \cdots \mid S_t = s]$$

$$= V_{\pi'}(s) \qquad \because V_\pi(s) \triangleq E\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]$$

$$\pi'(s) = \arg\max_a q_{\pi'}(s, a)$$

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

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

Since model is known i.e. expectation are calculated. The exploration is not used

Policy improvement:

Form a new policy which is greedy wrt value f^n of old policy. Suppose new policy is such that

$$V_\pi(s) = V_{\pi'}(s) \quad \forall \; s \in S$$

In this case π & π' are both optimal. (2) 4

Also V_π is the optimal value f^n

$$V_{\pi'}(s) = \max_a E\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= \max_a E\left[R_{t+1} + \gamma V_{\pi'}(S_{t+1}) \mid S_t = s, A_t = a\right]$$

$$= \max_a \sum_{s', r} p(s', s | s, a)\left[r + \gamma V_{\pi'}(s')\right]$$

Thus $V_{\pi'}$ is the soln to bellman eqn.

$= V_{\pi'} = V_*$ optimal reward.

One can instead of deterministic policies, consider stochastic policies.

Suppose $A(s) \equiv$ set of feasible actions in state s

Let $A(s) = \{a_1, a_2, \ldots a_n\}$

Let's assume that a_i, a_j, a_k are the maximising actions

then a stochastic policy

$$\pi'(a|s) = \begin{cases} a_i & \text{w.p } p_1 \\ a_j & \text{w.p } p_2 \\ a_k & \text{w.p } p_3 \end{cases}$$

s.t $p_1 + p_2 + p_3 = 1$, $p_1, p_2, p_3 \geq 0$

Thus for non-optimal, i.e., those in $A(s)$ not $\{a_i, a_j, a_k\}$

we assign zero prob.

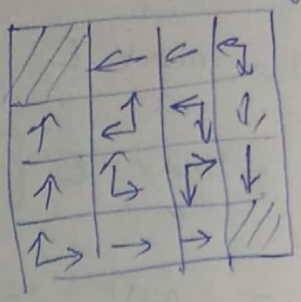
grid example

Random policy (π)

V_π

0	-14	-20	-22
-14	-18	-20	-20
-20	-20	-18	-14
-22	-20	-14	0

π' greedy



Question: What is $V_{\pi'}$?

$V_0 =$

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

$V_1 =$

0	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	0

$V_2 =$

0	-1	-2	-2
-1	-2	-3	-2
-2	-3	-2	-1
-3	-2	-1	0

$V_3 =$

0	-1	-2	-3
-1	-2	-3	-2
-2	-3	-2	-1
-3	-2	-1	0

V_4

0	-1	-2	-3
-1	-2	-3	-2
-2	-3	-2	-1
-3	-2	-1	0

ϵ has converged.

Find π'' - the greedy policy on $V_{\pi'}$

13/oct/2018

(notes in MBROI Book)