HW #2: RETRIEVE THE STATE-VALUE FUNCTION OF FIGURE 3.2

OBSERVE THAT THE BELLMAN EQUATION (3.14) MAY BE WRITTEN IN MATRIX FORM AS

$\boxed{v = Av + b}$ WHERE THE ELEMENT OF A CORRESPONDING TO STATE PAIR (s, s') IS

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

(FOR $\gamma = 1$, THIS MATRIX IS (RIGHT) STOCHASTIC, MEANING THAT EACH ROW IS A DISTRIBUTION. FOR SUCH MATRICES, ALL EIGENVALUES λ_i HAVE $|\lambda_i| \leq 1$)

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

ESPECIALLY FOR THE GRIDWORLD

$$A_{s,s'} = \gamma \sum_{a} \sum_{s'} \pi(a|s)\, p(s'|s,a) =$$

$$\gamma \sum_{a} \pi(a|s)\, 1_{s' = s'(a,s)}$$

$$b_s = \sum_{a} \pi(a|s)\, r(s,a)$$

THEREFORE:

$$v = Av + b \Leftrightarrow \quad Iv - Av = b \Leftrightarrow$$

$$v = (I - A)^{-1} b$$

THIS IS THE SOLUTION THAT THE ANN MUST PROVIDE

3.6 OPTIMAL POLICIES AND OPTIMAL VALUE FUNCTIONS

DEFINITION 1) A POLICY π' IS BETTER OR EQUAL THAN π IF

$v_{\pi'}(s) \geq v_{\pi}(s) \quad \forall s \in S$. SO THERE IS A PARTIAL ORDERING (AS OPPOSED TO A TOTAL ORDERING) AMONG POLICIES

2) A POLICY π' IS OPTIMAL IF IT IS BETTER THAN OR EQUAL TO ALL OTHER POLICIES.

THEOREM: THERE IS ALWAYS AT LEAST ONE OPTIMAL POLICY

COMMENTS: 1) NOT OBVIOUS AT ALL

2) PROOF IS BY SHOWING THE BELLMAN OPTIMALITY CONDITION (TO BE SHOWN LATER) IS A CONTRACTION MAPPING.

WE DEFINE

$$v_*(s) \triangleq \max_{\pi} v_{\pi}(s) = E_*[G_t \mid S_t = s] \quad \forall s \in S$$

$$q_*(s, a) \triangleq \max_{\pi} q_{\pi}(s, a) = E_*[G_t \mid S_t = s, A_t = a]$$
$$\forall s \in S, \forall a \in \mathcal{A}(s)$$

OPTIMAL RETURN IF WE TAKE ACTION a AND THEN BEHAVE OPTIMALLY

THE TWO ARE CONNECTED:

$$q_*(s,a) = E_*\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

$$= E_*\left[R_{t+1} \mid S_t = s, A_t = a\right]$$

$$+ \gamma \sum_{s',r} p(s',r \mid s,a) E_*\left[G_{t+1} \mid \cancel{S_t = s, A_t = a}, S_{t+1} = s', \cancel{R_{t+1} = r}\right]$$

$$= E_*\left[R_{t+1} \mid S_t = s, A_t = a\right]$$

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

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

$$\Rightarrow \boxed{q_*(s,a) = E_*\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = a, A_t = a\right]} \quad (A)$$

(INTUITIVELY CLEAR THAT THIS SHOULD HOLD)

WE CAN ALSO GIVE v_* IN TERMS OF q

$$\boxed{v_*(s) = \max_{a \in A(s)} q_*(s,a)} \quad (B)$$

INDEED, SKETCH OF PROOF:

$$v_*(s) = \max_\pi E_\pi\left[G_t \mid S_t = s\right] =$$

$$= \max_\pi \sum_{a \in A(s)} \pi(a \mid s) E_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

$$\leq q_{\pi_*}(s,a),$$

ACHIEVED BY MAXIMIZING THESE AND THEN PICKING ITS COEFFICIENT UNITY

SO WE PROVED $\leq$.

WE CAN EXCLUDE THE STRICT EQUALITY BY CONTRADICTION

BELLMAN OPTIMALITY CONDITION FOR $v_*(s)$

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

$$\stackrel{A}{=} \max_{a \in A(s)} E_* \left[R_{t+1} + \gamma v_*(S_{t+1}) \,\middle|\, S_t = s, A_t = a \right]$$

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

(3.19)

$$\Rightarrow \boxed{v_*(s) = \max_{a \in A(s)} \sum_{s',r} p(s',r|s,a) \left[r + \gamma v_*(s') \right]}$$

(nonlinear SYSTEM) (INTUITIVELY CLEAR. COMPARE WITH (3.14))

BELLMAN OPTIMALITY CONDITION FOR $q_*(s,a)$

$$q_*(s,a) \stackrel{A}{=} E_* \left[R_{t+1} + \gamma v_*(S_{t+1}) \,\middle|\, S_t = s, A_t = a \right]$$

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

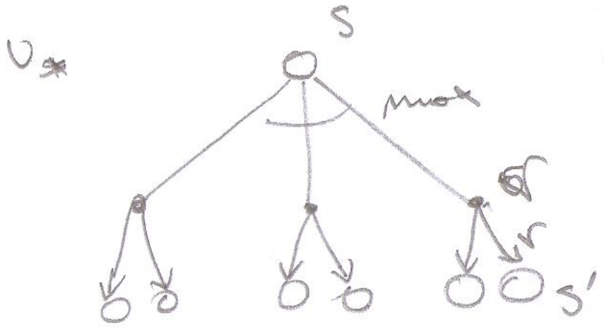
$$\stackrel{B}{=} \sum_{s',r} p(s',r|s,a) \left[r + \gamma \max_{a'} q(s',a') \right]$$

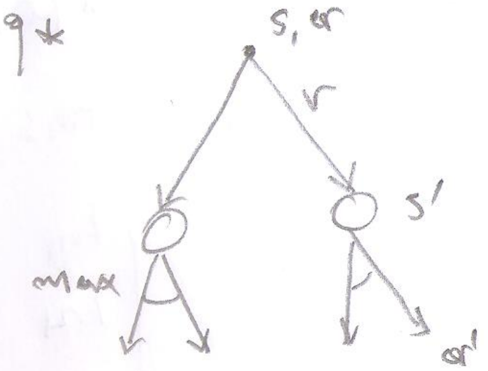
(ALSO INTUITIVELY CLEAR) (AGAIN, nonlin SYSTEM)

COMMENTS: 1) IN BOTH CASES, WE HAVE AS MANY EQUATIONS AS WE HAVE UNKNOWNS, BUT NOW SYSTEM IS NONLINEAR.

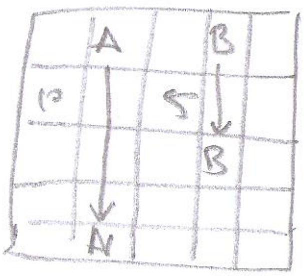
2) IF WE SOLVE THE EQUATIONS (EITHER SET), THE OPTIMAL POLICY IS TRIVIAL TO FIND.

3) BOTH SETS CAN BE DESCRIBED IN TERMS OF BACKUP DIAGRAMS.





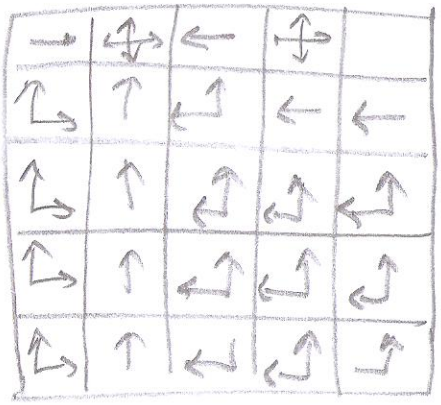
EXAMPLE 3.8 GRIDWORLD



10+14.4

22	24.4	22	19.4	17.5
19.8	22	19.8	17.8	16.
17.8	19.8	17.8	16	14.4
16	17.8	16.0	14.4	13
14.4	16	14.4	13	11.7

OBSERVE THAT THESE NUMBERS INDEED SATISFY BELLMAN'S OPTIMALITY CONDITION



CHAPTER 4: DYNAMIC PROGRAMMING

4.1 POLICY EVALUATION

PROBLEM: GIVEN A POLICY π, FIND ITS VALUE FUNCTION $v_\pi(s)$ FOR ALL $s \in \mathcal{S}$

WE HAVE BELLMAN' EQUATION

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

WE CAN WRITE THIS AS:

$$\begin{pmatrix} v_\pi(1) \\ v_\pi(2) \\ \vdots \\ v_\pi(n) \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix} \begin{pmatrix} v_\pi(1) \\ v_\pi(2) \\ \vdots \\ v_\pi(n) \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

$$\Leftrightarrow \quad v = Av + b \Leftrightarrow Iv - Av = b \Leftrightarrow (I - A)v = b$$

$$\Leftrightarrow \quad \boxed{v = (I-A)^{-1} b}$$

SO WE CAN FIND THE VALUE FUNCTION BY SOLVING A LINEAR SYSTEM.

THERE IS ANOTHER WAY, WHICH WORKS BETTER AND IS MORE GENERALIZABLE:

ITERATIVE POLICY EVALUATION ALGORITHM

1) WE SET $V_0(s)$ ARBITRARILY

(BUT IF THERE IS A TERMINAL STATE S_T, THEN WE SET $V_0(S_T) = 0$, BECAUSE THIS IS THE RIGHT VALUE, AND THE ALGORITHM WILL NOT UPDATE IT)

2) SET $v_{t+1}(s) = \sum_{a} \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_t(s')\right]$

THEOREM IF $\gamma < 1$ OR EVENTUAL TERMINATION IS GUARANTEED (FOR EPISODIC TASKS) THEN ALGORITHM IS WILL CONVERGE TO SOLUTION

SKETCH OF PROOF FOR $\gamma < 1$ CASE:

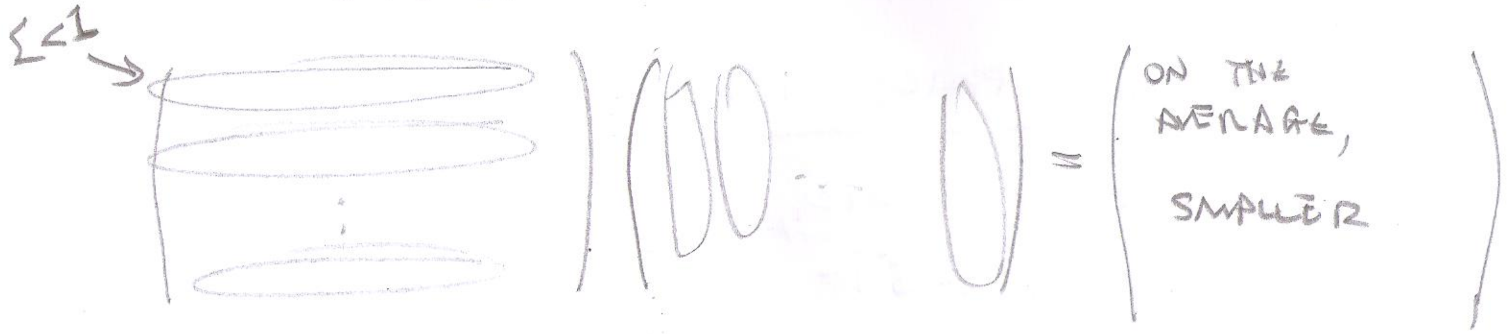
$$V_1 = AV_0 + b, \qquad V_2 = A(AV_0 + b) + b = A^2V_0 + Ab + b$$

$$V_3 = A(AV_0 + b) + b = A^3V_0 + A^2b + Ab + b \ \ldots$$

$$V_t = (A^{t-1} + A^{t-2} + \ldots + A + I)b + A^tV_0$$

HOWEVER, $\gamma < 1 \Rightarrow$ NORM $|A| < 1$

INDEED CONSIDER A^2:



IT FOLLOWS THAT

$$V_\infty = \left(\sum_{t=0}^{\infty} A^t\right) b$$

HOWEVER, WE KNOW THAT:

$$\sum_{t=0}^{\infty} A^t = (I - A)^{-1} \qquad \text{(NEUMANN SERIES)}$$

WHICH CAN BE PROVEN SIMILARLY TO

$$\sum_{t=0}^{\infty} r^t = \frac{1}{1-r}$$

WE CAN IMPROVE SPEED OF CONVERGENCE IF WE USE VALUES THE MOMENT WE COMPUTE THEM:

ITERATIVE POLICY EVALUATION (pg. 75)

INPUT: π, θ (FOR TERMINATION CONDITION),

$V(s)$ (ARBITRARY, BUT $V(\text{TERMINAL}) = 0$)

LOOP:

$\Delta \leftarrow 0$

LOOP FOR EACH $s \in \mathcal{S}$:

$v \leftarrow V(s)$

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

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

UNTIL $\Delta < \theta$

4.2 POLICY IMPROVEMENT

GIVEN A SPECIFIC POLICY, WE KNOW THAT:

$$\left.\begin{aligned} v_\pi(s) &\overset{\Delta}{=} \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s] \\ &= \sum_{a} \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right] \end{aligned}\right\} \quad (1)$$

$$q_\pi(s,a) \overset{\Delta}{=} \mathbb{E}_\pi\left[G_t | S_t = s, A_t = a\right]$$

$$= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a\right]$$

$$= \mathbb{E}_\pi\left[R_{t+1} | S_t = s, A_t = a\right] + \gamma \mathbb{E}_\pi\left[G_{t+1} | S_t = s, A_t = a\right]$$

$$= E[R_{t+1} \mid S_t = s, A_t = a]$$

$$+ \gamma \sum_{s',r} p(s', r \mid s, a) \, E_\pi[G_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s', R_{t+1} = r]$$

$$= E[R_{t+1} \mid S_t = s, A_t = a]$$

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

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

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

$$\Rightarrow \boxed{\begin{aligned} q_\pi(s, a) &= E[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')] \end{aligned}} \quad (2)$$

LET US WRITE (1) AND (2) FOR <u>DETERMINISTIC</u> POLICIES

$$v_\pi(s) = \sum_{s',r} p(s', r \mid s)[r + \gamma v_\pi(s')] \quad (1')$$

NEW NOTATION

$$q_\pi(s, \pi(s)) = \sum_{s',r} p(s', r \mid s, \pi(s))[r + \gamma v_\pi(s')] \quad (2')$$

ALSO OBSERVE THAT THEN

$$v_\pi(s) = q_\pi(s, \pi(s))$$

POLICY IMPROVEMENT THEOREM (SPECIAL CASE)

LET π, π' DETERMINISTIC POLICIES SUCH THAT:

$$\forall s \in \mathcal{S} \quad q_\pi(s, \pi'(s)) \geq v_\pi(s) \quad \text{(A)}$$

THEN IT FOLLOWS THAT

$$v_{\pi'}(s) \geq v_\pi(s) \quad \text{(B)}$$

MOREOVER, IF (A) IS STRICT FOR SOME s, THEN IT IS ALSO STRICT FOR THIS s.

(OBSERVE THAT IT MAKES SENSE!)

(COMPARE MM)

PROOF:

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

$$= \sum_{s', r} P_{S_{t+1}, R_{t+1} \mid S_t, A_t}(s', r \mid s, \pi'(s)) \left[r + \gamma v_\pi(s')\right]$$

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

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

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

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

(NOTE:

$$E[X + E[Y \mid Z]] = E[E[E[X \mid Z] + E[Y \mid Z]]] =$$

$$E[E[X + Y \mid Z]]$$
)