How good is a policy? $\rightarrow$ look at value / utility

$$V(s_0, s_1, \ldots s_T) = \sum_{t=0}^{T} R(s_t)$$

$$V(s_0, s_1, \ldots s_T) = \sum_{t=0}^{T} \gamma^t R(s_t) \quad \gamma \in [0,1]$$

$$V(s)^\pi = E_{Pr(s_0, s_1, \ldots \mid s_0 = s, \pi)} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \right]$$

$$\pi^*_{(s)} = \underset{\pi}{\operatorname{argmax}}\ V(s)^\pi \longrightarrow \pi^*_{(s)} = \underset{a \in A(s)}{\operatorname{argmax}} \sum_{s'} P(s' \mid s, a)\, Value(s')$$

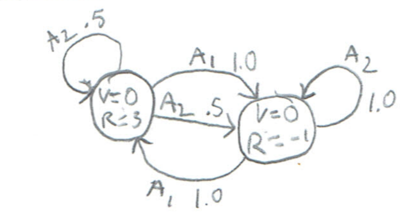
$$V(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s' \mid s, a) V(s') \qquad \text{Bellman equation}$$

For $i = 1$ to <stopping criteria>

$$V_{i+1}(s) \longleftarrow R(s) + \gamma \max_{a \in A(s)} \sum_{s'} P(s' \mid s, a) V(s') \qquad \text{Value iteration}$$

$i \leftarrow i+1$

Ex 1



$S = \{s_1, s_2\}$

$A = \{a_1, a_2\}$

$R(s_1) = 3$

$R(s_2) = -1$

$\gamma = .5$

$$T = \left[\left[\left[P(s_1 \mid s_1, a_1), P(s_1 \mid s_1, a_2) \right], \left[P(s_1 \mid s_2, a_1), P(s_1 \mid s_2, a_2) \right] \right], \left[\left[P(s_2 \mid s_1, a_1), P(s_2 \mid s_1, a_2) \right], \left[P(s_2 \mid s_2, a_1), P(s_2 \mid s_2, a_2) \right] \right] \right]$$

$i = 1, \ V_0(s_1) = 0, \ V_0(s_2) = 0$

$$V_1(s_1) = R(s_1) + \gamma \max_{a \in A(s)} \sum_{s'} P(s' \mid s_1, a) V(s')$$

$$= 3 + .5 * \max \begin{bmatrix} \begin{pmatrix} P(s' = s_1 \mid s_1, a = a_1) V(s_1) \\ + \\ P(s' = s_2 \mid s_1, a = a_1) V(s_2) \end{pmatrix} \\ \begin{pmatrix} P(s' = s_1 \mid s_1, a = a_2) V(s_1) \\ + \\ P(s' = s_2 \mid s_1, a = a_2) V(s_2) \end{pmatrix} \end{bmatrix} \begin{matrix} \rightarrow 0 \cdot V(s_1) \rightarrow 0 \cdot 0 = 0 \\ \rightarrow 1.0 \cdot V(s_2) \rightarrow 1.0 \cdot 0 = 0 \\ \rightarrow .5 \cdot V(s_1) \rightarrow .5 \cdot 0 = 0 \\ \rightarrow .5 \cdot V(s_2) \rightarrow .5 \cdot 0 = 0 \end{matrix} \begin{matrix} \left. \begin{matrix} \\ \end{matrix} \right\} 0 + 0 = 0 \\ \left. \begin{matrix} \\ \end{matrix} \right\} 0 + 0 = 0 \end{matrix}$$

$$= 3$$

$$\boxed{\begin{matrix} V_1(s_1) = 3 \\ V_1(s_2) = -1 \end{matrix}}$$

$\longrightarrow$ * when $V_0(s) = 0$, $V_1(s) = R(s)$, therefore $V_1(s_2) = R(s_2) = -1$

$$T = \left[\begin{array}{c} \begin{array}{cccc} 111 & 112 & 121 & 122 \end{array} \\ \left[[0, .5], [1.0, 0]\right], \\ \left[[1.0, .5], [0, 1.0]\right] \\ \begin{array}{cccc} 211 & 212 & 221 & 222 \end{array} \end{array} \right]$$

$i=2$, $V_1(s_1)=3$, $V_1(s_2)=-1$

$$V_2(s_1) = R(s_1) + \gamma \max_{a\in A(s)} \left[\sum_{s'}^{s'\in S} P(s'|s_1,a)V(s') \right]$$

$$= 3 + .5 \cdot \max \begin{bmatrix} P(s'=s_1|s_1,a=a_1)V(s_1) + P(s'=s_2|s_1,a=a_1)V(s_2), \\ P(s'=s_1|s_1,a=a_2)V(s_1) + P(s'=s_2|s_1,a=a_2)V(s_2) \end{bmatrix}$$

s' number, state s number, action number

$$= 3 + .5 \cdot \max \begin{bmatrix} T(1,1,1)V(s_1) + T(2,1,1)V(s_2), \\ T(1,1,2)V(s_1) + T(2,1,2)V(s_2) \end{bmatrix}$$

$$= 3 + .5 \cdot \max(0\cdot 3 + 1.0\cdot -1,\ .5\cdot 3 + .5\cdot -1)$$
$$= (-1,\ 1)$$

$$= 3 + .5(1) = 3.5$$

$$\boxed{V_2(s_1) = 3.5}$$

$$V_2(s_2) = R(s_2) + \gamma \max_{a\in A(s)} \left[\sum_{s'}^{s'\in S} P(s'|s_2,a)V(s') \right]$$

$$= -1 + .5 \max \begin{bmatrix} P(s'=s_1|s_2,a=a_1)V(s_1) + P(s'=s_2|s_2,a=a_1)V(s_2), \\ P(s'=s_1|s_2,a=a_2)V(s_1) + P(s'=s_2|s_2,a=a_2)V(s_2) \end{bmatrix}$$

$$= -1 + .5 \max \begin{bmatrix} T(1,2,1)V(s_1) + T(2,2,1)V(s_2), \\ T(1,2,2)V(s_1) + T(2,2,2)V(s_2) \end{bmatrix}$$

$$= -1 + .5\cdot\max(1.0\cdot 3.5 + 0\cdot -1,\ 0\cdot 3.5 + 1.0\cdot -1)$$
$$= (3.5,\ -1)$$

$$= -1 + .5\cdot 3.5 = .75$$

$$\boxed{V_2(s_2) = .75}$$

$i \in$ epochs (iteration)

$j \in |S|$ (current state)

$k \in |S|$ (next state)

$n \in |A|$ (current action)

$T(1,1,1)=0 \quad T(2,1,1)=1.0$

$T(1,1,2)=.5 \quad T(2,1,2)=.5$

$T(1,2,1)=1.0 \quad T(2,2,1)=0$

$T(1,2,2)=0 \quad T(2,2,2)=1.0$

$V_2(s_1)=3.5$

$V_2(s_2)=.75$

$$V_i(s_j)=R(s_j)+\gamma \max_{a_n}^{a_n \in A(s)}\left[\sum_{s_k} P(s_k|s_j,a_n)V(s_k)\right]$$

$$V_i(s_j)=R(s_j)+\gamma \max \left[\begin{array}{l} \underbrace{\underset{k=1,\,n=1}{P(s_k=s_1|s_j,a_n=a_1)V(s_k)}+\underset{k=2,\,n=1}{P(s_k=s_2|s_j,a_n=a_1)V(s_k)}}_{\text{Summation for each possible next state}} \\ P(s_k=s_1|s_j,a_n=a_2)V(s_k)+P(s_k=s_2|s_j,a_n=a_2)V(s_k) \end{array}\right]$$

for each action

$$V_i(s_j)=R(s_j)+\gamma\max\left[\begin{array}{l} T(\overset{k}{1},j,\overset{n}{1})V(s_{\overset{k}{1}})+T(\overset{k}{2},j,\overset{n}{1})V(s_{\overset{k}{2}}), \\ T(1,j,2)V(s_1)+T(2,j,2)V(s_2) \end{array}\right]$$

$$V_3(s_{\overset{j=1}{1}})=3+.5\cdot\max\left[\begin{array}{l} \overset{k=1,j=1,n=1}{T(1,1,1)V(s_1)}+\overset{k=2,j=1,n=1}{T(2,1,1)V(s_2)}, \\ \overset{k=1,j=1,n=2}{T(1,1,2)V(s_1)}+\overset{k=2,j=1,n=2}{T(2,1,2)V(s_2)} \end{array}\right]$$

$$V_3(s_1)=3+.5\cdot\max\left[\underset{.75}{0\cdot3.5+1.0\cdot.75},\ \underset{2.125}{.5\cdot3.5+.5\cdot.75}\right]$$

$$\boxed{V_3(s_1)=4.0625}$$

$$V_3(s_2)=-1+.5\cdot\max\left[\begin{array}{l} T(1,2,1)V(s_1)+T(2,2,1)V(s_2), \\ T(1,2,2)V(s_1)+T(2,2,2)V(s_2) \end{array}\right]$$

$$=-1+.5\cdot\max\left[\underset{4.0625}{1.0\cdot4.0625+0\cdot.75},\ \underset{.75}{0\cdot4.0625+1.0\cdot.75}\right]$$

$$\boxed{V_3(s_2)=1.03125}$$

<stopping criteria> $\approx |V_i - V_{i-1}| < .1$ or $.05$ etc

$V_0(S_1 \ldots S_{|S|}) = 0$

for $i = 1$ to <stopping criteria>

    for $j = 1$ to $|S|$ // for all states

        vals $\leftarrow \{\}$

        for $n = 1$ to $|A|$ // for all valid actions in S_j

            $\text{vals}_n \leftarrow 0$

            for $k = 1$ to $|S|$ // for all reachable states from S_j

                $\text{vals}_n \leftarrow \text{vals}_n + T(k, j, n)\, V_{i-1}(S_k)$

        $V_i(S_j) \leftarrow R(S_j) + \gamma \cdot \max(\text{vals})$

Value iteration

---

now what? How to use values to generate policy?