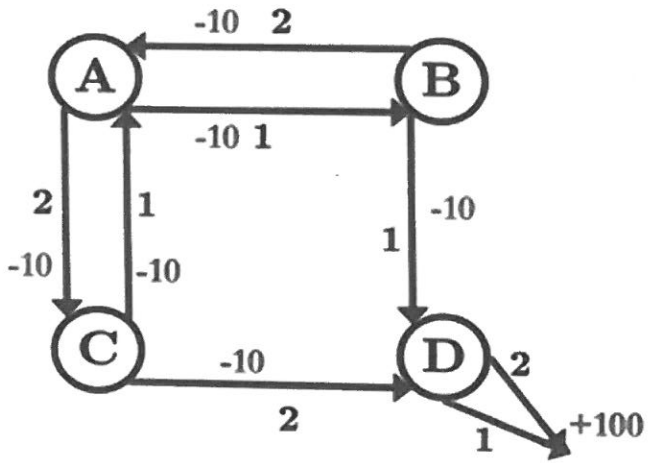
EXAMPLES



Goal = Terminal state

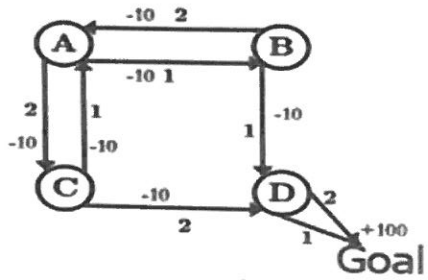
- 4 states
- 2 possible actions in each state. [E.g in A: 1) go to B or 2) go to C]
- P(s' | s, a) = (0.9 , 0.1) with 10% we go to a wrong direction

Rewards associated with each state-action pair are given on the arcs.

Calculating state values for a policy π_1:

π_1 : always choosing Action 1

$$\pi(1 \mid A) = 1 \qquad \pi(2 \mid A) = 0$$
$$\pi(1 \mid B) = 1 \qquad \pi(2 \mid B) = 0$$
$$\pi(1 \mid C) = 1 \qquad \pi(2 \mid C) = 0$$
$$\pi(1 \mid D) = 1 \qquad \pi(2 \mid D) = 0$$



$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} P(s'|s,a)[R(s,a) + \gamma V^{\pi}(s')]$$

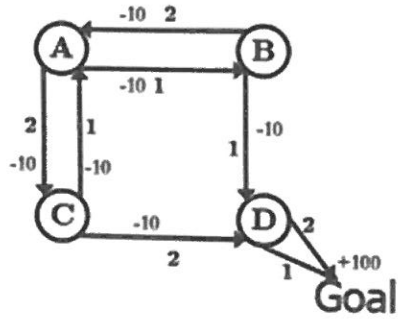
$$V^{\pi}(D) = 100$$

$$\begin{aligned}
V^{\pi}(B) = \quad & 1 \cdot 0.9 \cdot [-10 + V^{\pi}(D)] \\
+ \quad & 1 \cdot 0.1 \cdot [-10 + V^{\pi}(A)] \\
V^{\pi}(C) = \quad & 1 \cdot 0.9 \cdot [-10 + V^{\pi}(A)] \\
+ \quad & 1 \cdot 0.1 \cdot [-10 + V^{\pi}(D)] \\
V^{\pi}(A) = \quad & 1 \cdot 0.9 \cdot [-10 + V^{\pi}(B)] \\
+ \quad & 1 \cdot 0.1 \cdot [-10 + V^{\pi}(C)]
\end{aligned}
\Longrightarrow
\begin{aligned}
V^{\pi}(A) &= 75.61 \\
V^{\pi}(B) &= 87.56 \\
V^{\pi}(C) &= 68.05 \\
V^{\pi}(D) &= 100
\end{aligned}$$

Calculating state values for a policy π_2:

π_2 : always choosing Action 2

$$\pi(1 \mid A) = 0 \qquad \pi(2 \mid A) = 1$$
$$\pi(1 \mid B) = 0 \qquad \pi(2 \mid B) = 1$$
$$\pi(1 \mid C) = 0 \qquad \pi(2 \mid C) = 1$$
$$\pi(1 \mid D) = 0 \qquad \pi(2 \mid D) = 1$$



Similarly as before:

$$V^{\pi 2}(A) = 75.61$$
$$V^{\pi 2}(B) = 68.05$$
$$V^{\pi 2}(C) = 87.56$$
$$V^{\pi 2}(D) = 100$$

Calculating state values for a policy π_3:

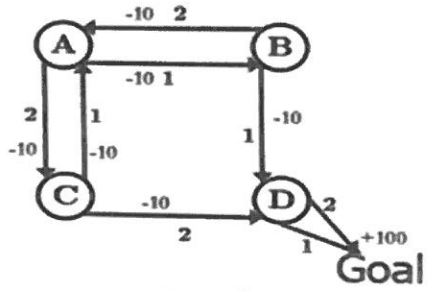
π_3 : mixed

$\pi(1 \mid A) = 0.4 \qquad \pi(2 \mid A) = 0.6$

$\pi(1 \mid B) = 1 \qquad \pi(2 \mid B) = 0$

$\pi(1 \mid C) = 0 \qquad \pi(2 \mid C) = 1$

$\pi(1 \mid D) = 1 \qquad \pi(2 \mid D) = 0$



$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} P(s'|s,a)[R(s,a) + \gamma V^{\pi}(s')]$$

$V^{\pi}(D) = 100$

$$\begin{aligned}
V^{\pi}(B) = \;& 1 \cdot 0.9 \cdot [-10 + V^{\pi}(D)] \\
+\; & 1 \cdot 0.1 \cdot [-10 + V^{\pi}(A)] \\
V^{\pi}(C) = \;& 1 \cdot 0.9 \cdot [-10 + V^{\pi}(D)] \\
+\; & 1 \cdot 0.1 \cdot [-10 + V^{\pi}(A)] \\
V^{\pi}(A) = \;& 0.4 \cdot 0.9 \cdot [-10 + V^{\pi}(B)] \\
+\; & 0.4 \cdot 0.1 \cdot [-10 + V^{\pi}(C)] \\
+\; & 0.6 \cdot 0.9 \cdot [-10 + V^{\pi}(C)] \\
+\; & 0.6 \cdot 0.1 \cdot [-10 + V^{\pi}(B)]
\end{aligned}
\quad \Longrightarrow \quad
\begin{aligned}
V^{\pi 3}(A) &= 77.78 \\
V^{\pi 3}(B) &= 87.78 \\
V^{\pi 3}(C) &= 87.78 \\
V^{\pi 3}(D) &= 100
\end{aligned}$$

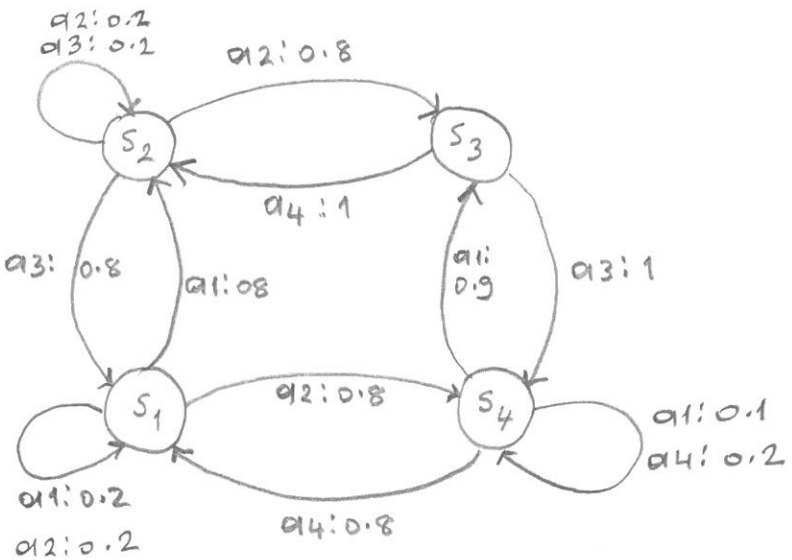
Comparing the 3 policies

	π_1	π_2	π_3
A	75.61	75.61	77.78
B	87.56	68.05	87.78
C	68.05	87.56	87.78
D	100	100	100

$\pi_1 \leq \pi_3$ and $\pi_2 \leq \pi_3 \Rightarrow \pi_3$ is optimal among these 3 policies

π_1 and π_2 are not comparable

Q.6. Consider the following MDP



S	r(s)
s_1	0
s_2	0
s_3	1
s_4	0

$\gamma = 0.5$

t	0	1	2	3	4
$V(s_1)$	0	0	0	0.18	0.206
$V(s_2)$	0	0	0.4	0.44	0.52
$V(s_3)$	0	1	1	1.225	1.2
$V(s_4)$	0	0	0.45	0.473	0.57

$$V^*(s_1) = 0 + \gamma . \max\{ 0.2 * V^*(s_1) + 0.8 * V^*(s_2);\ 0.2 * V^*(s_1) + 0.8 * V^*(s_4) \}$$

$$V^*(s_2) = 0 + \gamma . \max\{ 0.2 * V^*(s_2) + 0.8 * V^*(s_3);\ 0.2 * V^*(s_2) + 0.8 * V^*(s_1) \}$$

$$V^*(s_3) = 1 + \gamma . \max\{ V^*(s_2),\ V^*(s_4) \}$$

$$V^*(s_4) = 0 + \gamma . \max\{ 0.1 * V^*(s_4) + 0.9 * V^*(s_3);\ 0.2 * V^*(s_4) + 0.8 * V^*(s_1) \}$$