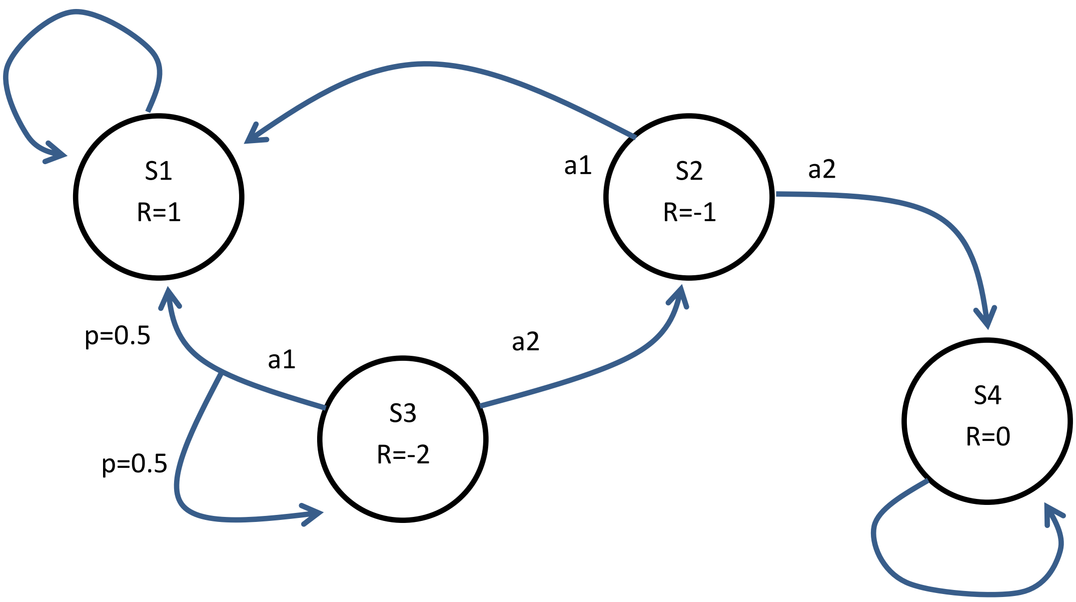
1. Value iteration

Consider the following four-state MDP:



Let the discount factor $\gamma = 0.9$.

a. Let π_1 denote a policy that always chooses action a_1. Write down an equation for the value function V^{π_1}, and solve it.

b. Starting from $V_0 = \{0,1,0,1\}$, run several iterations of the value iteration algorithm. For each iteration, calculate the greedy policy.

c. When changing the discount factor to $\gamma = 0.4$, and running value iteration until convergence, the optimal policy is $\pi^*(s_2) = a_1, \quad \pi^*(s_3) = a_1$. Explain.

d. Find the minimal γ for which $\pi^*(s_3) = a_2$.

Solution

a. Using the Bellman equation for a fixed policy $V^{\pi_1} = r + \gamma P^{\pi_1} V^{\pi_1}$, where

$$r = \begin{pmatrix} 1 \\ -1 \\ -2 \\ 0 \end{pmatrix}, \text{ and } P^{\pi_1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

Therefore we have

$$\begin{pmatrix} 1-\gamma & 0 & 0 & 0 \\ -\gamma & 1 & 0 & 0 \\ -0.5\gamma & 0 & 1-0.5\gamma & 0 \\ 0 & 0 & 0 & 1-\gamma \end{pmatrix} V^{\pi_1} = \begin{pmatrix} 1 \\ -1 \\ -2 \\ 0 \end{pmatrix}.$$

Solving gives:

$$V^{\pi_1}(s_1) = (1-\gamma)^{-1} = 10$$
$$V^{\pi_1}(s_2) = -1 + \gamma(1-\gamma)^{-1} = 8$$
$$V^{\pi_1}(s_3) = \frac{-2 + 0.5\gamma(1-\gamma)^{-1}}{1-0.5\gamma} = 4.54$$
$$V^{\pi_1}(s_4) = 0$$

b. iteration 1:

$$V_1(s_1) = 1 + \gamma V_0(s_1) = 1$$
$$V_1(s_2) = \max\{-1 + \gamma V_0(s_1), -1 + \gamma V_0(s_4)\} = -1 + \gamma = -0.1$$
$$V_1(s_3) = \max\{-2 + \gamma(0.5V_0(s_1) + 0.5V_0(s_3)), -2 + \gamma V_0(s_2)\} = -2 + \gamma = -1.1$$
$$V_1(s_4) = 0 + \gamma V_1(s_4) = \gamma = 0.9$$
$$\pi_1(s_2) = a_2$$
$$\pi_1(s_3) = a_2$$

Iteration 2:

$$V_2(s_1) = 1 + \gamma V_1(s_1) = 1 + \gamma = 1.9$$
$$V_2(s_2) = \max\{-1 + \gamma V_1(s_1), -1 + \gamma V_1(s_4)\} = -1 + \gamma = -0.1$$
$$V_2(s_3) = \max\{-2 + \gamma(0.5V_1(s_1) + 0.5V_1(s_3)), -2 + \gamma V_1(s_2)\} = -2 - 0.05\gamma = -2.045$$
$$V_2(s_4) = 0 + \gamma V_1(s_4) = \gamma^2 = 0.81$$
$$\pi_2(s_2) = a_1$$
$$\pi_2(s_3) = a_2$$

…

Iteration 200:

$V^*(s_1) = 10$
$V^*(s_2) = 8$
$V^*(s_3) = 5.2$
$V^*(s_4) = 0$
$\pi^*(s_2) = a_1$
$\pi^*(s_3) = a_2$

c. When γ decreases, the immediate negative rewards outweigh the potential positive ones in the future.

d. It is clear that the optimal action in s_2 is a_1. Let π_2 denote a policy that chooses a_2 in s_3 and a_1 in s_2 . To find the threshold γ we compare $V^{\pi_1}(s_3)$ with $V^{\pi_2}(s_3)$.

From a previous calculation we have $V^{\pi_1}(s_3) = \dfrac{-2 + 0.5\gamma(1-\gamma)^{-1}}{1 - 0.5\gamma}$, and note that

$V^{\pi_2}(s_3) = -2 + \gamma V^{\pi_1}(s_2) = -2 + \gamma\left(-1 + \gamma(1-\gamma)^{-1}\right)$.

Solving $V^{\pi_1}(s_3) = V^{\pi_2}(s_3)$ gives the threshold $\gamma = 0.5$.

2. Operator notation:

For an MDP with N states and actions $a \in A$, recall the definition of the Bellman operator $T : \mathbb{R}^N \to \mathbb{R}^N$

$$(TJ)(s) = \min_{a \in A}\left\{ r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) J(s') \right\}$$

a. Write down $(T^2 J)(s)$ explicitly, and relate it to a finite-horizon dynamic programming problem.

b. An operator T is said to have a *monotonicity* property if $J \le \bar{J} \Rightarrow TJ \le T\bar{J}$, where the inequality holds element-wise. Show that the Bellman operator is monotone.

c. Show that if T is monotone then T^k is also monotone.

d. Let e denote a vector of ones. Show that $\left(T^k(J + ce)\right)(s) = \left(T^k J\right)(s) + \gamma^k c$.

e. Show that T^k is a γ^k contraction (in the sup-norm).

Solution:

a. We have

$$
\begin{aligned}
\left(T^2 J\right)(s) &= \min_{a\in A}\left\{r(s,a)+\gamma\sum_{s'\in S} p(s'|s,a)TJ(s')\right\} \\
&= \min_{a_1\in A}\left\{r(s,a_1)+\gamma\sum_{s'\in S} p(s'|s,a_1)\min_{a_2\in A}\left\{r(s',a_2)+\gamma\sum_{s''\in S} p(s''|s',a_2)J(s'')\right\}\right\} \\
&= \min_{a_1,a_2\in A}\mathbb{E}\left[r(s,a_1)+\gamma r(s',a_2)+\gamma^2 J(s'')\right]
\end{aligned}
$$

which is exactly the dynamic programming algorithm for a 2-stage discounted problem with initial state s, reward r, and terminal reward $\gamma^2 J$.

b. This may be seen intuitively from (a), but here we calculate it explicitly. Assume $J(s)\le \bar{J}(s)$ for all $s\in S$. We have

$$
\begin{aligned}
(TJ)(s) &= \min_{a\in A}\left\{r(s,a)+\gamma\sum_{s'\in S} p(s'|s,a)J(s')\right\} \\
&= r(s,a^*)+\gamma\sum_{s'\in S} p(s'|s,a^*)J(s') \\
&\le r(s,\bar{a}^*)+\gamma\sum_{s'\in S} p(s'|s,\bar{a}^*)J(s') \\
&\le r(s,\bar{a}^*)+\gamma\sum_{s'\in S} p(s'|s,\bar{a}^*)\bar{J}(s') \\
&= \left(T\bar{J}\right)(s)
\end{aligned}
$$

c. We have $J\le \bar{J} \Rightarrow TJ\le T\bar{J} \Rightarrow T(TJ)\le T\left(T\bar{J}\right) \Rightarrow \ldots \Rightarrow T^k J\le T^k\bar{J}$.

d. We have

$$
\begin{aligned}
\left(T(J+ce)\right)(s) &= \min_{a\in A}\left\{r(s,a)+\gamma\sum_{s'\in S} p(s'|s,a)\left(J(s')+c\right)\right\} \\
&= \min_{a\in A}\left\{r(s,a)+\gamma\sum_{s'\in S} p(s'|s,a)J(s')\right\}+\gamma c \\
&= TJ(s)+\gamma c
\end{aligned}
$$

And therefore

$$T^2(J+ce)=T(TJ+\gamma ce)=T^2 J+\gamma^2 ce \;,$$

And by induction the result follows.

e. For some J and $\bar{J}$ let $c=\max_s\left\{J(s)-\bar{J}(s)\right\}$. We have that for all $s\in S$

$$J(s)-c\le \bar{J}(s)\le J(s)+c$$

Thus, using the monotonicity property we have

$$T^k(J-ce)\le T^k\left(\bar{J}\right)\le T^k(J+ce)$$

And using the result of (d) we have

$$T^k(J) - \gamma^k c \le T^k(\bar{J}) \le T^k(J) + \gamma^k c$$

Therefore $\max_s \{T^k J(s) - T^k \bar{J}(s)\} \le \gamma^k c = \gamma^k \max_s \{J(s) - \bar{J}(s)\}$.