למידה במערכות דינמיות אביב תש"פ

תרגיל בית 3

מגישים:

יאיר נחום 034462796

דר ערבה 205874951

1.

a) It is easy to see that the optimal policy must be to reach state n as fast as we can since the reward is 0 on every other state. Thus, the policy is to go right on every stage. In stage n we have no other option than to go back to 0.

b) For the fixed policy we can apply policy evaluation by solving |S| equations (the same as calculating inverse matrix $(I - \gamma P^{\pi})^{-1} r^{\pi}$). Also, we notice that in this case the MDP is deterministic once we know the action by the policy:

$$\forall s \in S, \qquad V^{\pi^*}(s_t) = r\big(s_t, \pi(s_t)\big) + \gamma \sum_{s_{t+1} \in S} P\big(s_{t+1} \big| s_t, \pi(s_t)\big) V^{\pi^*}(s_{t+1})$$

$$= r\big(s_t, \pi(s_t)\big) + \gamma V^{\pi^*}(s_{t+1})$$

$$\Rightarrow V^{\pi^*}(s_t) = \begin{cases} s_t < n, \gamma V^{\pi^*}(s_{t+1}) \\ s_t = n, 1 + \gamma V^{\pi^*}(s_{t+1}) \end{cases}$$

We can solve these equations by noticing the connection between the first stage and the last stage:

$V^{\pi^*}(1) = \gamma V^{\pi^*}(2) = .. = \gamma^{n-1} V^{\pi^*}(n) \; and \; V^{\pi^*}(n) = 1 + \gamma V^{\pi^*}(1)$

$$\Rightarrow V^{\pi^*}(1) = \frac{\gamma^{n-1}}{1-\gamma^n}$$

$$\Rightarrow V^{\pi^*}(s) = \frac{\gamma^{n-s}}{1-\gamma^n}$$

c) According to total probability on the actions, we can calculate the transitions matrix as follows:

$$P^{\pi}(s'|s) = \sum_{\alpha \in A} P^{\pi}(s', a|s) = \sum_{\alpha \in A} P^{\pi}(s'|s, a)\pi(a|s)$$

for each state other than n, the policy is stochastic with 0.5 chance for every action.

In order to calculate the stationary distribution of the states, one needs to solve the vector equation:

$$d^T P^{\pi} = d^T$$

Or per state probability:

$$d_j = \sum_{i \in S} d_i {P^{\pi}}_{ij}$$

According to the formula above for the transition probabilities ${P^{\pi}}_{ij}$ we have n equations and also the demand:

$$\sum_{i \in S} d_i = 1$$

We note that the dynamics are deterministic once the action is selected, thus, the transition probability is the policy probability.

$$d_1 = 0.5(d_2 + d_3 + .. + d_{n-1}) + d_n$$

$$d_2 = d_1$$

$$d_k = 0.5 d_{k-1}, k \in \{3,4, .. n\}$$

Again, by applying the recursion of the equations and using the normalization to a probability vector we get:

$$1 = \sum_{i \in S} d_i = d_1 + d_1 + 0.5d_1 + 0.5^2 d_1 + .. + 0.5^{n-2} d_1 = d_1(1 + \frac{1(0.5^{n-1} - 1)}{0.5 - 1})$$

$$= d_1(1 + 2(1 - 0.5^{n-1})) = d_1(3 - 0.5^{n-2})$$

$$\Rightarrow d_1 = \frac{1}{(3 - 0.5^{n-2})}$$

$$d_2 = \frac{1}{(3 - 0.5^{n-2})}$$

$$d_k = \frac{0.5^{k-2}}{(3 - 0.5^{n-2})}, k \in \{3,4,..n\}$$

d) FPVI specifies $V_{n+1}^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P^{\pi}(s' | s, \pi(s)) V_n^{\pi}(s')$

In matrix description $V_{n+1}^{\pi} = r^{\pi} + \gamma P^{\pi} V_n^{\pi} = \sum_{k=0}^{n} (\gamma P^{\pi})^k r^{\pi} + (\gamma P^{\pi})^{n+1} V_0^{\pi}$

In our case, since we started with $V_0^{\pi} = 0$ then we get: $V_{n+1}^{\pi} = \sum_{k=0}^{n} (\gamma P^{\pi})^k r^{\pi}$

The transition matrix can be calculated using:

$$P^{\pi}(s'|s) = \sum_{\alpha \in A} P^{\pi}(s', a|s) = \sum_{\alpha \in A} P^{\pi}(s'|s, a) \pi(a|s)$$

On state 1 we have only one action possible, so we can only move to state 2.

On other states other than n, we select an action randomly with probability 0.5 for each action and then continue the dynamics deterministically.

Thus:

$$P^{\pi}(s' = i + 1 | s = i) = \sum_{\alpha \in A} P^{\pi}(s'|s, a) \pi(a|s) = 0.5 \sum_{\alpha \in A} \mathbb{I}(\alpha \, from \, i \, to \, i + 1)$$

$$= 0.5$$

The same for going from i state to 1.

On the n state we have no option than going back to 1 state.

When:

$$r^{\pi} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, P^{\pi} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0.5 & 0 & 0.5 & 0 & \cdots & 0 \\ 0.5 & 0 & 0 & 0.5 & \cdots & 0 \\ 0.5 & \vdots & \vdots & \vdots & \ddots & 0.5 \\ 1 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

Therefore, If $V_0^{\pi} = 0$, then:

$$V_1^{\pi} = r^{\pi} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

$$V_2^\pi = r^\pi + \gamma P^\pi V_1^\pi = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} + \gamma \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0.5 & 0 & 0.5 & 0 & \cdots & 0 \\ 0.5 & 0 & 0 & 0.5 & \cdots & 0 \\ 0.5 & \vdots & \vdots & \vdots & \ddots & 0.5 \\ 1 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0.5\gamma \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0.5\gamma \\ 1 \end{pmatrix}$$

$$V_3^\pi = r^\pi + \gamma P^\pi V_2^\pi = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} + \gamma \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0.5 & 0 & 0.5 & 0 & \cdots & 0 \\ 0.5 & 0 & 0 & 0.5 & \cdots & 0 \\ 0.5 & \vdots & \vdots & \vdots & \ddots & 0.5 \\ 1 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0.5\gamma \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0.25\gamma^2 \\ 0.5\gamma \\ 1 \end{pmatrix}$$

From FPVI convergence Proposition 5.2, we get that $V_\infty^\pi = \lim_{n\to\infty} V_n^\pi = V^\pi$

BTW, V^π can be calculated directly from $(I - \gamma P^\pi)^{-1} r^\pi$as we proved in Lemma 5.2 in the lecture.

e) We can apply recursive expectation calculation. First, let's observe what happens in simple cases like n=2,3,4:

Let's denote in T_k the time (expectation of it) to reach state n from state k.

We want to find T_1

In case of n=2:

$T_2 = 0$

$T_1 = 1 + T_2$

$\Rightarrow T_1 = 1$

As there is only one possible transition to state 2.

In case of n=3:

$T_3 = 0$

$T_2 = 1 + 0.5T_1 + 0.5T_3$

$T_1 = 1 + T_2$

$\Rightarrow T_1 = 1 + 1 + 0.5T_1$

$\Rightarrow T_1 = 4$

Another way to calculate it is by explicit total expectation:

$$T_1 = 1 + T_2 = 1 + E[E[T|start\_state = 2, next\_state]|start\_state = 2]$$
$$= 1 + p \cdot 1 + (1-p)(1 + T_1) = 2 + pT_1$$

When in our case, the p denotes the probability to move from state i to state i+1.

In case of n=4:

$T_4 = 0$
$T_3 = 1 + 0.5T_1 + 0.5T_4$
$T_2 = 1 + 0.5T_1 + 0.5T_3$
$T_1 = 1 + T_2$
$\Rightarrow T_1 = 1 + 1 + 0.5T_1 + 0.5(1 + 0.5T_1) = 2.5 + 0.75T_1$
$\Rightarrow T_1 = 10$

In the general case:

$$T_1 = 1 + (1 + 0.5+..+0.5^{n-2}) + 0.5T_1(1 + 0.5+..+0.5^{n-3}) =$$
$$1 + \frac{1(0.5^{n-2} - 1)}{0.5 - 1} + \frac{0.5T_1(0.5^{n-2} - 1)}{0.5 - 1} = 3 - 0.5^{n-3} + T_1(1 - 0.5^{n-2})$$
$$\Rightarrow T_1 = \frac{3-0.5^{n-3}}{0.5^{n-2}} = 3 \cdot 2^{n-2} - 2$$

for $n \geq 2$

f) In order to go through all state actions pairs, we will define a policy such that at the first time we encounter some state i we get back to 1 from it. Otherwise, we continue to next state i+1.
We note that going over state 1 +action right, we do it already for reaching state 2.
Thus, the number of steps needed to go over all state action pairs can be calculated as follows:

$$N = \underbrace{(1 + 1)}_{(state,action):\ (1,right)\ \&\ (2,left)} + \underbrace{(2 + 1)}_{(state,action):\ (2,right)\ \&\ (3,left)}$$
$$+ \underbrace{(3 + 1)}_{(state,action):\ (3,right)\ \&\ (4,left)} + \cdots + \underbrace{(n - 1 + 1)}_{(state,action):\ (n-1,right)\ \&\ (n,left)}$$
$$= n - 1 + (1 + 2+..+n - 1) = n - 1 + \frac{n(n - 1)}{2} = \frac{(n + 2)(n - 1)}{2}$$

and we can check it's correct for all n using induction (base: $n = 2 \rightarrow\ N = 2$).

g) Since we want to maximize the rewards which are exponentially reduced as times goes by due to discount factor, the optimal policy is to select the path that gives us as much as possible consecutive rewards in the near future.
Meaning, we will go from stage 1 to n and back to 1 in order to collect these consecutive rewards (that are reduced constantly by discount factor) and then collect rewards as we did on the previous section (1 → 2 → 1 → 2 → 3 → 1 ...).
If we look at the summary of steps, we get the same amount as in previous section (just in a different order of execution in order to collect the maximum reward)

$$N = \underbrace{(n - 1 + 1)}_{(state,action):\ (1,right)\ \&\ (2,right)\&...\&(n-1,right)\&(n,left)} + \underbrace{(1 + 1)}_{(2,left)} + \underbrace{(2 + 1)}_{(3,left)} +..+ \underbrace{(n - 2 + 1)}_{(n-1,left)}$$
$$= n - 1 + (1 + 2+..+n - 1) = n - 1 + \frac{n(n - 1)}{2} = \frac{(n + 2)(n - 1)}{2}$$

h) As we saw in section d. in case n=3 the rewards vector and transition matrix (under the stochastic policy is as follows):

$$r^{\pi} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \; P^{\pi} = \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 1 & 0 & 0 \end{pmatrix}$$

Applying the closed solution of the 3 equations:
$V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi}$ we get:

$$V^{\pi} = \begin{pmatrix} 1 & -\gamma & 0 \\ -0.5\gamma & 1 & -0.5\gamma \\ -\gamma & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & -0.5 & 0 \\ -0.25 & 1 & -0.25 \\ -0.5 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} =$$

$$\begin{pmatrix} 1.23076\ldots & 0.61538\ldots & 0.15384\ldots \\ 0.46153\ldots & 1.23076\ldots & 0.30769\ldots \\ 0.61538\ldots & 0.30769\ldots & 1.07692\ldots \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.15384 \\ 0.30769 \\ 1.07692 \end{pmatrix}$$

2.

a) The state space:
$S = \{All\ subgroups\ of\ the\ group \{1,2 \ldots N\}\}$
Therefore, we have $|S| = 2^N$
The state's defines the remaining jobs that were not served yet.
The start state is $s_0 = \{1,2, \ldots, N\}$

The action space:
$A = \{1,2 \ldots N\}$
The action index defines the job that the server tries to serve at that time index.
It cannot be an index that doesn't belong to the state at that time index.

$$a_t \in A_t = s_t \in S$$

Transition probabilities:
As stated, We note that we cannot have an action that is not part of the current state.
If we define some state as $J = \{some\ subset\ of\ \{1,2,.,,N\}\}$. We can move from $J \cup \{k\}$ to J by selecting $a_t = k$. If the job was not served (with probability μ_k) we stay in the same state J.

$$P(s' = J | s = J \cup \{k\}, a = k) = \mu_k$$
$$P(s' = J \cup \{k\} | s = J \cup \{k\}, a = k) = 1 - \mu_k$$

For all $a_t = k \in A_t = s_t$. And all other probabilities are 0 (we can't move to other states if we took action k).

The total cost can be computed as follows:

$$\sum_{t=0}^{\mathrm{T}} \boldsymbol{C_t(s,a)} = \sum_{t=0}^{\mathrm{T}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \mathbb{I}(k \in s_t) c_k$$

Meaning, only if we have in current time index an unhandled job we pay with its cost.
We can see that the state cost doesn't depend on action we take next, but only on the current state $\boldsymbol{C_t(s,a)} = \boldsymbol{C_t(s)}$.
At some point in time T, we reach the terminal state in which there are no jobs to

handle. This is ensured as all probabilities to handle some job are greater than 0. Thus we will reach the terminal state at infinity with probability 1 (absorbing state). The cost in that terminal state is 0.
We can also think of it with a termination state that is the empty set (no more jobs to process).

The bellman equation is therefore:

$$V(s_t) = \min_{a_t \in A_t} \left\{ C(s_t) + \sum_{s_{t+1} \in S} P(s_{t+1} | s_t, a_t) V(s_{t+1}) \right\} , \quad s_t \in S$$

We can write it as:

$$V(s_t) = C(s_t) + \min_{a_t \in A_t} \{ \mu_{a_t} V(s_{t+1}) + (1 - \mu_{a_t}) V(s_t) \} , \quad s_t \in S$$

When s_{t+1} is $s_t / \{a_t\}$

b) The optimal policy according to Bellman equation is:

$$\pi^*(s_t) = \underset{a_t \in A_t}{argmin} \{ \mu_{a_t} V(s_{t+1}) + (1 - \mu_{a_t}) V(s_t) \} , \quad s_t \in S$$

We need to check the suggested optimal policy:

$$i^* = \underset{i \in A_t}{argmax} \{ c_i \mu_i \} , \quad s_t \in S$$

by policy evaluation.
If we develop the policy evaluation equation according to this selected action. we get:

$$V(s_t) = C(s_t) + \mu_{i^*} V(s_{t+1}) + (1 - \mu_{i^*}) V(s_t), \qquad s_t \in S$$

$$\mu_{i^*} V(s_t) = C(s_t) + \mu_{i^*} V(s_{t+1}) , \quad s_t \in S$$

We also remember that s_{t+1} is $s_t / \{i^*\}$. Thus:

$$\mu_{i^*} V(s_t) = C(s_t) + \mu_{i^*} V(s_t / \{i^*\})$$

$$V(s_t) = \frac{C(s_t)}{\mu_{i^*}} + V(s_t / \{i^*\})$$

And also we can see that:

$$(***) \; V\left(s_t / \{i^{*\}}\right) = V(s_t) - \frac{C(s_t)}{\mu_{i^*}}$$

If we continue recursively with i^{**} as optimal for $V(s_t / \{i^*\})$ we get:

$$V(s_t) = \frac{C(s_t)}{\mu_{i^*}} + \frac{C(s_t / i^*)}{\mu_{i^{**}}} + V(s_t / \{i^*, i^{**}\})$$

and if we continue this way until the termination state we get:

$$V(s_t) = \frac{C(s_t)}{\mu_{i^*}} + \frac{C(s_t / i^*)}{\mu_{i^{**}}} + \frac{C(s_t / \{i^*, i^{**}\})}{\mu_{i^{***}}} + .. + 0$$

We can sort the indexes chosen by the given policy and denote $i^* = z_1, i^{**} = z_2$..
Which match the selection according to:

$$c_{z_1} \mu_{z_1} \geq c_{z_2} \mu_{z_2} \geq c_{z_3} \mu_{z_3} ..$$

If we assume k jobs not handled in start stage, we can reorder the summation order of the cost as follows:

$$V(s_t) = \frac{c_{z_1} + c_{z_2} + .. + c_{z_k}}{\mu_{z_1}} + \frac{c_{z_2} + .. + c_{z_k}}{\mu_{z_2}} + \frac{c_{z_3} + .. + c_{z_k}}{\mu_{z_3}} + .. + \frac{c_{z_k}}{\mu_{z_k}}$$

$$= \frac{c_{z_1}}{\mu_{z_1}} + (\frac{c_{z_2}}{\mu_{z_1}} + \frac{c_{z_2}}{\mu_{z_2}}) + (\frac{c_{z_3}}{\mu_{z_1}} + \frac{c_{z_3}}{\mu_{z_2}} + \frac{c_{z_3}}{\mu_{z_3}}) + ..$$

And from (***) we get:

$$V(s_t/i^*) = \frac{c_{z_1}}{\mu_{z_1}} + (\frac{c_{z_2}}{\mu_{z_1}} + \frac{c_{z_2}}{\mu_{z_2}}) + (\frac{c_{z_3}}{\mu_{z_1}} + \frac{c_{z_3}}{\mu_{z_2}} + \frac{c_{z_3}}{\mu_{z_3}}) + .. - \frac{c_{z_1} + c_{z_2} + .. + c_{z_k}}{\mu_{z_1}}$$

Next, we put things together in the bellman equation we started from and apply the first action selection according to policy(selected action is z_1):

$$V(s_t) = C(s_t) + \underset{a_t \in A_t}{min} \{\mu_{a_t} V(s_{t+1}) + (1 - \mu_{a_t}) V(s_t)\}, \ s_t \in S$$

$$0 = c_{z_1} + c_{z_2} + .. + c_k + \mu_{z_1} V(s_{t+1}) - \mu_{z_1} V(s_t), \ s_t \in S$$

If we put our solution to the V according to the given policy we get a valid equation:

$$c_{z_1} + c_{z_2} + .. + c_k + \mu_{z_1} \underbrace{\left[\frac{c_{z_1}}{\mu_{z_1}} + \left(\frac{c_{z_2}}{\mu_{z_1}} + \frac{c_{z_2}}{\mu_{z_2}} \right) + \left(\frac{c_{z_3}}{\mu_{z_1}} + \frac{c_{z_3}}{\mu_{z_2}} + \frac{c_{z_3}}{\mu_{z_3}} \right) + .. - \frac{c_{z_1} + c_{z_2} + .. + c_{z_k}}{\mu_{z_1}} \right]}_{V(s_{t+1})}$$

$$- \mu_{z_1} \underbrace{\left[\frac{c_{z_1}}{\mu_{z_1}} + \left(\frac{c_{z_2}}{\mu_{z_1}} + \frac{c_{z_2}}{\mu_{z_2}} \right) + \left(\frac{c_{z_3}}{\mu_{z_1}} + \frac{c_{z_3}}{\mu_{z_2}} + \frac{c_{z_3}}{\mu_{z_3}} \right) + .. \right]}_{V(s_t)}$$

$$= c_{z_1} + c_{z_2} + .. + c_k + \mu_{z_1} \left[- \frac{c_{z_1} + c_{z_2} + .. + c_k}{\mu_{z_1}} \right] = 0$$

We conclude that the given policy solves the bellman equation and thus optimal.

3. We need to show that for a fixed policy the DP operator is not a contraction operator with respect to the Euclidean norm.

We define the MDP as **one action per state**, and from that action selected by the only existing policy, we define the transitions as depicted in the hint.

We follow the hint and look at simple vectors of 2 components and configurable transition probabilities.

We now build the equations related to the DP operator with L2 norm on some 2 vectors.

$$V' = \begin{pmatrix} v'_1 \\ v'_2 \end{pmatrix}, V = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

The L2 norm of the difference between these 2 vectors squared is:

$$\|V' - V\|_2{}^2 = (v'_1 - v_1)^2 + (v'_2 - v_2)^2$$

We define the probabilities as in the example and fixed rewards per state (we will see these are removed when subtracting between DP operator results).

The operator over vector V on some state s_t is as follows:

$$(T^\pi V)(s_t) = r\big(s_t, \pi(s_t)\big) + \gamma \sum_{s_{t+1} \in S} P\big(s_{t+1} \big| s_t, \pi(s_t)\big) V^\pi(s_{t+1})$$

Therefore, for state s_1:

$$(T^\pi V)(s_1) = r_{s_1} + \gamma((1-p_1)V(s_1) + p_1 V(s_2)) = r_{s_1} + \gamma((1-p_1)v_1 + p_1 v_2)$$

And in the same way for state s_2:

$$(T^\pi V)(s_2) = r_{s_2} + \gamma((1-p_2)v_2 + p_2 v_1)$$

In vector annotation:

$$T^\pi V = \begin{pmatrix} r_{s_1} + \gamma((1-p_1)v_1 + p_1 v_2) \\ r_{s_2} + \gamma((1-p_2)v_2 + p_2 v_1) \end{pmatrix}$$

We apply the same on $V'$and get:

$$T^\pi V' = \begin{pmatrix} r_{s_1} + \gamma((1-p_1)v'_1 + p_1 v'_2) \\ r_{s_2} + \gamma((1-p_2)v'_2 + p_2 v'_1) \end{pmatrix}$$

The difference between the vectors is:

$$T^\pi V' - T^\pi V = \gamma \begin{pmatrix} (1-p_1)(v_1 - v'_1) + p_1(v_2 - v'_2) \\ (1-p_2)(v_2 - v'_2) + p_2(v_1 - v'_1) \end{pmatrix}$$

Therefore, applying L2 norm squared:

$$\|T^\pi V' - T^\pi V\|_2{}^2$$
$$= \gamma^2[((1-p_1)(v_1 - v'_1) + p_1(v_2 - v'_2))^2$$
$$+ ((1-p_2)(v_2 - v'_2) + p_2(v_1 - v'_1))^2]$$

In order to show the claim about L2, we just need to find some V, V', γ, p_1, p_2 such that:

$$\|T^\pi V' - T^\pi V\|_2{}^2 > \|V' - V\|_2{}^2$$

$\gamma = 0.95, p_1 = 0, p_2 = 1, V' = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, V = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$$\|V' - V\|_2{}^2 = 1$$

$$\|T^\pi V' - T^\pi V\|_2{}^2 = \gamma^2\left[\left((1)(1) + 0(0)\right)^2 + \left((0)(0) + 1(1)\right)^2\right] = \gamma^2 \cdot 2 = 1.805$$

$$\Rightarrow \|T^\pi V' - T^\pi V\|_2{}^2 = \gamma^2 \cdot 2 = 1.805 \ > 1 = \|V' - V\|_2{}^2$$

■

4.

1. The Bellman equations for the value function v^π, v^*:

$$v^\pi(s) = c\big(s, \pi(s)\big) + \underbrace{\gamma}_{=1} \sum_{s'} P\big(s'\big|s, \pi(s)\big) v^\pi(s') \quad \forall s \in S/\{0\}$$

For terminal state -

$$v^\pi(s = 0) = c\big(0, \pi(0)\big) = 0$$

$$v^*(s) = \min_a \left\{ c(s,a) + \underbrace{\gamma}_{=1} \sum_{s'} P(s'|s,a) v^*(s') \right\} \quad \forall s \in S/\{0\}$$

For terminal state -

$$v^*(s = 0) = \min_a\{c(0,a)\} = 0$$

2. The Bellman operators T^*, T^π :

$$(T^\pi(v))_{(s)} = c\big(s,\pi(s)\big) + \sum_{s'} P(s'|s,\pi(s))v(s')$$

$$(T^*(v))_{(s)} = \min_a \left\{ c(s,a) + \sum_{s'} P(s'|s,a)v(s') \right\}$$

3. Without the assumption of all stationary policies being proper, for some improper policy π we get - $J^\pi(s) = \mathbb{E}^\pi[\sum_{t=0}^{\infty} c(s_t,a_t)\,|s_0 = s] \to \infty$.

Since there is zero probability of reaching the terminal state, therefore the sum is infinite.

Also, there is no discount factor to obtain convergence of the infinite sum.

4. Proof –

Let's consider a new SSP with same transitions and costs all equal -1, except for the terminal state, 0. Let $\hat{J}(s)$ be the optimal value from state s in the new SSP and- $\xi(s) = -\hat{J}(s)$.

Since all costs equal -1 and cost of terminal state is 0, all values $J(s)$ from $s \in S/\{0\}$ will uphold-

$$J(s) \le -1$$

The optimal value $\hat{J}(s)$ also upholds the inequality above, therefore –

$$\boxed{\xi(s) = -\hat{J}(s) \ge 1}$$

a). The Bellman equations for $\hat{J}(s)$ –

$$\underbrace{\hat{J}(s)}_{-\xi(s)} = \min_a \left\{ \underbrace{c(s,a)}_{=-1} + \sum_{s'} P(s'|s,a)\underbrace{\hat{J}(s')}_{-\xi(s')} \right\} \qquad \forall s \in S/\{0\}$$

$$\to\; -\xi(s) = -1 + \min_a \left\{ -\sum_{s'} P(s'|s,a)\xi(s') \right\} \;\to\; -\xi(s) + 1 = \min_a \left\{ -\sum_{s'} P(s'|s,a)\xi(s') \right\}$$

$$\underbrace{\to}_{for\ some\ policy\ \pi} \quad -\ \xi(s) + 1 \le -\sum_{s'} P(s'|s,\pi(s))\xi(s')$$

$$\to \boxed{\sum_{s'} P(s'|s,\pi(s))\xi(s') \le\ \xi(s) - 1}$$

b). From result: $\xi(s) \geq 1$, we get: $\xi(s) > \xi(s) - 1 \geq 0$. Therefore-

$$\beta = \max_{s'} \left\{ \frac{\xi(s') - 1}{\xi(s')} \right\} < 1$$

$$\xi(s) - 1 = \xi(s)\left(1 - \frac{1}{\xi(s)}\right) = \xi(s)\left(\frac{\xi(s) - 1}{\xi(s)}\right)$$

$$\leq \max_{s} \left\{ \frac{\xi(s) - 1}{\xi(s)} \right\} \xi(s) = \underbrace{\max_{s'} \left\{ \frac{\xi(s') - 1}{\xi(s')} \right\}}_{=\beta} \xi(s)$$

$$\rightarrow \boxed{\xi(s) - 1 \leq \beta\, \xi(s)}$$

Returning to the original SSP problem-

Let J_1 and J_2 be two elements in $\mathbb{R}^S$, which uphold - $\|J_1 - J_2\|_\xi = d$. Let π_1^* be the policy such that - $T^{\pi_1^*}(J_1) = T^*(J_1)$

Then-

$$T^*(J_2) - T^*(J_1) = T^*(J_2) - T^{\pi_1^*}(J_1) \leq T^{\pi_1^*}(J_2) - T^{\pi_1^*}(J_1)$$

The Bellman equation for $T^{\pi_1^*}$ -

$$(T^{\pi_1^*}(J))_{(s)} = c\big(s, \pi_1^*(s)\big) + \sum_{s'} P(s'|s, \pi_1^*(s))J(s')$$

Therefore, for each state s –

$$\big(T^*(J_2)\big)_{(s)} - \big(T^*(J_1)\big)_{(s)} \leq \left(T^{\pi_1^*}(J_2)\right)_{(s)} - \left(T^{\pi_1^*}(J_1)\right)_{(s)}$$

$$= \underbrace{c\big(s, \pi_1^*(s)\big) + \sum_{s'} P\big(s'\big|s, \pi_1^*(s)\big)J_2(s')}_{(T^{\pi_1^*}(J_2))_{(s)}} \underbrace{- c\big(s, \pi_1^*(s)\big) - \sum_{s'} P\big(s'\big|s, \pi_1^*(s)\big)J_1(s')}_{(T^{\pi_1^*}(J_1))_{(s)}}$$

$$= \sum_{s'} P\big(s'\big|s, \pi_1^*(s)\big)\big(J_2(s') - J_1(s')\big)$$

We get-

$$\big|(T^*(J_1))_{(s)} - (T^*(J_2))_{(s)}\big| \leq \left| \sum_{s'} P\big(s'\big|s, \pi_1^*(s)\big)\big(J_2(s') - J_1(s')\big) \right|$$

$$\leq \sum_{s'} P\big(s'\big|s, \pi_1^*(s)\big)|J_2(s') - J_1(s')|$$

$$\leq \sum_{s'} P\big(s'\big|s, \pi_1^*(s)\big) \max_{s' \in S}|J_2(s') - J_1(s')|$$

$$\leq \sum_{s'} P\big(s'\big|s, \pi_1^*(s)\big) \underbrace{\frac{\max_{s' \in S}|J_2(s') - J_1(s')|}{\xi(s')}}_{=\|J_1 - J_2\|_\xi = d} \xi(s') = d \sum_{s'} P\big(s'\big|s, \pi_1^*(s)\big)\xi(s')$$

$$\rightarrow \left|(T^*(J_1))_{(s)} - (T^*(J_2))_{(s)}\right| \leq d \sum_{s'} P(s'|s,\pi_1^*(s))\xi(s') \underbrace{\leq}_{(a)} d(\xi(s)-1) \underbrace{\leq}_{(b)} d\beta\xi(s)$$

$$\rightarrow \frac{\left|(T^*(J_1))_{(s)} - (T^*(J_2))_{(s)}\right|}{\xi(s)} \leq d\beta \qquad \forall s \in S$$

$$\rightarrow \max_{s\in S} \frac{\left|(T^*(J_1))_{(s)} - (T^*(J_2))_{(s)}\right|}{\xi(s)} = \left\|(T^*(J_1))_{(s)} - (T^*(J_2))_{(s)}\right\|_{\xi} \leq d\beta$$

$$\rightarrow \boxed{\left\|(T^*(J_1))_{(s)} - (T^*(J_2))_{(s)}\right\|_{\xi} \leq \beta\|J_1 - J_2\|_{\xi}}$$

Since $\beta \in [0,1)$, we have that T^* is contracting with respect to the norm $||\cdot||_{\xi}$.

For a proper policy π, we can view T^{π} as the optimal Bellman operator T^* in a new problem where the action space for each state s is $A(s) \equiv \pi(s)$. For this new problem we get the desired result-

$$\rightarrow \boxed{\left\|(T^{\pi}(J_1))_{(s)} - (T^{\pi}(J_2))_{(s)}\right\|_{\xi} \leq \beta\|J_1 - J_2\|_{\xi}}$$

T^{π} is contracting with respect to the norm $||\cdot||_{\xi}$.

■

5. A counter example for the given claim-

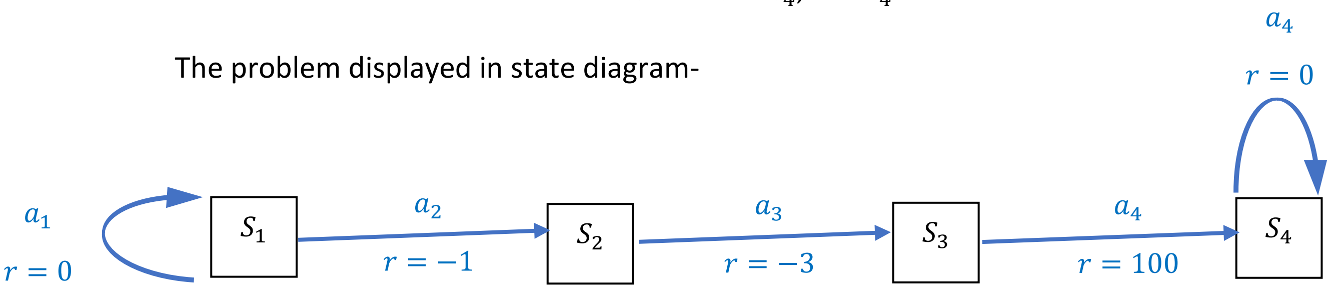
Let us define a problem consisting the state space- $S = \{s_1, s_2, s_3, s_4\}$
and action space- $A(s_1) = \{a_1, a_2\}$, $A(s_2) = \{a_3\}$, $A(s_3) = \{a_4\}$,$A(s_4) = \{a_4\}$
We define γ as 1.
Let the transition probabilities be defined –

$$P(s_1|s_1,a_1) = 1$$
$$P(s_2|s_1,a_2) = 1$$
$$P(s_3|s_2,a_3) = 1$$
$$P(s_4|s_3,a_4) = 1$$
$$P(s_4|s_4,a_4) = 1$$

And all other transitions are with zero probability.
Let the reward function be-

$$r(s,a) = \begin{cases} 0 & s = s_1, a = a_1 \\ -1 & s = s_1, a = a_2 \\ -3 & s = s_2, a = a_3 \\ 100 & s = s_3, a = a_4 \\ 0 & s = s_4, a = a_4 \end{cases}$$

The problem displayed in state diagram-



Let us apply the Value Iteration algorithm with a specific initial value vector-

a. Set arbitrary initial value- $V_0 = \begin{pmatrix} V_0(s_1) \\ V_0(s_2) \\ V_0(s_3) \\ V_0(s_4) \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 0 \\ 0 \end{pmatrix}$

   We get the initial greedy policy (using argmax):

   $\bar{\pi}_0 = \begin{pmatrix} \bar{\pi}_0(s_1) \\ \bar{\pi}_0(s_2) \\ \bar{\pi}_0(s_3) \\ \bar{\pi}_0(s_4) \end{pmatrix} = \begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_4 \end{pmatrix}$ (which is BTW the optimal policy for this MDP.

b. For $n = 1,2 \ldots$

$$V_{n+1}(s) = \max_{\mathrm{a \in A(s)}} \left\{ r(s,a) + \sum_{s'} P(s'|s,a) V_n(s') \right\}$$

$$n = 1:\ V_1 = \begin{pmatrix} \max\left\{ \underbrace{0+0}_{for\,a_1}, \underbrace{-1+2}_{for\,a_2} \right\} \\ -3+0 \\ 100+0 \\ 0+0 \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ 100 \\ 0 \end{pmatrix}$$

and the greedy policy is – $\bar{\pi}_1 = \begin{pmatrix} \bar{\pi}_1(s_1) \\ \bar{\pi}_1(s_2) \\ \bar{\pi}_1(s_3) \\ \bar{\pi}_1(s_4) \end{pmatrix} = \begin{pmatrix} a_1 \\ a_3 \\ a_4 \\ a_4 \end{pmatrix}$

- $n = 2:\ V_2 = \begin{pmatrix} \max\left\{ \underbrace{0+1}_{for\,a_1}, \underbrace{-1+-3}_{for\,a_2} \right\} \\ -3+100 \\ 100+0 \\ 0+0 \end{pmatrix} = \begin{pmatrix} 1 \\ 97 \\ 100 \\ 0 \end{pmatrix}$

and the greedy policy is - $\bar{\pi}_2 = \begin{pmatrix} \bar{\pi}_2(s_1) \\ \bar{\pi}_2(s_2) \\ \bar{\pi}_2(s_3) \\ \bar{\pi}_2(s_4) \end{pmatrix} = \begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_4 \end{pmatrix}$

- $n = 3:\ V_3 = \begin{pmatrix} \max\left\{ \underbrace{0+1}_{for\,a_1}, \underbrace{-1+97}_{for\,a_2} \right\} \\ -3+100 \\ 100+0 \\ 0+0 \end{pmatrix} = \begin{pmatrix} 96 \\ 97 \\ 100 \\ 0 \end{pmatrix}$

and the greedy policy is - $\bar{\pi}_3 = \begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_4 \end{pmatrix}$

On the next value iteration the values doesn’t change and we get the optimal policy $\pi^* = \begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_4 \end{pmatrix}$.

We can see that along the VI algorithm we started with the optimal policy $\bar{\pi}_0 = \begin{pmatrix} a_2 \\ a_3 \\ a_4 \\ a_4 \end{pmatrix}$ and then changed it to some bad policy $\bar{\pi}_1 = \begin{pmatrix} \bar{\pi}_1(s_1) \\ \bar{\pi}_1(s_2) \\ \bar{\pi}_1(s_3) \\ \bar{\pi}_1(s_4) \end{pmatrix} = \begin{pmatrix} a_1 \\ a_3 \\ a_4 \\ a_4 \end{pmatrix}$ (due to temporal close immediate rewards) and then back to a better policy.

If we take $\bar{\pi}_0$ vs $\bar{\pi}_1$and compute (by FPVI) their values (after convergence) we shall see that

$$\rightarrow\ V^{\bar{\pi}_0}(s_1) > V^{\bar{\pi}_1}(s_1)$$

And that the policies got worse.

The $V^{\bar{\pi}_0}(s_1)$ is actually the optimal policy as we saw in the VI we've run.

Thus, $V^{\bar{\pi}_0}(s_1)\ =\ 96$

If we stick with $\bar{\pi}_1$,the $V^{\bar{\pi}_1}(s_1)$ value would be 0 (according to the FPVI below):

$$V^{\bar{\pi}_1}{}_{n+1}(s) = r\big(s, \bar{\pi}_1(s)\big) + \sum_{s'} P\big(s'\big|s, \bar{\pi}_1(s)\big)V_n(s')$$

$$V_0 = \begin{pmatrix} V_0(s_1) \\ V_0(s_2) \\ V_0(s_3) \\ V_0(s_4) \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 0 \\ 0 \end{pmatrix}, V_1 = \begin{pmatrix} 0+0 \\ -3+0 \\ 100+0 \\ 0+0 \end{pmatrix} = \begin{pmatrix} 0 \\ -3 \\ 100 \\ 0 \end{pmatrix},\ V_2 = \begin{pmatrix} 0+0 \\ -3+100 \\ 100+0 \\ 0+0 \end{pmatrix} = \begin{pmatrix} 0 \\ 97 \\ 100 \\ 0 \end{pmatrix}$$

When $V_2 = V^{\bar{\pi}_1}$ is the last iteration as the value function doesn't change any more.