1 Question 1

1.1

- for i=T,T-1,...,0:
  - $\underline{J}^{\pi,t}(s) = \begin{cases} \min_{(a_0,s_0:P(a_0,s_0|s_0=s))>0}[r(a_0,s_0)] & t=0 \\ \min_{(a_0,s_0,s_1):P(a_0,s_0,s_1|s_0=s)>0}[r(a_0,s)+\underline{J}^{\pi,t-1}(s_1)] & t\neq 0 \end{cases}$

1.2

1. Set $\overline{J}^{*,0}(s) = \max\limits_{(a_0:P(a_0,s_0|s_0=s))>0}[r(a_0,s_0)] \quad \forall s$
2. for t=1:T:
   (a) $\overline{J}^{*,t}(s) = \max\limits_{(a_0,s_1):P(a_0,s_0,s_1|s_0=s)>0}[r(a_0,s)+\overline{J}^{*,t-1}(s_1)] \quad \forall s$
3. return $\overline{J}^{*,T}$

1.3

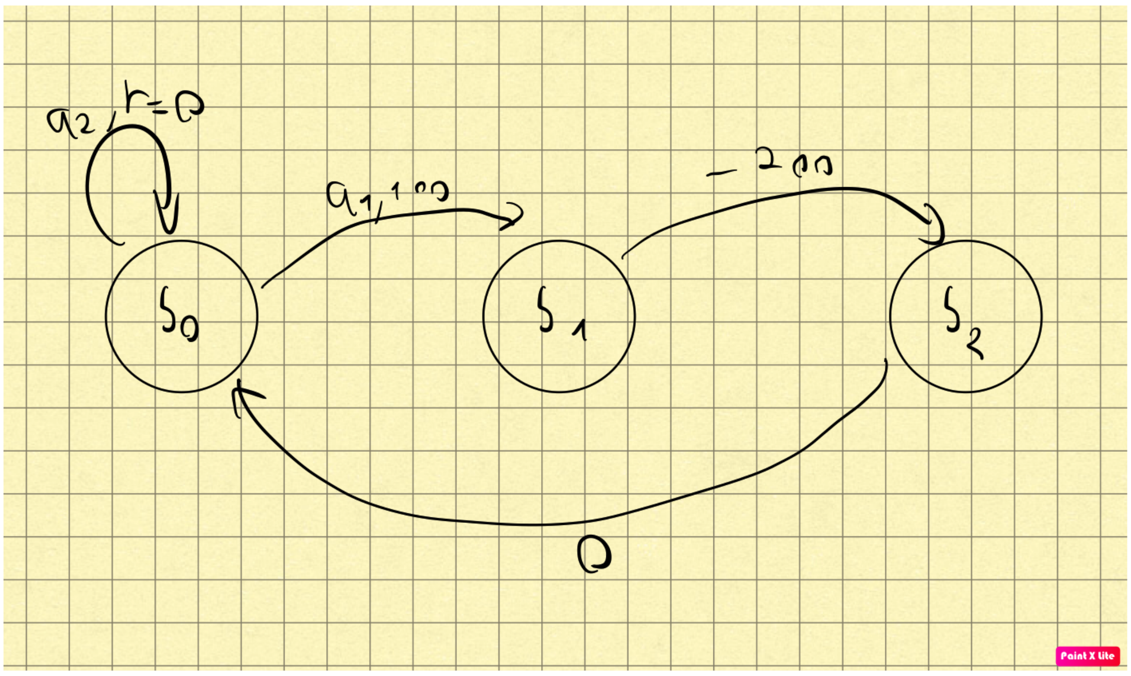

Figure 1: Counter example MDP

Set $T=2$, when we start at s_2 the first action is to move to s_0 and then we will do a_1.
when we start at s_0 we will do a_2 twice, so there is no deterministic optimal policy for all initial stages.

1.4

1.4.1

The dynamic programming equation will be: $\underline{J}^{\pi'}(s) = \sup\limits_{a} \min\limits_{s_1:P(s_1|s,a)>0}[r(s,a)+\gamma\underline{J}^{\pi}(s_1)]$

1.4.2

Let π, π' be two different policies and denote $J^{\pi} = J_1, J^{\pi'} = J_2$. Lets denote s as the state who satisfies: $|T\{J_1\}(s) - T\{J_2\}(s)| = \|T\{J_1\} - T\{J_2\}\|_\infty$.
Without loss of generality lets assume that we have $TJ_1(s) - TJ_2(s) > 0$.
Denote:
$a^* = \arg\max\limits_{a} \min\limits_{s_1:P(s_1|s,a)>0}[r(s,a)+\gamma\underline{J}_1(s_1)]$
$s_1^* = \arg\min\limits_{s_1:P(s_1|s,a)>0}[r(s,a)+\gamma\underline{J}_2(s_1)]$
Hence, we get:

$$\begin{aligned} T\{J_1\}(s) - T\{J_2\}(s) &= \sup_a \min_{s_1:P(s_1|s,a)>0}[r(s,a)+\gamma J_1(s_1)] - \sup_a \min_{s_1:P(s_1|s,a)>0}[r(s,a)+\gamma J_2(s_1)] \\ &\leq \min_{s_1:P(s_1|s,a^*)>0}[r(s,a^*)+\gamma J_1(s_1)] - \min_{s_1:P(s_1|s,a^*)>0}[r(s,a^*)+\gamma J_2(s_1)] \\ &\leq r(s,a^*)+\gamma J_1(s_1^*) - r(s,a^*) - \gamma J_2(s_1^*) = \gamma\cdot(J_1(s_1^*) - J_2(s_1^*)) \\ &\leq \|J_1 - J_2\|_\infty \end{aligned}$$

Hence, we got that the dynamic programming operator is a γ contraction with respect to the $\|\cdot\|_\infty$ norm.

2 Question 2

2.1

In order to define MDP we need to define the the stages, actions, transition probabilities and costs.
Let us define the set of N jobs $J = \{1, 2, ..., N\}$. The set of stages will be all the possible combinations in J - the power set of J (the empty set is also included and this is the terminal state). For each stage the possible actions will be to pick one of the remaining jobs. Assume we are in some stage S, with probability μ_i we will move to stage S/i, with probability $1 - \mu_i$ we will stay at S. When arriving to some stage S the instant cost will be $c(S) = -\sum\limits_{i \in S} c_i$.

The Bellman optimality equation is:

$$\begin{aligned} V^\pi(s) &= \min_a [r(s, a) + \gamma \sum_{s'} p(s'|s, a) \cdot V^\pi(s')] \\ &= \min_{i \in s} [c(s) + \gamma(\mu_i \cdot V^\pi(s \setminus i) + (1 - \mu_i) \cdot V^\pi(s))] \\ &= c(s) + \min_{i \in s} [(\mu_i \cdot V^\pi(s \setminus i) + (1 - \mu_i) \cdot V^\pi(s))] \qquad (1) \end{aligned}$$

in (1) we set $\gamma = 1$ because we want all the jobs to be finished.

2.2

Assuming we are on a certain stage s and we have M jobs left where $M \leq N$ left then $s = \{j_1, j_2, ..., j_M\}, j_k \in [N]$. We are using the following policy $\pi(s) = argmin_{i \in S}[\mu_i \cdot c_i]$. Define $\pi(s) = a_1$.The Bellman equation is:

$$V^\pi(s) = c(s) + \mu_{a_1} \cdot V^\pi(s \setminus a_1) + (1 - \mu_{a_1}) \cdot V^\pi(s)$$
$$V^\pi(s) = \frac{1}{\mu_{a_1}} c(s) + V^\pi(s \setminus a_1)$$

Using this result and $\pi(s \setminus a_1) = a_2$ we have :

$$V^\pi(s \setminus a_1) = \frac{1}{\mu_{a_2}} c(s \setminus a_1) + V^\pi(s \setminus (a_1, a_2))$$

after M steps we have stage s':

$$V^\pi(s \setminus (a_1, ..., a_{M-1})) = \frac{1}{\mu_{a_M}} c(s \setminus (a1, ..., a_{M-1}) + V^\pi(s \setminus (a_1, a_2, ..., a_M)) = \frac{1}{\mu_{a_M}} c(s \setminus (a1, ..., a_{M-1}) + V^\pi(\phi)$$

$V^\pi(\phi) = 0$ because all jobs are done:

$$V^\pi(s \setminus (a_1, ..., a_{M-1})) = \frac{1}{\mu_{a_M}} c(s \setminus (a1, ..., a_{M-1})$$

Now we can calculate $V^\pi(s)$ based on c expression only:

$$V^\pi(s) = \sum_{i=1}^{M} \frac{1}{\mu_{a_i}} c(s \setminus (..., a_i))$$

We can write it also as:

$$V^\pi(s) = -(\frac{1}{\mu_{a_1}})c_{a_1} - (\frac{1}{\mu_{a_1}} + \frac{1}{\mu_{a_2}})c_{a_2} - ... - (\frac{1}{\mu a_1} + ... + \frac{1}{\mu_{a_M}})c_{a_M}$$

and therefor:

$$V^\pi(s \setminus \{a_k\}) - V^\pi(s) = \frac{1}{\mu_{a_k}} c(actions\ before\ a_k) + c_{a_k} \sum_{a \in \{actions\ after\ a_k\}} \frac{1}{\mu_a}$$

observing the bellman optimailty equation we have:

$$V^{\pi*}(s) = c(s) + \min_{i \in s} [(\mu_i \cdot V^{\pi*}(s \setminus i) + (1 - \mu_i) \cdot V^{\pi*}(s))]$$

$$\begin{aligned} c(s) &= \max_{i \in s} [\mu_i (V^{\pi*}(s) - V^{\pi*}(s \setminus a_i)) \\ &= \max_{i \in s} [\mu_i (\frac{1}{\mu_{a_i}} c(actions\ before\ a_i) + c_{a_i} \sum_{a \in \{actions\ after\ a_i\}} \frac{1}{\mu_a})] \\ &= \max_{i \in s} [(c(actions\ before\ a_i) + \mu_i c_{a_i} \sum_{a \in \{actions\ after\ a_i\}} \frac{1}{\mu_a})] \end{aligned}$$

therefore we will choose i s.t $\mu_i \cdot c_i$ is maximal.

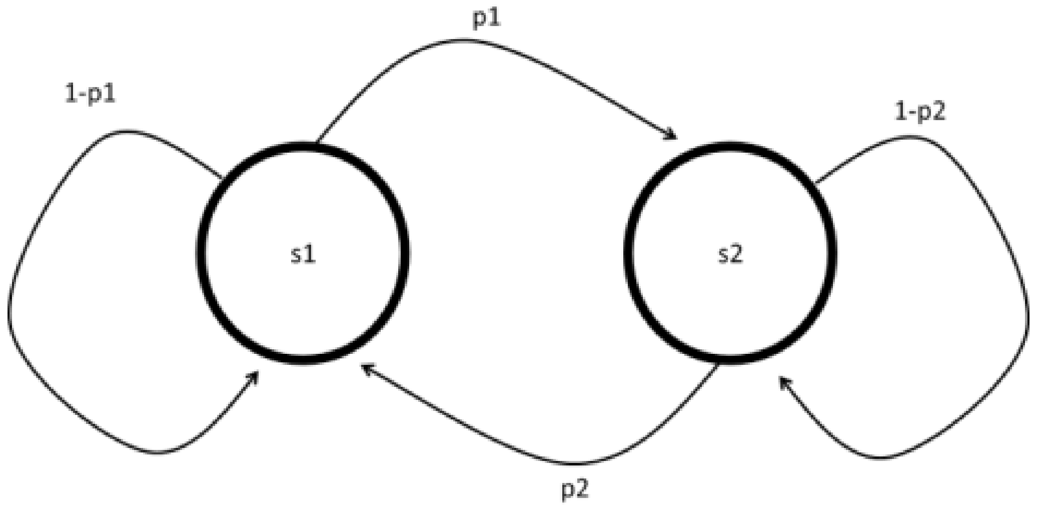

Figure 2: The proposed MDP

3 Question 3

Lets assume we have MDP as proposed in the hint [2].

Lest denote $i \in \{1, 2\}$ and $j = 3 - i$.
So, $(T^{\pi}(J))(s_i) = r(s_i, \pi(s_i)) + \gamma\left((1 - p_i)J(s_i) + p_i J(s_j)\right)$.

$$\begin{aligned}
&\Rightarrow (T^{\pi}(J_1))(s_i) - (T^{\pi}(J_2))(s_i) = \gamma\left((1 - p_i)(J_1(s_i) - J_2(s_i)) + p_i(J_1(s_j) - J_2(s_j))\right) \\
&\Rightarrow \|T^{\pi}(J_1) - T^{\pi}(J_2)\|_2^2 = \left(\gamma\left((1 - p_1)(J_1(s_1) - J_2(s_1)) + p_1(J_1(s_2) - J_2(s_2))\right)\right)^2 \\
&\qquad + \left(\gamma\left((1 - p_2)(J_1(s_2) - J_2(s_2)) + p_2(J_1(s_1) - J_2(s_1))\right)\right)^2
\end{aligned} \tag{2}$$

Setting up $p_1 = 1$, $p_2 = 0$, we get:

$$\begin{aligned}
\Rightarrow \|T^{\pi}(J_1) - T^{\pi}(J_2)\|_2^2 &= \left(\gamma(J_1(s_2) - J_2(s_2))\right)^2 + \left(\gamma\left(J_1(s_2) - J_2(s_2)\right)\right)^2 \\
&= 2\gamma^2\left(J_1(s_2) - J_2(s_2)\right)^2
\end{aligned} \tag{3}$$

Taking: $J_1(s_2) \neq J_2(s_2), J_1(s_1) = J_2(s_1), \gamma = 1$, we get:

$$\|T^{\pi}(J_1) - T^{\pi}(J_2)\|_2^2 = 2\left(J_1(s_2) - J_2(s_2)\right)^2 > \left(J_1(s_2) - J_2(s_2)\right)^2 = \|J_1 - J_2\|_2^2$$

Hence, we got a contradiction to the γ contraction.

4 Question 4

4.1

$$v^{\pi}(s) = \begin{cases} c(s, \pi(s)) + \sum\limits_{s'} \mathbb{P}(s'|s, \pi(s)) \cdot v^{\pi}(s') & , s \neq 0 \\ 0 & , s = 0 \end{cases}$$

$$v^{*}(s) = \begin{cases} \min\limits_{a \in \mathcal{A}(s)} \left[c(s, a) + \sum\limits_{s'} \mathbb{P}(s'|s, a) \cdot v^{*}(s')\right] & , s \neq 0 \\ 0 & , s = 0 \end{cases}$$

4.2

$$T_{\pi}\left(v\right)\left(s\right) = c(s, \pi(s)) + \sum_{s'} \mathbb{P}(s'|s, \pi(s)) \cdot v(s')$$

$$T\left(v\right)\left(s\right) = \min_{a \in \mathcal{A}(s)} \left[c(s, a) + \sum_{s'} \mathbb{P}(s'|s, a) \cdot v(s')\right]$$

4.3

It is necessary to assume proper policy in order to guarantee that the termination state would be reached with non zero probability.

4.4

4.4.1

As mentioned at the guidelines, lets consider new SSP with the same transitions and with costs all equal to -1, except the terminal state with 0. Lets define $\hat{J}(s)$ to be the optimal value from state s in the new SSP problem. We will define: $\xi(s) = -\hat{J}(s)$.

$$\begin{aligned}
\hat{J}(s) &= \min_a \left[c(s,a) + \sum_{s'} \mathbb{P}(s'|s,a) \cdot \hat{J}(s') \right] \\
&= -1 + \min_a \left[\sum_{s'} \mathbb{P}(s'|s,a) \cdot \hat{J}(s') \right] \\
&\leq -1 + \sum_{s'} \mathbb{P}(s'|s,a) \cdot \hat{J}(0) = -1 \\
&\Rightarrow \xi(s) \geq 1
\end{aligned}$$

4.4.2

$$\begin{aligned}
\hat{J}(s) &= \min_a \left[c(s,a) + \sum_{s'} \mathbb{P}(s'|s,a) \cdot \hat{J}(s') \right] \\
&\leq -1 + \sum_{s'} \mathbb{P}^{\pi}(s'|s) \cdot \hat{J}(s') \\
&= -1 - \sum_{s'} \mathbb{P}^{\pi}(s'|s) \cdot \xi(s') \\
&\Rightarrow \xi(s) - 1 \geq \sum_{s'} \mathbb{P}^{\pi}(s'|s) \cdot \xi(s')
\end{aligned}$$

4.4.3

We proved that $\xi(s) \geq 1 \quad \forall s$

$$\begin{aligned}
&\Rightarrow 0 \leq \xi(s) - 1 < \xi(s) \\
&\Rightarrow \frac{\xi(s) - 1}{\xi(s)} < 1 \quad \forall s \\
&\Rightarrow \beta = \max_{s'} \frac{\xi(s') - 1}{\xi(s')} < 1 \\
&\beta \cdot \xi(s) = \xi(s) \cdot \max_{s'} \frac{\xi(s') - 1}{\xi(s')} \geq \xi(s) \cdot \frac{\xi(s) - 1}{\xi(s)} = \xi(s) - 1 \\
&\Rightarrow \beta \cdot \xi(s) \geq \xi(s) - 1
\end{aligned}$$

4.4.4

$$\begin{aligned}
|T_\pi(J_1)(s) - T_\pi(J_2)(s)| &= |\sum_{s'} \mathbb{P}^{\pi}(s'|s)(J_1(s') - J_2(s'))| \\
\leq \sum_{s'} \mathbb{P}^{\pi}(s'|s)|J_1(s') - J_2(s')| &= \sum_{s'} \mathbb{P}^{\pi}(s'|s) \frac{|J_1(s') - J_2(s')|}{\xi(s')} \cdot \xi(s') \\
\leq \sum_{s'} \mathbb{P}^{\pi}(s'|s) \max_{\hat{s}} \left[\frac{|J_1(\hat{s}) - J_2(\hat{s})|}{\xi(\hat{s})} \right] \cdot \xi(s') &= \|J_1 - J_2\|_\xi \cdot \sum_{s'} \mathbb{P}^{\pi}(s'|s) \cdot \xi(s') \\
&\leq \|J_1 - J_2\|_\xi \cdot (\xi(s) - 1) \leq \|J_1 - J_2\|_\xi \cdot \beta \cdot \xi(s) \\
\Rightarrow \frac{|T_\pi(J_1)(s) - T_\pi(J_2)(s)|}{\xi(s)} &\leq \beta \cdot \|J_1 - J_2\|_\xi \quad \forall s \\
\Rightarrow \max_s \left[\frac{|T_\pi(J_1)(s) - T_\pi(J_2)(s)|}{\xi(s)} \right] &\leq \beta \cdot \|J_1 - J_2\|_\xi \\
\Rightarrow \|T_\pi(J_1) - T_\pi(J_2)\|_\xi &\leq \beta \cdot \|J_1 - J_2\|_\xi
\end{aligned}$$

Hence T_π is a β contraction with respect to the defined $\|\cdot\|_\xi$ norm.

5 Question 5

Lets look at the following MDP:
With $\gamma = 0.8$.
For policy $\hat{\pi}(s_1) = a_1$ we get: $v^{\hat{\pi}}(s_1) = \sum_{t=0}^{\infty} \gamma^t \cdot 2 = \frac{2}{1-\gamma} = 10$.
And for policy $\tilde{\pi}(s_1) = a_2$ we get: $v^{\tilde{\pi}}(s_1) = 0.8 \cdot 5 = 4$.
Now, assume we start with $V_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

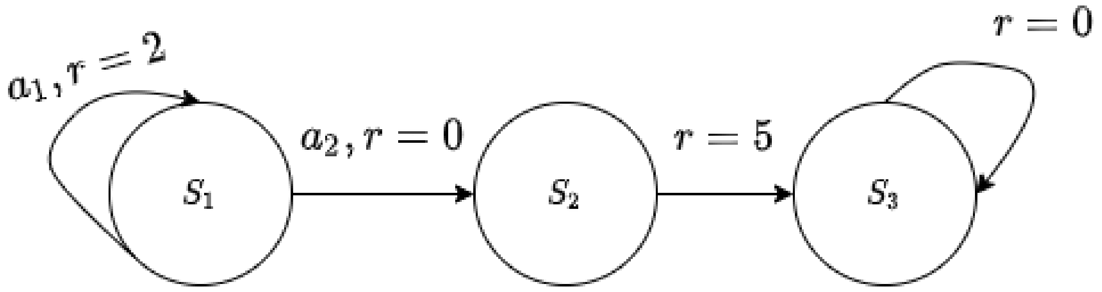

Figure 3: example MDP

At the first update state we get: $V_1 = \begin{pmatrix} \max(2,0) \\ 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 0 \end{pmatrix}$. And the current greedy policy would be $\pi_1(s_1) = a_1$. At the next update step we would get: $V_2 = \begin{pmatrix} \max(2 + 0.8 \cdot 2, 0.8 \cdot 5) \\ 5 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \\ 0 \end{pmatrix}$.
And the greedy policy would be: $\pi_2(s_1) = a_2$.
Therefor as we can see $V^{\pi_1}(s_1) > v^{\pi_2}(s_1)$.
And so, we can deduce that in this sort of algorithms we are not guarenteed to have a sequence of improving policies.