HW3 046203 RL

Submitters:
Submitter 1: David Valensi 342439643

Submitter 2: Yaniv Galron 206765646

Question 1: Worst Case Reward

1. Let's define $\underline{J}^{\pi,t',T}(s) = \min_{\{s_i,a_i\}_{t'}^{T}:P_\pi(\dots|s_{t'}=s)} \sum_{t=t'}^{T} r(s_t,a_t)$

   Our goal is to find $\underline{J}^{\pi,0,T}(s) = \underline{J}^{\pi,T}(s)$

   Basis:

   For $t' = T$: $\underline{J}^{\pi,t'=T,T}(s) = \min_{\{s_T,a_T\}:P_\pi(\dots|s_T=s)} \sum_{t=T}^{T} r(s_t,a_t) = \min_{\{s_T,a_T\}:P_\pi(\dots|s_T=s)} r(s_T,a_T) =$

   $$= \min_{\{s_T\}:P_\pi(\dots|s_T = s)} r(s_T)$$

   $Backward\ recursion:\ For\ t' = T-1, \dots:$

   $For\ all\ s \in S, we\ compute\ the\ following:$

   $$\underline{J}^{\pi,t',T}(s) = \min_{\{s_i,a_i\}_{t'}^{T}:P_\pi(\dots|s_{t'} = s)} \sum_{t'}^{T} r(s_t,a_t) =$$

   $$\underline{J}^{\pi,t',T}(s) = \min_{\{a_{t'}\in\pi(s_{t'}):\pi(a_{t'}|s_{t'})>0\},\ \{s_{t'+1}:P_\pi(s_{t'+1}|s_{t'},a_{t'})>0\}} [\,r(s_{t'},a_{t'}) + \underline{J}^{\pi,t'+1,T}(s_{t'+1})]$$

   Finally, we return $\underline{J}^{\pi,0,T}(s)$.

2. For a given MDP, let's define the following policy $\forall s \in S, a \in A,\ \pi(s,a) = \frac{1}{|A|}$, which takes each action with uniform probability. This policy enables every single transition (which exist in the MDP) with some positive probability: $P_\pi(s_{t'+1}|s_{t'},\pi(s_{t'})) > 0$

   This way, $\bar{J}^{*,T}(s) = \sup_{\pi'} \bar{J}^{\pi',T}(s) = \bar{J}^{\pi,T}(s)$. This is right since the supremum works on every possible stationary stochastic policy and we defined some stationary stochastic policy which enables all transitions in the given MDP. It means that our π will cover every possible sequence $\{s_i,a_i\}_{t'}^{T}: P_\pi(\dots|s_{t'} = s)$ in the given MDP.

   Thus, we can use the previous DP algorithm with maximum instead of minimum to find: $\bar{J}^{\pi,T}(s) = \max_{\{s_i,a_i\}_{t'}^{T}:P_\pi(\dots|s_{t'}=s)} \sum_{t=t'}^{T} r(s_t,a_t)$ by using the following backward for example:

   $$\bar{J}^{\pi,t',T}(s) = \max_{\{a_{t'}\in\pi(s_{t'}):\pi(a_{t'}|s_{t'})>0\},\ \{s_{t'+1}:P_\pi(s_{t'+1}|s_{t'},a_{t'})>0\}} [\,r(s_{t'},a_{t'}) + \bar{J}^{\pi,t'+1,T}(s_{t'+1})]$$

   Or similarly:

   $$\bar{J}^{\pi,T}(s) = \max_{\{a\in\pi(s):\pi(a|s)>0\},\ \{s':P(s|s,a)>0\}} \left[r(s,a) + \bar{J}^{\pi,T-1}(s')\right]$$

3. Yes, there exists a deterministic policy that attains $\bar{J}^{*,T}(s)$.

   Let's denote $\tau = (s_0,a_0,s_1,a_1,\dots,s_T,a_T)\ s.t. \sum_{t=t'}^{T} r(s_t,a_t) = \bar{J}^{*,T}(s)$

   We define the deterministic policy:

   $$\pi\ s.t.\ \pi(s) = a\ where\ a,s' \in \operatorname*{argmax}_{\{a\in\pi(s):\pi(a|s)>0\},\ \{s':P(s'|s,a)>0\}} \left[r(s,a) + \bar{J}^{\pi,T-1}(s')\right]$$

   Let's prove that $\bar{J}^{*,T}(s) = \bar{J}^{\pi,T}(s)$,by induction for all t.

   Basis: $t = 0$: $\bar{J}^{\pi,0}(s) = r(s) = \bar{J}^{*,0}(s)$

   Step: $Let's\ assume\ that\ \bar{J}^{*,t}(s) = \bar{J}^{\pi,t}(s)\ for\ some\ t.$

Then $\bar{J}^{\pi,t+1}(s) = r\big(s,\pi(s)\big) + \max\limits_{\{s':P\left(s'\middle|s,\pi(s)\right)>0}\bar{J}^{\pi,t}(s') =$

From the definition of $\pi(s)$:

$$= \max_{\{a\in\pi(s):\pi(a|s)>0\},\ \{s':P(s'|s,a)>0\}}\left[r(s,a) + \bar{J}^{\pi,t}(s')\right] =$$

$$=_{(induction\ hypothesis)} \max_{\{a\in\pi(s):\pi(a|s)>0\},\ \{s':P(s'|s,a)>0\}}\left[r(s,a) + \bar{J}^{*,t}(s')\right] = \bar{J}^{*,t+1}(s)$$

$\Rightarrow \bar{J}^{*,T}(s) = \bar{J}^{\pi,T}(s)$.

4. i. The supremum is taken over all policies, so the actions are maximizing the expression, but by definition of $\underline{J}$, we consider the worst-case transition in the MDP
   Thus, the DP equations are

$$\underline{J}^*(s) = \max_{a\in A}\min_{\{s':P(s'|s,a)>0\}}\left[r(s,a) + \gamma\underline{J}^*(s')\right]$$

   ii. It results the following DP operator:

$$T^*\underline{J}\ (s) = \max_{a\in A}\min_{\{s':P(s'|s,a)>0\}}\left[r(s,a) + \gamma\underline{J}^*(s')\right]$$

   We must prove that

$$\left\|T^*\underline{J_1}\ (s) - T^*\underline{J_2}\ (s)\right\|_\infty \le \gamma\left\|\underline{J_1} - \underline{J_2}\right\|_\infty$$

   As in class, let's take

$$a_1 = \operatorname*{argmax}_{a\in A}\min_{\{s':P(s'|s,a)>0\}}\left[r(s,a) + \gamma\underline{J_1}^*(s')\right]$$

   For all state s, it holds:

$$T^*\underline{J_2}\ (s) = \max_{a\in A}\min_{\{s':P(s'|s,a)>0\}}\left[r(s,a) + \gamma\underline{J_2}^*(s')\right] = \min_{\{s':P(s'|s,a_2)>0\}}\left[r(s,a_2) + \gamma\underline{J_2}^*(s')\right]$$

$$T^*\underline{J_1}\ (s) = \max_{a\in A}\min_{\{s':P(s'|s,a)>0\}}\left[r(s,a) + \gamma\underline{J_1}^*(s')\right] \ge \min_{\{s':P(s'|s,a_2)>0\}}\left[r(s,a_2) + \gamma\underline{J_1}^*(s')\right]$$

$$\Rightarrow T^*\underline{J_1}\ (s) - T^*\underline{J_2}\ (s)$$

$$\le \min_{\{s':P(s'|s,a_2)>0\}}\left[r(s,a_2) + \gamma\underline{J_1}^*(s')\right] - \min_{\{s':P(s'|s,a_2)>0\}}\left[r(s,a_2) + \gamma\underline{J_2}^*(s')\right]$$

$$= \gamma\left(\min_{\{s':P(s'|s,a_2)>0\}}\left[\underline{J_1}^*(s')\right] - \min_{\{s':P(s'|s,a_2)>0\}}\left[\underline{J_2}^*(s')\right]\right)$$

$$\le (*)\gamma\max_{s}\left(\underline{J_1}^*(s) - \underline{J_2}^*(s)\right) = \gamma\left\|\underline{J_1} - \underline{J_2}\right\|_\infty$$

   (*) As we learned in class, the difference between two function minima is smaller or equal to the maximal difference between two functions.
   Similarly, with a_2, we can prove that

$$T^*\underline{J_2}\ (s) - T^*\underline{J_1}\ (s) \le \gamma\left\|\underline{J_1} - \underline{J_2}\right\|_\infty$$

$$\Rightarrow \left\|T^*\underline{J_1}\ (s) - T^*\underline{J_2}\ (s)\right\|_\infty \le \gamma\left\|\underline{J_1} - \underline{J_2}\right\|_\infty$$

Thus, T^* is a contraction.

iii. (Bonus) T^* is a contracting. We remember that from Banach-fixed-point Theorem, there exists a unique solution $\underline{J}^*$ to the equation $T^*\underline{J} = \underline{J}$.

Thus, we can define the following stationary deterministic policy

$$\forall s, \pi(s) = arg\max_{\mathrm{a}\in\mathrm{A}} \min_{\{s':\mathrm{P}(s'|s,a)>0\}} \left[r(s,a) + \gamma \underline{J}^*(s')\right]$$

Such that by definition we attain the optimal value $\underline{J}^*$:

$$\underline{J}^\pi(s) = \min_{\{s':\mathrm{P}\left(s'\middle|s,\pi(s)\right)>0\}} \left[r\left(s,\pi(s)\right) + \gamma \underline{J}^*(s')\right] = \underline{J}^*(s)$$

We can write in operator notations.

$$\underline{J}^* = T^*\underline{J}^* = r\left(s,\pi(s)\right) + \gamma P_\pi \underline{J}^* = T^\pi \underline{J}^*$$

$I.e. \pi\ is\ the\ optimal\ policy\ and\ attains\ \underline{J}^*.$

Question 2 - The $c\mu$ rule

a. we define the following MDP for the server problem:

State space: state S_t will hold all jobs which are still unfinished in the system.

The state space is actually $S = P(\{1 \dots n\})$ - the power set of $\{1 \dots n\}$. Thus $|S| = 2^n$

Action space: action a_t will mark the job chosen by the server. The action space is $A = \{1 \dots n\}$

Thus $|A| = n$

Transition probabilities: when choosing an action $a_t = i$ there are two options - the job is completed and it is taken out of the system or the job is uncompleted and is putted back in the system. Thus

$$P(s \mid S_t = B, a_t = i) = \begin{cases} \mu_i & s = B/\{i\} \\ 1-\mu_i & s = B \\ 0 & \text{else} \end{cases}$$

cost: cost $c_t(S_t, a_t)$ will be the total cost of all jobs which are still in the system. we notice that the cost doesnt depend on the action or the time.

we have

$$C_t(S_t, a_t) = C_t(S_t) = \sum_{i=1}^{n} I(i \in S_t) C_i$$

The total cost we wish to minimize is

$$J^{\pi}(S = \{1 \dots n\}) = E^{\pi, S}\left(\sum_{t=0}^{\infty} c(S_t)\right)$$

we also notice that once we reached ~~out~~ the state $S_t = \{\}$ the process ends and the cost for all $t \geq \tau$ is $C_t(S_t) = 0$

Bellman equation:

$$V(S) = \min_{a = i \in A}\left\{c(S) + \sum_{s'} P(s' \mid s, i) V(s')\right\} = c(S) + \min_{a = i \in A}\left\{\mu_i V(S/\{i\}) + (1-\mu_i) V(S)\right\}$$

b. we wish to prove that the stationary policy $j^* = \arg\max_{i} \{c_i \mu_i\}$ is an optimal policy. we know that a policy is optimal if its value function satisfies the Bellman equation. we will compute the value function for the above policy and show that it does.

for convenience we mark $S = \{i_1 \dots i_k\}$ s.t. the jobs $i_1 \dots i_k$ are in an ordered way s.t $\mu_{i_1} c_{i_1} \geq \dots \geq \mu_{i_k} c_{i_k}$

In the suggested policy and using the above marking we take job i_1:

$$V(S) = V(\{i_1 \dots i_k\}) = c(S) + \mu_{i_1} V(S/\{i_1\}) + (1-\mu_{i_1}) V(S)$$

$$\Rightarrow V(S) = \frac{c(S)}{\mu_{i_1}} + V(S/\{i_1\}) = \frac{1}{\mu_{i_1}} \sum_{j=1}^{k} c_{i_j} + V(S/\{i_1\})$$

Applying the recursion to $V(S_k)$ we get

$$V(S) = \frac{1}{\mu_{i_1}} \sum_{j=1} c_{i_j} + V(S/\{i_1\}) = \dots = \frac{1}{\mu_{i_1}} \sum_{j=1} c_{i_j} + \dots + \frac{1}{\mu_{i_k}} \sum_{j=k} c_{i_j}$$

Similarly for $V(S/\{i_a\})$ we take c_{i_a} out from each summand and cancel the $\frac{1}{\mu_{i_a}}$ term

$$V(S/\{i_a\}) = V(S) - \left(c_{i_a} \sum_{j=1}^{a-1} \frac{1}{\mu_{i_j}} + \frac{1}{\mu_{i_a}} \sum_{j=a}^{k} c_{i_j}\right)$$

now we show that for each state S the Bellman equation holds:

$$V(S) = c(S) + \min_{i_a \in A} \{\mu_{i_a} V(S/\{i_a\}) + (1-\mu_{i_a}) V(S)\}$$

$$= c(S) + V(S) + \min_{i_a \in A} \{\mu_{i_a} V(S/\{i_a\}) - \mu_{i_a} V(S)\}$$

$$c(S) = -\min_{i_a \in A} \{\mu_{i_a} (V(S/\{i_a\}) - V(S))\} = \max_{i_a \in A} \{\mu_{i_a} (V(S) - V(S/\{i_a\}))\}$$

$$= \max_{i_a \in A} \left\{ \mu_{i_a} C_{i_a} \sum_{j=1}^{a-1} \frac{1}{\mu_{i_j}} + \sum_{j=a}^{K} C_{i_j} \right\}$$

The maximum above is achieved for $i_a = i_1$ and thus we get:

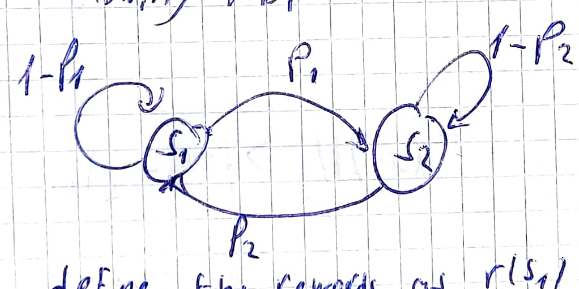
$$= \mu_{i_1} C_{i_1} \sum_{j=1}^{1-1} \frac{1}{\mu_{i_j}} + \sum_{j=1}^{K} C_{i_j} = \sum_{j=1}^{K} C_{i_j} = C(S)$$ ∎

Question 3 - DP operator not contracting in Euclidean norm

The fixed operator T^{π}:

$$(T^{\pi}(J))(s) = r(s, \pi(s)) + \gamma \sum_{s'} P(s' | s, \pi(s)) J(s')$$

We wish to prove that T^{π} is not necessarily a contraction in the Euclidean norm. As hinted, we will use the following MDP:



We also define the rewards as $r(s_1) = r_1$ $r(s_2) = r_2$
we will show that for appropriate P_1, P_2, γ, there exist J_1, J_2 such that:

$$\| T^{\pi}(J_1) - T^{\pi}(J_2) \|_2 \geq \| J_1 - J_2 \|_2$$

First for the above MDP we have:

$$(T^{\pi}(J))(s_i) = r_i + \gamma((1-P_i)J(s_i) + P_i J(s_{3-i}))$$

$$\Rightarrow (T^{\pi}(J_1) - T^{\pi}(J_2))(s_i) = \gamma((1-P_i)(J_1(s_i) - J_2(s_i)) + P_i(J_1(s_{3-i}) - J_2(s_{3-i})))$$

$$T^{\pi}(J_1) - T^{\pi}(J_2) = \begin{pmatrix} \gamma((1-p_1)(J_1(s_1)-J_2(s_1)) + p_1(J_1(s_2)-J_2(s_2))) \\ \gamma((1-p_2)(J_1(s_2)-J_2(s_2)) + p_2(J_1(s_1)-J_2(s_1))) \end{pmatrix}$$

$$\|T^{\pi}(J_1) - T^{\pi}(J_2)\|_2^2 = \gamma^2\Big(((1-p_1)^2+p_2^2)(J_1(s_1)-J_2(s_1))^2$$
$$+2(p_1(1-p_1)+p_2(1-p_2))(J_1(s_1)-J_2(s_1))(J_1(s_2)-J_2(s_2))$$
$$+((1-p_2)^2+p_1^2)(J_1(s_2)-J_2(s_2))^2\Big)$$

we take the following $J_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ $J_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

for these we have:

$$\|J_1 - J_2\|_2^2 = 1$$

$$\|T^{\pi}(J_1) - T^{\pi}(J_2)\|_2^2 = \gamma^2((1-p_2)^2+p_1^2) = (*)$$

lets take $p_1 = 1$, $p_2 = 0$, $\gamma = 0.8$:

$$(*) = 1.28 > 1 = \|J_1 - J_2\|_2^2$$

We have found that for a given J_1 J_2 we get

$$\|T^{\pi}(J_1) - T^{\pi}(J_2)\|_2 > \|J_1 - J_2\|_2$$

Thus T^{π} is not a contraction under the Euclidean norm.

Question 4 – Stochastic Shortest Path

1. We define for this MDP the value function v^{π} as we learned in class:
   $\tau = \inf\{t \geq 0 \; s.t. \, s_t = 0 \; (termination\ state)\}$
   $V_{ssp}^{\pi}(s) = \mathbb{E}^{\pi}(\sum_{t=0}^{\tau-1} c(s_t, a_t) + r_G(0)|s_0 = s) \;=_{(*)}\; \mathbb{E}^{\pi}(\sum_{t=0}^{\tau-1} c(s_t, a_t))|s_0 = s)$

   *Since terminal state is cost free, once the state 0 is reached, the best way to behave is to stay at goal state since $c(0, a) = 0, \forall a \; and \; c(s, a) > 0, \forall s \neq 0, a$
   The Bellman Equations are
   $$V^{\pi}(s) = \begin{cases} c\big(s, \pi(s)\big) + \sum_{s'} p(s'|s, \pi(s))V^{\pi}(s'), & if \; s \neq 0 \\ 0, & if \; s = 0 \end{cases}$$
   For the optimal policy, it holds:
   $$V^{*}(s) = \begin{cases} \min_{a} \left\{ c(s, a) + \sum_{s'} p(s'|s, a)V^{\pi}(s') \right\}, & if \; s \neq 0 \\ 0, & if \; s = 0 \end{cases}$$
2. For a fixed stationary policy $\pi: S \to A$, the Bellman operator $T^{\pi}: \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ is:
   $$\big(T^{\pi}(V)\big)(s) = \begin{cases} c\big(s, \pi(s)\big) + \sum_{s'} p(s'|s, \pi(s))V(s'), & if \; s \neq 0 \\ 0, & if \; s = 0 \end{cases}$$
   The Bellman Operator is:
   $$\big(T^{*}(V)\big)(s) = \begin{cases} \min_{a} c(s, a) + \sum_{s'} p(s'|s, a))V(s'), & if \; s \neq 0 \\ 0, & if \; s = 0 \end{cases}$$
3. The cost function is positive $c(s, a) > 0, \forall s \neq 0, a. \, And \; \gamma = 1$ (given).
   For a non-proper policy, it holds that
   $\tau = \inf\{t \geq 0 \; s.t. \, s_t = 0\} = \infty \Rightarrow J^{*}(s) = \mathbb{E}^{*}(\sum_{t=0}^{\tau=\infty} \gamma^t c(s_t, a_t))|s_0 = s) = \infty$
   I.e. the optimal cost function J^{*} is not finite.
4. First, we consider the SSP problem with same transitions but with costs
   $c(s) = -1 \; \forall s \neq 0 \, , c(0) = 0.$
   For any policy we have $J^{\pi}(s) = \mathbb{E}^{\pi}(\sum_{t=0}^{\tau-1} c(s_t, a_t))|s_0 = s) \leq 0 \; (*)$ since $c \leq 0$.
   $\hat{J}(s)$, the optimal value from state s, holds that
   $$\hat{J}(s \neq 0) = \min_{a}\left\{ c(s, a) + \sum_{s'} p(s'|s, a)\hat{J}(s') \right\}$$
   $$= -1 + \min_{a}\left\{ \sum_{s'} p(s'|s, a)\hat{J}(s') \right\} \leq \; -1 + \sum_{s'} p(s'|s, a)\hat{J}(s'), a \in A$$

   $$\hat{J}(s \neq 0) \leq \; -1 + \sum_{s'} p(s'|s, a) \; \max_{s''} \hat{J}(s'') \leq -1 + \max_{s''} \hat{J}(s'') \leq_{*} -1$$
   Thus, defining $\xi(s) = -\hat{J}(s) \Rightarrow \xi(s) \; \geq_{**} 1.$

Let's write the Bellman Optimality Equation for $\hat{J}(s)$, the optimal value from state s.

$$\hat{J}(s \neq 0) = \min_a \left\{ c(s,a) + \sum_{s'} p(s'|s,a)\hat{J}(s') \right\}$$

a) $For\ any\ stationary\ policy\ \pi, we\ have\ that$:

$$\min_a \left\{ \sum_{s'} p(s'|s,a)\hat{J}(s') \right\} \leq \sum_{s'} p(s'|s,\pi(s))\hat{J}(s')$$

$$\Rightarrow \hat{J}(s \neq 0) = -1 + \min_a \left\{ \sum_{s'} p(s'|s,a)\hat{J}(s') \right\} \leq -1 + \sum_{s'} p\big(s'\big|s,\pi(s)\big)\hat{J}(s')$$

$$= -1 + \sum_{s'} p^{\pi}(s'|s)\hat{J}(s')$$

$$\Rightarrow \xi(s) = -\hat{J}(s) \geq +1 - \sum_{s'} p^{\pi}(s'|s)\hat{J}(s') = 1 + \sum_{s'} p^{\pi}(s'|s)\xi(s')$$

$$\Rightarrow \xi(s) - 1 \geq \sum_{s'} p^{\pi}(s'|s)\xi(s')$$

b) $\xi(s) - 1 = \frac{\xi(s)-1}{\xi(s)}\xi(s) \leq \max_{s'} \frac{\xi(s')-1}{\xi(s')}\xi(s) = \beta\xi(s)$

$**: \forall s'\colon\ \xi(s') > \xi(s') - 1\ \geq 0 \Rightarrow \beta = \max_{s'} \frac{\xi(s') - 1}{\xi(s')} < 1$

Joining a+b, we have: $\sum_{s'} p^{\pi}(s'|s)\xi(s') \leq \xi(s) - 1 \leq \beta\xi(s) \quad (***)$

Now, $\forall J_1, J_2 \in \mathbb{R}^S$: $|T_\pi J_1(s) - T_\pi J_2(s)\ | =$

$$\left| c\big(s,\pi(s)\big) + \sum_{s'} p\big(s'\big|s,\pi(s)\big)J_1(s') - c\big(s,\pi(s)\big) + \sum_{s'} p\big(s'\big|s,\pi(s)\big)J_2(s') \right| =$$

$$\left| \sum_{s'} p\big(s'\big|s,\pi(s)\big)J_1(s') - \sum_{s'} p\big(s'\big|s,\pi(s)\big)J_2(s') \right|$$

$$= \left| \sum_{s'} p\big(s'\big|s,\pi(s)\big)\left(J_1(s') - J_2(s')\right) \right| \leq \sum_{s'} p\big(s'\big|s,\pi(s)\big)\big|J_1(s') - J_2(s')\big|$$

$$= \sum_{s'} p\big(s'\big|s,\pi(s)\big)\xi(s')\frac{\left(\big|J_1(s') - J_2(s')\big|\right)}{\xi(s')}$$

$$\leq \sum_{s'} p\big(s'\big|s,\pi(s)\big)\xi(s') \max_{s''} \frac{\left(\big|J_1(s'') - J_2(s'')\big|\right)}{\xi(s'')}$$

$$= \left\|J_1 - J_2\right\|_\xi \sum_{s'} p\big(s'\big|s,\pi(s)\big)\xi(s') \leq_{***} \left\|J_1 - J_2\right\|_\xi \beta\xi(s)$$

$$\Rightarrow \frac{|T_\pi J_1(s) - T_\pi J_2(s)\,|}{\xi(s)} \leq \left\|J_1 - J_2\right\|_\xi \beta \qquad (from\ **: \xi(s) \geq 1\ s.t. we\ can\ divide)$$

The previous inequality holds for all states s. Thus, we can write

$\|T^\pi(J_1) - T^\pi(J_2)\|_\xi = \max_s \frac{|T_\pi J_1(s) - T_\pi J_2(s)\,|}{\xi(s)} \leq \ \|J_1 - J_2\|_\xi \beta$

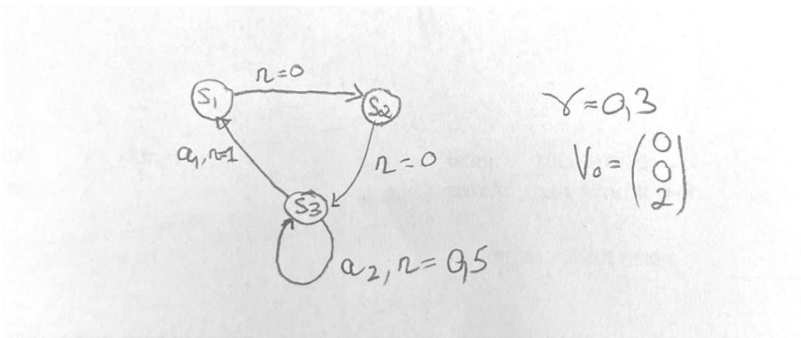
We proved that T^π is a contraction operator with the weighted maximum norm that's defined, and the contraction coefficient is β.

Question 5:

As we learned in class, the Value Iteration algorithm produces Value functions and thus greedy policies which are not necessarily better than the previous steps policies. However, there is a convergence guarantee to the optimal V^* and π^*.

We will show a <u>counter example</u> to the claim "VI algorithm produces a sequence of V_i and greedy policy π_i w.r.t. V_i s.t. $V^{\pi_i} \geq V^{\pi_{i-1}}\ \forall i$"

For the following MDP:



With $V_0 = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}, \gamma = 0.3$

Of course, for this MDP we have $\pi_i(s_1) = \pi_i(s_2) = \sim (any\ action)$

We use the VI updates:

$V_{i+1}(s) = \max\limits_{a}\{r(s,a) + \gamma \sum_{s'} p(s'|s,a)V_i(s')\}$.

$V_1(s_1) = 0 + \gamma 0 = 0$

$V_1(s_2) = 0 + 2\gamma = 2 * 0.3 = 0.6$

$V_1(s_3) = \max\limits_{a \in \{a_1,a_2\}}\{1 + \gamma * V_0(s_1), 0.5 + \gamma * V_0(s_3)\} = \max\limits_{a \in \{a_1,a_2\}}\{1 + 0.3 * 0, \qquad 0.5 + 0.3 * 2\}$

$= 1.1$

$\Rightarrow \pi_1(s_3) = a_2$

$V_2(s_1) = 0 + 0.3 * V_1(s_2) = 0.3 * 0.6 = 0.18$

$V_2(s_2) = 0 + 0.3 * 1.1 = 0.33$

$V_2(s_3) = \max\limits_{a \in \{a_1,a_2\}}\{1 + 0.3 * V_1(s_1), 0.5 + 0.3 * V_1(s_3)\} = \max\limits_{a \in \{a_1,a_2\}}\{1, 0.83\} = 1$

$\Rightarrow \pi_2(s_3) = a_1$

We observe for this sequence that

$V_1(s_3) = 1.1 < 2 = V_0(s_3)$

$V_2(s_3) = 1 < 1.1 = V_1(s_3)$

In other words, we see twice that $V^{\pi_i} < V^{\pi_{i-1}}$ $for\ some\ s$ as needed.