Dry3 - 046203

Zohar Rimon

03.06.2020

1 Simple MDP

1.1

Obviously, the optimal policy in this case is:

$$\mu(s) = \begin{cases} \text{left}, & \text{if s = n} \\ \text{right}, & \text{else} \end{cases}$$

This policy will be the one to visit the last state the fastest, this is important because we have a discounted reward.

1.2

The definition of V:

$$V^{\pi}(s) \triangleq E^{\pi}\left(\sum_{t=0}^{\infty}\gamma^{t}r\left(s_{t},a_{t}\right)|s_{0}=s\right)$$

In our case, $\pi = \mu$, is deterministic and the dynamics are determenstic as well, thus:

$$V^{\mu}(s) = \sum_{t=0}^{\infty}\gamma^{t}r\left(s_{t},\mu(s_{t})\right), s_{0}=s$$

The first time we will visit the n'th state depends on the initial state s_0:

$$t_0 = n - s_0$$

Afterwards we will visit the n'th state every n steps and get a reward of 1 when we do, so:

$$V^{\mu}(s) = \gamma^{n-s_0} + \sum_{i=1}^{\infty}\gamma^{i\cdot n+t_0}\cdot 1 = \gamma^{n-s_0} + \sum_{i=1}^{\infty}\gamma^{i\cdot n+n-s_0} = \gamma^{n-s_0}\cdot\sum_{i=0}^{\infty}\gamma^{i\cdot n}$$

We know that $\gamma < 1$, so:

$$V^{\mu}(s) = \frac{\gamma^{n-s_0}}{1-\gamma^{n}}$$

The calculation could have been done with the Bellman equation for the discounted value function as well.

1.3

The transition matrix:

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0.5 & 0 & 0.5 & 0 & \dots & 0 \\ 0.5 & 0 & 0 & 0.5 & \dots & 0 \\ 0.5 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

The stationary distribution holds:

$$d = d \cdot P, \quad d_i \geq 0, \quad \sum d_i = 1$$

Plugging in the transition matrix:

$$\begin{cases} d_1 = 0.5 \cdot (d_2 + d_3 + \cdots + d_{n-1} + 2 \cdot d_n) \\ d_2 = d_1 \\ d_i = 0.5 \cdot d_{i-1}, \quad \forall i \in [3, \dots n] \end{cases}$$

Thus:

$$\begin{cases} d_1 = 0.5 \cdot (d_2 + d_3 + \cdots + d_{n-1} + 2 \cdot d_n) \\ d_2 = d_1 \\ d_i = 0.5^{i-2} \cdot d_1, \quad \forall i \in [3, \dots n] \end{cases}$$

Plugging the result into the normalization constraint:

$$\begin{aligned} \sum_{i=1}^{n} d_i = d_1 + d_1 + \sum_{i=3}^{n} 0.5^{i-2} d_1 = d_1 \cdot \left(2 + \sum_{i=1}^{n-2} 0.5^i\right) = d_1 \cdot \left(1 + \sum_{i=0}^{n-2} 0.5^i\right) = d_1 \cdot \left(1 + \frac{1 - 0.5^{n-1}}{0.5}\right) \\ = d_1 \cdot \left(3 - 0.5^{n-2}\right) = 1 \end{aligned}$$

Thus:

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

Plugging back the result:

$$\begin{cases} d_1 = \frac{1}{3-0.5^{n-2}} \\ d_2 = \frac{1}{3-0.5^{n-2}} \\ d_{i+1} = \frac{0.5^{i-2}}{3-0.5^{n-2}}, \quad \forall i \in [3, \dots n] \end{cases}$$

1.4

The iterative process in the value iteration algorithm (the probability here is due to the stochastic manner of the policy):

$$V_{n+1}(s) = E^{\pi}\left[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V_n(s')\right], \quad \forall s \in S$$

The expectancy over π is due to the fact that π is stochastic.
In out case $v^0(s) = 0$, so:

$$v^1(s) = E^{\pi}[r(s,a)] = \begin{cases} 1 & \text{if s = n} \\ 0 & \text{else} \end{cases}$$

$$v^2(s) = E^{\pi}[r(s,a)] + \gamma E^{\pi}\left[\sum_{s' \in S} p(s'|s,a) v^1(s')\right] = \begin{cases} 1 & \text{if s = n} \\ 0 + \gamma(0.5 \cdot 1 + 0.5 \cdot 0) & \text{if s=n-1} \\ 0 & \text{else} \end{cases}$$

$$= \begin{cases} 1 & \text{if s = n} \\ 0.5 \cdot \gamma & \text{if s=n-1} \\ 0 & \text{else} \end{cases}$$

$$v^3(s) = E^{\pi}[r(s,a)] + \gamma E^{\pi}\left[\sum_{s' \in S} p(s'|s,a) v^2(s')\right] = \begin{cases} 1 & \text{if s = n} \\ 0 + 0.5 \cdot \gamma \cdot v^2(n) & \text{if s = n-1} \\ 0 + 0.5 \cdot \gamma \cdot v^2(n-1) & \text{if s = n-2} \\ 0 & \text{else} \end{cases}$$

$$= \begin{cases} 1 & \text{if s = n} \\ 0.5 \cdot \gamma & \text{if s=n-1} \\ 0.25 \cdot \gamma^2 & \text{if s=n-2} \\ 0 & \text{else} \end{cases}$$

We know that:

$$v^n(s) = E^\pi \left(\sum_{t=0}^{n-1} \gamma^t r\left(s_t, a_t\right) + \gamma^n v^0\left(s_n\right) | s_0 = s \right)$$

Where E^π is due to the fact that the policy is stochastic. Now let's look at the difference between $v^n(s)$ and $V^\pi(s)$ (in our case v^0 is 0, but the following hold for the general case, as seen in class):

$$v^\pi(s) - v^n(s) = E^\pi \left(\sum_{t=n}^{\infty} \gamma^t r\left(s_t, a_t\right) | s_0 = s \right)$$

the reward is bounded and $\gamma < 1$, so obviously, for $n \longrightarrow \infty$ the difference between the two value function will go to zero. Concluding:

$$v^\infty(s) = \lim_{n \to \infty} \left(v^n(s)\right) = V^\pi(s), \quad \forall s \in S$$

1.5

We will denote the expectancy of the first visit to the i state as $E[N_i]$. We will denote the random variable (Bernoulli), representing an indicator of a successful transition from state i to i+1 at the first visit of state i, and using the law of total expectation:

$$\begin{aligned} E[N_{i+1}] &= E[E[N_{i+1}|T]] = 0.5 \cdot E[N_{i+1}|T=1] + 0.5 \cdot E[N_{i+1}|T=0] \\ &= 0.5\left(\cdot E[N_i|T=1] + 1\right) + 0.5 \cdot \left(E[N_i] + 1 + E[N_{i+1}]\right) \\ &= E[N_i] + 1 + 0.5 \cdot E[N_{i+1}] \end{aligned}$$

Extracting $E[N_{i+1}]$:

$$E[N_{i+1}] = 2 \cdot \left(E[N_i] + 1\right)$$

We also know that $E[N_2] = 1$, so:

$$\begin{aligned} E[N_n] &= 2\left(E[N_{n-1}] + 1\right) = 2\left(2\left(E[N_{n-2}] + 1\right) + 1\right) \\ &= 2^2\left(E[N_{n-2}] + 1 + 0.5\right) = \cdots = 2^{n-2}\left(E[N_2] + 1 + 0.5 + 0.5^2 + \cdots + 0.5^{n-3}\right) \\ &= 2^{n-2}\left(1 - 2 \cdot \left(0.5^{n-2} - 1\right)\right) = 3 \cdot 2^{n-2} - 2 \end{aligned}$$

1.6

In order to reach every state with minimal steps, the policy that we will take is:

$$\pi_t(s) = \begin{cases} \text{left}, & \text{if first time at s} \\ \text{right}, & \text{else} \end{cases}$$

After visiting the i state, the number of steps it will take to visit the i+1 state is i+1 and we will have one more step in order to go back to the first state (total of i+2). The special cases are the first state, which only has one possible transition and the last state (which does not require to go back to after the transition to the first state). Thus the total number of steps in order to visit every transition is:

$$N = 1 + 3 + 4 + \cdots + (n-1) + 1 = \frac{n(1+n)}{2} - 1$$

1.7

Now we act in order to maximize the discounted return. Thus, we first will go right at each state, this is due to the fact that if we will go left the first state will not result in a reward now (because we already visited it). Next we would visit every left transition (to the first state). We can see that we will get the same number of steps as the previous section:

$$N = \frac{n(1+n)}{2} - 1$$

1.8

We know that a V^π is a unique solution to the linear system of equations:

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$$

Where:

$$r^\pi = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, V^\pi = \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 1 & 0 & 0 \end{pmatrix}$$

Thus:

$$V^\pi = \left(I - 0.5 \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 1 & 0 & 0 \end{pmatrix} \right)^{-1} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{16}{13} & \frac{8}{13} & \frac{2}{13} \\ \frac{6}{13} & \frac{16}{13} & \frac{4}{13} \\ \frac{8}{13} & \frac{4}{13} & \frac{14}{13} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{2}{13} \\ \frac{4}{13} \\ \frac{14}{13} \end{pmatrix}$$

2 The cμ rule

2.1

The states of the MDP will be subsets of the jobs, representing the jobs that did not finish running. The action for each state is the job that we try to run on the server (needs to be a part of the state's subset). There is a $1 - \mu_i$ probability to stay in the same state (if the run was unsuccessful) and a μ_i probability to move to a new state (which is the same as the previous state but without the job that we just finished). The cost at time t is the sum of the costs of the left jobs at time t (which are the jobs in s_t)
The bellman equation for the problem:

$$V(s) = \min_{a \in A} \left\{ C(s) + \sum_{s' \in S} p\left(s'|s,a\right) V\left(s'\right) \right\} = \min_{a \in s} \left\{ C(s) + \mu_a \cdot V(s/a) + (1 - \mu_a) \cdot V(s) \right\} = C(s) + \min_{a \in s} \left\{ \mu_a \cdot V(s/a) + (1 - \mu_a) \cdot V(s) \right\}$$

Thus, the optimal policy π^* holds:

$$\pi^*(s) \in \arg\min_{a \in A} \left\{ \mu_a \cdot V(s/a) + (1 - \mu_a) \cdot V(s) \right\}$$

2.2

We will calculate the value function for the policy π, which is choosing i as: $i^* = \arg\max_i c_i \mu_i$. In order to do so, we will sort (descending order) the states with respect to $\mu_i \cdot c_i$ and denote this new order as $n_{i\,i=1}^{n}$.

$$V^\pi(s) = C(s) + \left\{ \mu_a \cdot V^\pi(s/a) + (1 - \mu_a) \cdot V^\pi(s) \right\}_{a=\pi(s)} = C(s) + \mu_{n_1} \cdot V^\pi(s/n_1) + (1 - \mu_{n_1}) \cdot V^\pi(s)$$

Thus:

$$\mu_{n_1} V^\pi(s) = C(s) + \mu_{n_1} \cdot V^\pi(s/n_1)$$
$$V^\pi(s) = \frac{C(s)}{\mu_{n_1}} + V^\pi(s/n_1)$$

Due to the fact that state n_2 is with maximal $\mu_i \cdot c_i$ in the set of s/n_1:

$$V^\pi(s/n_1) = \frac{C(s/n_1)}{\mu_{n_2}} + V^\pi(s/\left(n_1 \ \& \ n_2\right))$$

Thus:

$$V^\pi(s) = \frac{C(s)}{\mu_{n_1}} + V^\pi(s/n_1) = \frac{C(s)}{\mu_{n_1}} + \frac{C(s/n_1)}{\mu_{n_2}} + V^\pi(s/\left(n_1 \ \& \ n_2\right))$$
$$= \frac{C(s/n_1) + C(n_1)}{\mu_{n_1}} + \frac{C(s/n_1)}{\mu_{n_2}} + V^\pi(s/\left(n_1 \ \& \ n_2\right))$$

We can continue in this manner to ac achieve:

$$V^\pi(s) = \left[\frac{C(n_1)}{\mu_{n_1}} + \left(\frac{C(n_2)}{\mu_{n_2}} + \frac{C(n_2))}{\mu_{n_1}} \right) + \cdots + V^\pi(\phi) \right]$$

And we know that $V^{\pi}(\phi) = 0$ (because there are no jobs left). So:

$$V^{\pi}(s) = \left[\frac{C(n_1)}{\mu_{n_1}} + \left(\frac{C(n_2)}{\mu_{n_2}} + \frac{C(n_2))}{\mu_{n_1}}\right) + \cdots + C(n_n)\left(\frac{1}{\mu_{n_1}} + \frac{1}{\mu_{n_2}} + \ldots \frac{1}{\mu_{n_n}}\right)\right]$$

Now we can check if this value function satisfies the bellman equation:

$$V(s) = C(s) + \min_{a \in s}\{\mu_a \cdot V(s/a) + (1-\mu_a) \cdot V(s)\}$$
$$0 = C(s) + \min_{a \in s}\{\mu_a \cdot V(s/a) + -\mu_a \cdot V(s)\}$$

We saw that:

$$V^{\pi}(s) = \frac{C(s)}{\mu_a} + V^{\pi}(s/a)$$
$$V^{\pi}(s/a) = V^{\pi}(s) - \frac{C(s)}{\mu_a}$$

Thus, plugging our value function in the bellman equation will result in:

$$0 = C(s) + \min_{a \in s}\left\{\mu_a\left(V^{\pi}(s) - \frac{C(s)}{\mu_a}\right) + -\mu_a \cdot V(s)\right\} = C(s) + \min_{a \in s}\left\{-\mu_a \frac{C(s)}{\mu_a}\right\}$$
$$= C(s) + \min_{a \in s}\{-C(s)\} = 0$$

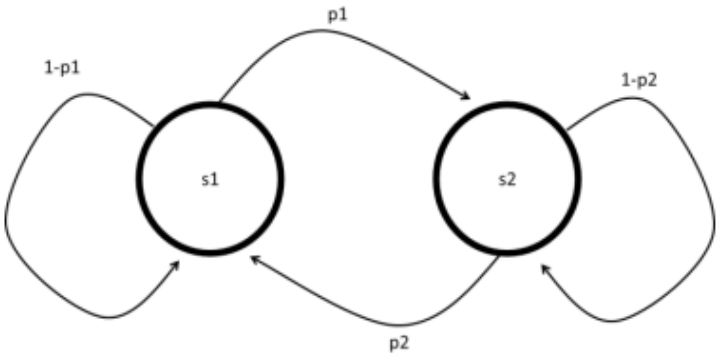
We reached a true statement, thus the value function for the suggested policy satisfies the bellman equation and the suggested policy is optimal.

3 DP operator not contracting in Euclidean norm

The DP operator that we saw in class:

$$\left(T^{\pi}(J)\right)(s) = r(s, \pi(s)) + \gamma \sum_{s' \in s} p\left(s'|s, \pi(s)\right) J\left(s'\right)$$

We proved that: T^{π} is a γ -contraction operator with respect to the max-norm, namely $\|T^{\pi}\left(V_1\right) - T^{\pi}\left(V_2\right)\|_{\infty} \leq \gamma \| V_1 - V_2 \|_{\infty}$ for all $V_1, V_2 \in R^{|S|}$
Consider the following MDP:



With rewards and transition matrix which are independent on the policy.
Denote:

$$V_1 = \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix}, V_2 = \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix}$$

Thus, applying the DP operator will result in:

$$T^{\pi}\left(V_1\right) = \begin{pmatrix} r(1) + \gamma \cdot [(1-p1) \cdot v_{11} + p1 \cdot v_{12}] \\ r(2) + \gamma \cdot [(1-p2) \cdot v_{12} + p2 \cdot v_{11}] \end{pmatrix}, T^{\pi}\left(V_2\right) = \begin{pmatrix} r(1) + \gamma \cdot [(1-p1) \cdot v_{21} + p1 \cdot v_{22}] \\ r(2) + \gamma \cdot [(1-p2) \cdot v_{22} + p2 \cdot v_{21}] \end{pmatrix}$$

Thus:

$$\left\|T^{\pi}\left(V_1\right)-T^{\pi}\left(V_2\right)\right\|_2^2=\left\|\left(\begin{array}{l} r(1)+\gamma\cdot[(1-p1)\cdot v_{11}+p1\cdot v_{12}] \\ r(2)+\gamma\cdot[(1-p2)\cdot v_{12}+p2\cdot v_{11}] \end{array}\right)-\left(\begin{array}{l} r(1)+\gamma\cdot[(1-p1)\cdot v_{21}+p1\cdot v_{22}] \\ r(2)+\gamma\cdot[(1-p2)\cdot v_{22}+p2\cdot v_{21}] \end{array}\right)\right\|_2^2$$
$$=\gamma^2\left\|\cdot\left(\begin{array}{l} [(1-p1)\cdot(v_{11}-v_{21})+p1\cdot(v_{12}-v_{22})] \\ [(1-p2)\cdot(v_{12}-v_{22})+p2\cdot(v_{11}-v_{21})] \end{array}\right)\right\|_2^2$$

And we can calculate the euclidean distance between $V_!$ and V_2:

$$\left\|V_1-V_2\right\|_2^2=\left(v_{11}-v_{21}\right)^2+\left(v_{12}-v_{22}\right)^2$$

For instance, if we will take $v_{12}=v_{21}=v_{22}=0$, we get:

$$\left\|T^{\pi}\left(V_1\right)-T^{\pi}\left(V_2\right)\right\|_2^2=\gamma^2\cdot\left((1-p1)^2\cdot v_{11}^2+p2^2\cdot v_{11}^2\right)$$

And:

$$\left\|V_1-V_2\right\|_2^2=v_{11}^2$$

Now, we will see when is the DP operator is not contracting under the euclidean norm. We will demand that:

$$\gamma^2\cdot\left((1-p1)^2\cdot v_{11}^2+p2^2\cdot v_{11}^2\right)=\left\|T^{\pi}\left(V_1\right)-T^{\pi}\left(V_2\right)\right\|_2^2>\left\|V_1-V_2\right\|_2^2=v_{11}^2$$

Thus:

$$\gamma^2\cdot\left((1-p1)^2+p2^2\right)>1$$

So we can see that for small p1 and big p2, the inequality will hold, for example, $\gamma^2=0.9, p1=0.1, p2=0.9$:

$$0.9\cdot(0.9^2+0.9^2)=1.458>1$$

In conclusion, we have found a setup that the DP operator does not contract the distance between V1 and V2, thus it is not a contraction operator under the euclidean norm.

4 Stochastic Shortest Path

4.1

First, for the terminal state, we already reached termination, so:

$$v^{\pi}(0)=v^*(0)=0$$

Thus, for any other state ($s>0$):

$$v^{\pi}(s)=\left\{c(s,a)+\gamma\sum_{s'\in S}p(s'|s,a)\cdot v^{\pi}(s')\right\}_{a=\pi(s)}=\left\{c(s,a)+\sum_{s'\in S/0}p(s'|s,a)\cdot v^{\pi}(s')\right\}_{a=\pi(s)}$$

And the optimal value function:

$$v^*(s)=\min_{a\in A}\left\{c(s,a)+\gamma\sum_{s'\in S}p(s'|s,a)\cdot v^*(s')\right\}=\min_{a\in A}\left\{c(s,a)+\sum_{s'\in S/0}p(s'|s,a)\cdot v^*(s')\right\}$$

4.2

As we saw in class:

$$\left(T^{\pi}(V)\right)(s)=c(s,\pi(s))+\gamma\sum_{s'\in S}p\left(s'|s,\pi(s)\right)V\left(s'\right)$$

$$\left(T^*(V)\right)(s)=\min_{a\in A}\left\{c(s,a)+\gamma\sum_{s'\in S}p\left(s'|s,a\right)V\left(s'\right)\right\}$$

And in our case:

$$(T^{\pi}(V))(s) = c(s,\pi(s)) + \sum_{s' \in S/0} p(s'|s,\pi(s))\, V(s')$$

$$(T^{*}(V))(s) = \min_{a \in A} \left\{ c(s,a) + \sum_{s' \in S/0} p(s'|s,a)\, V(s') \right\}$$

4.3

Let's assume that J^* is finite and it holds $T^* \cdot J^* = J^*$, but the there is a non-proper stationary policy π. Thus:

$$(T^{*}(J^{*}))(s) = \min_{a \in A} \left\{ c(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)\, J^{*}(s') \right\}$$

We know that there is a probability of zero to reach the terminal state and the costs of the rest of the states are greater than zero. Thus, we can see that we have an infinite sum of bounded from below values (bounded by the minimum cost), this sum will reach infinity, and we reach a contradiction to the fact that J^* is finite.

4.4

In our case the proof that the DP operator is a contraction operator cannot follow from the general proof we saw in class due to the fact that $\gamma = 1$. In our case the contraction property will result from the fact that all of the stationary policies are proper and after reaching the terminal state we will not have any further cost.
We will look at a new SSp: each state, beside the terminal one, has a cost of -1.
Define $\hat{J}(s)$ as the optimal value function from every state is the new SSP and $\xi(s) = -\hat{J}(s)$.
This optimal value function satisfies the bellman equation:

$$\hat{J}(s) = \min_{a \in A} \left\{ c(s,a) + \sum_{s' \in S/0} p(s'|s,a) \cdot \hat{J}(s') \right\}$$

For any state except the terminal state (which will hold $\hat{J}(s) = \xi(s) = 0$):

$$\hat{J}(s) = \min_{a \in A} \left\{ -1 + \sum_{s' \in S/0} p(s'|s,a) \cdot \hat{J}(s') \right\} \leq -1$$

Thus:

$$\xi(s) \geq 1$$

Writing again the bellman equation and plugging $\hat{J}(s) = -\xi(s)$:

$$-\xi(s) = \min_{a \in A} \left\{ -1 - \sum_{s' \in S/0} p(s'|s,a) \cdot \xi(s') \right\}$$

Thus for any stationary $\pi(s)$:

$$-\xi(s) \leq -1 - \sum_{s' \in S/0} p(s'|s,\pi(s)) \cdot \xi(s')$$

$$\sum_{s' \in S/0} p(s'|s,\pi(s)) \cdot \xi(s') = \sum_{s' \in S} p(s'|s,\pi(s)) \cdot \xi(s') \leq \xi(s) - 1$$

Now we will show the second part of the section. We can see that the wanted property is equivalent to:

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

Which obviously holds. Further, we know that $\xi \geq 1$, so:

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

Here we assumed that $\xi \neq 1$, in the case that it is, we will get $\beta = 0 < 1$

4.5

We will gather all of the results from the previous sub-section in order to prove that the DP operator is a contracting operator for the SSP problem.
From the results of the previous subsection we can draw the conclusion that:

$$\sum_{s' \in S} p(s'|s,\pi(s)) \cdot \xi(s') \leq \xi(s) - 1 \leq \max_{s'} \frac{\xi\left(s'\right) - 1}{\xi\left(s'\right)} \cdot \xi\left(s\right)$$

Now we will show the contraction property, under the weighted maximum norm:
Let $s \in S$:

$$\begin{aligned}
|T_\pi J_1(s) - T_\pi J_2(s)| &= \left| c(s,\pi(s)) + \sum_{s' \in S} p\left(s'|s,\pi(s)\right) J_1\left(s'\right) - c(s,\pi(s)) - \sum_{s' \in S} p\left(s'|s,\pi(s)\right) J_2\left(s'\right) \right| \\
&= \left| \sum_{s' \in S} p\left(s'|s,\pi(s)\right) J_1\left(s'\right) - \sum_{s' \in S} p\left(s'|s,\pi(s)\right) J_2\left(s'\right) \right| \\
&= \left| \sum_{s' \in S} p\left(s'|s,\pi(s)\right) \left(J_1\left(s'\right) - J_2\left(s'\right)\right) \right| \\
&\leq \sum_{s' \in S} p\left(s'|s,\pi(s)\right) \left|J_1\left(s'\right) - J_2\left(s'\right)\right| \\
&= \sum_{s' \in S} p\left(s'|s,\pi(s)\right) \frac{\xi(s')}{\xi(s')} \left|J_1\left(s'\right) - J_2\left(s'\right)\right| \\
&\leq \sum_{s' \in S} p\left(s'|s,\pi(s)\right) \cdot \xi(s') \max_{s'' \in S} \frac{\left|J_1\left(s''\right) - J_2\left(s''\right)\right|}{\xi(s')} \\
&= \max_{s'' \in S} \frac{\left|J_1\left(s''\right) - J_2\left(s''\right)\right|}{\xi(s'')} \sum_{s' \in S} p\left(s'|s,\pi(s)\right) \cdot \xi(s') \\
&\leq \beta ||J_1 - J_2||_\xi \cdot \xi(s)
\end{aligned}$$

Thus:

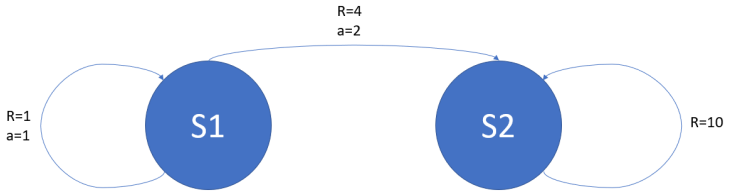
$$\frac{|T_\pi J_1(s) - T_\pi J_2(s)|}{\xi(s)} \leq \beta ||J_1 - J_2||_\xi$$

This is true for every $s \in S$, so:

$$\max_{s \in S} \frac{|T_\pi J_1(s) - T_\pi J_2(s)|}{\xi(s)} = ||T_\pi J_1 - T_\pi J_2||_\xi \leq \beta ||J_1 - J_2||_\xi$$

5 Value function of the greedy policy

We will show that π_i are not necessary improving, using an example. Notice, that this is not in contradiction to the theorem about the monotonic descent of the value functions in the value iteration process because in the value iteration, the value function are arbitrary vectors and does not represent a real policy's value function (in particular, not the greedy policy).
Consider the following MDP with deterministic dynamics:



With $\gamma = 0.9999$ (we will take γ as equal to one in our calculation, it will not make any different because we are only looking at two steps of the algorithm).Assume that:

$$V^0 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$$

Thus:

$$\pi^0 = \arg\max_{a \in 1,2} \left[r(s,a) + \gamma V^0(s') \right] = \arg\max_{a \in 1,2} \left[1+5, 4+1 \right] = 1$$

$$V^{\pi^0} = \begin{pmatrix} 15 \\ 10 \end{pmatrix}$$

$$V^1(s) = \max_{a \in 1,2} \left[r(s,a) + \gamma V^0(s') \right] = \begin{pmatrix} 6 \\ 11 \end{pmatrix}$$

The second iteration:

$$\pi^1 = \arg\max_{a \in 1,2} \left[r(s,a) + \gamma V^1(s') \right] = \arg\max_{a \in 1,2} \left[1+6, 4+11 \right] = 2$$

$$V^{\pi^1} = \begin{pmatrix} 14 \\ 10 \end{pmatrix}$$

We saw an example of a setup that holds:

$$V^{\pi^1}(s=1) < V^{\pi^0}(s=0)$$

So the sequence π_i is not necessarily improving.