EN.520.637: Foundations of Reinforcement Learning Homework 3

Ha Manh Bui (CS Department)
hbui13@jhu.edu

Fall 2023

1 Problem 1

You are in a casino! You start with \$10 and will play until you lose it all or as soon as you reach \$30. You can choose to play two slot machines: 1) slot machine A costs \$10 to play and will return \$20 with probability 0.1 and \$0 otherwise; and 2) slot machine B costs \$20 to play and will return \$30 with probability 0.4 and \$0 otherwise. Until you are done, you will choose to play machine A or machine B in each turn.

(a) Compute the expected reward you gain from playing machine A and B one time, respectively.
Let A be the action random variable we play, $A = a$ and $A = b$ represent playing either machine A or B. Let R be the reward random variable representing the money we gain from that play, then we have the expected reward if we play machine A one time is

$$\mathbb{E}[R|A = a] = \sum_r rP(R = r|A = a) = (20 - 10) * 0.1 + (-10) * 0.9 = -8,$$

and the expected reward if we play machine B one time is

$$\mathbb{E}[R|A = b] = \sum_r rP(R = r|A = b) = (30 - 20) * 0.4 + (-20) * 0.6 = -8.$$

(b) We can model this as an MDP. Let the state be the current money you have. Let the action be playing either machine A or B once, and the reward be the money you gain from that play. Write down the state space and the action space. Then draw a diagram for this MDP.

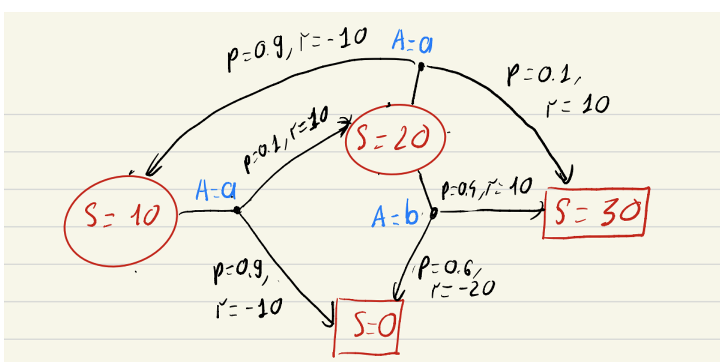

Figure 1: The diagram corresponds to the MDP in Question 1 (b).

Because we start with \$10 and we will play until we lose or reach \$30, we have the following possible action

1. Play machine A and lose, then have \$0 and end.

2. Play machine A and win, then have \$20.

   (a) Continue to play machine A and win, then have \$30 and end.

   (b) Continue to play machine A and lose, then have \$10. This is back to (1).

   (c) Continue to play machine B and lose, then have \$0 and end.

   (d) Continue to play machine B and win, then have \$30 and end.

Let S be the random variable representing the state of current money, then we have the corresponding state space $\mathcal{S} = \{0, 10, 20, 30\}$. This yields the corresponding action space as follows

$$\begin{cases} \mathcal{A}(S=0) = \{\emptyset\} \\ \mathcal{A}(S=10) = \{a\} \\ \mathcal{A}(S=20) = \{a, b\} \\ \mathcal{A}(S=30) = \{\emptyset\}. \end{cases}$$

The diagram for this MDP is in Fig. 1.

(c) Explain why all possible policies $\pi(a|s)$ for this MDP can be uniquely defined by $\beta \in [0,1]$, where β is the probability of choosing slot machine A when you have \$20. Now consider such a policy π_β. Compute v_{π_β} for all the non-terminal states. What is the optimal policy?

Because we only can select machine A if we have \$10, the possible policy at this stage is only to play machine A. Therefore, if β is the probability of choosing slot machine A when we have \$20, we have all possible policies $\pi(a|s)$ as follows

$$\begin{cases} \pi(A=a|S=10) = 1, \\ \pi(A=a|S=20) = \beta, \\ \pi(A=b|S=20) = 1-\beta. \end{cases}$$

As a consequence, all possible policies $\pi(a|s)$ for this MDP can be uniquely defined by $\beta \in [0,1]$.
Consider policy π_β, we have the corresponding expected return at stage $S = 20$ is

$$\begin{aligned} v_{\pi_\beta}(S=20) &= \mathbb{E}_{\pi_\beta}\left[q_{\pi_\beta}(S,A)|S=20\right] = \beta q_{\pi_\beta}(S=20, A=a) + (1-\beta) q_{\pi_\beta}(S=20, A=b) \\ &= \beta\left[\mathbb{E}[R|S=20, A=a] + \sum_{s' \in \{10,30\}} p(s'|S=20, A=a) v_{\pi_\beta}(s')\right] \\ &+ (1-\beta)\left[\mathbb{E}[R|S=20, A=b] + \sum_{s' \in \{0,30\}} p(s'|S=20, A=b) v_{\pi_\beta}(s')\right] \\ &= \beta\left[(20-10)*0.1 + (-10)*0.9 + 0.9 * v_{\pi_\beta}(S=10) + 0.1*0\right] \\ &+ (1-\beta)\left[(30-20)*0.4 + (-20)*0.6 + 0.6*0 + 0.4*0\right] \\ &= \beta\left[-8 + 0.9 * v_{\pi_\beta}(S=10)\right] + (1-\beta)*(-8), \end{aligned}$$

and the expected return at stage $S = 10$ is

$$\begin{aligned} v_{\pi_\beta}(S=10) &= \mathbb{E}_{\pi_\beta}\left[q_{\pi_\beta}(S,A)|S=10\right] = q_{\pi_\beta}(S=10, A=a) \\ &= \mathbb{E}[R|S=10, A=a] + \sum_{s' \in \{0,20\}} p(s'|S=10, A=a) v_{\pi_\beta}(s') \\ &= (20-10)*0.1 + (-10)*0.9 + 0.9*0 + 0.1 * v_{\pi_\beta}(S=20) = -8 + 0.1 * v_{\pi_\beta}(S=20) \\ &= -8 + 0.1 * \left\{\beta\left[-8 + 0.9 * v_{\pi_\beta}(S=10)\right] + (1-\beta)*(-8)\right\}. \end{aligned}$$

This yields

$$v_{\pi_\beta}(S=10) = \frac{-8.8}{1-0.09\beta} \quad \text{and} \quad v_{\pi_\beta}(S=20) = -8 + 0.9\beta * \frac{-8.8}{1-0.09\beta},$$

hence, when $\beta = 0$, then the maximum of $v_{\pi_\beta}(S=10) = -8.8$ and $v_{\pi_\beta}(S=20) = -8$. As a consequence, the corresponding optimal policy is $\pi^*(A=a|S=10) = 1$ and $\pi^*(A=b|S=20) = 1$.

(d) If we now assume that "slot machine B costs \$20 to play and will return \$30 with probability $0 < \eta < 1$ and \$0 otherwise". What value of η ensures that any policy is an optimal policy?
In order for any policy to be optimal policy, the $\pi^*(A=a|S=20) = \pi^*(A=b|S=20)$, this equivalent to $q_\pi(S=20, A=a) = q_\pi(S=20, A=b) = v_\pi(S=20)$. When the probability to get \$30 with machine B is η, $0 < \eta < 1$, then the following Equation must holds

$$
\begin{aligned}
& q_\pi(S=20, A=a) = q_\pi(S=20, A=b) = v_\pi(S=20) \\
& \Leftrightarrow [(20-10)*0.1 + (-10)*0.9 + 0.9*v_\pi(S=10)] = [(30-20)*\eta + (-20)*(1-\eta)] = v_\pi(S=20) \\
& \Leftrightarrow -8 + 0.9*[-8 + 0.1*v_\pi(S=20)] = 30\eta - 20 = v_\pi(S=20).
\end{aligned}
$$

This yields $v_\pi(S=20) = \frac{-1520}{91}$ and $\eta = \frac{10}{91}$.

2 Problem 2

(a) What are the equations analogous to (1), (2), and (3), but for action-value functions instead of state-value functions? Start with $q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$, show all your derivations.

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] \qquad (1) \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right], \qquad (2)
\end{aligned}
$$

and

$$
\begin{aligned}
v_{k+1}(s) &= \mathbb{E}_\pi\left[R_{t+1} + \gamma v_k(S_{t+1})|S_t = s\right] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_k(s')\right]. \qquad (3)
\end{aligned}
$$

First, recall the return function with Infinite Horizon T

$$
G_t = \sum_{k=t}^{\infty} \gamma^{k-t} R_{k+1} = R_{k+1} + \sum_{k=t+1}^{\infty} \gamma^{k-t} R_{k+1} = R_{k+1} + \gamma \underbrace{\sum_{k=t+1}^{\infty} \gamma^{k-(t+1)} R_{k+1}}_{G_{t+1}}.
$$

Then, we have the q function as follows

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s, A_t = a] \\
&= \mathbb{E}_\pi\left[\mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s, A_t = a, S_{t+1}, A_{t+1}]|S_t = s, A_t = a\right] \\
&= \mathbb{E}_\pi\left[R_{t+1} + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1}, A_{t+1}]|S_t = s, A_t = a\right] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \\
&= \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s',a')],
\end{aligned}
$$

and

$$
\begin{aligned}
q_{k+1}(s,a) &= \mathbb{E}_\pi[R_{t+1} + \gamma q_k(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \\
&= \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a'} \pi(a'|s') q_k(s',a')].
\end{aligned}
$$

(b) Equation (3) can be viewed as $v_{k+1} = \mathcal{T}_\pi(v_k)$, where $\mathcal{T}_\pi$ is an operator acting on the value function v. Analogous to this, from (a) we can define an iteration $q_{k+1} = \mathcal{T}_\pi^q(q_k)$ to compute the action value function for policy π, where $\mathcal{T}_\pi^q$ is an operator acting on the action value function q. Show $\mathcal{T}_\pi^q$ is γ-contracting

w.r.t. $||.||_\infty$, where $0 < \gamma < 1$ is the discounting factor.
First, recall $\forall q \in \mathbb{R}^{nm}$, where n is the number of states and m is the number of actions, we have

$$\begin{aligned}
q_{k+1}(s,a) &= \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a'} \pi(a'|s') q_k(s',a')] \\
&= \sum_{s',r} p(s',r|s,a) r + \gamma \sum_{s',r} p(s',r|s,a) \sum_{a'} \pi(a'|s') q_k(s',a') \\
&= \sum_{s',r} p(s',r|s,a) r + \gamma \sum_{s',a'} p(s',a'|s,a) q_k(s',a') \\
&= r(s,a) + \gamma P_\pi^q q = R_\pi^q + \gamma P_\pi^q q,
\end{aligned}$$

for some $R_\pi^q \in \mathbb{R}^{nm}$, $P_\pi^q \in \mathbb{R}^{nm \times nm}$. Therefore, $\forall q, q'$, we get

$$\begin{aligned}
||\mathcal{T}_\pi^q(q) - \mathcal{T}_\pi^q(q')||_\infty &= ||R_\pi^q + \gamma P_\pi^q q - R_\pi^q - \gamma P_\pi^q q'||_\infty \\
&= \gamma ||P_\pi^q q - P_\pi^q q'||_\infty = \gamma ||P_\pi^q (q - q')||_\infty \le \gamma ||q - q'||_\infty,
\end{aligned}$$

where $0 < \gamma < 1$. As a consequence, we obtain $\mathcal{T}_\pi^q$ is γ-contracting w.r.t. $||.||_\infty$, where $0 < \gamma < 1$ is the discounting factor.

3 Problem 3

Consider a single-server queueing system where L customers are waiting to get the service. At any time step, you can choose to serve μ customers, where $\mu \in \{0, 1, \cdots, L\}$. At each time t, after deciding to serve μ customers, there is a cost $h(\mu)$ and an additional cost $c(i)$ for having i customers remaining in the queue. The idea is that one should be able to cut down on customer waiting costs, by choosing to serve more customers at a time, so that the service is optimally traded-off.

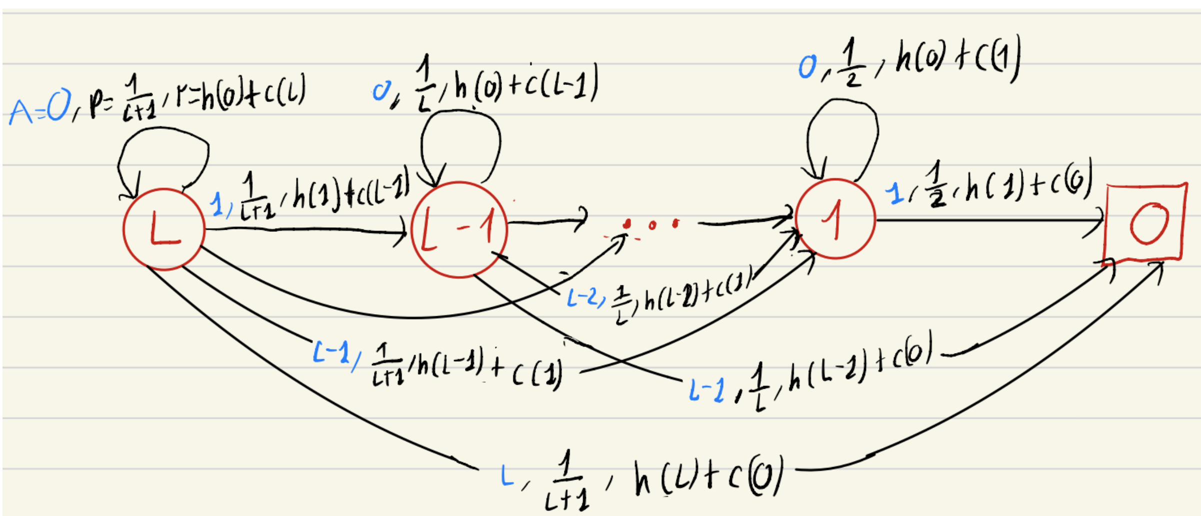

Figure 2: The transition diagram in Question 3 (b).

(a) Now formulate this problem as an infinite horizon DP problem (assume a discount factor $0 < \gamma < 1$). What is the set of all possible states $\mathcal{S}_k$ at stage k? What are the set of all possible actions $\mathcal{A}_k(s_k)$ at stage k given a particular state?
This problem can be formalized as an infinite horizon DP problem with discount factor γ where S is the state random variable representing the number of customers in the queuing system, then we have all possible states $\mathcal{S}_k$ at stage k is

$$\mathcal{S}_k = \{0, 1, \cdots, L\}.$$

Because at each stage k, we have s_k customers remaining in the queue, so the corresponding actions $\mathcal{A}_k(s_k)$ representing the choice of customers to serve are

$$\mathcal{A}_k(s_k) = \{0, \cdots, s_k\}.$$

(b) Under your DP problem, draw the transition diagram and specify the state-reward distribution

$$p(s', r|s, a) = \mathbb{P}(S_{k+1} = s', R_{k+1} = r|S_k = s, A_k = a).$$

Let R_{k+1} be the reward as the cost of action A_k at stage k. Because at each stage k, after deciding to serve μ customers, there is a cost $h(\mu)$ and an additional cost $c(i)$ for having i customers remaining in the queue, then we have the transition diagram in Fig. 2 and state-reward distribution at stage $S_k = s$, $\forall s \in \mathcal{S}$ as follows

$$\mathbb{P}(S_{k+1} = s - \mu, R_{k+1} = h(\mu) + c(s - \mu)|S_k = s, A_k = \mu) = \frac{1}{s+1}, \forall \mu \in \{0, \cdots, s\}.$$

(c) For simplicity let's assume $\mu \in \{0, 1\}$, $c(i) = ci$, and $h(0) = 0$. Compute the value functions $v_{\pi_j}(s), \forall s \in \mathcal{S}$, $j \in \{1, 2\}$ for the policies

1. π_1: always serve when there are customers in line ($\mu = 1$) and don't serve when there are no customers ($\mu = 0$).
2. π_2: always refuse to serve (service rate $\mu = 0$).

Consider the policy π_1, we have the corresponding value function at stage $S_k = s$, $\forall s \in \mathcal{S} \setminus \{0\}$ (i.e., there are customers in the line) is

$$\begin{aligned}
v_{\pi_1}(s) &= \sum_a \pi_1(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_{\pi_1}(s')\right] \\
&= p(S' = s - 1, R = h(1) + c(s - 1)|S = s, A = 1) \left[h(1) + c(s - 1) + \gamma v_{\pi_1}(s - 1)\right] \\
&= h(1) + c(s - 1) + \gamma v_{\pi_1}(s - 1),
\end{aligned}$$

and at stage $S_k = 0$ (i.e., there are no customers) is

$$\begin{aligned}
v_{\pi_1}(0) &= \sum_a \pi_1(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_{\pi_1}(s')\right] \\
&= p(S' = 0, R = h(0) + c(0)|S = 0, A = 0) \left[h(0) + c(0) + \gamma v_{\pi_1}(0)\right] \\
&= h(0) + c(0) + \gamma v_{\pi_1}(0) = \gamma v_{\pi_1}(0).
\end{aligned}$$

Consider the policy π_2, we have the corresponding value function at stage $S_k = s$, $\forall s \in \mathcal{S}$ is

$$\begin{aligned}
v_{\pi_2}(s) &= \sum_a \pi_2(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_{\pi_2}(s')\right] \\
&= p(S' = s, R = h(0) + c(s)|S = s, A = 0) \left[h(0) + c(s) + \gamma v_{\pi_2}(s)\right] \\
&= h(0) + c(s) + \gamma v_{\pi_2}(s) \\
&= c(s) + \gamma v_{\pi_2}(s).
\end{aligned}$$

(d) Show that if

$$\frac{c}{1-\gamma} > h(1)$$

holds, then policy π_1 dominates the policy π_2, i.e., $v_{\pi_1}(s) \geq v_{\pi_2}(s)$, $\forall s \in \mathcal{S}$, where γ is the discounting factor.

When $s \neq 0$, we have

$$\begin{aligned}
v_{\pi_1}(s) &= h(1) + c(s - 1) + \gamma v_{\pi_1}(s - 1) \\
&= h(1) + cs - c + \gamma v_{\pi_1}(s - 1),
\end{aligned}$$

since $\frac{c}{1-\gamma} > h(1)$, i.e., $h(1) - c < h(1)\gamma$, we get

$$v_{\pi_1}(s) = h(1) + cs - c + \gamma v_{\pi_1}(s-1) < cs + \gamma\left[h(1) + v_{\pi_1}(s-1)\right] < \underbrace{cs + \gamma v_{\pi_2}(s)}_{v_{\pi_2}(s)}.$$

On the other hand, when $s = 0$, we have $v_{\pi_1}(0) = \gamma v_{\pi_1}(0)$ and $v_{\pi_2}(0) = \gamma v_{\pi_2}(0)$. As a result, we obtain

$$v_{\pi_1}(s) \le \underbrace{cs + \gamma v_{\pi_2}(s)}_{v_{\pi_2}(s)}, \forall s \in \mathcal{S},$$

combining with the fact that we want to minimize $v_\pi(s)$, then the policy π_1 dominates the policy π_2.

4 Problem 4

Equivalency between a discounted problem and one with a geometric horizon. Consider an undiscounted MDP $\mathcal{M}$ with action space $\mathcal{A}$, and state space $\mathcal{S} \cup \{z\}$ where z denotes an absorbing, terminal state i.e.,:

$$p(z|z,a) = 1 \quad \forall a \in \mathcal{A}$$
$$r(z,a) = 0 \quad \forall a \in \mathcal{A}$$

Note that transitions do not depend on time. Furthermore, assume that at each step there is a positive probability of going to the termination state:

$$P(z|s,a) = 1 - \gamma \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

The goal is to maximize the cumulative undiscounted reward:

$$\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} R_{t+1} | S_0 = s\right] \quad s \in \mathcal{S}.$$

(a) Consider starting at $t = 0$ at state s. Let T be the time until transitioning to the absorbing state z. Show that T is a geometrically distributed random variable. What is its parameter?
We have the MDP starting at $t = 0$ at state s and z is the terminal state, i.e., the MDP with the horizon T will stop when we reach state $S_T = z$. This is equivalent to the probability distribution of the number $T - 1$ of failure, i.e., $S_k \neq z, \forall k \in \{0, 1 \cdots, T-1\}$ before the first success $S_T = z$, supported on the set $\{0, 1, \cdots, T\}$. This yields T is a random variable that follows Geometric distribution.
On the other hand, at state s we have the positive probability that reaching z is

$$p(z|s,a) = 1 - \gamma, \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

this equivalent to

$$p(S_T = z) = \gamma^{T-1}(1-\gamma),$$

i.e., $T \sim \mathit{Geom}(1-\gamma)$.

(b) For this MDP, let:

$$(T_\pi v)(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\left(r(s,a) + \sum_{s' \in \mathcal{S} \cup \{z\}} P(s'|s,a)v(s')\right)$$

$$(T^* v)(s) = \max_{a \in \mathcal{A}}\left(r(s,a) + \sum_{s' \in \mathcal{S} \cup \{z\}} P(s'|s,a)v(s')\right)$$

be the Bellman operator for a fixed policy π and the optimal one, respectively. Show that these operators are equivalent to those from a *discounted* MDP $\bar{\mathcal{M}}$ with state space $\mathcal{S}$, no terminal state, and transition probabilities:

$$\bar{P}(s'|s,a) = \frac{1}{\gamma} P(s'|s,a)$$

Consider MDP $\mathcal{M}$, we have he Bellman operator for a fixed policy π of $\mathcal{M}$ is

$$(T_\pi v)(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + \sum_{s'\in\mathcal{S}\cup\{z\}} p(s'|s,a)v(s') \right) = \sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + p(z|s,a)v(z) + \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right),$$

and the Bellman operator for the optimal one of $\mathcal{M}$ is

$$(T^* v)(s) = \max_{a\in\mathcal{A}} \left(r(s,a) + \sum_{s'\in\mathcal{S}\cup\{z\}} p(s'|s,a)v(s') \right) = \max_{a\in\mathcal{A}} \left(r(s,a) + p(z|s,a)v(z) + \sum_{s'\in\mathcal{S}} P(s'|s,a)v(s') \right).$$

Since $p(z|z,a) = 1$ and $r(z,a) = 0$, $\forall a \in \mathcal{A}$, we get $v(z) = 0$. Therefore, we obtain,

$$\begin{aligned}
(T_\pi v)(s) &= \sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + p(z|s,a)v(z) + \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + (1-\gamma)*0 + \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right), \qquad (4)
\end{aligned}$$

and

$$\begin{aligned}
(T^* v)(s) &= \max_{a\in\mathcal{A}} \left(r(s,a) + \sum_{s'\in\mathcal{S}\cup\{z\}} p(s'|s,a)v(s') \right) = \max_{a\in\mathcal{A}} \left(r(s,a) + (1-\gamma)*0 + \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right) \\
&= \max_{a\in\mathcal{A}} \left(r(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right). \qquad (5)
\end{aligned}$$

Consider a discounted MDP $\bar{\mathcal{M}}$ with state $\mathcal{S}$, no terminal state, and transition probability $\bar{p}(s'|s,a) = \frac{1}{\gamma}p(s'|s,a)$, by definition, we have the Bellman operator for a fixed policy π of $\bar{\mathcal{M}}$ is

$$\begin{aligned}
(\bar{T_\pi} v)(s) &= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s',r} \bar{p}(s',r|s,a) \left[r + \gamma v(s') \right] \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + \sum_{s'\in\mathcal{S}} \frac{1}{\gamma} p(s'|s,a)\gamma v(s') \right) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \left(r(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right) \\
&= \text{Eq. 4 (so the Bellman operator for a fixed policy } \pi \text{ between } \mathcal{M} \text{ and } \bar{\mathcal{M}} \text{ are similar)},
\end{aligned}$$

and the Bellman operator for the optimal one of $\bar{\mathcal{M}}$ is

$$\begin{aligned}
(\bar{T^*} v)(s) &= \max_{a\in\mathcal{A}} \left(\sum_{s',r} \bar{p}(s',r|s,a) \left[r + \gamma v(s') \right] \right) \\
&= \max_{a\in\mathcal{A}} \left(r(s,a) + \sum_{s'\in\mathcal{S}} \frac{1}{\gamma} p(s'|s,a)\gamma v(s') \right) \\
&= \max_{a\in\mathcal{A}} \left(r(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)v(s') \right) \\
&= \text{Eq. 5 (so the Bellman operator for the optimal one between } \mathcal{M} \text{ and } \bar{\mathcal{M}} \text{ are similar)}.
\end{aligned}$$