Introduction to Machine Learning

Fall 2019

University of Science and Technology of China

Lecturer: Jie Wang | Homework 7

Posted: Dec. 9, 2019 | Due: Dec. 23, 2019

Name: San Zhang | ID: PBXXXXXXXX

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Properties of expectation and variance 10pts

Let X, Y, and Z be random variables. Show that the following results hold.

1. (5pts) The tower property holds, i.e.,

$$\mathrm{E}[X|Y] = \mathrm{E}[\mathrm{E}[X|Y,Z]|Y].$$

2. (5pts) The variance decomposition formula holds, i.e.,

$$\mathrm{Var}[X] = \mathrm{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathrm{E}[X|Y]].$$

Hint: if you do not know measure theory well, you can assume that X, Y, and Z are continuous random variables.

Solution: 1. Let $f_{X|Y,Z}(x)$ be the conditional probability density function of the random variable X, conditioned on the particular event $Y \wedge Z$ with $P(Y \cap Z) > 0$. Note that $\mathrm{E}[X|Y,Z] = \int_{-\infty}^{\infty} x f_{X|Y,Z}(x)dx$. We have

$$\begin{aligned}
\mathrm{RHS} = \mathrm{E}[\mathrm{E}[X|Y,Z]|Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y,Z}(x)dx f_{Z|Y}(z)dz \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X|Y,Z}(x) f_{Z|Y}(z)dzdx \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} \frac{f_{X,Y,Z}(x,y,z)}{f_{Y,Z}(y,z)} \cdot \frac{f_{Y,Z}(y,z)}{f_Y(y)}dzdx \\
&= \int_{-\infty}^{\infty} x \frac{f_{X,Y}(x,y)}{f_Y(y)}dx \\
&= \int_{-\infty}^{\infty} x f_{X|Y}(x)dx \\
&= \mathrm{E}[X|Y] = \mathrm{LHS},
\end{aligned}$$

which completes the proof.

2. The tower property implies $\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X|Z]]$, i.e., the law of iterated expectation.

Then by the property of variance, we have

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\
&= \mathrm{E}[\mathrm{E}[X^2|Y]] - (\mathrm{E}[\mathrm{E}[X|Y]])^2 \\
&= \mathrm{E}[\mathrm{E}[X^2|Y] - (\mathrm{E}[X|Y])^2 + (\mathrm{E}[X|Y])^2] - (\mathrm{E}[\mathrm{E}[X|Y]])^2 \\
&= \mathrm{E}[\mathrm{Var}[X|Y]] + \mathrm{E}[(\mathrm{E}[X|Y])^2] - (\mathrm{E}[\mathrm{E}[X|Y]])^2 \\
&= \mathrm{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathrm{E}[X|Y]],
\end{aligned}$$

which completes the proof.

■

Exercise 2: Properties of transition matrix 30pts

A matrix is nonnegative (positive) if all its entries are nonnegative (positive). A right (left) stochastic matrix is a square nonnegative matrix with each row (column) adds up to one. Without loss of generality, we study the right stochastic matrix in this exercise. Suppose that $T \in \mathbb{R}^{n\times n}$ is a right stochastic matrix.

1. (5pts) Show that T has a eigenvalue 1.
2. (10pts) Let λ be one of T's eigenvalues. Show that $|\lambda| \leq 1$.
3. (5pts) Show that $I - \gamma T$ is invertible, where $I \in \mathbb{R}^{n\times n}$ is the identity matrix and $\gamma \in (0, 1)$.
4. (10pts) We now show that $(I - \gamma T)^{-1} = \sum_{i=0}^{\infty}(\gamma T)^i$.
   (a) For $\mathbf{x} \in \mathbb{R}^n$, we define the infinity norm by
   $$\|\mathbf{x}\|_\infty = \max_i |x_i|.$$
   The induced norm of matrix $M \in \mathbb{R}^{m\times n}$ is
   $$\|M\|_\infty = \max_{\|\mathbf{x}\|_\infty \leq 1} \|M\mathbf{x}\|_\infty.$$
      i. Show that $\|M\|_\infty = \max_i \sum_{j=1}^n |m_{i,j}|$.
      ii. Show that $\|cM\|_\infty = |c|\|M\|_\infty$ for any $c \in \mathbb{R}$.
      iii. Show that $\|AB\|_\infty \leq \|A\|_\infty\|B\|_\infty$ holds for any $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$.
   (b) Show that the sequence $I, \gamma T, \sum_{i=0}^2(\gamma T)^i, \ldots, \sum_{i=0}^n(\gamma T)^i, \ldots$ is Cauchy. (Hint: a matrix sequence $\{A_p\}$ is Cauchy in $(\mathbb{R}^{m\times n}, \|\cdot\|_\infty)$, if given any $\epsilon > 0$, there is an integer $N \geq 1$ such that $\|A_p - A_q\|_\infty < \epsilon$ whenever $p, q \geq N$.)
   (c) Combining the result in the last part and the fact that $\mathbb{R}^{n\times n}$ is complete, we can conclude that $\sum_{i=0}^{\infty}(\gamma T)^i$ converges to a matrix which we denote by L. Show that $(I - \gamma T)^{-1} = L$. (Hint: you need to show that $(I - \gamma T)L = \lim_{n\to\infty}\sum_{i=0}^n(I - \gamma T)(\gamma T)^i$.)

Solution:

1. Let $\mathbf{x} = \begin{bmatrix}1, & 1, & \cdots, & 1\end{bmatrix}^T \in \mathbb{R}^n$. Then $T\mathbf{x} = \mathbf{x} \Rightarrow \lambda = 1$ is an eigenvalue of T.
2. We show it by contradiction. Assume $\exists\, |\lambda| > 1$, s.t. $T\mathbf{x} = \lambda\mathbf{x}$. Let $|x_k| = \max_i |x_i|$.Note that $|x_k| > 0$ as $\mathbf{x} \neq 0$. Since T is a right stochastic matrix, every entry of $\lambda\mathbf{x}$ is a convex conmbination of the entries of $\mathbf{x}$. Let $t_{i,j}$ be a entry in the i^{th} row, j^{th} column of T. We have
   $$|\lambda x_i| = |t_{i,1}x_1 + t_{i,2}x_2 + \cdots + t_{i,n}x_n| \leq |x_k|, \quad \text{where } t_{i,j} \in [0, 1]\,, \forall\, i, j \in [n].$$
   Letting $i = k$, we get $|\lambda||x_k| \leq |x_k|$, which is a contradiction to $|\lambda| > 1$.

3. Since $I - \gamma T = -\gamma(T - \frac{1}{\gamma} I)$ and $\frac{1}{\gamma} > 1$, we conclude that $\frac{1}{\gamma}$ is not a eigenvalue of T, i.e., $\det(T - \frac{1}{\gamma} I) \neq 0$.

   Hence

$$\det(I - \gamma T) = (-\gamma)^n \det(T - \frac{1}{\gamma} I) \neq 0,$$

   which implies that $I - \gamma T$ is invertible.

4. (a) i.

$$\begin{aligned}
\|M\|_\infty &= \max_{\|\mathbf{x}\|_\infty \leq 1} \|M\mathbf{x}\|_\infty \\
&= \max_{\|\mathbf{x}\|_\infty \leq 1} \max_i |\sum_{j=1}^n m_{i,j} x_j| \\
&= \max_i \max_{\|\mathbf{x}\|_\infty \leq 1} \sum_{j=1}^n |m_{i,j} x_j| \\
&= \max_i \sum_{j=1}^n \left(|m_{i,j}| \cdot \max_{\|\mathbf{x}\|_\infty \leq 1} |x_j| \right) \\
&= \max_i \sum_{j=1}^n |m_{i,j}|,
\end{aligned}$$

   which completes the proof.

   ii. $\|cM\|_\infty = \max_i \sum_{j=1}^n |cm_{i,j}| = |c| \max_i \sum_{j=1}^n |m_{i,j}| = |c| \|M\|_\infty$.

   iii. First,

$$\text{RHS} = \left(\max_i \sum_{j=1}^n |a_{i,j}| \right) \left(\max_s \sum_{j=1}^n |b_{s,j}| \right).$$

   Next,

$$\begin{aligned}
\text{LHS} &= \max_i \sum_{j=1}^n |\sum_{k=1}^n a_{i,k} b_{k,j}| \\
&\leq \max_i \sum_{j=1}^n \sum_{k=1}^n |a_{i,k} b_{k,j}| \\
&= \max_i \sum_{k=1}^n \left(|a_{i,k}| \sum_{j=1}^n |b_{k,j}| \right) \\
&\leq \left(\max_i \sum_{k=1}^n |a_{i,k}| \right) \left(\max_s \sum_{j=1}^n |b_{s,j}| \right) = \text{RHS}.
\end{aligned}$$

   Hence $\|AB\|_\infty \leq \|A\|_\infty \|B\|_\infty$.

(b) From the properties of the right stochastic matrix, we know $\|T\|_\infty = 1$ and $\|aT_1 + bT_2\|_\infty = |a| + |b|$, where $T_1, T_2 \in \mathbb{R}^{n\times n}$ are right stochastic matrices.

Next we show T^m is still a right stochastic matrix for any $m \in \mathbb{N}$ once T is a right stochastic matrix:

For two right stochastic matrices $A, B \in \mathbb{R}^{n\times n}$, we have $(AB)_{i,j} = \sum_{k=1}^n a_{i,k}b_{k,j} \geq 0$. Thus the i^{th} row adds up to

$$\sum_{j=1}^n \sum_{k=1}^n a_{i,k}b_{k,j} = \sum_{k=1}^n \left(a_{i,k} \sum_{j=1}^n b_{k,j} \right) = \sum_{k=1}^n a_{i,k} = 1, \quad i = 1, \cdots, n,$$

which implies that AB is a right stochastic matrix. It follows that T^m, $m \in \mathbb{N}$ is a right stochastic matrix.

For all $\varepsilon > 0$, setting $N = \lceil \log_\gamma(1-\gamma)\varepsilon \rceil$, for the sequence $\{T_k\}_{k=0}^\infty = \{\sum_{i=0}^k (\gamma T)^i\}_{k=0}^\infty$, we have

$$\begin{aligned}
\|T_p - T_q\|_\infty &= \| \sum_{i=q+1}^p (\gamma T)^i \|_\infty \\
&= \|\gamma^{q+1}T^{q+1} + \cdots + \gamma^p T^p\|_\infty \\
&= \gamma^{q+1} + \cdots + \gamma^p \\
&= \frac{\gamma^{q+1} - \gamma^p}{1-\gamma} \\
&< \frac{\gamma^q}{1-\gamma} \\
&\leq \frac{\gamma^N}{1-\gamma} \\
&\leq \varepsilon, \forall p, q \geq N.
\end{aligned}$$

Therefore, the sequence $\{\sum_{i=0}^k (\gamma T)^i\}_{k=0}^\infty$ is Cauchy.

(c) We have

$$\begin{aligned}
(I-\gamma T)L &= (I-\gamma T)\lim_{n\to\infty}\sum_{i=0}^{n}(\gamma T)^i \\
&= \lim_{n\to\infty}\sum_{i=0}^{n}(I-\gamma T)(\gamma T)^i \\
&= \lim_{n\to\infty}\sum_{i=0}^{n}\left[(\gamma T)^i-(\gamma T)^{i+1}\right] \\
&= L-\lim_{n\to\infty}\sum_{i=1}^{n+1}(\gamma T)^i \\
&= L-\lim_{n\to\infty}\left(\sum_{i=0}^{n+1}(\gamma T)^i-I\right) \\
&= L-L+I \\
&= I.
\end{aligned}$$

Hence $(I-\gamma T)^{-1}=L=\sum_{i=0}^{\infty}(\gamma T)^i$.

■

Exercise 3: Grid World with a Given Policy 30pts

Consider the grid world shown in Figure 1. The finite state space is $\mathcal{S} = \{s_i : i = 1, 2, \ldots, 11\}$ and the finite action space is $\mathcal{A} = \{\text{up, down, left, right}\}$.
State transition probabilities After the agent picks and performs a certain action, there are four possibilities for the next state: the destination state, the current state, the states to the right and left of the current state. If the states are reachable, the corresponding probabilities are 0.8, 0.1, 0.05, and 0.05, respectively; otherwise, the agent stays where it is. The game will terminate if the agent arrives at s_{10} (loss) or s_{11} (win).
Reward After the agent picks and performs a certain action at its current state, it receives rewards of 100, -100, and 0, if it arrives at states s_{11}, s_{10}, and all the other states, respectively.
Policy In Figure 1, the arrows show the policy $\pi : \mathcal{S} \to \mathcal{A}$ for the agent. The random variable S_t is the state at time t under the policy π.

1. (5pts) Find the matrix $M \in \mathbb{R}^{11\times 11}$ with $m_{i,j} = \mathbf{P}(S_{t+1} = s_i | S_t = s_j)$, i.e., the conditional probability of the agent moving from s_j to s_i .
2. Suppose that the initial state distribution is uniform distribution, that is $\mathbf{P}(S_0 = s_i) = 1/11$, $i = 1, \ldots, 11$.
   (a) (5pts) Find the distributions $\mathbf{P}(S_1)$ and $\mathbf{P}(S_2)$ by following the policy π.
   (b) (5pts) Show that the agent would finally arrive at either s_{10} or s_{11}, i.e.,
   $$\lim_{t\to\infty} \mathbf{P}(S_t = s_i) = 0,\ i = 1, \ldots, 9.$$
   (c) (5pts) Please find $\lim_{t\to\infty} \mathbf{P}(S_t = s_{10})$ and $\lim_{t\to\infty} \mathbf{P}(S_t = s_{11})$.
3. (5pts) Find the value function corresponding to π, where the discount factor $\gamma = 0.9$.
4. **Bonus** (10pts) Show that the result in (2b) holds for any initial probabilities we choose for $\mathbf{P}(S_0 = s_i)$, $i = 1, \ldots, 11$.

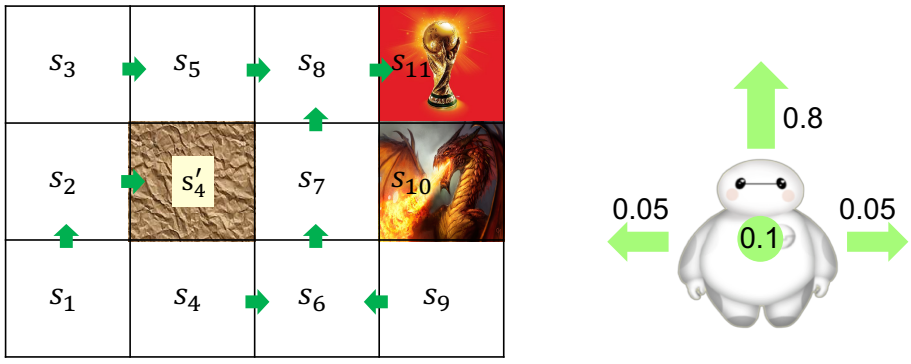

Figure 1: Illustration of a grid world with a policy.

Solution: 1. According to the problem, we get

$$M = \begin{bmatrix} 0.15 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.8 & 0.9 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.05 & 0.15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.05 & 0 & 0 & 0.2 & 0 & 0.05 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & 0 & 0.1 & 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.8 & 0.15 & 0.05 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0.8 & 0.15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.05 & 0 & 0 & 0.15 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.05 & 0 & 0.05 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 1 \end{bmatrix}.$$

2.

(a) Denote $\mathbf{P}_k = \begin{bmatrix} \mathbf{P}(S_k = s_1), & \mathbf{P}(S_k = s_1), & \ldots, & \mathbf{P}(S_k = s_1 1) \end{bmatrix}^T \in \mathbb{R}^{11}$ and $\mathbf{P}_k^{(i)}$ as the i^{th} entry. We have

$$\mathbf{P}_1 = M\mathbf{P}_0 = \begin{bmatrix} 0.01818182 \\ 0.15909091 \\ 0.01818182 \\ 0.02727273 \\ 0.09090909 \\ 0.15454545 \\ 0.09090909 \\ 0.15909091 \\ 0.01818182 \\ 0.1 \\ 0.16363636 \end{bmatrix},$$

$$\mathbf{P}_2 = M\mathbf{P}_1 = M^2\mathbf{P}_0 = \begin{bmatrix} 0.01068182 \\ 0.15863636 \\ 0.01068182 \\ 0.01409091 \\ 0.03272727 \\ 0.05181818 \\ 0.14522727 \\ 0.16931818 \\ 0.01045455 \\ 0.10545455 \\ 0.29090909 \end{bmatrix}.$$

(b) Since M is diagonalizable, we have $M = S\Lambda S^{-1}$, where the column vectors of S are eigenvectors of M.

Thus

$$\begin{aligned}\mathbf{P}_k &= M\mathbf{P}_{k-1} = \ldots = M^k\mathbf{P}_0 \\ &= S\Lambda S^{-1}S\Lambda S^{-1}\cdots S\Lambda S^{-1}\mathbf{P}_0 \\ &= S\Lambda^k S^{-1}\mathbf{P}_0.\end{aligned}$$

Note that M is a left stochastic matrix, then we know that M^k is a left stochastic matrix from exercise 2. By the property of left stochastic matrix, all the eigenvalues of M are no more than 1.

Thus

$$\begin{aligned}\lim_{k\to\infty}\mathbf{P}_k &= \lim_{k\to\infty} S\Lambda^k S^{-1}\mathbf{P}_0 \\ &= S\cdot\mathrm{diag}(1^\infty, 1^\infty, 0, \cdots)\cdot S^{-1}\mathbf{P}_0 \\ &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{v}_{10} \\ \mathbf{v}_{11}\end{bmatrix}\mathbf{P}_0 \\ &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0.1221 \\ 0.8779\end{bmatrix}.\end{aligned}$$

where $\mathbf{v}_{10} = [0.0107, 0.0073, 0.0039, 0.0657, 0.0037, 0.0657, 0.0623, 0.0037, 0.1207, 1, 0]$ and $\mathbf{v}_{11} = [0.9893, 0.9927, 0.9961, 0.9343, 0.9963, 0.9343, 0.9377, 0.9963, 0.8793, 0, 1]$.

Thus

$$\lim_{t\to\infty}\mathbf{P}(S_t = s_i) = \mathbf{P}_\infty^{(i)} = 0,\ i = 1, \ldots, 9.$$

That is, the agent would finally arrive at either s_{10} or s_{11}.

(c) We know that $\lim_{t\to\infty}\mathbf{P}(S_t = s_{10}) = \mathbf{P}_\infty^{(10)} = 0.1221$, and $\lim_{t\to\infty}\mathbf{P}(S_t = s_{11}) = \mathbf{P}_\infty^{(11)} = 0.8779$.

3. First we have $R = (E[r(s_1, \pi(s_1))], E[r(s_2, \pi(s_2))], \cdots, E[r(s_{11}, \pi(s_{11}))])^T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -5 \\ 80 \\ -5 \\ 0 \\ 0 \end{bmatrix}$.

   Next, by $V = R + \gamma M^\top V$, we have $V = (I - \gamma M^\top)^{-1} R = \begin{bmatrix} 21.2152 \\ 21.9747 \\ 71.5671 \\ 56.2074 \\ 84.6065 \\ 64.0139 \\ 74.4246 \\ 96.3573 \\ 47.5029 \\ 0 \\ 0 \end{bmatrix}$.

4. For any $\mathbf{P}_0$, we have

$$\begin{aligned} \lim_{t \to \infty} \mathbf{P}_t &= M^\infty \mathbf{P}_0 \\ &= \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{v}_{10} \\ \mathbf{v}_{11} \end{bmatrix} \mathbf{P}_0. \end{aligned}$$

   Thus

$$\lim_{t \to \infty} \mathbf{P}(S_t = s_i) = 0,\ i = 1, \ldots, 9.$$

   Q.E.D.

■

Exercise 4: Optimal Policy 40pts

Consider the grid world problem described in Exercise 3. Let π^* be the optimal policy, V^* the corresonding value function, and $\gamma = 0.9$.

1. (10pts) Please derive the Bellman Equation in terms of the value function V^* and the Q function, respectively.
2. (10pts) Please choose one of the algorithms we introduced in class to find π^* and V^* respectively and write their pseudocode (hand in your code if you have one).
3. (20pts) Please design a reward scheme such that following the resulting optimal policy will never lose. Specifically, you need to derive the resulting optimal policy and show

$$\lim_{t\to\infty} \mathbf{P}(S_t = s_i) = 0,\ i = 1, \ldots, 10,$$

whenever $\mathbf{P}(S_0 = s_{10}) = \mathbf{P}(S_0 = s_{11}) = 0$.

Solution: 1. Starting from an arbitrary state s_t, the expected cumulative reward by following the policy π^* is given by

$$\begin{aligned} V^*(s_t) &= E[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots \mid S_t = s_t] \\ &= E[r(s_t, \pi^*(s_t))] + \gamma E[R_{t+1} + \gamma R_{t+2} + \cdots \mid S_t = s_t] \\ &= E[r(s_t, \pi^*(s_t))] + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s_t, \pi^*(s_t)) V^*(s'). \end{aligned}$$

The Q function is defined as

$$Q(s, a) = E[r(s, a)] + \gamma \sum_{s'} P(s' \mid s, a) V^*(s'),$$

where s' denotes the probable state following s by performing action a. We have

$$\begin{aligned} \pi^*(s) &= \mathbf{argmax}_a\, Q(s, a), \\ V^*(s) &= \max_a Q(s, a). \end{aligned}$$

Hence

$$Q(s, a) = E[r(s, a)] + \gamma \sum_{s'} \left(P(s' \mid s, a) \max_{a'} Q(s', a') \right).$$

2. Use policy iteration as follows:

Algorithm 1 policy iteration

Initialize: $\pi \leftarrow \pi_0$, $\pi' \neq \pi_0$.

while $(\pi \neq \pi')$ **do**
  $V \leftarrow (I - \gamma T^{\pi})^{-1} R^{\pi}$
  $\pi' \leftarrow \pi$
  for $s \in \mathcal{S}$ **do**
    $\pi(s) \leftarrow \mathbf{argmax}_a \, E[r(s,a)] + \gamma \sum_{s'} P(s'|s,a) V(s')$
  end for
end while

Therefore we derive that

$$\pi^*(s_i) = \begin{cases} \text{up}, & i = 1 \\ \text{up}, & i = 2 \\ \text{right}, & i = 3 \\ \text{right}, & i = 4 \\ \text{right}, & i = 5 \\ \text{up}, & i = 6 \\ \text{up}, & i = 7 \\ \text{right}, & i = 8 \\ \text{left}, & i = 9 \end{cases}$$

and

$$V^*(s_i) = \begin{cases} 56.857, & i = 1 \\ 64.795, & i = 2 \\ 73.795, & i = 3 \\ 56.207, & i = 4 \\ 84.606, & i = 5 \\ 64.014, & i = 6 \\ 74.425, & i = 7 \\ 96.357, & i = 8 \\ 47.503, & i = 9 \end{cases}$$

3. Reward scheme: the agent will receive 100,-1000,0, if it arrives at s_{11}, s_{10} and all the other states, respectively.

Then by Policy iteration algorithm, we can derive the optimal policy:

$$\pi^*(s_i) = \begin{cases} \text{up}, & i = 1 \\ \text{up}, & i = 2 \\ \text{right}, & i = 3 \\ \text{left}, & i = 4 \\ \text{right}, & i = 5 \\ \text{left}, & i = 6 \\ \text{left}, & i = 7 \\ \text{right}, & i = 8 \\ \text{down}, & i = 9 \end{cases}$$

which is shown by figure 2.

The corresponding transition matrix T is

$$T = \begin{bmatrix} 0.15 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.8 & 0.2 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.8 & 0.15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.05 & 0 & 0 & 0.2 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.15 & 0.05 & 0 & 0.05 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.05 & 0.9 & 0.05 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0.05 & 0.15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.95 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 1 \end{bmatrix}.$$

Next we prove $\lim_{t\to\infty} \mathbf{P}(S_t = s_i) = 0$, $i = 1, \cdots, 10$ whenever $\mathbf{P}(S_0 = s_{10}) = \mathbf{P}(S_0 = s_{11}) = 0$:

Similar to exercise 3.4, for any initial probabilities $\mathbf{P}_0$, we have

$$\begin{aligned} \lim_{t\to\infty} \mathbf{P}_t &= T^{\infty}\mathbf{P}_0 \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \end{bmatrix} \mathbf{P}_0 \\ &= \left[0, \ \cdots, \ 0, \ \mathbf{P}_0^{(10)}, \ \mathbf{P}_0^{(1)} + \mathbf{P}_0^{(2)} + \cdots + \mathbf{P}_0^{(9)} + \mathbf{P}_0^{(11)}\right]^{\top}. \end{aligned}$$

Hence

$$\lim_{t\to\infty} \mathbf{P}(S_t = s_i) = 0,\ i = 1, \ldots, 10,$$

whenever $\mathbf{P}(S_0 = s_{10}) = \mathbf{P}(S_0 = s_{11}) = 0$.

■

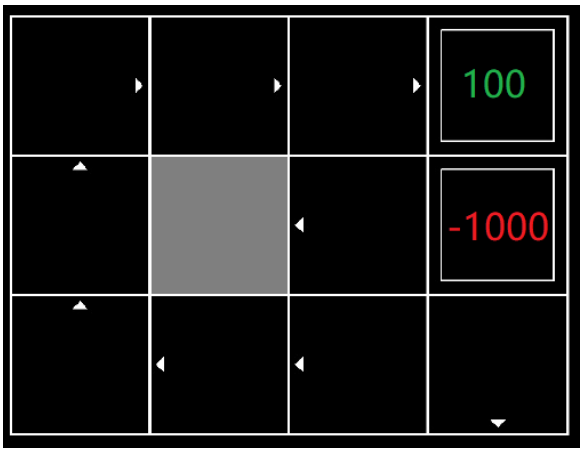

Figure 2: a never lose policy

Exercise 5: Policy Iteration 20pts

Consider a Markov Decision Process with bounded rewards and finite state-action pairs. The transition probability is $\mathbf{P}[s'|s,a]$ and the reward function is $r(s,a)$. Let $\pi : \mathcal{S} \to \mathcal{A}$ be a deterministic policy.

1. (10pts) Let $Q^\pi(s,a)$ be the accumulated reward by performing the action a first and then following the policy π. Find the Bellman Equation for Q^π.

2. (10pts) Consider a new policy π' given by

$$\pi'(s) = \mathop{\mathbf{argmax}}_{a\in\mathcal{A}} Q^\pi(s,a).$$

Note that if $\mathbf{argmax}_{a\in\mathcal{A}} Q^\pi(s,a)$ is not unique, we can choose one action arbitrarily. Show that $V^{\pi'}(s) \geq V^\pi(s)$ for all $s \in \mathcal{S}$.

Solution: 1.

$$Q^\pi(s,a) = \mathrm{E}[r(s,a)] + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a)V^\pi(s').$$

2. Define P^π as the transition matrix for policy π with demension $|\mathcal{S}| \times |\mathcal{S}|$, whose (s,s')-th is

$$[P^\pi]_{s,s'} = \mathbf{P}(s'|s,\pi(s)).$$

Similarly define R^π as the reward vector for policy π with dimension $|S| \times 1$, whose s-th entry is

$$[R^\pi]_s = r(s,\pi(s)).$$

Define V^π as the value vector for policy π with demension $|S| \times 1$, whose s-th entry is

$$[V^\pi]_s = V^\pi(s) = Q^\pi(s,\pi(s)).$$

Since

$$\pi'(s) = \mathop{\mathbf{argmax}}_{a\in\mathcal{A}} Q^\pi(s,a),$$

we have

$$\begin{aligned} Q^\pi(s,\pi'(s)) &\geq Q^\pi(s,\pi(s)) \\ &= V^\pi(s) \\ \Rightarrow R^{\pi'} + \gamma P^{\pi'} V^\pi &\geq V^\pi. \end{aligned} \tag{1}$$

Consider a sequence of non-stationary policies $\{\pi_i\}_{i\geq 0}$, where $\pi_0 = \pi$, $\pi_\infty = \pi'$. For any intermediate i, π_i is the non-stationary policy that follows π' for the first i time-steps and switches to π for the remainder of the trajectory. Then

$$V^{\pi_i} = \sum_{j=0}^{i-1} (\gamma P^{\pi'})^j R^{\pi'} + (\gamma P^{\pi'})^i V^{\pi},$$

where $\sum_{j=0}^{-1} (\gamma P^{\pi'})^j R^{\pi'} = 0$.

Note that $\lim_{i\to\infty} V^{\pi_i} = \sum_{j=0}^{\infty} (\gamma P^{\pi'})^j R^{\pi'} = V^{\pi'}$, since $\lim_{i\to\infty} (\gamma P^{\pi'})^i V^{\pi} = 0$.

Hence

$$\begin{aligned}
V^{\pi'} - V^{\pi} &= \lim_{i\to\infty} V^{\pi_i} - V^{\pi_0} \\
&= \sum_{i=0}^{\infty} (V^{\pi_{i+1}} - V^{\pi_i}) \\
&= \sum_{i=0}^{\infty} (\gamma P^{\pi'})^i R^{\pi'} + (\gamma P^{\pi'})^{i+1} V^{\pi} - (\gamma P^{\pi'})^i V^{\pi} \\
&= \sum_{i=0}^{\infty} (\gamma P^{\pi'})^i \left(R^{\pi'} + \gamma P^{\pi'} V^{\pi} - V^{\pi} \right) \\
&\geq 0,
\end{aligned}$$

where the last inequality comes from (1).

■