Solution Set #3

Warning: Homeworks will not be graded if submitted after the deadline. For all problems, show detailed reasoning.

1. (Two-armed bandit problem) *Consider a two-armed bandit problem, where the reward for the first arm is Bernoulli(p) and the reward for the second arm is Bernoulli(q), where* $0 < p, q < 1$*. Assume* $A_1 = 1$ *and* $A_2 = 2$*, i.e., you choose the first arm at time 1 and choose the second arm at time 2. Knowing the outcomes of the two tries, which arm should you choose at time* $t = 3$ *to maximize the chance of getting the reward of one at* $t = 3$*? Hint) You can also use a random strategy. Since the values of* p *and* q *are assumed to be unknown, try to maximize the minimum of the two expected rewards, one assuming* p *and* q *are swapped and the other assuming they are not. Maximizing the minimum of the two rewards will make your strategy symmetric and not dependent on the values of* p *and* q*.*

Solution Let a, b, c, and d denote the probabilities of choosing arm 1 when the two outcomes are $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$, respectively. Then, the maximum $R_{\max}$ of the minimum of the two expected rewards is given as follows.

$$\begin{aligned} R_{\max} = \max_{0\le a,b,c,d\le 1} \min\{&(1-p)(1-q)(ap+(1-a)q) + (1-p)q(bp+(1-b)q) \\ &+ p(1-q)(cp+(1-c)q) + pq(dp+(1-d)q), \\ &(1-q)(1-p)(aq+(1-a)p) + (1-q)p(bq+(1-b)p) \\ &+ q(1-p)(cq+(1-c)p) + qp(dq+(1-d)p)\} \end{aligned}$$

This is upperbounded as follows.

$$\begin{aligned} R_{\max} &\overset{(a)}{\le} \frac{1}{2} \max_{0\le a,b,c,d\le 1} (1-p)(1-q)(ap+(1-a)q) + (1-p)q(bp+(1-b)q) \\ &\qquad + p(1-q)(cp+(1-c)q) + pq(dp+(1-d)q) \\ &\qquad + (1-q)(1-p)(aq+(1-a)p) + (1-q)p(bq+(1-b)p) \\ &\qquad + q(1-p)(cq+(1-c)p) + qp(dq+(1-d)p) \\ &= \frac{1}{2} \max_{0\le a,b,c,d\le 1} (1-p)(1-q)(p+q) + q(bp+(1-b)q) + p(cp+(1-c)q) \\ &\qquad + p(bq+(1-b)p) + q(cq+(1-c)p) - pq(p+q) \\ &= \frac{1}{2} \max_{0\le a,b,c,d\le 1} p+q+(c-b)(p-q)^2 \\ &= \frac{p+q+(p-q)^2}{2}, \end{aligned}$$

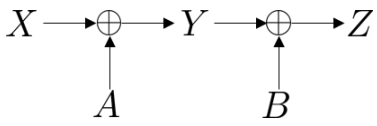
where (a) follows since $\min(x,y) \le \frac{x+y}{2}$. We can also show $R_{\max}$ is lowerbounded as follows.

$$\begin{aligned} R_{\max} &\overset{(b)}{\ge} \min\{(1-p)(1-q)(p+q)/2 + (1-p)q^2 + p^2(1-q) + pq(p+q)/2, \\ &\qquad\qquad (1-q)(1-p)(q+p)/2 + (1-q)p^2 + q^2(1-p) + qp(q+p)/2\} \\ &= (1-p)(1-q)(p+q)/2 + (1-p)q^2 + p^2(1-q) + pq(p+q)/2 \\ &= \frac{p+q+(p-q)^2}{2}, \end{aligned}$$

where (b) follows by setting $a = d = \frac{1}{2}$, $b = 0$, and $c = 1$. Therefore, by combining the above two bounds, we conclude that the maximum reward is $R_{\max} = \frac{p+q+(p-q)^2}{2}$, which is achieved by setting $a = d = \frac{1}{2}$, $b = 0$, and $c = 1$, i.e., choosing arms 1 or 2 with probability $\frac{1}{2}$ each when the outcomes are $(0,0)$ or $(1,1)$, choosing arm 1 when the outcomes are $(1,0)$ and choosing arm 2 when the outcomes are $(0,1)$.

2. (Markov property) *Find an example of three binary random variables X, Y, Z such that X is Bernoulli($\frac{1}{5}$), Y is Bernoulli($\frac{2}{5}$), Z is Bernoulli($\frac{7}{15}$), and $X - Y - Z$ form a Markov chain in that order. Hint) If X is Bernoulli(p), Y is Bernoulli(q), and X and Y are independent, then $X \oplus Y$ is Bernoulli($p(1-q) + (1-p)q$), where $X \oplus Y$ is the XOR of X and Y.*

Solution Let X, A and B be mutually independent random variables such that X is Bernoulli$\left(\frac{1}{5}\right)$ and A and B are Bernoulli$\left(\frac{1}{3}\right)$. Also, we define Y as $X \oplus A$ and Z as $Y \oplus B$. Then, Y is Bernoulli($\frac{2}{5}$) and Z is Bernoulli($\frac{7}{15}$).



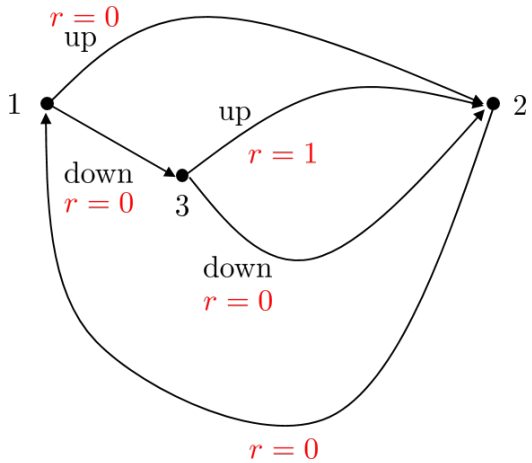
Then,

$$\begin{aligned}
\Pr(Z = z | X = x, Y = y) &= \Pr(Y \oplus B = z | X = x, Y = y) \\
&= \Pr(B = z \oplus y | X = x, Y = y) = \Pr(B = z \oplus y) \\
\Pr(Z = z | Y = y) &= \Pr(Y \oplus B = z | Y = y) = \Pr(B = z \oplus y | Y = y) = \Pr(B = z \oplus y)
\end{aligned}$$

Since $\Pr(Z = z | X = x, Y = y) = \Pr(Z = z | Y = y)$ $\forall x, y, z \in \{0, 1\}$, $X - Y - Z$ form a Markov chain.

3. (Value functions) *Find the state-value function $v_\pi(s)$ for the continuing task given in the figure in the right side in page 8 of lecture notes #16. Assume π is an optimal policy and $0 < \gamma < 1$.*

Solution



The Bellman equation in page 3 of lecture notes # 16 is,

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')]$$

And the transition probability of the given environment is given by:

$$p(s', r|s, a) = \begin{cases} 1 & \text{if } s' = 2,\ r = 0,\ s = 1,\ a = \text{up} \\ 1 & \text{if } s' = 3,\ r = 0,\ s = 1,\ a = \text{down} \\ 1 & \text{if } s' = 1,\ r = 0,\ s = 2 \\ 1 & \text{if } s' = 2,\ r = 1,\ s = 3,\ a = \text{up} \\ 1 & \text{if } s' = 2,\ r = 0,\ s = 3,\ a = \text{down} \\ 0 & \text{otherwise} \end{cases}$$

Also, it is obvious that the optimal policy is given by,

$$\pi(a|s) = \begin{cases} 1 & \text{if } a = \text{down},\ s = 1 \\ 1 & \text{if } a = \text{up},\ s = 3 \\ 0 & \text{otherwise} \end{cases}$$

Then, the Bellman equation becomes

$$\begin{pmatrix} v_\pi(1) \\ v_\pi(2) \\ v_\pi(3) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 & \gamma \\ \gamma & 0 & 0 \\ 0 & \gamma & 0 \end{pmatrix} \begin{pmatrix} v_\pi(1) \\ v_\pi(2) \\ v_\pi(3) \end{pmatrix}$$

Its solution is given by

$$\begin{pmatrix} v_\pi(1) \\ v_\pi(2) \\ v_\pi(3) \end{pmatrix} = \frac{1}{1-\gamma^3} \begin{pmatrix} \gamma \\ \gamma^2 \\ 1 \end{pmatrix}$$

4. (Maze game) *What is the minimum number of value iterations needed to learn to find the shortest path in the* 20×20 *maze game in page 19 in lecture notes #15? Assume* $v_0(s)$*'s are initialized to zero, the starting state is* $(1, 9)$ *(i.e., the 9th cell from left in the first row), and the terminal state is* $(20, 20)$ *(i.e., right bottom corner cell). The reward is 1 when the terminal state is reached from* $(20, 19)$ *and is 0 for all other transitions. Possible actions are up, down, left, and right. Assume the state is unchanged if an action is taken whose movement is blocked by an wall. If the action is not blocked by a wall, then you move by one cell. Assume* $0 < \gamma < 1$.

Solution In page 12 of lecture notes #16,

$$\begin{aligned} v_{k+1}(s) &= \max_a \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s, A_t = a] \\ &= \max_a \sum_{s', r} p(s', r|s, a)[r + \gamma v_k(s')] \end{aligned}$$

In the maze environment, there exists a single transition that produces nonzero reward, when we take right action at state (20,19). So, at the first step of value iteration, value of (20,19) is updated to nonzero. Then, at the next step, the value of (19,19) is updated.

$$\begin{array}{cccc} v_0(20,20) = 0 & v_1(20,20) = 0 & v_2(20,20) = 0 & v_3(20,20) = 0 \\ v_0(20,19) = 0 & v_1(20,19) = 1 & v_2(20,19) = 1 & v_3(20,19) = 1 \\ v_0(19,19) = 0 & v_1(19,19) = 0 & v_2(19,19) = \gamma & v_3(19,19) = \gamma \\ v_0(19,18) = 0 & v_1(19,18) = 0 & v_2(19,18) = 0 & v_3(19,18) = \gamma^2 \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

The shortest path is learned when $v(1,9)$ is updated. Then, the minimum number of value iterations needed to learn the shortest path is equal to the length of the shortest path from (1,9) to (20,20), i.e. 222.