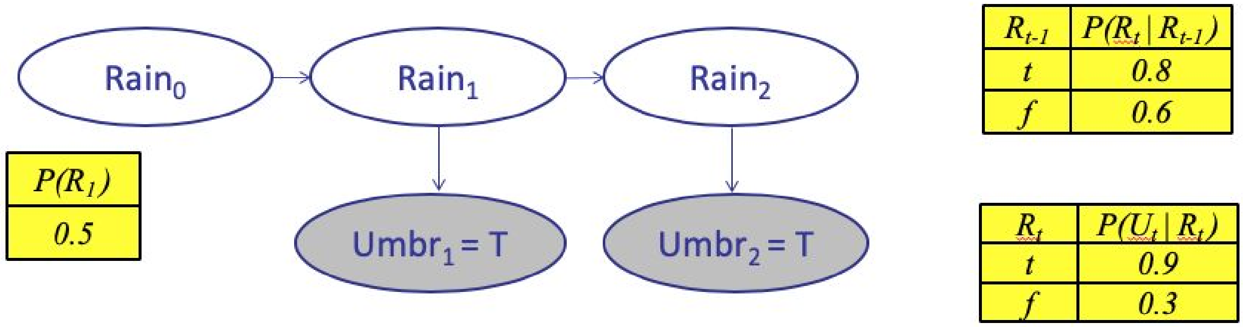
Question 1. Given an HMM with parameters specified below, calculate P(Rain$_2$ | Umbr$_1$, Umbr$_2$), i.e. that it's rainy on day 2 after observing umbrellas on day 1 and day 2.

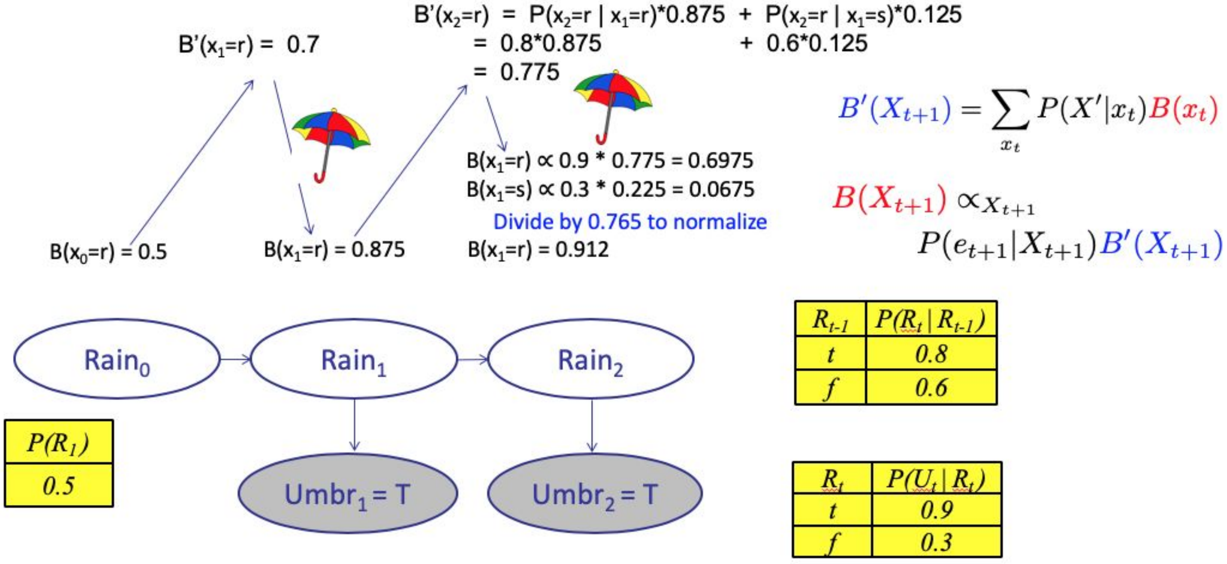


P(R$_1$)
0.5

R$_{t-1}$	*P(R$_t$ | R$_{t-1}$)*
t	*0.8*
f	*0.6*

R$_t$	*P(U$_t$ | R$_t$)*
t	*0.9*
f	*0.3*

Answer: 0.912



B'(x$_2$=r) = P(x$_2$=r | x$_1$=r)*0.875 + P(x$_2$=r | x$_1$=s)*0.125
= 0.8*0.875 + 0.6*0.125
= 0.775

B'(x$_1$=r) = 0.7

$$B'(X_{t+1}) = \sum_{x_t} P(X'|x_t)B(x_t)$$

B(x$_1$=r) ∝ 0.9 * 0.775 = 0.6975
B(x$_1$=s) ∝ 0.3 * 0.225 = 0.0675
Divide by 0.765 to normalize

$$B(X_{t+1}) \propto_{X_{t+1}} P(e_{t+1}|X_{t+1})B'(X_{t+1})$$

B(x$_0$=r) = 0.5

B(x$_1$=r) = 0.875

B(x$_1$=r) = 0.912

P(R$_1$)
0.5

R$_{t-1}$	*P(R$_t$ | R$_{t-1}$)*
t	*0.8*
f	*0.6*

R$_t$	*P(U$_t$ | R$_t$)*
t	*0.9*
f	*0.3*

Question 2:

A. No, they are dependent – R is a common cause.
B. Yes, they are independent – knowing R blocks the common cause.
C. No, dependent. – the common cause is still blocked, but since we now know the value of S, there is now an active path thru S.

Q4. [17 pts] Probability and Bayes Nets

(a) [2 pts] Suppose $A \perp\!\!\!\perp B$. Determine the missing entries (x, y) of the joint distribution $P(A, B)$, where A and B take values in $\{0, 1\}$.

$$P(A = 0, B = 0) = 0.1$$
$$P(A = 0, B = 1) = 0.3$$
$$P(A = 1, B = 0) = x$$
$$P(A = 1, B = 1) = y$$

$x =$ ___.15___, $y =$ ___.45___

Note that $y/x = P(A = 1, B = 1)/P(A = 1, B = 0) = P(A = 0, B = 1)/P(A = 0, B = 0) = P(B = 1)/P(B = 0) = 3$ So $y = 3x$ and $x + y = 0.6$. Solve for x, y.

(b) [3 pts] Suppose $B \perp\!\!\!\perp C \mid A$. Determine the missing entries (x, y, z) of the joint distribution $P(A, B, C)$.

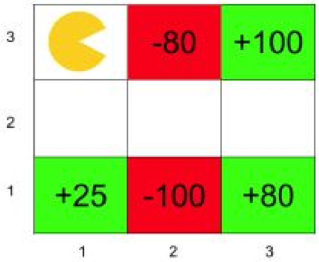
$$P(A = 0, B = 0, C = 0) = 0.01$$
$$P(A = 0, B = 0, C = 1) = 0.02$$
$$P(A = 0, B = 1, C = 0) = 0.03$$
$$P(A = 0, B = 1, C = 1) = x$$
$$P(A = 1, B = 0, C = 0) = 0.01$$
$$P(A = 1, B = 0, C = 1) = 0.1$$
$$P(A = 1, B = 1, C = 0) = y$$
$$P(A = 1, B = 1, C = 1) = z$$

$x =$ ___0.06___, $y =$ ___0.07___, $z =$ ___0.7___

First use the same observation about ratios as above to get that $x = 0.03 \cdot \frac{0.02}{0.01} = 0.06$. Then we have that $0.01 + 0.02 + 0.03 + 0.06 + 0.01 + 0.1 + y + z = 1$ so $y + z = 0.77$. The same observation about ratios gives $z/y = 10$. Solving, we get $y = 0.07, z = 0.7$.

Q8. [8 pts] Q-Learning Strikes Back

Consider the grid-world given below and Pacman who is trying to learn the optimal policy. If an action results in landing into one of the shaded states the corresponding reward is awarded during that transition. All shaded states are terminal states, i.e., the MDP terminates once arrived in a shaded state. The other states have the *North, East, South, West* actions available, which deterministically move Pacman to the corresponding neighboring state (or have Pacman stay in place if the action tries to move out of the grad). Assume the discount factor $\gamma = 0.5$ and the Q-learning rate $\alpha = 0.5$ for all calculations. Pacman starts in state (1, 3).



(a) [2 pts] What is the value of the optimal value function V^* at the following states:

$V^*(3,2) =$ ___100___ $V^*(2,2) =$ ___50___ $V^*(1,3) =$ ___12.5___

The optimal values for the states can be found by computing the expected reward for the agent acting optimally from that state onwards. Note that you get a reward when you transition *into* the shaded states and not *out* of them. So for example the optimal path starting from (2,2) is to go to the +100 square which has a discounted reward of $0 + \gamma * 100 = 50$. For (1,3), going to either of +25 or +100 has the same discounted reward of 12.5.

(b) [3 pts] The agent starts from the top left corner and you are given the following episodes from runs of the agent through this grid-world. Each line in an Episode is a tuple containing (s, a, s', r).

Episode 1	Episode 2	Episode 3
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), S, (2,1), -100	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0
	(3,2), N, (3,3), +100	(3,2), S, (3,1), +80

Using Q-Learning updates, what are the following Q-values after the above three episodes:

Q((3,2),N) = ___50___ Q((1,2),S) = ___0___ $Q((2,2),E) =$ ___12.5___

Q-values obtained by Q-learning updates - $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(R(s,a,s') + \gamma \max_{a'} Q(s',a'))$.

(c) Consider a feature based representation of the Q-value function:

$$Q_f(s,a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$$

$f_1(s)$: The x coordinate of the state $\quad$ $f_2(s)$: The y coordinate of the state

$$f_3(N) = 1,\ f_3(S) = 2,\ f_3(E) = 3,\ f_3(W) = 4$$

(i) [2 pts] Given that all w_i are initially 0, what are their values after the first episode:

$w_1 =$ ___-100___ $w_2 =$ ___-100___ $w_3 =$ ___-100___

Using the approximate Q-learning weight updates: $w_i \leftarrow w_i + \alpha[(R(s,a,s') + \gamma \max_{a'} Q(s',a')) - Q(s,a)]f_i(s,a)$. The only time the reward is non zero in the first episode is when it transitions into the -100 state.

(ii) [1 pt] Assume the weight vector w is equal to $(1,1,1)$. What is the action prescribed by the Q-function in state $(2,2)$?

___West___

The action prescribed at (2,2) is $\max_a Q((2,2),a)$ where $Q(s,a)$ is computed using the feature representation. In this case, the Q-value for *West* is maximum $(2+2+4=8)$.