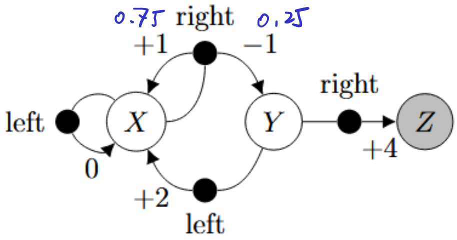
1. **Three state MDP**[1]

Consider the MDP below, in which there are three states, $\mathcal{S} = \{X, Y, Z\}$, two actions, $\mathcal{A} = \{\text{left}, \text{right}\}$, and the rewards on each transition are as indicated by the numbers. Note that if action *right* is taken in state X, then the transition may be either to X with a reward of $+1$ or to Y with a reward of -1. These two possibilities occur with probabilities 0.75 (for the transition to X) and 0.25 (for the transition to state Y). The state Z is a terminal state, i.e., all transitions from Z are back to Z with a reward of 0. The initial state is always X.



(a) $P_0 = (1, 0, 0)$

(b) For what combinations of inputs $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, $r \in \{4, 2, 1, 0, -1\}$ is the dynamics distribution $p(s', r|s, a)$ of this MDP non-zero? Note that the distribution is discrete since the states, actions, and rewards are discrete. Write down the probabilities for these combinations.

Hint: There should be seven combinations with non-zero probability.

$p(X, 0 | X, \text{left}) = 1$; $p(X, +1 | X, \text{right}) = 0.75$

$p(Y, -1 | X, \text{left}) = 0.25$

$p(X, +2 | Y, \text{left}) = 1$; $p(Z, +4 | Y, \text{right}) = 1$

$p(Z, 0 | Z, \text{left}) = 1$; $p(Z, 0 | Z, \text{right}) = 1$

(c) Write down $\mathcal{P}(s'|s, a)$ and $\mathcal{R}(s, a)$ for all $s, s' \in \mathcal{S}, a \in \mathcal{A}$. The reward function can be derived from the dynamics distribution considered in part (b) using the formula from the lecture.

$s = X$:

$P(X|X, \text{left}) = 1$, $P(Y|X, \text{left}) = 0$, $P(Z|X, \text{left}) = 0$

$P(X|X, \text{right}) = 0.25$, $P(Y|X, \text{right}) = 0.75$, $P(Z|X, \text{right}) = 0$

$s = Y$:

$P(X|Y, \text{left}) = 1$, $P(Y|Y, \text{left}) = 0$, $P(Z|Y, \text{left}) = 0$

$P(X|Y, \text{right}) = 0$, $P(Y|Y, \text{right}) = 0$, $P(Z|Y, \text{right}) = 1$

$s = Z$:

$P(X|Z, \text{left}) = 0$, $P(Y|Z, \text{left}) = 0$, $P(Z|Z, \text{left}) = 1$

$P(X|Z, \text{right}) = 0$, $P(Y|Z, \text{right}) = 0$, $P(Z|Z, \text{right}) = 1$

$s = X$

$R(X, \text{left}) = 0$

$R(X, \text{right}) = (+1) \cdot 0.75 + (-1) \cdot 0.25 = 0.5 = \sum r \sum p(s', r | s, a)$

$s = Y$

$R(Y, \text{left}) = +2$

$R(Y, \text{right}) = +4$

$s = Z$

$R(Z, \text{left}) = 0$

$R(Z, \text{right}) = 0$

(d) Consider the two deterministic policies π_1 and π_2:

$$\pi_1(X) = \text{right} \qquad \pi_2(X) = \text{left}$$
$$\pi_1(Y) = \text{right} \qquad \pi_2(Y) = \text{right}$$

Write down a typical trajectory for policy π_1, i.e., make up a sequence of states, actions, and rewards that is likely to occur. What happens if you do this for π_2?

π_1:

	$S_0 = X$, $a_0 = \pi_1(X) = \text{right}$, $r_1 = +1$	π_2: stays at X forever
$p = 0.75$	$S_1 = X$, $a_1 = \pi_1(X) = \text{right}$, $r_2 = -1$	$S_0 = X$, $a_0 = \pi_2(X) = \text{left}$, $r = 0$
$p = 0.25$	$S_2 = Y$, $a_2 = \pi_1(Y) = \text{right}$, $r_3 = +4$	$S_1 = X$, same
$p = 1$	$S_3 = Z$	