1. **Discounted returns**

(a) Assume you observe a sequence of five rewards

$$R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2$$

until you reach a terminal state, i.e. a state that always transitions back to itself with a reward of 0. Calculate the returns $G_0, \ldots, G_5$ for a discount factor of $\gamma = 0.5$.

$$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5 + \sum_{k=0}^{\infty} \gamma^{5+k} \cdot 0$$

$$= -1 + 0.5 \times 2 + 0.25 \times 6 + 0.125 \times 3 + 0.0625 \times 2$$

$$G_5 = \sum_{k=0}^{\infty} \gamma^k \cdot \underset{R_{5+k+1}}{\underset{\Downarrow}{0}}$$

$$G_4 = R_5 = 2$$

$$G_3 = R_4 + \gamma R_5 = 3 + 1 = 4$$

$$G_2 = 6 + 2 = 8$$

$$G_1 = 2 + 4 = 6$$

$$G_0 = -1 + 3 = 2$$

(b) Assume an MDP produces an infinite sequence of rewards of 5, i.e.

$$R_1 = 5, R_2 = 5, R_3 = 5, \ldots$$

Calculate the return G_0 for the discount factors $\gamma \in \{0, 0.2, 0.5, 0.9, 0.95, 0.99, 0.999\}$. What would happen if the discount factor was $\gamma = 1$?

Hint: You can use the closed form of a special case of the power series.

$$G_0 = 5 \sum_{k=0}^{\infty} \gamma^k$$

$\cancel{\gamma = 0, \; G_0 = 0 \; 5}$ only look at the next step

$$\gamma \geq 0, \quad G_0 = \frac{5}{1-\gamma}$$

$$\gamma = 0.2 \quad G_0 = \frac{5}{0.8} = 6.25$$

$$\gamma = 0.5 \quad G_0 = \frac{5}{0.5} = 10$$

$$\gamma = 0.9 \quad G_0 = \frac{5}{0.1} = 50$$

$$\gamma = 0.95 \quad G_0 = \frac{5}{0.05} = 100$$

$$\gamma = 0.99 \quad G_0 = \frac{5}{0.01} = 500$$

$$\gamma = 0.999 \quad G_0 = \frac{5}{0.001} = 5000$$

importance of future increases as γ increase

↳ make sure the return is properly bounded

" one trajectory to the end, no feedback in between "

tic-tak-toe

Real case (sparse reward environment)
e.g.

(c) Assume you observe a sequence of $T > 1$ rewards

$$R_1 = 0, R_2 = 0, \ldots, R_{T-1} = 0, R_T$$

$\Rightarrow$ get reward at the end of the dialogue

until you reach a terminal state. Note that all rewards except R_T are zero. How can you choose γ such that the initial return G_0 is equal to $\epsilon > 0$? Calculate these γ for the following situations:

i. $\epsilon = 0.1$, $R_T = 1$, $T = 10$
ii. $\epsilon = 0.1$, $R_T = 1$, $T = 50$
iii. $\epsilon = 0.01$, $R_T = 1$, $T = 50$

time $\uparrow$ factor $\uparrow$

return $\downarrow$ factor $\downarrow$

$G_0 = \gamma^{T-1} R_T \qquad \gamma = \sqrt[T-1]{\frac{G_0}{R_T}} = \sqrt[T-1]{\frac{\epsilon}{R_T}}$

$G_T = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

$G_0 = R_1 + \gamma R_2 + \cdots + \gamma^{T-1} R_T$

$\Updownarrow$

$\gamma = \sqrt[T-1]{\frac{\epsilon}{R_T}}$

i. $\gamma = \sqrt[9]{0.1} \approx 0.7743$

ii. $\gamma = \sqrt[49]{0.1} \approx 0.9541$

iii. $\gamma = \sqrt[49]{0.01} \approx 0.9103$

2. **Value functions**

$v_\pi(s) = \sum_a \pi(a|s)\left(R(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\pi(s')\right)$

(a) For any given MDP, policy π and *terminal state* E, what is $v_\pi(E)$? All transitions from a terminal state are back to itself with a reward of 0.

$p(G_t = 0 \mid S_t = E) = 1 \qquad v_\pi(E) = E[G_t \mid S_t = E]$

$= 0 \cdot p(G_t = 0 \mid S_t = E) = 0$

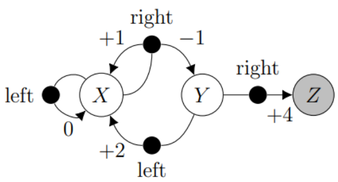
$v_\pi(E) = \sum_a \pi(a|E)(\underset{0}{R(s,a)} + \gamma\, \underset{1}{p(E|E,a)} v_\pi(E))$

$\Rightarrow v_\pi(E) = \gamma v_\pi(E)$

$\gamma < 1 \Rightarrow v_\pi(E) = 0$

(b) Consider the MDP and policy π_1 from the previous exercise set. Note that if action *right* is taken in state X, then the transitions to X and Y occur with probabilities 0.75 and 0.25, respectively. The deterministic policy π_1 is defined as $\pi_1(X) = \text{right}$, $\pi_1(Y) = \text{right}$.

value functions are policy dependent



Calculate the values of states X and Y under policy π_1, i.e. $v_{\pi_1}(X)$ and $v_{\pi_1}(Y)$, using a discount factor of $\gamma = 0.9$.

Hint: Start with the value of Y.

$v_{\pi_1}(Y) = E[G_t \mid S_t = Y] = 4 \cdot p(G_t = 4 \mid S_t = Y) = 4$

for $v_\pi(X)$, $\quad g_0 = \sum_{k=0}^{\infty} \gamma^k \cdot (-1) \qquad p = \lim_{i \to \infty} 0.75^i = 0$

$T = 0, 1, \cdots, \quad g_T = \sum_{k=0}^{T-1} \gamma^k (-1) + 4\gamma^T \qquad p = 0.75^{T-1} \cdot 0.25$

$v_{\pi_1}(X) = \sum_{T=0}^{\infty} \left(-\sum_{k=0}^{T-1} \gamma^k + 4\gamma^T\right) \cdot 0.75^{T-1} \cdot 0.25$

$\Downarrow$

$\frac{1 - \gamma^{T-1}}{1 - \gamma} + 4\gamma^T$

$v_{\pi_1}(X) = \sum_{k=0}^{\infty} 0.75^k \cdot 0.25\, v_{\pi_1}(Y) + \sum_{k=0}^{\infty} \left(\frac{1 - \gamma^{T-1}}{1 - \gamma}\right) \cdot 0.75^b \cdot 0.25$

$S \in \{x, y, z\}$

$A = \{left, right\}$

$\pi_1(right|X) = 1, \ \pi_1(left|X) = 0$

$\pi_1(right|Y) = 1, \ \pi_1(left|Y) = 0$

$v_{\pi_1}(Z) = 0$

$$v_{\pi_1}(Y) = \overset{1}{\overset{\Uparrow}{\pi_1(right|Y)}}(4 + r(P(Z|Y, right)\, v_{\pi_1}(Z)))$$

$$= 4$$

$$v_{\pi_1}(X) = \pi_1(right|X)(0.5 + r(P(X|X, right)\, v_{\pi_1}(X) + P(Y|X, right)\, v_{\pi_1}(Y)))$$

$R(s,a) = 0.5$

✓ how to calculate $v_{\pi_1}(x)$ in general: we "evaluate" the policy iteratively

$0.325\, v_{\pi_1}(X) = 1.4$

(c) We may add "-1" reward to intermediate step.