CS449/CS549 Computational Learning Quiz 4 Fall 2023 (take home).
Due: 12/05/23 (in class, submit hardcopy). *You may consult class materials and other additional references, but your submitted answers must be your own work.*

Markov Decision Process: infinite play.
Let $S = \{s_0, s_1, s_2\}$ be a set of states, $A = \{a, b, c\}$ be a set of actions, $\{P_a, P_b, P_c\}$ be a set of Markov chains, and $R : S \to \mathbb{R}$ be a reward function given by

$$P_a = \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}, \qquad P_b = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad P_c = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \qquad R = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \gamma = \frac{1}{2}$$

Our MDP is $M = (S, A, \{P_a, P_b, P_c\}, R, s_0, \gamma)$. A policy is a map $\pi : S \to A$.

1. Consider the `random-play` policy $\pi_a = \begin{bmatrix} a \\ a \\ a \end{bmatrix}$. Compute its value function V_{π_a}.
2. (T/F) `random-play` is bad: π_a is not optimal. *Hint*: apply the fixed-point iterator Φ.
3. Consider a suspiciously better policy $\pi_o = \begin{bmatrix} b \\ a \\ c \end{bmatrix}$. Compute its value function V_{π_o}.
4. (T/F) The suspicion is well-founded: π_o is better than `random-play`.
5. (T/F) The suspicion is more than well-founded: π_o is optimal. *Hint*: apply the fixed-point iterator Φ.
6. (CS549) (T/F) Policy Iteration is better than Value Iteration on π_a.