Problem 1

For this problem, we will work with a reward function operating on transitions, \( R : S \times A \times S \rightarrow \mathbb{R} \). We are given an infinite-horizon, discounted MDP \( M = \langle S, A, R, T, \gamma \rangle \) but we will actually solve a MDP \( M' \) with an augmented reward function \( M' = \langle S, A, R', T, \gamma \rangle \) where \( R'(s, a, s') = R(s, a, s') + F(s, a, s') \). To provide some motivation, think of a scenario where \( R \) produces values of 0 for most transitions; a bonus reward function \( F : S \times A \times S \rightarrow \mathbb{R} \) that produces non-zero values could provide us more immediate feedback and help accelerate the learning speed of our agent. In this problem, we will focus on a particular type of reward bonus \( F(s, a, s') = \gamma \phi(s') - \phi(s) \), for some arbitrary function \( \phi : S \rightarrow \mathbb{R} \) and \( \forall (s, a, s') \in S \times A \times S \).

1. Let \( Q^*_M, Q^*_M' \) denote the optimal action-value functions of MDPs \( M \) and \( M' \), respectively. Using the Bellman equation, prove that \( Q^*_M(s, a) - \phi(s) = Q^*_M'(s, a) \) and then use this fact to conclude that \( \pi^*_M(s) = \pi^*_M(s), \forall s \in S \).
2. Consider running $Q$-learning in each MDP $\mathcal{M}$ and $\mathcal{M}'$ which requires, for each MDP, initial values $Q^0_{\mathcal{M}}(s,a)$ and $Q^0_{\mathcal{M}'}(s,a)$. Let $q_{\text{init}} \in \mathbb{R}$ be a real value such that

$$Q^0_{\mathcal{M}}(s,a) = q_{\text{init}} + \phi(s), \quad Q^0_{\mathcal{M}'}(s,a) = q_{\text{init}}.$$

At any moment in time, the current $Q$-value of any state-action pair is always equal to its initial value plus some $\Delta$ value denoting the total change in the $Q$-value across all updates:

$$Q_{\mathcal{M}}(s,a) = Q^0_{\mathcal{M}}(s,a) + \Delta Q_{\mathcal{M}}(s,a), \quad Q_{\mathcal{M}'}(s,a) = Q^0_{\mathcal{M}'}(s,a) + \Delta Q_{\mathcal{M}'}(s,a).$$

Show that if $\Delta Q_{\mathcal{M}}(s,a) = \Delta Q_{\mathcal{M}'}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, then show that these two $Q$-learning agents yield identical updates for any state-action pair.