Homework 5 Challenge Reflection:

10.5) Conditions for convergence are satisfied.

Therefore, the Q-values converge to expected return

$$Q^{S1}(S, \rightarrow) = \frac{\text{sum of rewards of episodes starting at S and going} \rightarrow}{\text{number of episodes starting at S and going} \rightarrow} = \frac{1+10}{2} = \boxed{\frac{11}{2}}$$

$$Q^{S1}(A, esc) = \frac{\text{sum of rewards of episodes starting at A and going esc}}{\text{number of episodes starting at A and going esc}} = \frac{1+10}{2} = \boxed{\frac{11}{2}}$$

10.6) Conditions for convergence are satisfied.

Therefore, the Q-values converge to expected return

$$Q^{S2}(S, \rightarrow) = \frac{\text{sum of rewards of episodes starting at S and going} \rightarrow}{\text{number of episodes starting at S and going} \rightarrow} = \frac{1+10+10}{3} = \frac{21}{3} = \boxed{7}$$

$$Q^{S2}(A, esc) = \frac{\text{sum of rewards of episodes starting at A and going esc}}{\text{number of episodes starting at A and going esc}} = \frac{1+10+10}{3} = \frac{21}{3} = \boxed{7}$$

10.7) $\boxed{Q^{S1}(S, \rightarrow)}$. S1 has the same distribution of returns as the true distribution, even though all possible transitions are not experienced. I.e. $Q^{S1}(S, \rightarrow) = 11/2 = V^*(S)$ whereas $Q^{S2}(S, \rightarrow) = 7 \neq V^*(S)$.

10.8) $\boxed{\text{MDP1 only}}$. This is because in deterministic MDPs, even with a learning rate of 1 (i.e. giving 100% weight to new samples and 0% weight to old samples), you still converge, because actions are deterministic. In stochastic MDPs, there is no such guarantee since resulting states from actions are not fixed.