Puterman 2.5 (a)

We have the following setup:

Horizon: $T = \{1, 2\}$. $(N = 2)$

States: $S = \{s_1, s_2\}$.

Actions: $A_{s_1} = \{a_{11}, a_{12}\}$, $A_{s_2} = \{a_{21}, a_{22}\}$

Rewards: $r_1(s_1, a_{11}) = 5$, $r_1(s_1, a_{12}) = 10$,
$r_1(s_2, a_{21}) = -1$, $r_1(s_2, a_{22}) = 2$
$r_N(s_1) = r_2(s_1) = 0$, $r_N(s_2) = r_2(s_2) = 0$. $\Big\}\ v(s) = 0$.

In the one-period model, we can apply eq. (2.2.1) from Puterman. Using eq. (2.2.1) we find the total expected reward for the different policies:

$$X_1 = s_1,\ d(s_1) = a_{11}: \quad r_1(s_1, a_{11}) + \sum_{j \in S} p_1(j \mid s_1, a_{11})\, v(j)$$

$$(\text{As } v(j) = 0,\ \forall j \in S) \quad = r_1(s_1, a_{11}) = 5.$$

$$X_1 = s_1,\ d(s_1) = a_{12}: \quad r_1(s_1, a_{12}) = 10,$$
$$X_1 = s_2,\ d(s_2) = a_{21}: \quad r_1(s_2, a_{21}) = -1,$$
$$X_1 = s_2,\ d(s_2) = a_{22}: \quad r_1(s_2, a_{22}) = 2.$$

Thus, the optimal policy $\pi^* = (d^*(s_1), d^*(s_2))$
$= (a_{12}, a_{22})$.