Learning to Play Optimally

Let us consider the following simple game for two agents:

r^i	a_1^2	a_2^2
a_1^1	$1.0, 2.0$	$0.0, 0.0$
a_2^1	$0.0, 0.0$	$2.0, 1.0$

The first number in each cell is a payoff for agent 1 and the second number for agent 2.

Teach the game for two Q-learning agents endowed with the normal single-agent Q-learning algorithm. As this is a stagegame, i.e. a game containing only one state, the Q-learning becomes:

$$Q_t^i(a_t^i) = (1 - \alpha_t)Q_t^i(a_t^i) + \alpha_t r^i(a_t^1, a_t^2),$$

where i is the learner identification, i.e. $i = 1, 2$. Apply pure random exploration for action selection during the learning. The learning rate parameter α_t can be fixed, e.g. 0.1. Is the resulting policy optimal? Change the exploration to the SoftMax-exploration policy. How is the resulting policy now? In addition, plot a graph demonstrating the convergence of the learning with both exploration strategies.

Extend the normal Q-learning rule for multiagent case, i.e.

$$Q_t^i(a_t^1, a_t^2) = (1 - \alpha_t)Q_t^i(a_t^1, a_t^2) + \alpha_t r^i(a_t^1, a_t^2).$$

Solve the game using this rule. Is the learning sensitive for selected exploration strategy? After the utility values have been learned, how can the optimal policy be solved?