# 2-Player Q-learning

Normal update equation:

$$Q[s,a] \leftarrow Q[s,a] + \alpha \left[ r + \gamma \max_{a'} Q[s',a'] - Q[s,a] \right]$$

Normally we always maximize our rewards. Consider ***2-player Q-learning*** with player A maximizing and player B minimizing (as in minimax).

Why does this break the update equation?

# 2-Player Q-learning

Player A's update equation:

$$Q[s,a] \leftarrow Q[s,a] + \alpha \left[ r + \gamma \min_{a'} Q[s',a'] - Q[s,a] \right]$$

Player B's update equation:

$$Q[s,a] \leftarrow Q[s,a] + \alpha \left[ r + \gamma \max_{a'} Q[s',a'] - Q[s,a] \right]$$

Player A's optimal policy output:

$$\pi(s) = \operatorname*{argmax}_{a} Q[s,a]$$

Player B's optimal policy output:

$$\pi(s) = \operatorname*{argmin}_{a} Q[s,a]$$