# Chapter 1 Exercises

Kyler Krenzke

1. **Exercise 1.1 (Self-Play): Suppose, instead of playing against a random opponent, the reinforcement learning algorithm described above played against itself, with both sides learning. What do you think would happen in this case? Would it learn a different policy for selecting moves?**

   Yes, it would learn a different policy for selecting moves. The values of all the actions would change over time as the opponent changes their strategy.

2. **Exercise 1.2 (Symmetries): Many tic-tac-toe positions appear different but are really the same because of symmetries. How might we amend the learning process described above to take advantage of this? In what ways would this change improve the learning process? Now think again. Suppose the opponent did not take advantage of symmetries. In that case, should we? Is it true, then, that symmetrically equivalent positions should necessarily have the same value?**

   The learning process could be amended to identify states as the same if they are symmetrically identical. This could improve the learning process by requiring less total states to converge if the opponent also treats symmetric states as identical. However, if the opponent plays differently on differing, symmetric boards, the agent is short-changing its state space and will not find the optimal solution.

3. **Exercise 1.3 (Greedy Play): Suppose the reinforcement learning player was greedy, that is, it always played the move that brought it to the position that it rated the best. Might it learn to play better, or worse, than a nongreedy player? What problems might occur?**

   A greedy player would never explore any actions with lower values which means the policy the player learns will be dependant on the player's inital policy.

4. **Exercise 1.4 (Learning from Exploration): Suppose learning updates occurred after all moves, including exploratory moves. If the step-size parameter is appropriately reduced over time (but not the tendency to explore), then the state values would converge to a different set of probabilities. What (conceptually) are the two sets of probabilities computed when we do, and when we do not, learn from exploratory moves? Assuming that we do continue to make exploratory moves, which set of probabilities might be better to learn? Which would result in more wins?**

   The first set of probabilities (with the tendency to explore being reduced) would converge to the a policy for the agent with a perfect transition model. The second set of probabilities (without the tendency to explore being reduced) would converge to the optimal policy for the agent if a move only has a 1-alpha probability. Assuming the agent does continue to make exploratory moves, the second set of probabilities would be better to learn.

5. **Exercise 1.5 (Other Improvements): Can you think of other ways to improve the reinforcement learning player? Can you think of any better way to solve the tic-tac-toe problem as posed?**

   The player could improve by using planning. The tic-tac-toe state space is small enough that direct planning would be a feasible solution.