Exercises - Chapter 1

Victoria Nagorski

Exercise 1.1: Self-Play

Suppose, instead of playing against a random opponent, the reinforcement learning algorithm described above played against itself. What do you think would happen in this case? Would it learn a different way of playing?

In the case of tic-tac-toe, the games would eventually result in a draw from both opponents. Reinforcement learning would discover the optimal actions to beat the opponent. However, if both agents are finding the optimal actions, it would ultimately lead to a draw.

Exercise 1.2: Symmetries

Many tic-tac-toe positions appear dierent but are really the same because of symmetries. How might we amend the reinforcement learning algorithm described above to take advantage of this? In what ways would this improve it? 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?

Symmetries would decrease the number of states, and thus lessen the amount of required calculations. If the opponent, however, did not take advantage of symmetries, then we might want to use the full board (unless acting like the board is symmetrical brings about the best outcome).

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. Would it learn to play better, or worse, than a non-greedy player? What problems might occur?

Worse long-run. It would never explore and learn better strategies. It might consistently choose A but C might give better reward. It would never know because it did not explore.

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, then the state values would converge to a set of probabilities. What 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?

Do $\rightarrow$ Probability would then effect score further behind with a potentially lower score because an exploratory move.
Do-Not $\rightarrow$ Would be more accurate. Update exploratory field based off of the new actions.
Probably not updating with the exploratory moves would produce the most wins.

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?

Online learning. If it could learn from multiple people- not just one opponent.