Chapter 1: Introduction - Exercises

Jacob Taylor Cassady

November 22, 2022

1 Self-Play

Suppose, instead of playing against a random opponent, the reinforcement learning algorithm described above played against itself, with both sides learning.

1.1 What do you think would happen in this case?

It would converge on always drawing.

1.2 Would it learn a different policy for selecting moves?

The policies for each side would converge on the same policy.

2 Symmetries

Many tic-tac-toe positions appear different but are really the same because of symmetries.

2.1 How might we amend the learning process described above to take advantage of this?

Enforce that all matching positions are updated when one is.

2.2 In what ways would this change improve the learning process?

It would speed up the time of convergence on the optimal policy.

2.3 Now think again. Suppose the opponent did not take advantage of symmetries. In that case, should we?

Yes! This would give us a faster rate of convergence; thus leading to more early wins. Sucks to suck, opponent!

2.4 Is it true, then, that symmetrically equivalent positions should necessarily have the same value?

Yes. See my answer two questions ago.

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.

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

It depends on the opponent! If the opponent switched their play style, the greedy player might not have updated the value of those states yet and consequently may perform less optimally. If the opponent had only one play style, the greedy algorithm would converge faster.

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.

4.1 What (conceptually) are the two sets of probabilities computed when we do, and when we do not, learn from exploratory moves?

The first set of probabilities, when the tendency to explore is reduced over time, the algorithm would not learn less from bad exploratory moves long term. On the other hand, when the tendency to explore is not reduced over time, the algorithm would learn more from poor exploratory moves.

4.2 Assuming that we do continue to make exploratory moves, which set of probabilities might be better to learn?

The set of probabilities where the tendency to explore is reduced over time.

4.3 Which would result in more wins?

The set of probabilities where the tendency to explore is reduced over time. As long as the rate of reduction is sufficiently low to allow for exploration at the beginning.

5 Other Improvements

5.1 Can you think of other ways to improve the reinforcement learning player?

You could assign higher values to certain moves at the point of intializing the table that are more likely to lead to a win such as the center square. You may even base these intial values on the amount of win conditions including that square.

5.2 Can you think of any better way to solve the tic-tac-toe problem as posed?

Not at the moment.