Name: _____

Tic-Tac-Toe Revised

3. (12 points) Tic-tac-toe is a paper-and-pencil game for two players, X and O, who take turns marking the spaces in a 3×3 grid. The player who succeeds in placing three of their marks sequentially in a horizontal, vertical, or diagonal row wins the game. The following example game is won by the first player, X:

In this question, we'll consider a "solitaire" version of tic-tac-toe, in which we assume:

- We are the X player;
- The O player is a fixed (but possibly stochastic) algorithm;
- The initial state of the board is empty, and X has the first move;
- We can select any of the nine squares on our turn;
- We don't know the strategy of the O player or the reward function used by O.

We place an X in an empty square, then an O appears in some other square, and then it's our turn to play again. We receive a +1 reward for getting three X's in a row, reward -1 if there are three O's in a row, and reward 0 otherwise. If we select a square that already has an X or an O in it, nothing changes and it's still our turn.

(a) We can model this problem as a Markov decision process in several different ways. Here are some possible choices for the state space.

- Jody suggests letting the state space be all possible 3 x 3 grids in which each square contains one of the following: a space, an O, and an X.
- Dana suggests using all possible 3 x 3 grids in which each square contains one of the three options (a space, an O, and an X), and there is an equal number of O's and X's.
- Chris suggests using all 3 x 3 tic-tac-toe game grids which appear in games where the players both employ optimal strategies.

i. Is Dana's suggestion better or worse for tabular Q learning than Jody's? Explain your answer.

ii. Is Chris' suggestion better or worse for tabular Q learning than Jody's? Explain your answer.

(b) Many states of the game are effectively the same due to symmetry.

i. Draw a pair of such states which are the same due to symmetry:

ii. Jordan suggests using a state-space that includes one state that stands for each set of board games that are equivalent due to symmetry. Would this be better or worse for learning than Jody's representation? Explain your answer.

(c) What is the action space of the MDP with Dana's state space definition?

(d) You get to sit and watch an expert player (who always makes optimal moves) play this game for a long time, and you observe the sequence of state-action pairs that occur in many games. Which of the following machine-learning problem formulations is most appropriate, for you to learn how to play the game? For the item you select, provide the specified additional information (where not "none").

1. supervised regression (describe the loss function)
2. supervised classification (describe the loss function)
3. reinforcement learning of a policy (none)
4. reinforcement learning of a value function (none)

Explain your answer.

(e) You get to interact with an implementation of this game for many game instances, selecting your actions, observing the results and rewards. Which of the following machine-learning problem formulations is most appropriate, for you to learn how to play the game? For the item you select, provide the specified additional information (where not "none").

1. supervised regression (describe the loss function)
2. supervised classification (describe the loss function)
3. reinforcement learning of a policy (none)
4. reinforcement learning of a value function (none)

Explain your answer.

(f) Barney wants to solve a tic-tac-toe problem that is exactly the same as the above game (i.e., three in a row/column/diagonal wins), except that it is played on a 100 x 100 grid.

i. Is it better for Barney to use tabular Q learning or neural-net Q learning? Explain.

ii. Suppose Barney were to use neural-net Q learning; would it help for him to start with a convolutional layer? If your answer is yes, describe four 3 x 3 convolutional filters that would be particularly helpful for this problem.

(g) Suppose you apply Q-learning to the 3x3 tic-tac-toe problem, and your actions always select an unfilled square. Bert suggests that it is okay to let the discount factor be 1. Is that true? Explain why or why not.