Homework 3 Tic-Tac-Toe

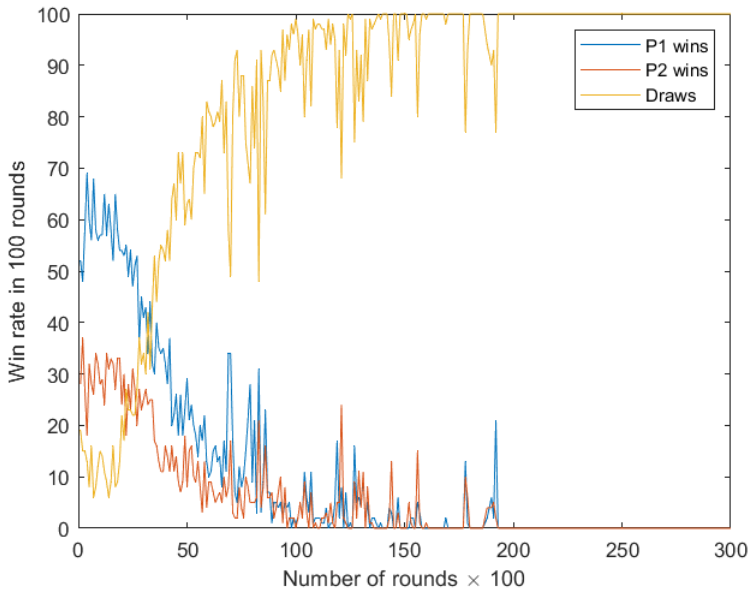

Figure 1: Winning rate respect to each 100 rounds

Since player 1 always start first, it's natural he will have a larger win rate at the beginning. The ϵ-greedy policy with $\epsilon = 0.95$ in the first 3000 rounds makes the plot slightly different from other's result, that non of the players plays with best solution before 3000 rounds. However it encourages the players to explore as much as possible to avoid keeping exploiting a local optimal strategy. Otherwise the final Q-table will not cover enough states, which will result in a failure.

As the training continues, both players gradually choose optimal action from the Q-table. Thus the wining rate decays and the draw rate increases until it's no longer possible for any player to win.

With the final Q-table, the agent will always find an optimal action to win, or at least draw, even when playing against human.