COMP 4900 Project

John Oommen

Abdulaah Emin

101130854

Q1.

I have implemented both schemes (L_{RI} and Tsetlin), my implementation is as follows:
Tsetlin:

Initialize n automata, they will independently choose an action based on their current state, they will then independently either receive a reward or a penalty based on the number of automata choosing action 1 (yes), each automaton will then update their state.

L_{RI}:

Initialize n automata, they will independently choose an action based on their current probability vector, they will then independently either receive a reward or a penalty based on the number of automata choosing action 1 (yes), each automaton will then update their probability vector ignoring penalties and updating on rewards.

I tested both schemes with 10 players (automatons) and changing θ^* to a different value (i.e., 5,8)

Using ensemble average, both schemes would have the correct number of players converge to "yes" with 95% accuracy.

Q2.

If g is a bi-modal function, then the Automata may converge to one of the two θ^*, depending on their initial states and the outcome of the first few rounds. To avoid this, once an automaton converges, we can have the automata remember its convergence and then have it explore the other action with some randomness. That way all the players (automatons) eventually converges to the same θ^* which will minimize the overall penalties.

Q3.

We can have both automatons play against each other in the same environment, for example have automata 1 play 100 games, and have automata 2 play 100 games, compare accuracies and convergence speed.

Q4.

DL_{RI} can be used for this context, since L_{RI} converges already, by making it discretized we are guaranteed convergences as well but at a faster speed.

Instead of multiplying a probability by λ, we instead divide the possible values of the probability vector, i.e. [0 , 0.25 , 0.5 , 0.75 , 1], this way reaching 1 is possible, and we move in steps.

We can use an estimator scheme that keeps track of previous action/reward combinations, that way we to estimate the value of theta based on the past votes and rewards and use the estimate to choose the action that is expected to maximize the referee's performance criterion.