A Comparison of SoftMax and ϵ-greedy Exploration Methods

In this project work your task is to implement a simple Q-learning agent exploring its environment in a 20×20 maze. In the maze, there exists a goal cell producing a large, say e.g. $+100.0$, reward value. The agent can move to the adjacent cells (4-neighborhood) and it should stay on the grid. It is enough to implement a system by using a tabular version of the Q-learning algorithm.

Let the agent to start its exploration from different, randomly selected, cells and implement two different exploration strategies: ϵ-greedy exploration and SoftMax action selection. In the SoftMax action selection the probability for each action $a \in A(s)$ in a state s is assigned in the following way:

$$P(a|s) = \frac{e^{\alpha Q(s,a)}}{\sum_{b \in A(s)} e^{\alpha Q(s,b)}}$$

In the equation, α coefficient controls the "greediness" of the action selection procedure: if $\alpha = 0$, action selection is purely random and if $\alpha \to \infty$ it approaches greedy action selection. In ϵ-greedy action selection an action with the highest Q-value is selected with the probability $1 - \epsilon$ and a random action is selected with the probability ϵ.

In this project work, let free parameters in both exploration strategies vary linearly during the learning process so that at the beginning exploration is almost random and at the end it is almost greedy. Are there differences in the convergence speed? How about the coverage of the exploration; which percentage of all states is actually visited during the learning?