Documentation of Q-learning tests on a toy problem

Corridor

This environment is a generalization of an environment called riverswim.

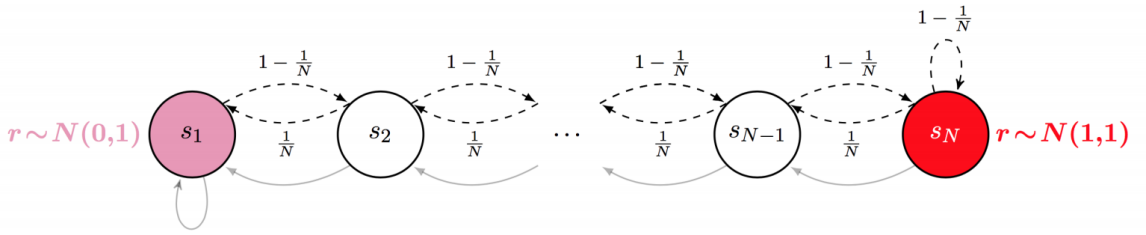

Figure 1: Chain Env

The corridor environment allows one to define:

- **N:** The environment consists of a N long corridor where the agent can move left or right.
- **K:** in a K number of states, the direction traveled when given is the opposite of the expected. I.e. action left will cause the agent to move right.
- **p:** Probability of success when moving right. 1-**p** probability of moving left instead.

Experiment setup

The current setup gives 1 reward for reaching the final node to the right, and 0 otherwise. The agent has N steps to try to reach the goal. The experiment is run with K=0 and p=1 to allow a simple linear model and simple calculation of the optimal reward.

The agent is then tested in the environment multiple times for an increasing value of N. The attempt is stopped when the agent reaches a running average of regret per episode that is lower than 1% of the optimal reward. When this happens we consider that the agent has "learned" the environment.

Experiment pseudocode

Agent

The agent is a vanilla Q-learning algorithm.

Critics

For each Critic pseudocode is provided for the key functionality of the critic.

e-greedy

Uses a linear regression method using SGD for parameter optimization.

Sample Target UBE Critic

Note that this implementation does not directly propagate the local uncertainty.

The action varaince calculation and update are best shown through their equations

Action Varaince

return state*sigma[action]*state

Update Action Varaince

```
\textbf{test}
test
```