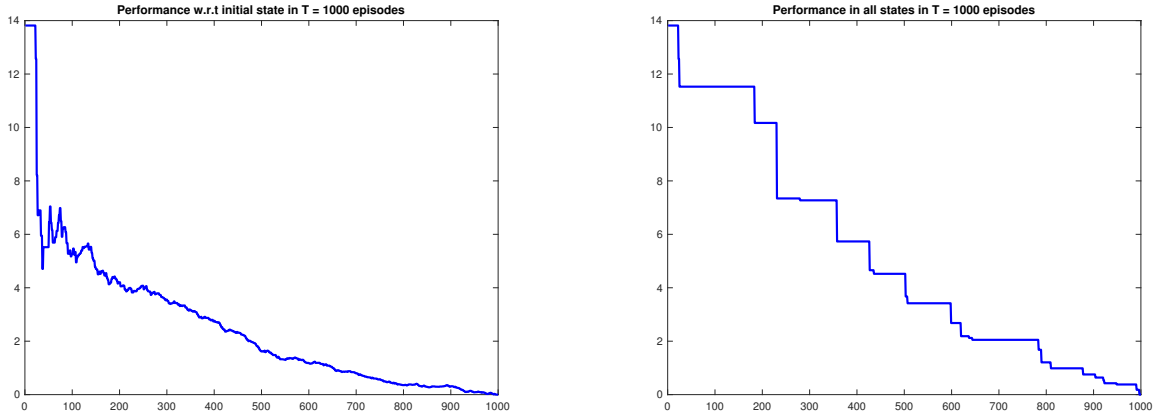
Master MVA
Dynamic Programming & Reinforcement Learning
TP2 - 27/11/2016
Achari Berrada Youssef

1 Q-Learning :

Q1: The exploration policy is the ϵ-greedy policy. I choose $\epsilon = 0.4$ so that we explore all the action-states. With probability $1 - \epsilon$, we choose $a_t = \text{argmax}\, Q(x_t, a)$, and with probability ϵ, we choose a random action. This exploration policy allows us to explore a large set of state-actions in order to satisfy the stochastic approximation requirement. I choose the learning rate to be :

$$\alpha(x, a) = \frac{1}{N(x, a)}$$

Performance evolution over $T = 1000$ episodes:



(a) Performance in state 1.

(b) Error Propagation.

Figure 1: Performance of the Q-Learning algorithm.

2 Stochastic Multi-Armed Bandits on simulated Data

Q2: