Reinforcement Learning Lecture: Homework 12

Ngo Anh Vien

MLR, University of Stuttgart

June 30, 2016

1 PoWER Implementation

The PoWER algorithm is described in slide 54. Here are some special setting for the cart-pole domain as in previous exercises (use similar setting, except the following points):

- $\epsilon \sim \mathcal{N}(0, \Sigma)$, use $\Sigma = \text{diag}(0.5, 0.5, 0.5, 0.5)$. Because of this, we can simplify W_t as

$$W_t = \phi(s_t)\phi(s_t)^\top / \big(\phi(s_t)^\top \phi(s_t)\big)$$

- use an unbiased sample of $Q(s, a)$. For each trajectory i,

$$Q(s_t^{[i]}, a_t^{[i]}) = \sum_{l=t}^{T_i} \gamma^{l-t} r_l^{[i]}$$

- each update uses 20 episodes (you can increase it to get a smoother plot)
- the maximum length of each episode is 5000
- $\gamma = 1.0$

Report the result over 300 updates of θ.