Problem Set 8 Cristian Rodriguez

8a) ε-greedy algorithm: policy = $\begin{cases} \text{explore} & \text{w.p. } \varepsilon \\ \text{exploit} & \text{otherwise} \end{cases}$

$$\rho(T) = \sum_{t=0}^{T-1} \left(r_{max}(t) - r(t) \right)$$

$P(r(t) \neq r_{max}(t)) = \frac{N-1}{N}$ with N = num slots.

average difference between $r(t)$ and $r_{max}(t)$ is $P_{max} - \frac{1}{N-1} \sum_{i=1}^{N-1} P_i$ and we explore εT times so.

$$\lim_{T\to\infty} \mathbb{E}[\rho(t)] = \lim_{T\to\infty} \left(\frac{N-1}{N} \cdot \left(P_{max} - \frac{1}{N-1} \sum_{i=1}^{N-1} P_i \right) \cdot \varepsilon \right)$$

$$= \frac{N-1}{N} \cdot \varepsilon \cdot \left(P_{max} - \frac{1}{N-1} \sum_{i=1}^{N-1} P_i \right)$$

$\Rightarrow$ To decrease this, we can increase epsilon which will allow us to explore faster and find the optimal slot quicker.

8b) In i), we are approximating the Q function with θ. The assignment in ii) is meant to achieve convergence, such that comparing Q to Q_θ has a significance. If we did not do it after several iterations, Q would be very similar to Q_θ.

8c) We may want to do this to verify the gradient descent approach. If the gradient descent is well-implemented, it should converge to the same spot alway. We may also find more local minimas.