AV499 & AVD871 – Tutorial questions for final exam

(covering the Reinforcement Learning portion of the syllabus)

1. The following questions from Sutton and Barto are from the 2nd edition of the text book, which is available online.
   a) Chapter 3 – Exercises 3.6, 3.7, 3.8, 3.9, 3.10
   b) Chapter 5 – Exercise 5.5, 5.9
   c) Chapter 6 – Example 6.5, Exercise 6.9
2. Review the pseudo code for the algorithms discussed in the class, including but not restricted to the following
   a) Value iteration
   b) Policy iteration
   c) Generalized policy iteration
   d) Monte carlo first visit, every visit
   e) Monte carlo on-policy and off-policy
   f) TD(0) estimation, SARSA(0) and Q-learning
   g) n-step TD estimation
   h) TD(lambda) and implementation using eligibility traces
   i) SARSA(lambda)
   j) Value function approximation methods
   k) Policy gradient and REINFORCE
3. Do all the proofs for the results discussed in the class, including but not restricted to the following
   a) Policy improvement
   b) epsilon-soft policy improvement
   c) Policy gradient theorem
4. Suppose $f(x) = x^2$. Note that the minimum value of $f(x)$ is achieved at $x^* = 0$. Note that x^* can be obtained using the following gradient descent approach:

$$x[k+1] = x[k] - \mu \frac{df(x)}{dx}\bigg|_{x=x[k]}, k \geq 1.$$

We expect that $\lim_{k\to\infty} x[k] = x^*$. Derive the number of steps N after which $x[k]$ would be in the interval $[x^* - \epsilon, x^* + \epsilon]$, i.e., $x[k] \in [x^* - \epsilon, x^* + \epsilon]$ for all $k \geq N$. Note that μ and ϵ are real positive numbers. Sometimes the gradient approach uses a computationally computed derivative, which is influenced by noise. Assume that we have a gradient descent approach where

$$x[k+1] = x[k] - \mu \frac{df(x)}{dx}\bigg|_{x=x[k]} + W[k], k \geq 1.$$

Here $W[k], k \in \{1, 2, 3, \dots\}$ is a set of IID Gaussian random variables with mean 0 and variance σ^2. Derive the probability that $x[N] \in [x^* - \epsilon, x^* + \epsilon]$ for the N that you have derived before.