Assignment #5 Reinforcement Learning

(note: all the answers should be typed in MS-word or Latex and the pdf file is submitted. No handwritten answers are accepted.)

1- The first episode of an agent interacting with an environment under policy π is as follows:

Timestep	Reward	State	Action
0		X	U1
1	16	X	U2
2	12	X	U1
3	24	X	U1
4	16	T	

Assume discount factor, γ=0.5, step size $\alpha = 0.1$ and q_π is initially zero.
What are the estimates of $q_\pi\,(X, U1)$ and $q_\pi\,(X, U2)$ using 2-step SARSA?

2- What is the purpose of introducing Control Variates in per-decision importance sampling?

3- In off-policy learning, what are the pros and cons of the Tree-Backup algorithm versus off-policy SARSA (comment on the complexity, exploration, variance, and bias, and others)?

4- Exercise 7.4 of the textbook (page 148).