Assignment Set #3

Problem 1 (30)
Consider the learning algorithm that is just like Q-learning except that instead of the maximum over next state-action pairs it uses the expected value, taking into account how likely each action is under the current policy. That is, consider the algorithm otherwise like Q-learning except with the update rule

\[ Q(s_t, a_t) = Q(s_t, a_t) + \alpha [r_{t+1} + \gamma E[Q(s_{t+1}, a_{t+1})|s_t] - Q(s_t, a_t)] \]

\[ = Q(s_t, a_t) + \alpha \left[ r_{t+1} + \gamma \sum_a \pi(s, a)Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \]

Answer the following:
1. Is this new method an on-policy or off-policy method?
2. What is the backup diagram for this algorithm?
3. Given the same amount of experience, would you expect this method to work better or worse than SARSA?
4. What are the practical considerations when implementing this algorithm compared to SARSA?

Problem 2 (20)
On slide 14 in lecture notes #10, the RMS curves for TD-learning appear to be decreasing and then increasing. Suggest an intuitive explanation for this behavior.

Problem 3 (20)
Refer to lecture notes #11, slide 13. Why do you think on-line methods performed better than off-line methods on the random-walk example? Why is it that in both cases the optimal value for \( \alpha \) is neither 0 nor 1?

Problem 4 (30)
The Bellman equation describes the relationship between the values of subsequent states for a given policy. It may be expressed as
\[ V^\pi(s) = E_{s'} \left[ p(s,a) + \gamma V^\pi(s') \right] = r(s,a) + \gamma \sum_{s'} p(s'|s,a)V^\pi(s') \quad \forall s \in S \]

Let us label the states \( s = 1,2,\ldots, N \), for which the Bellman equation is a set of linear relationships. We further use the following labels:

\[
V^\pi = [V^\pi(1), V^\pi(2), \ldots, V^\pi(N)]^T
\]

\[
r^\pi = [r(1, \pi(1)), r(2, \pi(2)), \ldots, r(N, \pi(N))]^T
\]

\[
P^\pi = \begin{bmatrix}
 p_{11}^\pi & p_{12}^\pi & \cdots & p_{1N}^\pi \\
 p_{21}^\pi & p_{22}^\pi & \cdots & p_{2N}^\pi \\
 \vdots & \vdots & \ddots & \vdots \\
 p_{N1}^\pi & p_{N2}^\pi & \cdots & p_{NN}^\pi
\end{bmatrix}
\]

\[
p_{ij}^\pi = p(s_j | s_i, a = \pi(s_i))
\]

Show that when \((I - \gamma P^\pi)\) is invertible, the value function is given by

\[
V^\pi = (I - \gamma P^\pi)^{-1} r^\pi
\]