CS 234 Midterm - Winter 2017-18

**Do not turn this page until you are instructed to do so.

Instructions

Please answer the following questions to the best of your ability. Read all the questions first before answering. You have **80** minutes to complete the exam. The exam is closed-book, closed-note and closed-internet. Additionally use of all electronic items during the exam is prohibited. However, you may use one letter-sized sheet (front and back) of notes as reference. All of the intended answers can be written well within the space provided. Good luck!

Stanford University Honor Code

The following is a statement of the Stanford University Honor Code:

1. The Honor Code is an undertaking of the students, individually and collectively: (1) that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading; (2) that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.
2. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.
3. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

By signing your name below, you acknowledge that you have abided by the Stanford Honor Code while taking this exam.

Signature:

Name:

SUNet ID:

Grading (For Midterm Graders Only)

Question #	1	2	3	4	5	6
Maximum Points	15	14	10	10	8	15
Student Grade						

Question 1 – Warm-Up [15pts]

A)

1. True
2. True
3. True
4. True
5. False
6. False
7. False

B)

1. The number of deterministic policies are $|A|^{|S|}$. The number of stochastic policies is uncountable.
2. No, REINFORCE will not necessarily converge to an optimal policy, even if the policy class can represent the optimal policy. It is only guaranteed to converge to a local minimum. In special cases, when the objective function is convex, i.e local minimum is the global minimum then one can guarantee convergence to an optimal policy.
3. SARSA for finite-state and finite-action MDPs will converge to the optimal function value when the following conditions are satisfied:
   - The policy sequence satisfies the conditions of GLIE (greedy in the limit of infinite exploration).
   - The step-sizes satisfy the Robbins-Munro conditions.
4. Maximizing the entropy of the distribution over the paths subject to the feature constraints from observed data implies we maximize the likelihood of the observed data under the maximum entropy (exponential family) distribution.

$$P(\tau_j|w) = \frac{1}{Z(w)} \exp(w^T \mu_{\tau_j}) = \frac{1}{Z(w)} \exp\left(\sum_{s_i \in \tau_j} w^T \mu_{\tau_j} x(s_i) \right) \tag{1}$$

The optimization problem being

$$\begin{aligned} &\max_P - \sum_\tau P(\tau) \log P(\tau) \\ &\text{s.t } \sum_\tau P(\tau) = 1 \quad . \end{aligned} \tag{2}$$

Question 2 – Commuting from Stanford [14 pts]

a) San Jose
b) $\pi^{ast}(S_1) = bus$, $\pi^{ast}(S_2) = bus$, $\pi^{ast}(S_3) = train$.
c) Yes
d) The optimal policy is to always take the bus. It differs because the MDP defined in c) is a non-Markovian representation of the actual MDP (in s_{12} there is a dependence on the last action taken)
e) Yes in the limit of infinite data. Monte Carlo does not rely on the Markov assumption.

Question 3 – Value Iteration [10 pts]

a) Consider a MDP $M(S, A, P, R, \gamma)$ with 2 states $S = \{S_1, S_2\}$. From each state there are 2 available actions $A = \{stay, go\}$. Choosing *"stay"* from any state leaves you in the same state and gives reward -1. Choosing *"go"* from state S_1 takes you to state S_2 deterministically giving reward -2, while choosing *"go"* from state S_2 ends the episode giving reward 3.

Let us initialize value iteration as $V_0 = [0, 0]$. Then $V_1 = [-1, 3]$ and $V_2 = [1, 3]$. We also have $V^* = [1, 3]$. Thus for state S_1, convergence is clearly not monotonic.

b) The Bellman operator is a contraction. This means that the maximum absolute value across all states of the error (difference of the value function compared to the optimal value function) must not increase between iterations of value iteration. However, for individual states it is possible that the error increases between iterations. In this example, $||V_0 - V^*||_\infty = 3$, $||V_1 - V^*||_\infty = 2$ and $||V_2 - V^*||_\infty = 0$, so clearly there is no contradiction.

Question 4 – MC Update vs TD Update [15 pts]

A) Monte Carlo update
a)

$$Q(s,a) \leftarrow Q(s,a) + \alpha(G_t - Q(s,a))$$

Here G_t is discounted sum of returns

b) G_t is sampled
c) $O(T)$ where T is the length of the episode

B) Temporal Difference update
a)

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha(r_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t))$$

b) $Q(s_{t+1},a_{t+1})$ is bootstrapping

Question 5 – Value Function Approximation [10 pts]

A) No (Depends) it does not imply that an optimal policy has been found.

First let us consider the case when linear approximation is not able to represent the value function with our chosen set of features. In that case even if we have found the global minimum of the function that is represented by our function approximator, there is no guarantee that the extracted greedy policy corresponding to the current Q will be an optimal policy.

Secondly let us consider the case when the linear approximator is able to represent the value function with our chosen set of features. Then it may be the case that the Q value is still changing with iterations and hence the extracted greedy policy will not in general be an optimal policy, as it can change in future iterations. It may also be the case that we have converged to a local minimum and in that case although the extracted greedy policy will not change with future iterations, it may not be an optimal policy. The only case where the extracted greedy policy is guaranteed to be an optimal policy is if we have converged to a global minimum.

B) I'd pick **D10**. First notice that **S100** and **D10** have the same number of parameters (in an order of magnitude sense). By Hornik's universal approximation theorem, we know that a single hidden layer can represent any continuous function. However the number of nodes needed can be extremely large. On the other hand, it has been observed that to represent the same function class one can reduce the number of nodes in each hidden layer while simultaneously increasing the number of hidden layers. Thus multiple hidden layers provide a more compact representation of the value function. This motivates our choice for **D10**, since the number of parameters in **S100** and **D10** are the same, and so we expect to be able to represent a larger function class with **D10**.

The number of trainable parameters for each network are the same order of magnitude for S100 and D10, but are two orders of magnitude lower for S2.

Question 6 – Alice in Wonderland [25pts]

A) $Q_1^*(s_1, \pi^*(s_1, 1)) - Q_1^*(s_1, \pi(s_1, 1))$

B) Relating the Value Function to the Q-Values

Let Q^π be the Q-values of policy π followed by Alice. Then

$$\begin{aligned}
V_1^*(s_1) - V_1^\pi(s_1) &= Q_1^*(s_1, \pi^*(s_1, 1)) - Q_1^\pi(s_1, \pi(s_1, 1)) && (3)\\
&= Q_1^*(s_1, \pi^*(s_1, 1)) - Q_1^*(s_1, \pi(s_1, 1)) + Q_1^*(s_1, \pi(s_1, 1)) - Q_1^\pi(s_1, \pi(s_1, 1)) && (4)\\
&= Q_1^*(s_1, \pi^*(s_1, 1)) - Q_1^*(s_1, \pi(s_1, 1)) + V_2^*(s_2) - V_2^\pi(s_2) && (5)\\
&\overset{(\star)}{=} Q_1^*(s_1, \pi^*(s_1, 1)) - Q_1^*(s_1, \pi(s_1, 1)) + \sum_{t=2}^{H} Q_t^*(s_t, \pi^*(s_t, t)) - Q_t^*(s_t, \pi(s_t, t)) && (6)\\
&= \sum_{t=1}^{H} Q_t^*(s_t, \pi^*(s_t, t)) - Q_t^*(s_t, \pi(s_t, t)) && (7)
\end{aligned}$$

where $(\star)$ follows from the inductive hypothesis.

C) Loss Due to Inaccurate Q-values

$$\begin{aligned}
V_1^*(s_1) - V_1^{\tilde{\pi}}(s_1) &= \sum_{t=1}^{H} Q_t^*(s_t, \pi^*(s_t,t)) - Q_t^*(s_t, \tilde{\pi}(s_t,t)) \\
&= \sum_{t=1}^{H} Q_t^*(s_t, \pi^*(s_t,t)) - \tilde{Q}_t(s_t, \pi^*(s_t,t)) + \tilde{Q}_t(s_t, \pi^*(s_t,t)) - Q_t^*(s_t, \tilde{\pi}(s_t,t)) \\
&\leq \sum_{t=1}^{H} Q_t^*(s_t, \pi^*(s_t,t)) - \tilde{Q}_t(s_t, \pi^*(s_t,t)) + \tilde{Q}_t(s_t, \tilde{\pi}(s_t,t)) - Q_t^*(s_t, \tilde{\pi}(s_t,t)) \\
&\leq \sum_{t=1}^{H} (\epsilon + \epsilon) = 2\epsilon H
\end{aligned}$$