**Mahindra University Hyderabad**
**École Centrale School of Engineering**
**Mock Paper for Minor-II**
**Subject: Reinforcement Learning and Autonomous Systems (CS 4122)**

**Time Duration: 1 hour 30 minutes** **Max. Marks: 30**

***Instructions:***

1) The submitted answer sheet should only contain your final answer and the **solution procedure**. **Use separate sheets for rough work**.
2) The paper has a sum total of **40 marks**. However, you can score a maximum of **30 marks**. Any **bonus marks** you score will be included in the next exam.

**Q1: A Variant of Double Q-Learning (10 marks, Difficulty: EASY)**

In Double Q-Learning we keep two estimates of the Q-function. But, the same concept of double Q-learning can be applied to the case when we keep three estimates of the Q-function. Let's call this *Triple Q-Learning*. Write a *neat psuedocode* for Triple Q-Learning. **I only need the psuedocode. You will loose points if you give any other explanation**.

**Q2: Bandit Setup (5 marks + 10 marks)**

Answer the following questions:

**(a) (Difficulty: EASY):** In any learning setup, any action should have a *finite probability* (however small that probability may be) of getting choosen at any time slot. How does policy gradient for contextual bandits ensures that this criteria is met? **Your answer must not exceed ¼ of a page.**

**(b) (Difficulty: HARD):** Consider an episodic RL setup where an episode always lasts for only two time slots. This RL setup is characterized by:

- State space $S$.
- Action space $A$. Action space does not change with the state.
- Average reward $r(x)$ where $x$ is the state. *Average reward does not depend on the action*.
- State transition probability is $P[x'|x, a]$ where $x$ and $a$ are the current state and action and $x'$ is the next state.

Your task is to *convert this RL setup into an EQUIVALENT contextual bandit setup* so that we can apply algorithms for contextual bandits. Answer the following questions related to this equivalent contextual bandit setup:

**(a)** What is the context space? **(2 marks)**

**(b)** What is the action space? **(2 marks)**

**(c)** What is the average reward for a context-action pair assuming that for the RL setup, the objective was to minimize the $\beta$-discounted reward? Your answer should be in terms of $S$, $A$, $r(x)$, and $P[x'|x, a]$. **(6 marks)**

**Q3: MDP to Speed-Up Airport Immigration Process (15 marks, Difficulty: MODERATE if you did programming assignment 2)**

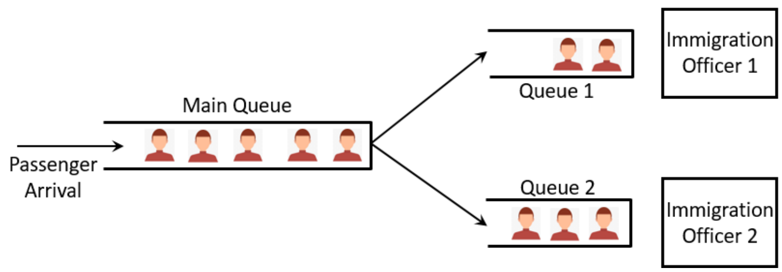


**Figure 2**

Figure 2 shows a queueing system for immigration check in an airport. It consists of a main queue where passengers arrive. At any time slot, a maximum of $M$ passengers can arrive. The probability that $m \in \{0,1,\cdots,M\}$ passengers arrive in a time slot is $p_m$. There are two immigration officers indexed 1 and 2. Each of the immigration officer has its own queue shown as queue 1 and queue 2 in Figure 2. A scheduler has to decide whether to send the passenger in the front of the main queue to either queue 1 and queue 2. *This decision once made can't be retracted*.

The immigration officer of queues 1 and 2 check the details of the passenger in the front of their respective queue. Officer $i$, where $i = 1,2$, will at least take $\tau_i$ time slots to finish checking a passenger. After $\tau_i$ time slots are over, the probability that officer $i$ will finish checking the passenger in a time slot is $\theta_i$.

Your objective is to minimize the discounted cost of the number of people in the three queues. Answer the following questions:

**(a)** What is the state and state space for this problem? **(4 marks)**

**(b)** What is the action and action space for this problem? **(2 marks)**

**(c)** Define the reward for all state-action pair for this problem? **(2 marks)**

**(d)** What is the Bellman optimality equation for this problem? **(7 marks)**