E1 222 Stochastic Models and Applications
Problem Sheet 3.6

1. Suppose you have to play the following game. You are going to be shown N prizes in sequence. At any time you can either accept the one that is being offered or reject it and choose to see the next prize. Once you reject a prize you cannot go back to it. At any time you are seeing a prize, all the information you have is the relative rank of the prize that you are being offered, with respect to all the ones that have gone by. That is, when you are seeing the third prize, you know how it ranks with respect to the first and second ones that you have already seen and rejected. Consider the following strategy. You fix an integer k between 1 and N. You reject the first k prizes and then accept the first one that you see which is better than all the ones rejected till that point. (If after the first k, in the remaining $N-k$ chances, you never see a prize that is better than all the ones you had rejected till then, then you would end up rejecting all the prizes). Assuming that all possible orderings of the N prizes are equally likely, calculate the probability that this strategy would get you the best prize. Based on this, suggest what is a good value of k to choose.

2. Consider a gambling machine which has two arms and we play by choosing one of the arms. Each time we choose an arm we get either a reward or a penalty. (Reward may be that you get back double the money you bet while penalty may be that you loose the money. Such a machine or model is called a two-armed bandit). These outcomes are stochastic. When we choose the first arm, we get reward with probability d_1 and penalty with probability $1-d_1$. Similarly, for the second arm we get reward with probability d_2 and penalty with probability $1-d_2$. However, we have no knowledge of the numbers d_1 and d_2 . We say that the first arm is optimal if $d_1 > d_2$. (If $d_2 > d_1$ we say second arm is optimal). We want a strategy of repetedly playing with the machine to find which arm is optimal. Consider a possible iterative method as follows. At iteration k, we choose first arm with probability $p(k)$ and second arm with probability $1-p(k)$. (We start with p(0) = 0.5). Define $b(k)$ as $b(k) = 1$ if our choice at k^{th} iteration resulted in reward and $b(k) = 0$ if our choice resulted in penalty. Now we change $p(k)$ as

follows:

$$\begin{aligned} p(k+1) &= p(k) + \lambda(1 - p(k)) \text{ if arm 1 is chosen and } b(k) = 1 \\ &= p(k) - \lambda p(k) \text{ if arm 2 is chosen and } b(k) = 1 \\ &= p(k) \text{ if } b(k) = 0 \end{aligned}$$

where $0 < \lambda < 1$ is a constant (known as the learning step size). The above algorithm has simple intuitive motivation. If you played arm-1 and got a reward, you are increasing $p(k)$ so that the probability of choosing arm-1 at $k+1$ is increased. If you plated arm-2 and got a reward then you are decreasing $p(k)$ so that probability of playing arm-2 at $k+1$ is increased. When you get a penalty you do nothing (because you have decided that you learn only from rewards). Since $p(k)$ has to be between 0 and 1, when you increase it, the increment is made proportional to $(1-p(k))$ and when you decrease it, decrement is made proportional to $p(k)$. Note that $\{p(k), k = 0, 1, \cdots\}$ is a sequence of random variables because by the above iterative algorithm, $p(k+1)$ depends on the random choice of arm at k and random value $b(k)$. Calculate $E[p(k+1)|p(k)]$. Use this to show that if $d_1 > d_2$ then $E[p(k+1)] > E[p(k)], \ \forall k$.

3. Consider a TV reality show like this. There are three closed doors. Behind two of them there is a goat while behind the third one there is a new car. The contestant has to choose one of the three doors. She/he gets the car for free if she/he chooses the door behind which there is a car. Now, after the contestant chooses a door, the host of the show opens one of the other two doors. There is a goat behind that door. Now the host offers the contestant a chance to change her/his selection. The question is should the contestant change her/his selection.

   So, we want to ask does the probability of winning increase or decrease or remains the same if the contestant changes her/his selection. For this we need to model the situation correctly. For the contestant we assume she/he chooses randomly among the three doors. That is a reasonable assumption. What about the process by which the host selects one of the two doors? Let us consider two cases: (a.) In the first we assume the host also chooses randomly from the other two doors. (b). In the other case we assume that the host would only choose a door behind which there is a goat. Thus, if there is only one of the two remaining

doors behind which there is a goat, the host chooses that. If there is a goat behind both the remaining doors then the host chooses one of them at random. In each of these two cases suggest what is a good strategy for the contestant.