

## Assignment 2: CS 215

Due: 2nd September before 11:55 pm

**All members of the group should work on all parts of the assignment. Copying across groups or from other sources is not allowed. We will adopt a zero-tolerance policy against any violation.**

### Submission instructions:

1. You should type out a report containing all the answers to the written problems in Word (with the equation editor) or using Latex, or write it neatly on paper and scan it. In either case, prepare a single pdf file.
2. The report should contain names and roll numbers of all group members on the first page as a header.
3. Put the pdf file and the code for the programming parts all in one zip file. The pdf should contain the names and ID numbers of all students in the group within the header. The pdf file should also contain instructions for running your code. Name the zip file as follows: A2-IdNumberOfFirstStudent-IdNumberOfSecondStudent-IdNumberofThirdStudent.zip. (If you are doing the assignment alone, the name of the zip file is A2-IdNumber.zip, if there are two students it should be A2-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip).
4. Upload the file on moodle BEFORE 11:55 pm on the due date. We will nevertheless allow and not penalize any submission until 10:00 am on the following day (i.e. 3rd September). No assignments will be accepted thereafter.
5. Note that only one student per group should upload their work on moodle, though all group members will receive grades.
6. Please preserve a copy of all your work until the end of the semester.

### Questions:

1. Let  $X_1, X_2, \dots, X_n$  be  $n > 0$  independent identically distributed random variables with cdf  $F_X(x)$  and pdf  $f_X(x) = F'_X(x)$ . Derive an expression for the cdf and pdf of  $Y_1 = \max(X_1, X_2, \dots, X_n)$  and  $Y_2 = \min(X_1, X_2, \dots, X_n)$  in terms of  $F_X(x)$ . [10 points]
2. We say that a random variable  $X$  belongs to a Gaussian mixture model (GMM) if  $X \sim \sum_{i=1}^K p_i \mathcal{N}(\mu_i, \sigma_i^2)$  where  $p_i$  is the ‘mixing probability’ for each of the  $K$  constituent Gaussians, with  $\sum_{i=1}^K p_i = 1; \forall i, 0 \leq p_i \leq 1$ . To draw a sample from a GMM, we do the following: (1) One of the  $K$  Gaussians is randomly chosen as per the PMF  $\{p_1, p_2, \dots, p_K\}$  (thus, a Gaussian with a higher mixing probability has a higher chance of being picked). (2) Let the index of the chosen Gaussian be (say)  $m$ . Then, you draw the value from  $\mathcal{N}(\mu_m, \sigma_m^2)$ . If  $X$  belongs to a GMM as defined here, obtain expressions for  $E(X)$ ,  $\text{Var}(X)$  and the MGF of  $X$ . Now consider a random variable of the form  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$  for each  $i \in \{1, 2, \dots, K\}$ . Define another random variable  $Z = \sum_{i=1}^K p_i X_i$  **where  $\{X_i\}_{i=1}^K$  are independent random variables**. Derive an expression for  $E(Z)$ ,  $\text{Var}(Z)$  and the PDF, MGF of  $Z$ . [2+2+2+2+2+2+3=15 points]
3. Using Markov’s inequality, prove the following one-sided version of Chebyshev’s inequality for random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$ :  $P(X - \mu \geq \tau) \leq \frac{\sigma^2}{\sigma^2 + \tau^2}$  if  $\tau > 0$ , and  $P(X - \mu \geq \tau) \geq 1 - \frac{\sigma^2}{\sigma^2 + \tau^2}$  if  $\tau < 0$ . [15 points]

4. Given stuff you've learned in class, prove the following bounds:  $P(X \geq x) \leq e^{-tx} \phi_X(t)$  for  $t > 0$ , and  $P(X \leq x) \leq e^{-tx} \phi_X(t)$  for  $t < 0$ . Here  $\phi_X(t)$  represents the MGF of random variable  $X$  for parameter  $t$ . Now consider that  $X$  denotes the sum of  $n$  independent Bernoulli random variables  $X_1, X_2, \dots, X_n$  where  $E(X_i) = p_i$ . Let  $\mu = \sum_{i=1}^n p_i$ . Then show that  $P(X > (1 + \delta)\mu) \leq \frac{e^{\mu(e^t - 1)}}{e^{(1+\delta)t\mu}}$  for any  $t \geq 0, \delta > 0$ . You may use the inequality  $1 + x \leq e^x$ . Further show how to tighten this bound by choosing an optimal value of  $t$ . [15 points]
5. Consider  $N$  independent random variables  $X_1, X_2, \dots, X_N$ , such that each variable  $X_i$  takes on the values 1, 2, 3, 4, 5 with probability 0.05, 0.4, 0.15, 0.3, 0.1 respectively. For different values of  $N \in \{5, 10, 20, 50, 100, 200, 500, 1000, 5000, 10000\}$ , do as follows:
- Plot the (empirically determined) distribution of the average of these random variables ( $X_{avg}^{(N)} = \sum_{i=1}^N X_i / N$ ) in the form of a histogram with 50 bins.
  - Empirically determine the CDF of  $X_{avg}^{(N)}$  using the `ecdf` command of MATLAB (this is called the empirical CDF). On a separate figure, plot the empirical CDF. On this, overlay the CDF of a Gaussian having the same mean and variance as  $X_{avg}^{(N)}$ . To get the CDF of the Gaussian, use the `normcdf` function of MATLAB.
  - Let  $E^{(N)}$  denote the empirical CDF and  $\Phi^{(N)}$  denote the Gaussian CDF. Compute the maximum absolute difference (MAD) between  $E^{(N)}(x)$  and  $\Phi^{(N)}(x)$  numerically, at all values  $x$  returned by `ecdf`. For this, read the documentation of `ecdf` carefully. Plot a graph of MAD as a function of  $N$ . [3+3+4 = 10 points]
6. Read in the images T1.jpg and T2.jpg from the homework folder using the MATLAB function `imread` and cast them as a double array using the code

```
im = double(imread('T1.jpg'));
```

These are magnetic resonance images of a portion of the human brain, acquired with different settings of the MRI machine. They both represent the same anatomical structures and are perfectly aligned (i.e. any pixel at location  $(x, y)$  in both images represents the exact same physical entity). Consider random variables  $I_1, I_2$  which denote the pixel intensities from the two images respectively. Write a piece of MATLAB code to shift the second image along the X direction by  $t_x$  pixels where  $t_x$  is an integer ranging from -10 to +10. While doing so, assign a value of 0 to unoccupied pixels. For each shift, compute the following measures of dependence between the first image and the *shifted version* of the second image:

- the correlation coefficient  $\rho$ ,
- a measure of dependence called quadratic mutual information (QMI) defined as  $\sum_{i_1} \sum_{i_2} (p_{I_1 I_2}(i_1, i_2) - p_{I_1}(i_1)p_{I_2}(i_2))^2$ , where  $p_{I_1 I_2}(i_1, i_2)$  represents the *normalized* joint histogram (i.e., joint pmf) of  $I_1$  and  $I_2$  ('normalized' means that the entries sum up to one).

For computing the joint histogram, use a bin-width of 10 in both  $I_1$  and  $I_2$ . For computing the marginal histogram, you need to integrate the joint histogram along one of the two directions respectively. You should write your own joint histogram routine in MATLAB - do not use any inbuilt functions for it. Plot a graph of the values of  $\rho$  versus  $t_x$ , and another graph of the values of QMI versus  $t_x$ .

Repeat exactly the same steps when the second image is a negative of the first image, i.e.  $I_2 = 255 - I_1$ .

Comment on all the plots. In particular, what do you observe regarding the relationship between the dependence measures and the alignment between the two images? Your report should contain all four plots labelled properly, and the comments on them as mentioned before. [25 points]

7. Derive the covariance matrix of a multinomial distribution using moment generating functions. You are not allowed to use any other method. Since a covariance matrix  $\mathbf{C}$  is square and symmetric, it is enough to derive expression for the diagonal elements  $C_{ii}$  and the off-diagonal elements  $C_{ij}, i \neq j$ . [10 points]