Assignment 4
CMPT 210

Due: In class on 2 August

(1) **[35 marks]** Suppose n balls are thrown randomly into m boxes. Each ball lands in each box with uniform probability. Define X_i be the r.v. equal to the number of balls that land in box i.

- What is the distribution of X_i? Compute $\mathbb{E}[X_i]$ and $\text{Var}[X_i]$. [15 marks]
- Are the X_i r.v's (i) mutually independent (ii) pairwise independent? Justify your reasoning. [5 marks]
- For $m = 500$, $n = 1000$, using the Chernoff bound, prove that,

  $$\Pr[X_i < 4] > 0.53$$

  [15 marks]

(2) **[55 marks]** If X is a random variable from a certain distribution, then the function

$$\phi(t) := \mathbb{E}[\exp^{tX}] := \sum_{x \in \text{Range}(X)} \exp(t\,x) \Pr[X = x]$$

is referred to as the *moment generating function* of the specific distribution.

The name *moment generating function* comes from the fact that derivatives of $\phi(t)$ can be used to generate the different moments of the distribution. For example,

$$\mathbb{E}[X] = \phi'(0) := \frac{d}{dt}[\phi(t)]\bigg|_{t=0}$$

$$\mathbb{E}[X^2] = \phi''(0) = \frac{d}{dt}[\phi'(t)]\bigg|_{t=0}$$

and so on. To see this, note that,

$$\phi'(t) = \frac{d}{dt}\left[\sum_{x \in \text{Range}(X)} \exp(t\,x) \Pr[X = x]\right] = \left[\sum_{x \in \text{Range}(X)} x \, \exp(t\,x) \Pr[X = x]\right]$$

$$\phi'(0) = \left[\sum_{x \in \text{Range}(X)} x \, \exp(0\,x) \Pr[X = x]\right] = \left[\sum_{x \in \text{Range}(X)} x \, \Pr[X = x]\right] = \mathbb{E}[X]$$

Similarly, we can show that for $q \geq 1$, $\mathbb{E}[X^q] = \frac{d^q}{dt^q}\phi(t)\Big|_{t=0}$.

Using the above definition,

- Prove that if $X \sim \text{Ber}(p)$, then $\phi(t) = p\,e^t + 1 - p$. [10 marks]

- Using this moment generating function, prove that if $X \sim \text{Ber}(p)$, then $\mathbb{E}[X] = p$ and $\text{Var}[X] = p\,(1-p)$. [15 marks]

  Similarly,

- Prove that if $X \sim \text{Bin}(n, p)$, then $\phi(t) = (p\,e^t + 1 - p)^n$. [15 marks]

- Using this moment generating function, prove that if $X \sim \text{Bin}(n, p)$, then $\mathbb{E}[X] = np$ and $\text{Var}[X] = n\,p\,(1-p)$. [15 marks]

(3) **[25 marks]** A herd of cows is stricken by an outbreak of cold cow disease. The disease lowers a cow's body temperature from normal levels, and a cow will die if its temperature goes below 90 degrees F. The disease epidemic is so intense that it lowered the average temperature of the herd to 85 degrees. Body temperatures as low as 70 degrees, but no lower, were actually found in the herd.

- Use Markov's Theorem to prove that at most 3/4 of the cows could survive. [15 marks]

- Suppose there are 400 cows in the herd. Show that the bound from the previous part is the best possible by giving an example set of temperatures for the cows so that the average herd temperature is 85 and 3/4 of the cows will have a high enough temperature to survive. [10 marks]

(4) **[20 marks]** If R is a non-negative random variable, then Markov's Theorem gives an upper bound on $\Pr[R \geq x]$ for any real number $x > \mathbb{E}[R]$. If b is a lower bound on R, then Markov's Theorem can also be applied to $R - b$ to obtain a possibly different bound on $\Pr[R \geq x]$.

- Show that if $b > 0$, applying Markov's Theorem to $R - b$ gives a tighter upper bound on $\Pr[R \geq x]$ than simply applying Markov's Theorem directly to R. [15 marks]

  Hence, the Markov Theorem we prove is tighter.

- What value of $b \geq 0$ gives the best bound? [5 marks]

(5) **[40 marks]** A computer program crashes at the end of each hour of use with probability p, if it has not crashed already. Let H be the number of hours until the first crash.

- What is the distribution of H? Compute $\mathbb{E}[H]$ and $\text{Var}[H]$. [10 marks]

- Use Chebyshev's Theorem to upper-bound $\Pr[|H - {}^{1}\!/_{p}| > \frac{x}{p}]$ for $x > 0$. [10 marks]

- Use the above bound to show that $\Pr[H > {}^{a}\!/_{p}] < \frac{1-p}{(a-1)^2}$. [5 marks]

- Compute the exact value of $\Pr[H > {}^{a}\!/_{p}]$. (For simplicity, assume that the constant a is divisible by p.) [10 marks]

- Compare the bound from Chebyshev's Theorem with the exact value. Which quantity is smaller? [5 marks]

(6) **[25 marks]** There is a fair coin and a biased coin that flips heads with probability 3/4. You are given one of the coins (with probability $\frac{1}{2}$), but you don't know which. To determine which coin was picked, your strategy will be to choose a number n and flip the picked coin n times. If the number of heads flipped is closer to $3n/4$ than to $n/2$, you will guess that the biased coin had been picked and otherwise you will guess that the fair coin had been picked. Use the Chebyshev Bound to find a value n so that with probability 0.95 your strategy makes the correct guess, no matter which coin was picked.

(7) **[25 marks]** Let us prove the one-sided Chebyshev's Theorem:

$$\Pr[R - \mathbb{E}[R] \geq x] \leq \frac{\mathrm{Var}[R]}{x^2 + \mathrm{Var}[R]}$$

For this, define $Y := (R - \mathbb{E}[R] + a)^2$ for $a \geq 0$. This implies that if $R - \mathbb{E}[R] \geq x$, then, $Y \geq (x + a)^2$. Using this reasoning,

- Apply the Markov bound to the r.v Y, and prove the following statement:

$$\Pr[R - \mathbb{E}[R] \geq x] \leq \frac{a^2 + \mathrm{Var}[R]}{(a + x)^2} \text{ [15 marks]}$$

- Prove the one-sided Chebyshev's Theorem by finding the best value of a (optimize w.r.t a to obtain the tightest bound). [10 marks]

(8) **[75 marks]** Implementing Randomized QuickSelect and Randomized QuickSort

(a) For a given input array A of n distinct elements, and $k \in \{1, n\}$, write a function in the language of your choice (preferably C or Python) to implement Randomized QuickSelect (from Lecture 14) to compute the k^{th} smallest element. [10 marks]

(b) Use the above function to implement an algorithm to sort the array A. [10 marks]

(c) Write a function that implements Randomized QuickSort (from Lecture 20) to sort the array A. [15 marks]

Print out your code and submit it with the assignment.

Use the following array of $n = 10$ in order to test the code. $A = [7, 3, 99, 4, 0, 34, 84, 9, 1, 456]$. We can compute the expected runtime for both algorithms by repeating the experiment for 100 independent runs (each run of the algorithm involves selecting a random pivot element p).

(i) Report the expected runtime of the functions for the subparts (a), (b), (c) above. [5 marks]

(ii) Compute the standard deviation in the runtime for the experiment above, and report the quantity $\mu + \sigma$ and $\mu - \sigma$ for each of the subparts (a), (b), (c) above. The $[\mu - \sigma, \mu + \sigma]$ is referred to as the *confidence interval* and is typically used to report the results of a randomized experiment. [15 marks]

In order to study the effect of n (size of the array) on the performance of each function written in parts (b) and (c) above, let us create a *scaling plot.*

- For this, we will generate random arrays of size n for $n \in \{5, 20, 50, 100, 500, 1000\}$. For each n, repeat the experiment in part (i) above for 50 times, and compute the average runtime across the 50 runs. Plot the average runtime with respect to n for each of parts (b) and (c). [12 marks]
- Which sorting algorithm is faster across values of n? Explain why? [8 marks]