COMP 690 Fall 2020 Assignment 1

Name: Andrew Cheng

Problem 1

It suffices to show that X takes on the values 0 and 1 with probability p and $1-p$, respectively. First we show that the probability mass cannot lie outside the interval $[0,1]$. Since X is strictly positive, we can just show $X \in (1,\infty)$ with probability 0. This will be done via contradiction. Suppose $P\{X > 1\} > 0$. Then there exists constants $C, \epsilon > 0$ such that

$$P\{X \geq 1+\epsilon\} = C$$

Now let $m > 1$. For arbitrary $\tilde{k} \in \mathbb{N}$, we have by Chebyshev,

$$\begin{aligned} P\{|X^{\tilde{k}} - p| \geq \sqrt{\frac{m \operatorname{Var}\{X^{\tilde{k}}\}}{C}} \\ &\leq \frac{\operatorname{Var}(X^{\tilde{k}})}{\operatorname{Var}(X^{\tilde{k}})} \times \frac{C}{m} \\ &< C \end{aligned} \tag{1}$$

Then we have

$$C = P\{X \geq 1+\epsilon\} = P\{X^k \geq (1+\epsilon)^k\} \tag{2}$$

for all $k \geq 1$ since X is non-negative. Now we want to get (1) into a form so that we may use Chebyshev to derive the needed contradiction. To this end, notice that we may choose $\tilde{k} \in \mathbb{N}$ such that

$$(1+\epsilon)^{\tilde{k}} > p + \sqrt{\frac{m \operatorname{Var}(X^k)}{C}} \tag{3}$$

for any $m > C$. We can do this since clearly $(1+\epsilon)^k < (1+\epsilon)^{k+1}$ and $\lim_{k\to\infty}(1+\epsilon) = \infty$, i.e., $(1+\epsilon)^k$ is strictly increasing in k. Continuing from (1),

$$\begin{aligned} C &= P\{X^{\tilde{k}} \geq (1+\epsilon)^{\tilde{k}}\} \\ &\leq P\{X^{\tilde{k}} > p + \sqrt{\frac{m \operatorname{Var}(X^k)}{C}}\} \\ &= P\left\{X^{\tilde{k}} - p > \sqrt{\frac{m \operatorname{Var}(X^k)}{C}}\right\} \\ &\leq P\left\{X^{\tilde{k}} - p > \sqrt{\frac{m \operatorname{Var}(X^k)}{C}} \text{ and } X^{\tilde{k}} - p < -\sqrt{\frac{m \operatorname{Var}(X^k)}{C}}\right\} \\ &= P\left\{|X^{\tilde{k}} - p| > \sqrt{\frac{m \operatorname{Var}(X^k}{C}}\right\} \\ &\leq \frac{C}{m} \\ &< C \end{aligned}$$

where the second last inequality follows from (1). Now we can assume that $P\{X \in [0,1]\} = 1$. To show that the probability must lie on the discrete set $\{0,1\}$, we partition the interval $[0,1]$ in countable intervals. Let's construct a sequence $\{x_k\}_{k\in\mathbb{Z}}$ consisting of endpoints of said intervals. Let $x_0 = \frac{1}{2}$ and α be any positive integer and construct the intervals in the following manner:

1. Define $x_{k-1} = (x_k)^{\frac{1}{\alpha}}$ $\forall k \in \mathbb{Z}_{\geq 0}$.
2. $x_{k+1} = (x_k)^{\alpha}$ $\forall k \in \mathbb{Z}_{\leq 0}$

Then $\lim_{k\to-\infty} x_k = 1$ and $\lim_{k\to\infty} = 0$ and $x_k < x_{k+1}$ for all k. Hence the intervals $[x_k, x_{k+1})$ are disjoint for all k. Since $\mathbb{Z}$ is countable, then it holds that

$$\bigcup_{k\in\mathbb{Z}} [x_k, x_{k+1}]$$

is a countable union. Clearly $[0,1] = \bigcup_{k\in\mathbb{Z}}[x_k, x_{k+1}) \cup \{0\} \cup \{1\}$ is also countable. Denote

$$p_k := \Big\{ P\{X \in [x_k, x_{k+1}]\} \Big\}$$

For contradiction assume $P\{X \in (0,1)\} > 0$. But then we have,

$$\begin{aligned} p &= E\{X\} \\ &\geq P\{X = 1\} + x_0 p_0 + x_1 p_1 + x_2 p_2 + \cdots x_{-1} p_{-1} + x_{-2} p_{-2} + \cdots \\ &> P\{X = 1\} + x_0 p_1 + x_1 p_2 + \cdots x_{-1} p_0 + x_{-2} p_{-1} + x_{-3} p_{-2} + \cdots \end{aligned}$$

where the last strict inequality holds since we showed $\{x_k\}_{k\in\mathbb{Z}}$ is strictly increasing, i.e. $x_0 < x_1$ and it must be that $X \in (x_k, x_{k+1}]$ for some k. By assumption we had $p_k > 0$ for some k. But then we have

$$\begin{aligned} p &> P\{X = 1\} + x_0 p_1 + x_1 p_2 + \cdots x_{-1} p_0 + x_{-2} p_{-1} + x_{-3} p_{-2} + \cdots \\ &= P\{X = 1\} + x_0^{2\alpha} p_{-1} + x_1^{2\alpha} p_0 + x_2^{2\alpha} p_1 + \cdots x_{-1}^{2\alpha} + x_{-2}^{2\alpha} p_{-3} \\ &\geq E\Big\{ X^{2\alpha} \Big\} \end{aligned}$$

but since $\alpha > 1$, $\alpha \in \mathbb{N}$ and we have $E\{X^k\} = p$ for any k, we yield

$$p = E\{X\} > E\{X^{2\alpha}\} = p$$

which is a contradiction. It follows that $P\{X \in \{0,1\}\} = 1$.

Problem 2

1. For this problem, I drew out the cases for $d = 1, 2, 3$, and realized that the geometric volume of $P\{X_1 + \cdots + X_d \leq 1\}$ correponds to the volume of a d-dimensional simplex. From convex analysis and linear algebra, we know that the volume of a d-dimensional simplex based in a d-dimensional unt cube is

$$\frac{1}{d!} \det \left(e_1, \cdots e_d \right)$$

where e_i, is the ith standard basis vector. The result follows.

2.

$$\begin{aligned}
F(x) &= P\{d \min\{X_1, \ldots, X_d < x\} \\
&= 1 - P\{d \min\{X_1, \ldots, X_d \geq x\} \\
&= 1 - P\{\min\{X_1, \ldots, X_d\} \geq \frac{x}{d}\} \\
&= 1 - P\{X_1 \geq \frac{x}{d}, \ldots, X_d \geq \frac{x}{d}\} \\
&= 1 - \prod_{i=1}^{d} P\{X_i \geq \frac{x}{d}\} \qquad \text{(Independence)} \\
&= 1 - \prod_{i=1}^{d} \left(1 - P\{X_i < \frac{x}{d}\}\right) \\
&= 1 - \prod_{i=1}^{d} \left(1 - \int_0^{\frac{x}{d}} du\right) \\
&= 1 - \left(1 - \frac{x}{d}\right)^d
\end{aligned}$$

Taking the limit with respect to d,

$$1 - \lim_{d \to \infty} \left(1 + (-\frac{x}{d})\right)^d = 1 - e^{-x}$$

which shows that this random variable variable is exponential with parameter 1 when $d = \infty$.

3. This proof will show that as the dimension of the hypercube grows, we should expect that the proportion of points lying near the surface of the cube to increase (contingent

on uniform sampling). Note $||X||_2 = \sqrt{x_1^2 + \cdots + x_d^2}$. We have

$$\begin{aligned}
P\left\{(1-\epsilon)\sqrt{\frac{d}{3}} \le ||x|| \le (1+\epsilon)\sqrt{\frac{d}{3}}\right\} &= P\left\{(1-\epsilon)^2\frac{d}{3} \le ||x||^2 \le (1+\epsilon)^2\frac{d}{3}\right\} \\
&= P\left\{(\epsilon^2 - 2\epsilon + 1)\frac{d}{3} \le ||x||^2 \le (\epsilon^2 + 2\epsilon + 1)\frac{d}{3}\right\} \\
&= P\left\{\frac{\epsilon^2 d}{3} - \frac{2\epsilon d}{+} 3\frac{d}{3} \le ||x||^2 \le \frac{\epsilon^2 d}{3} + \frac{2\epsilon d}{3} + \frac{d}{3}\right\} \\
&\ge P\left\{\frac{\epsilon^2 d}{3} - \frac{2\epsilon d}{3} \le ||x||^2 - \frac{d}{3} \le -\frac{\epsilon^2 d}{3} + \frac{2\epsilon d}{3}\right\} \\
&= P\left\{\left| ||x||^2 - \frac{d}{3}\right| \le \frac{2\epsilon d}{3} - \frac{\epsilon^2 d}{3}\right\}
\end{aligned}$$

We can show $E||d||^2 = \frac{d}{3}$:

$$\begin{aligned}
E||x||^2 &= E\{\sum_{i=1}^{d} X_i^2\} \\
&= \sum_{i=1}^{d} E\{X_i^2\} \\
&= \sum_{i=1}^{d} \int_{-1}^{1} u^2 du \\
&= \frac{d}{3}
\end{aligned}$$

We had,

$$\begin{aligned}
P\left\{(1-\epsilon)\sqrt{\frac{d}{3}} \le ||x|| \le (1+\epsilon)\sqrt{\frac{d}{3}}\right\} &\ge P\left\{\left| ||x||^2 - \frac{d}{3}\right| \le \frac{2\epsilon d}{3} - \frac{\epsilon^2 d}{3}\right\} \\
&= P\left\{\left| ||x||^2 - E||x||^2\right| \le \frac{2\epsilon d}{3} - \frac{\epsilon^2 d}{3}\right\} \\
&\ge 1 - \frac{\text{Var}\left(||x||^2\right)}{\left(\frac{2\epsilon}{3} - \frac{\epsilon^2}{e}\right)^2 d^2} \\
&\xrightarrow{d \to \infty} 1
\end{aligned}$$

Where the last inequality follows from Chebyshev. We will now justify the last line above. It suffices to show that $\text{Var}(||x||^2) = \mathcal{O}(d)$. Computing this quantity

$$\begin{aligned}
\text{Var}(||X||^2) &= \text{Var}(X_1^2 + \cdots X_d^2) \\
&= \sum_{i=1}^{d} \text{Var}(X_d^2) \\
&= \sum_{i=1}^{d} \left[E\{x^4\} - \big(E\{X^2\}\big)^2 \right] \\
&= d\left[\int_{-1}^{1} u^4 du - \frac{1}{9} \right] \\
&= \frac{4d}{45} \\
= \mathcal{O}(d)
\end{aligned}$$

Problem 3

(a) It is clear that the distribution that $\sum_{i=1}^{d} X_i$ is Binomial with d trials and $p = \frac{1}{2}$. Intuitively, we can view the sum as a sum of independent fair coin tosses (i.e. d independent fair Bernoulli trials).

(b) We can view the hamming distance as a $Bin(d, \frac{1}{2})$ random variable. In fact, $H(X,Y) = \sum_{i=1}^{d} |X_i - Y_i|$. Now define a new random variable $Z_i = |X_i - Y_i|$. Notice that

$$Z_i = \begin{cases} 1, & \text{if } (X_i = 1, Y_i = 0) \vee (X_i = 0, Y_i = 1) \\ 0, & \text{otherwise} \end{cases}$$

which implies

$$Z_i = \begin{cases} 1, & \text{w.p. } \frac{1}{2} \\ 0, & \text{w.p. } \frac{1}{2} \end{cases}$$

This effectively shows that $Z_i \sim Bernoulli(\frac{1}{2})$ for $i = 1, \ldots, d$. Since Z_i's are independent we have that $\sum_{i=1}^{d} Z_i \sim Binomial(d, \frac{1}{2}) \implies E\{\sum_{i=1}^{d} Z_i\} = \frac{d}{2}$. Invoking the central limit theorem yields:

$$\sum_{i=1}^{d} Z_i - \frac{d}{2} \xrightarrow{d \to \infty} \mathbf{N}(0, d\sigma^2)$$

where $\sigma^2 = \frac{1}{4}$ is the variance of Z_i. Now using the fact that if $X \sim \mathbf{N}(0, \sigma^2)$ then $|X|$ has a half-normal distribution with parameter σ^2. Hence, $\lim_{d\to\infty} \left| \sum_{i=1}^{d} Z_i - \frac{d}{2} \right|$ is a half-normal distribution with parameter $\sqrt{d\sigma^2}$. It follows that

$$
\begin{aligned}
E\left|\lim_{d\to\infty}\sum_{i=1}^{d} Z_i - \frac{d}{2}\right| &= E\left|\lim_{d\to\infty} H(X,Y) - \frac{d}{2}\right| \\
&= \sigma\sqrt{\frac{2d}{\pi}} \\
&= \frac{1}{4}\sqrt{\frac{2d}{\pi}} \\
&= \mathcal{O}(\sqrt{d})
\end{aligned}
$$