Core Project 1.1 - Random Binary Expansions

January 25, 2022

Contents

1 Question 1

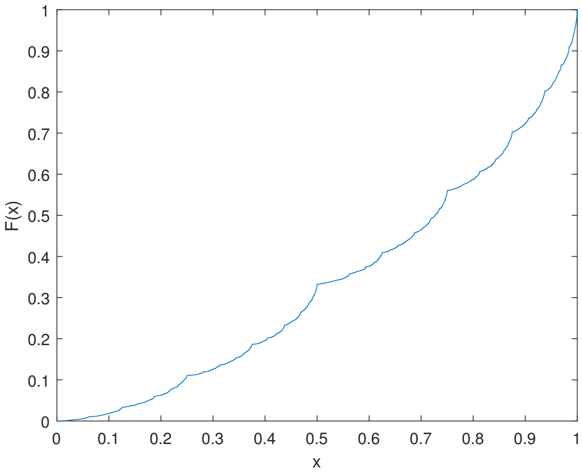


Figure 1: Empirical cumulative distribution function with $p = \frac{2}{3}$, $n = 30$ and $N = 10000$

Qualitatively, at $N = 10000$, the graphs are very reproducible, and the symmetry and fractal nature of the graph is clear. This is not the case for $N = 1000$ as seen in Figure (2).

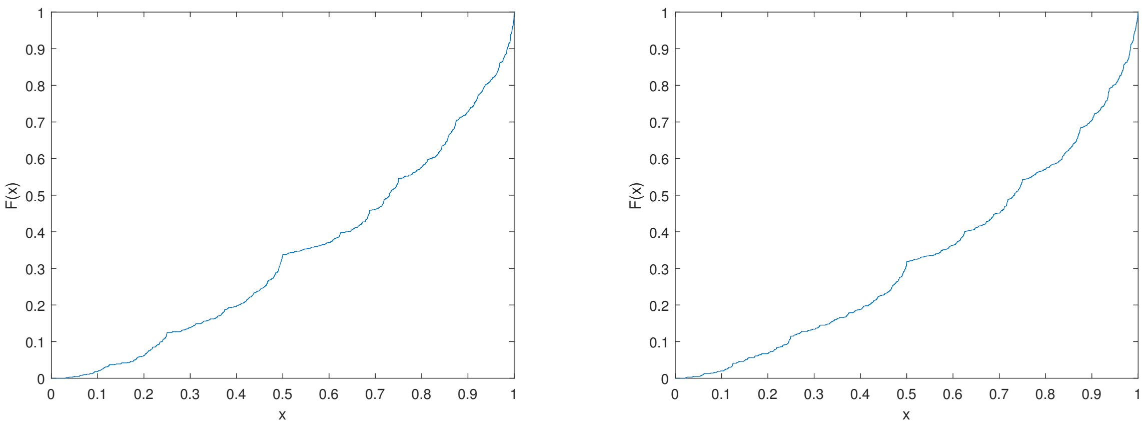


Figure 2: Two runs at $N = 1000$, demonstrating that the graphs are not entirely reproducible

Statistically, if we want a 0.99 probability of our estimate of $F(x)$ to be within 0.02 of the actual distribution at each x, we can find a minimum N. (Assuming n is large enough that X_j^n is distributed like X)

$$S_N := \frac{\hat{F}(x)}{N} = \sum_{j=1}^{N} \mathbf{1}(X_j \leq x)$$

This is a sum of iid random Bernoulli variables, so can be manipulated using the central limit theorem. Since $S_N \sim \text{Bin}(N, p)$, where $p = F(x)$:

$$S_N = Np + \sqrt{Np(1-p)} \cdot Z$$

where Z is the standard normal distribution.

$$\mathbb{P}(|\hat{F}(x) - F(x)| \leq 0.02) = \mathbb{P}(\sqrt{\frac{p(1-p)}{N}}|Z| \leq 0.02)$$

Choosing a worst variance when $p = \frac{1}{2} \implies \sqrt{p(1-p)} = \frac{1}{2}$,

$$0.02 \cdot 2 \cdot \sqrt{N} \geq \Phi^{-1}(0.995) \implies N \geq 4147$$

2 Question 2

In order for $\{X \leq x\}$, the first digit (let's say the lth digit) at which x and X differ must satisfy $U_l < x_l$. The event that no digits differ has probability zero, so we do not consider it.

$$\{X \leq x\} = \bigsqcup_{k=1}^{n} \{(U_1 = x_1, \dots, U_{k-1} = x_{k-1}, U_k = 0)\mathbf{1}(x_k = 1)\} \tag{1}$$

Note that this is a disjoint union, so the probability of the LHS is the sum of probabilities of each event on the RHS. By independence:

$$\begin{aligned} F(x) = \mathbb{P}(X \leq x) &= \sum_{k=1}^{n} \mathbb{P}((U_1 = x_1, \dots, U_{k-1} = x_{k-1}, U_k = 0)\mathbf{1}(x_k = 1)) \\ &= \sum_{k=1}^{n} x_k \mathbb{P}(U_1 = x_1) \dots \mathbb{P}(U_{k-1} = x_{k-1})\mathbb{P}(U_k = 0) \\ &= \sum_{k=1}^{n} q x_k \prod_{i=1}^{k-1} p\mathbf{1}(x_i = 1) q\mathbf{1}(x_i = 0) = \sum_{k=1}^{n} x_k p^{\ell_k} q^{k-\ell_k} \end{aligned} \tag{2}$$

Where $\ell_k = \sum_{i=1}^{k-1} \mathbf{1}(x_i = 1)$.

In the code (7.2.3), this is implemented like so:

$$F(x) = \sum_{k=1}^{n} q x_k \prod_{i=1}^{k-1} (p x_i + q(1 - x_i))$$

3 Question 3

The relevant code used to produce these graphs are found here: 7.2.1 and 7.2.3

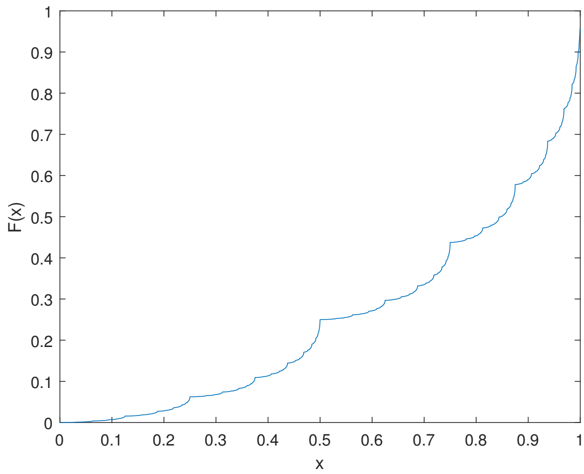


Figure 3: $F(x)$ plotted with exact values sampled at regular intervals of $2^{-n}, n = 11$, with $p = \frac{3}{4}$

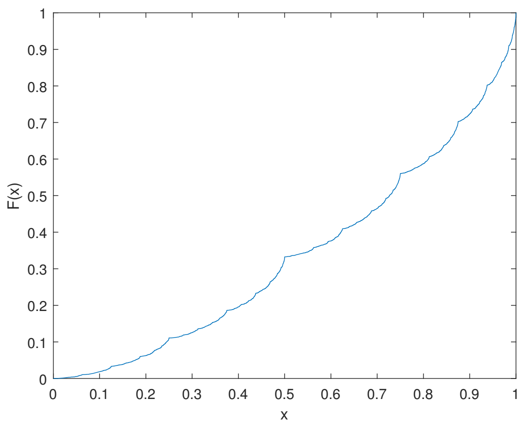
3.1 Comparison of graphs

When we compare the exact graph in Figure 3 to the empirical plot at the same p-value (Figure 4b), there is no noticeable difference, and in fact

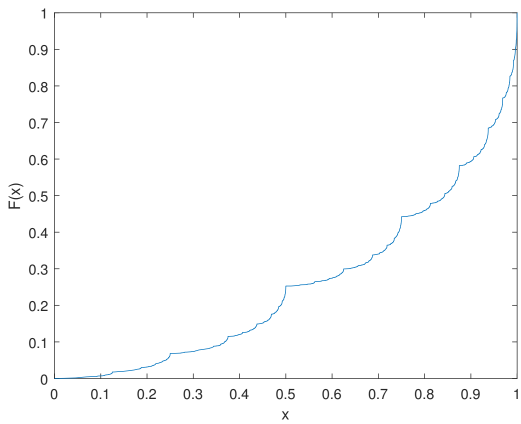
$$\max_x |F_{\text{empirical}}(x) - F_{\text{exact}}(x)| = 0.0065436$$

so we can have some confidence in the formula obtained in Question 2.

When we compare the exact graph to the empirical plot in Question 1 (Figure 4a), there are some quantitative differences. For ease of reference, we say $F_p(x)$ refers to the value of F for a given p-value.



(a) Empirical plot of $p = \frac{2}{3}$



(b) Empirical plot of $p = \frac{3}{4}$

Figure 4: Empirical plots of $F(x)$ using the Monte Carlo method of Question 1

$F_{\frac{3}{4}}(x)$ increases at a slower rate immediately after x-values which have a short finite binary expansion in both graphs, but this effect is far more pronounced when $p = \frac{3}{4}$. The plot looks rougher, qualitatively. This can be seen by measuring the gradient of a chord after a small change h in x, as demonstrated in Table 1.

Table 1: Comparison of gradients of chords between $(x, F(x))$ and $(x+h, F(x+h))$, where $h = 10 \times 2^{-11}$

x	$\frac{F(x+h)-F(x)}{h}$	
	$p = \frac{3}{4}$	$p = \frac{2}{3}$
0.5	0.0111328125	0.0763027316
0.625	0.0333984375	0.152605463259329
0.8215	0.1001953125	0.305210926518657

Also, $\forall x \in (0,1), F_{\frac{3}{4}}(x) < F_{\frac{2}{3}}(x)$. This is expected, because if each digit is more likely to be a 1, then the probability that the whole number is smaller than any given x will be smaller than when $p = \frac{2}{3}$

3.2 Complexity

3.2.1 Monte Carlo

With the Monte Carlo simulation, seen in 7.1.1, each U_n is generated in $O(1)$ time, so each X^n is generated in $O(n)$. As there are N samples, the algorithm is $O(nN)$.

3.2.2 Exact formula

The algorithm used to calculate for finite expansions can be found at 7.2.1 and 7.2.3. $N = 2^n$ where N denotes the number of samples. Each sample is made in $O(n)$ time. So then in total the time complexity is $O(nN) = O(n2^n)$ again.

4 Question 4

In this question, we say c is a number with binary expansion of size n, and $N > n$. We use the definition of continuity at y:

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ s.t. } |x - y| < \delta \implies |F(y) - F(x)| < \varepsilon$$

As a cumulative probability distribution, $F(x)$ is increasing, so

$$\begin{aligned} x \in (y - \delta, y) &\implies |F(x) - F(y)| \leq |F(y - \delta) - F(y)| \\ x \in (y, y + \delta) &\implies |F(x) - F(y)| \leq |F(y + \delta) - F(y)| \end{aligned}$$

so it is sufficient to show that

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ s.t. } |F(y \pm \delta) - F(y)| < \varepsilon \tag{3}$$

Lemma 1. *Given* $\varepsilon > 0, \exists \delta_1 > 0$ *s.t.* $|F(c + \delta_1) - F(c)| < \varepsilon$

Proof.

$$x = c + 2^{-N} = \sum_{i=1}^{N} \frac{x_i}{2^i} \implies x_k = \begin{cases} c_k & 0 \leq k \leq n \\ 0 & n < k < N \\ 1 & k = N \end{cases}$$

Using the formula found in (2), and the same ℓ_k notation, note that $\ell_N = \ell_n + 1$:

$$F(x) - F(c) = \sum_{k=1}^{N} (x_k p^{\ell_k} q^{k-\ell_k} - c_k p^{\ell_k} q^{k-\ell_k}) = p^{\ell_N} q^{N-\ell_N} = p^{\ell_n+1} q^{N-\ell_n-1} \tag{4}$$

For sufficiently large N, (4) $< \varepsilon$ because $q \in (0, 1)$. So choose $\delta_1 = 2^{-N}$ □

Lemma 2. *Given* $\varepsilon > 0, \exists \delta_2 > 0$ *s.t.* $|F(c - \delta_2) - F(c)| < \varepsilon$

Proof.

$$x = c - 2^{-N} \implies x_k = \begin{cases} c_k & 0 \leq k < n \\ 0 & k = n \\ 1 & n < k \leq N \end{cases}$$

Note that $\ell_k = \ell_n + k - n - 1, \forall k > n$

$$\begin{aligned} F(x) - F(c) &= \sum_{k=1}^{N} (x_k p^{\ell_k} q^{k-\ell_k} - c_k p^{\ell_k} q^{k-\ell_k}) \\ &= -p^{\ell_n} q^{n-\ell_n} + \sum_{k=n+1}^{N} p^{\ell_k} q^{k-\ell_k} \\ &= p^{\ell_n} q^{n-\ell_n} \left(-1 + q \sum_{k=0}^{N-n-1} p^k \right) \\ &= -p^{N-n+\ell_n} q^{n-\ell_n} \end{aligned} \tag{5}$$

For sufficiently large N, (4) $< \varepsilon$ because $p \in (0, 1)$. So choose $\delta_2 = 2^{-N}$ □

Theorem 1. *F is continuous at c, where c has finite binary expansion*

Proof. We can choose $\delta = \min(\delta_1, \delta_2)$ to satisfy the continuity equation (3) for any $\varepsilon > 0$ □

Theorem 2. *F is continuous everywhere (as suggested by plots)*

Proof. Let $C \subset [0,1]$ be the set of numbers with finite binary representations, or with infinite trailing ones (because they can be replaced by trailing zeroes, wlog). Therefore $\forall x \in [0,1] \setminus C, \forall N, \exists m > N$ s.t. $x_m = 0$

We can get bounds $c < x < d$ (where $c, d \in C$) simply by truncating its binary representation at the mth zero, where $m > N$, and choosing 0 or 1 respectively for the mth digit.

Clearly, $d - c \leq 2^{-N}$, so by noting that $F(x)$ is an increasing function and appealing to equation (4), we can choose N large enough such that

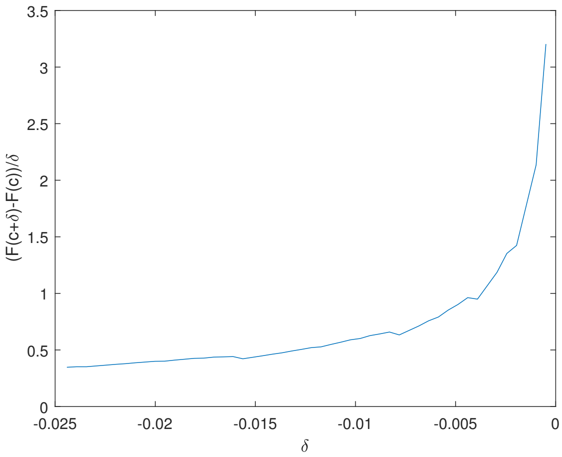
$$F(d) - F(c) \leq F(c + 2^{-N}) - F(c) < \varepsilon$$

After choosing $\delta = \min(d - x, x - c)$, we have that

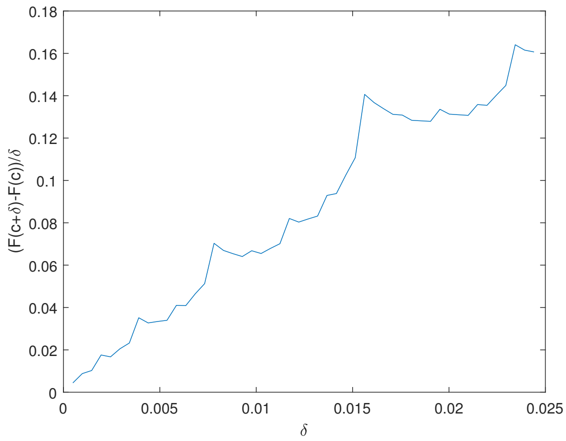
$$|x - y| < \delta \implies |F(x) - F(y)| \leq F(c) - F(d) < \varepsilon$$

$\square$

5 Question 5



(a) $\delta \in [0, 50 \times 2^{-11}]$



(b) $\delta \in [-50 \times 2^{-11}, 0]$

Figure 5: Plots of $\frac{F(c+\delta)-F(c)}{\delta}$ for $\delta \in [-50 \times 2^{-11}, 50 \times 2^{-11}]$ and $c = 0.5625$

In Figure 5 are two graphs separating the δ into positive and negative values, so that the detail of the graph can be seen for positive δ, which is otherwise hidden because of the relatively large values that arise in negative δ.

I chose this range of δ because it gives about 100 data points which is sufficient to see a trend, but the values of δ remain small enough that gradients of chords are close to the value of the derivative (if it exists).

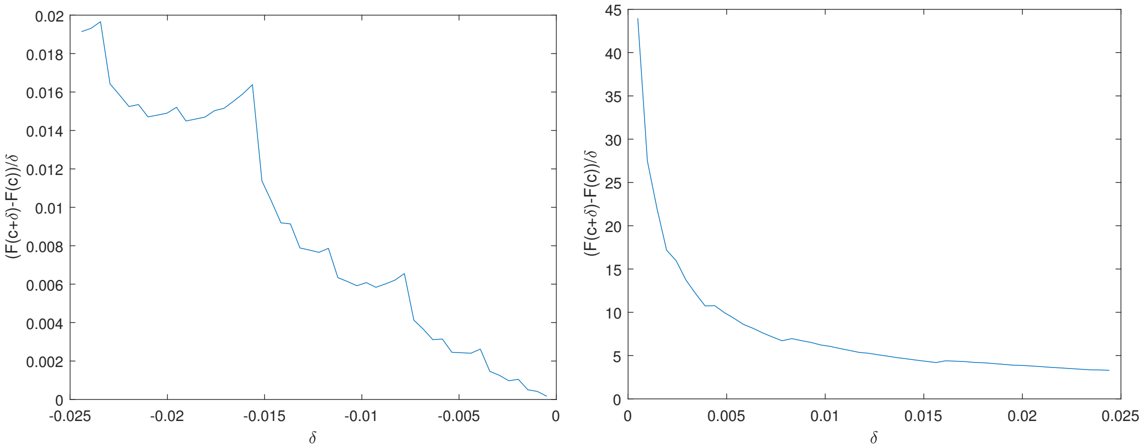
These graphs suggest that as $\delta \downarrow 0$, we can expect the value to tend to zero, and so F is right-differentiable at c, however as $\delta \uparrow 0$, the gradients of chords seem to be unbounded, and so the limit is $+\infty$, and F is not left-differentiable at c.

6 Question 6

Conjecture 1. *For arbitrary $c \in (0,1)$ with finite binary expansion of length n,*

1. *If $p > \frac{1}{2}$*
   (a) *F is right-differentiable at c with limit 0*
   (b) *F is not left-differentiable at c*
2. *If $p < \frac{1}{2}$*
   (a) *F is not right-differentiable at c*
   (b) *F is left-differentiable at c with limit 0*
3. *If $p = \frac{1}{2}$*
   (a) *F is differentiable at c with derivative 1*

6.1 Plots supporting conjecture



(a) $\delta \in [0, 50 \times 2^{-11}]$ (b) $\delta \in [-50 \times 2^{-11}, 0]$

Figure 6: Plots of $\frac{F(c+\delta)-F(c)}{\delta}$ for $\delta \in [-50 \times 2^{-11}, 50 \times 2^{-11}]$ and $c = 0.5$ for $p = \frac{1}{5}$

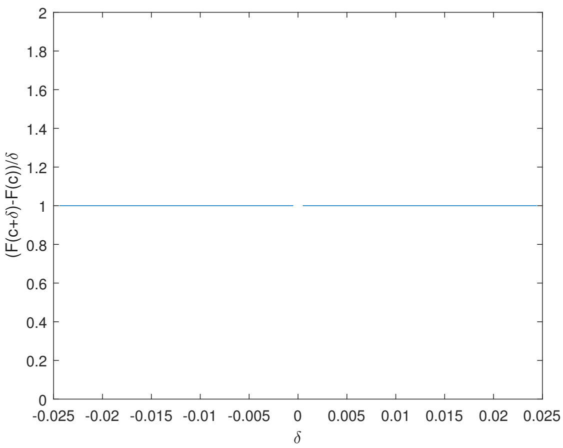


Figure 7: Plots of $\frac{F(c+\delta)-F(c)}{\delta}$ for $\delta \in [-50 \times 2^{-11}, 50 \times 2^{-11}]$ and $c = 0.875$ for $p = \frac{1}{2}$

1. Figure 5 supports this part of the conjecture.

2. Figure 6 is calculated with different p and c values and support this part of the conjecture. Up to a scaling factor, the graphs are incredibly similar to those in 5

3. Figure 7 is calculated with yet another c, and seems to not only have right and left limits at 1, but in fact is constant with value 1. (Note that the gap is just because the function is undefined at $\delta = 0$)

6.2 Proof of conjecture

Proof. For notational ease, throughout this proof we let $g(x) = \frac{F(c+x)-F(c)}{x}$

1. If $p > \frac{1}{2}$:

   (a) $F(x)$ is right-differentiable at c with right derivative 0 is equivalent to:

   $$\forall \varepsilon > 0, \exists \delta > 0 \text{ s.t. } \forall x \in (0, \delta), |g(x)| < \varepsilon$$

   First we show that $\forall \varepsilon > 0, \exists N$ s.t. $\forall m \geq N, |g(2^{-m})| < \frac{\varepsilon}{2}$.
   By (4):
   $$g(2^{-N}) = (2q)^N p^{\ell_n+1} q^{-\ell_n-1} \tag{6}$$
   As $q < \frac{1}{2}$, this can be made arbitrarily small, so choose N where $\forall m \geq N, |g(2^{-m})| < \frac{\varepsilon}{2}$
   Take $\delta = 2^{-N}$.
   Given some $x \in (0, \delta), \exists m \geq N$ s.t. $2^{-m-1} < x < 2^{-m}$, so then:

   $$F(c+x) \leq F(c+2^{-m})$$
   $$g(x) \leq \frac{F(c+x)-F(c)}{2^{-m-1}} \leq \frac{F(c+2^{-m})-F(c)}{2^{-m-1}} = 2g(2^{-m}) < \varepsilon$$

   We have shown that $F(x)$ is right-differentiable at c with right derivative zero.

   (b) Non-existence of left limit is equivalent to saying

   $$\forall L > 0, \exists \varepsilon > 0 \text{ s.t. } \forall \delta > 0, \exists x \in (0, \delta) \text{ s.t. } |g(-x) - L| \geq \varepsilon$$

   Given arbitrary L, let $\varepsilon = 1$. By (5),

   $$g(-2^{-N}) = -2^N p^{N-n+\ell-1} q^{n-\ell+1} \tag{7}$$

   Since $2p > 1$, this can be made arbitrarily large and negative.
   So given some $M, \exists N$ s.t. $g(-2^{-N}) < -M$
   We can freely choose $M > 1 + L$. $\forall \delta, \exists k \geq M$ s.t. $x = 2^{-k} \in (0, \delta)$
   By construction, $\exists x \in (0, \delta)$ s.t. $|g(-x) - L| \geq 1$
   This contradicts the existence of L. $F(x)$ is not left-differentiable at c.

2. For $p < \frac{1}{2}$, we can use the same arguments as the ones made above, except this time, $2p < 1$ means that (7) can be made arbitrarily small and $2q > 1$ means that (6) can be made arbitrarily large. This means that exactly the opposite results hold.

   This is sufficient to prove the result, but actually something stronger and more interesting is true: if we let $\mathbb{P}_a(A)$ be the probability of event A happening when $p = a$, we have that

   Lemma 3.

   $$\mathbb{P}_{1-a}(X \leq x) = \mathbb{P}_a(X \geq 1-x) \iff F_{1-a}(x) = -F_a(1-x)$$

   Proof. To prove this, let us extend our reasoning in Question 2 to x with infinite binary expansion. There is a small caveat with some x having non-unique binary expansions, but we will solve this problem by just choosing the trailing zeros expansion, and these events happen with probability zero so do not affect the cdf:

$$\{X \leq x\} = \bigsqcup_{k=1}^{\infty} \{(U_1 = x_1, \ldots, U_{k-1} = x_{k-1}, U_k = 0)\mathbf{1}(x_k = 1)\} \tag{8}$$

and similarly,

$$\{X \geq x\} = \bigsqcup_{k=1}^{\infty} \{(U_1 = x_1, \ldots, U_{k-1} = x_{k-1}, U_k = 1)\mathbf{1}(x_k = 0)\} \tag{9}$$

$$1 - x = \sum_{i=1}^{\infty} \frac{1}{2^i} - \sum_{i=1}^{\infty} \frac{x_i}{2^i} = \sum_{i=1}^{\infty} \frac{1 - x_i}{2^i} \tag{10}$$

Combining (9) and (10), we get:

$$\{X \geq 1 - x\} = \bigsqcup_{k=1}^{\infty} \{(U_1 \neq x_1, \ldots, U_{k-1} \neq x_{k-1}, U_k = 1)\mathbf{1}(x_k \neq 0)\} \tag{11}$$

Since $\mathbb{P}_a(U_i \neq x_i) = \mathbb{P}_{1-a}(U_i = x_i)$ we must have $\mathbb{P}_{1-a}(8) = \mathbb{P}_a(11)$ $\square$

This means that the cumulative distribution of $p = 1 - a$ is just a rotation of π degrees around $(\frac{1}{2}, \frac{1}{2})$. So then we can use the chain rule to get the conjectured result for $p < \frac{1}{2}$ using what was proven for $p > \frac{1}{2}$

3. In the case that $p = \frac{1}{2}$, we see that for general x, by the same reasoning as that used in question 2,

$$\begin{aligned}
\mathbb{P}(X \leq x) &= \sum_{k=1}^{\infty} x_k \mathbb{P}(U_1 = x_1) \ldots \mathbb{P}(U_{k-1} = x_{k-1})\mathbb{P}(U_k = 0) \\
&= \sum_{k=1}^{\infty} x_k \left(\frac{1}{2}\right)^k \\
&= x
\end{aligned}$$

$F(x) = x$ is differentiable with $F'(x) = 1, \forall x \in (0, 1)$

$\square$

7 Code

7.1 Question 1

7.1.1 FEmp.m

```
function [F] = FEmp(p,n,N)
% Monte Carlo simulation to generate X^n
XnVec=zeros(N,1);
for j=1:N
    randVec=rand(n,1);
    Un=zeros(n,1);
%     Make a vector U_n of p-biased coin flips.
    for i=1:n
        if randVec(i)<p
            Un(i)=1;
        end
    end
%     Create X_n in the appropriate way.
    Xn=0;
    for i=1:n
        Xn=Xn+Un(i)/(2^i);
    end
    XnVec(j)=Xn;
end
F=@(x) 1/N*indicatorSum(x,XnVec,N)
end
```

7.1.2 indicatorsum.m

```
function[sum]=indicatorSum(x,XnVec,N)
    sum=0;
    for j=1:N
        if XnVec(j)<=x
            sum=sum+1;
        end
    end
end
```

7.1.3 Question1.m

```
% Plots the empirical function
format long
n=30;
N=10000;
p=2/3;
F=FEmp(p,n,N);
fplot(F,[0,1])
xlabel('x')
ylabel('F(x)')
```

7.2 Question 3

7.2.1 FFormula.m

```
function[sum]=FFormula(xVec,p,n)
% Implements the formula for F found in Question 2.
q=1-p;
sum=0;
prod=q;
for i=1:n
    sum=sum+xVec(i)*prod;
    prod=prod*(xVec(i)*p+(1-xVec(i))*q);
end
end
```

7.2.2 FExact.m

```
function [indexList,FList] = FExact(p,n)
% Generates a list of samples and an index list for ease of plotting the graph.
FList=zeros(2^n-1,1);
indexList=zeros(2^n-1,1);
for i=1:2^n-1
    indexList(i)=i/(2^n);
    xVec=dec2binVec(i,n);
    FList(i)=FFormula(xVec,p,n);
end
end
```

7.2.3 Question3.m

```
% Plots the lists generated in FExact.m
format long;
p=3/4;
n=11;
[indexList, FList]=FExact(p,n);
plot(indexList,FList)

xlabel('x')
ylabel('F(x)')
```

7.3 Question 5

7.3.1 gradChord.m

```
function [indexList,gradList] = gradChord(FList,c,delta1,delta2)
% gradChord Returns a list of values of (F(c+delta)-F(c))/delta
%Also returns an index list of scaled delta values to help plot the graph
delta1scaled=delta1*2^(11);
delta2scaled=delta2*2^(11);
indexList=zeros(delta2scaled-delta1scaled+1,1);
gradList=zeros(delta2scaled-delta1scaled+1,1);
for i=delta1scaled:delta2scaled
    gradList(i-delta1scaled+1)=(FList(c*2^11+i)-FList(c*2^11))/(i*2^(-11));
    indexList(i-delta1scaled+1)=i*2^(-11);
end
end
```

7.3.2 Question5.m

```
% Plots the gradChord function for specified c, p and n.
format long;
p=3/4;
n=11;
[otherIndexList, FList]=FExact(p,n);
c=0.5625;
[indexList,gradList]=gradChord(FList,c,-100*2^(-11),0);
plot(indexList,gradList)
xlabel('\delta')
ylabel('(F(c+\delta)-F(c))/\delta')
saveas(gcf,'C:\Users\Neel\Documents\CATAM\Core 1 - Random Binary Expansions\MATLAB\Figures\q5ifig','epsc')
```