Mathemtical Tripos, Part 1B
Computational project

Random Binary Expansions

January 21, 2020

Question 1

General description

Identify the coin tosses are identically independent Bernoulli distributions with probability p, then the first program should be a function of three variables, namely n, N and p; A random sample is generated by the summation formula given in the problem description.
The second half of the question is about approximating the distribution function by Monte Carlo simulation. I wrote a program to obtain the empirical distribution function for n=30, p=2/3 as requested, and chose a suitable N by error analysis.
The program to find the EDF can be found at **Program 1** below
The program to generate a random sample can be found at **Program 2** below

Error analysis

For this specific Monte Carlo simulation, the error can be quantified by the standard deviation of the empirical distribution function $\hat{\mathrm{F}}$, i.e.:

$$Random\ error = \sigma_{\hat{F}} = \sqrt{F(1-F)/N} = O(N^{-1/2}) \tag{1}$$

It is reasonable to make the error approximately 0.01, and we should choose N=10000.

Graph

The range of x was chosen as $[0, 1]$ with interval 0.01. The empirical distribution function is 1 for $x \geq 1$ and 0 for $x \leq 1$.

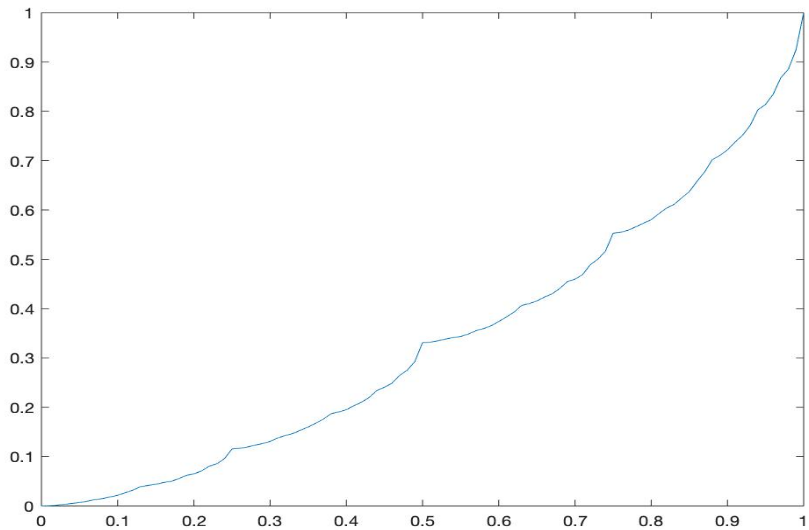

Figure 1.1: Plot of $\hat{F}\ for\ x \in [0, 1]$

Question 2

Without loss of generality, we shall write x as

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

Calculation

In this and the following question, only consider binary expansion without an infinite tail of 1's.

Any number smaller than or equal to x has the first n digits of binary expansion being a smaller number than x, so it is justified to consider only the first n digits . The relation: $x \sim y,\ x, y \in [0, 1]$ if they have the same first n digits in binary expansion is an equivalence relation, which is easy to verify.

For numbers less than or equal to x, there are exactly

$$\sum_{i=1}^{n} 2^n \frac{x_i}{2^i} = 2^n x$$

equivalence classes.
Consider digit-wise, let k denote the number of 1's in its first n digits, or u denote the number of 0's in its first n digits. The probability of obtaining such an equivalence class is $p^k q^{n-k}$ *or* $p^{n-u} q^u$. Hence the distribution function is as follows:
For x as given,

$$F(x) = \sum_{i=1}^{2^n x} (q)^n (\frac{p}{q})^{k_i}$$

alternatively,

$$F(x) = \sum_{i=1}^{2^n x} (p)^n (\frac{q}{p})^{u_i}$$

Where k_i, u_i are the numbers of 1's and 0's in the first n digits of the ith equivalence class representation respectively.

Question 3

The graph of F can be found at **Figure 2**.
The comparison graph is **Figure 3**.
The program is **Program 3**.

From **Figure 3**, it is obvious that the empirical distribution function gives an overestimate of the value of distribution function at each of x with a finite binary expansion. F is smoother than $\hat{F}$, and both functions are continuous and increasing by plot.
Quantitatively, observe that if the first digit of x is 0, i.e. $x < 1$, then

$$F(\frac{x}{2}) = qF(x)$$

, otherwise

$$F(\frac{x+1}{2}) = pF(x) + q$$

Hence,

$$F(x) = \begin{cases} qF(2x) & 0 < x < 0.5 \\ pF(2x-1) + q & 0.5 < x < 1 \end{cases}$$

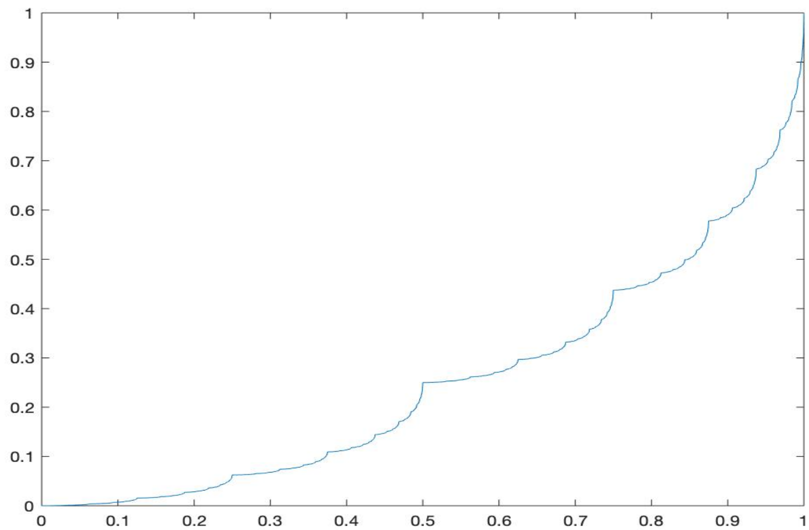

Figure 3.1: Plot of $F\ for\ x \in [0, 1]$

Time complexity

The time complexity of the loop of Program 1 for general n, N is $\mathcal{O}(nN)$, and that of Program 3 is $\mathcal{O}\left(\sum_{i=1}^{2^n} y(i)\right) = \mathcal{O}\left(2^{2n-1} + 2^{n-1}\right) = \mathcal{O}\left(2^{2n-1}\right)$. Therefore, for n large, program 3 takes much more time.

Question 4

Claim 1. *If $c \in [0,1]$ has a finite binary expansion, then F is continuous at c.*

Proof. By 1A probability, for X an random variable,

$$\mathbb{P}(\{x = c\}) = F_X(c) - F_X(c^-)$$

and hence if $\mathbb{P}(\{X = c\}) = 0$, X has a continuous distribution function. For c with a finite binary expansion as in Question 3, $0 < p, q < 1$,

$$\mathbb{P}(\{X = c\}) = \lim_{k \to \infty} \prod_{i=1}^{n} (x_i p + (1 - x_i) q)^{x_i} q^k = 0$$

Therefore F is continuous at c. ∎

The plot suggests that F is continuous elsewhere.

Claim 2. *F is continuous.*

Proof. Write

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

, then

$$\mathbb{P}(X = c) = \prod_{i=1}^{\infty} (x_i p + (1 - x_i) q)^{x_i} = p^m q^n$$

for some particular m,n. If $0 < p, q < 1$, it is then enough to show that either m or n is infinite, which is true as x can have an infinite binary distribution anyway (e.g. for c can write $x = \sum_{i=1}^{\infty} \frac{x_i}{2^i} \ with \ x_i = 0 \ for \ i > n)$. ∎

Question 5

The plots below suggest that F is right-differentiable at c, and the derivative is 0. The choice of delta is small enough compared with the value of c. In particular: The proportion of greatest value of delta is $\frac{2^{-11}}{2^{-1} + 2^{-4}} \simeq 0.1\%$, which is suitable.

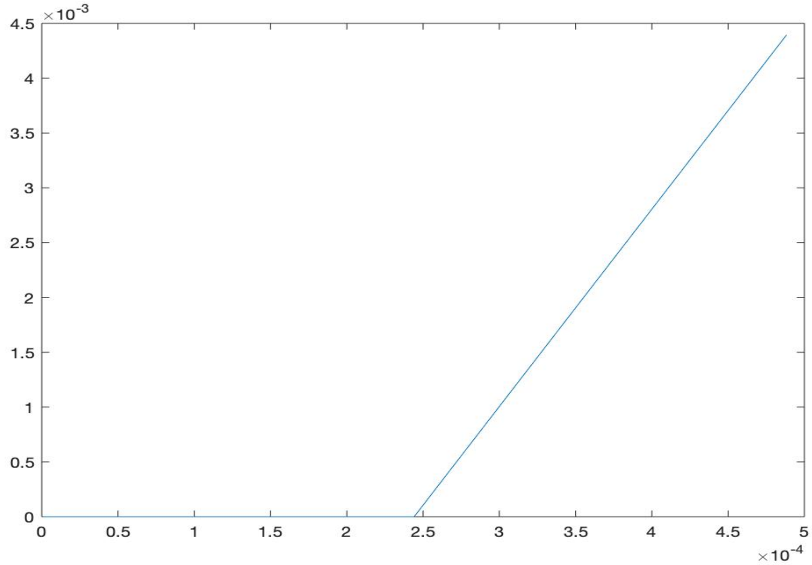

Figure 5.1: Output of Q5(x, 3/4, 11, 9/16) , note that it vanishes at x close to 0.

Question 6

Conjecture 1. *F is right-differentiable at an arbitrary point c with a finite binary expansion.*

Proof. **k_i, u_i used below are as in Question 2.**
If c has a finite binary expansion as in Question 3, then c=$\frac{m}{2^n}$ for some $m \in \mathbb{N}$. Then

$$\lim_{l\to\infty} \frac{F(c+2^{-(n+l)})-F(c)}{2^{-(n+l)}} = \lim_{l\to\infty} \frac{F(\frac{2^l m+1}{2^{n+l}})-F(\frac{2^l m}{2^{n+l}})}{2^{-(n+l)}} = \lim_{l\to\infty} (2q)^{n+l}(\frac{p}{q})^{k_{2^l m+1}}$$

where $k_{2^l m+1}$ is a constant for a fixed c. Hence, the limit exists and is finite when $q < \frac{1}{2}$. Similarly, by approaching from the left,

$$\lim_{l\to\infty} \frac{F(\frac{2^l m-1}{2^{n+l}})-F(\frac{2^l m}{2^{n+l}})}{-2^{-(n+l)}} = \lim_{l\to\infty} (2p)^{n+l}(\frac{q}{p})^{u_{2^l m}}$$

where $u_{2^l m}$ is smaller than n for a fixed c, as the binary expansion of the $2^l m$ th number is of value $m - \frac{1}{2^{n+l}}$ which has at least l 1's out of the n+l

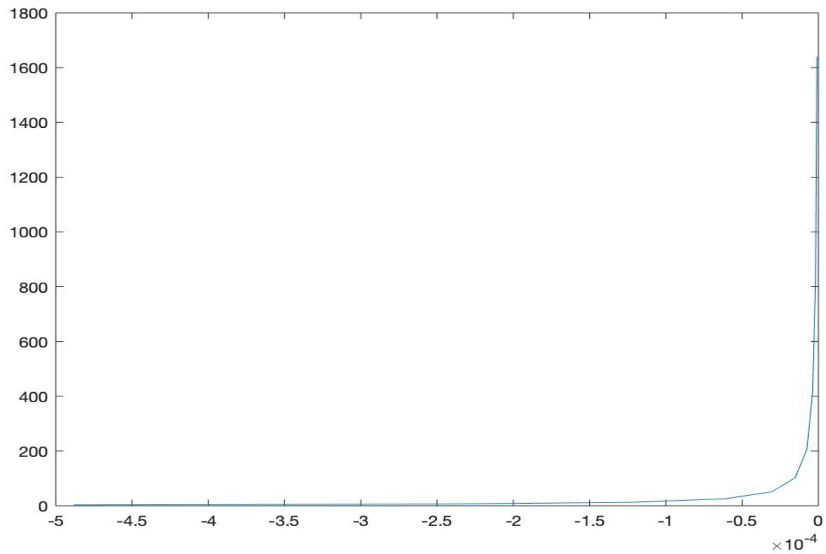

Figure 5.2: Output of Q5(y, 3/4, 11, 9/16)

digits. Hence the limit exists and is finite when $p < \frac{1}{2}$. Therefore, F is left-differentiable when $p < \frac{1}{2}$, and right-differentiable when $p > \frac{1}{2}$. At $p = \frac{1}{2}$, by the distribution function, it is a straight line joining 0 & 1. $\square$

Graphic support

At $c = \frac{10}{16}$, with p=3/7, **Input values:**

```
x = zeros(1, 10);
for i = 1: 10
    x(i) = (2^(-(10+i)) + 10/16);
end

y = zeros(1, 10);
for i = 1: 10
    y(i) = (-2^(-(i+10)) + (10/16));
end
```

The following graphs **6.1**, **6.2** and **6.3** support the conjecture.

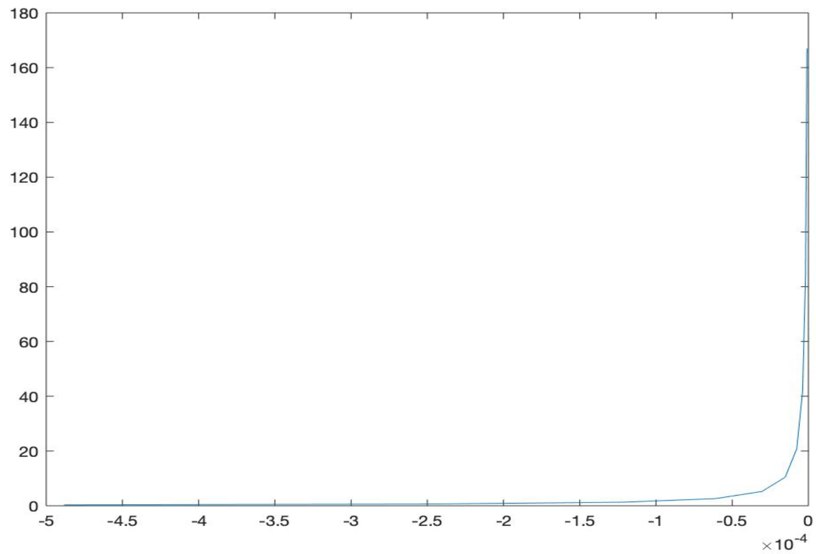

Figure 6.2: Output of Q5 (y, 3/7, 11, 10/16)

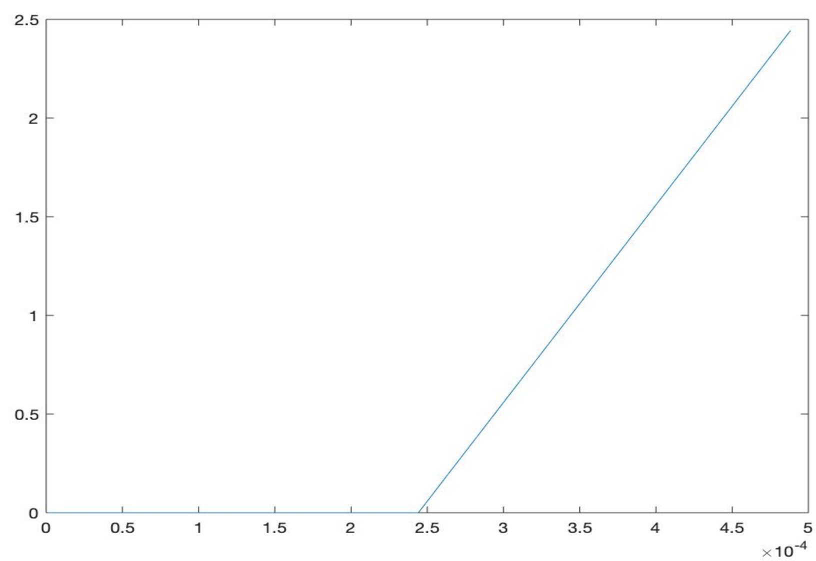

Figure 6.1: Output of Q5 (x, 3/7, 11, 10/16)

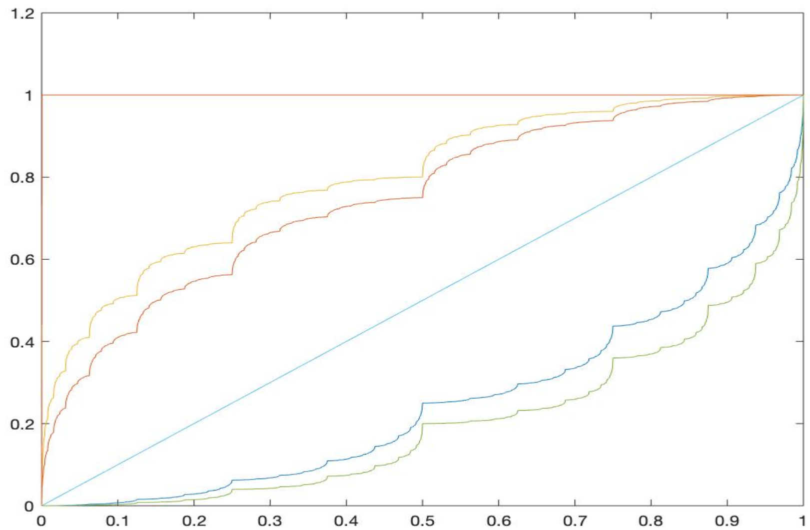

Figure 6.3: Output of Q3 with p=1,0,1/2,1/4,1/5,3/4,4/5

Program 1: Empirical distribution function

```
function [F] = EDF(x, X)
% Empirical distribution function(EDF)
% By Monte Carlo Simulation,
% obtain the EDF.
F = zeros(1, length(x));
% Initialize F.
for i = 1: length(x)
    for j = 1: length(X)
        if X(j) <= x(i)
       F(i) = F(i) + 1;
        end
    end
end
F = F./length(X);
return
```

Program 2: Random Sample

```
function [X] = RandomSample(n,N,p)
```

```
% Sample Generator:
% Generates a random N by n matrix
% with each row a finite
% sequence of 0 and 1's.
U = zeros(N,n);
% Initialize the matrix
for i = 1:N
  for j = 1:n
      U(i,j) = binornd(1,p);
  end
end
X = zeros(1,N);
% Initialize the random sample
for i=1:N
    for j = 1:n
    X(1,i) = X(1,i) + 2^(-j)*U(i,j);
    end
end
return
```

Program 3: Distribution Function

```
function [G] = Q3(x, p, n)
% The distribution function for some values of x
% x is a matrix of finite 0-1 rows;
% F is calculated as shown in report;
% In command window input Q3((0:2^(-11):1), 3/4, 11) to generate the graph
y = x*2^n;
q = 1-p;
z = de2bi((0:1:2^(n+1)));
% here use range up to 2^(n+1) as there is input value greater than 1
F = zeros(1, size(y, 2));
k = sum(z' == 1);
for i = 1: size(y,2)
   for j = 1: y(i)
       F(i) = F(i) + (q)^(n)*(p/q)^(k(j));
   end
end
G = plot(x, F);
return
```

Program 4: Question 5

```
function [G] = Q5(x, p, n, c)
% The distribution function for some values of x
% F is calculated as shown in report;
y = x*2^n;
z = de2bi((0:1:2^(n)));
% here use range up to 2^(n+1)
% as there is input value greater than 1
F = zeros(1, size(y, 2));
delta = x-c;
q = 1-p;
k = sum(z' == 1);

Fc = zeros(1, size(y, 2));
for i = 1: 10
    for j = 1: (2^n)*(c)
        Fc(i) = Fc(i) + (q)^(n)*(p/q)^(k(j));
    end
end
% function value at c
for i = 1: size(y,2)
   for j = 1: y(i)
       F(i) = F(i) + (q)^(n)*(p/q)^(k(j));
   end
end
for i = 1: size(y,2)
    F(i) = (F(i)-Fc(i))/delta(i);
end
G = plot(delta, F);
return
```

Input values:

```
x = zeros(1, 10);
for i = 1: 10
    x(i) = (2^(-(10+i)) + 9/16);
end

y = zeros(1, 10);
for i = 1: 10
```

```
    y(i) = (-2^(-(i+10)) + (9/16));
end
```