New CS 473 Spring 2015
Homework 6 Problem 2

Arash Khatibi (khatibi2)
Liang Tao (ltao3)
Chen Zhang (czhang49)

Define the following variables:
L : The L-th item in the stream.
n : The total number of items in the stream, which is unknown.
Pr(L): The probability that the L-th item is returned as the sample.

1 Part A

For an item numbered L (the L-th item), the possibility to pick up this item is $\frac{1}{L}$ and the possibility that this item is not picked up (line is not executed) is $\frac{L-1}{L}$. This is true because for the L-th item, we are randomly choosing a number from 1 to L, and if the result is 1, the line is executed, otherwise not. So if an item is returned as the sample, it means that line is executed at this item and after this item, the line is not executed until the algorithm ends.
Therefore, the probability that the first item is returned as the sample is

$$\begin{aligned} Pr(1) &= 1 \times \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \ldots \times \frac{n-2}{n-1} \times \frac{n-1}{n} \\ &= \frac{1}{n} \end{aligned}$$

Similarly, the probability that the L-th item will be returned as the sample is

$$\begin{aligned} Pr(L) &= \frac{1}{L} \times \frac{L}{L+1} \times \frac{L+1}{L+2} \ldots \times \frac{n-2}{n-1} \times \frac{n-1}{n} \\ &= \frac{1}{n} \end{aligned}$$

This proves that the item returned by GETONESAMPLE(S) is chosen randomly from S.

2 Part B

Define the following function:
F(L) : $F(L) = 1$ if at item L, line is executed. $F(L) = 0$ if at item L, line is not executed.
Therefore, the expected number of line executions can be defined as $E[F(1)+F(2)+F(3)+\ldots+F(n)]$. Since the $F(L)$ function is independent of each other, the following equation holds true:

$$\begin{aligned} E[F(1)+F(2)+F(3)+\ldots+F(n-1)+F(n)] &= E[F(1)]+E[F(2)]+E[F(3)]+\ldots+E[F(n-1)]+E[F(n)] \\ &= 1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n-1}+\frac{1}{n} \end{aligned}$$

Thus the expected number of line executions is the harmonic function.

3 Part C

At item L, if the line is executed for the last time, it means that the L-th item is returned as the sample of the stream S. Since in Part A, we proved that the probability of each item returned as the sample is $\frac{1}{n}$, the expected value of L can be calculated as:

$$\begin{aligned} E[L] &= 1 \times \frac{1}{n} + 2 \times \frac{1}{n} + 3 \times \frac{1}{n} + \ldots + (n-1) \times \frac{1}{n} + n \times \frac{1}{n} \\ &= \frac{1}{n} \times (1 + 2 + 3 + \ldots + (n-1) + (n)) \\ &= \frac{n+1}{2} \end{aligned}$$

4 Part D

Because we know that at item 1, line is executed. So when the line is executed for the second time at another item L, it means that the first situation is that from item 1 to L-1, line is not executed and at item L, line is executed. The second situation is that line is only executed at item 1 and not executed until algorithm ends. Therefore the expected value of L can be calculated from:

$$\begin{aligned} E[L] =& (2 \times \frac{1}{2}) + (3 \times \frac{1}{2} \times \frac{1}{3}) + (4 \times \frac{1}{2} \times \frac{2}{3} \times \frac{1}{4}) + \ldots + (n \times \frac{1}{2} \times \frac{2}{3} \times \ldots \times \frac{n-2}{n-1} \times \frac{1}{n}) \\ &+ (n \times \frac{1}{2} \times \frac{2}{3} \times \ldots \times \frac{n-2}{n-1} \times \frac{n-1}{n}) \end{aligned}$$

The last part of the equation corresponds to the case that the algorithm ends before line is executed for the second time. The rest of the elements in the equations corresponds to the line is executed for the second time before the algorithm ends.
Rearrange the equation we have the following result:

$$E[L] = 1 + \sum_{L=2}^{L=n-1} \frac{1}{L-1}$$