New CS 473: Algorithms, Spring 2015
Homework 7

Due Tuesday, March 31, 2015 at 5pm

All homework must be submitted electronically via Moodle as separate PDF files, one for each numbered problem. Please see the course web site for more information.

---

1. ***Reservoir sampling*** is a method for choosing an item uniformly at random from an arbitrarily long stream of data; for example, the sequence of packets that pass through a router, or the sequence of IP addresses that access a given web page. Like all data stream algorithms, this algorithm must process each item in the stream quickly, using very little memory.

```
GetOneSample(stream S):
  ℓ ← 0
  while S is not done
      x ← next item in S
      ℓ ← ℓ + 1
      if Random(ℓ) = 1
          sample ← x        (⋆)
  return sample
```

At the end of the algorithm, the variable ℓ stores the length of the input stream S; this number is *not* known to the algorithm in advance. If S is empty, the output of the algorithm is (correctly!) undefined.

Consider an arbitrary non-empty input stream S, and let n denote the (unknown) length of S.

(a) Prove that the item returned by GetOneSample(S) is chosen uniformly at random from S.

(b) What is the *exact* expected number of times that GetOneSample(S) executes line ($\star$)?

(c) What is the *exact* expected value of ℓ when GetOneSample(S) executes line ($\star$) for the *last* time?

(d) What is the *exact* expected value of ℓ when either GetOneSample(S) executes line ($\star$) for the *second* time (or the algorithm ends, whichever happens first)?

(e) Describe and analyze an algorithm that returns a subset of k distinct items chosen uniformly at random from a data stream of length at least k. The integer k is given as part of the input to your algorithm. Prove that your algorithm is correct.

For example, if $k = 2$ and the stream contains the sequence $\langle \spadesuit, \heartsuit, \diamondsuit, \clubsuit \rangle$, the algorithm would return the subset $\{\diamondsuit, \spadesuit\}$ with probability $1/6$.

New CS 473 Spring 2015 — Homework 7 Problem 1

Name:	NetID:
Name:	NetID:
Name:	NetID:

(a) Prove that the item returned by GetOneSample(S) is chosen uniformly at random from S.

(b) What is the *exact* expected number of times that GetOneSample(S) executes line ($\star$)?

(c) What is the *exact* expected value of ℓ when GetOneSample(S) executes line ($\star$) for the *last* time?

(d) What is the *exact* expected value of ℓ when either GetOneSample(S) executes line ($\star$) for the *second* time (or the algorithm ends, whichever happens first)?

(e) Describe and analyze an algorithm that returns a subset of k distinct items chosen uniformly at random from a data stream of length at least k. The integer k is given as part of the input to your algorithm. Prove that your algorithm is correct.
