Dealing with Massive Data February 8, 2011

Homework 1: Due February 28, 8pm

For each of the problems below you must prove your answer correct. You are encouraged to discuss problems with each other in small groups (2-3 people), as long as you list all discussion partners on your problem set. Discussion of homework problems may include brainstorming and verbally walking through possible solutions, but should not include one person telling the others how to solve the problem. In addition, each person must write up their solutions entirely on their own; you may not look at another student's written solutions. Moreover, all materials you consult must be appropriately acknowledged.

Question 1 (Probability)

You are given a stream $\mathcal{Y} = y_1, y_2, \ldots$ on n elements $[1, n]$ s.t. for any element $j \in [1, n]$, $\textbf{Prob}[y_i = j] = {}^1\!/_n$. In other words, each of the n elements is equally likely to appear in every position. Give the best bound you can on a threshold t so that all n elements have appeared among the first t elements of the stream $(y_1, y_2, \ldots, y_t)$ with probability at least $1 - {}^1\!/_{n^2}$.

Question 2 (Sampling)

In class we mentioned the idea of sampling from a stream. In this question we specify how to do this formally. Suppose you are given a stream of unknown length $\mathcal{X} = x_1, x_2, \ldots$. Describe an algorithm that maintains a single sample S so that at every time $t \geq 1$,

$$\textbf{Prob}\left[S = \{i\}\right] = \frac{f_i}{t},$$

where f_i is the number of times i appeared in the stream before time t (formally, $f_i = \sum_{j \leq t : x_j = i} 1$). Note that this is the same as saying that S is a random sample from elements in $x_1, \ldots, x_t$. Prove the correctness of your algorithm.

Question 3 (Counting)

As we saw in lecture 3, the Count-Sketch algorithm provides an $\epsilon\sqrt{F_2}$ approximation guarantee, where $F_2 = \sum_j f_j^2$ is the second frequency moment. In this question you will show how to approximate F_2.

(a) Let $g(\cdot)$ be a hash function $g : [1, n] \to \{+1, -1\}$. Consider the following algorithm. Maintain a counter Z, and upon seeing element x_i, increment the counter $Z \leftarrow Z + g(x_i)$. Show that $\mathbb{E}[Z^2] = F_2$.

(b) Show that the variance of the estimate, $\mathbf{var}[Z^2] \leq 2F_2^2$.

(c) Show how to boost the success probability of a single such counter to get an estimate $\hat{F}_2$ so that

$$\mathbf{Prob}\left[|\hat{F}_2 - F_2| \geq \epsilon F_2\right] \leq \delta.$$

Question 4 (More Counting)

In class in lecture 3 we showed that $Z \cdot g(x_j)$ is an unbiased estimate of f_j where $Z = \sum_{x_i \in \mathcal{X}} g(x_i)$ is the global counter.

(a) Compute the variance of the estimate $\mathbf{var}[Z \cdot g(x_j)]$.

(b) Suppose we computed k such estimates independently. Derive a new unbiased estimator with $1/k$ the variance of the estimate in part (a).

(c) Show how to use the estimator in part (b) to derive a new estimate to f_j that has an error of more than $\epsilon\sqrt{F_2}$ with probability at most $1 - \delta$.

(d) How much space does your final estimator use? How much time does it take to update the estimate after reading a new element? How does it compare with the estimator derived in class?