CS 49/149
Fall 2011
Data Stream Algorithms

Homework 1
Due Wed Oct 12, 5:00pm

Prof. Amit Chakrabarti
Computer Science Department
Dartmouth College

General Instructions: Please write concisely, but rigorously, and show your calculations explicitly, as we do in class. Each problem is worth 5 points, and only "nearly flawless" solutions will earn full credit.

Honor Principle: You are allowed to discuss the problems and exchange solution ideas with your classmates. But when you write up any solutions for submission, you must work alone. You may refer to any textbook you like, including online ones. However, you may not refer to published or online solutions to the specific problems on the homework. *If in doubt, ask the professor for clarification!*

Constructive Examples of 2-Universal Hashing

Let X and Y be finite sets and let Y^X denote the set of all functions from X to Y. We will think of these functions as "hash" functions. [The term "hash function" has no formal meaning; strictly speaking, one should say "family of hash functions" or "hash family" as we do here.] A family $\mathscr{H} \subseteq Y^X$ is said to be 2-universal if the following property holds, with $h \in_R \mathscr{H}$ picked uniformly at random:

$$\forall\, x, x' \in X \;\forall\, y, y' \in Y \left(x \neq x' \;\Rightarrow\; \Pr_h \left[h(x) = y \wedge h(x') = y' \right] = \frac{1}{|Y|^2} \right) .$$

We shall give two examples of 2-universal hash families from the set $X = \{0,1\}^n$ to the set $Y = \{0,1\}^k$ (with $k \leq n$).

1. Treat the elements of X and Y as column vectors with 0/1 entries. For a matrix $A \in \{0,1\}^{k \times n}$ and vector $b \in \{0,1\}^k$, define the function $h_{A,b} : X \to Y$ by $h_{A,b}(x) = Ax + b$, where all additions and multiplications are performed mod 2.

   Prove that the family of functions $\mathscr{H} = \{h_{A,b} : A \in \{0,1\}^{k \times n}, b \in \{0,1\}^k\}$ is 2-universal.

2. Identify X with the finite field $GF(2^n)$. For elements $a, b \in X$, define the function $g_{a,b} : X \to Y$ as follows:

   $$\begin{aligned} g_{a,b}(x) &= \text{rightmost } k \text{ bits of } f_{a,b}(x), \quad \text{where} \\ f_{a,b}(x) &= ax + b, \quad \text{with addition and multiplication performed in } GF(2^n). \end{aligned}$$

   Prove that the family of functions $\mathscr{G} = \{g_{a,b} : a, b \in GF(2^n)\}$ is 2-universal. Is the family $\mathscr{G}$ better or worse than $\mathscr{H}$ in any sense? Why? [Note: If you are unfamiliar with finite fields, we should discuss this topic outside of class.]

Applications of 2-Universal Hashing

3. Let $\mathscr{H} \subseteq Y^X$ be a 2-universal hash family, with $|Y| = cM^2$, for some constant $c > 0$. Suppose we use a random function $h \in_R \mathscr{H}$ to hash a stream σ of elements of X, and suppose that σ contains at most M distinct elements. Prove that the probability of a collision (i.e., the event that two distinct elements of σ hash to the same location) is at most $1/(2c)$.

4. Recall that we said in class that the buffer B in the BJKST Algorithm for DISTINCT-ELEMENTS can be implemented cleverly by not directly storing the elements of the input stream in B, but instead, storing the hash values of these elements under a secondary hash function whose range is of size cM^2, for a suitable M.

   Using the above result, flesh out the details of this clever implementation. (One important detail that you must describe: how do you implement the buffer-shrinking step?) Plug in $c = 3$, for a target collision probability bound of $1/(2c) = 1/6$, and figure out what M should be. Compute the resulting upper bound on the space usage of the algorithm. It should work out to

   $$O\left(\log n + \frac{1}{\varepsilon^2} \left(\log \frac{1}{\varepsilon} + \log \log n \right) \right) .$$

On Grading: Never forget that the purpose of a course is to broaden and deepen your knowledge, not to "collect" points. With this in mind, please refrain from writing up half-baked ideas in the hope of getting "some" credit.