Your name
Your student number

CS 466/666: ALGORITHM DESIGN AND ANALYSIS
PROBLEM SET 3

DUE ON: OCTOBER 18, 2017

ACKNOWLEDGEMENTS.

Acknowledge all collaborators with whom you discussed any of the problems in this problem set and all external sources that you may have used in the completion of the assignment. See the website for the full homework policy.

1. PROBLEM 1: HASHING AND CHAINING

(a) Show that we don’t even need linked lists when the number of hash values is large enough.

**Theorem 1.** Let $H = \{h : U \rightarrow [5n^2]\}$ be a universal family of hash functions. For any sequence $x_1, \ldots, x_n \in U$ of $n$ items, when $h \in H$ is drawn uniformly at random then with probability at least $\frac{9}{10}$ all the hash values $h(x_1), \ldots, h(x_n)$ are distinct.

**Proof.** Enter your answer here. □

(b) A family $H$ of hash functions mapping $U$ to $[n]$ is pairwise independent if for every distinct $x, x' \in U$ and every $y \in [n]$, $\Pr_{h \in H}[h(x) = h(x') = y] = \frac{1}{n^2}$.¹

**Theorem 2.** Let $H = \{h : U \rightarrow [n]\}$ be a pairwise independent family of hash functions. For any sequence $x_1, \ldots, x_n \in U$ of $n$ items, when $h \in H$ is drawn uniformly at random then with probability at least $\frac{99}{100}$, every $y \in [n]$ satisfies $|\{i \leq n : h(x_i) = y\}| \leq 10\sqrt{n}$.

**Proof.** Enter your answer here. □

(c) (Bonus.) Show that the bound in the last part is tight.

**Theorem 3.** There exists a pairwise independent family of hash functions $H = \{h : U \rightarrow [n]\}$ and a sequence $x_1, \ldots, x_n \in U$ of $n$ items such that when $h \in H$ is drawn uniformly at random then with probability at least $\frac{9}{10}$, there is some value $y \in [n]$ such that $|\{i \leq n : h(x_i) = y\}| = \Omega(\sqrt{n})$.

**Proof.** Enter your answer here. □

¹As an aside: how does this definition compare to universality?
2. Problem 2: Bloom filters

Recall that we say a randomized algorithm solves the Set Membership problem if, after seeing \( n \) objects \( x_1, \ldots, x_n \in U \) and then receiving a query \( a \in U \), it (i) correctly answers YES when \( a = x_i \) for some \( i \leq n \), and (ii) correctly answers NO with probability at least \( \frac{2}{3} \) when \( a \) is distinct from the \( n \) elements \( x_1, \ldots, x_n \). In class, we saw a simple algorithm that solves the Set Membership problem:

**Algorithm 1:** Simple set membership algorithm

- Draw \( h : U \to [n] \) uniformly at random from the universal family \( H \);
- Initialize \( T \in \{0, 1\}^{3n} \) to zero;
- for \( i = 1, \ldots, n \) do
  - Set \( T[h(x_i)] = 1 \);
- On query \( a \in U \), output YES iff \( T[h(a)] = 1 \);

That algorithm also requires \( 3n \) bits of memory to store the table \( T \). Show that it is possible to modify the algorithm to use a smaller hash table.

(a) Establish the following theorem (by defining the algorithm and completing its analysis).

**Theorem 4.** For every pairwise independent\(^2\) family of hash functions \( H = \{h : U \to [2\sqrt{3}n]\} \), there is a variant of Algorithm 1 that solves the Set Membership problem with a hash table \( T \) of size \( [2\sqrt{3}n] \).

**Proof.** Enter your answer here. \( \square \)

(b) Is this actually an improvement over the simple set membership algorithm?

Enter your answer here.

(c) **(Bonus.)** Show that there is some constant \( c < 3 \) such that when \( H_{\text{all}} = \{h : U \to [cn]\} \) is the family that includes all functions mapping \( U \) to \( [cn] \), then there is another variant of Algorithm 1 that solves the Set Membership problem with a hash table \( T \) of size \( [cn] \).

Enter your theorem and proof here.

\(^2\)See Problem 1 for the definition of pairwise independent families of hash functions.
3. Problem 3: Almost-universal families of hash functions

Chernoff’s bound gives exponentially small probabilities that the sum of independent random variables ends up being far from its expected value. One of the forms of this bound is as follows.

**Lemma 1** (Chernoff (lower) bound). Let \( X = X_1 + X_2 + \cdots + X_t \) where \( X_1, \ldots, X_t \) are independent random variables that are each drawn from the same distribution, and let \( \mu = \mathbb{E}[X] \). Then for any \( \delta > 0 \),

\[
\Pr[X < (1 - \delta)\mu] \leq e^{-\mu\delta^2/2}.
\]

(a) Use Chernoff’s bound to prove the following strengthening of the claim made in class:

**Theorem 5.** There exists a \( \frac{3}{4} \)-universal family of hash functions \( H = \{ h : \{0, 1\}^n \to \{0, 1\} \} \) of size \( |H| = 16n \).

**Proof.** Enter your answer here. \( \square \)

**Remark.** Note also that the family obtained in this way is of size linear in \( n \), not \( O(n^2) \), as we originally claimed. Does this make a significant difference in applications, or at least in the ones we have seen?

(b) **(Bonus.)** For any prime \( p \geq 2 \) and integers \( m \geq n \geq 2 \), define the hash family \( H_{p,n} = \{ h_a : [m] \to [n] \}_{0 \leq a < p} \) by setting

\[
h_a(x) = (a \cdot x \pmod{p}) \pmod{n}
\]

for each \( a = 0, 1, 2, \ldots, p - 1 \). Prove that \( H_{p,n} \) is \( \frac{2}{n} \)-universal.

**Theorem 6.** The family of hash functions \( H_{p,n} \) defined above is \( \frac{2}{n} \)-universal.

**Proof.** Enter your answer here. \( \square \)

**Extra-bonus question.** Modify the definition of \( H_{p,n} \) to obtain a (truly) universal family of hash functions.