Pairwise Independence and Sampling

A set of random variables $X_1, X_2, \ldots$ is said to be pairwise independent if for all $i \neq j$, $X_i$ and $X_j$ are independent.

Example: $X$ and $Y$ be independent r.v. each taking values -1 and 1 with probability 1/2 and let $Z = XY$.

**Theorem 1.** Let $n$ be a prime number and $\mathbb{Z}_n$ denote the field of integers modulo $n$. For $a$ and $b$ chosen independently and uniformly at random from $\mathbb{Z}_n$, let

$$Y_i = ai + b \pmod{n}$$

Then for $i \neq j \pmod{n}$, $Y_i$ and $Y_j$ are uniformly distributed on $\mathbb{Z}_n$ and pairwise independent.

**Proof.** Let $i \neq j$. Then in the field $\mathbb{Z}_n$, for any given fixed values of $y_i$ and $y_j$ we can solve the equations:

$$y_i = ai + b \pmod{n}$$
\[ y_j = a j + b (mod \ n) \]

uniquely for \( a \) and \( b \).

\[ a = (y_i - y_j)((i - j)^{-1} mod \ n) mod \ n \]

\[ b = (y_i - ak) mod \ n \]

That is, there is 1-1 correspondence between pairs \((a, b)\) and \((y_i, y_j)\).

If we pick \((a, b)\) uniformly at random, any pair \((y_i, y_j)\) is equally likely. \(\square\)
Theorem 2. Let $X_1, X_2, \ldots, X_m$ be pairwise independent r.v.s and $X = \sum_{i=1}^{m} X_i$. Then

$$Var(X) = \sum_{i=1}^{m} Var(X_i).$$

Covariance of two random variables is defined as:

$$Cov(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$$

$$= E(X_iX_j) - E[X_i]E[X_j]$$

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

If $X_i$ and $X_j$ are independent $Cov(X_i, X_j) = 0$.

In general, $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{i<j} Cov(X_i, X_j)$
Application

We consider a class of randomized algorithms called RP.

An RP algorithm $A$ for deciding whether input strings $x$ belong to a language $L$ is as follows:

Given $x$, $A$ picks a random $r$ from $\mathbb{Z}_n = \{0, \ldots, n-1\}$ for some prime $n$. Computes a binary function $f(x, r)$ such that:

If $x \in L$, then $f(x, r) = 1$ for at least half the possible values of $r$.

If $x \notin L$ then $f(x, r) = 1$ for all choices of $r$.

Repeating the algorithm $t$ times using $t$ random samples from $\mathbb{Z}_n$ the probability of error is at most $2^{-t}$. Uses $\Omega(t \log n)$ random bits.
Probability Amplification

How much can reduction in error by using only 2 samples from $\mathbb{Z}_n$?

Choose $a$ and $b$ independently and uniformly at random.

Let $r_i = a_i + b(mod \ n)$ and compute $f(x, r_i)$, $1 \leq i \leq t$.

What is the error probability?

$Y = \sum_{i=1}^{t} f(x, r_i)$

$E[Y] \geq t/2$ and $Var(Y) \leq t/4$.

$Pr(Y = 0) \leq Pr(|Y - E[Y]| \geq t/2) \leq 1/t$.

Improvement over the naive bound of $1/4$. 
Given a set of possible keys $U$, such that $|U| = u$ and a table of $m$ entries, a **Hash function** $h$ is a mapping from $U$ to $M = \{1, ..., m\}$.

A **collision** occurs when two hashed elements have $h(x) = h(y)$.

**Definition 1.** A hash function $h : U \rightarrow M$ is **perfect** for a set $S$ if it causes no collisions for pairs in $S$.

For any given $S$ such that $|S| \leq m$ there is a perfect hash function.

For any $S$ such that $|S| > m$ there is **no** perfect hash function.
Chaining

$h(.)$ - hash function.

A table $T[1..m]$ such that $T[k]$ is a pointer to a linked list of all the elements hashed to $T[k]$.

Insert $k$: add $k$ to the linked list $T[h(k)]$.

Search/delete $k$: search (+ delete) in $T[h(k)]$.

The cost is proportional to the length of the link lists.
Hash Functions

\[ h(k) = k \mod m \]

\[ h(k) = (ak + b) \mod m, \]

\[ H = \{ h(k) \mid 1 \leq a \leq m - 1, \ 0 \leq b \leq m - 1 \} \]

If \( m \) not a prime, let \( p > m \) be a prime

\[ h(k) = (((ax + b) \mod p) \mod m) \]
Analysis of Hashing with Chaining

Let $n$ be the number of keys stored in the table.

The load factor $\alpha = \frac{n}{m}$.

Worst case insert time $O(1)$.

Worst case search/delete time $O(n)$.

For simple probabilistic analysis:

**Simple Uniform Assumption:** Keys are hashed to uniformly random and independent locations.

Assume that $h(.)$ is computed in $O(1)$ time.
Theorem 3. In a hash table in which collisions are resolved by chaining, under the assumption of simple uniform hashing,

1. An unsuccessful search takes $\Theta(1 + \alpha)$ expected time.

2. A successful search takes $\Theta(1 + \alpha)$ expected time.
Universal Hash Functions

Definition 2. A family $H$ of hash functions from $U$ to $M$ is universal (2-universal) if for all $x, y \in U$, such that $x \neq y$, and for a randomly chosen function $h$ from $H$

$$Pr(h(x) = h(y)) \leq \frac{1}{m}.$$ 

Let $H$ be the set of all functions from $U$ to $M$, then $H$ is universal.

**Problem:** There are $m^u$ functions from $U$ to $M$ - requires $u \log m$ bits to choose, represent and store as a table.
**Theorem 4.** Assume that we hash $n$ keys to a table of size $m$, $n \leq m$, using a hash function $h$ chosen at random from a universal family of hash functions, and we resolve collisions by chaining. Then searching for a key takes expected time at most $1 + \alpha$.

**Proof.** For every pair of distinct keys $k$ and $l$, $X_{kl} = 1$ iff $h(k) = h(l)$, else 0.

$$E[X_{kl}] \leq 1/m.$$ 

For each key $k$, define the r.v. $Y_k$ that equals the number of keys other than $k$ that hash to the same slot as $k$:

$$Y_k = \sum_{l \in T, l \neq k} X_{kl}$$

Thus $E[Y_k] = \sum_{l \in T, l \neq k} E[X_{kl}] \leq \sum_{l \in T, l \neq k} 1/m$

If key $k \notin T$: $E[Y_k] \leq n/m = \alpha$

If key $k \in T$: $E[Y_k] \leq (n - 1)/m + 1 < 1 + \alpha \quad \Box$
Corollary 1. Using universal hashing and collision resolution by chaining in a table with $m$ slots, it takes expected time $\Theta(s)$ to handle any sequence of $s$ INSERT, SEARCH, and DELETE operations containing $O(m)$ INSERT operations.
Constructing universal hash functions

Choose a prime number $p$ such that $0 \leq k \leq p - 1$.

Let $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$ and $\mathbb{Z}_p^* = \{1, 2, \ldots, p-1\}$.

Let

$$h_{a,b}(k) = ((ak + b) \mod p) \mod m$$

be a hash function for any $a \in \mathbb{Z}_p^*$ and any $b \in \mathbb{Z}_p$.

Define the set of hash functions:

$$\mathcal{H}_{p,m} = \{h_{a,b} : a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p\}$$
Theorem 5.  $\mathcal{H}_{p,m}$ is universal.

Proof. Consider two distinct keys $k$ and $l$ from $\mathbb{Z}_p \setminus \{\}$. For a given hash function $h_{a,b}$ we let

$$r = (ak + b) \mod p$$

$$s = (al + b) \mod p$$

$$r \neq s.$$

Moreover, each distinct pair $(a, b)$ with $a \neq 0$ yields a distinct pair $(r, s)$ with $r \neq s$.

Thus the probability that keys $k$ and $l$ collide is equal to the probability that $r = s \mod m$ with $r \neq s$.

For a given $r$, the number of values of $s$ such that $r \neq s$ and $r = s \mod m$ is at most $\lceil p/m \rceil - 1 \leq (p + m - 1)/m - 1 = (p - 1)/m$.

Thus probability that $s$ collides with $r$ is $\leq 1/m$. $\square$