Lecture 8 - Message Authentication and Universal Hashing

Jan Bouda

FI MU

April 28, 2010

Part I

k-wise independent random variables

k-wise independence

Definition

Random variables $X_1, X_2, \ldots, X_n$ are k-**wise independent** iff for any $I \subseteq \{1, \ldots, n\}$ with $|I| \leq k$ and for any values x_i, $i \in I$, it holds that

$$\mathcal{P}\left(\bigwedge_{i \in I} X_i = x_i\right) = \prod_{i \in I} \mathcal{P}(X_i = x_i). \tag{1}$$

For $k = 2$ we say that random variables are **pairwise independent**.

Advantage of pairwise independent random variables is that they require much less randomness to construct, in contrast to independent random variables.

Constructing Pairwise Independent Bits

Let $X_1, \dots, X_b$ be uniformly distributed independent random variables on $\{0,1\}$. Let $S_j \subseteq \{1,\dots,b\}$, $S_j \neq \emptyset$ be a nonempty set of indices, there are $2^b - 1$ such subsets. Let us define random variables

$$Y_j = \bigoplus_{i \in S_j} X_i \tag{2}$$

as the XOR of X_i's.

Theorem

Random variables $Y_1, Y_2, \dots, Y_{2^b-1}$ are uniform and pairwise independent.

Constructing Pairwise Independent Bits

Proof.

First we have to show that Y_j is uniform for any j. We will do so using the principle of deferred decision. Let $z = \max S_j$. Then

$$Y_j = \left(\bigoplus_{i \in S_j \setminus \{z\}} X_i \right) \oplus X_z. \tag{3}$$

Suppose we know values of all X_i, $i \in S_j \setminus \{z\}$. Then the value of Y_j is determined by the value of X_z, and the probabilities are $\mathcal{P}(Y_j = 0) = \mathcal{P}(Y_j = 1) = 1/2$. $\square$

Constructing Pairwise Independent Bits

Proof.

Next we have to show the pairwise independence. Consider any Y_k and Y_l together with the corresponding index sets S_k and S_l. Assume WLOG that $z \in S_l \setminus S_k$ and let us calculate

$$\mathcal{P}(Y_l = d | Y_k = c) \tag{4}$$

for any $c, d \in \{0, 1\}$. We use again the principle of deferred decision. Suppose that we know all values of X_i, $i \in (S_k \cup S_l) \setminus \{z\}$. This completely determines the value of S_k, but we need X_z to determine the value of S_l. This gives

$$\mathcal{P}(Y_l = d | Y_k = c) = \mathcal{P}(Y_l = d) = \frac{1}{2} \tag{5}$$

for any $c, d \in \{0, 1\}$ showing the pairwise independence. $\square$

Constructing Pairwise Independent Integers

In a much analogous way we may construct pairwise independent random variables $Y_0, Y_1, \ldots, Y_{p-1}$ uniformly taking integer values modulo p (for some prime p). We need two independent uniform random variables X_1 and X_2 over $\{1, \ldots, p-1\}$ and set

$$Y_i = X_1 + iX_2 \bmod p \text{ for } i = 0, \ldots, p-1. \tag{6}$$

Theorem

Random variables $Y_0, Y_1, \ldots, Y_{p-1}$ are uniform and pairwise independent.

Constructing Pairwise Independent Integers

Proof.

By the principle of deferred decisions, random variables Y_i are uniform. Given X_2, all uniformly distributed values of X_1 imply uniform distribution on all possible values of Y_i.

Consider any pair of random variables Y_i and Y_j. We would like to show that, for any $a, b \in \{1, \ldots, p-1\}$,

$$\mathcal{P}(Y_i = a \vee Y_j = b) = \frac{1}{p^2}. \tag{7}$$

The event $[Y_i = a] \cup [Y_j = b]$ is equivalent to

$$X_1 + iX_2 \equiv a \pmod{p} \text{ and } X_1 + jX_2 \equiv b \pmod{p}. \tag{8}$$

□

Constructing Pairwise Independent Integers

Proof.

We have a system of two linear equations with the unique solution

$$X_2 = \frac{b-a}{j-i} \bmod p \text{ and } X_1 = a - \frac{i(b-a)}{j-i} \bmod p. \tag{9}$$

X_1 and X_2 are uniform and independent, determining the probability of this event to be $\frac{1}{p^2}$ as desired. □

This proof can be easily extended to show that it suffices to have

Part II

Graphs: Finding Large Cuts

Probabilistic method

The following theorem is a special case of the probabilistic method. It establishes the fact, that there is at least one value in $\mathbf{Im}(X)$ greater or equal to $E(X)$ and at least one value smaller or equal to $E(X)$.

Theorem

Suppose we have a random variable X with $E(X) = \mu$. Then $\mathcal{P}(X \leq \mu) > 0$ and $\mathcal{P}(X \geq \mu) > 0$.

Probabilistic Method

Proof.

Recall that

$$\mu = E(X) = \sum_{x \in \mathbf{Im}(X)} x\mathcal{P}(X = x).$$

If $\mathcal{P}(X \geq \mu) = 0$, we have

$$\begin{aligned} \mu = \sum_{x \in \mathbf{Im}(x)} x\mathcal{P}(X = x) &= \sum_{x \in \mathbf{Im}(X), x < \mu} x\mathcal{P}(X = x) \\ < \sum_{x \in \mathbf{Im}(X), x < \mu} \mu\mathcal{P}(X = x) &= \mu, \end{aligned}$$

obtaining a contradiction. □

Proof.

Similarly for $\mathcal{P}(X \leq \mu) = 0$ we have

$$\mu = \sum_{x \in \mathbf{Im}(x)} x\mathcal{P}(X = x) = \sum_{x \in \mathbf{Im}(X), x > \mu} x\mathcal{P}(X = x)$$
$$> \sum_{x \in \mathbf{Im}(X), x > \mu} \mu\mathcal{P}(X = x) = \mu,$$

□

Existence of Large Cuts

Given a (not oriented) graph $G = (V, E, f)$ a cut of the graph is a partitioning V into two sets A and $B = V \setminus A$. Weight of the cut is the sum of weights of edges connecting A and B, i.e.

$$\sum_{\substack{\{u,v\}\in E \\ u\in A, v\in B}} f(\{u, v\}).$$

Here we assume that the weight of every edges is equal to 1. The problem of finding maximum cut is NP-hard.
We show, using the probabilistic method, that the values of the maximal cut is at least $|E|/2$.

Theorem

Given a graph $G = (V, E)$ with n nodes and m edges, there is partitioning of V into two disjoint sets A and B such that $m/2$ edges connect a node in A and a node in B.

Existence of Large Cuts

Proof.

Construct sets A and B in the way that you assign each node in V independently and and uniformly either to A or to B. Let $\{e_1, e_2, \ldots e_m\}$ be arbitrary enumeration of the edges of G. For $i = 1, \ldots, m$ we define

$$X_i = \begin{cases} 1 & \text{if edge } i \text{ connects A to B,} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The probability that a particular edge connects A and B is $1/2$ giving

$$E(X_i) = \frac{1}{2}, \tag{11}$$

since for $e_i = \{u, v\}$

$$E(X_i) = \mathcal{P}(X_i = 1) = \mathcal{P}(u \in A \wedge v \in B) + \mathcal{P}(u \in B \wedge v \in A).$$

Using independence of the node assignment we have
$\mathcal{P}(u \in A \wedge v \in B) = \mathcal{P}(u \in B \wedge v \in A) = \mathcal{P}(u \in A)\mathcal{P}(v \in B) = 1/4.$ □

Existence of Large Cuts

Proof.

Let $c(A, B)$ be a random variable (function of A and B) denoting the value of the cut corresponding to A and B. Then

$$E(c(A,B)) = E\left(\sum_{i=1}^{m} X_i\right) = \sum_{i=1}^{m} E(X_i) = \frac{m}{2}. \tag{12}$$

Using the previous theorem we obtain the required result. □

A **Las Vegas** algorithm is a randomized algorithm that always gives correct results. We will use the last theorem to design a Las Vegas algorithm that finds a cut of the size at least $m/2$.

Finding Large Cuts

Require: Graph $G = (V, E)$, $V = \{v_1, \dots, v_n\}$

repeat
  $A \leftarrow \emptyset$
  $B \leftarrow \emptyset$
  $r = (r_1, \dots, r_n) \xleftarrow{\text{independently and randomly}} \{0,1\}^n$
  for $i = 1, \dots, n$ **do**
    if $r_i = 0$ **then**
      $A \leftarrow A \cup \{v\}$
    else
      $B \leftarrow B \cup \{v\}$
    end if
  end for
until $c(A, B) \geq m/2$ $\quad \triangleright$ $c(A, B)$ can be evaluated in polynomial time

Finding Large Cuts

Theorem

The expected number E of the repeat cycle executions is at most $\lceil m/2 \rceil$.

Proof.

Let

$$p = \mathcal{P}\left(c(A,B) \geq \frac{m}{2}\right). \tag{13}$$

Then

$$\begin{aligned}\frac{m}{2} =& E(c(A,B)) \\ =& \sum_{i \leq m/2-1} i\mathcal{P}(c(A,B)=i) + \sum_{i \geq m/2} i\mathcal{P}(c(A,B)=i) \\ \leq& (1-p)\left(\frac{m}{2}-1\right) + pm.\end{aligned} \tag{14}$$

$\square$

Finding Large Cuts

Proof.

Finally,

$$p \geq \frac{1}{m/2+1}. \tag{15}$$

Recalling that we are looking for the expected value of a geometric distribution we have

$$E = \frac{1-p}{p} \leq \frac{m/2}{m/2+1}\frac{m/2+1}{1} = m/2. \tag{16}$$

□

Derandomizing the algorithm

Consider now a modified version of the algorithm, where the bits r_i are chosen pairwise independently, but (not necessarily) independently.

- Recall that the only place where we use independence of respective bits r_i is Equation (11), where pairwise independence is sufficient.
- The aforementioned algorithm works with pairwise independent bits as well.
- Let the pairwise independent bits $r_1, \ldots, r_n$ be generated from uniform random bits $X_1, \ldots, X_b$, with $b = \lceil \log_2(n+1) \rceil$, using the aforementioned procedure.
- The algorithm with this random input finds cut of size at least $m/2$ with probability at least $p \geq \frac{1}{m/2+1}$.
- Using the probabilistic method principle, there is an assignment of values $x_1, \ldots, x_b$ to $X_1, \ldots, X_b$ such that the algorithm with this assignment returns a cut of the desired size.

Finally, it suffices to run algorithm sequentially for all $2^{\lceil \log_2(n+1) \rceil}$ possible inputs. Therefore, such an algorithm runs in time $O(mn)$.

Part III

Variance of Pairwise Independent Random Variables

Variance of a Sum

Lemma

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\left(X_i\right) + 2\sum_{i<j} Cov(X_i, X_j).$$

Variance of a Sum

Proof.

We know that this equation holds for n=2. Let us assume that it holds for $n \leq n_0$ and we will show that it holds for $n_0 + 1$.

$$
\begin{aligned}
Var\left(\sum_{i=1}^{n_0+1} X_i\right) &= E\left(\left[\sum_{i=1}^{n_0} X_i + X_{n_0+1} - E\left(\sum_{i=1}^{n_0} X_i + X_{n_0+1}\right)\right]^2\right) \\
&= E\left(\left[\sum_{i=1}^{n_0} X_i + X_{n_0+1} - E\left(\sum_{i=1}^{n_0} X_i\right) - E\left(X_{n_0+1}\right)\right]^2\right) \\
&= E\left(\left[\sum_{i=1}^{n_0} X_i - E\left(\sum_{i=1}^{n_0} X_i\right) + X_{n_0+1} - E\left(X_{n_0+1}\right)\right]^2\right) \\
&= \cdots = Var\left(\sum_{i=1}^{n_0} X_i\right) + Var\left(X_{n_0+1}\right) + 2Cov\left(\sum_{i=1}^{n_0} X_i, X_{n_0+1}\right).
\end{aligned}
$$

Variance of a Sum

Proof.

To complete the proof, observe that

$$Cov\left(\sum_{i=1}^{n_0} X_i, X_{n_0+1}\right) = \sum_{i=1}^{n_0} Cov\left(X_i, X_{n_0+1}\right). \tag{17}$$

□

Variance and Pairwise Independence

Theorem

Let $X = \sum_{i=1}^{n} X_i$, where X_i are pairwise independent. Then

$$Var(X) = \sum_{i=1}^{n} Var(X_i). \tag{18}$$

Theorem directly follows from the fact that the covariance $Cov(X_i, X_j) = 0$ for (pairwise) independent random variables X_i and X_j.

Part IV

Wegman-Carter Hashing

Universal hashing

Definition

Let A and B be sets such that $|A| > |B|$. A family H of hash functions $h : A \to B$ is k-**universal** iff for any $x_1, x_2, \ldots, x_k \in A$ and a hash function $h \in H$ randomly and uniformly chosen from H it holds that

$$\mathcal{P}(h(x_1) = h(x_2) = \cdots = h(x_k)) \leq \frac{1}{|B|^{k-1}}. \tag{19}$$

Applications of k-universal classes are mainly in database hashing and randomness extractors (see later lectures).

Definition

Let A and B be sets such that $|A| > |B|$. A family H of hash functions $h : A \to B$ is **strongly** k-**universal** iff for any $x_1 \neq x_2 \neq \cdots \neq x_k \in A$, any $y_1, y_2, \ldots, y_k \in B$ and a hash function $h \in H$ randomly and uniformly chosen from H it holds that

Universal hashing

For any fixed elements $a_1 \neq a_2 \neq \cdots \neq a_k \in A$ and h selected uniformly from some strongly k-universal hashing family, we have that the induced random variables $X_i = h(a_i)$, $i = 1, \ldots, k$ are k-wise independent.

Following this the strongly k-universal classes are sometimes called k-**wise independent classes** of hash functions. The original name of (strongly) k-universal classes introduce by Wegman and Carter is (strongly) universal$_k$, but we find the k-universal to be more preferable.

The most important application of strongly k-universal classes is that they establish a perfectly secure message authentication (details provided during the practice lectures).

Note that any strongly k-universal H is k-universal as well. Also, strongly k-universal H is strongly l-universal for any $l \leq k$ and k-universal H is l-universal for any $l \leq k$.

Universal Hashing: Example

Let $A = \{0, 1, \ldots, m-1\}$ and $B = \{0, 1, \ldots, n-1\}$ with $m \geq n$. Let $p \geq m$ be some prime. Consider the class of hash functions

$$h_{a,b}(x) = ((ax + b) \bmod p) \bmod n. \tag{21}$$

Let

$$H = \{h_{a,b} | 1 \leq a \leq p-1, 0 \leq b \leq p\}, \tag{22}$$

stressing that $a \neq 0$.

Theorem

H is 2-universal.

Universal Hashing: Example

Proof.

We count the number of function from H for which two distinct elements x_1 and x_2 from A collide. $x_1 \neq x_2$ implies

$$ax_1 + b \not\equiv ax_2 + b \pmod{p},$$

since the opposite occurs only if $a(x_1 - x_2) \equiv 0 \pmod{p}$. However, we know that neither $a \equiv 0 \pmod{p}$ nor $x_1 - x_2 \equiv 0 \pmod{p}$, what implies the equation.

With fixed x_1 and x_2, For every pair $u \neq v \in B$ there exists exactly one pair a, b such that $ax_1 + b \equiv u \pmod{p}$ and $ax_2 + b \equiv v \pmod{p}$. $\square$

Universal Hashing: Example

Proof.

Solving the system of two linear equations we obtain the unique solution

$$a = \frac{v - u}{x_2 - x_1} \bmod p \qquad (23)$$

$$b = u - ax_1 \bmod p. \qquad (24)$$

Since there is exactly one hash function for each pair (a, b), we have there is exactly one hash function in H such that

$$ax_1 + b \equiv u \pmod{p} \text{ and } ax_2 + b \equiv v \pmod{p}.$$

We have that the number of collisions equals to the number of pairs (u, v) from $\{0, \ldots, p-1\}$ satisfying $u \neq v$ and $u \equiv v \pmod{n}$. For each choice of u there are at most $\lceil p/n \rceil - 1$ possible values of v. □

Universal Hashing: Example

Proof.

Together we have that there are at most

$$p(\lceil p/n \rceil - 1) \leq p\left(\frac{p+(n-1)}{n} - \frac{n}{n}\right) = \frac{p(p-1)}{n}.$$

such pairs. Therefore, the collision probability is

$$P(h_{a,b}(x_1) = h_{a,b}(x_2)) \leq \frac{p(p-1)/n}{p(p-1)} = \frac{1}{n}.$$

$\square$