Constructing k-wise Independent Variables

1 Definitions

Recall the setting from last time. We have $X_1, \cdots, X_n$, which are random variables taking values in some set T, and specified by a distribution $D : T^n \to [0, 1]$. D is *pairwise independent* if for all $1 \leq i < j \leq n$, $t_1, t_2 \in T$

$$\mathbb{P}_{X_1,\cdots,X_n \sim D}[X_i = t_1, X_j = t_2] = \mathbb{P}[X_i = t_1]\mathbb{P}[X_j = t_2].$$

We also defined pairwise independence for hash functions $h_s : U \to T$ but if we order U as $\{u_1, \cdots, u_n\}$, then we can get equivalent definitions by taking $X_i = h_s(u_i)$ for all i.

Pairwise independence generalizes to a stronger notion, and we call the resulting scheme *k-wise independence*. D is k-wise independent if for all $i_1, i_2, \cdots, i_k$ (all unique) and $t_1, \cdots, t_k \in T$

$$\mathbb{P}_{X_1,\cdots,X_n \sim D}[X_{i_1} = t_1, \cdots, X_{i_k} = t_k] = \mathbb{P}[X_{i_1} = t_1] \cdots \mathbb{P}[X_{i_k} = t_k].$$

2 A (More) Specific Construction

Last time we discussed a class of pairwise independent hash functions over finite fields. Since not everyone is necessarily comfortable with finite fields, we'll go over a more concrete construction which requires only elementary facts of $\mod p$ arithmetic.

2.1 Modulo Prime Fields

Let p be a prime. Then $\mathbb{Z}_p = \{0, 1, \cdots, p-1\}$ is a field with the operations addition and multiplication $\mod p$. Let our random seed be $s = (a, b) \in \mathbb{Z}_p \times \mathbb{Z}_p$ drawn uniformly at random. Then our hash function $h_s : \mathbb{Z}_p \to \mathbb{Z}_p$ performs the familiar operation

$$h_s(x) = ax + b \mod p.$$

Let us check that this is in fact pairwise independent. Let $x_1, x_2, t_1, t_2 \in \mathbb{Z}_p$ s.t. $x_1 \neq x_2$. What is the probability that $h_s(x_1) = t_1$ and $h_s(x_2) = t_2$? This is the probability that

$$\begin{aligned} a &= (t_1 - t_2)(x_1 - x_2)^{-1} \mod p \\ b &= (t_1 x_2 - t_2 x_1)(x_1 - x_2)^{-1} \mod p \end{aligned}$$

where $q^{-1} \in \mathbb{Z}_p$ is the unique multiplicative inverse of q. Note that this is guaranteed to exist if and only if q is non-zero, and we satisfy this condition in the above expressions since $x_1 \neq x_2$. Since a and b are drawn uniformly and independently from $\mathbb{Z}_p$, the probability that they both take on these values is $1/p^2$.

2.2 Extending to Polynomials

Having generated pairwise independent variables, is it possible to extend this scheme to k-wise independence? The answer is yes. Let p be a prime, and $k \geq 1$ be an integer. Let our random seed be $s = (a_0, a_1, \cdots, a_{k-1}) \in \mathbb{Z}_p^k$ drawn uniformly at random. Then our hash function is given by

$$h_s(x) = \sum_{i=0}^{k-1} a_i x^i \mod p.$$

We can see this is k-wise independent since if we take $x_1, \cdots, x_k \in \mathbb{Z}_p$ (all unique) and $t_1, \cdots, t_k \in \mathbb{Z}_p$, the following system of equations has a unique solution for $a_0, \cdots, a_{k-1}$.

$$\begin{aligned}
\sum_{i=0}^{k-1} a_i x_1^i &\equiv t_1 \mod p \\
\sum_{i=0}^{k-1} a_i x_2^i &\equiv t_2 \mod p \\
&\vdots \\
\sum_{i=0}^{k-1} a_i x_k^i &\equiv t_k \mod p
\end{aligned}$$

The reason that this has a unique solution is because if we write

$$V = \begin{bmatrix} 1 & x_1 & \cdots & x_1^{k-1} \\ 1 & x_2 & \cdots & x_2^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_k & \cdots & x_k^{k-1} \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{bmatrix}$$

then because V is a Vandermonde matrix with $x_i \neq x_j$ for all $i \neq j$, V's determinant is nonzero and $VA = T$ has a unique solution for A. Thus, since the a_i's are chosen uniformly and independently, the probability that they satisfy this equation is $1/p^k$.

3 Time/Space Tradeoff

Note that if we consider the random seed as being a string of bits that we must query to hash our values, then to hash a family of n values using the above schemes, we must store $O(k \log n)$ bits of the random seed and query the whole seed, i.e. $O(k \log n)$ bits, to compute the hash function. It would be desirable to store and query a fewer number of bits in order to compute a k-wise independent hash function. The following negative result shows us that we cannot get something for nothing.

3.1 A Lower Bound

We give a simple combinatorial lower bound to show that n pairwise independent hash functions (to $\{0, 1\}$) which are each computed using only q queries must have a random seed of at least $m = n^{\Omega(1/q)}$ bits. Let us say that each hash function f_i takes as input the random seed $r = r_1 \cdots r_m$, but only accesses a subset S_i of q bits of r.

3.1.1 A (Super) Simple Argument

Note that if there are two functions f_i and f_j such that they are the same function that access the same subset of bits, then $f_1, \cdots, f_n$ are not pairwise independent since, for example,

$$\mathbb{P}(f_i(r) = 0, f_j(r) = 1) = 0.$$

Thus, at the very least, we need m and q to be large enough so that there are enough functions to avoid this problem.

How many functions are there with a random seed of m bits and q queries allowed? There are $\binom{m}{q}$ ways of choosing the subset that a function will depend on, and 2^{2^q} ways of choosing a function from $\{0,1\}^q \to \{0,1\}$. Thus, for constant q, we will need

$$\binom{m}{q} 2^{2^q} \geq n \longrightarrow m = n^{\Omega(1/q)}.$$

3.1.2 A (Less) Simple Argument

The above bound is non-trivial only for $q = o(\log \log n)$. Can we do better? Yes, we were too generous with the number of pairwise independent functions over $\{0,1\}^q$.

For now, let's consider the equivalent set of functions, $\{f : \{0,1\}^q \to \{-1,1\}\}$. We can associate with each such function f a vector $v_f \in \{-1,1\}^{2^q}$.

Let f and g be two functions. We claim that pairwise independence implies $< v_f, v_g >= 0$. To see this, note that if f and g are pairwise independen then they agree on exactly half of their inputs. But this is equivalent to having $< v_f, v_g >= 0$.

Thus in order to have N functions over $\{0,1\}^q$, we at least need their associated vectors in $\{-1,1\}^{2^q}$ to be othogonal. Since the dimension of this space is 2^q, we see that there are 2^q pairwise independent functions over a set of q bits. Thus our original bound improves to

$$\binom{m}{q} 2^q \geq n.$$

Thus, the bound becomes non-trivial for $q = o(\log n)$, but remains asymptotically the same for constant q.

Now that we have found a lower bound, is there any k-wise independent hashing scheme which only makes q queries and stores this many random bits? I.e. is there a matching upper bound? The answer turns out to be yes.

3.2 An Upper Bound

Consider a bipartite graph $G = (X, R, E)$ where $X = \{x_1, \cdots, x_n\}$, $R = \{r_1, \cdots, r_m\}$, and each x_i has $q \geq 1$ neighbors in R. We say that G is k-unique if for all subsets $T \subset X$ s.t. $|T| \leq k$, there is an $x^* \in T$ s.t. x^* has a neighbor in R that no other vertex in T has as a neighbor.

Denote by S_i the set of all of x_i's neighbors. And define

$$x_i = \bigoplus_{r \in S_i} r.$$

Proposition 1. *If G is k-unique, then $x_1, \cdots, x_n$ are k-wise independent.*

Proof. It is enough to show that for any subset of size $\leq k$ of X, the variables are linearly independent when written as linear functions of $r_1, \cdots, r_m$. We will prove this by induction of $s \leq k$, the size of the subset T.

Base Case: $s = 1$. If $T = \{x_i\} = \{\oplus_{r \in S_i} r\}$, then since $|S_i| = q \geq 1$, this is a non-constant subspace, and thus linearly independent.

Induction hypothesis: Assume for $s < k$.

Induction step: We will show the statement still holds when $|T| = s + 1 \leq k$. Since G is k-unique and $|T| \leq k$, there is an element of T, call it x', that has a neighbor r' that none of the other elements of T have as a neighbor. By the induction hypothesis, $T \setminus \{x'\}$ are linearly independent, and they do not depend on r', which x' does depend on. Thus, T is linearly independent. □

So we have shown that if we have a G that is k-unique, we have n k-wise independent variables. Which G's are k-unique? It turns out that a random G is unique with high probability, so long as m is large enough. Our random process goes as follows:

1. For each $x \in X$:
    2. Choose q elements from R uniformly, independently, and with replacement.
    3. Add the edges from x to the chosen elements

What is the probability that G is not k-unique? This is the probability that there exists a subset of size k of X s.t. all kq outgoing edges land in a subset of size $\frac{kq}{2}$. Algebraically (and through Sterling's approximation),

$$\begin{aligned}\sum_{\substack{T \subset X: \\ |T|=k}} \sum_{\substack{S \subset R: \\ |R|=\frac{kq}{2}}} \left(\frac{kq}{2m}\right)^{kq} &= \binom{n}{k}\binom{m}{\frac{kq}{2}}\left(\frac{kq}{2m}\right)^{kq} \\ &\approx n^k m^{kq/2}\left(\frac{kq}{2m}\right)^{kq} \\ &= n^k m^{-kq/2}\left(\frac{kq}{2}\right)^{kq}\end{aligned}$$

Thus, in order to make this probability less than $\epsilon > 0$, it is sufficient to have

$$m \geq \left(\frac{1}{\epsilon}\right)^{2/kq} n^{2/q}\left(\frac{kq}{2}\right)^2.$$