In this chapter we introduce and apply a limited notion of independence, known as k-wise independence, focusing in particular on the important case of pairwise independence. Applying limited dependence can allow us to reduce the amount of randomness used by a randomized algorithm, in some cases enabling us to convert a randomized algorithm into an efficient deterministic one. Limited dependence is also used in the design of universal and strongly universal families of hash functions, giving space- and time-efficient data structures. We consider why universal hash functions are effective in practice and show how they lead to simple perfect hash schemes. Finally, we apply these ideas to the design of effective and practical approximation algorithms for finding frequent objects in data streams, generalizing the Bloom filter data structure introduced in Chapter 5.

13.1. Pairwise Independence

Recall that in Chapter 2 we defined a set of events $E_1, E_2, \ldots, E_n$ to be mutually independent if, for any subset $I \subseteq [1, n]$,

$$\Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr(E_i).$$

Similarly, we defined a set of random variables $X_1, X_2, \ldots, X_n$ to be mutually independent if, for any subset $I \subseteq [1, n]$ and any values x_i, $i \in I$,

$$\Pr\left(\bigcap_{i \in I} X_i = x_i\right) = \prod_{i \in I} \Pr(X_i = x_i).$$

Mutual independence is often too much to ask for. Here, we examine a more limited notion of independence that proves useful in many contexts: k-wise independence.

Definition 13.1:

1. *A set of events $E_1, E_2, \ldots, E_n$ is k-*wise independent *if, for any subset $I \subseteq [1, n]$ with $|I| \le k$,*

$$\Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr(E_i).$$

2. *A set of random variables $X_1, X_2, \ldots, X_n$ is k-wise independent if, for any subset $I \subseteq [1, n]$ with $|I| \le k$ and for any values x_i, $i \in I$,*

$$\Pr\left(\bigcap_{i \in I} X_i = x_i\right) = \prod_{i \in I} \Pr(X_i = x_i).$$

3. *The random variables $X_1, X_2, \ldots, X_n$ are said to be* pairwise independent *if they are 2-wise independent. That is, for any pair i, j and any values a, b,*

$$\Pr((X_i = a) \cap (X_j = b)) = \Pr(X_i = a)\Pr(X_j = b).$$

13.1.1. *Example: A Construction of Pairwise Independent Bits*

A random bit is uniform if it assumes the values 0 and 1 with equal probability. Here we show how to derive $m = 2^b - 1$ uniform pairwise independent bits from b independent, uniform random bits $X_1, \ldots, X_b$.

Enumerate the $2^b - 1$ nonempty subsets of $\{1, 2, \ldots, b\}$ in some order, and let S_j be the jth subset in this ordering. Set

$$Y_j = \bigoplus_{i \in S_j} X_i,$$

where $\bigoplus$ is the exclusive-or operation. Equivalently, we could write this as

$$Y_j = \sum_{i \in S_j} X_i \bmod 2.$$

Lemma 13.1: *The Y_j are pairwise independent uniform bits.*

Proof: We first show that, for any nonempty set S_j, the random bit

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

is uniform. This follows easily using the principle of deferred decisions (see Section 1.3). Let z be the largest element of S. Then

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

Suppose we reveal the values for X_i for all $i \in S_j - \{z\}$. Then it is clear that the value of X_z determines the value of Y_j and that Y_j will take on the values 0 and 1 with equal probability.

Now consider any two variables Y_k and Y_ℓ with their corresponding sets S_k and S_ℓ. Let z be an element of S_ℓ that is not in S_k and consider, for any values $c, d \in \{0, 1\}$,

$$\Pr(Y_\ell = d \mid Y_k = c).$$

We claim, again by the principle of deferred decisions, that this probability is $1/2$. For suppose that we reveal the values for X_i for all i in $(S_k \cup S_\ell) - \{z\}$. Even though this determines the value of Y_k, the value of X_z will determine Y_ℓ. The conditioning on the value of Y_k therefore does not change that Y_ℓ is equally likely to be 0 or 1. Hence

$$\begin{aligned}\Pr((Y_k = c) \cap (Y_\ell = d)) &= \Pr(Y_\ell = d \mid Y_k = c) \Pr(Y_k = c) \\ &= 1/4.\end{aligned}$$

Since this holds for any values of $c, d \in \{0, 1\}$, we have proven pairwise independence.

■

Pairwise independence is much weaker than mutual independence. For example, we can use Chernoff bounds to evaluate the tail distribution of a sum of independent random variables, but we cannot directly apply a Chernoff bound if the X_i are only pairwise independent. However, pairwise independence is strong enough to allow for easy calculation of the variance of the sum, which allows for a useful application of Chebyshev's inequality.

Theorem 13.3: *Let $X = \sum_{i=1}^{n} X_i$, where the X_i are pairwise independent random variables. Then*

$$\mathbf{Var}[X] = \sum_{i=1}^{n} \mathbf{Var}[X_i].$$

Proof: We saw in Chapter 3 that

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

where

$$\mathbf{Cov}(X_i, X_j) = \mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])] = \mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\mathbf{E}[X_j].$$

Since $X_i, X_2, \dots, X_n$ are pairwise independent, it is clear (by the same argument as in Theorem 3.3) that for any $i \neq j$ we have

$$\mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\mathbf{E}[X_j] = 0.$$

Therefore,

$$\mathbf{Var}[X] = \sum_{i=1}^{n} \mathbf{Var}[X_i]. \qquad \blacksquare$$

Applying Chebyshev's inequality to the sum of pairwise independent variables yields the following.

Corollary 13.4: *Let $X = \sum_{i=1}^{n} X_i$, where the X_i are pairwise independent random variables. Then*

$$\Pr(|X - \mathbf{E}[X]| \geq a) \leq \frac{\mathbf{Var}[X]}{a^2} = \frac{\sum_{i=1}^{n} \mathbf{Var}[X_i]}{a^2}.$$