

5

Hashing

Hashing is a basic computer science technique used in many different contexts, from dictionary data structures to load balancing and symmetry breaking, to cryptography and complexity theory. In the next few lectures we will study the following:

- Desirable properties of hash families
- Constructions of hash families with these properties
- Applications of hash functions in various contexts

5.1 Maintaining a Dictionary

To understand the desired properties, let us keep one application in mind. We want to maintain a dictionary. We have a large universe of “keys” (say the set of all strings of length at most 80 using the Roman alphabet), denoted by U . The actual dictionary (say the set of all English words) is some subset S of this universe. S is typically much smaller than U . The operations we want to implement are:

- $\text{add}(x)$: add the key x to S .
- $\text{query}(q)$: is the key $q \in S$?
- $\text{delete}(x)$: remove the key x from S .

In some cases, we don’t care about adding and removing keys, we just care about fast query times—e.g., the actual English dictionary does not change (or changes so gradually that we can afford to rebuild the data structure when it changes). This is called the *static case*. Another special case is when we just add keys: the *insertion-only case*. The general case is called the *dynamic case*, or for emphasis, the *fully-dynamic case*.

5.1.1 Desired Properties

In this lecture, let $[N]$ denote the numbers $\{0, 1, 2, \dots, N - 1\}$. One natural approach is to choose a hash function $h : U \rightarrow [M]$, and store the key $x \in S$ at (or near) the location $h(x)$. What do we want from

hash functions:

- (i) *Small probability of distinct keys colliding*: if $x \neq y \in S$ then $\Pr[h(x) = h(y)]$ is “small”.
- (ii) *Small range*: we want the hash table size M to be small. At odds with first desired property.
- (iii) *Small number of bits to store the hash function h* .
- (iv) *h is easy to compute*.

Here it’s important to ask: *what is the probability in above expressions taken over?* There are two choices: (a) we could want that two *random keys* $x \neq y$ do not collide very often (i.e., $\Pr_{x,y \in U|x \neq y}[h(x) = h(y)] \leq \text{blah}$). But the keys in our dictionary are typically not random keys, so such a guarantee may be useless. Instead, (b) *we could choose the hash function randomly* (from some set H of hash functions) and want that for every pair of distinct elements x and y ,

$$\Pr_{h \leftarrow H} [h(x) = h(y)] \leq \text{blah}.$$

This latter approach is the one we take.

An Important Note: We will assume that the dictionary S is chosen “adversarially”, we have no control over it. We choose h randomly from the family H . This is the only randomness in the process. Of course the adversary does not see h . Then we look at the performance of our random h on this worst-case S . It’s like we’re playing a game, and both of us are choosing our actions simultaneously, and we want our minimax behavior to be as good as possible.

5.1.2 An Ideal: The Perfectly Random Hash Function

Consider the completely random hash function: for each $e \in S$, we choose a uniformly random location in $[M]$, and set $h(e)$ to be that location. Clearly, this has great properties:

- Low collision probability: $\Pr_h[h(a) = h(b)] = 1/M$ for any $a \neq b$, since having fixed where a maps to, there is a $1/M$ chance that b is mapped to the same location.
- In fact, even conditioned on knowing where any set of elements $A \subseteq S$ maps to, the position of any $e \in S \setminus A$ is still random:

$$\Pr_h[h(e) = \alpha \mid \wedge_{a \in A} h(a) = \alpha_a] = 1/M.$$

The problem? Storing this hash function requires storing $\lg_2 M$ bits for each $e \in S$ and hence $|S| \lg_2 M$ overall. Moreover, it is not clear how to compute $h(\cdot)$ fast, other than doing a table-lookup.

However, perfectly random hash functions are good to keep in mind: we often develop algorithms assuming we have a perfectly random hash function. Then we see what properties we needed for

Specifically, we are asking: what is the *random experiment*? What is the *sample space*?

In practice people use often hashing schemes based on cryptographic hash functions like MD5 and SHA (of which there are many variants), or Google’s CityHash/FarmHash. These hash functions are deterministic, and so can only give the former kind of guarantee (if at all). If you do use a fixed hash function, you may have to think through why there will be few collisions on the dataset used in your application (or else, you would cross your fingers and hope it will be so).

the analysis (e.g., low collision probability, or small sets of elements behave as though they are independent), and find good hash functions that have these properties.

5.2 Universal Hashing

The definition of *universal hashing* tries to capture the most basic desired property that distinct keys do not collide too often. It was proposed by Carter and Wegman.

Carter and Wegman (1979)

Definition 5.1. A family H of hash functions mapping U to $[M]$ is called *universal* if for any two keys $x \neq y \in U$, we have

$$\Pr_{h \leftarrow H} [h(x) = h(y)] \leq \frac{1}{M}.$$

Make sure you understand the definition. This condition must hold for *every pair* of distinct keys, and the randomness is over the choice of the actual hash function h from the set H .

5.2.1 A Single-Bit Construction

The first construction of a universal hash family is the following. Consider the (trivial?) case where $|U| = 2^u$ and $M = 2$. The hash functions are defined as follows.

Take a uniformly random u -bit vector a . For $x \in U$, view x as a u -bit vector in $\{0, 1\}^u$, and define

$$h(x) := a \cdot x$$

where the calculations are done modulo 2.

Since the hash function is completely defined by the vector a , there are 2^u hash functions in this family H_{basic} .

Theorem 5.2. *The family of hash functions H_{basic} defined above is universal for $M = 2$.*

Proof. Consider $x \neq y \in \{0, 1\}^u$. We claim that the probability that $h(x) = h(y)$ is at most $1/M = 1/2$. Since h is a linear map, $h(x) = h(y) \iff h(x - y) = \vec{0}$. Equivalently, we want to show that for any non-zero vector $z \in \{0, 1\}^u$,

$$\Pr_{h \leftarrow H} [h(z) = \vec{0}] = \Pr[a \cdot z = \vec{0}] \leq 1/2.$$

Note that $a \cdot z = \sum_{i \in [u]} a_i z_i$. Say that z_{i^*} equals 1 (since z is non-zero, there is at least one such coordinate). Now fix all the entries of a except entry i^* . For $a \cdot z$ to be zero, it must be the case that $a_{i^*} = \sum_{i \neq i^*} a_i z_i$. But a_{i^*} is a random bit, so it equals the value on the right with probability $1/2$. Hence the probability of $Az = \vec{0}$ is

$$1/2 = 1/M. \quad \square$$

5.2.2 A Construction with Larger Tables

Of course, a hash table of size $M = 2$ is not very useful. But we can immediately get larger hash tables by the simple idea of concatenating many independently drawn hash functions from the basic hash family. For $M = 2^m$, do the following:

Take an $u \times m$ matrix A and fill it with random bits. For $x \in U$, view x as a u -bit vector in $\{0, 1\}^u$, and define

$$h(x) := Ax$$

where the calculations are done modulo 2.

Since the hash function is completely defined by the matrix A , there are 2^{um} hash functions in this family H_{matrix} .

Theorem 5.3. *The family of hash functions H_{matrix} is universal for $M = 2^m$.*

Proof. We can view a hash function from this family as obtained by taking m independent hash functions from H_{basic} and concatenating their results. The probability of collision in each bit is $1/2$, so the probability of colliding in *all* m bits is $1/2^m$. \square

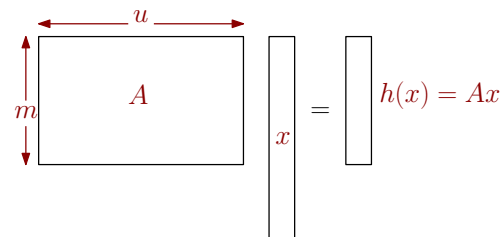
BTW, note that $h(\vec{0}) = \vec{0}$, so picking a random function from the hash family H_{matrix} does not map each key to a random place. (Indeed, the definition of universality does not require this.) It just ensures that the probability of pairwise collisions is small.

5.3 Application #1: Dictionaries

The condition of universality may not seem like much, but it gives a lot of power. As mentioned above, one of the main applications of universal hashing is to dictionary data structures. When many keys hash to the same location, the hash table can store only one of them. So we need some way of “resolving” these collisions, and storing these extra keys. There are many solutions, which you’ve probably seen before.

Hashing with separate chaining: An easy way to resolve collisions, also easy to analyze, but it may increase the space usage of the data structure. Here we maintain a linked list of all the “additional” keys. So the lookup time at location i becomes proportional to $|\{x \in S \mid h(x) = i\}|$, the number of keys in the dictionary S that map to i . Hence, when we perform a lookup on key q , we will spend expected time proportional to

$$E_{h \leftarrow H} [|\{x \in S \mid h(x) = h(q)\}|] = \sum_{x \in S} \Pr_{h \leftarrow H} [h(x) = h(q)] \leq \frac{|S|}{M}.$$



Hence, with a table of size $M = N = |S|$, lookups take expected constant time. (Also observe that item deletion is easy with separate chaining.)

Hashing with open addressing: What are other ways of resolving collisions? One way that requires no extra space is *open addressing*, where colliding keys are stored in the array itself. Where? That depends. The most basic idea is *linear probing*: When you are inserting x and $h(x)$ is occupied, you look for the smallest index i such that $(h(x) + 1) \bmod M$ is free, and store $h(x)$ there. When querying for q , you look at $h(q)$ and scan linearly until you find q or an empty space. Observe that deletions are not quite as simple any more. It is known that linear probing can also be done in expected constant time, but universal hashing does not suffice to prove this bound: 5-universal hashing is necessary and sufficient.

One can use other probe sequences: e.g., not probe each location but choose some step size s and look at

$$h(x), h(x) + s \pmod{M}, h(x) + 2s \pmod{M}, \dots$$

Or *quadratic probing*, where you look at

$$h(x), h(x) + 1 \bmod M, h(x) + 4 \bmod M, \dots, h(x) + i^2 \bmod M.$$

Or you can use a random pattern for each key, chosen according to its own hash function.

One can also try to store multiple (usually a constant number of) keys in the same table location. And there's a different approach called *cuckoo hashing*, which we will discuss later in this chapter.

5.4 Application #2: Perfect Hashing

The results for separate chaining mentioned above hold in expectation (over the choice of the hash function). Can we hope for worst-case bounds? For a static dictionary S with $|S| = N$, there is an elegant solution that gives worst-case constant lookup time, and uses only tables of total size $O(N)$. And it only uses universal hashing, combined with a two-level hashing idea. Here's how.

First, we claim that if we hash a set S of size N into a table of size $O(N)$ using a universal hash family, with probability at least $1/2$ no location will have more than $O(\sqrt{N})$ keys mapped to it. Why? For $x, y \in S$, let C_{xy} be the indicator random variable for whether $h(x) = h(y)$, i.e., they "collide". The total number of collisions is $C = \sum_{x \neq y \in S} C_{xy}$, and its expectation is

$$E[C] = E \left[\sum_{x \neq y \in S} C_{xy} \right] = E \sum_{x \neq y \in S} E[C_{xy}] \leq \binom{N}{2} \frac{1}{M}. \quad (5.1)$$

The necessity was shown by Mihai Patrascu and Mikkel Thorup, the sufficiency was shown by Anna Pagh, Rasmus Pagh and Milan Ruzic.

Python implements its built-in dictionary data structure using hashing with open addressing: the hash function merely uses $h(x) = x \bmod M$, and the considers the probe sequence $j_0 \leftarrow h(x)$ and $j_i \leftarrow 5j_{i-1} + 1 \bmod M$. There are other bells and whistles too such as adding on an additional perturbation, see this [stackexchange post](#) or the [source code](#) for details.

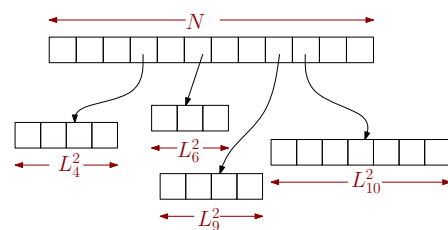
If we allow ourselves a table of size $M = \Omega(N^2)$, this is easy, because by the tightness of the birthday paradox—or the calculation in (5.1)—we could ensure that all keys map to distinct locations. But that is a lot of wasted space.

For $M = N$, say, this is at most $N/2$. So by Markov's inequality, we have $\Pr[C \geq N] \leq 1/2$. Moreover, if some location did have $\sqrt{2N}$ keys hashing to it, that location itself would result in $\binom{\sqrt{2N}}{2} \geq N$ collisions. Hence, with probability at least half, the maximum load at any location is at most $\sqrt{2N}$.

In fact, things are even better than that. If the load at location i is L_i , then the total number of collisions is $\sum_{i \in [M]} \binom{L_i}{2}$. And by the argument above, this is smaller than N (with probability $1/2$). Hence $\sum_i L_i^2 \leq 3N$. Fix this first-level hash function $h^* : U \rightarrow [N]$.

Now we can take all the L_i keys that map into location i of the main table, build a special second-level table for them of size $M_i = O(L_i^2)$, and use the calculation (5.1) with $M = O(L_i^2)$ to argue that using a universal hash family for this second-level hashing from these L_i keys to $[M_i]$ will map all of them into separate locations. So we can choose a good hash function h_i^* for the keys that h^* maps to location i .

To look up q , we look at location $i = h^*(q)$, and then check location $h_i^*(q)$ —this takes two hash function evaluations. Total space: N for the first table, then $\sum_i O(L_i^2)$ for the second level tables, which is again $O(N)$. (All space is measured in the number of keys.) We also need to store the hash functions, of course, which adds linear overhead.



5.5 Pairwise and k -wise Independent Hashing

A couple years after their original paper, Carter and Wegman proposed a stronger requirement which we will call *pairwise-independent*.¹ Let us define a general notion of being *k -wise-independent*.

Definition 5.4. A family H of hash functions mapping U to $[M]$ is called *k -wise-independent* if for any k distinct keys $x_1, x_2, \dots, x_k \in U$, and any k values $\alpha_1, \alpha_2, \dots, \alpha_k \in [M]$ (not necessarily distinct), we have

$$\Pr_{h \leftarrow H} [h(x_1) = \alpha_1 \wedge h(x_2) = \alpha_2 \wedge \dots \wedge h(x_k) = \alpha_k] \leq \frac{1}{M^k}.$$

Such a hash family is also called *k -wise independent*. The case $k = 2$ is called *pairwise independent*.

The following facts about k -wise independent hash families are simple to prove.

Fact 5.5. Suppose H is a k -wise independent family for $k \geq 2$. Then

- H is also $(k - 1)$ -wise independent.
- For any $x \in U$ and $\alpha \in [M]$, $\Pr[h(x) = \alpha] = 1/M$.
- H is universal.

¹ They called it *strongly universal* but this terminology does not naturally generalize to k -tuples. (They use *k -strongly universal* to mean something else, so that doesn't help.)

From part (c) above, we see that 2-wise independence is indeed at least as strong a condition as universality. And one can check that the construction in Section 5.2.2 is not 2-wise independent (since then it would also be 1-wise independent by Fact 5.5(a), but $\Pr[h(\vec{0}) = \vec{0}] = 1 \neq 1/M$). In the next section we give some constructions of 2-wise independent and k -wise independent hash families.

5.5.1 Construction #1: A Variant on a Familiar Theme

The first construction is a simple modification of the universal hash family we saw in Section 5.2.2 for the case where $|U| = 2^u$ and $M = 2^m$.

Take an $u \times m$ matrix A and fill it with random bits. Pick a random m -bit vector $b \in \{0,1\}^m$. For $x \in U = \{0,1\}^u$, define

$$h(x) := Ax + b$$

where the calculations are done modulo 2.

The hash function is defined by the matrix A containing um random bits, and vector b containing m random bits, there are $2^{(u+1)m}$ hash functions in this family H .

Claim 5.6. The family H is 2-wise independent.

Proof. Exercise. □

5.5.2 Construction #2: Using Fewer Bits

In the above construction, describing the hash function requires $O(um)$ bits. A natural question is whether we can do better. Indeed we can. Here is a related construction:

Take an $u \times m$ matrix A . Fill the first row $A_{1,*}$ and the first column $A_{*,1}$ with random bits. For any other entry i, j for $i > 1$ and $j > 1$, define $A_{i,j} = A_{i-1,j-1}$. So all entries in each "northwest-southeast" diagonal in A are the same.

Also pick a random m -bit vector $b \in \{0,1\}^m$. For $x \in U = \{0,1\}^u$, define

$$h(x) := Ax + b$$

where the calculations are done modulo 2.

Hence the hash family H consists of $2^{(u+m-1)+m}$ hash functions, one for each choice of A and b . You will prove that this family H is 2-wise independent as part of your homework. Here we need $O(u+m)$ random bits, and hence the space to store the hash function is comparable to the space to store a constant number of elements from U and $[M]$. Much better than $O(um)$!

5.5.3 Construction #3: Using Finite Fields

Suppose we want to map the universe $U = \{0, 1\}^u$ to $[M] = \{0, 1\}^m$. For this construction, we will work with the Galois field $GF(2^u)$ (and we associate strings in U with elements of the field in the natural way). First, we construct a 2-wise independent map from U to U as follows.

Pick two random numbers $a, b \in GF(2^u)$. For any $x \in U$, define

$$h(x) := ax + b$$

where the calculations are done over the field $GF(2^u)$.

To prove 2-wise independence, note that for $x_1 \neq x_2 \in U$,

$$\begin{pmatrix} h(x_1) \\ h(x_2) \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

To calculate $\Pr[h(x_1) = \alpha_1 \wedge h(x_2) = \alpha_2]$, we get

$$\Pr \left[\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \right] = \Pr \left[\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \end{pmatrix}^{-1} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \right]$$

where the matrix is invertible because $x_1 \neq x_2$, and we're working over a field. But since a, b are chosen randomly, the chance that each of them equals some specified values is at most $1/2^u \times 1/2^u = 1/2^{2u}$, which is $1/|U|^2$ as desired for 2-wise independence.

That's cute. On the other hand, we hashed $U \rightarrow U$, which does not seem useful. But now we could truncate the last $u - m$ bits of the hash value to get a hash family mapping $[2^u]$ to $[2^m]$ for $m \leq u$; you can check this is 2-wise independent too.

5.5.4 Construction #4: k -universal Hashing

The construction for k -universal is not very different; let's consider hashing $GF(2^u) \rightarrow GF(2^u)$ once again.

Pick k random numbers $a_0, a_1, \dots, a_{k-1} \in GF(2^u)$. For any $x \in U$, define

$$h(x) := a_0 + a_1x + a_2x^2 + \dots + a_{k-1}x^{k-1}$$

where the calculations are done over the field $GF(2^u)$.

The proof of k -universality is similar to that above; this is something you'll show in the homework. (In fact you could use any finite field $GF(p^s)$ you want.)

5.6 Optional: Other Hashing Schemes

While the above properties (universality and k -universality) are the most popular to prove that algorithms work well, here are some other hashing schemes which are commonly used, and have other good features.

5.6.1 Simple Tabulation Hashing

One proposal that has been around for some time (even considered by Carter and Wegman in their 1979 paper on universal hashing) is that of tabulation hashing. In this case, imagine $U = [k]^u$ and $M = 2^m$.

Tabulation Hashing. Initialize a 2-dimensional $u \times k$ array T with each of the uk entries having a random m -bit string. Then for the key $x = x_1x_2 \dots x_u$, define its hash as

$$h(x) := T[1, x_1] \oplus T[2, x_2] \oplus \dots \oplus T[u, x_u].$$

Note that the hash function is completely defined by the table, which contains $u \cdot k \cdot m$ random bits. Hence the size of this hash family is 2^{kmu} . Is this any good? We can look at the independence properties of this family, for one.

Theorem 5.7. *The hash family H for tabulation hashing is 3-wise independent but not 4-wise independent.*

However, this is one case where independence properties of the hash family do not capture how good it is. A recent paper of Patrascu and Thorup showed that the performance of many natural applications (linear probing, cuckoo hashing, balls-into-bins) using tabulation hashing almost matches the performance of these applications using truly random hash functions. An extension called *twisted tabulation* gives a better behavior for some applications.

Patrascu and Thorup (2010)

Patrascu and Thorup (2012)

5.6.2 A 5-universal Variant of Tabulation Hashing

Thorup and Zhang show that if we just write $x = x_1x_2$, and use the hash function

$$h(x) = T[1, x_1] \oplus T[2, x_2] \oplus T[3, x_1 + x_2]$$

which is slight variant on simple tabulation, then we get 5-universality. Recall that 5-universal is good for some applications, like for hashing with linear probing.

5.6.3 A Practical Almost-Universal Scheme

One hashing scheme that is not universal (but almost is), and is very easily implementable is as follows. As usual, we are hashing $U \rightarrow [M]$. Consider the common case where both $|U|$ and M are powers of 2; i.e., $|U| = 2^u$ and $M = 2^m$.

Pick a random **odd** number a in $[M]$. Define

$$h_a(x) := (ax \bmod U) \operatorname{div} (U/M)$$

Note that this construction clearly gives us an answer in $[M]$. It is also easy to implement: e.g., the div operation can be implemented by shifting to the right $u - m$ times. But is this any good? It turns out the collision probability is only twice as bad as ideal.

Theorem 5.8. *For the hash family H defined as above, for $x \neq y \in U$,*

$$\Pr_{h \leftarrow H} [h(x) = h(y)] \leq \frac{2}{M}.$$

(The proof is not very difficult, you should try it as a bonus problem.)

5.6.4 Further Reading on Fast/Practical Hashing

There has been a lot of work on making hashing fast and practical, while also maintaining good provable properties – and also to understand why certain hashing schemes work well in practice. Check out papers by, e.g., Martin Dietzfelbinger, Rasmus Pagh, Mikkel Thorup, and Mihai Patrascu, and the references therein.

5.7 Bloom Filters

A central application of hashing is for dictionary data structures, as we saw earlier. In some cases it is acceptable to have a data structure that occasionally has mistakes.

A *Bloom filter* is one such data structure. It has the feature that it only has false positives (it may report that a key is present when it is not, but never the other way). Compensating for this presence of errors is the fact that it is simple and fast. A common application of a Bloom filter is as a “filter” (hence the name): if you believe that most queries are not going to belong to the dictionary, then you could first use this data structure on the queries: if the answer is No you know for sure the query key is absent. And if the answer is Yes you can use a slower data structure to confirm.

Here’s the data structure.

Burton H. Bloom (1970). Space/time trade-offs in hash coding with allowable errors

For example, the Google Chrome browser uses a Bloom filter to maintain its list of potentially malicious websites: it’s OK to have a few false positives, but fast lookups are essential.

Keep an array T of M bits, initially all entries are zero. Moreover, you have k hash functions $h_1, h_2, \dots, h_k : U \rightarrow [M]$; for this analysis assume they are completely random hash functions.

To add a key $x \in S \subseteq U$ to the dictionary, set bits $T[h_1(x)], T[h_2(x)], \dots, T[h_k(x)]$ to 1.

Now, when a query comes for key $x \in U$, check if all the entries $T[h_i(x)]$ are set to 1; if so, answer Yes else answer No.

Just that simple. Note that if the key x was in the dictionary S , all those bits would be on, and hence we would always answer Yes. However, it could be that other keys have caused all the k bits in positions $h_1(x), h_2(x), \dots, h_k(x)$ to be set. What is the probability of that?

5.7.1 The Analysis

As usual, assume that $|S| = N$. For any key in S , h_1 does not hash this key to the location $\ell \in [M]$ with probability $(1 - 1/M)$. If the bit $T[\ell] = 0$, the same must be true for all N keys, and all k hash functions—this happens with probability

$$\left(1 - \frac{1}{M}\right)^{kN} \approx e^{-kN/M}.$$

Denote this probability by p . So each location is 0 with probability p , and hence the expected fraction of zeros in the table is p . One can show, by a concentration bound, that the fraction of zeros is close to p with very high probability; see the survey of Broder and Mitzenmacher.

Now for a false positive on some query x , the bits

$$T[h_1(x)], T[h_2(x)], \dots, T[h_k(x)]$$

in all the k random locations must be set. Since there are a p fraction of zeros in the table, this happens with probability

$$(1 - p)^k \approx (1 - e^{-kN/M})^k. \quad (5.2)$$

Just to get a sense of the numbers, suppose $M = 2N$. Then the false positive probability is about $(1 - e^{-k/2})^k$ —minimizing this as a function of k gives us a false positive rate of 38%; for $M = 8N$ this falls to 2%. In general, taking derivatives tells us that the optimal setting of k is $k = (\ln 2) \cdot (M/N)$, which gives false-positive probability of $(0.6185)^{M/N}$. In other words, if the false-positive probability is ϵ , then the number of bits we use is $M \approx 1.44N \log(1/\epsilon)$ —about $1.44 \log(1/\epsilon)$ bits per entry.

Bloom filters often arise in applications because of their simplicity and wide applicability; see the survey by Broder and Mitzenmacher on many applications in networking.

Broder and Mitzenmacher (2003)

The best possible space usage in this model is $\log(1/\epsilon)$ bit per key, so Bloom filters are off by 44%. See a paper by Anna Pagh, Rasmus Pagh and S. Srinivasa Rao for an optimal data structure.

5.8 Cuckoo Hashing

Cuckoo hashing is a form of hashing without any false positives/negatives.

It was invented by Rasmus Pagh and Flemming Rodler. Due to its simplicity, and its good performance in practice, it has become very popular algorithm. Again, we want to maintain a dictionary $S \subseteq U$ with N keys.

Pagh and Rodler (2004)

Take two tables T_1 and T_2 , both of size $M = O(N)$, and two hash functions $h_1, h_2 : U \rightarrow [M]$ from hash family H .²

When an element $x \in S$ is inserted, if either $T_1[h_1(x)]$ or $T_2[h_2(x)]$ is empty, put the element x in that location. If both locations are occupied, say y is the element in $T_1[h_1(x)]$, then place x in $T_1[h_1(x)]$, and “bump” y from it. Whenever an element z is bumped from one of its locations $T_i[h_i(z)]$ (for some $i \in \{1, 2\}$), place it in the other location $T_{3-i}[h_{3-i}(z)]$. If an insert causes more than $6 \log N$ bumps, we stop the process, pick a new pair of hash functions, and rehash all the elements in the table.

² We assume that H is fully-random, but you can check that choosing H to be $O(\log N)$ -universal suffices.

If x is queried, if either $T_1[h_1(x)]$ or $T_2[h_2(x)]$ contains x say Yes else say No.

If x is deleted, remove it from whichever of $T_1[h_1(x)]$ or $T_2[h_2(x)]$ contains it.

Note that deletes and lookups are both constant-time operations. It only remains to bound the time to perform inserts. It turns out that inserts are not very expensive, since we usually perform few bumps, and the complete rebuild of the table is extremely rare. This is formalized in the following theorem.

Theorem 5.9. *The expected time to perform an insert operation is $O(1)$ if $M \geq 4N$.*

See [these notes](#) from Erik Demaine’s class for a proof. [We’ll sketch it on the board.] You can also see these notes on [Cuckoo Hashing for Undergraduates](#) by Rasmus Pagh for a different proof.

5.8.1 Discussion of Cuckoo Hashing

In the above description, we used two tables to make the exposition clearer; we could just use a single table T of size $4M$ and two hash functions h_1, h_2 , and a result similar to Theorem 5.9 will still hold. Now this starts to look more like the two-choices paradigm from the previous section: the difference being that we are allowed to move balls around, and also have to move these balls on-the-fly.

One question that we care about is the occupancy rate: the theorem says we can store N objects in $4N$ locations and get constant insert time and expected constant insert time. That is only using 25% of the memory! Can we do better? How close to 100% can we get? It turns out that you can get close to 50% occupancy, but better than 50% causes the linear-time bounds to fail. What happens if we use d

hash functions instead of 2? With $d = 3$, experimentally it seems that one can get more than 90% occupancy and still linear-time bounds hold. And what happens when we are allowed to put, say, two or four items in each location? Again there are experimental conjectures, but the theory is not fleshed out yet. See [this survey](#) for more open questions and pointers.

Moreover, do we really need $O(\log N)$ -universal hash functions? Such functions are expensive to compute and store. Patrascu and Thorup showed that simple tabulation hashing (see §5.6.1) gives performance very similar to that of truly-random hash functions. [Cohen and Kane](#) show that we cannot get away with 6-universal hash functions.

[Patrascu and Thorup \(2010\)](#)