Recap of this week’s lectures:

- Hashing: Universal hashing, and constructions.
- Perfect hashing: dictionary lookup in constant worst-case time.
- The Data Streaming model
- Heavy hitters, both without and with deletions.

**Hashing:** A *universal* hash family $\mathcal{H}$ from $U$ to $[m] := \{0, 1, \ldots, m - 1\}$ is a set of hash functions $\mathcal{H} = \{h_1, h_2, \ldots, h_k\}$ each mapping $U$ to $[m]$, such that for any $a \neq b \in U$, when you pick a random function from $\mathcal{H}$,

$$\Pr[h(a) = h(b)] \leq \frac{1}{m}.$$  

Also, a $\ell$-*universal* hash family $\mathcal{H}$ from $U$ to $[m] := \{0, 1, \ldots, m - 1\}$ is a set of hash functions $\mathcal{H} = \{h_1, h_2, \ldots, h_k\}$ each mapping $U$ to $[m]$, such that for any distinct $a_1, \ldots, a_\ell \in U$, and for any $\alpha_1, \ldots, \alpha_\ell \in [m]$, when you pick a random function from $\mathcal{H}$,

$$\Pr[h(a_1) = \alpha_1 \text{ and } \ldots \text{ and } h(a_\ell) = \alpha_\ell] = \frac{1}{m^\ell}.$$  

1. Show that a 2-universal hash family is a universal hash family.

**Solution:** We know that for any $a \neq b$ and any $\alpha \in [m]$, $\Pr[h(a) = h(b) = \alpha] = \frac{1}{m^2}$ by definition of 2-universal. So

$$\Pr[h(a) = h(b)] = \sum_{\alpha \in [m]} \Pr[h(a) = h(b) = \alpha] = \sum_{\alpha \in [m]} \frac{1}{m^2} = \frac{1}{m}.$$
2. Show that a $k$-universal hash family is a $\ell$-universal hash family for any $\ell \leq k$.

**Solution:** We show this for $\ell = k - 1$ and then can use induction. Indeed,

\[
\Pr[h(a_1) = \alpha_1 \text{ and } \ldots \text{ and } h(a_\ell) = \alpha_\ell] = \sum_{\alpha_{\ell+1} \in [m]} \Pr[h(a_1) = \alpha_1 \text{ and } \ldots \text{ and } h(a_\ell) = \alpha_\ell \text{ and } h(a_{\ell+1}) = \alpha_{\ell+1}] = \sum_{\alpha_{\ell+1} \in [m]} \frac{1}{m^{\ell+1}} = \frac{1}{m^\ell}.
\]

3. Is this hash family from $U = \{a, b\}$ to $\{0, 1\}$ (i.e., $m = 2$) universal? 1-universal? 2-universal?

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$h_2$</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Solution:** It’s universal because $a$ and $b$ collide under only one of the two functions, so they collide with probability $1/2$. It’s not 1-universal because $b$ hashes to value 0 with probability 1, whereas it should hash with probability half. And hence it is not $\ell$-universal for any $\ell \geq 1$.


<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$h_2$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$h_3$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$h_4$</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Solution:** This is 2-universal (can verify, for each pair of elements, that all 4 possibilities for $\alpha, \beta$ each occur once) and therefore it is universal. Also, 2-universal implies 1-universal. But it’s not 3-universal (e.g., can’t get all three elements to simultaneously hash to 1).
5. Suppose that an adversary knows the hash family $H$ and controls the keys we hash, and the adversary wants to force a collision. In this problem part, suppose that $H$ is universal. The following scenario takes place: we choose a hash function $h$ randomly from $H$, keeping it secret from the adversary, and then the adversary chooses a key $x$ and learns the value $h(x)$. Can the adversary now force a collision? In other words, can it find a $y \neq x$ such that $h(x) = h(y)$ with probability greater than $\frac{1}{m}$?

If so, write down a particular universal hash family in the same format as in part (b), and describe how an adversary can force a collision in this scenario. If not, prove that the adversary cannot force a collision with probability greater than $\frac{1}{m}$.

**Solution:** By adding one extra key to our previous example, we can construct a scenario where the adversary can force a collision. On a universe $U = x, y, z$, consider the following family $H$:

<table>
<thead>
<tr>
<th></th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$h_2$</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$H$ is still a universal hash family: $x$ and $y$ collide with probability $\frac{1}{2}$, $x$ and $z$ collide with probability $\frac{1}{2}$, and $y$ and $z$ collide with probability $0 < \frac{1}{2}$.

The adversary can determine whether we have selected $h_1$ or $h_2$ by giving us $x$ to hash. If $h(x) = 0$, then we have chosen $h_1$, and the adversary then gives us $y$. Otherwise, if $h(x) = 1$, we have chosen $h_2$ and the adversary gives us $z$.

6. Answer the question from #5, but supposing that $H$ is 2-universal, not just universal.

**Solution:** With a 2-universal hash family, the adversary cannot force a collision with probability better than $\frac{1}{m}$. Essentially, knowing $h(x)$ gives the adversary no information about $h(y)$ for any other key $y$.

We can prove this formally using conditional probabilities. Suppose we choose a random hash function $h \in H$, and then the adversary forces us to hash some key $x$ and learns the value $h(x) = X$. Then the adversary gives us any key $y \neq x$, hoping to cause a collision. By definition of 2-universality, we have that for any $x$ and $y$ with $x \neq y$:

$$\Pr_{h \in H}[h(y) = h(x)|h(x) = X] = \frac{\Pr_{h \in H}[h(y) = h(x) \wedge h(x) = X]}{\Pr_{h \in H}[h(x) = X]} = \frac{1/m^2}{1/m} = \frac{1}{m}.$$  

Therefore, no matter which $x$ the adversary chooses first, and which $h(x) = X$ value it learns, the probability of any particular $y$ colliding with $x$ is only $\frac{1}{m}$.
Streaming.

**Sampling:** Given a number \(k\), you want to maintain a random sample of size \(k\) from the stream. I.e., for each \(n \geq k\), the set you have at time \(n\) should be a random subset of the prefix \(a_{[1:n]}\), each of the \(\binom{n}{k}\) subsets of size \(k\) from this prefix should be equally likely.

1. For \(k = 1\), show that the algorithm: pick the first element. When faced with the \(n\)th element, with prob. \(1/n\) discard the element in your hand and pick the new element, and with prob. \(1 - 1/n\) keep the element in hand.

**Solution:** We claim that for any element from \(e \in a_{[1..n]}\), we have \(e\) in our hand after step \(n\) w.p. \(1/n\). The proof is inductive. The base case is easy. Consider the case for element \(e\) at some time \(n\).

- If \(e \in a_{[1..n-1]}\), then we have it at time \(n-1\) w.p. \(1/(n-1)\). Now at time \(n\), the chance we don’t discard it is \((n-1)/n\), so multiplying we get \(1/n\).
- If \(e = a_n\), then the chance we picked it at time \(n\) is \(1/n\).

2. Give an algorithm for general \(k\). (What would you do when faced with the \(n\)th element? With what probability should you pick this element? Which element should you drop?)

**Solution:** The algorithm: pick the \(n\)th item with probability \(k/n\), and drop a uniformly random item from the current set.

Claim: for any \(k\)-sized set \(S\) from \(e \in a_{[1..n]}\), we have \(S\) after step \(n\) w.p. \(1/(\binom{n}{k})\). The proof is again inductive. The base case for \(n = k\) is easy. Consider the case for element \(e\) at some time \(n > k\).

- If \(S \subseteq a_{[1..n-1]}\), then we have it at time \(n-1\) w.p. \(1/(\binom{n-1}{k})\). Now at time \(n\), the chance we don’t destroy it is \((n-1-k)/n\), so multiplying we get \(1\).
- If \(a_n \in S\), then look at \(S_i = (S \setminus \{a_n\}) \cup \{a_i\}\) where \(a_i \notin S\). There are \(n-k\) such sets \(S_i\). For each of the \(n-k\) sets, inductively the chance we have it at time \(n-1\) is \((n-1-k)/n\), then we have to drop \(a_i\) (w.p. \(1/k\)), and add \(a_n\) (w.p. \(1/n\)). Overall we get

\[
(n-k) \times \frac{k!(n-1-k)!}{(n-1)!} \times \frac{1}{k} \times \frac{1}{n} = \frac{1}{\binom{n}{k}}.
\]

This idea is called Reservoir Sampling.
Missing Numbers: Suppose I give you a stream of \( n - 1 \) elements, which contains all the numbers from 1 thru \( n \) except one of them. (The numbers do not appear in sorted order.) Clearly you can figure out the missing number by storing all \( n - 1 \) numbers and looking for the missing number. How can you output the missing number with only \( O(\log n) \) space? What if there are two missing numbers: can you again use only \( O(\log n) \) space?

**Solution:** For one missing number, you can store the sum of all numbers seen so far. Then finally subtract that from \( \frac{n(n+1)}{2} \) to get the missing number. For two, you can store, e.g., the sum, and the sum of their squares. Then you'll know \( a + b \) and \( a^2 + b^2 \), and can solve for the answer.

You could also have stored the sum and product of the numbers seen, but that requires more space. The product of the numbers could be as large as \( \Omega(n!) \) which requires \( \Omega(n \log n) \) bits to store.
Min-Hashing.
In min-hashing we created an estimator with “one-sided error”: our estimate was always an overestimate. I.e., for the target value \( v \) we created a random variable \( X \) such that \( \Pr[X \geq v] = 1 \). Suppose \( \mathbb{E}[X] = \mu \).

1. Show that \( \Pr[X \geq 2\mu] \leq \frac{1}{2} \). (“The probability of one estimate being too large is at most 50%.”)

**Solution:** Markov’s inequality.

2. Use this to show that if we take \( k \) independent copies \( X_1, X_2, \ldots, X_k \) of the r.v. \( X \), then \( \Pr[\min_{i=1}^{k} X_i \geq 2\mu] \leq 2^{-k} \).

**Solution:** In order to get such a high estimate, we must get unlucky all \( k \) times. The prob of that is \( 2^{-k} \).

3. Show that \( k = \log(1/\delta) \) gives \( 2^{-k} = \delta \). (If we want error probability \( 2^{-100} \), take the minimum of 100 independent estimates.)

**Solution:** \( 2^{-k} = 2^{-\log 1/\delta} = \delta \).