Lecture 15: Universal Hashing: Minimizing Collisions


- **$k$-wise Independence**
  - Intuition: First $k$ inputs are answered uniformly at random
  - Formally: For all distinct $x_1, \ldots, x_k \in \mathcal{D}$ and $y_1, \ldots, y_k \in \mathcal{R}$ we have
    \[
    \Pr \left[ h(x_1) = y_1, h(x_2) = y_2, \ldots, h(x_k) = y_k : h \leftarrow \mathcal{H} \right] = \frac{1}{|\mathcal{R}|^k}
    \]
  - One Construction: The set of all degree $< k$ polynomials.
2-wise Independence/Pairwise Independence

- Special case of $k = 2$ mentioned above
- Formally: For all distinct $x_1, x_2 \in D$ and $y_1, y_2 \in R$ we have

$$\Pr \left[ h(x_1) = y_1, h(x_2) = y_2 : h \leftarrow \mathcal{H} \right] = \frac{1}{|R|^2}$$

- One Construction: Linear functions
Universal Hash Function Family

- Intuition: Probability of Collision is low
- Formally: For all distinct \( x_1, x_2 \in D \) we have

\[
\mathbb{P} \left[ h(x_1) = h(x_2) : h \xleftarrow{\$} \mathcal{H} \right] \leq \frac{1}{|\mathcal{R}|}
\]

- Construction: Any 2-wise independent hash function family is also universal (we proved this result). The collision probability

\[
\mathbb{P} \left[ h(x_1) = h(x_2) : h \xleftarrow{\$} \mathcal{H} \right] = \frac{1}{|\mathcal{R}|}
\]

in this case.
Recall IV

- **Constructing Better Universal Hash Function Families**
  - We know that if the range is larger (or same size) than the domain, then we can achieve collision probability $P[h(x_1) = h(x_2) : h \leftarrow \mathcal{H}] = 0$ for every distinct $x_1, x_2 \in D$ (we saw that any one-one function achieves this)
  - When the range is smaller than the domain, we saw that any 2-wise independent hash function family achieves collision probability $P[h(x_1) = h(x_2) : h \leftarrow \mathcal{H}] = \frac{1}{|R|}$
  - When the range is smaller than the domain, can we have collision probability $P[h(x_1) = h(x_2) : h \leftarrow \mathcal{H}] < \frac{1}{|R|}$ for all distinct $x_1, x_2 \in D$?
    - In the previous lecture we saw that we can construct one hash function family $\mathcal{H}$, for $|D| = 4, |R| = 2$ such that the collision probability is $= \frac{1}{3} < \frac{1}{2}!$
    - Can we have even lower collision probabilities? In this lecture we shall prove that a lower collision probability is impossible!
Let the size of the domain $\mathcal{D}$ be $N$
Let the size of the range $\mathcal{R}$ be $M$
Suppose we have $M < N$

We shall prove the following theorem

**Theorem (Collision Lower Bound)**

Let $\mathcal{H}$ be a hash function family such that the domain of the function is $\mathcal{D}$ and the range of the functions is $\mathcal{R}$. There exists distinct $x_1^*, x_2^* \in \mathcal{D}$ such that

$$\Pr[h(x_1^*) = h(x_2^*): h \leftarrow \mathcal{H}] \geq \frac{N}{M} - \frac{1}{N - 1}$$

Note that for $M = 2$ and $N = 4$, the bound is $1/3$. The hash function family from the previous lecture achieves this bound.
Let us fix a hash function \( h \in \mathcal{H} \)

Suppose the range is the set \( \{y_1, y_2, \ldots, y_M\} \)

Let \( n_i \) be the size of the set \( \{x: x \in \mathcal{D}, h(x) = y_i\} \), for \( i \in \{1, 2, \ldots, M\} \). That is, \( n_1 \) inputs maps to \( y_1 \), \( n_2 \) inputs maps to \( y_2 \), and so on ...

The intuition of this is pictorially represented below
Proof of the Lower-bound II

- Let us count the number (represented by $\#\text{col}_h$) of entries \(\{x_1, x_2\}\), where \(x_1, x_2\) are distinct elements from the domain \(\mathcal{D}\), such that \(h(x_1) = h(x_2)\)

**Claim**

\[
\#\text{col}_h = \sum_{i=1}^{M} \binom{n_i}{2}
\]

**Proof.**

- Note that the number of distinct \(\{x_1, x_2\}\) that collide at \(y_1\) is \(\binom{n_1}{2}\)
- Note that the number of distinct \(\{x_1, x_2\}\) that collide at \(y_2\) is \(\binom{n_2}{2}\)
- And, so on ...
- Adding these entries, we get the total number of distinct \(\{x_1, x_2\}\) that collide
\begin{itemize}
  \item Note that \( n_i \geq 0 \) and \( \sum_{i=1}^{M} n_i = N \)
  \item We are interested in lower-bounding the expression \( \sum_{i=1}^{M} \binom{n_i}{2} \)
  \item Consider the following manipulation

\[
\sum_{i=1}^{M} \binom{n_i}{2} = \sum_{i=1}^{M} \frac{n_i(n_i - 1)}{2} = \sum_{i=1}^{M} \frac{n_i^2 - n_i}{2} = \sum_{i=1}^{M} \frac{n_i^2}{2} - \frac{N}{2}
\]
\end{itemize}
We are interested in lower-bounding $\sum_{i=1}^{M} n_i^2$ under the constraint $n_i \geq 0$ and $\sum_{i=1}^{M} n_i = N$.

So our task is to look at all the solutions to the equations: $n_i \geq 0$ (for all $i \in \{1, \ldots, M\}$) and $\sum_{i=1}^{M} n_i = N$. And minimize $\sum_{i=1}^{M} n_i^2$.

For $M = 2$, we have the following picture for intuition. The THICK RED line is the set of all feasible solutions. The quantity $n_1^2 + n_2^2$ measures the distance of the solution from the origin. The dotted lines represent this distance for various solutions.

Using the AM-GM inequality, one can show that the minimum is achieved when all the coordinates of the solution are equal.
Proof of the Lower-bound V

\[ n_1 + n_2 = N \]

Universal Hashing
Proof of the Lower-bound VI

- So, the solution where \( n_1 = n_2 = \cdots = n_M \) and \( \sum_{i=1}^{M} n_i = N \) is
  \[
  \left( \frac{N}{M}, \frac{N}{M}, \cdots, \frac{N}{M} \right)
  \]

- For this feasible solution, we have:
  \[
  \sum_{i=1}^{M} n_i^2 = \sum_{i=1}^{M} \left( \frac{N}{M} \right)^2 = \frac{N^2}{M}
  \]

- Therefore, we get

Claim

\[
\#\text{col}_h \geq \frac{\frac{N^2}{M} - N}{2}
\]

Universal Hashing
Suppose \( \mathcal{H} = \{h_1, \ldots, h_K\} \). Then, the total number (represented by \( \#\text{col}_\mathcal{H} \)) of entries \( \{h, x_1, x_2\} \), where \( x_1, x_2 \) are distinct elements from the domain \( \mathcal{D} \), \( h \in \mathcal{H} \), and \( h(x_1) = h(x_2) \) is

\[
\#\text{col}_\mathcal{H} \geq K \left( \frac{N^2}{M} - N \right)
\]

**Proof.**

- For each \( h \), we have shown earlier that \( \#\text{col}_h \geq \left( \frac{N^2}{M} - N \right) \).

- Summing over all \( h \in \mathcal{H} \), we get this result.
Let us define $P$ be the set of all distinct $\{x_1, x_2\}$ such that $x_1, x_2 \in D$. Note that $|P| = \binom{N}{2} = N(N - 1)/2$.

Suppose we perform the following experiment:

1. Sample $(x_1, x_2) \leftarrow P$
2. Sample $h \leftarrow H$
3. Output 1 if $h(x_1) = h(x_2)$; otherwise output 0

Let us denote the output of this experiment by $Z$.

Let us calculate expected outcome of $Z$. 

Universal Hashing
Proof of the Lower-bound IX

Consider the following manipulation

\[
\mathbb{E} \left[ Z : (x_1, x_2) \leftarrow \mathcal{P}, h \leftarrow \mathcal{H} \right] = \mathbb{P} \left[ Z = 1 : (x_1, x_2) \leftarrow \mathcal{P}, h \leftarrow \mathcal{H} \right]
\]

\[
= \frac{\#\text{col}_\mathcal{H}}{|\mathcal{P}| \cdot |\mathcal{H}|}
\]

\[
\geq K \left( \frac{\frac{N^2}{M} - N}{2} \right)
\]

\[
\geq \frac{N(N-1)}{2} \cdot K
\]

\[
= \frac{\frac{N}{M} - 1}{N - 1}
\]
Proof of the Lower-bound X

- So, we get the following result

**Claim**

\[ \mathbb{E} \left[ Z : (x_1, x_2) \leftarrow \mathcal{P}, h \leftarrow \mathcal{H} \right] \geq \frac{N}{M} - 1 \]

- Note that the above expression is identical to the following statement:
  For \((x_1, x_2) \leftarrow \mathcal{P},\) we have \(\mathbb{E} \left[ Z : h \leftarrow \mathcal{H} \right] \geq \frac{N}{M} - 1 \)

- By Pigeon-hole Principle, we get: There exists \((x_1^*, x_2^*) \in \mathcal{P}\) such that
  \[ \mathbb{E} \left[ Z : h \leftarrow \mathcal{H} \right] \geq \frac{N}{M} - 1 \]

Universal Hashing
So, for this choice of $x_1^*$ and $x_2^*$ the collision probability is

$$\mathbb{P} \left[ h(x_1^*) = h(x_2^*) : h \leftarrow \mathcal{H} \right] \geq \frac{N}{M} - 1$$

This completes the proof of the theorem.
Given domain of size $N$ and range of size $M$, where $M < N$ and $M$ divides $N$

Can we design universal hash functions such that for all distinct $x_1, x_2 \in \mathcal{D}$ we have

$$\Pr[h(x_1) = h(x_2) : h \leftarrow \mathcal{H}] = \frac{\frac{N}{M} - 1}{N - 1} = \frac{1}{M} \cdot \frac{N - M}{N - 1}$$

This implies that we have to achieve equality at every step of the proof of the collision lower-bound theorem

- We have to ensure $n_1 = n_2 = \cdots = n_M$
- We have to ensure that the “average” collision probability for every $(x_1, x_2)$ is identical

This problem will be posed in the homework
“Better(?) than $k$-wise Independence”

- Note that when defining $k$-wise Independence we stated that the probability of a hash function mapping $x_1 \mapsto y_1$, $x_2 \mapsto y_2$, \ldots, and $x_k \mapsto y_k$ is

$$\frac{1}{|\mathcal{R}|^k}$$

- Why did we not write $\leq \frac{1}{|\mathcal{R}|^k}$?

- Is it even possible to get $< \frac{1}{|\mathcal{R}|^k}$?

- In the homework you will prove that for any hash function family, there exists distinct $x_1, \ldots, x_k$ and $y_1, \ldots, y_k$ such that

$$\mathbb{P}\left[h(x_1) = y_1, \ldots, h(x_k) = y_k : h \leftarrow \mathcal{H}\right] \geq \frac{1}{|\mathcal{R}|^k}$$

- So, there is no way to get $< \frac{1}{|\mathcal{R}|^k}$. The bound $\leq \frac{1}{|\mathcal{R}|^k}$ would be equivalent to the bound $= \frac{1}{|\mathcal{R}|^k}$. 

Universal Hashing
Appendix: Inequality Proof I

Suppose $n_1, \ldots, n_M$ are positive numbers such that $n_1 + \cdots + n_M = N$. Then the following claim holds.

**Claim**

$$n_1^2 + \cdots + n_M^2 \geq N^2 / M$$

**Proof.**

- We shall use AM-GM inequality to prove this result.
- AM-GM inequality states that, for non-negative $a$ and $b$, the following holds.

$$\frac{a + b}{2} \geq \sqrt{ab}$$

Moreover, the equality holds if and only if $a = b$. 

Universal Hashing
Consider the following manipulation of the original expression

\[ \sum_{i=1}^{M} n_i^2 = (n_1 + \cdots + n_M)^2 - \sum_{1 \leq i < j \leq M} 2n_in_j \]

\[ = N^2 - \sum_{1 \leq i < j \leq M} 2n_in_j, \]

\[ \geq N^2 - \sum_{1 \leq i < j \leq M} (n_i^2 + n_j^2), \]

\[ = N^2 - (M - 1) \sum_{1 \leq i \leq M} n_i^2 \]

Rearranging, we get

\[ M \sum_{i=1}^{M} n_i^2 \geq N^2 \]
This gives the inequality of the claim. Equality holds if and only if $n_i = n_j$, for all $1 \leq i < j \leq M$. This holds if and only if $n_1 = n_2 = \cdots = n_M$.