CSE5401 Lecture 4

Hashing

1. Pairwise independence + $O(1)$-time dictionaries
2. Power of two choices

Recall the dictionary data structure that maintains a set S
Two operations: insert(x): adds x to S
lookup(x): returns whether $x \in S$ (or not)
Standard solution is to use a hash table —
hash function h maps x to integer in $\{0, 1, \ldots, N-1\}$.
"Ideal" h maps x to a uniform random index indep,
though we can't afford to actually choose h this way.
Still useful to consider: how many elements can we insert with no collisions w.p. $1-\delta \in (0,1)$? $n \approx \sqrt{2N \ln \frac{1}{1-\delta}}$

No collisions w.p. $1 \cdot \left(1-\frac{1}{N}\right) \cdot \left(1-\frac{2}{N}\right) \cdots \left(1-\frac{n-1}{N}\right) = 1-\delta$

$\hookrightarrow \ln(1-\delta) = \sum_{i=1}^{n-1} \ln\left(1-\frac{i}{N}\right) = \sum_{i=1}^{n-1} \left(-\frac{i}{N} + O\left(\left(\frac{i}{N}\right)^2\right)\right) = -\frac{n(n-1)}{2N} + o\left(\frac{n^3}{N^2}\right)$

$o(1)$ for $n \in O(\sqrt{N})$

$\hookrightarrow n \approx \sqrt{2N \ln \frac{1}{1-\delta}}$ ("Birthday paradox")

We can match this using pairwise indep. hash fns.

Defn: The family of RVs $\{X_1, \ldots, X_n\}$ are pairwise indep. if for all $i \neq j$ and x_i, x_j $\Pr[X_i = x_i \text{ and } X_j = x_j] = \Pr[X_i = x_i] \cdot \Pr[X_j = x_j]$.

Here, H will be chosen uniformly at random from a family $\mathcal{H}$, and $\{H(x_1), \ldots, H(x_n)\}$ will be pairwise independent and uniform.

Theorem A pairwise independent collection of $\leq \sqrt{2N\delta}$ RVs is mutually distinct w.p. $1-\delta$.

Proof: For each pair (i,j) define the collision indicator $I_{ij} = \begin{cases} 1 \text{ if } X_i = X_j \\ 0 \text{ o.w.} \end{cases}$

The total # collisions $\sum_{i \neq j} I_{ij}$ is $0 \Leftrightarrow X_i \neq X_j$ for all $i \neq j$.

So, it suffices to show this is < 1 w.p. $1-\delta$.

The expected # is $\mathbb{E}\left[\sum_{i\neq j} I_{ij}\right] = \sum_{i \neq j} \mathbb{E}[I_{ij}] = \sum_{i \neq j} \sum_{x} \Pr[X_i = x, X_j = x]$

$= \binom{n}{2} \cdot \frac{N}{N^2} < \frac{n^2}{2N}$

Observe: if $n \leq \sqrt{2N\delta}$, $\leq \delta$

Thus, when there is ≥ 1 collision, $\sum_{i \neq j} I_{ij} > \frac{1}{\delta} \mathbb{E}\left[\sum_{i \neq j} I_{ij}\right]$

Markov's Inequality gives $\Pr\left[\sum_{i \neq j} I_{ij} > \frac{1}{\delta} \mathbb{E}\left[\sum_{i \neq j} I_{ij}\right]\right] < \delta$

$\Vert\Vert$ some $X_i = X_j$ ▨

There are many possible constructions of pairwise indep hash function families. A simple and efficient one breaks the input x into blocks of $\log_2 N$ bits each — suppose every x has length $\ell \log_2 N$ (so: ℓ blocks). Then $\mathcal{H} = \{h_{\vec{a},b} : a_1, \ldots, a_\ell, b \in \{0, \ldots, N-1\}\}$ given by $h_{\vec{a},b}(x_1, x_2, \ldots, x_\ell) = \sum_{i=1}^{\ell} a_i x_i + b \bmod N$ is pairwise independent.

Proof: Suppose $h_{\vec{a},b}(x_1 \ldots x_\ell) = y$ and $h_{\vec{a},b}(x'_1, \ldots, x'_\ell) = y'$, where some $x_i \neq x'_i$. How many choices for a_i & b satisfy these eqns given $\vec{a}_{-i}$? We have the system $\begin{bmatrix} x_i & 1 \\ x'_i & 1 \end{bmatrix} \begin{bmatrix} a_i \\ b \end{bmatrix} = \begin{bmatrix} y - \sum_{j \neq i} a_j x_j \\ y' - \sum_{j \neq i} a_j x'_j \end{bmatrix} \bmod N$

It has a unique solution: $a_i = (x_i - x'_i)^{-1}(y - y' - \sum_{j \neq i} a_j (x_j - x'_j)) \bmod N$, $b = y - \sum_{i=1}^{\ell} a_i x_i \bmod N$. (Diagonalizable mod N) Therefore,

$$\Pr_{\vec{a},b}\left[h_{\vec{a},b}(x_1 \ldots x_\ell) = y \text{ and } h_{\vec{a},b}(x'_1, \ldots, x'_\ell) = y'\right] = \frac{1}{N^2}$$ ▨

Power of Two Choices

Obtaining $O(1)$ access time with a set of size n

requires a table of size $\Omega(n^2)$, which we might not be happy with. It turns out that if the table has size $O(n)$, then even an ideal hash fn has some bin with $\Theta\left(\frac{\log n}{\log\log n}\right)$ items whp, and we might as well be using a balanced tree. We can do <u>much</u> better by using <u>two</u> hash functions and placing the item in the bin with fewer items on insert.

<u>Theorem</u> Suppose for $X_1, \ldots, X_n$, $H_1(X_i)$ and $H_2(X_i)$ are indep. and uniform on $\{0, \ldots, n-1\}$ and X_i is placed in whichever of the two contains fewer of $X_1, \ldots, X_{i-1}$. (Ties are broken at random.) Then with probability $1 - o(\frac{1}{n})$, every index has $O(\log\log n)$ elements.

<u>(leading constant is $\frac{1}{\ln 2}$ ...)</u>

<u>Proof:</u> Let B_i be the RV that counts the # of indices that contain $\geq i$ elements at the conclusion. Observe that before an element is placed in a index with $i+1$ (or more) elements, it must be that both H_1 & H_2 had at least i elements. There were at most B_i such indices, so the probability that an element is placed in such a index is $\leq \left(\frac{B_i}{n}\right)^2$, and hence $\mathbb{E}[B_{i+1} \mid B_i \leq \beta_i] \leq n \cdot \left(\frac{\beta_i}{n}\right)^2$: this is because $B_{i+1} \leq \sum_{\text{element } t} I[H_1(X_t) \text{ \& } H_2(X_t) \text{ contain} \geq i \text{ elts.}]$, where distinct t are independent. We can therefore use the Chernoff bound:

<u>Chernoff Bound</u> Let $X_1, \ldots, X_n$ be independent RVs taking values in the range $[0,1]$. Then $\Pr[\sum_i X_i > (1+\varepsilon)\mathbb{E}[\sum_i X_i]] \leq e^{-\frac{\mathbb{E}[\sum X_i]}{3\varepsilon^2}}$ <u>and</u> $\Pr[\sum_i X_i < (1-\varepsilon)\mathbb{E}[\sum_i X_i]] \leq e^{-\frac{\mathbb{E}[\sum X_i]}{2\varepsilon^2}}$

Conditioning on "$B_i \le \beta_i$" breaks the independence but, supposing this occurs w.p. $\ge 1 - \frac{1}{n^2}$,

$$\mathbb{E}\left[\sum_t I[H_1 + H_2 \text{ contain} \ge i]\right] \le \underbrace{\mathbb{E}\left[\sum_t I[H_1 + H_2 \text{ contain} \ge i] \mid B_i \le \beta_i\right]}_{n \cdot \left(\frac{\beta_i}{n}\right)^2} \cdot 1 + n \cdot \frac{1}{n^2}$$

where, as long as $\mathbb{E}[\sum_t I[H_1 + H_2 \text{ contain} \ge i]] \ge 24 \ln n$, the Chernoff bound gives $B_{i+1} \le \frac{3}{2} n \left(\left(\frac{\beta_i}{n}\right)^2 + \frac{1}{n^2}\right)$ w.p. $1 - \frac{1}{n^2}$. Notice, $B_2 \le \frac{n}{2}$ w.p. 1. Now, how large can be the smallest i^* s.t. $\mathbb{E}[\sum_t I[H_1 + H_2 \text{ contain} \ge i^*]] < 24 \ln n$?

We have $\frac{B_3}{n} \le \frac{3}{2}\left(\frac{1}{2^2} + \frac{1}{n^2}\right)$, $\frac{B_4}{n} \le \frac{3}{2}\left(\left(\frac{3}{2}\right)^2\left(\frac{1}{2}\right)^4 + O\left(\frac{1}{n^2}\right)\right)$, ... $\frac{B_{i+2}}{n} \le \underbrace{\left(\frac{3}{2}\right)^{2^i - 1}}_{\frac{2}{3}\left(\frac{3}{4}\right)^{2^i}} \cdot \left(\frac{1}{2}\right)^{2^i} + O\left(\frac{i}{n^2}\right)$

so for $i = O(\log \log n)$, the upper bound is $< \frac{24 \ln n}{n}$.

Now, whatever the real expectation is, $B_{i^*} \le 36 \ln n$ w.p. $1 - \frac{1}{n^2}$. (Comparing to "rounded up" probabilities.) To conclude, we'll show $B_{i^*+2} < 1$ w.p. $1 - O(\frac{1}{n})$: if $B_{i^*+2} \ge 1$, at least two elements must have had both indices contain $\ge i^*$ elements. By a union bound over all $\binom{n}{2}$ pairs of elements, the total prob. of this occurring is $\le \binom{n}{2}\left(\frac{36 \ln n}{n}\right)^4 \le O\left(\frac{1}{n}\right)$. Finally, if $B_{i^*+2} < 1$, indeed, all indices have $O(\log \log n)$ elts., as needed ▨

Our analysis here used the Chernoff bound, which assumes that the hash values are truly uniform and independent. The bound is useful as-is for load balancing when we wish to assign clients/jobs/etc. to servers as they arrive; then we can simply make two independent choices for each arrival and

we don't need to store them. This doesn't work for a dictionary structure. Is pairwise independence enough? Unfortunately, no. But a generalization suffices:

Def^n The RVs $\{X_1, \dots, X_n\}$ are k-wise independent if for all sets $S \subseteq \{1, \dots, n\}$ of size k and assignments $X_i = x_i$ for $i \in S$,

$$\Pr[X_i = x_i \text{ for all } i \in S] = \prod_{i \in S} \Pr[X_i = x_i].$$

We needed the Chernoff bound to conclude $\Pr[\sum_i I_i > \frac{3}{2}\mathbb{E}[I_i]] < \frac{1}{n^2}$ for our independent indicators. Recall that we prove our tail bounds by transforming the inequality, e.g., Chebyshev squares $X - \mathbb{E}[X]$; Chernoff exponentiates $\sum_i I_i > (1+\varepsilon)\mathbb{E}[\sum_i I_i] \Leftrightarrow e^{\sum_i I_i} > e^{(1+\varepsilon)\cdot \mathbb{E}[\sum_i I_i]}$; then we use Markov's inequality to get a high-probability bound. For k-wise independence, we can use a k^{th} power:

$$\sum_i I_i > (1+\varepsilon)\mathbb{E}[\sum_i I_i] \Leftrightarrow \left(\sum_i I_i\right)^k > (1+\varepsilon)^k \mathbb{E}[\sum_i I_i]^k \overset{\text{k-wise indep.}}{=} (1+\varepsilon)^k \mathbb{E}[(\sum_i I_i)^k]$$

For non-negative I_i's or even k, we can now invoke Markov:

Theorem (k-wise independence tail bound) For a k-wise indep. collection $\{I_1, \dots, I_n\}$ (nonneg. or even k) and any $\varepsilon > 0$,

$$\Pr\left[\sum_i I_i > (1+\varepsilon)\mathbb{E}[\sum_i I_i]\right] < \left(\frac{1}{1+\varepsilon}\right)^k.$$

For $\varepsilon = \frac{1}{2}$, what value for k suffices to get $\frac{1}{n^2}$ prob.?

$\left(\frac{2}{3}\right)^k = \frac{1}{n^2} \Leftrightarrow 2\log n = k \log\frac{3}{2} \Leftrightarrow k = \frac{2}{\log\frac{3}{2}} \log n$. In general, for any $\varepsilon > 0$ and any $\frac{1}{n^d}$ $d \in \mathbb{N}$, $k = O(\log n)$

So, how do we get k-wise independence? A similar construction:

Theorem The family $\mathcal{H} = \{h_{a_0, a_1, \dots a_{k-1}}(x) = \sum_{i=0}^{k-1} a_i x^i \bmod N : a_i \in \{0, \dots, N-1\}\}$ is k-wise independent (for prime N).

Proof Suppose $x_1, \dots, x_k$ are all distinct. Then the matrix $\begin{bmatrix} x_1^{k-1} & x_1^{k-2} & \cdots & x_1 & 1 \\ \vdots & & & & \\ x_k^{k-1} & x_k^{k-2} & \cdots & x_k & 1 \end{bmatrix}$ has full rank mod N: indeed, if we subtract x_1 times the neighboring column, we get $\begin{bmatrix} 0 & \cdots & 0 & 1 \\ x_2^{k-2}(x_2-x_1) & \cdots & (x_2-x_1) & 1 \\ \vdots & & \vdots & \vdots \\ x_k^{k-2}(x_k-x_1) & \cdots & (x_k-x_1) & 1 \end{bmatrix}$ and subtracting x_2 gives $\begin{bmatrix} 0 & \cdots & & & 1 \\ 0 & \cdots & & & 1 \\ \vdots & & & & \vdots \\ x_k^{k-3}(x_k-x_2)(x_k-x_1) & \cdots & (x_k-x_2)(x_k-x_1) & (x_k-x_1) & 1 \end{bmatrix}$ and so on, thus giving a matrix with determinant $\prod_{i<j} (x_j - x_i) \bmod N$ which is nonzero since N is prime and $0 < (x_j - x_i) \bmod N < N$. Thus, the system $\begin{bmatrix} x_1^{k-1} & \cdots & x_1 & 1 \\ \vdots & & \vdots & \vdots \\ x_k^{k-1} & \cdots & x_k & 1 \end{bmatrix} \begin{bmatrix} a_{k-1} \\ \vdots \\ a_0 \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}$ has a unique solution mod N:

Hence, $\Pr[H(x_1) = y_1, \dots, H(x_k) = y_k] = \frac{1}{N^k} = \prod_{i=1}^{k} \Pr[H(x_i) = y_i]$ ∎