COMS 6998: Randomness in Computing — Sep 24, 2012

Perfect Hashing and Chebyshev's Inequality

Instructor: *Xi Chen* — Scribes: *Clément Canonne*

1 Perfect Hashing

1.1 Last week

Remember the definition of a t-universal hashing family, for a given *universe* $\mathcal{U}$ (set of elements):

Definition 1 (Universal hashing family)**.** *A family H of functions from $\mathcal{U}$ to $[t]$ is said to be a t-*universal hashing family *if, for any two distinct elements $k, k' \in \mathcal{U}$, $\mathbb{P}\{\, h(k) = h(k') \,\} \leq \frac{1}{t}$, when h is drawn uniformly at random from H.*

Let $S = \{k_1, \dots k_n\} \subset \mathcal{U}$ be a set of *keys*: we are interested in testing membership to S. Last week, we proved the following lemma, where H_m denotes a m-universal hashing family:

Lemma 2. *If $h \in H_m$ is picked uniformly at random, where $m = 2\binom{n}{2}$, the probability that no two keys from S collide is at least $\frac{1}{2}$.*

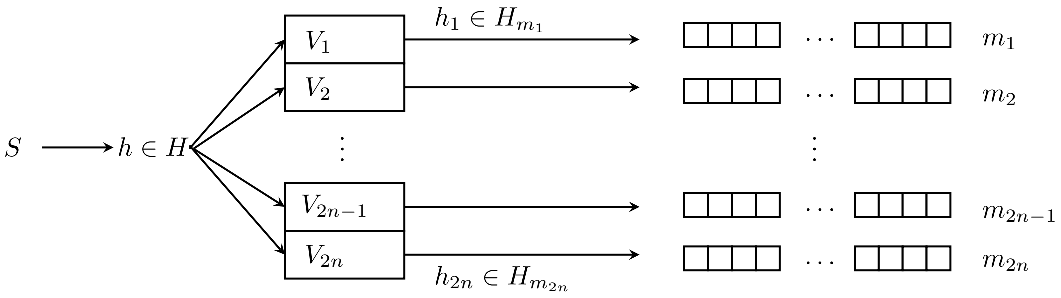
Problem: this is not good enough, as we need H to have size quadratic on n, to achieve only a probability of success $\frac{1}{2}$.

1.2 New attempt: two-layer scheme

We start with a table of size $2n$, and draw a hash function $h \sim H_{2n}$ uniformly at random. Let V_i be the set of keys that h assigns to slot i, and $t_i = |V_i|$ (note that $\sum_{j=1}^{2n} t_j = n$). Assume we have the following requirement for the integers $(t_j)_{1 \leq j \leq 2n}$:

$$\binom{t_1}{2} + \binom{t_2}{2} + \cdots + \binom{t_{2n}}{2} \leq n \tag{1}$$

then the *two-layer hashing scheme* illustrated below only requires linear space $\mathrm{O}(n)$. This scheme first maps k_i ($1 \leq i \leq n$) to a slot V_j (j in $\{1, \dots, 2n\}$) according to h, and computes the number t_i of elements each slot V_i contains. If (1) is not satisfied, it samples another $h \sim H_{2n}$ and start over; otherwise, for each slot V_j, it finds another hash function $h_i \sim H_{m_i}$, with $m_i = 2\binom{t_i}{2}$:



Proof. To see why (1) holds with high (constant) probability for the function h picked at the first step, let us define $X_{ij} = \begin{cases} 1 & \text{if } h(k_i) = h(k_j) \\ 0 & \text{otherwise} \end{cases}$. Then,

$$\underbrace{\mathbb{E}\sum_{ij} X_{ij}}_{\mathbb{E}\sum_{i=1}^{2n}\binom{t_i}{2}} = \sum_{ij} \mathbb{E}X_{ij} \leq \binom{n}{2}\frac{1}{2n} < \frac{n}{4}$$

It follows that, with Markov's inequality, $\mathbb{P}\left\{ \sum_{i=1}^{2n} \binom{t_i}{2} > n \right\} < \frac{1}{4}$.

We still need to show that for each $i \in [2n]$, by picking h_i from H_{m_i} uniformly at random, there is, whp, no conflict between keys in V_i. This was proved in the last lecture that if m_i is set to be $2\binom{t_i}{2}$, h_i induces no collision with probability at least 1/2. As a result, we can draw randomly an h_1 from H_{m_1}, check if there is any collision between keys in V_1. If so, draw a hash function from H_{m_1} again. In expectation we only need to try twice in order to find an h_1 that induces no collision between keys in V_1. Then we move to V_2 and find a hash function h_2, so on and so forth until we find a hash function h_i for every $i \in [2n]$. Using Markov's inequality, it can be shown that with probability at least 1/2, the total hash functions we draw is at most $8n$. It follows that such a two-level hash table can, with high constant probability, be constructed efficiently. □

2 Variance and (new) concentration inequalities

Definition 3 (Moments and variance)**.** *Let X be a r.v., and $k \in \mathbb{N}$. Then, whenever this expectation is defined, the* k^{th} moment of X *is the quantity* $\mathbb{E}X^k$.
The variance *of X is then defined as* $\operatorname{Var} X = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \mathbb{E}X^2 - (\mathbb{E}X)^2$*, and its* standard deviation *as* $\operatorname{std}(X) = \sqrt{\operatorname{Var} X}$.

The variance of a random variable measures how it varies around its expectation; the more it is, the more the random variable usually deviates from its mean – for instance, a r.v. with variance 0 is almost surely constant.

2.1 Applications to concentration inequalities

We are interested in bounding the *tail probabilities*, that is, for $t > 0$, the quantities

$$\mathbb{P}\{\, X > \mathbb{E}X + t \,\} \qquad \mathbb{P}\{\, X < \mathbb{E}X - t \,\} \qquad \mathbb{P}\{\, |X - \mathbb{E}X| > t \,\}$$

Theorem 4 (Chebyshev's Inequality)**.** *If X is a r.v., such that $\operatorname{Var} X$ is well-defined, then, $\forall t > 0$,*

$$\mathbb{P}\{\, |X - \mathbb{E}X| > t \,\} \leq \frac{\operatorname{Var} X}{t^2}$$

Proof. Define the r.v. $Y = (X - \mathbb{E}X)^2$: we have $\mathbb{E}Y = \operatorname{Var} X$, and

$$\mathbb{P}\{\, |X - \mathbb{E}X| > t \,\} = \mathbb{P}\Big\{\, \sqrt{Y} > t \,\Big\} = \mathbb{P}\Big\{\, Y > t^2 \,\Big\} \underset{\text{(Markov)}}{\leq} \frac{\mathbb{E}Y}{t^2} = \frac{\operatorname{Var} X}{t^2}$$

□

2.2 Coupon Collector's problem: beyond Markov's loose bound

Recall we had $\mathbb{E}X = \sum_{i=1}^{n} \mathbb{E}X_i = n\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) = nH_n$, where H_n denotes the harmonic series ($H_n = \ln n + \gamma + \mathrm{o}(1)$), and X is the random variable which takes as value the number of draws before having seen at least once every number from 1 to n.
Markov's inequality gives

$$\mathbb{P}\{\, X > 2nH_n \,\} < \frac{1}{2} \tag{2}$$

which is not very tight. Let us apply Chebyshev's inequality – for this, we first need to compute $\operatorname{Var} X$.

Linearity of variance? It would be great if, by chance, it happened that $\operatorname{Var} X = \operatorname{Var} \sum_{i=1}^{n} X_i =_{(?)} \sum_{i=1}^{n} \operatorname{Var} X_i$. Unfortunately, *the variance is not linear in general:* $\operatorname{Var}(X + Y) \neq \operatorname{Var} X + \operatorname{Var} Y$. However, it is true under some additional assumption:

Theorem 5. *Suppose X, Y are two* independent *random variables. Then,*

$$\operatorname{Var}(X + Y) = \operatorname{Var} X + \operatorname{Var} Y$$

In general, if $(X_j)_{1 \leq j \leq m}$ is a family of pairwise independent *r.v., then* $\operatorname{Var} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \operatorname{Var} X_i$.

Proof. Suppose X, Y are two r.v. for which the variance is defined, and with expectations respectively

μ_X, μ_Y. Then,

$$\begin{aligned}\operatorname{Var}(X+Y) &= \mathbb{E}\Big[(X+Y)^2\Big] - (\mathbb{E}[X+Y])^2 \underset{\text{(linearity)}}{=} \mathbb{E}\Big[X^2\Big] + \mathbb{E}\Big[Y^2\Big] + 2\mathbb{E}[XY] - (\mathbb{E}X + \mathbb{E}Y)^2 \\ &= \mathbb{E}\Big[X^2\Big] + \mathbb{E}\Big[Y^2\Big] + 2\mathbb{E}[X]\,\mathbb{E}[Y] - \mu_X^2 - \mu_Y^2 - 2\mu_X\mu_Y \\ &= \underbrace{\mathbb{E}\Big[X^2\Big]}_{} + \overbrace{\mathbb{E}\Big[X^2\Big]}^{} + 2\mu_X\mu_Y - \underbrace{\mu_X^2}_{} - \overbrace{\mu_Y^2}^{} - 2\mu_X\mu_Y \\ &= \operatorname{Var} X + \operatorname{Var} Y\end{aligned}$$

where we used the fact that if two r.v. are independent, the expectation of their product is the product of their expectations.
By induction, the result extends to finite families of r.v. □

Observation 6. *The converse is not true:* $(\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y) \not\Rightarrow$ *(X, Y independent).*

Fact 7. *If* $\operatorname{Var} X$ *is defined, then, for all* $\lambda \in \mathbb{R}$, $\operatorname{Var}(\lambda X) = \lambda^2 \operatorname{Var} X$.

Lemma 8 (Variance of a Geometric Law)**.** *Let* $X \sim \operatorname{Geom}(p)$*, for some* $p \in (0,1]$*. Then,* $\operatorname{Var} X = \frac{1-p}{p^2}$.

Back to the coupon collector: recall that $X_i \sim \operatorname{Geom}(p_i)$, with $p_i = 1 - \frac{i-1}{n} = \frac{n-i+1}{n}$, and that the X_i's are pairwise independent. It follows that

$$\begin{aligned}\operatorname{Var} X &= \sum_{i=1}^{n} \operatorname{Var} X_i = \sum_{i=1}^{n} \frac{1-p_i}{p_i^2} = \sum_{i=1}^{n} \frac{i-1}{n} \cdot \frac{n^2}{(n-i+1)^2} = n \sum_{i=1}^{n} \frac{i-1}{(n-i+1)^2} \\ &= n \sum_{k=1}^{n} \frac{n-k}{k^2} = n^2 \sum_{k=1}^{n} \frac{1}{k^2} - n \sum_{k=1}^{n} \frac{1}{k} \leq \frac{\pi}{6} n^2 - nH_n \qquad \left(\text{as } \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi}{6}\right)\end{aligned}$$

Applying Chebyshev's inequality, we get

$$\mathbb{P}\{\, X > 2nH_n \,\} < \frac{\frac{\pi}{6} n^2 - nH_n}{4n^2 H_n^2} = \mathrm{O}\left(\frac{1}{\log^2 n}\right) \tag{3}$$

If still not optimal, this is clearly a strong improvement over (2).

Direct approach towards a bound: following a "naive" approach here actually gives a tighter bound, as we shall see. Let E_i be the event *"the i^{th} ball is never drawn in the first $2nH_n$ rounds"*. Clearly,

$$\mathbb{P}\{\, X > 2nH_n \,\} = \mathbb{P} \bigcup_{i=1}^{n} E_i \underset{\substack{\text{Union}\\ \text{bound}}}{\leq} \sum_{i=1}^{n} \mathbb{P}\, E_i$$

and

$$\begin{aligned}
\mathbb{P}\, E_i &= \left(1-\frac{1}{n}\right)^{2nH_n} = \mathrm{e}^{\ln\left(1-\frac{1}{n}\right)\cdot 2nH_n} = \exp\_\ 2nH_n\left(\frac{1}{n}+\frac{1}{2n^2}+\mathrm{o}\!\left(\frac{1}{n^2}\right)\right)\\
&= \exp\_\ 2\left(\ln n+\gamma+\mathrm{o}(1)\right)\left(1+\frac{1}{2n}+\mathrm{o}\!\left(\frac{1}{n}\right)\right) = \exp\_\ 2\left(\ln n+\gamma+\mathrm{o}(1)\right)\\
&= \frac{1}{n^2}\,\mathrm{e}^{-2\gamma+\mathrm{o}(1)} \underset{n\to\infty}{\sim} \frac{\mathrm{e}^{-2\gamma}}{n^2} = \Theta\!\left(\frac{1}{n^2}\right)
\end{aligned}$$

Therefore,

$$\mathbb{P}\{\, X > 2nH_n \,\} \leq n\cdot\frac{1}{n^2}\,\mathrm{e}^{-2\gamma+\mathrm{o}(1)} < \frac{1}{n} \qquad \text{for } n \text{ big enough.} \tag{4}$$

2.3 Two-point sampling

Suppose that, for some language $L \in \mathsf{RP}$, we have an algorithm $\mathcal{A}$ such that (for some prime p):

- if $x \in L$, then, for at least half of the integers q from $\mathbb{Z}_p$, $\mathcal{A}(x,q)=1$;
- if $x \notin L$, then, for all $q \in \mathbb{Z}_p$, $\mathcal{A}(x,q)=0$.

The goal is to *amplify* this $\frac{1}{2}$ probability. An obvious way to do so would be as follows:

```
Draw uniformly at random $q_1,\dots,q_k$ from $\mathbb{Z}_p$
if $\exists i \in [k]$, $\mathcal{A}(x,q_i)=1$ then
  return 1
else
  return 0
end if
```

The probability of failure is then shrunk to at most $\frac{1}{2^k}$. However, this method is good only if we have access to an unlimited supply of randomness (random numbers are "free"). What if we only have access to *two* random $a,b \in \mathbb{Z}_p$?

Idea Run on $\mathcal{A}(x, ak+b)$, for several values of k.
$\rightsquigarrow$Problem: the $(ak+b)_k$ are no longer independent! That is, they are not *mutually independent*... however, it is easy to see that they remain *pairwise independent*. As we shall see, this is enough:

Define $X_k = \mathbb{1}_{\{ak+b \text{ is a witness}\}}$ and $X=\sum_{k=1}^t X_k$, where $q\in\mathbb{Z}_p$ is said to be a *witness for x* if $\mathcal{A}(x,q)=1$. By assumption, $\mathbb{P}\{\, X_k = 1\,\} \geq \frac{1}{2}$, and we also have $\operatorname{Var} X_k = \mathbb{P}\{\, X_k=1\,\}\,(1-\mathbb{P}\{\, X_k=1\,\}) \leq \frac{1}{4}$. Therefore, by pairwise independence, $\operatorname{Var} X = \operatorname{Var}\sum_{k=1}^t X_k \leq \frac{t}{4}$, and applying Chebyshev's inequality yields

$$\mathbb{P}\{\, X = 0\,\} \leq \mathbb{P}\{\, |X-\mathbb{E}X| \geq \mathbb{E}X\,\} \leq \frac{t/4}{(t/2)^2} = \frac{1}{t}$$

Therefore, after t runs of the algorithms, we reduced the one-sided probability of failure from $\frac{1}{2}$ to $\frac{1}{t}$ (not as good as the exponential improvement for truly independent q_i's, but already a huge boost).

3 Randomized median algorithm

Here is described an algorithm with, on input elements $a_1, \dots, a_n$ from an ordered set, finds the median of the a_i's with probability $1 - \mathrm{o}(1)$ and running time $2n + \mathrm{o}(n)$. Even better, the algorithm is guaranteed to return either the correct answer, or FAIL.

Idea let S be a string of size $|S| = n$ (for convenience, we assume all elements from S are distinct). First, we sample $n^{3/4}$ elements from S, uniformly at random with replacement: this defines a (multi)set R. With high probability, the median of this sample R will be "close" to the real median. More precisely:

```
Draw $R$ by sampling independently $n^{3/4}$ elements from $\mathcal{U}(S)$ (uniform distribution)
Let $\ell$ be the $(\frac{1}{2}n^{3/4} - \sqrt{n})^{\text{th}}$ smallest element in $R$     {can be found in time $\mathrm{o}(n)$, e.g. by sorting $R$}
Let $r$ be the $(\frac{1}{2}n^{3/4} + \sqrt{n})^{\text{th}}$ smallest element in $R$
Compare $\ell$ with all elements from $S$ to compute $\operatorname{rank}_S(\ell)$     {$n-1$ comparisons}
Compare $r$ with all elements from $S$ to compute $\operatorname{rank}_S(r)$     {$n-1$ comparisons}
Define $C = \{\, a \in S \mid \ell \leq a \leq r \,\}$
if $\operatorname{rank}_S(\ell) > \frac{n}{2}$ then
  return FAIL     {Event $E_1$}
else if $\operatorname{rank}_S(r) < \frac{n}{2}$ then
  return FAIL     {Event $E_2$}
else if $|C| > 4n^{3/4}$ then
  return FAIL     {Event $E_3$}
else
  Find the $(\frac{n}{2} - \operatorname{rank}_S(\ell))^{\text{th}}$ element $\hat{m}$ in $C$     {e.g. by sorting $C$: still $\mathrm{o}(n)$-time}
  return $\hat{m}$
end if
```

Probability of failure By an union bound, $\mathbb{P}\{\,\mathsf{FAIL}\,\} \leq \mathbb{P}\,E_1 + \mathbb{P}\,E_2 + \mathbb{P}\,E_3$. We bound each term separately (hereafter, $\operatorname{med}(S)$ will be used to denote the true median of S):

- First, observe that $\mathbb{P}\,E_1 \leq \mathbb{P}\Big\{\,\#(\text{samples} < \operatorname{med}(S)) \leq \frac{1}{2}n^{3/4} - \sqrt{n}\,\Big\}$. For $m = n^{\frac{3}{4}}$, define $X_1, \dots, X_m$, where $X_i = \mathbb{1}_{\left\{i^{\text{th}}\text{ sample} < \operatorname{med}(S)\right\}}$. By definition, linearity and independence,

$$\mathbb{P}\{\,X_i = 1\,\} = \frac{1}{2}, \qquad \operatorname{Var} X_i = \frac{1}{4}, \qquad \mathbb{E}\sum_{i=1}^{m} X_i = \frac{m}{2} \qquad \text{and} \qquad \operatorname{Var}\sum_{i=1}^{m} X_i = \frac{m}{4}$$

  Hence,

$$\mathbb{P}\,E_1 \leq \mathbb{P}\left\{\,\sum_{i=1}^{m} X_i < \frac{m}{2} - \sqrt{n}\,\right\} \leq \mathbb{P}\left\{\,\left|\sum_{i=1}^{m} X_i - \mathbb{E}\sum_{i=1}^{m} X_i\right| > \sqrt{n}\,\right\} \underset{\text{(Chebyshev)}}{\leq} \frac{m}{4\sqrt{n}^2} = \frac{1}{4n^{1/4}} \tag{5}$$

- Similarly,

$$\mathbb{P}\,E_2 \leq \frac{1}{4n^{1/4}} \tag{6}$$

- As $E_3 = \{\text{more than } 4m \text{ points from } S \text{ fall in } \{\ell, \dots, r\}\}$, we have $E_3 \subset E_3^1 \cup E_3^2$, where

$$E_3^1 = \{\text{at least } 2m \text{ points from } S \text{ are} > \operatorname{med}(S) \text{ and fall in } \{\ell, \dots, r\}\}$$
$$E_3^2 = \{\text{at least } 2m \text{ points from } S \text{ are} < \operatorname{med}(S) \text{ and fall in } \{\ell, \dots, r\}\}$$

  (if $\operatorname{med}(S) \notin \{\ell, \dots, r\}$, in particular, we have either E_3^1 or E_3^2)

  - Let d be the $(\frac{n}{2} + 2m)^{\text{th}}$ smallest point in the whole string S. Since $E_3^1 \subseteq \{r \geq d\}$,

$$\begin{aligned} \mathbb{P}\, E_3^1 &\leq \mathbb{P}\left\{ \#\,(\text{samples in } \{d, \dots, \max S\}) > \frac{m}{2} - \sqrt{n} \right\} \\ &\leq \mathbb{P}\left\{ \left| \sum_{i}^{m} Y_i - \mathbb{E} \sum_{i=1}^{m} Y_i \right| > \sqrt{n} \right\} \\ &\underset{(\text{Chebyshev})}{\leq} \frac{m}{4n} = \frac{1}{4n^{1/4}} \end{aligned} \tag{7}$$

    where Y_i is defined as the indicator r.v. $\mathbb{1}_{\left\{i^{\text{th}} \text{ sample } \in \{d, \dots, \max S\}\right\}}$.

  - by symmetry, the same analysis applies to E_3^2.

Putting (5), (6) and (7) together, we end up with

$$\mathbb{P}\{\,\mathsf{FAIL}\,\} \leq \frac{1}{4n^{1/4}} + \frac{1}{4n^{1/4}} + 2 \cdot \frac{1}{4n^{1/4}} = \frac{1}{n^{1/4}} = o(1)$$

□