An Applications: Hashing

Balls = Elements

Bins = Table entries

assuming hashing with chaining and perfectly random hash functions

(to be relaxed later)

1 Occupancy Problems

m Balls placed independently and uniformly at random into

n Bins

X_i number of balls in bin i

More Application

Parallel disk scheduling:

bins = disks

balls = blocks to be read/written

randomization with random mapping logical address space $\rightarrow$ disks

Parallel Processing

bins = processors

balls = data items

random placement of data saves coordination overhead

Questions

- Is there a collision? $\mathbb{P}\left[\exists i : X_i > 1\right]$
  $\leadsto$ perfect hashing, birthday paradoxon
- How many balls are needed to cover all bins $\neg\exists i : X_i = 0$?
  $\leadsto$ Coupon Collector's problem
- Number of empty bins $\left|\left\{i : X_i = 0\right\}\right|$
  $\leadsto$ amount of memory "wasted"
- Maximum occupancy $L_{\max} = \max_{i=1}^{m} X_i$
  $\leadsto$ sth like worst case access time
- How much overflow if bins have capacity K?
  $\leadsto$ overflow area for external hashing

More Accurate Calculation

$\mathcal{E}_i$ = Ball i does not collide with balls $1..i-1$.

$$\mathbb{P}\left[\bigwedge_{i=1}^{m}\mathcal{E}_i\right] = \mathbb{P}[\mathcal{E}_1]\,\mathbb{P}[\mathcal{E}_2 \mid \mathcal{E}_1]\,\mathbb{P}[\mathcal{E}_3 \mid \mathcal{E}_1 \wedge \mathcal{E}_2]\cdots\mathbb{P}\left[\mathcal{E}_m \mid \bigwedge_{i=1}^{m-1}\mathcal{E}_i\right]$$

$$= 1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{m-1}{n}\right) \rightsquigarrow 23 \text{ by brute force}$$

$$= \prod_{i=1}^{m-1}\left(1-\frac{i}{n}\right) \qquad \text{Power Series } e^{-x} \geq 1-1/x$$

$$\leq \prod_{i=1}^{n-1}\exp\left(-\frac{i}{n}\right) = \exp\left(-\frac{m(m-1)}{2n}\right)$$

solve $\exp\left(-\frac{m(m-1)}{2\cdot 365}\right) \leq \frac{1}{2}$

$\rightsquigarrow m \geq 22.99994315$

1.1 The Birthday Paradoxon

$n = 365$, How big should m be so that $\mathbb{P}[\exists i : X_i \geq 2] > 1/2$

(It pays to bet).

A rough estimate:

$Y_{ij} = 1$ if balls i and j ($i < j$) collide, $Y_{ij} = 0$ otherwise.

$\mathbb{E}[Y_{ij}] = \mathbb{P}[Y_{ij} = 1] = \frac{1}{n}$.

There are $\frac{m(m-1)}{2}$ pairs of balls.

$$\Rightarrow \mathbb{E}\left[\sum_{i<j} Y_{ij}\right] = \frac{m(m-1)}{2n}$$

i.e., at 28 people the expected number of collisions is 1.

The Geometric Distribution

$\mathbb{P}[Y = k] = p(1-p)^{k-1}$

$$\mathbb{E}[Y] = \sum_{j=1}^{\infty} j\mathbb{P}[Y = j]$$

$$= \sum_{j=1}^{\infty}\mathbb{P}[Y \geq j] = \sum_{j=1}^{\infty}(1-p)^{j-1} = \frac{1}{1-(1-p)} = \frac{1}{p}$$

The geometric distribution is memoryless, i.e.,

$$\mathbb{P}[Y = k+\ell \mid Y > \ell] = \mathbb{P}[Y = k] \ .$$

1.2 The Coupon Collectors Problem

Let X = number of balls until all bins have a ball

T_j is the number balls until a new bin is filled when j bins are already filled.

$X = \sum_{j=0}^{n-1} T_j$

Success probability: $p_j = \frac{n-j}{n}$

T_j is geometrically distributed with parameter $p_j \Rightarrow \mathbb{E}[T_j] = \frac{1}{p_j}$

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{j=0}^{n-1} T_j\right] = \sum_{j=0}^{n-1}\mathbb{E}[T_j] = \sum_{j=0}^{n-1}\frac{n}{n-j}$$

$$= n\sum_{j=0}^{n-1}\frac{1}{n-j} = n\sum_{j=1}^{n}\frac{1}{j} = nH_n = n\ln n + \Theta(1)$$

1.4 Maximum Occupancy

we want to bound $\mathbb{E}[L_{\max}] = \mathbb{E}\left[\max_{j=1}^{n} X_j\right]$

Fix a bin, say bin 1

$Y_i := 1$ iff ball i is allocated to bin 1, i.e., $X_1 = \sum_{i=1}^{m} Y_i$

Suppose $\mathbb{P}[X_1 > k^*] \leq n^{-3}$ for some threshold k^*

Observation:

$$\mathbb{P}[L_{\max} > k^*] = \mathbb{P}[\exists j : X_j > k^*] \leq \sum_{j=1}^{n} \mathbb{P}[X_j > k^*] = n\mathbb{P}[X_1 > k^*] \leq n^{-2}$$

Hence, $\mathbb{E}[L_{\max}] =$

$$\sum_{k=0}^{m} k\mathbb{P}[L_{\max} = k] \leq k^* + m\mathbb{P}[X > k^*] \leq k^* + \frac{m}{n^2} = k^* + \frac{\mathbb{E}[X_1]}{n}$$

1.3 Number of Empty Bins

Let $Y = |\{i : X_i = 0\}|$

Exercise: Show that $\mathbb{E}[Y] \leq n \exp\left(-\frac{m}{n}\right)$

2.2 Chernoff Bound

Let $X_1, \ldots, X_n$ denote independent 0-1 random variables with $\mathbb{P}[X_i = 1] = p, q = 1 - p, X = \sum_{i=1}^{n} X_i$

$$\forall t > 0 : \mathbb{P}[X > (p+t)n] \leq \left(\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}\right)^n$$

2 Tail Bounds

Often large deviations of random variable from their expectation are unlikely.

We want to make this observation quantitative.

2.1 Markov's Inequality

$$\forall Y \in \mathbb{R}^{+\Omega} : \forall t > 0 : \mathbb{P}[Y > t] \leq \frac{\mathbb{E}[Y]}{t}$$

Proof. We show $\mathbb{E}[Y] \geq t\mathbb{P}[Y > t]$.

$$\mathbb{E}[Y] = \int_0^t y\phi(y)dy + \int_t^\infty y\phi(y)dy$$
$$\geq \int_t^\infty y\phi(y)dy \geq \int_t^\infty t\phi(y)dy = t\int_t^\infty \phi(y)dy = t\mathbb{P}[Y > t]$$

□

$$\mathbb{P}[X > (p+t)n] \leq \left(\frac{(pe^{\lambda}+q)}{e^{\lambda(p+t)}}\right)^n$$

rhs is minimized for $\lambda = \ln(\frac{q(p+t)}{p(q-t)})$. Maple:

`solve(diff((p*E^L+q)/E^(L*(p+t)),L)=0,L);`

We get

$$\frac{pe^{\lambda}+q}{e^{\lambda(p+t)}} \rightsquigarrow \frac{p\frac{q(p+t)}{p(q-t)}+q}{\left(\frac{q(p+t)}{p(q-t)}\right)^{p+t}} = \left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t-1}\left(\frac{q(p+t)}{(q-t)}+q\right)$$

$$= \left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}$$

Maple: `simplify((q-t)/q*(q*(p+t)/(q-t)+q));` $\rightsquigarrow 1$

■

Proof approach from [Dubhashi/Panconesi 98]

consider arbitrary $\lambda > 0$.

$\mathbb{P}[X > (p+t)n] = \mathbb{P}\left[e^{\lambda X} > e^{\lambda(p+t)n}\right] \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda(p+t)n}}$ "turbo-charged Markov"

$\mathbb{E}\left[e^{\lambda X}\right]$ is the moment generating function

$$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda\sum_i X_i}\right] = \mathbb{E}\left[\prod_i e^{\lambda X_i}\right] = \prod_i \mathbb{E}\left[e^{\lambda X_i}\right]$$

$$= \prod_i (pe^{\lambda}+qe^0) = (pe^{\lambda}+q)^n,$$

i.e,

$$\mathbb{P}[X > (p+t)n] \leq \left(\frac{pe^{\lambda}+q}{e^{\lambda(p+t)}}\right)^n$$

and we are still free to choose λ.

Chernoff Hoeffding Bound

Let $X_1, \ldots, X_n$ denote independent random variables $\Omega \to [0,1]$ with $\frac{1}{n}\sum_i \mathbb{E}[X_i] = p$, $q = 1-p$, $X = \sum_{i=1}^n X_i$

$$\forall t > 0 : \mathbb{P}[X > (p+t)n] \leq \left(\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{q}{q-t}\right)^{q-t}\right)^n \quad (1)$$

Proof: ... not here

in a sense the case of identically distributed 0-1 variables is the worst case.

More Convenient Formats

Let $X_1, \ldots, X_n$ denote independent random variables $\Omega \to [0,1]$, $X = \sum_{i=1}^n X_i$

$$\mathbb{P}[X > (1+\delta)\mathbb{E}[X]] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}[X]} \text{ for } 0 < \delta \quad (2)$$

$$\leq \exp\left(-\frac{\delta^2\mathbb{E}[X]}{3}\right) \text{ for } 0 < \delta < 1 \quad (3)$$

$$\mathbb{P}[X > (1-\delta)\mathbb{E}[X]] \leq \exp\left(-\frac{\delta^2\mathbb{E}[X]}{2}\right) \text{ for } 0 < \delta < 1 \quad (4)$$

$$\mathbb{P}[|X - \mathbb{E}[X]| > t] \leq 2\exp\left(-\frac{2t^2}{n}\right) \quad (5)$$

Good Balance

$$\mathbb{E}[L_{\max}] \stackrel{!}{\leq} (1+\delta+\mathcal{O}(n^{-1}))\frac{m}{n}+\mathcal{O}(1) \text{ for } m \geq \frac{9}{\delta^2} n \ln n$$

We show $\mathbb{P}\left[X_1 > (1+\delta)\frac{m}{n} \leq \frac{1}{nm}\right]$ for suff. large n

$\mathbb{E}[X_1] = \frac{m}{n}$. Chernoff Bound (3):

$$\mathbb{P}\left[X_1 > (1+\delta)\frac{m}{n}\right] \leq \exp\left(-\frac{\delta^2 m}{3n}\right) \leq \exp\left(-\frac{\delta^2 9 n \ln n}{3n\delta^2}\right)$$
$$= \exp(-3\ln n) = n^{-3} \ .$$

True threshold is more like $\frac{2}{\delta^2} n \ln(n)$?

2.3 Back To Maximum Occupancy

Let $L_{\max} = \max_{j=1}^{n} X_j$.

$$\mathbb{E}[L_{\max}] \leq \begin{cases} 1+o(1) & \text{If } m = o(\sqrt{n}) \text{ Birthday paradoxon} \\ \frac{\ln n}{\ln\ln n}+\mathcal{O}(1) & \text{if } m = \Theta(n) \text{ we outline } \frac{4\ln n}{\ln\ln n} \\ (1+\delta)\frac{m}{n}+\mathcal{O}(1) & \text{if } m \geq \frac{9}{\delta^2} n\ln n \text{ see below} \\ \frac{m}{n}+\mathcal{O}\left(\sqrt{\frac{m}{n}\log n}\right) & \text{if } m = \Omega(n\log n) \text{ use bound (3)} \end{cases}$$

The Case $m = n$ A First Attempt

What about $\mathbb{E}[L_{\max}] \leq \ln n$?

$\mathbb{E}[X_1] = 1$.

Chernoff Bound (2):

$$\mathbb{P}[X_1 > \ln n] \leq \left(\frac{e^{\ln n - 1}}{(\ln n)^{\ln n}}\right)^1 \text{ use } x^x = e^{\ln(x)x}$$
$$= \frac{n}{e n^{\ln\ln n}} \leq n^{1-\ln\ln n}$$

Did we overshoot $L_{\max}$ by a factor $\ln\ln n$?

The Case $m = n$

$\mathbb{E}[X_1] = 1$. Chernoff Bound (2):

$$\mathbb{P}\left[X_1 > \tfrac{4\ln n}{\ln\ln n}\right] \leq \frac{e^{\frac{4\ln n}{\ln\ln n}-1}}{\frac{4\ln n}{\ln\ln n}^{\frac{4\ln n}{\ln\ln n}}} \text{ use } x^x = e^{\ln(x)x}$$
$$= \exp\left(\frac{4\ln n}{\ln\ln n} - 1 - \ln\left(\frac{4\ln n}{\ln\ln n}\right)\frac{4\ln n}{\ln\ln n}\right)$$
$$= \exp\left(\frac{4\ln n}{\ln\ln n} - 1 - (\ln 4 + \ln\ln n - \ln\ln\ln n)\frac{4\ln n}{\ln\ln n}\right)$$
$$= \exp\left(\frac{4\ln n}{\ln\ln n} - 1 - \left(\frac{4\ln 4\ln n}{\ln\ln n} + 4\ln n - \frac{4\ln\ln\ln(n)\ln n}{\ln\ln n}\right)\right)$$
$$\leq \exp\left(-4\ln n - 1 + \frac{4\ln\ln\ln(n)\ln n}{\ln\ln n}\right) = n^{-e+o(1)} \leq n^{-2}$$

for sufficiently large n ■

Analysis

Theorem 1. $\forall$ *constants* α, β : *if* $s = \alpha n^{2/3} \ln^{1/3} n$ *and* $g = \sqrt{s \ln(\beta s)}$, *then fastSelect performs an expected number of* $n + \mathcal{O}\left(n^{2/3} \ln^{1/3} n\right)$ *memory accesses and* $n + \min(k, n-k) + \mathcal{O}\left(n^{2/3} \ln^{1/3} n\right)$ *comparisons.*

Proof outline:$\mathbb{E}\left[|\text{candidates}|\right]$

$$= 2g\frac{n}{s} = 2n\frac{\sqrt{s\ln(\beta s)}}{s} = 2n\sqrt{\frac{\ln(\beta s)}{s}}$$

$$= 2n\sqrt{\frac{\ln(\beta\alpha n^{2/3}\ln^{1/3} n)}{\alpha n^{2/3}\ln^{1/3} n}} = \mathcal{O}\left(n\sqrt{\frac{\ln n}{n^{2/3}\ln^{1/3} n}}\right) = \mathcal{O}\left(n^{2/3}\ln^{1/3} n\right)$$

$X_i := 1$ iff sample$[i] \le e_k$; $X := \sum_i X_i$.
The X_i are independent and $\mathbb{P}[X_i = 1] = k/n$.
Use Chernoff bound 5 . . .

3 The "Ultimate" Selection Algorithm

Function fastSelect(a: array of size n, k) -- Assume $k \le n/2$
  sample := $\langle a[i_0], \ldots, a[i_{s-1}]\rangle$ for random indices i_j
  ℓ := qSelect(sample, $\lfloor \frac{ks}{n} \rfloor - g$) and $-\infty$ if $\lfloor \frac{ks}{n} \rfloor - g < 0$
  r := qSelect(sample, $\lceil \frac{ks}{n} \rceil + g$)) and ∞ if $\lceil \frac{ks}{n} \rceil + g \ge n$
  candidates := $\langle\rangle$
  smaller := 0
  foreach $c := a[i]$, $i := 0$ **to** $n-1$ **do**
    if $c \le r$ **then**
      if $c \ge l$ **then** add c to candidates
      else smaller := smaller+1
  if smaller$\ge k \vee$ smaller $+ |\text{candidates}| < k$ **then**
    return qSelect(a, k) -- $e_k \notin$ candidates
  else return qSelect(candidates, $k-$smaller)

Example

Let $Y_1, \ldots, Y_n$ denote the bins choosen in an n balls into n bins game.
$f(y_1, \ldots, y_n)$:=# of empty bins if balls are placed into bins $y_1, \ldots, y_n$.
$\mathbb{E}[f(Y_1, \ldots, Y_n)] = n\left(1 - \frac{1}{n}\right)^n \approx n/e$
Bounded Difference Inequality $\rightsquigarrow$

$$\mathbb{P}\left[|f(X_1, \ldots, X_n) - n\left(1 - \frac{1}{n}\right)^n| \ge \sqrt{\alpha n}\right] \le 2\exp -\frac{\alpha n}{n} = 2e^{-\alpha}$$

3.1 More General Tail Bounds

Theorem 2 (The Method of Bounded Differences).
Consider independent random variables $X_i : \Omega \to M_i$, *and* $f : M_1 \times \cdots \times M_n \to \mathbb{R}$ *with the Lipshitz condition*

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, y_i, \ldots, x_n)| \le c_i \ .$$

then

$$\mathbb{P}\left[|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \ge t\right] \le 2\exp -\frac{t^2}{\sum_i c_i^2}$$

Proof: not here.