CSC 5450 Randomness and Computation 2013

Week 2: Moments and Deviations

Plan

① Balls and bins, the coupon collector's problem

② Markov inequality, Chebyshev inequality, finding median

③ Heuristic arguments, power of two choices.

Balls and Bins

It is a simple random process that underlies some basic phenomenon.

We have m balls and n bins. Each ball is thrown to a uniformly random bin independently.

We would like to study what does a typical situation look like.

There are many questions one can ask to understand the distribution of the balls into the bins.

We start from some easy calculations.

Expected Number of Balls in a Bin.

Let $B_{i,j}$ be the indicator variable that ball j is in bin i.

Then $E[\#\text{ balls in bin } i] = E[\sum_{j=1}^{m} B_{i,j}] = \sum_{j=1}^{m} E[B_{i,j}] = \sum_{j=1}^{m} \Pr[\text{ball } j \text{ in bin } i] = \sum_{j=1}^{m} \frac{1}{n} = \frac{m}{n}.$

In particular, when $m=n$, the expected number of balls in a bin is one.

Do we expect that most bins have about one ball most of the time?

In this famous ping-pong machine, there are five balls and five bins. There are some needles above the bins to make the ball's location pretty "random". You put in one dollar and play this game. You'd get a small prize if there are four balls in a row or in a column. You'd get two small prizes if there are five balls in a row or in a column. So, do you expect the answer to the above question to be yes?

Expected number of empty bins [MU 5.3]

Let Y_i be the indicator variable that bin i is empty.

Then, $E[Y_i] = \Pr(\text{bin } i \text{ is empty}) = (1-\frac{1}{n})^m \approx e^{-\frac{m}{n}}$ (using $1-x \leq e^{-x}$ and $1-x \approx e^{-x}$ for small x).

Then, $E[Y_i] = \Pr(\text{ bin } i \text{ is empty}) = (1-\frac{1}{n})^m \approx e^{-\frac{m}{n}}$ (using $1-x \le e^{-x}$ and $1-x \approx e^{-x}$ for small x).

So, $E[\#\text{ of empty bins}] = E[\sum_{i=1}^{n} Y_i] = \sum_{i=1}^{n} E[Y_i] = n \cdot e^{-\frac{m}{n}}$.

When $m=n$, we expected to see that a $\frac{1}{e}$-fraction of the bins are empty.

Just looking at the expectation of a more "global" variable gives a better understanding of the distribution.

Maximum Load [MU 5.2] What is the maximum number of balls in a bin typically?

A simpler question is for what m do we expect to see two balls in a bin (a "collison" occurs).

The birthday paradox is the case when $n=365$ (ignoring Feb 29).

The probability that there are no collision in the first m balls is :

$$(1-\frac{1}{n})(1-\frac{2}{n})\ldots(1-\frac{m-1}{n}) \le e^{-\frac{1}{n}} e^{-\frac{2}{n}} \cdots e^{-\frac{m-1}{n}} = e^{-\frac{(m-1)m}{2n}} \approx e^{-\frac{m^2}{2n}}.$$

This probability would be smaller than $1/2$ when $m = \sqrt{2n \ln 2}$.

For $n=365$, it says that when $m \ge 22.49$, the probability that the maximum load is at least two is at least $1/2$. This estimate is very close to the exact answer.

To summarize, we expect to see a collison when $m = \Theta(\sqrt{n})$. This observation is useful in different places (e.g. hashing, analyzing a heuristic for factoring integers, etc.)

An intuitive explanation is that there are m^2 pairs of possible collisons, and we expect some collisons would occur when $m^2 \approx n$, instead of the incorrect intuition that collisons would occur only when $m \approx n/2$.

The maximum load when $m=n$

The probability that a bin has at least k balls is at most $\binom{n}{k}(\frac{1}{n})^k$.

It is often that we have to deal with binomial coefficients.

Some useful bounds are : $(\frac{n}{k})^k < \binom{n}{k} < \frac{n^k}{k!} < (\frac{ne}{k})^k$. The proofs are left as exercises.

Using this bound, the above probability is at most $(\frac{ne}{k})^k(\frac{1}{n})^k = \frac{e^k}{k^k}$.

By the union bound, $\Pr[\text{ some bin has at least } k \text{ balls }] \le n \cdot \frac{e^k}{k^k} = e^{\ln n + k - k\ln k}$.

We would like to estimate the smallest k such that this probability is small enough.

In other words, we want the minimum k such that $k \ln k > \ln n$.

Setting $k = 3 \ln n / \ln \ln n$ would do (simple calculations, see MU 5.2.1).

Therefore, with high probability, the maximum load is at most $O(\ln n / \ln\ln n)$.

Chain Hashing [MU 5.5.1]

A hash table is a data structure to store a dictionary to support quick searching.

Think of the scenario where we have n IP addresses to store.

To support quick searching, an easy way is to open an array of size 256^4, but the problem is that this is space-inefficient if $n << 256^4$.

A hash function maps elements from a universe into $[0, n-1]$, e.g. $h: (\text{IP address}) \to [0, n-1]$, so that we can use an array of size n to store the n IP addresses.

An ideal hash function would behave like a random function, so that with high probability it works for any set of n IP addresses.

The above analysis shows that the maximum load is at most $O(\ln n / \ln\ln n)$, and this gives an upper bound of the worst case search time, assuming we do chain hashing, i.e. storing the values with the same hash value by a list (or a chain).

Of course, there are other issues about hashing that we have not discussed (e.g. construction of a hash function with small storage and fast evaluation time, etc).

We will study hashing to some depth later in the course.

Coupon Collector [MU 2.4.1]

Question: For what m do we expect to have no empty bins?

(picture from "oldcake.net"...)

Let X be the number of balls thrown until there are no empty bins.

Let X_i be the number of balls thrown when there are exactly i nonempty bins.

Then $E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$.

How to compute $E[X_i]$? Note that each X_i is a geometric random variable with parameter $p = \frac{i}{n}$.

Recall that a geometric random variable Y is given by the distribution that $\Pr(Y=k) = (1-p)^{k-1}p$.

In words, Y is the number of trials until the first success, when the success probability is p.

The expected value of a geometric random variable Y with parameter p is $\frac{1}{p}$.

There are at least three ways to see it:

① direct calculation from the definition with a differentition trick (see Ross, [1] of L1).

② use $E[Y] = \sum_{i=1}^{\infty} i \Pr(Y=i) \underset{\substack{\uparrow \\ \text{change of summation}}}{=} \sum_{i=1}^{\infty} \Pr(Y \geq i) = \sum_{i=1}^{\infty} (1-p)^{i-1} = 1/(1-(1-p)) = \frac{1}{p}$.

③ use conditional probability to argue that $E[Y] = p + (1-p)(E[Y]+1) = (1-p)E[Y]+1$, hence $E[Y] = 1/p$.

Anyway, we have $E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} 1/(i/n) = \sum_{i=1}^{n} \frac{n}{i} \approx n \ln n$ (again we see the Harmonic number).

It would be much cheaper for many coupon collectors to collaborate; we may do this in homework 1.

Tail Probabilities

It is often not enough to just have the expected value of a random variable.

In many situations we would like to determine whether the value is close to its expected value.

Markov inequality [MU 3.1]

As an example, we proved that the expected running time of randomized quicksort is at most $2n\ln n$.

What is the probability that its running time is at least $4n\ln n$?

Well, since the running time is non-negative, this probability can't be bigger than $1/2$, otherwise it would contradict the expected running time is at most $2n\ln n$.

This is essentially Markov's inequality.

Theorem Let X be a non-negative discrete random variable. Then $\Pr(X \geq a) \leq E[X]/a$ for all $a > 0$.

Proof $E[X] = \sum_i i \Pr(X=i) \geq \sum_{i \geq a} i \Pr(X=i) \geq a \sum_{i \geq a} \Pr(X=i) = a \Pr(X \geq a)$. □

In general, if we do not have additional information about the random variable, this is best possible.

But it is usually very weak in applications, e.g. Let X be the number of heads in n fair coin flips. Markov inequality just says that $\Pr(X \geq 3n/4) \leq E[X]/(3n/4) = (n/2)/(3n/4) = \frac{2}{3}$, although this probability is exponentially small (topic of next week).

Moments and Variance [MU 3.2]

To give a better bound on the deviation from its expected value, we need to know more information.

The k-th moment of a random variable X is $E[X^k]$, e.g. second moment is $E[X^2]$.

The variance of a random variable is defined as $\mathrm{Var}[X] = E[(X-E[X])^2] = E[X^2 - 2X \cdot E[X] + E[X]^2]$
$= E[X^2] - E[X]^2$.

The standard deviation of X is defined as $\sigma[X] = \sqrt{\mathrm{Var}[X]}$.

The covariance of two random variables X, Y is defined as $\mathrm{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])]$.

- if $\mathrm{Cov}(X,Y) > 0$, then we say that X and Y are positively correlated.
- if $\mathrm{Cov}(X,Y) < 0$, then we say that X and Y are negatively correlated.

The following are some facts collected from [MU]; the proofs are straightforward (see [MU 3.2])

① $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$.

② If X and Y are independent, then $E[XY] = E[X]E[Y]$.

③ If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$ and $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Bucket Sort [MU 5.2.2]

Suppose we have $n = 2^m$ elements to be sorted, each is uniformly random from $[0, 2^k)$.

Create n buckets, each takes 2^{k-m} numbers (put numbers into bins by looking the first m bits).

We can sort each bucket and concatenate the numbers to get a sorted sequence.

Let's say we just use a quadratic time algorithm to sort the number in each bucket.

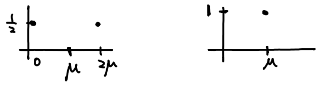
Then the running time is $\sum_i E[X_i^2]$, where X_i is the number of balls in bin i.

Note that $E[X_i] = np = 1$ (where $p = 1/n$) and $\mathrm{Var}[X_i] = np(1-p) = 1 - \frac{1}{n}$.

Then $E[X_i^2] = \mathrm{Var}[X_i] + (E[X_i])^2 = (1 - \frac{1}{n}) + 1 < 2$.

Therefore, $\sum_i E[X_i^2] \le 2n$, the expected running time is at most $2n$ steps.

Chebyshev inequality [MU 3.3]

We would like to distinguish distributions that are concentrated around its expectation and those that are not, e.g. 

.

One possible test is to compute $E[X^2]$ and see how far it is from $(E[X])^2$.

The Chebyshev inequality provides such a bound.

Theorem For any $a > 0$, $\Pr(|X - E[X]| \ge a) \le \mathrm{Var}[X]/a^2$. by Markov

Theorem For any $a>0$, $\Pr(|X-E[X]| \geq a) \leq \mathrm{Var}[X]/a^2$.

Proof $\Pr(|X-E[X]| \geq a) = \Pr((X-E[X])^2 \geq a^2) \overset{\text{by Markov}}{\leq} E[(X-E[X])^2]/a^2 = \mathrm{Var}[X]/a^2$. □

Corollary For any $t>1$, $\Pr(|X-E[X]| \geq t\cdot\sigma[X]) \leq 1/t^2$, and

$\Pr(|X-E[X]| \geq t\cdot E[X]) \leq \mathrm{Var}[X]/(t^2(E[X])^2)$.

Let's do some examples of applying Chebyshev's inequality.

Coin Flips Let X be the number of heads in n fair coin flips. Again we try to bound $\Pr(X \geq 3n/4)$.

Let X_i be the indicator variable that the i-th coin is head. Then $E[X] = E[\sum_i X_i] = \sum_i E[X_i] = n/2$.

To apply Chebyshev's inequality, we want to compute $\mathrm{Var}[X] = \mathrm{Var}[\sum_i X_i] = \sum_i \mathrm{Var}[X_i]$ by independence.

Now, $\mathrm{Var}[X_i] = E[(X_i - E[X_i])^2] = \frac{1}{2}(1-\frac{1}{2})^2 + \frac{1}{2}(0-\frac{1}{2})^2 = \frac{1}{4}$. (in general $\mathrm{Var}[X_i] = p(1-p)$)

So, $\Pr(X \geq 3n/4) \leq \Pr(|X-E[X]| \geq n/4) \leq (n/4)/(n/4)^2 = 4/n$.

Coupon Collector Revisited

Again, let X be the number of balls thrown until no empty bins, and let X_i be the number of balls thrown when there are exactly i empty bins. Then $X = \sum_{i=1}^{n} X_i$, and X_i is a geometric random variable.

To apply Chebyshev's inequality, we'd like to compute $\mathrm{Var}[X] = \mathrm{Var}[\sum_i X_i] = \sum_i \mathrm{Var}[X_i]$ by independence.

What is $\mathrm{Var}[Y]$ where Y is a geometric random variable with parameter p?

$E[Y^2] = \Pr(\text{first trial success}) \cdot E[Y^2 \mid \text{first trial success}] + \Pr(\text{first trail fail}) \cdot E[Y^2 \mid \text{first trial fail}]$

$= p \cdot 1 + (1-p)\, E[(Y+1)^2]$ (by the memoryless property, see MU Lemma 2.8)

$= (1-p)\cdot E[Y^2] + 2(1-p)E[Y] + (1-p) + p = (1-p)E[Y^2] + (2-p)/p$.

This implies that $E[Y^2] = (2-p)/p^2$ and thus $\mathrm{Var}[Y] = (1-p)/p^2 \leq 1/p^2$. (see MU 3.3.1 for a direct calculation)

Hence, $\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[X_i] \leq \sum_{i=1}^{n} (1/(i/n)^2) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} = O(n^2)$.

(Actually, $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi}{6}$. See wiki: Basel problem)

By Chebyshev's inequality, $\Pr(X \geq 2n\ln n) \leq \Pr(|X - n\ln n| \geq n\ln n) \leq O(n^2)/(n\ln n)^2 = O(1/(\ln n)^2)$.

This is much stronger than Markov's inequality, which would just give $\Pr(X \geq 2n\ln n) \leq 1/2$.

Remarks:

① Recall that $E[\#\text{ of empty bins}] \approx n e^{-\frac{m}{n}}$. Let $m = n\ln n + cn$.

Then this expectation is at most $n e^{-\ln n - c} = e^{-c}$. When $c = \ln n$, this is at most $1/n$.

By Markov's inequality, $\Pr(\#\text{ of empty bins} \geq 1) \leq E[\#\text{ empty bins}] = 1/n$.

This shows that the above bound by Chebyshev's inequality is not so tight (why?).

② The Chebyshev's inequality is most useful when we only have the second moment or the second moment is easier to compute and is enough, e.g. second moment method, data streaming, etc.

Finding Median (MU 3.4) : a more advanced example using Chebyshev's inequality.

Problem: Find the median of n numbers in linear time. Let S be the input set.

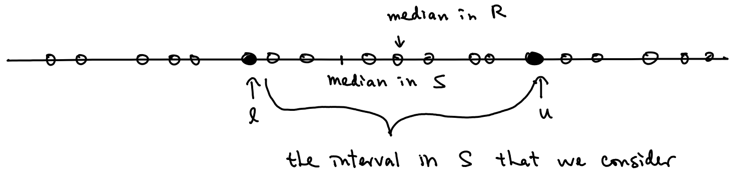
Idea: ① Sample a subset R of S of sublinear size.

② The median of R should be close to the median of S.

③ To be safe, take the $(\frac{|R|}{2} - K)$-th and $(\frac{|R|}{2} + K)$-th numbers of R be a lower bound and an upper bound of the median, called them l and u.

④ Only consider those numbers between l and u in S, and locate the median.

The number K should be chosen that the median is between l and u, but not so many other numbers are between l and u.



circles represent the numbers in R.

Algorithm ① Pick a set R of $n^{3/4}$ numbers, each independently and uniformly from the input S. Sort R.

② Let l be the $(n^{3/4}/2 - \sqrt{n})$-th number in R, and u be the $(n^{3/4}/2 + \sqrt{n})$-th number in R.

③ Consider $C = \{x \in S \mid l \leq x \leq u\}$. Get the positions of l and u in S, denoted by $pos(l)$, $pos(u)$.

④ If $pos(l) > n/2$ or $pos(u) < n/2$, FAIL (because the median is not in between)

If $|C| > 4n^{3/4}$, FAIL (because there are too many numbers in between)

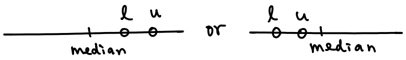
⑤ Sort the numbers in C, and determine the median in S.

Running time: Only step ③ is linear time. All other steps are sublinear time.

Analysis: So we just need to bound the failure probability of the algorithm.

There are two bad events.

Bad event $\mathcal{E}_1$:



Just consider the first subcase; the other subcase is symmetric.

This bad event occurs because we chose too many numbers greater than the median in R. More precisely, we chose at least $n^{3/4}/2 + \sqrt{n}$ numbers greater than the median in R.

What is the expected number of numbers greater than the median? Each random number is greater than the median with probability at most $1/2$. So, this expected number is $n^{3/4}/2$.

Let Y be the number of numbers in R greater than the median, and Y_i be the indicator variable the i-th sample is greater than the median.

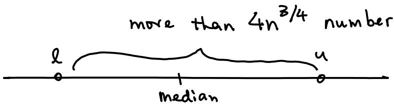
So, to bound this failure probability, we'd like to bound $\Pr[|Y-E[Y]| > \sqrt{n}]$.

We will use Chebyshev's inequality, and to do we will compute $Var[Y] = Var[\sum_i Y_i] = \sum_i Var[Y_i]$.

Note that $Var[Y_i] = E[(Y_i - E[Y_i])^2] = \frac{1}{2}(1-\frac{1}{2})^2 + \frac{1}{2}(0-\frac{1}{2})^2 = \frac{1}{4}$, and so $Var[Y] = n^{3/4}/4$.

By Chebyshev's inequality, $\Pr[|Y-E[Y]| > \sqrt{n}] \leq Var[Y]/(\sqrt{n})^2 = 1/(4n^{1/4})$, a very small number.

Bad event $\mathcal{E}_2$:



It means that there are at most $2\sqrt{n}$ samples in these $>4n^{3/4}$ numbers, because there are at most $2\sqrt{n}$ samples between l and u.

Let p be the probability that a random sample is between l and u, then $p > (4n^{3/4})/n = 4/n^{1/4}$.

Let Z be the number of samples between l and u, and Z_i be the indicator variable that the i-th sample is between l and u.

Then $E[Z] = E[\sum_i Z_i] = \sum_i E[Z_i] = n^{3/4}p$.

And $Var[Z] = Var[\sum_i Z_i] = \sum_i Var[Z_i] = n^{3/4}(p(1-p)^2 + (1-p)(0-p)^2) = n^{3/4}p(1-p) < n^{3/4}p$

By Chebyshev's inequality, $\Pr(\mathcal{E}_2) \leq \Pr[|Z-E[Z]| \geq \frac{1}{2}E[Z]] \leq Var[Z]/(\frac{1}{2}E[Z])^2$

$$= n^{3/4}p/(\tfrac{1}{4}n^{6/4}p^2)$$

$$= 4/(n^{\frac{3}{4}}p) \leq 1/\sqrt{n}.$$

Therefore, the total failure probability is $O(n^{-1/4})$, very small.

Heuristic Arguments

We showed that the maximum load is $O(\ln n/\ln\ln n)$ with high probability. Is it tight?

We showed that the probability of having an empty bin after $n\ln n + cn$ balls is at most e^{-c}. Is it tight?

Here we will use a heuristic argument to get some intuition.

Maximum Load (MU 5.4 & 5.3)

Let p_r be the probability that a bin has r balls.

Then $p_r = \binom{m}{r}\left(\frac{1}{n}\right)^r\left(1-\frac{1}{n}\right)^{m-r} = \frac{1}{r!}\,\frac{m(m-1)\cdots(m-r+1)}{n^r}\left(1-\frac{1}{n}\right)^{m-r}$

Assuming $m = n >> r$. Then $p_r = \frac{1}{r!}\,\frac{n(n-1)\cdots(n-r+1)}{n^r}\left(1-\frac{1}{n}\right)^{n-r} \approx \frac{1}{r!}\cdot 1 \cdot e^{-1} = \frac{1}{er!}$.

We further assume that all bins are independent (while not true, intuitively not too far off).

Then no bin has exactly r balls is at most $\left(1-\frac{1}{er!}\right)^n \le e^{-n/(er!)}$.

If this probability is very small, i.e. $e^{-n/(er!)} \le n^{-2}$, then with high probability there will be some bin with at least r balls.

For $e^{-n/(er!)} \le n^{-2}$ to hold, it suffices to set $-n/(er!) \le -2\ln n \Leftrightarrow r! \le n/(2e\ln n)$

$$\Leftrightarrow \ln r! \le \ln n - \ln\ln n - \ln(2e). \quad (*)$$

By Stirling's approximation that $r! \le e\sqrt{r}\left(\frac{r}{e}\right)^r \le r\left(\frac{r}{e}\right)^r$, (see MU Lemma 5.8 for the first inequality)

$\ln(r!) \le r\ln r - r + \ln r$

Set $r = \ln n / \ln\ln n$.

Then $\ln r! \le \frac{\ln n}{\ln\ln n}(\ln\ln n - \ln\ln\ln n) - \frac{\ln n}{\ln\ln n} + (\ln\ln n - \ln\ln\ln n)$

$\le \ln n - \ln n/\ln\ln n$ (since the sum of the remaining three terms is less than zero)

$\le \ln n - \ln\ln n - \ln(2e)$.

This shows that $(*)$ holds when $r = \ln n/\ln\ln n$, and therefore there exists some bin with load $\Omega(\ln n/\ln\ln n)$.

Coupon Collector (MU 5.4.1)

To estimate the probability that some bin is empty after $n\ln n + cn$ balls, again we use

$$p_r = \frac{1}{r!}\,\frac{m(m-1)\cdots(m-r+1)}{n^r}\left(1-\frac{1}{n}\right)^{m-r} \approx \frac{1}{r!}\left(\frac{m}{n}\right)^r e^{-m/n}.$$

For $m = n\ln n + cn$, $p_0 \approx e^{-c}/n$.

So, the probability of having some empty bin is $\approx 1 - \left(1 - e^{-c}/n\right)^n \approx 1 - e^{-e^{-c}} = 1 - \frac{1}{e^{\frac{1}{e^c}}}$.

When c is a large positive constant, this is very close to zero.

When c is a large negative constant, this is very close to one.

This is a "sharp" threshold phenomenon, for which we expect the event happens when there are very close to $n\ln n$ balls.

Poisson Approximation (Optional) [MU 5.4] (see also 2011 L2)

Why can we assume independence in previous arguments?

No, we can not, but we can make it precise by using the Poisson approximation technique.

Recall that $p_r = \binom{m}{r} \left(\frac{1}{n}\right)^r \left(1-\frac{1}{n}\right)^r \approx e^{-m/n} (m/n)^r / r!$.

Think of m/n as the mean.

Define a Poisson random variable with parameter μ by the probability distribution $Pr(X=j) = e^{-\mu} \mu^j / j!$.

Then p_r is just a Poisson random variable with $\mu = m/n$.

Note that it is a probability distribution, with expected value μ, and it is a good approximation of binomial random variables (MU Thm 5.5).

Let $X_i^{(m)}$ be the number of balls in bin i when m balls are thrown, and $Y_i^{(m)}$ be a Poisson random variable with mean m/n.

A main difference between the distributions $(X_1^{(m)}, X_2^{(m)}, \ldots, X_n^{(m)})$ and $(Y_1^{(m)}, Y_2^{(m)}, \ldots, Y_n^{(m)})$ is that $\sum_i Y_i^{(m)}$ may not be equal to m.

There are two key points in the proof:

① Conditioned on $\sum_i Y_i^{(m)} = m$. Then the two distributions are the same.

② $\sum_i Y_i^{(m)} = m$ happens with reasonable probability.

Combining these two points, if we can give a good upper bound in the Poisson distribution, we can give a (just slightly bigger) upper bound on the original distribution.

For the maximum load problem and the coupon collector problem, we can give a very small upper bound on the bad events in the Poisson distribution, the heuristic arguments can be made precise.

It is usually not easy to deal with dependent random variables.

Power of Two Choices (Optional) [MU 14.1]

Now we know that when n balls are thrown into n bins. Then the maximum load is $\Theta(\ln n / \ln\ln n)$ w.h.p.

Consider the following variant when each ball is thrown we pick two random bins and put the ball in the bin with fewer balls.

Surprisingly this simple modification can significantly reduce the maximum load to $O(\ln\ln n)$!

The intuition is simple. A ball is of height i if it is the i-th ball put in the bin. Suppose we can

The intuition is simple. A ball is of height i if it is the i-th ball put in the bin. Suppose we can bound the number of bins with at least i balls by β_i, over the entire course of the process. What should be β_{i+1}? A ball is of height $i+1$ if the two random bins both have at least i balls. This happens with probability at most $(\beta_i/n)^2$. Hence $\frac{\beta_{i+1}}{n} \le \left(\frac{\beta_i}{n}\right)^2$. Solving the recurrence gives that β_j becomes $O(\ln n)$ when $j = O(\ln\ln n)$. At this point the number is too small to apply concentration inequalities for the induction, but it is easy to finish the proof from there.

We will use the following Chernoff bound (to be proved next week)

$\Pr(B(n,p) \ge 2np) \le e^{-np/3}$, where $B(n,p)$ is the binomial random variable with n trials and success prob. p.

Let $\beta_4 = n/4$ and $\beta_{i+1} = 2\beta_i^2/n$.

Let $\mathcal{E}_i$ be the event that after all n balls are thrown the number of bins with at least i balls is $\le \beta_i$.

Note that β_4 holds with probability 1.

We will prove that if $\mathcal{E}_i$ holds then $\mathcal{E}_{i+1}$ holds with high probability (until β_i becomes too small).

In the following we condition on the event $\mathcal{E}_i$.

Let $Y_t = 1$ if the t-th ball has height at least $i+1$.

Then $\Pr(Y_t = 1) \le \beta_i^2/n$.

Let $p_i = \beta_i^2/n$. Then $\Pr\left(\sum_{t=1}^{n} Y_t > k\right) \le \Pr(B(n,p_i) > k)$.

So $\Pr(\#\text{ bins with at least } i+1 \text{ balls} > k) \le \Pr(\#\text{ balls with height at least } i+1 > k \mid \mathcal{E}_i)$

$$= \Pr\left(\sum_{t=1}^{n} Y_t > k \mid \mathcal{E}_i\right)$$

$$\le \Pr(B(n,p_i) > k)$$

Set $k = \beta_{i+1} = 2np_i$, then the above probability $\le \dfrac{\Pr(B(n,p_i) > 2np_i)}{\Pr(\mathcal{E}_i)} \le \dfrac{1}{\Pr(\mathcal{E}_i)\cdot e^{p_i\cdot n/3}}$

This implies that $\Pr(\neg\mathcal{E}_{i+1} \mid \mathcal{E}_i) \le \dfrac{1}{n^2 \Pr(\mathcal{E}_i)}$ as long as $p_i \cdot n \ge 6\ln n$.

So $\Pr(\neg\mathcal{E}_{i+1}) = \Pr(\neg\mathcal{E}_{i+1} \mid \mathcal{E}_i)\Pr(\mathcal{E}_i) + \Pr(\neg\mathcal{E}_{i+1} \mid \neg\mathcal{E}_i)\Pr(\neg\mathcal{E}_i)$

$\le \Pr(\neg\mathcal{E}_{i+1} \mid \mathcal{E}_i)\Pr(\mathcal{E}_i) + \Pr(\neg\mathcal{E}_i).$ $\le \frac{1}{n^2} + \Pr(\neg\mathcal{E}_i)$ as long as $p_i \cdot n \ge 6\ln n$.

To finish the proof we need two more steps. First is to prove that $p_i \cdot n < 6\ln n$ in $O(\ln\ln n)$ steps. And second is to handle the case when $p_i \cdot n < 6\ln n$.

The first step is easy. A simple induction can show that $\beta_{i+4} \le n/2^{2^i}$ and therefore

The first step is easy. A simple induction can show that $\beta_{i+1} \leq n/2^{2^i}$ and therefore $p_i \cdot n < 6 \ln n$ in $O(\ln \ln n)$ steps. And thus $\Pr(\neg \mathcal{E}_i) \leq O(\ln \ln n)/n^2$ in this step.

The second step is also easy. By Chernoff bound we can show that whp there are at most $O(\ln n)$ bins with at least $\Omega(\ln \ln n)$ balls at this stage. Then those bins are just too few that we can finish the argument by naive bound ($\left(\frac{\ln n}{n}\right)^2$ to reach one step higher and there are at most $\binom{n}{2}$ ways of choosing two balls) to show that there are at most 2 bins with one more ball. And then no more bin with one more ball.

This concludes the proof (sketch).

<u>**Remarks :**</u>

① The bound is tight. One can prove that there exists some bin with $\Omega(\ln \ln n)$ balls with high prob.

② One can hope to use <u>two</u> hash functions to improve the worst case search time to $O(\ln \ln n)$.

③ What about $d \geq 3$ choices?