Concentration of a random variable around its mean

Curs 2018

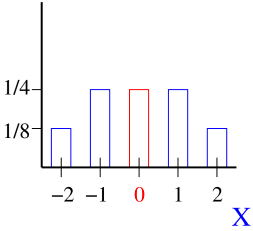
Expectation does not suffice

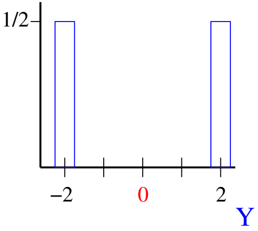
The expected value of a random variable is a nice single number to *tag* the random variable, but it leaves out most of the important properties of the r.v.

Consider r.v X with $X(\Omega) = \{-2, -1, 0, 1, 2\}$ and PMF:
$p(-2) = \frac{1}{8}, p(-1) = \frac{1}{4}, p(0) = \frac{1}{4}, p(1) = \frac{1}{4}, p(2) = \frac{1}{8}$.

and consider r.v. Y with $Y(\Omega') = \{-2, 0, 2\}$ and PMF:
$p_Y(-2) = \frac{1}{2}, p_Y(2) = \frac{1}{2}$.

Note that $\mathbf{E}[X] = 0 = \mathbf{E}[Y]$, but p_X is totally different from p_Y.





Deviation of a r.v. from its mean

- Consider the deterministic Quicksort algorithm on n-size inputs. Let $T(n)$ be a r.v. counting the number of steps of Quicksort on a specific input with size n
- Its worst case complexity is $O(n^2)$, but its average complexity is $O(n \lg n)$.
- It does not give information about the behavior of the algorithm on a particular input.
- Given an algorithm, for any input x of size $|x| = n$, how close is $T(x)$ to $\mathbf{E}[T(n)]$.

Deviation of a r.v. and Concentration

- For ex.: If $\mathbf{E}[T(n)] = 10$, then 10 is an average running time on "most inputs" to the algorithm. We want to assure, that for most inputs, their value $\mathbf{E}[T(n)]$ is concentrated around 10.
- Once we get $\mathbf{E}[T(n)]$, we like to make sure that the probability of having instances for which $|\mathbf{E}[T(n)] - T(n)|$ is large, is going to be very small.
- Intuitively it seems clear from the definition of $\mathbf{E}[]$, if for the above running time, we get an instance e for which $T(e) = 10^9$, and $\mathbf{E}[T(n)] = 10$, the probability of selecting that specific e is going to be quite small, as its contribution to the average 10 is $10^9 \mathbf{Pr}\left[T(n) = 10^9\right]$.

Markov's inequality

Lemma If $X \geq 0$ is a r.v, for any constant $a > 0$,

$$\mathbf{Pr}\left[X \geq a\right] \leq \frac{\mathbf{E}\left[X\right]}{a}.$$

Proof Given the r.v. $X \geq 0$ define the indicator r.v.

$$Y = \begin{cases} 1 & \text{if } X \geq a \text{ true} \\ 0 & \text{otherwise} \end{cases}$$

Notice $Y \leq X/a$, so $\mathbf{E}\left[Y\right] = \mathbf{Pr}\left[Y = 1\right] = \mathbf{Pr}\left[X \geq a\right]$ and $\mathbf{E}\left[Y\right] = \mathbf{Pr}\left[Y = 1\right] \leq \mathbf{E}\left[\frac{X}{a}\right] = \frac{\mathbf{E}[X]}{a}$. □

Alternative expression for Markov: Taking $b = a/\mathbf{E}\left[X\right]$

Corollary If $X \geq 0$ is a r.v, for any constant $b > 0$,

$$\mathbf{Pr}\left[X \geq b\mathbf{E}\left[X\right]\right] \leq \frac{1}{b}.$$

Markov's inequality: Proving Union-Bound

Assume we have events $\{A_i\}$ and let $\mathbf{Pr}\left[A_i\right]$ be the probability that A_i occurs.

Define X as the r.v. counting how many events occur.
Note $\mathbf{Pr}\left[X \geq 1\right] = \mathbf{Pr}\left[\cup_{i=1}^{n} A_i\right]$

To compute $\mathbf{E}\left[X\right]$ define the i.r.v. $X_i = 1$ if A_i occurs, $X_i = 0$ otherwise. Note, $\mathbf{Pr}\left[X_i = 1\right] = \mathbf{Pr}\left[A_i\right]$.

As $X = \sum_{i=1}^{n} X_i \Rightarrow \mathbf{E}\left[X\right] = \sum_{i=1}^{n} \mathbf{E}\left[X_i\right] = \sum_{i=1}^{n} \mathbf{Pr}\left[A_i\right]$

Using Markov: $\mathbf{Pr}\left[\cup_{i=1}^{n} A_i\right] = \mathbf{Pr}\left[X \geq 1\right] \leq \mathbf{E}\left[X\right] = \sum_{i=1}^{n} \mathbf{Pr}\left[A_i\right]$.

Markov could be too weak

Consider the randomized hiring algorithm. We computed that the expected number of pre-selected students is $\mathbf{E}[X] = \lg n$. We also know there are instances for which $X = n$.

We would like to compute that the probability of selecting a "bad instance" is very small.

Using Markov, for any constant b, $\mathbf{Pr}[X \geq b \lg n] \leq 1/b$. (for ex. $b = 100$)

The problem with Markov is that it does not bound away the probability of *bad cases* as a function of the input size.

With High Probability

In the randomized algorithms, we aim to obtain results with high probability, the probability that the complexity of the algorithm for any input is ”near” the expected value, tends to 1 as the size n grows.

An event that occurs with high probability (whp) is one that happens with probability $\geq 1 - \frac{1}{n}$, so that as $n \to \infty$.

The parameter n is usually the size of the inputs, or the size of the combinatorial structure, ...

Variance

Given a r.v. X its variance measures the spread of its distribution.

Given X, with $\mu = \mathbf{E}[X]$, define its variance by:

$$\mathbf{Var}[X] = \mathbf{E}\left[(X-\mu)^2\right]$$

Usually it is more easy to use: $\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2$

Proof

$$\begin{aligned}\mathbf{Var}[X] &= \mathbf{E}\left[(X-\mu)^2\right] = \mathbf{E}\left[X^2 - 2\mu\mathbf{E}[X] + \mu^2\right] \\ &= \mathbf{E}\left[X^2\right] - 2\mu\underbrace{\mathbf{E}[X]}_{\mu} + \mu^2 = \mathbf{E}\left[X^2\right] - \mu^2 \qquad \square\end{aligned}$$

Further properties of the Variance

- $\mathbf{Var}\left[X\right] \geq 0$ as by Jensen's inequality, for any r.v X, $\mathbf{E}\left[X^2\right] \geq \mathbf{E}\left[X\right]^2$.
- $\mathbf{Var}\left[X\right] = 0$ iff $X =$ constant.
  Proof $(\Leftarrow)$ If $X = c$ then $\mathbf{E}\left[X\right] = c \Rightarrow \mathbf{Var}\left[X\right] = 0$.
  $(\Rightarrow)$ If $\mathbf{Var}\left[X\right] = 0 \Rightarrow \mathbf{E}\left[X^2\right] = \mathbf{E}\left[X\right]^2 \Rightarrow \mathbf{E}\left[X\right] = c$.
- $\mathbf{Var}\left[cX\right] = c^2\mathbf{Var}\left[X\right]$.
  Proof
  $\mathbf{Var}\left[cX\right] = \mathbf{E}\left[(cX)^2\right] - \mathbf{E}\left[cX\right]^2 = c^2\mathbf{E}\left[X^2\right] - (c\mathbf{E}\left[X\right])^2$

Computing $\mathbf{Var}[X]$

Given a r.v. X on Ω, such that $X(\Omega) = \{x_1, x_2, \dots, x_n\}$ with PDF $(p_X(x_1), \dots, p_X(x_1))$, we first compute $\mu = \mathbf{E}[X] = \sum_{i=1}^n x_i p_X(x_i)$. Then, use one of the following methods:

1. Use $\mathbf{Var}[X] = \mathbf{E}\left[(X-\mu)^2\right]$: For each x_i compute $(x_i - \mu)^2$, and then $\mathbf{Var}[X] = \sum_{i=1}^n (x_i - \mu)^2 p_X(x_i)$
2. Use $\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2$: For each x_i compute x_i^2, then $\mathbf{E}\left[X^2\right] = \sum_{i=1}^n x_i^2 p_i$.

EX.: Consider r.v. X with $X(\Omega) = \{1, 3, 4\}$ and PMF: $p_X(1) = \frac{1}{4}, p_X(3) = \frac{1}{4}, P_X(5) = \frac{1}{2}$. Then $\mu = 7/2$.

1. $\mathbf{Var}[X] = \frac{1}{4}(3 - \frac{7}{2})^2 + \frac{1}{4}(5 - \frac{7}{2})^2 + \frac{1}{2}(1 - \frac{7}{2})^2 = \frac{11}{4}$
2. $X^2(\Omega) = \{1, 9, 25\}$, so $\mathbf{E}\left[X^2\right] = \frac{1}{4} + \frac{9}{4} + \frac{25}{2} = 15$
   $\mathbf{Var}[X] = 15 - (\frac{7}{2})^2 = \frac{11}{4}$

Examples

Consider r.v. Y with $X(\Omega) = \{-2, 0, 2\}$ and PMF:
$p_Y(-2) = \frac{1}{2}, p_Y(2) = \frac{1}{2}$.
Therefore, the values $(X - \mu)^2$ are $(-2 - 0)^2$ and $(2 - 0)^2$
$\Rightarrow \mathbf{Var}[X] = \frac{1}{2}4 + \frac{1}{2}4 = 4$
Notice in this case $\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] = 4$

You win 100€ with probability $= 1/10$, otherwise you win 0€. Let X be a r.v. counting your earnings. What is $\mathbf{Var}[X]$?
$\mu = 100/10 = 10$. Therefore, $\mathbf{E}\left[X^2\right] = \frac{1}{10}(100^2) = 1000$, and as $\mu^2 = 100$, so $\mathbf{Var}[X] = 900$.

$\mathbf{Var}\left[\right]$ is not necessarily linear

Let $X_1, \ldots, X_n$ be independent r.v., then

$$\mathbf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{Var}\left[X_i\right].$$

We prove the particular case that if X and Y are independent $\mathbf{Var}\left[X+Y\right] = \mathbf{Var}\left[X\right] + \mathbf{Var}\left[Y\right]$

$$\begin{aligned}
\mathbf{Var}\left[X+Y\right] &= \mathbf{E}\left[(X+Y)^2\right] - \left(\mathbf{E}\left[X+Y\right]\right)^2 \\
&= \mathbf{E}\left[X^2\right] + \mathbf{E}\left[Y^2\right] + 2\mathbf{E}\left[XY\right] - \left(\mathbf{E}\left[X\right]\right)^2 - \left(\mathbf{E}\left[Y\right]\right)^2 - 2\mathbf{E}\left[X\right]\mathbf{E}\left[Y\right] \\
&= \mathbf{E}\left[X^2\right] - \left(\mathbf{E}\left[X\right]\right)^2 + \mathbf{E}\left[Y^2\right] - \left(\mathbf{E}\left[Y\right]\right)^2 + 2\underbrace{\left(\mathbf{E}\left[XY\right] - \mathbf{E}\left[X\right]\mathbf{E}\left[Y\right]\right)}_{\mathbf{E}[XY]=\mathbf{E}[X]\mathbf{E}[Y]}
\end{aligned}$$

Variance of some basic distributions

1. If $X \in B(p, n)$ then $\mathbf{Var}\left[X\right] = pqn$, where $q = (1 - p)$.
2. If $X \in P(\lambda)$ then $\mathbf{Var}\left[X\right] = \lambda$.
3. If $X \in G(p)$ then $\mathbf{Var}\left[X\right] = \frac{q}{p^2}$.

Proof

(1.-) Let $X = \sum_{i=1}^{n} X_i$, where X_i is an indicator r.v s.t. $X_i = 1$ with probability p

Then, $\mathbf{Var}\left[X_i\right] = \mathbf{E}\left[X_i^2\right] - \mathbf{E}\left[X\right]^2 = (p \cdot 1^2 + q \cdot 0 - p^2 = p(1 - p)$, as all X_i are independent, $\mathbf{Var}\left[X\right] = \sum_{i=1}^{n} \mathbf{Var}\left[X_i\right] = np(1 - p)$.

A more natural measure of spread: Standard Deviation

Why we did not define $\mathbf{Var}[X] = \mathbf{E}[|X-\mu|]$?
To be sure we are averaging only non-negative values.

But as we defined the variance, we are using squared units!

Recall the example X a r.v. counting the wins, when you win 100€ with probability $= 1/10$, otherwise you win 0€. We got $\mathbf{Var}[X] = 900€^2$.
To convert the numbers back to re-scale, we take the square root.

The Standard Deviation of a r.v. X is defined as

$$\sigma[X] = \sqrt{\mathbf{Var}[X]}.$$

In the previous example, to convert the spread from €^{2} to €,
$\sigma[X] = \sqrt{900} = 30$ €.

Chebyshev's Inequality

Pafnuty Chebyshev (XIXc)
If you can compute the **Var** [] then you can compute σ and get better bounds for concentration of any r.v. (positive or negative).

Theorem Let X be a r.v. with μ and $\sigma > 0$, then for any $a > 0$

$$\mathbf{Pr}\left[|X-\mu| \geq a\sigma\right] \leq \frac{1}{a^2}.$$

Note that $|X-\mu| \geq a\sigma \Leftrightarrow (X \geq a\sigma + \mu) \cup (X \geq \mu - a\sigma)$.

Proof As the r.v. $|X-\mu| \geq 0$, we can apply Markov to it:

$$\begin{aligned}\mathbf{Pr}\left[|X-\mu| \geq a\sigma\right] &= \mathbf{Pr}\left[(X-\mu)^2 \geq a^2\sigma^2\right] \qquad \text{(by Markov)}\\ &\leq \frac{\mathbf{E}\left[(X-\mu)^2\right]}{a^2\sigma^2} = \frac{\mathbf{Var}\left[X\right]}{a^2\mathbf{Var}\left[X\right]} = \frac{1}{a^2} \qquad \square\end{aligned}$$

More on Chebyshev's Inequality

We had: $\mathbf{Pr}\left[|X-\mu| \geq a\sigma\right] \leq \frac{1}{a^2}$.

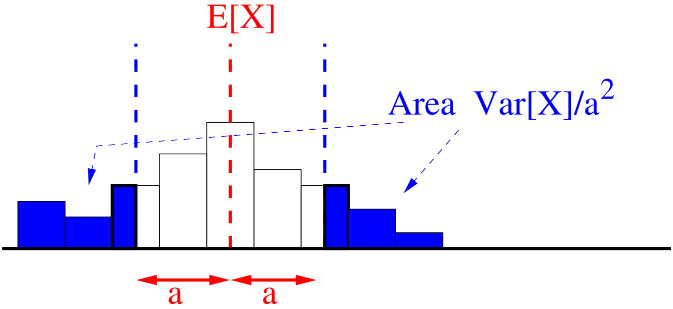
Alternative equivalent statement:

$$\forall b > 0, \mathbf{Pr}\left[|X-\mu| \geq b\right] \leq \frac{\mathbf{Var}\left[X\right]}{b^2}.$$

Proof As before: $\mathbf{Pr}\left[(X-\mu)^2 \geq b^2\right] \leq \frac{\mathbf{E}\left[(X-\mu)^2\right]}{b^2}$.

Chebyshev's Inequality: Picture

$$\mathbf{Pr}\left[|X-\mu| \geq a\right] \leq \frac{\mathbf{Var}\left[X\right]}{a^2}.$$



An easy application

Let flip n-times a fair coin, give an upper bound on the probability of having at least $\frac{3n}{4}$ heads.

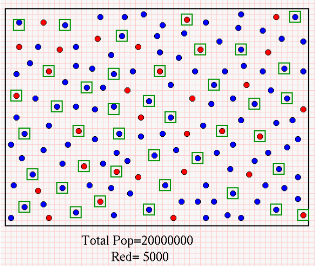
Let $X \in B(n, 1/2)$, then, $\mu = n/2, \mathbf{Var}[X] = n/4$.

We want to bound $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right]$.

- Markov: $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq \frac{\mu}{3n/4} = 2/3$.
- Chebyshev's: We need the value of a s.t. $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq \mathbf{Pr}\left[|X - \frac{n}{2}| \geq a\right] \Rightarrow a = \frac{3n}{4} - \frac{n}{2} = \frac{n}{4}$.

  $\mathbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq \mathbf{Pr}\left[|X - \frac{n}{2}| \geq \frac{n}{4}\right] \leq \frac{\mathbf{Var}[X]}{(n/4)^2} = \frac{4}{n}$.

Sampling



- Given a large population Σ, $|\Sigma| = n$, we wish to estimate the proportion p of elements in Σ, with a given property.
- Sampling: Take a random sample S with size $m << n$ and observe p^{-} in S.
- Sometimes, if n is large, the obvious estimator $m \times p^{-}$ is sufficiently good, i.e. it is sharply concentrated.
- Many times getting the random sample S is non-trivial.

Streaming

Stream: Get m elements $S = \{x_1, \dots x_m\}$, from a universe $S = \{a, b, \cdots\}$ each element x_t at time t, for $1 \leq t \leq m$. It may happen that two or more elements in S are the same repeated element from Σ.

Given a population Σ, and a stream $S = \{x_1, \dots x_m\}$, for any $a \in \Sigma$ the frequency f_a as the number of times a appear in S.

Some common problems in streaming:

- How would you get an unbiased random sample from a stream of incoming data, when we don't know the size of the population, until we have seen the last element of the stream. (This is the reservoir sampling problem)
- Finding the elements with larger frequency. Those are called the heavy hitters.
- Find the number of distinct elements from Σ in S.

Finding the median of n elements

- Recall that, given a set S with n distinct elements, the median of S is the $\lceil n/2 \rceil$ larger element in S.
- We can use Quickselect to find the median with expected time $O(n)$. Even there is a linear time deterministic algorithm, which in practice for large instance works worst than Quick-select.
- We present another randomized algorithm to find the median m in S, which is based in sampling.
- The purpose of this algorithm is to introduce the technique of filtering large data by sampling small amount of the data.

Finding the median of n elements: A Filtering Data algorithm

INPUT: An unordered set $S = \{x_1, x_2, \dots x_n\}$, with $n = 2k + 1$ elements.
OUTPUT: The median, which is the $k + 1$ largest element in S.
For any element y define the $\text{rank}(y) = |\{x \in S | x \leq y\}|$.

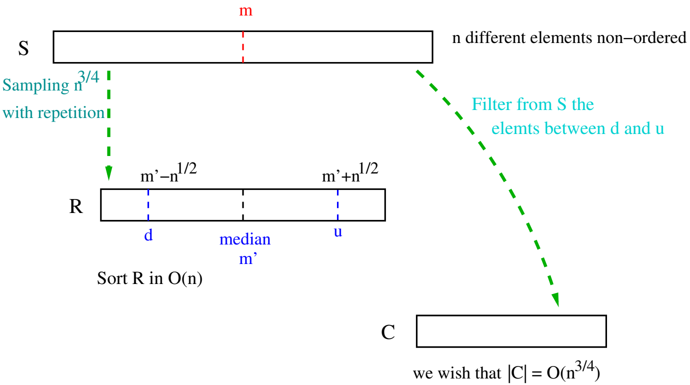
The idea of the filtering Algorithm is to sample with replacement a "small" subset of C elements from S, so we can order C in $O(n)$ time (linear with respect to the size of S).
Then the algorithm find the median of the elements in C and either return it as the median in S or return failure

We will prove that whp the algorithm finds the median of S, in linear time.

Outline of the algorithm

1. Let $\tilde{S}$ be the ordered set S (we do not know $\tilde{S}$).
2. Find elements $d, u \in S$ s.t. $d < m < u$ and distance between d and u in $\tilde{S}$ is $< n/\lg n$.
3. To find d and u sample with replacement S to get a multi-set R, with $|R| = O(\lceil n^{3/4} \rceil)$. Notice $\lceil n^{3/4} \rceil < n/\lg n$. Find $u, d \in R$ s.t. m will be close to median in R).
4. Filter-out the elements $x \in S$, which are $< d$ or $> u$ to form a set $C = \{x \in S | d \leq x \leq u\}$.
5. Sort elements in C in $O(n)$, and find m.
6. Prove that w.h.p. the algorithm succeeds.

Outline of the algorithm



Things that can be wrong:

C too large,

$m \notin C$,

$m \in C$ but no the median in C.

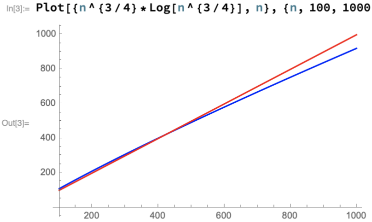
Randomized Median algorithm

1. Sample $\lceil n^{3/4} \rceil$ elements from S, u.a.r., independently, and with replacement.
2. Sort R in $O(n)$
3. Set $d = \lfloor (\frac{n^{3/4}}{2} - \sqrt{n}) \rfloor$-smallest element in R
4. Set $d = \lfloor (\frac{n^{3/4}}{2} + \sqrt{n}) \rfloor$-greatest element in R
5. Compute $C = \{x \in S | d \leq x \leq u\}$, $l_d = |\{x \in S | x < d\}|$ and $l_u = |\{x \in S | x > u\}|$ (cost $= \Theta(n)$).
6. If $l_d > \frac{n}{2}$ or $l_d > \frac{n}{2}$ OUTPUT FAIL ($m \notin C$)
7. If $|C| \leq 4n^{3/4}$ sort C, otherwise OUTPUT FAIL.
8. OUTPUT the $(\lfloor \frac{n}{2} \rfloor - l_d + 1)$-smallest element in sorted C, that should be m.

Complexity and correctness of the Randomized Median algorithm

Theorem: The Randomized Median algorithm terminates in $O(n)$ steps. If the algorithm does not output FAIL, then it outputs the median m of S.

Proof: As asymptotically $n^{3/4}\lg(n^{3/4}) \leq n$, using Mergesort on R takes $O(\frac{n}{\lg n}\lg(\frac{n}{\lg n})) = O(n)$.
The only incorrect answer is that it outputs an item, but $m \notin C$, but if so, it would fail in step 6, as either $l_d > n/2$ or $l_u > n/2$. □



```
In[3]:= Plot[{n^{3/4}*Log[n^{3/4}], n}, {n, 100, 1000}, PlotStyle → {Blue, Red}]
Out[3]=
```

?

Figure: $n^{3/4}\lg(n^{3/4})$ versus n

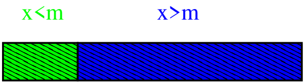
Bounding the probability of output FAIL

Theorem: The Randomized Median algorithm finds m with probability $\geq 1 - \frac{1}{n^{1/4}}$, i.e. w.h.p.
Proof (Highlights): Consider the following 3 events:
E_1: $d > m$,
E_2: $u < m$,
E_3: $|C| > 4n^{3/4}$.
Then, the algorithm outputs FAIL iff one of the three events occurs, i.e.

$$\mathbf{Pr}\left[\text{FAILS}\right] = \mathbf{Pr}\left[E_1 \cup E_2 \cup E_3\right] \leq \mathbf{Pr}\left[E_1\right] + \mathbf{Pr}\left[E_2\right] + \mathbf{Pr}\left[E_3\right]$$

Bounding $\mathbf{Pr}\left[E_1\right]$

Consider R ordered, where R is obtained by sampling $n^{3/4}$ elements from S



- $d > m$, when the green block has size $< \lfloor n^{3/4}/2 - \sqrt{n} \rfloor$.
- Let $Y = |\{x \in R \,|\, x \leq m\}|$, then $\mathbf{Pr}\left[E_1\right] = \mathbf{Pr}\left[Y < n^{3/4}/2 - \sqrt{n}\right]$.
- For $1 \leq j \leq n^{3/4}$, define the i.r.v. $Y_j = 1$ iff integer in j-th. position in R is $\leq m$.
- Then $Y = \sum_{j=1}^{n^{3/4}} Y_j$, moreover as the sampling is with replacement, then each Y_j is independent.

As $m =$ median of S ($|S| = n$), then we have $\frac{(n-1)}{2} + 1$ elements in S that are $\leq m$.

Bounding $\mathbf{Pr}\left[E_1\right]$ and $\mathbf{Pr}\left[E_2\right]$

- $\mathbf{Pr}\left[Y_j = 1\right] = \frac{(n-2)/2+1}{n} = \frac{1}{2} + \frac{1}{2n}$,
- so $Y \in B(n^{3/4}, \frac{1}{2} + \frac{1}{2n})$.
- Then $\mathbf{E}\left[Y_i\right] \geq 1/2 \Rightarrow \mathbf{E}\left[Y\right] \geq \frac{n^{3/4}}{2}$,
- xs $\mathbf{Var}\left[Y\right] = n^{3/4}(\frac{1}{2} + \frac{1}{2n})(\frac{1}{2} - \frac{1}{2n}) \leq \frac{n^{3/4}}{4}$.

Using Chebyshev:

$$\begin{aligned}\mathbf{Pr}\left[E_1\right] &= \mathbf{Pr}\left[Y < \frac{n^{3/4}}{4} - \sqrt{n}\right] \\ &\leq \mathbf{Pr}\left[\left|Y - \mathbf{E}\left[Y\right]\right| \geq \sqrt{n}\right] \leq \frac{\mathbf{Var}\left[Y\right]}{(\sqrt{n})^2} = \frac{1}{4n^{1/4}} \quad \square\end{aligned}$$

In the same way we can compute $\mathbf{Pr}\left[E_2\right] \leq \frac{1}{4n^{1/4}}$

Bounding $\mathbf{Pr}[E_3]$

Recall C is obtained directly from S by filtering, using the values d and u obtained in R.

For C to have $> 4n^{3/4}$ keys either of the following events must happen:

1. A: At least $> 2n^{3/4}$ items in C are $> m$.
2. B: At least $> 2n^{3/4}$ items in C are $< m$.

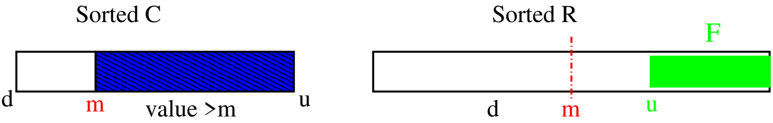
Then,

$$\mathbf{Pr}[E_3] \leq \mathbf{Pr}[A \cup B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B] .$$

Bounding $\mathbf{Pr}[A]$

Event A happens when there are $2n^{3/4}$ element in C, which are $> m$
$\Rightarrow$ the rank(u) in $tildeS$ is $\geq n/2 + 2n^{3/4}$

Then in R, any element in F has rank $\geq n/2 + 2n^{3/4}$



We will prove that $\mathbf{Pr}\left[\bar{A}\right] = 1 - O(1/n) \to 1$.

Bounding $\mathbf{Pr}[A]$

- Let $X = \#$ selected items in R that are in F (have rank $\geq n/2 + 2n^{3/4}$)
- Then $\mathbf{Pr}[A] \leq \mathbf{Pr}\left[X \geq \lfloor n^{3/2}/2 - \sqrt{n} \rfloor\right]$.
- For $1 \leq j \leq n^{3/4}$, define the i.r.v.$X_j = 1$ iff the jth item in R also is in F.
- Note $X = \sum_{j=1}^{n^{3/4}} X_j$ and $\mathbf{Pr}[X_j = 1] = \frac{1}{2} - \frac{2}{n^{1/4}} + \frac{1}{n}$.
- So $\mathbf{E}[X] = \frac{n^{3/4}}{2} - 2n^{1/2} + n^{1/4}$ and $\mathbf{Var}[X] \leq n^{3/4}/4$

$$\begin{aligned}
\mathbf{Pr}[A] &\leq \mathbf{Pr}\left[X \geq \lfloor \frac{n^{3/2}}{2} - n^{1/2} \rfloor\right] \leq \mathbf{Pr}\left[X \geq \frac{n^{3/4}}{2} - 2n^{1/2} + n^{1/4}\right] \\
&\leq \mathbf{Pr}\left[X \geq \mathbf{E}[X] + n^{1/2} - 1 - n^{1/4}\right] \\
&\leq \mathbf{Pr}\left[|X - \mathbf{E}[X]| \geq n^{1/2} - 1 - n^{1/4}\right] = O(\frac{1}{n^{1/4}}). \quad \square
\end{aligned}$$

Bounding $\mathbf{Pr}[B]$ and finishing the proof

In the same way we can compute $\mathbf{Pr}[B] = O(\frac{1}{n^{1/4}})$

To end the whole proof, we also proved that
$\mathbf{Pr}[E_3] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B] = O(\frac{1}{n^{1/4}})$

$\Rightarrow \mathbf{Pr}[\text{algorithm fails}] = \mathbf{Pr}[E_1 \cup E_2 \cup E_3] \leq^{\text{UB}} O(\frac{1}{n^{1/4}})$.

Therefore,
$\mathbf{Pr}[\text{algorithm succeeds}] = 1 - \mathbf{Pr}[\text{algorithm fails}] \geq 1 - \frac{1}{n^{1/4}}$
i.e. w.h.p. the Randomized Median algorithm finds the correct *m*
□

Why we need more concentration bounds?

- Remember that given a random variable, we are trying to determine how concentrate it is, i.e. that the probability of hitting a random instance which deviates far from the expectation μ, is small.
- We aim to have random variables (events) which are concentrated around its mean with high probability.
- We saw that if $X \geq 0$ Markov can give an indication that there are values very far away from its mean, but in general is to weak for proving strong concentration results.
- Chebyshev's inequality can give stronger results for concentration of X around μ, but we must compute $\mathbf{Var}\left[X\right]$, which could be difficult.

Chernoff Bounds

Sergei Bernstein (1924), Wassily Hoeffding (1964), Herman Chernoff (1952)

The Chernoff bound can be used when the random variable X is the sum of several independent random variables, what is known as Poisson trials, where each X_i can have a different distribution p_i. The particular case where all p_i are equal is the Bernouilli trials.

Theorem (Ch-1) Let $\{X_i\}_{i=0}^{n}$ be independent Poisson trials, with $\mathbf{Pr}\left[X_i = 1\right] = p_i$, a. Then if $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbf{E}\left[X\right]$, we have

1. $\mathbf{Pr}\left[X < (1-\delta)\mu\right] < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$, for $\delta \in (0,1)$.
2. $\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] < \left(\frac{e^{-\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$ for any $\delta > 0$.

Weaker Chernoff, but easy to use

Corollary (Ch-2) Let $\{X_i\}_{i=0}^{n}$ be independent Poisson trials, with $\mathbf{Pr}\left[X_i = 1\right] = p_i$. Then if $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbf{E}\left[X\right]$, we have

1. $\mathbf{Pr}\left[X < (1-\delta)\mu\right] \leq e^{-\mu\delta^2/2}$, for $\delta \in (0,1)$.
2. $\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] \leq e^{-\mu\delta^2/3}$, for $\delta \in (0,1]$.

An immediate corollary to the previous result:

Corollary (Ch-3) Let $\{X_i\}_{i=0}^{n}$ be independent Poisson trials, with $\mathbf{Pr}\left[X_i = 1\right] = p_i$. Then if $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbf{E}\left[X\right]$ and $\delta \in (0,1)$, we have

$$\mathbf{Pr}\left[|X-\mu| \geq \delta\mu\right] \leq 2e^{-\mu\delta^2/3}.$$

Sketch Proof of (3) using (2):

$$\begin{aligned}\mathbf{Pr}\left[|X-\mu| \geq \delta\mu\right] &= \mathbf{Pr}\left[X < (1-\delta)\mu\right] + \mathbf{Pr}\left[X \geq (1+\delta)\mu\right] \\ &\leq e^{-\mu\delta^2/2} + e^{-\mu\delta^2/3} \leq 2e^{-\mu\delta^2/3}\end{aligned}$$

Proof of Ch-2.1 using Ch-1.1

From (Ch-2.1) We must prove that for $\delta \in (0,1)$, we have $\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \leq e^{-\mu\delta^2/2}$.

Let $f(\delta) = \ln\left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} - \ln\left(e^{-\mu\delta^2/2}\right)$
$f(\delta) = -\delta - (1-\delta)\ln(1-\delta) + \delta^2/2 \leq 0.$

Differenciating 2 times $f(\delta)$: $f'(\delta) = \ln(1-\delta) + \delta$ and
$f''(\delta) = \frac{1}{1-\delta} + 1 \leq 0$

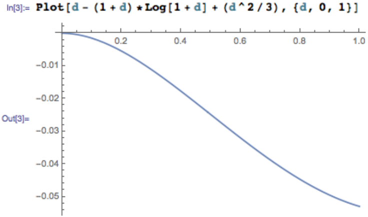
$\Rightarrow f"(\delta) < 0$ in $(0,1)$ and as $f'(0) = 0$, then $f'(\delta) \leq 0$ in $[0,1)$, i.e. $f(\delta)$ is non-increasing in $[0,1)$.

As $f(0) = 0 \Rightarrow f(\delta) \leq 0$ for $\delta \in (0,1)$.

Proof of Ch-2.2 using Ch-1.2

From (Ch-2.2) We must prove that for $\delta \in (0,1)$, we have $\left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \leq e^{-\delta^2/3}$.

Taking logs: $f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \delta^2/3 \leq 0$. Differentiating 2 times $f(\delta)$, and using the same argument as above, we see $f(\delta) \leq 0$ in $(0,1]$.



```
In[3]:= Plot[d - (1 + d) *Log[1 + d] + (d^2 / 3), {d, 0, 1}]
```

An easy application (Cont.)

Let flip n-times a fair coin, give an upper bound on the probability of having at least $\frac{3n}{4}$ heads.
Recall Let $X \in B(n, 1/2)$, then, $\mu = n/2, \textbf{Var}\left[X\right] = n/4$.

We want to bound $\textbf{Pr}\left[X \geq \frac{3n}{4}\right]$.

- Markov: $\textbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq \frac{\mu}{3n/4} = 2/3$.
- Chebyshev's: $\textbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq \textbf{Pr}\left[|X - \frac{n}{2}| \geq \frac{n}{4}\right] \leq \frac{\textbf{Var}[X]}{(n/4)^2} = \frac{4}{n}$.
- Chernoff: We want $\textbf{Pr}\left[X \geq \frac{3n}{4}\right]$. Using Ch-2.2,
  $\textbf{Pr}\left[X \geq \frac{3n}{4}\right] = \textbf{Pr}\left[X \geq (1+\delta)\frac{n}{2}\right] \Rightarrow (1+\delta)\frac{3}{2} \Rightarrow \delta = \frac{1}{2}$
  $\therefore \textbf{Pr}\left[X \geq \frac{3n}{4}\right] \leq e^{-\mu\delta^2/3} = e^{-\frac{n}{24}}$.

If $n = 100$, Cheb. $= 0.04$, Chernoff $= 0.0155$
If $n = 10^6$, Cheb. $= 4 \times 10^{-6}$, Chernoff $= 2.492 \times 10^{-18095}$

Another example

Toss n times a fair coin, what is the probability of having $m < n/2$ heads?

Let $X = \#$ heads, then $\mu = n/2$ and $\mathbf{Var}[X] = n/4$.

1. Markov: $\mathbf{Pr}[X < n/2] = 1 - \mathbf{Pr}[X \geq n/2]$, but $\mathbf{Pr}[X \geq n/2] \leq \frac{n/2}{n/2} = 1$. So $\mathbf{Pr}[X < n/2] \geq 0$. No Information
2. Chebyshev: How would you compute $\mathbf{Pr}[X < n/2]$ using $\mathbf{Pr}\left[|X - \frac{n}{2}| \geq c\right]$?
3. Chernoff: We want $\mathbf{Pr}[m < n/2]$ using $\mathbf{Pr}[X < (1-\delta)\mu]$. As $\mu = n/2$, we make $m = (1-\delta)\frac{n}{2} \Rightarrow \delta = (1 - \frac{2m}{n})$ $\therefore \mathbf{Pr}[m < n/2] = \mathbf{Pr}\left[X < (1 - (1 - \frac{2m}{n}))\frac{n}{2}\right] < e^{-\frac{n}{4}(1-\frac{2m}{n})^2}$

   If we flip the coin 100 times, then $\mathbf{Pr}[X < 10] < e^{-16} \sim 0.0000001$.

Proof of Chernoff-1: Upper tail

Note if for a r.v. X, and $a > 0$ and for any $t > 0$ we have

$$\boxed{(e^{tX} \geq e^{ta}) \Leftrightarrow (X \geq a)}$$

Therefore $\mathbf{Pr}[X \geq a] = \mathbf{Pr}\left[e^{tX} \geq e^{ta}\right] \underbrace{\leq}_{\text{Markov}} \frac{\mathbf{E}\left[e^{tX}\right]}{e^{ta}}$.

$$\mathbf{Pr}[X \geq (1+\delta)\mu] = \mathbf{Pr}\left[e^{tX} \geq e^{t(1+\delta)\mu}\right] \underbrace{\leq}_{\text{Markov}} \frac{\mathbf{E}\left[e^{tX}\right]}{e^{t(1+\delta)\mu}} \qquad (*)$$

$$\mathbf{E}\left[e^{tX}\right] = \mathbf{E}\left[e^{t(\sum_{i=1}^{n} X_i)}\right] = \mathbf{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] \underbrace{=}_{\text{Ind.}X_i} \prod_{i=1}^{n} \mathbf{E}\left[e^{tX_i}\right].$$

$$\mathbf{E}\left[e^{tX_i}\right] = p_i e^t + (1-p_i)e^0 = p_i(e^t - 1) + 1 < e^{p_i(e^t-1)}.$$

$$\therefore \prod_{i=1}^{n} \mathbf{E}\left[e^{tX_i}\right] < \prod_{i=1}^{n} e^{p_i(e^t-1)} = e^{\sum_{i=1}^{n} p_i(e^t-1)} \underbrace{=}_{e^t=\Theta(1)} e^{\mu(e^t-1)}.$$

From (*): $\mathbf{Pr}[X \geq (1+\delta)\mu] < \frac{e^{\mu(e^t-1)}}{e^{t(1+\delta)\mu}} = e^{\mu(e^t-1-t-\delta t)}$

Proof of Chernoff-1: Upper tail

We got $\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] < e^{\mu(e^t-1-t-\delta t)}$.
To get a tight bound we have to choose t s.t. it minimizes the above expression.
i.e. we have to derivate wrt t: $\frac{\mathrm{d}(e^t-1-t-\delta t)}{\mathrm{d}t} = 0 \Rightarrow t = \ln(\delta+1)$
Substituting in the above equation:

$$\begin{aligned}\mathbf{Pr}\left[X \geq (1+\delta)\mu\right] &< e^{\mu((\delta+1)-1-\ln(\delta+1)-\delta\ln(\delta+1))} \\ &= \left(\frac{e^{\delta+1-1}}{e^{(\delta+1)\ln(\delta+1)}}\right)^{\mu} = \left(\frac{e^{\delta}}{(\delta+1)^{\delta+1}}\right)^{\mu}. \quad \square\end{aligned}$$

Proof of Chernoff-2: Lower tail

As before, we write inequality as inequality in exponents, multiplied by a $t > 0$, which we minimized to get the sharp bound.

As before we use Markov, but the inequality would be reversed:

$$\mathbf{Pr}\left[X < (1-\delta)\mu\right] = \mathbf{Pr}\left[e^{-tX} > e^{-t(1-\delta)\mu}\right] \le \frac{\mathbf{E}\left[e^{-tX}\right]}{e^{-t(1-\delta)\mu}}.$$

As $X = \sum X_i$, where $\{X_i\}$ are independent, then $e^{-tX} = \prod_{i=1}^{n} e^{-tX_i}$,

$$\Rightarrow \mathbf{E}\left[e^{-tX}\right] = \mathbf{E}\left[\prod_{i=1}^{n} e^{-tX_i}\right] = \prod_{i=1}^{n} \mathbf{E}\left[e^{-tX_i}\right].$$

But $\mathbf{E}\left[e^{-tX_i}\right] = p_i e^{-t} + (1-p_i)e^{0} = p_i e^{-t} + (1-p_i) =$
$1 - p_i(1-e^{-t}) \underbrace{\le}_{e^{-t} \ge 1-t} e^{-p_i(1-e^{-t})} \le e^{p_i(e^{-t}-1)}$

$$\Rightarrow \prod_{i=1}^{n} \mathbf{E}\left[e^{-tX_i}\right] < \prod_{i=1}^{n} e^{p_i(e^{-t}-1)} = e^{\sum_i p_i(e^{-t}-1)} = e^{(\mu(e^{-t}-1))}$$

So $\mathbf{Pr}\left[X < (1-\delta)\mu\right] < \frac{e^{(\mu(e^{-t}-1))}}{e^{-t(1-\delta)\mu}} = e^{\mu(e^{-t}+t-t\delta-1)}$.

Proof of Chernoff-2: Lower tail

We have to minimize wrt t: $\mathbf{Pr}\left[X < (1-\delta)\mu\right] < e^{\mu(e^{-t}+t-t\delta-1)}$.
$\frac{\mathrm{d}\mu(e^{-t}+t-t\delta-1)}{\mathrm{d}t} = 0 \Rightarrow t = \ln\frac{1}{1-\delta}$.
Substituting back into the above equation,

$$\begin{aligned}\mathbf{Pr}\left[X < (1-\delta)\mu\right] &< e^{\mu((-e^{\ln(1/(1-\delta))})+(1-\delta)\ln(1/(1-\delta))-1)} \\ &= e^{\mu((1-\delta)+(1-\delta)(\ln(1)-\ln(1-\delta))-1)} \\ &= e^{\mu((1-\delta)-1+1/((1-\delta)^{1-\delta})} = \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}\end{aligned}$$

□

Powerful Technique: Chernoff + Union-Bound

Assume we have an event $A = \cup_{i=1}^{n} A_i$, where the $\{A_i\}_{i=1}^{n}$ are NOT independent, and we want to prove that the probability that A has a bad instance $\to 0$ (it is tiny).

The technique consists in:

1. Use Chernoff to prove that for each A_i the probability of a bad instance is very small, for each A_i of the n ones, i.e. we compute that $\mathbf{Pr}\left[A_i \text{ is bad}\right]$ is very small,
2. use Union-Bound to prove $\mathbf{Pr}\left[A \text{ is bad}\right] = \mathbf{Pr}\left[\cup_{i=1}^{n} A_i \text{ is bad}\right] \leq \sum_{i=1}^{n} \mathbf{Pr}\left[A_i \text{ is bad}\right]$ is very small.

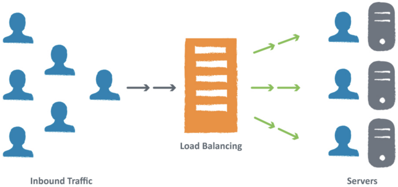
Notice, that means that we need $\mathbf{Pr}\left[A_i \text{ is bad}\right] = \omega(1/n)$, so the sum does not affect $\mathbf{Pr}\left[A \text{ is bad}\right]$.

Load balancing problem

Suppose we have k servers and n jobs, $n >> k$. Assume n jobs stream sequentially but very quickly, we have to assign each job to a server, where each job take a while to process. We are interested in to keep similar load in each servers. We want to have an algorithm that on the fly distribute the jobs into the servers, to balance the load between them, as much as we can.

Random algorithm for load balancing

We want to prove "how close" our algorithm achieve a load balance to the perfect load balance $= n/k$,
i.e. prove that w.h.p., the maximum load of all the servers is near n/k



Randomized Algorithm: Assign independently each job to a random server, with probability $= 1/k$.

Load balancing: correctness

For $(1 \leq i \leq k)$ let X_i be a r.v. counting the number of jobs handled by server i. (notice they are not indicator r.v.)

For each $X_i \in B(n, \frac{1}{k}) \Rightarrow \mathbf{E}[X_i] = \frac{n}{k}$ clear?

But $(X_1, \ldots, X_k)$ are not independent, as

$$\underbrace{\mathbf{Pr}[(X_1 = n) \cap \cdots \cap (X_k = n)]}_{=0} \neq (\underbrace{\mathbf{Pr}[X_1 = n] \cdots \mathbf{Pr}[X_k = n]}_{=(\frac{1}{k})^{kn}}.$$

Let M be a r.v. counting the maximum load among all the k servers. $M = \max\{X_1, \ldots, X_n\}$

We want to show $\mathbf{Pr}\left[M \geq \frac{n}{k} + \gamma\right]$ very small, for a not too large γ.

Correctness-2

For any $1 \leq i \leq k$ define the bad event B_i as $B_i \equiv X_i \geq \frac{n}{k} + \gamma$,

Define the event $B = \cup_{i=1}^{k} B_i$, i.e B is the event $M \geq \frac{n}{k} + \gamma$.

We aim to show that $\mathbf{Pr}\left[B\right] \leq \frac{1}{k^2}$, $\Rightarrow \mathbf{Pr}\left[\bar{B}\right] > 1 - \frac{1}{k^2}$.

Notice by independence, for all $1 \leq i \leq k$ we have the same value of $\mathbf{Pr}\left[B_i\right]$. therefore, let $\mathbf{Pr}\left[B_i\right] = \mathbf{Pr}\left[X_i \geq \frac{n}{k} + \gamma\right] = \beta$.

To get $\mathbf{Pr}\left[B\right] \leq \frac{1}{k^2}$, using Union Bound:
$\mathbf{Pr}\left[B\right] \leq \sum_{i=1}^{k} \mathbf{Pr}\left[B_i\right] = k\beta$, which we need $= \frac{1}{k^2}$, $\Rightarrow$ we need $\mathbf{Pr}\left[B_1\right] = \beta \leq \frac{1}{k^3}$.

W.o.l.g let us compute $\mathbf{Pr}\left[B_1\right]$.

Showing that $\mathbf{Pr}\left[B_1\right] \leq 1/k^3$

As $X_1 \in B(n, \frac{1}{k})$, then $X_1 = \sum_{j=1}^{n} I_j$, where I_j is i.r.v. that is 1 if job j goes to server 1. So $\mathbf{Pr}\left[I_j = 1\right] = \frac{1}{k}$.
$\Rightarrow \mathbf{E}\left[X_1\right] = \mu = \sum_{i=1}^{n} \sum_{i=1}^{n} \frac{1}{k} = \frac{n}{k}$.

We use Ch-2.2 to bound $\mathbf{Pr}\left[B_1\right] = \mathbf{Pr}\left[X_1 \geq \mu + \gamma\right]$.

$$\mathbf{Pr}\left[X_1 \geq (1+\delta)(\tfrac{n}{k})\right] = \mathbf{Pr}\left[X_1 \geq \left(\frac{n}{k} + \underbrace{\frac{\delta n}{k}}_{\gamma}\right)\right] \leq e^{-\frac{\delta^2 \mu}{3}}$$

We need to take values of δ and γ, to make everything work.

Choosing values of δ and γ so everything works

We know $n >> k$, we want $\delta < 1$ and $\mathbf{Pr}[B_1] \leq 1/k^3$, then we can make

$$\frac{1}{k^3} = e^{-\frac{\mu\delta^2}{3}}.$$

Taking ln in both sides: $\mu\delta^2 = 9 \ln k \Rightarrow \delta = 3\sqrt{\ln k}\sqrt{k/n}$.

As $\gamma = \frac{\delta n}{k} \Rightarrow \gamma = 3\sqrt{\ln k}\sqrt{n/k}$.

Therefore, $\mathbf{Pr}[B_1] = \mathbf{Pr}\left[X_1 \geq \mu + 3\sqrt{\frac{n \ln k}{k}}\right] \leq \frac{1}{k^3}$,

and $\mathbf{Pr}[B] \leq \frac{1}{k^2}$. □

The final result

We have proved that the simple randomized algorithm to allocate n jobs to k servers, with $n \geq 9k \ln k$, we get that the algorithm produces a load balancing, where the probability of having a bad event, i.e. at least of of the loads in one server deviates more that $3\sqrt{\ln k}\sqrt{n/k}$ from the expected is $1/k^3$.

Therefore. w.h.p. the randomized algorithm will keep the load concentrated around n/k.

Consequences

In practice, how good is that bound ?

Pretty good! If $n = 10^6$ and $k = 10^3$, $n/k = 10^3$ and $\gamma = 250$. So the result ⇒ w.h.p. , the maximum load is ≤ 1250.

There are better algorithms to the load distribution's problem, but they use more advanced probability techniques, as the power of two choices.

Chernoff: More Sampling

(See also section 4.2.3 in MU book)

We want to poll a sample of size n from a large population of N individuals, about the if they like or they do not like, a given product (answer yes/no).

We want to estimate the real fraction p $(0 < p < 1)$ of the population N, that likes the product, i.e. $p = \#\text{yes votes}/N$.

For that, we sample u.a.r. n persons, i.e. with replacement, and want to know how large n should be so the sampling yields an estimation $\tilde{p} = \#\text{yes answers}/n$ of the vote intention, which is "accurate" and has a high "confidence".

Sampling: Accuracy and confidence

- Accuracy: It is difficult to pinpoint exactly the value of p, so we consider a $\delta > 0$ (the accuracy), and define an interval $[\tilde{p} - \delta, \tilde{p} + \delta]$, such that $\mathbf{Pr}\left[p \in [\tilde{p} - \delta, \tilde{p} + \delta]\right]$ is very high.
- Confidence: choosing γ as small as possible so that $\mathbf{Pr}\left[p \in [\tilde{p} - \delta, \tilde{p} + \delta]\right] \geq 1 - \gamma$, where $1 - \gamma$ is the confidence.

Notice we have to tune the values of n, δ and γ as to optimize the accuracy δ with as high as possible confidence $1 - \gamma$.

In a poll, we want to be able to say things like:
This poll is 3% accurate, 19 times out of 20.
Which mean that with confidence $1 - \gamma = 19/20 = 95\%$, the outcome on the whole population N is $\pm 3\%$ of our obtained prediction $\tilde{p}$, i.e. the accuracy is $\delta = 0.03$.

Sampling

Let n be the selected number of people that we poll. Define a set of independent r.v. $\{X_i\}_{i=1}^n$, where each $X_i = 1$ if the i-th person would vote for the product, otherwise $X_i = 0$.

Let $X = \sum_{i=1}^n X_i$, then $X \in B(n, p)$ and X count the number of people that likes the product

Define our "guess" $\tilde{p}$ as $X = \tilde{p}n$.

We want to compute how large do we have to make n to have a good "accuracy" δ with high "confidence" $1 - \gamma$.

Sampling Theorem

Sampling Theorem: Suppose we use independent, uniformly random samples (with replacement) to compute an estimate $\tilde{p}$, for p. If the number of samples we use is n, satisfies $n \geq \frac{3}{\delta^2} \ln \frac{2}{\gamma}$, then we can assert that:

$$\mathbf{Pr}\left[p \in [\tilde{p} - \delta, \tilde{p} + \delta]\right] \geq 1 - \gamma.$$

Proof: Recall $X = \tilde{p}n$ has $\mathbf{E}[X] = np$.
But given a particular sampling of n people, we find that exactly $X = n\tilde{p}$ people like the product,
we have to find values of δ and γ s.t.:

$$\mathbf{Pr}\left[p \in [\tilde{p} - \delta, \tilde{p} + \delta]\right] = \mathbf{Pr}\left[np \in [n(\tilde{p} - \delta), n(\tilde{p} + \delta)]\right] \geq 1 - \gamma.$$

Proof of the sampling theorem

If $p \notin [\tilde{p} - \delta, \tilde{p} + \delta]$ is because either,

- $p < \tilde{p} - \delta \Rightarrow X = n\tilde{p} > n(p + \delta) = \mu(1 + \delta/p)$, or
- $p > \tilde{p} + \delta \Rightarrow X = n\tilde{p} < n(p - \delta) = \mu(1 - \delta/p)$.

Using Ch-2, we get

$$\begin{aligned}\mathbf{Pr}\left[p \notin [\tilde{p} - \delta, \tilde{p} + \delta]\right] &= \mathbf{Pr}\left[X < np\left(1 - \delta/p\right)\right] + \mathbf{Pr}\left[X > np\left(1 + \delta/p\right)\right] \\ &< e^{-n\delta^2/2p} + e^{-n\delta^2/3p}\end{aligned}$$

As $p \leq 1$ we get

$$\mathbf{Pr}\left[p \in [\tilde{p} - \delta, \tilde{p} + \delta]\right] = 1 - \mathbf{Pr}\left[p \notin [\tilde{p} - \delta, \tilde{p} + \delta]\right] \geq 1 - 2e^{-n\delta^2/3}.$$

But if we want confidence $1 - \delta$, then $\gamma \geq 2e^{-\frac{n\delta^2}{3}}$
$\Rightarrow \frac{2}{\gamma} \leq e^{\frac{n\delta^2}{3}} \Rightarrow \frac{2}{\gamma} \leq \frac{n\delta^2}{3} \Rightarrow n \geq \frac{3}{\delta^2} \ln \frac{2}{\gamma}$ □

Sampling Theorem: Some comments

In the previous example, $\delta = 3\%$ and confidence 95% i.e. $\gamma = 1/20$, then we need $n \geq \lceil \frac{3}{0.02^2} \ln \frac{2}{1/20} \rceil = 12297$ people giving valid answers.

- Notice in the Sampling Theorem, the number of samples n does not depend on the size N of the total population. (i.e. the number of samples you need to get a certain accuracy and a certain confidence only depends on that accuracy and confidence).
- Computing a high accuracy could be costly in the number n of samples, because of the $1/\delta^2$ term. We should design the sampling to tune between accuracy and a realistic sampling of people.
- Getting really high confidence is cheap: because of the ln, it hardly costs anything to get a very small δ.