Lecture 3: “Occupancy, Moments and Deviations”

Sotiris Nikoletseas
Professor

CEID - ETY Course
2017 - 2018

1. Some basic inequalities (I)

(i) $\left(1+\frac{1}{n}\right)^n \leq e$

<u>Proof:</u> It is: $\forall x \geq 0$: $1+x \leq e^x$. For $x=\frac{1}{n}$, we get $\left(1+\frac{1}{n}\right)^n \leq \left(e^{\frac{1}{n}}\right)^n = e$

(ii) $\left(1-\frac{1}{n}\right)^{n-1} \geq \frac{1}{e}$

<u>Proof:</u> It suffices that $\left(\frac{n-1}{n}\right)^{n-1} \geq \frac{1}{e} \Leftrightarrow \left(\frac{n}{n-1}\right)^{n-1} \leq e$

But $\frac{n}{n-1} = 1+\frac{1}{n-1}$, so it suffices that $\left(1+\frac{1}{n-1}\right)^{n-1} \leq e$ which is true by (i).

1. Some basic inequalities (II)

(iii) $n! \geq \left(\frac{n}{e}\right)^n$

Proof: It is obviously $\dfrac{n^n}{n!} \leq \displaystyle\sum_{i=0}^{\infty} \frac{n^i}{i!}$

But $\displaystyle\sum_{i=0}^{\infty} \frac{n^i}{i!} = e^n$ from Taylor's expansion of $f(x) = e^x$.

(iv) For any $k \leq n$: $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$

Proof: Indeed, $k \leq n \Rightarrow \frac{n}{k} \leq \frac{n-1}{k-1}$
Inductively $k \leq n \Rightarrow \frac{n}{k} \leq \frac{n-i}{k-i}, (1 \leq i \leq k-1)$
Thus $\left(\frac{n}{k}\right)^k \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-(k-1)}{k-(k-1)} = \frac{n^{\underline{k}}}{k!} = \binom{n}{k}$
For the right inequality we obviously have $\binom{n}{k} \leq \frac{n^k}{k!}$
and by (iii) it is $k! \geq \left(\frac{k}{e}\right)^k$

(i) Boole's inequality (or union bound)

Let random events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$. Then

$$Pr\left\{\bigcup_{i=1}^{n} \mathcal{E}_i\right\} = Pr\{\mathcal{E}_1 \cup \mathcal{E}_2 \cup \cdots \cup \mathcal{E}_n\} \leq \sum_{i=1}^{n} Pr\{\mathcal{E}_i\}$$

Note: If the events are disjoint, then we get equality.

(ii) Expectation (or Mean)

Let X a random variable with probability density function (pdf) $f(x)$. Its expectation is:

$$\mu_x = E[X] = \sum_x x \cdot Pr\{X = x\}$$

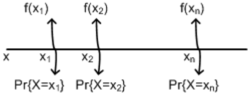
If X is continuous, $\mu_x = \int_{-\infty}^{\infty} x f(x)\, dx$

2. Preliminaries

(ii) Expectation (or Mean)

Properties:

- $\forall X_i \; (i = 1, 2, \ldots, n) : E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$
  This important property is called "linearity of expectation".
- $E[cX] = cE[X]$, where c constant
- if X, Y stochastically independent, then $E[X \cdot Y] = E[X] \cdot E[Y]$
- Let $f(X)$ a real-valued function of X. Then $E[f(x)] = \sum_{x} f(x) Pr\{X = x\}$



(iii) Markov's inequality

Theorem: Let X a non-negative random variable. Then, $\forall t > 0$

$$Pr\{X \geq t\} \leq \frac{E[X]}{t}$$

Proof:
$$E[X] = \sum_{x} x Pr\{X = x\} \geq \sum_{x \geq t} x Pr\{X = x\}$$
$$\geq \sum_{x \geq t} t Pr\{X = x\} = t \sum_{x \geq t} Pr\{X = x\} = t\ Pr\{X \geq t\}$$

Note: Markov is a (rather weak) concentration inequality, e.g.

$Pr\{X \geq 2E[X]\} \leq \frac{1}{2}$

$Pr\{X \geq 3E[X]\} \leq \frac{1}{3}$

etc

(iv) Variance (or second moment)

- Definition: $Var(X) = E[(X-\mu)^2]$, where $\mu = E[X]$
  i.e. it measures (statistically) deviations from mean.

- Properties:
  - $Var(X) = E[X^2] - E^2[X]$
  - $Var(cX) = c^2 Var(X)$, where c constant.
  - if X, Y independent, it is $Var(X+Y) = Var(X) + Var(Y)$

Note: We call $\sigma = \sqrt{Var(X)}$ the standard deviation of X.

(v) Chebyshev's inequality

Theorem: Let X a r.v. with mean $\mu = E[X]$. It is:

$$Pr\{|X - \mu| \geq t\} \leq \frac{Var(X)}{t^2} \qquad \forall t > 0$$

Proof: $Pr\{|X - \mu| \geq t\} = Pr\{(X - \mu)^2 \geq t^2\}$

From Markov's inequality:

$$Pr\{(X - \mu)^2 \geq t^2\} \leq \frac{E[(X-\mu)^2]}{t^2} = \frac{Var(X)}{t^2}$$

Note: Chebyshev's inequality provides stronger (than Markov's) concentration bounds, e.g.

$Pr\{|X - \mu| \geq 2\sigma\} \leq \frac{1}{4}$

$Pr\{|X - \mu| \geq 3\sigma\} \leq \frac{1}{9}$

etc

3. Occupancy - importance

- occupancy procedures are actually stochastic processes (i.e, random processes in time). Particularly, the occupancy process consists in placing randomly balls into bins, one at a time.
- occupancy problems/processes have fundamental importance for the analysis of randomized algorithms, such as for data structures (e.g. hash tables), routing etc.

3. Occupancy - definition and basic questions

- general occupancy process: we uniformly randomly and independently put, one at a time, m distinct objects ("balls") each one into one of n distinct classes ("bins").
- basic questions:
  - what is the maximum number of balls in any bin?
  - how many balls are needed so as no bin remains empty, with high probability?
  - what is the number of empty bins?
  - what is the number of bins with k balls in them?
- Note: in the next lecture we will study the coupon collector's problem, a variant of occupancy.

3. Occupancy - the case $m = n$

Let us randomly place $m = n$ balls into n bins.

Question: What is the maximum number of balls in any bin?

Remark: Let us first estimate the expected number of balls in any bin.

For any bin i $(1 \leq i \leq n)$ let $X_i = \#$ balls in bin i.
Clearly $X_i \sim B(m, \frac{1}{n})$ (binomial)
So $E[X_i] = m\frac{1}{n} = n\frac{1}{n} = 1$

We however expect this "mean" (expected) behaviour to be highly improbable, i.e.,

- some bins get no balls at all
- some bins get many balls

3. Occupancy - the case $m = n$

Theorem 1. With probability at least $1 - \frac{1}{n}$, no bin gets more than $k^* = \frac{3 \ln n}{\ln \ln n}$ balls.

Proof: Let $\mathcal{E}_j(k)$ the event "bin j gets k or more balls". Because of symmetry, we first focus on a given bin (say bin 1). It is
$\Pr\{\text{bin 1 gets exactly } i \text{ balls}\} = \binom{n}{i}\left(\frac{1}{n}\right)^i\left(1 - \frac{1}{n}\right)^{n-i}$
since we have a binomial $B(n, \frac{1}{n})$. But

$\binom{n}{i}\left(\frac{1}{n}\right)^i\left(1 - \frac{1}{n}\right)^{n-i} \leq \binom{n}{i}\left(\frac{1}{n}\right)^i \leq \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i$
(from basic inequality iv)

$$\text{Thus } Pr\{\mathcal{E}_1(k)\} \leq \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i \leq \left(\frac{e}{k}\right)^k \cdot \left(1 + \frac{e}{k} + \left(\frac{e}{k}\right)^2 + \cdots\right) =$$
$$= \left(\frac{e}{k}\right)^k \frac{1}{1 - \frac{e}{k}}$$

3. Occupancy - the case $m = n$

Now, let $k^* = \left\lceil \frac{3\ln n}{\ln\ln n} \right\rceil$. Then:

$$Pr\{\mathcal{E}_1(k^*)\} \leq \left(\frac{e}{k^*}\right)^{k^*} \frac{1}{1-\frac{e}{k^*}} \leq 2\left(\frac{e}{\frac{3\ln n}{\ln\ln n}}\right)^{k^*}$$

since it suffices $\frac{1}{1-\frac{e}{k^*}} \leq 2 \Leftrightarrow \frac{k^*}{k^*-e} \leq 2 \Leftrightarrow k^* \leq 2k^* - 2e \Leftrightarrow$

$\Leftrightarrow k^* \geq 2e$ which is true.

But $2\left(\frac{e}{\frac{3\ln n}{\ln\ln n}}\right)^{k^*} = 2\left(e^{1-\ln 3-\ln\ln n+\ln\ln\ln n}\right)^{k*}$

$$\leq 2\left(e^{-\ln\ln n+\ln\ln\ln n}\right)^{k*} \leq 2\exp\left(-3\ln n + 6\ln n\frac{\ln\ln\ln n}{\ln\ln n}\right)$$

$$\leq 2\exp(-3\ln n + 0.5\ln n) = 2\exp(-2.5\ln n) \leq \frac{1}{n^2}$$

for n large enough.

3. Occupancy - the case $m = n$

Thus,

$$Pr\{\text{any bin gets more than } k^* \text{ balls}\} = Pr\left\{\bigcup_{j=1}^{n} \mathcal{E}_j(k^*)\right\}$$

$$\leq \sum_{j=1}^{n} Pr\{\mathcal{E}_j(k^*)\} \leq nPr\{\mathcal{E}_1(k^*)\} \leq n\frac{1}{n^2} = \frac{1}{n} \text{ (by symmetry) } \square$$

3. Occupancy - the case $m = n \log n$

- We showed that when $m = n$ the mean number of balls in any bin is 1, but the maximum can be as high as $k^* = \frac{3 \ln n}{\ln \ln n}$
- The next theorem shows that when $m = n \log n$ the maximum number of balls in any bin is more or less the same as the expected number of balls in any bin.
- Theorem 2. When $m = n \ln n$, then with probability $1 - o(1)$ every bin has $O(\log n)$ balls.

3. Occupancy - the case $m = n$ - An improvement

- If at each iteration we randomly pick d bins and throw the ball into the bin with the smallest number of balls, we can do much better than in Theorem 1:
- Theorem 3. We place $m = n$ balls sequentially in n bins as follows:
  For each ball, $d \geq 2$ bins are chosen uniformly at random (and independently). Each ball is placed in the least full of the d bins (ties broken randomly). When all balls are placed, the maximum load at any bin is at most $\frac{\ln \ln n}{\ln d} + O(1)$, with probability at least $1 - o(1)$ (in other words, a more balanced balls distribution is achieved).

3. Occupancy - tightness of Theorem 1

Theorem 1 shows that when $m = n$ then the maximum load in any bin is $O\left(\frac{\ln n}{\ln \ln n}\right)$, with high probability. We now show that this result is tight:

Lemma 1: There is a $k = \Omega\left(\frac{\ln n}{\ln \ln n}\right)$ such that bin 1 has k balls with probability at least $\frac{1}{\sqrt{n}}$.

Proof: $Pr[k \text{ balls in bin } 1] = \binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k}$

$$\geq \left(\frac{n}{k}\right)^k \frac{1}{n^k}\left(1-\frac{1}{n}\right)^{n-k} \qquad \text{(from basic inequality iv)}$$

$$= \left(\frac{1}{k}\right)^k\left(1-\frac{1}{n}\right)^{n-k} \geq \left(\frac{1}{k}\right)^k\left(\frac{1}{2e}\right) = \frac{1}{2e}\left(\frac{1}{k}\right)^k \quad (\text{for } n \geq 2)$$

3. Occupancy - tightness of Theorem 1

By putting $k = \frac{c \ln n}{\ln \ln n}$ we get

$Pr\{\frac{c \ln n}{\ln \ln n}$ balls in bin $1\} \geq \frac{1}{2e} \left(\frac{\ln \ln n}{c \ln n}\right)^{\frac{c \ln n}{\ln \ln n}} \geq \left(\frac{1}{c \ln n}\right)^{\frac{c \ln n}{\ln \ln n}}$
(for $n \geq 4$)

$= \left(\frac{1}{c 2^{\ln \ln n}}\right)^{\frac{c \ln n}{\ln \ln n}} = \frac{1}{c 2^{\ln \ln n \frac{c \ln n}{\ln \ln n}}} = \frac{1}{c 2^{c \ln n}} = \frac{1}{c n^c} = \Omega(n^{-c})$

Setting $c = \frac{1}{2}$ we get $Pr\{\frac{c \ln n}{\ln \ln n}$ balls in bin $1\} \geq \Omega(\frac{1}{\sqrt{n}})$ □

3. Occupancy - the case $m = n \log n$

Towards a proof of Theorem 2. We use the following bound.

Theorem (Chernoff bound). Let X a r.v.:

$$X = \sum_{i=1}^{n} X_i = X_1 + \cdots + X_n$$

where for all i ($1 \leq i \leq n$) the X_i's are independent and

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$

Let $E[X] = np = \mu$.
Then, $\forall \delta > 0$

$$Pr\{X \geq \mu(1+\delta)\} \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \Box$$

3. Occupancy - the case $m = n \log n$

When placing $m = n \log n$ balls into n bins let

$$X_i = \begin{cases} 1, & \text{if ball } i \text{ lands in bin 1 (prob=}\frac{1}{n}\text{)} \\ 0, & \text{else} \end{cases}$$

and

$$X = \sum_{i=1}^{m} X_i = \#\text{ of balls in bin 1.}$$

Then

$$\mu = E[X] = m\frac{1}{n} = \ln n$$

.

3. Occupancy - the case $m = n \log n$

Let us estimate the probability that bin 1 receives more than e.g. $10 \ln n$ balls

- by the Markov inequality:

  $Pr\{X \geq 10 \ln n\} \leq \frac{\ln n}{10 \ln n} = \frac{1}{10}$ (the bound is not strong)

3. Occupancy - the case $m = n \log n$

- by the Chebyshev's inequality:

  X is actually binomial, i.e. $X \sim B(m, \frac{1}{n})$ thus its variance is $Var(X) = m\left(\frac{1}{n}\right)\left(1 - \frac{1}{n}\right) = \frac{m}{n} - \frac{m}{n^2} \leq \frac{m}{n}$

  Thus $Pr\{X \geq \frac{m}{n} + k\} \leq Pr\{|X - \frac{m}{n}| \geq k\} \leq \frac{Var(X)}{k^2} \leq \frac{m}{nk^2}$

  For $m = n \ln n \Rightarrow \frac{m}{n} = \ln n$ and for $k = 9 \ln n$ we have $Pr\{X \geq 10 \ln n\} = Pr\{X \geq \ln n + 9 \ln n\} \leq \frac{n \ln n}{n 81 \ln^2 n} = \frac{1}{81 \ln n}$

  (a bound which is better than the one by Markov's inequality)

3. Occupancy - the case $m = n \log n$

Let us estimate the probability that bin 1 receives more than e.g. $10 \ln n$ balls

- by Chernoff bound:

  $Pr\{X \geq 10 \ln n\} = Pr\{X \geq (1+9) \ln n\} \leq \left(\frac{e^9}{10^{10}}\right)^{\ln n} \leq \frac{1}{n^{10}}$
  (much stronger)

  Thus,
  $Pr\{\exists \text{ bin with more than } 10 \ln n \text{ balls }\} \leq n \frac{1}{n^{10}} = n^{-9}$
  $\Rightarrow Pr\{\text{all bins have less than } 10 \ln n \text{ balls}\} \geq 1 - n^{-9}$

3. Occupancy - the case $m = n \log n$

A similar bound applies to the "low tail", i.e. the probability that there exists a bin with less than, say, $\frac{1}{10} \ln n$ balls tends to zero, as n tends to infinity. Overall, there is high concentration around the mean value of $\ln n$ balls per bin.

3. Occupancy - the case $m = n \log n$

Note: The corresponding bounds (for any bin) by Markov's inequality and Chebychev's inequality are trivial:

- by Markov we get $\leq \frac{n}{10}$
- by Chebyshev we get $\leq \frac{n}{81 \ln n}$

3. Occupancy - all balls in distinct bins

Let the experiment of sequentially putting m balls randomly in n bins.

Problem: How large m can be so that the probability of all balls being placed in distinct bins remains high?

3. Occupancy - all balls in distinct bins

For $2 \le i \le m$, let $\mathcal{E}_i$= "the ith ball lands in a bin not occupied by the first $i-1$ balls". The desired probability is:

$$Pr\{\bigcap_{i=2}^{m} \mathcal{E}_i\} = \prod_{i=2}^{m} Pr\{\mathcal{E}_i | \bigcap_{j=2}^{i-1} \mathcal{E}_j\} =$$

$$Pr\{\mathcal{E}_2\} Pr\{\mathcal{E}_3|\mathcal{E}_2\} Pr\{\mathcal{E}_4|\mathcal{E}_2\mathcal{E}_3\} \cdots Pr\{\mathcal{E}_m|\mathcal{E}_2 \ldots \mathcal{E}_{m-1}\}$$

But

$$Pr\{\mathcal{E}_i | \bigcap_{j=2}^{i-1} \mathcal{E}_j\} = 1 - \frac{i-1}{n} \le e^{-\frac{i-1}{n}}$$

$$Pr\{\bigcap_{i=2}^{m} \mathcal{E}_i\} \le \prod_{i=2}^{m} e^{-\frac{i-1}{n}} = e^{-\sum_{i=2}^{m} \frac{i-1}{n}} = e^{-\frac{1}{n}\sum_{i=1}^{m-1} i} = e^{-\frac{m(m-1)}{2n}}$$

3. Occupancy - all balls in distinct bins

Thus, when $m = \lceil \sqrt{2n} + 1 \rceil$ then this probability is at most $\frac{1}{e}$ while when m increases the probability decreases rapidly.

Note: This is similar to the classic "birthday paradox" in probability theory.