Chapter 4

The Occupancy and Coupon Collector problems

By Sariel Har-Peled, May 29, 2013[1]

4.1 Preliminaries

Definition 4.1.1 (Variance and Standard Deviation). For a random variable X, let $\mathbf{V}[X] = \mathbf{E}\Big[(X - \mu_X)^2\Big] = \mathbf{E}\Big[X^2\Big] - \mu_X^2$ denote the ***variance*** of X, where $\mu_X = \mathbf{E}[X]$. Intuitively, this tells us how concentrated is the distribution of X.

The ***standard deviation*** of X, denoted by σ_X is the quantity $\sqrt{\mathbf{V}[X]}$.

Observation 4.1.2. *(i)* $\mathbf{V}[cX] = c^2\,\mathbf{V}[X]$.

(ii) For X and Y independent variables, we have $\mathbf{V}[X + Y] = \mathbf{V}[X] + \mathbf{V}[Y]$.

Definition 4.1.3 (Bernoulli distribution). Assume, that one flips a coin and get 1 (heads) with probability p, and 0 (i.e., tail) with probability $q = 1 - p$. Let X be this random variable. The variable X is has *Bernoulli distribution with parameter* p. Then $\mathbf{E}[X] = p$, and $\mathbf{V}[X] = pq$.

Definition 4.1.4 (Binomial distribution). Assume that we repeat a Bernoulli experiments n times (independently!). Let $X_1, \ldots, X_n$ be the resulting random variables, and let $X = X_1 + \cdots + X_n$. The variable X has the *binomial distribution* with parameters n and p. We denote this fact by $X \sim B(n, p)$. We have

$$b(k; n, p) = \mathbf{Pr}[X = k] = \binom{n}{k} p^k q^{n-k}.$$

Also, $\mathbf{E}[X] = np$, and $\mathbf{V}[X] = npq$.

[1]

Observation 4.1.5. *Let $C_1, \ldots, C_n$ be random events (not necessarily independent). Than*

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} C_i\right] \leq \sum_{i=1}^{n} \mathbf{Pr}[C_i] .$$

(This is usually referred to as the **union bound***.) If $C_1, \ldots, C_n$ are* disjoint *events then*

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} C_i\right] = \sum_{i=1}^{n} \mathbf{Pr}[C_i] .$$

Lemma 4.1.6. *For any positive integer n, we have:*
(i) $(1 + 1/n)^n \leq e$.
(ii) $(1 - 1/n)^{n-1} \geq e^{-1}$.
(iii) $n! \geq (n/e)^n$.
(iv) For any $k \leq n$, we have: $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$.

Proof: (i) Indeed, $1 + 1/n \leq \exp(1/n)$, since $1 + x \leq e^x$, for $x \geq 0$. As such $(1 + 1/n)^n \leq \exp(n(1/n)) = e$.

(ii) Rewriting the inequality, we have that we need to prove$\left(\frac{n-1}{n}\right)^{n-1} \geq \frac{1}{e}$. This is equivalence to proving $e \geq \left(\frac{n}{n-1}\right)^{n-1} = \left(1 + \frac{1}{n-1}\right)^{n-1}$, which is our friend from (i).

(iii) Indeed,

$$\frac{n^n}{n!} \leq \sum_{i=0}^{\infty} \frac{n^i}{i!} = e^n,$$

by the Taylor expansion of $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. This implies that $(n/e)^n \leq n!$, as required.

(iv) Indeed, for any $k \leq n$, we have $\frac{n}{k} \leq \frac{n-1}{k-1}$ since $kn - n = n(k-1) \leq k(n-1) = kn - k$. As such, $\frac{n}{k} \leq \frac{n-i}{k-i}$, for $1 \leq i \leq k - 1$. As such,

$$\left(\frac{n}{k}\right)^k \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-i}{k-i} \cdots \frac{n-k+1}{1} = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

As for the other direction, we have

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \frac{n^k}{\left(\frac{k}{e}\right)^k} = \left(\frac{ne}{k}\right)^k,$$

by (iii). ■

4.2 Occupancy Problems

Problem 4.2.1. We are throwing m balls into n bins randomly (i.e., for every ball we randomly and uniformly pick a bin from the n available bins, and place the ball in the bin picked). What is the maximum number of balls in any bin? What is the number of bins which are empty? How many balls do we have to throw, such that all the bins are non-empty, with reasonable probability?

Let X_i be the number of balls in the ith bins, when we throw n balls into n bins (i.e., $m = n$). Clearly,

$$\mathbf{E}[X_i] = \sum_{j=1}^{n} \mathbf{Pr}\Big[\text{The } j\text{th ball fall in } i\text{th bin}\Big] = n \cdot \frac{1}{n} = 1,$$

by linearity of expectation. The probability that the first bin has exactly i balls is

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

This follows by Lemma 4.1.6 (iv).

Let $C_j(k)$ be the event that the jth bin has k or more balls in it. Then,

$$\mathbf{Pr}[C_1(k)] \leq \sum_{i=k}^{n} \left(\frac{e}{i}\right)^i \leq \left(\frac{e}{k}\right)^k \left(1 + \frac{e}{k} + \frac{e^2}{k^2} + \ldots\right) = \left(\frac{e}{k}\right)^k \frac{1}{1 - e/k}.$$

Let $k^* = \lceil (3\ln n)/\ln\ln n \rceil$. Then,

$$\begin{aligned}
\mathbf{Pr}[C_1(k^*)] &\leq \left(\frac{e}{k^*}\right)^{k^*} \frac{1}{1 - e/k^*} \leq 2\left(\frac{e}{(3\ln n)/\ln\ln n}\right)^{k^*} = 2(\exp(1 - \ln 3 - \ln\ln n + \ln\ln\ln n))^{k^*} \\
&\leq 2\Big(\exp(-\ln\ln n + \ln\ln\ln n)\Big)^{k^*} \\
&\leq 2\exp\left(-3\ln n + 6\ln n \frac{\ln\ln\ln n}{\ln\ln n}\right) \leq 2\exp(-2.5\ln n) \leq \frac{1}{n^2},
\end{aligned}$$

for n large enough. We conclude, that since there are n bins and they have identical distributions that

$$\mathbf{Pr}\Big[\text{any bin contains more than } k^* \text{ balls}\Big] \leq \sum_{i=1}^{n} C_i(k^*) \leq \frac{1}{n}.$$

Theorem 4.2.2. *With probability at least* $1 - 1/n$, *no bin has more than* $k^* = \left\lceil \frac{3\ln n}{\ln\ln n} \right\rceil$ *balls in it.*

Exercise 4.2.3. Show that for $m = n\ln n$, with probability $1 - o(1)$, every bin has $O(\log n)$ balls.

It is interesting to note, that if at each iteration we randomly pick d bins, and throw the ball into the bin with the smallest number of balls, then one can do much better. We currently do not have the machinery to prove the following theorem, but hopefully we would prove it later in the course.

Theorem 4.2.4. *Suppose that n balls are sequentially places into n bins in the following manner. For each ball,* $d \geq 2$ *bins are chosen independently and uniformly at random (with replacement). Each ball is placed in the least full of the d bins at the time of placement, with ties broken randomly. After all the balls are places, the maximum load of any bin is at most* $\ln\ln n/(\ln d) + O(1)$, *with probability at least* $1 - o(1/n)$.

Note, even by setting $d = 2$, we get considerable improvement. A proof of this theorem can be found in the work by Azar *et al.* [ABKU00].

4.2.1 The Probability of all bins to have exactly one ball

Next, we are interested in the probability that all m balls fall in distinct bins. Let X_i be the event that the ith ball fell in a distinct bin from the first $i-1$ balls. We have:

$$\begin{aligned}\mathbf{Pr}\left[\bigcap_{i=2}^{m} X_i\right] &= \mathbf{Pr}[X_2]\prod_{i=3}^{m}\mathbf{Pr}\left[X_i \,\middle|\, \bigcap_{j=2}^{i-1} X_j\right] \leq \prod_{i=2}^{m}\left(\frac{n-i+1}{n}\right) \leq \prod_{i=2}^{m}\left(1-\frac{i-1}{n}\right)\\ &\leq \prod_{i=2}^{m} e^{-(i-1)/n} \leq \exp\left(-\frac{m(m-1)}{2n}\right),\end{aligned}$$

thus for $m = \left\lceil \sqrt{2n}+1\right\rceil$, the probability that all the m balls fall in different bins is smaller than $1/e$.

This is sometime referred to as the ***birthday paradox***. You have $m = 30$ people in the room, and you ask them for the date (day and month) of their birthday (i.e., $n = 365$). The above shows that the probability of all birthdays to be distinct is $\exp(-30 \cdot 29/730) \leq 1/e$. Namely, there is more than 50% chance for a birthday collision, a simple but counterintuitive phenomena.

4.3 The Markov and Chebyshev inequalities

We remind the reader that for a random variable X assuming real values, its *expectation* is $\mathbf{E}[Y] = \sum_y y \cdot \mathbf{Pr}[Y = y]$. Similarly, for a function $f(\cdot)$, we have $\mathbf{E}[f(Y)] = \sum_y f(y) \cdot \mathbf{Pr}[Y = y]$.

Theorem 4.3.1 (Markov Inequality). *Let Y be a random variable assuming only non-negative values. Then for all $t > 0$, we have*

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

Proof: Indeed,

$$\begin{aligned}\mathbf{E}[Y] &= \sum_{y\geq t} y\,\mathbf{Pr}[Y=y] + \sum_{y<t} y\,\mathbf{Pr}[Y=y] \geq \sum_{y\geq t} y\,\mathbf{Pr}[Y=y]\\ &\geq \sum_{y\geq t} t\,\mathbf{Pr}[Y=y] = t\,\mathbf{Pr}[Y\geq t].\end{aligned}$$

■

Markov inequality is tight, as the following exercise testifies.

Exercise 4.3.2. For any (integer) $k > 1$, define a random positive variable X_k such that $\mathbf{Pr}\Big[X_k \geq k\,\mathbf{E}[X_k]\Big] = \frac{1}{k}$.

Theorem 4.3.3 (Chebyshev inequality). $\mathbf{Pr}[|X - \mu_X| \geq t\sigma_X] \leq \frac{1}{t^2}$, *where* $\mu_X = \mathbf{E}[X]$ *and* $\sigma_X = \sqrt{\mathbf{V}[X]}$.

Proof: Note that

$$\mathbf{Pr}\Big[\,|X-\mu_X| \geq t\sigma_X\Big] = \mathbf{Pr}\Big[(X-\mu_X)^2 \geq t^2\sigma_X^2\Big].$$

Set $Y = (X - \mu_X)^2$. Clearly, $\mathbf{E}[Y] = \sigma_X^2$. Now, apply Markov inequality to Y. ■

4.4 The Coupon Collector's Problem

There are n types of coupons, and at each trial one coupon is picked in random. How many trials one has to perform before picking all coupons? Let m be the number of trials performed. We would like to bound the probability that m exceeds a certain number, and we still did not pick all coupons.

Let $C_i \in \{1, \ldots, n\}$ be the coupon picked in the ith trial. The jth trial is a success, if C_j was not picked before in the first $j-1$ trials. Let X_i denote the number of trials from the ith success, till after the $(i+1)$th success. Clearly, the number of trials performed is

$$X = \sum_{i=0}^{n-1} X_i.$$

Now, the probability of X_i to succeed in a trial is $p_i = (n-i)/n$, and X_i has the geometric distribution with probability p_i. As such $\mathbf{E}[X_i] = 1/p_i$, and $\mathbf{V}[X_i] = q/p^2 = (1-p_i)/p_i^2$.

Thus,

$$\mathbf{E}[X] = \sum_{i=0}^{n-1} \mathbf{E}[X_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = nH_n = n(\ln n + \Theta(1)) = n \ln n + O(n),$$

where $H_n = \sum_{i=1}^{n} 1/i$ is the nth Harmonic number.

As for variance, using the independence of $X_0, \ldots, X_{n-1}$, we have

$$\begin{aligned} \mathbf{V}[X] &= \sum_{i=0}^{n-1} \mathbf{V}[X_i] = \sum_{i=0}^{n-1} \frac{1-p_i}{p_i^2} = \sum_{i=0}^{n-1} \frac{1-(n-i)/n}{\left(\frac{n-i}{n}\right)^2} = \sum_{i=0}^{n-1} \frac{i/n}{\left(\frac{n-i}{n}\right)^2} = \sum_{i=0}^{n-1} \frac{i}{n}\left(\frac{n}{n-i}\right)^2 \\ &= n \sum_{i=0}^{n-1} \frac{i}{(n-i)^2} = n \sum_{i=1}^{n} \frac{n-i}{i^2} = n\left(\sum_{i=1}^{n} \frac{n}{i^2} - \sum_{i=1}^{n} \frac{1}{i}\right) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH_n. \end{aligned}$$

Since, $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i^2} = \pi^2/6$, we have $\displaystyle\lim_{n\to\infty} \frac{\mathbf{V}[X]}{n^2} = \frac{\pi^2}{6}$.

This implies a weak bound on the concentration of X, using Chebyshev inequality, but this is going to be quite weaker than what we implied we can do. Indeed, we have

$$\mathbf{Pr}\left[X \geq n \log n + n + t \cdot n\frac{\pi}{\sqrt{6}}\right] \leq \mathbf{Pr}\Big[\,|X - \mathbf{E}[X]| \geq t\,\mathbf{V}[X]\Big] \leq \frac{1}{t^2},$$

for any t.

Stronger bounds will be shown in the next lecture.

4.5 Notes

The material in this note covers parts of [MR95, sections 3.1,3.2,3.6]

Bibliography

[ABKU00] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal. Balanced allocations. *SIAM J. Comput.*, 29(1):180–200, 2000.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.