Exercises of Chapter 5

Chuang-Chieh Lin*

Department of Computer Science and Information Engineering,
National Chung Cheng University, Ming-Hsiung, Chiayi 621, Taiwan.

Exercise 5.8. *Our analysis of Bucket sort in Section 5.2.2 (in [3]) assumed that n elements were chosen independently and uniformly at random from the range $[0, 2^k)$. Suppose instead that n elements are chosen independently from the range $[0, 2^k)$ according to a distribution with the property that any number $x \in [0, 2^k)$ is chosen with probability at most $a/2^k$ for some fixed constant $a > 0$. Show that, under these conditions, Bucket sort still requires linear expected time.*

Solution. Suppose that we have a set of $n = 2^m$ elements to be sorted and that each element is an integer chosen independently from the range $[0, 2^k)$, where $k \geq m$, *according to a distribution with the property that any number $x \in [0, 2^k)$ is chosen with probability at most $a/2^k$ for some constant $a > 0$.* Using Bucket sort, we can sort these n numbers in two stages. In the first stage, we place the elements into n buckets. The jth bucket holds all elements whose first m binary digits corresponds to the number j. For example, if $n = 2^{10}$, bucket 3 contains all elements whose first 10 binary digits are 0000000011. When $j < l$, the elements of the jth bucket all come before the elements in the lth bucket in the sorted order. Assuming that each element can be placed in the appropriate bucket in $O(1)$ time, this stage requires only $O(n)$ time. The probability that a chosen element is placed into jth bucket (note that j has m digits) is at most

$$\frac{a}{2^k} \cdot 2^{k-m} = \frac{a}{2^m} = \frac{a}{n},$$

since an element chosen from the range $[0, 2^k)$ has k digits and the first m of them are fixed to be j (i.e., there are 2^{k-m} numbers whose first m digits are j). Thus number of elements that land in a specific bucket follows a binomial distribution $B(n, a/n)$. Buckets can be implemented using linked lists.

In the second stage, each bucket is sorted using any standard quadratic time algorithm. Concatenating the sorted lists from each bucket in order gives us the sorted order for the elements. It remains to show that the expected time spent in the second stage is only $O(n)$.

Under the assumed input distribution, Bucket sort falls naturally into *balls-and-bins model*: the elements are balls, buckets are bins, and each ball falls into a certain bin with probability at most $a/2^k$ for some constant $a > 0$. Let X_j be the number of elements that

*Email address: lincc@cs.ccu.edu.tw

land in the jth bucket. The time to sort the jth bucket is then at most $c(X_j)^2$ for some constant c. The expected time spent sorting in the second stage is at most

$$\mathbf{E}\left[\sum_{j=1}^{n} c(X_j)^2\right] = c\sum_{i=1}^{n} \mathbf{E}[X_j^2] = cn\mathbf{E}[X_1^2],$$

where the first equality follows from the linearity of expectations and the second follows from symmetry, as $\mathbf{E}[X_j^2]$ is the same for all buckets. By using the results of Section 3.2.1 (page 48 in [3]), we have

$$\mathbf{E}[X_1^2] \leq n(n-1)\left(\frac{a}{n}\right)^2 + n \cdot \frac{a}{n} = \left(1 - \frac{1}{n}\right) \cdot a^2 + a,$$

therefore, the expected running time of Bucket sort is $cn\mathbf{E}[X_1^2] \leq cn \cdot ((1-1/n)a^2 + a) = O(n)$, which is still linear. □

Exercise 5.15. *We consider another way to obtain Chernoff-like bounds in the setting of balls and bins without using Theorem 5.7 (page 101 in [3]). Consider n balls thrown randomly into n bins. Let $X_i = 1$ if the ith bin is empty and 0 otherwise. Let $X = \sum_{i=1}^{n} X_i$. Let Y_i, $i = 1, \ldots, n$, be independent Bernoulli random variables that are 1 with probability $p = (1-1/n)^n$. Let $Y = \sum_{i=1}^{n} Y_i$.*

(a) Show that $\mathbf{E}[X_1 X_2 \ldots X_k] \leq \mathbf{E}[Y_1 Y_2 \ldots Y_k]$ for any $k \geq 1$.

*(b) Show that $\mathbf{E}[e^{tX}] \leq \mathbf{E}[e^{tY}]$ for all $t \geq 0$. (*Hint: *Use the expansion for e^x and compare $\mathbf{E}[X^k]$ to $\mathbf{E}[Y^k]$.)*

(c) Derive a Chernoff bound for $\mathbf{Pr}[X \geq (1+\delta)\mathbf{E}[X]]$.

Solution. (a) Since X_i's are indicator random variables, we have

$$\mathbf{E}[X_1 X_2 \ldots X_k] = 1 \cdot \mathbf{Pr}\left[\bigcap_{i=1}^{k} \{X_i = 1\}\right] = \left(1 - \frac{k}{n}\right)^n.$$

Similarly, we have

$$\mathbf{E}[Y_1 Y_2 \ldots Y_k] = 1 \cdot \mathbf{Pr}\left[\bigcap_{i=1}^{k} \{Y_i = 1\}\right] = \left(1 - \frac{1}{n}\right)^{nk}.$$

Since $1 - k/n \leq (1-1/n)^k$ (by Bernoulli's inequality or Taylor expansion of $(1 - 1/n)^k$), we have

$$\left(1 - \frac{k}{n}\right)^n \leq \left(1 - \frac{1}{n}\right)^{nk},$$

that is, $\mathbf{E}[X_1 X_2 \ldots X_k] \leq \mathbf{E}[Y_1 Y_2 \ldots Y_k]$, for any $k \geq 1$.

(b) By the Taylor expansion of e^{tX}, we have

$$\mathbf{E}[e^{tX}] = \mathbf{E}\left[1 + tX + \frac{(tX)^2}{2!} + \ldots\right] = \sum_{j \geq 0} \mathbf{E}\left[\frac{(tX)^j}{j!}\right].$$

and

$$\mathbf{E}[e^{tY}] = \sum_{j\geq 0} \mathbf{E}\left[\frac{(tY)^j}{j!}\right].$$

Moreover,

$$\begin{aligned}
\mathbf{E}[X^k] &= \mathbf{E}[(X_1 + \ldots X_k)^k] \\
&= \mathbf{E}\left[\sum_{r_1+\ldots+r_k=k} X_1^{r_1} X_2^{r_2} \ldots X_k^{r_k}\right] \\
&= \sum_{r_1+\ldots+r_k=k} \mathbf{E}[X_1 \ldots X_k] \quad \text{(by linearity of expectation)} \\
&\leq \sum_{r_1+\ldots+r_k=k} \mathbf{E}[Y_1 \ldots Y_k] \quad \text{(by the result of (a)} \\
&= \mathbf{E}[(Y_1 + \ldots Y_k)^k] \\
&= \mathbf{E}[Y^k],
\end{aligned}$$

for every $k \geq 1$, thus we have $\mathbf{E}[e^{tX}] \leq \mathbf{E}[e^{tY}]$.

(c) Note that it can be easily derived that $\mathbf{E}[X] = \mathbf{E}[Y] = n \cdot (1 - 1/n)^n$. Let $p = (1 - 1/n)^n$. We have the following inequality:

$$\begin{aligned}
\mathbf{Pr}[X \geq (1+\delta)\mathbf{E}[X]] &= \mathbf{Pr}[e^{tX} \geq e^{t(1+\delta)\mathbf{E}[X]}] \\
&\leq \frac{\mathbf{E}[e^{tX}]}{e^{t(1+\delta)\mathbf{E}[X]}} \\
&\leq \frac{\mathbf{E}[e^{tY}]}{e^{t(1+\delta)\mathbf{E}[Y]}} \quad \text{(by the result of (b))} \\
&\leq \frac{(pe^t + (1-p))^n}{e^{t(1+\delta)np}} \\
&\leq \frac{e^{np(e^t-1)}}{e^{t(1+\delta)np}}.
\end{aligned}$$

Let $t = \ln(1+\delta)$, we have the following Chernoff-like bound:

$$\mathbf{Pr}[X \geq (1+\delta)\mathbf{E}[X]] \leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^{\mathbf{E}[X]}.$$

□

Exercise 5.18. *An undirected graph on n vertices is* disconnected *if there exists a set of $k < n$ vertices such that there is no edge between this set and the rest of the graph. Otherwise, the graph is said to be* connected. *Show that there exists a constant c such that if $N > cn \log n$ then, with probability $O(e^{-n})$, a graph randomly chosen from $G_{n,N}$ is connected.*

Solution. The problem can be viewed as *throwing $2N$ balls into n bins*, where balls and bins are edges and vertices of $G_{n,N}$ respectively. However, since each edge has two endpoints, each edge is like throwing two balls at once into two different bins.

Let A_i be the event that vertex i is disconnected from the other $n-1$ vertices, then from the balls-and-bins model of analysis, we have

$$\mathbf{Pr}[A_i] = \left(1 - \frac{(n-1)}{\binom{n}{2}}\right)^N = \left(1 - \frac{2}{n}\right)^N.$$

What we want to have is actually the probability $\mathbf{Pr}[\overline{\bigcup_{i\in V} A_i}]$, which can be calculated as follows.

$$\begin{aligned}
\mathbf{Pr}\left[\overline{\bigcup_{i\in V} A_i}\right] &= \mathbf{Pr}[\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_n}] \\
&\leq \mathbf{Pr}[\overline{A_1}] \\
&= 1 - \left(1 - \frac{2}{n}\right)^N \\
&\leq 1 - \left(1 - \frac{2}{n}\right)^{cn\log n} \\
&\approx 1 - e^{-2c\ln n} \text{ (we assume that } \log n = \ln n) \\
&= 1 - n^{-2c} \\
&\leq e^{-n^{-2c}}.
\end{aligned}$$

Actually, there might be something missing or wrong with this exercise. Note that we cannot obtain

$$\mathbf{Pr}[\overline{A_1} \cap \overline{A_2} \cap \ldots \cap \overline{A_n}] = \prod_{i=1}^{n} \mathbf{Pr}[\overline{A_i}],$$

since A_i's are not mutually independent! The following are derived the comments from Dr. Ton Kloks.

Since $N = cn\log n$, we may assume that each edge of G appears with probability $p = cn\log n/\binom{n}{2}$. $\mathbf{Pr}$[The probability that G is disconnected] $\geq \mathbf{Pr}$[some vertex v is isolated] $= q^{n-1}$. Then we have $\mathbf{Pr}[G \text{ is disconnected}] \geq (1-p)^n = n^{-c'} \cdot (1 + O(\log n/n))$ for some constant c'. But the statements of the exercise makes no sense. If N is big enough, say $\Theta(n^2)$, then almost surely G is connected, but it says that the probability goes to zero when n gets big enough.

Please refer to Erdős and Rényi's paper [1], which shows that G is connected with probability $e^{-e^{-2y}}$ when n approaches to infinity, where $N = (n/2)\log n + yn + o(n)$. Taking $y = -c\log n$ we get that, for $N = (1/2-c)n\log n$, G is connected with probability $e^{-n^{2c}}$. Obviously we need $c < 1/2$.

In addition, please also refer to Gilbert's result [2], which shows that a graph is connected with probability EXACTLY $1 - nq^{n-1} + O(n^2 q^{3n/2})$, where p is the edge probability and $q = 1-p$. Yet by further calculations, we still cannot obtain the desired probability $O(e^{-n})$. obtained. □

References

[1] P. Erdős and A. Rényi: The Evolution of Random Graphs". *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960) 17–61.

[2] E. N. Gilbert: Random Graphs. *Annals of Mathematical Statistics* **30** (1959) 1141–1144.

[3] M. Mitzenmacher and E. Upfal: *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.