CIS 786: Approximation and Randomized Algorithms September 24, 2004

Lecture 4

Lecturer: Artur Czumaj Scribe: Nitesh B. Guinde

Disclaimer: *These notes have not been subjected to the usual scrutiny for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Balls and bins games

Many problems in randomized algorithms can be described in a framework of *balls and bins games* (see, e.g., [CLRS01, Chapter 5.4], [MR95, Chapter 3.1], [MRS01], and [RS98]). In this framework, one is aiming at distributing the balls in a set of bins according to certain rules, and in order to minimize certain parameters. In this lecture we will present some such games in details.

Consider a scenario where we throw balls into bins randomly (independently and uniformly at random, i.u.r.). Let m be the number of balls and n be the number of bins.

We throw m balls into n bins.
$\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\ldots\ldots\bullet\bullet\bullet\bullet\bullet\bullet \to m$ balls
$\sqcup\sqcup\sqcup\sqcup\sqcup\sqcup\ldots\ldots\ldots\ldots\sqcup\sqcup\sqcup\sqcup \to n$ bins

We are interested in finding the distribution of the load in the bins, where the *load of a bin* is the number of balls in the bin. It is possible to find a quite precise characterization of the entire probability distribution of the loads of the bins, but in this lecture we will focus our attention only on the most natural or the most important characterization. We will study solutions to the following questions:

(a) *How many bins are empty?*

(b) *How many balls we have to throw or are needed to have all the bins non-empty?*

(c) *How many balls needed to have a bin with load at least 2?*

(d) *What is the maximum load?*

1.1 Number of empty bins

Since balls are thrown randomly there is a possibility of some bins not containing the ball after all the m balls have been thrown. We have to find out the expected number of bins which will remain empty with reasonably high probability. Let X be a random variable indicating the number of empty bins, that is,

$$X_i = \begin{cases} 1 & \text{if bin i is empty;} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\mathbf{E}[X_i] \;=\; \mathbf{Pr}[\text{bin } i \text{ is empty}] \;=\; \prod_{j=1}^{n} \left(1 - \mathbf{Pr}[\text{ball } j \text{ went into bin } i]\right) \;=\; \left(1 - \frac{1}{n}\right)^m \quad .$$

This yields,

$$\mathbf{E}[X] \;=\; \sum_{i=0}^{n} \mathbf{E}[X_i] = n \cdot \left(1 - \frac{1}{n}\right)^m \quad .$$

Observe also that we could use the bounds above together with the union bound to obtain,

$$\mathbf{Pr}[\exists \text{ empty bin}] \;\le\; \sum_{i=1}^{n} \mathbf{Pr}[\text{bin } i \text{ empty}] \;\le\; n \cdot \left(1 - \frac{1}{n}\right)^m \;\le\; n \cdot e^{-\frac{m}{m}} \quad .$$

It should be noted that for $m \approx n$ we have $\mathbf{Pr}[\exists \text{ empty bin}] \sim \frac{n}{e}$, which gives probability greater than 1, and thus the bound is useless in this case. Therefore, in order to use this bound, we need to have $n \cdot e^{-\frac{m}{m}} \ll 1$, in which case we must have $m \gg n \cdot \ln n$. On the other hand, we see from the bound above that if $m \gg n \cdot \ln n$ then the probability that these exists an empty bin is very small. (For more discussion, see e.g., [MR95, Chapter 3.1].)

1.2 Number of balls needed to have all bins non-empty

This is identical to the *coupon collector problem* (see, e.g., [MR95, Chapter 3.6.2] or [CLRS01, Chapter 5.4.2]): for a given set of coupons of n types how many times we have to choose the coupons (we do it at random and with replacement) to collect coupons of every type. In this setting, the bins can be thought as of coupons and chosen coupons to be the bins. The time when all the coupons (types) are collected is identical to the event that all the n bins are non empty.

How many trials are needed to achieve this goal is to be found. Let X be the random variable which denotes the number of steps (number of balls thrown into the bins) needed to get at least one ball in each of the n bins, that is, X be a r.v. indicating the number of steps so that every bin has a ball. Our goal is to find $\mathbf{E}[X]$ and $\mathbf{Var}[X]$.

Let Y_i denote the number of balls taken to get for the first time i bins non-empty. Observe that $X = Y_n$.

Let us define random variables $X_1, X_2, \ldots, X_n$ such that $X_1 = Y_1$ and $X_i = Y_i - Y_{i-1}$ for $i \ge 2$. In other words, X_i denotes the number of balls we have to throw to get i non-empty bins if we start at the state when $i-1$ bins are already non-empty. Observe that

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

Therefore our goal is to analyze the expected value of $\sum_{i=1}^{n} X_i$.

Let us consider a fixed i and suppose that we already threw the balls so that exactly $i-1$ bins are non-empty. Then, the value of X_i is equal to the number of balls we have to throw to obtain a new non-empty bin. We observe that we will succeed in a single step, if the first new selected ball will choose an empty bin. Since we have already $i-1$ non-empty bins, the probability that will happen is equal to $1 - \frac{i-1}{n}$. Similarly, we will succeed in two steps, if we do not succeed in the first step and we succeed in the second step. The probability that this will happen is equal to $\frac{i-1}{n} \cdot (1 - \frac{i-1}{n})$. We can generalize this observation to arbitrary number of steps k, to obtain,

$$\mathbf{Pr}[X_i = k] = \left(\frac{i-1}{n}\right)^{k-1} \cdot \left(1 - \frac{i-1}{n}\right) \ .$$

Such a probability distribution of X_i is called a *geometric distribution* with the success probability $p = 1 - \frac{i-1}{n}$ (see Appendix A.3). Since we know that for a random variable Z with geometric distribution with the success probability p we have $\mathbf{E}[Z] = \frac{1}{p}$ and $\mathbf{Var}[Z] = \frac{1-p}{p^2}$, we obtain

$$\mathbf{E}[X_i] = \frac{n}{n-i+1} \qquad \text{and} \qquad \mathbf{Var}[X_i] = \frac{n \cdot (i-1)}{(n-i+1)^2} \ ,$$

and therefore,

$$\mathbf{E}[X] \ = \ \sum_{i=1}^{n} \mathbf{E}[X_i] \ = \ \sum_{i=1}^{n} \frac{n}{n-i+1} \ = \ n \cdot \sum_{i=1}^{n} \frac{1}{n-i+1} \ = \ n \cdot \sum_{j=1}^{n} \frac{1}{j} \ = \ n \cdot H_n \ = \ n \cdot (\ln n + o(1)) \ .$$

Similarly, since the random variables $X_1, X_2, \dots, X_n$ are independent, we obtain

$$\begin{aligned} \mathbf{Var}[X] \ &= \ \sum_{i=1}^{n} \mathbf{Var}[X_i] \ = \ \sum_{i=1}^{n} \frac{n \cdot (i-1)}{(n-i+1)^2} \ = \ \sum_{j=1}^{n} \frac{n \cdot (n-j)}{j^2} \ = \ \sum_{j=1}^{n} \frac{n^2}{j^2} - \sum_{j=1}^{n} \frac{n}{j} \\ &= \ n^2 \cdot \sum_{j=1}^{n} \frac{1}{j^2} - n \cdot H_n \ . \end{aligned}$$

Now, since it is known that $\sum_{j \ge 1} \frac{1}{j^2} = \frac{\pi^2}{6}$, we obtain

$$\mathbf{Var}[X] \ = \ n^2 \cdot \sum_{j=1}^{n} \frac{1}{j^2} - n \cdot H_n \ \approx \ \frac{\pi^2}{6} \cdot n^2 - n \cdot (\ln n + o(1)) \ = \ \frac{\pi^2}{6} \cdot n^2 + o(n^2) \ .$$

Now, we can use Chebyshev inequality to bound the probability that $X \gg n \ln n$ or $X \ll n \ln n$. For any λ, since $\mathbf{Var}[X] = \frac{\pi^2}{6} \cdot n^2 + o(n^2)$, if we set $t = \frac{1}{\sqrt{\lambda}}$, then by Chebyshev inequality we obtain,

$$\mathbf{Pr}[|X - n \cdot \ln n| \ge (\lambda + \mathcal{O}(1)) \cdot n] \ \le \ \mathbf{Pr}[|X - \mathbf{E}[X]| \ge t \cdot \sqrt{\mathbf{Var}[X]}] \ \le \ \frac{1}{t^2} \ = \ \frac{1}{\lambda} \ .$$

This implies that the value of X is highly concentrated around its mean.

1.3 Number of bins with load greater than 1

This problem is traditionally called the *birthday problem* (where the 365 days correspond to the 365 bins and birthday of each of the n people is chosen independently and uniformly from all the 365 days ignoring leap years). The goal is to find number of days on which more than 1 person have same birthday, i.e., the number of bins with load greater than 1. (See also [CLRS01, Chapter 5.4.1]).

Let X be the random variable denoting the number of bins with load greater than or equal to 2. It is easy to see that

$$X = \sum_{i=1}^{n} X_i \ ,$$

where

$$X_i = \begin{cases} 1 & \text{if bin } i \text{ has at least 2 balls ;} \\ 0 & \text{otherwise.} \end{cases}$$

Observe that,

$$\mathbf{Pr}[X_i = 1] = \mathbf{Pr}[\text{bin } i \text{ has at least 2 balls}] = 1 - \mathbf{Pr}[X_i \text{ has either 1 or 0 balls}] \ .$$

We analyze the probabilities by observing that the load of (number of balls in) every bin is a random variable having binomial distribution (see Appendix A.2). The probability that a bin has exactly k balls is equal to

$$\binom{m}{k} \cdot \left(\frac{1}{n}\right)^k \cdot \left(1 - \frac{1}{n}\right)^{m-k} \ .$$

Therefore we obtain

$$\begin{aligned} \mathbf{Pr}[X_i = 1] &= 1 - \mathbf{Pr}[X_i \text{ has either 1 or 0 balls}] \\ &= 1 - \binom{m}{1} \cdot \left(\frac{1}{n}\right)^1 \cdot \left(1 - \frac{1}{n}\right)^{m-1} - \binom{m}{0} \cdot \left(\frac{1}{n}\right)^0 \cdot \left(1 - \frac{1}{n}\right)^m \\ &= 1 - \frac{m}{n} \cdot \left(1 - \frac{1}{n}\right)^{m-1} - \left(1 - \frac{1}{n}\right)^m \ . \end{aligned}$$

Now, we will use some basic approximations, where we will assume that $n \gg m$.

$$\begin{aligned} \mathbf{Pr}[X_i = 1] &= 1 - \frac{m}{n} \cdot \left(1 - \frac{1}{n}\right)^{m-1} - \left(1 - \frac{1}{n}\right)^m \approx 1 - \frac{m}{n} \cdot \left(1 - \frac{1}{n}\right)^m - \left(1 - \frac{1}{n}\right)^m \\ &= 1 - \left(1 + \frac{m}{n}\right) \cdot \left(1 - \frac{1}{n}\right)^m \approx 1 - \left(1 + \frac{m}{n}\right) \cdot \left(1 - \frac{m}{n}\right) \\ &= 1 - \left(1 - \left(\frac{m}{n}\right)^2\right) = \left(\frac{m}{n}\right)^2 \ . \end{aligned}$$

By our discussion above, this implies the following:

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{Pr}[X_i = 1] \approx n \cdot \left(\frac{m}{n}\right)^2 = \frac{m^2}{n} \ .$$

Notice that this implies that if the number of balls is greater than $\sqrt{n}$, then we expect to see at least bin with more than one ball.

One could obtain a similar result by analyzing another random variable, that we denote by Y, which is equal to the number of *conflicts*, that is, the number of pairs of balls that are in the same bin. It is easy to see that if we define $Y_{i,j}$ such that

$$Y_{i,j} = \begin{cases} 1 & \text{balls } i \text{ and } j \text{ are at the same bin and } i \neq j \\ 0 & \text{otherwise,} \end{cases}$$

then $Y = \sum_{i=1}^{m} \sum_{j=i+1}^{m} Y_{i,j}$. Furthermore, it is easy to see that the probability that two balls will be allocated to the same bin is exactly $\frac{1}{n}$. Therefore,

$$\mathbf{E}[Y] \ = \ \sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathbf{E}[Y_{i,j}] \ = \ \binom{m}{2} \cdot \frac{1}{n} \ = \ \frac{m \cdot (m-1)}{2 \cdot n} \ .$$

Also this result implies that if the number of balls is greater than $\sqrt{n}$, then we expect to see at least one bin with more than one ball.

1.4 Maximum Load

The problems discussed before have been extensively studied in mathematics for many decades, while the problem of analyzing the maximum load has been extensively studied only very recently, because of its many applications in computer science. Indeed, the problem of finding maximum load has many applications in hashing, online load balancing, data allocation, routing, etc.

Hashing: This problem may be modelled to analyze the efficiency of hashing-algorithms. In the case of separate chaining in hashing, all keys that hash to the same location in the table are stored in a linked list. It is clear that the lengths of these lists are a measure for the complexity. For a well chosen hash-function (i.e. a hash-function which assigns the keys to all locations in the table with the same probability), the lengths of the lists have exactly the same distribution as the number of balls in a bin.

For more information about hashing, see [CLRS01, Chapter 11] and [MR95, Chapter 8.4].

Online Load Balancing: In parallel and distributed computing the load balancing problem is the following scenario: consider n database-servers and m requests which arise independently at different clients and which may be handled by any server. The problem is to assign the requests to the servers in such a way that all servers handle about the same number of requests.

For more information about this and similar applications, see, e.g., [ABKU99, SEK00].

1.4.1 Analysis of the max-load

Once we have seen that the problem of estimating maximum-load has many applications, let us begin the analysis of the max-load. Let X_i denote the load of bin i, $1 \le i \le n$. Our goal is to analyze $\max_{1 \le i \le n} X_i$.

Of course, we have

$$\mathbf{E}[X_i] = \frac{m}{n} \ .$$

Furthermore, we can write X_i as the sum of random variables $Y_1 + \dots + Y_m$, where

$$Y_i = \begin{cases} 1 & \text{if ball } j \text{ is in bin } i; \\ 0 & \text{otherwise.} \end{cases}$$

Since by the union bound, for every $L \ge 0$ we have

$$\mathbf{Pr}[\text{max-load } \ge L] \ = \ \mathbf{Pr}[\exists_{i:1 \le i \le n} \ X_i \ge L] \ \le \ \sum_{i=1}^{n} \mathbf{Pr}[X_i \ge L] \ ,$$

in order to estimate the probability that the maximum load is larger than some value, it is enough to focus attention on a single bin and show that for this bin, the probability that its load is greater than L is very small.

We will consider two main cases in our analysis, depending on whether $m \ge 2\,n \log_2 n$ or not.

Heavy loaded case: $m \ge 2\,n \log_2 n$.

We will analyze the probability distribution of the load of a fixed bin i using Chernoff bound. By Chernoff bound, we have

$$\mathbf{Pr}\big[X_i \ge (1+\delta) \cdot \mathbf{E}[X_i]\big] \le \Big(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\Big)^{\frac{m}{n}} \ .$$

Let us assume that $\delta \ge e - 1$. Then, we have,

$$\frac{e^\delta}{(1+\delta)^{(1+\delta)}} \ = \ \frac{1}{1+\delta} \cdot \left(\frac{e}{1+\delta}\right)^\delta \ \le \ \frac{1}{1+\delta} \ \le \ \frac{1}{2} \ .$$

Therefore, we have

$$\mathbf{Pr}\big[X_i \ge (1+\delta) \cdot \mathbf{E}[X_i]\big] \ \le \ \Big(\frac{e^\delta}{(1+\delta)^{1+\delta}}\Big)^{\frac{m}{n}} \ \le \ 2^{-\frac{m}{n}} \ \le \ \frac{1}{n^2} \ ,$$

where the last inequality holds for all $m \ge 2\,n \log_2 n$.

Therefore, we use the inequality above together with the union bound to obtain,

$$\mathbf{Pr}\big[\text{max-load } \ge \frac{e \cdot m}{n}\big] \ = \ \mathbf{Pr}\big[\exists_{i:1 \le i \le n} \ X_i \ge \frac{e \cdot m}{n}\big] \ \le \ \sum_{i=1}^{n} \mathbf{Pr}\big[X_i \ge \frac{e \cdot m}{n}\big] \ \le \ n \cdot \frac{1}{n^2} \ = \ \frac{1}{n} \ .$$

Hence, we have proven the following theorem.

Theorem 1 *For every $m \ge 2n \log_2 n$, if we throw m balls into n bins i.u.r., then the maximum load of the bin is smaller than $\frac{e \cdot m}{n}$ with probability at least $1 - \frac{1}{n}$.*

Actually, one could obtain a stronger bound, that the maximum load is $\frac{m}{n} + \Theta(\sqrt{\frac{m \cdot \log n}{n}})$ with high probability.

Lightly loaded case: $m < 2n \log_2 n$.

Now, we analyze the other case, when m is not much bigger than n. Since X_i is the sum of m independent 0-1 random variables, it has binomial probability distribution with parameters m and $1/n$, which we denote by $\mathbb{B}(m, 1/n)$. Furthermore, let us observe that for any k, $0 \le k \le m$, we have

$$\mathbf{Pr}\big[\mathbb{B}(m, \tfrac{1}{n}) \ge k\big] \;\le\; \binom{m}{k} \cdot \frac{1}{n^k} \;\;.$$

To see this, we notice that in order to have at least k successes in the binomial trials, we have to select k moments where the success will happen, and the probability that we will have successes in all these moments is n^{-k}. (Observe that this leads to the inequality rather then the equality.)

Therefore, we have[1]

$$\mathbf{Pr}[X_i \ge k] \;=\; \mathbf{Pr}\big[\mathbb{B}\left(m, \tfrac{1}{n}\right) \ge k\big] \;\le\; \binom{m}{k} \cdot \frac{1}{n^k} \;\le\; \left(\frac{m\,e}{k}\right)^k \cdot \frac{1}{n^k} \;\le\; \left(\frac{m\,e}{n\,k}\right)^k \;\;.$$

Our goal is to obtain $\mathbf{Pr}[X_i \ge k] \le \frac{1}{n^2}$ and we want to use the fact (used in previous lecture) that if $K \ge 2 \ln N / \ln\ln N$ then $K^K \ge N$ (this also requires that $N > e$, because otherwise $\ln\ln N$ could be either zero or even could be negative). Therefore, since

$$\left(\frac{m\,e}{n\,k}\right)^k \;\le\; \frac{1}{n^2}$$

is equivalent to

$$\left(\frac{n\,k}{m\,e}\right)^{\frac{n\,k}{m\,e}} \;\ge\; n^{\frac{2\,n}{m\,e}} \;\;,$$

we have that (for $m < \frac{2}{e} n \ln n$) if

$$\frac{n\,k}{m\,e} \;\ge\; \frac{2 \cdot \ln\left(n^{\frac{2\,n}{m\,e}}\right)}{\ln\ln\left(n^{\frac{2\,n}{m\,e}}\right)} \;=\; \frac{4\,n}{m\,e} \cdot \frac{\ln n}{\ln\left(\frac{2\,n}{m\,e} \ln n\right)} \;\;,$$

or in other words, if

$$k \;\ge\; \frac{4 \cdot \ln n}{\ln\left(\frac{2\,n}{m\,e} \ln n\right)} \;\;,$$

[1] Exercise: try to prove that for every $N \ge K$, we have $(N/K)^K \le \binom{N}{K} \le (N\,e/K)^K$; this bound is used here.

then

$$\mathbf{Pr}[X_i \geq k] \;\leq\; \frac{1}{n^2} \enspace .$$

Thus, since by the union bound we have,

$$\mathbf{Pr}[\exists_{1 \leq i \leq n}\ X_i \geq k] \;\leq\; \sum_{i=1}^{n} \mathbf{Pr}[X_i \geq k] \;\leq\; n \cdot \frac{1}{n^2} \;=\; \frac{1}{n} \enspace .$$

we can conclude with the following theorem.

Theorem 2 *For every $m < \frac{2}{e} n \ln n$, if we throw m balls into n bins i.u.r., then the maximum load of the bin is with probability at least $1 - \frac{1}{n}$ smaller than*

$$\frac{4 \cdot \ln n}{\ln\left(\frac{2\,n}{m\,e} \ln n\right)} \enspace .$$

Other analysis of the lightly loaded case — using Chernoff bound

Actually, one could also use Chernoff bound to prove the result almost identical to Theorem 2.

Let us remind the Chernoff bound, that for a set of m independent 0-1 random variables $X_1, \dots, X_m$, if $X = \sum_{i=1}^{m} X_i$ and we denote by $\mu = \mathbf{E}[X]$, then for every $\delta > 0$, we have

$$\mathbf{Pr}[X > (1+\delta) \cdot \mathbf{E}[X]] \;\leq\; \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \enspace .$$

Therefore, if we fix a single bin j, and we denote by X_i the indicator random variable whether ball i is allocated to bin j, then we have $X = \sum_{i=1}^{m} X_i$ to denote the load of bin j. Therefore, we can use the Chernoff bound above (with $\tau = 1 + \delta$) to obtain

$$\mathbf{Pr}[\text{load of bin } j \text{ is greater than } \tau \cdot \mu] \quad\leq\quad \left(\frac{e^{\tau-1}}{\tau^{\tau}}\right)^{\mu} \quad<\quad \left(\frac{e}{\tau}\right)^{\tau \cdot \mu} \enspace .$$

Next, we observe that our goal is to prove that the probability that the load of bin j is greater than $\tau \cdot \mu$ is smaller than $\frac{1}{n^2}$. To prove this, it is enough to choose τ such that $\left(\frac{e}{\tau}\right)^{\tau \cdot \mu} \leq \frac{1}{n^2}$, or equivalently, that $n^2 \leq \left(\frac{\tau}{e}\right)^{\tau \cdot \mu}$. Since the last inequality is equivalent to $n^{2\,e/\mu} \leq (\tau/e)^{\tau/e}$, and since we know that if $K \geq 2 \cdot \ln N / \ln \ln N$ then $K^K \geq N$, then we can choose (provided that $m < 2\,e\,n\, \ln n$)

$$\tau \;\geq\; \frac{4\,e^2}{\mu} \cdot \frac{\ln n}{\ln(\frac{2\,e\,n}{m} \cdot \ln n)} \enspace ,$$

to conclude that

$$\mathbf{Pr}[\text{load of bin } j \text{ is greater than } \tau \cdot \mu] \;\leq\; \frac{1}{n^2} \enspace ,$$

or in other words, that

$$\mathbf{Pr}\Big[\text{load of bin } j \text{ is greater than } 4 \cdot e^2 \cdot \frac{\ln n}{\ln(\frac{2\,e\,n}{m} \cdot \ln n)}\Big] \;\leq\; \frac{1}{n^2} \enspace .$$

This, by the union bound, implies that for $m < 2\,e\,n\, \ln n$, the maximum load is upper bounded by $4 \cdot e^2 \cdot \frac{\ln n}{\ln(\frac{2\,e\,n}{m} \cdot \ln n)}$ with probability at least $1 - 1/n$.

2 Choosing two random bins for each ball

Until now we were randomly selecting one bin. We will now analyze an algorithm which gives *two random choices* of bins (see, e.g., [ABKU99, BCSV00, CRS04, MRS01, Vöc99]).

The basic idea is based on the observation that if for every ball we will select more than one bin, then it might be possible to improve the maximum load with respect to the process studied in the previous section.

2.1 Online variant

Consider an online process where instead of one random choice, for every ball we try to get two random choices. We sequentially allocate balls into bins, and for each ball, we select two bin i.u.r. and we allocate it in the bin (out of the two selected once) with smaller load at the given moment. If the loads of the two bins are the same, then the tie is resolved arbitrarily.

Azar et. al. [ABKU99] proved that after allocating n balls in the n bins, the maximum load in the bins will be $\log_2 \log_2 n + \Theta(1)$ with high probability, see also [MRS01]. Later, in [BCSV00], it was shown that if one allocates m balls in the n bins, where m might be arbitrarily big, then the maximum load is $\frac{m}{n} + \log_2 \log_2 n + \Theta(1)$ with high probability.

What's when more then two random choices are done. If we do more than two random choices then we also obtain an improvement, but the improvement is very little: for d random choices, $d > 1$, the maximum load is $\frac{m}{n} + \log_d \log_2 n + \Theta(1)$ with high probability.

Application to hashing. If we apply this to hashing with chaining with two (random) hash functions then we obtain a static dictionary that performs the search in $\mathcal{O}(\log \log n)$ time with high probability (see [ABKU99] for more details).

2.2 Always go left — Vöcking's modification

Vöcking [Vöc99] suggested the following improvement to the online variant presented above. Instead of choosing two random bins from all the bins we divide the set of bins into two halves (left and right) and then choose a random bin from each of these halves. Now selection of the bin is done by comparing the left half and the right half. The bin from the lesser weighed half is chosen by the ball. In case the weights of the two halves are equal then the *left half is chosen.* Rather surprisingly, an improvement on the max load is obtained: the max load in the bin is further decreased to

$$\frac{\ln \ln n}{2 \cdot \ln(\frac{1+\sqrt{5}}{2})} + \Theta(1) \;\approx\; 0.72 \; \log_2 \log_2 n + \Theta(1)$$

with high probability. (In [BCSV00] this analysis is extended to the case when m balls are allocated, and then the maximum load is $\frac{m}{n} + 0.72 \; \log_2 \log_2 n + \Theta(1)$ with high probability.)

2.3 Off-line variant

The on-line variant of the process discussed above is certainly easy to implement and has found many applications (see e.g., [ABKU99, MRS01]). However, for some applications, there is no need to make the decision online. Now, we will discuss the off-line version of this scheme.

We consider the scheme in which at the beginning for each ball we choose two bins i.u.r. Then, by looking at all the balls choices, we allocate each bin to one of the selected bins to minimize the maximum-load.

Sanders et al. [SEK00] showed that a maximum-flow approach can be used to find an allocation that minimizes the maximum load; the running time is $\mathcal{O}(m\,n\,\log(n\,m))$. Furthermore, two probabilistic results about the value of the maximum load can be shown:

- If $m > c\,n\,\log n$ for certain constant c, then with high probability it is possible to allocate the balls so that the maximum load is $\lceil \frac{m}{n} \rceil$ which is a *perfect allocation* [CRS04]. (It is worth to notice that such a result is impossible for $m = o(n \ln n)$.)
- For every n and m, with high probability one can find an allocation of the balls so that the maximum load is $\lceil \frac{m}{n} \rceil + 1$; this is at most 1 off the perfect allocation [SEK00] (see also [ABKU99] and [CRS04]).

Application to hashing. These results for the off-line allocations have also immediate application to static dictionaries that uses the hashing with chaining approach. Let us assume that we want to store n elements in a hash table of length n. For each element in the dictionary, we use two random hash functions. By looking at the locations of all hash functions, we can find in polynomial time allocation that ensures that each cell in the has table will have the chain of length 2, with high probability. Therefore, this scheme leads to the static dictionary with search time $\mathcal{O}(1)$, with high probability (in fact, at most 4 look-ups are need, with high probability).

Some more applications are discussed in [SEK00].

References

[ABKU99] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal. Balanced allocations. *SIAM Journal on Computing*, 29(1): 180–200, 1999.

[BCSV00] P. Berenbrink, A. Czumaj, A. Steger, and B. Vöcking. Balanced allocations: The heavily loaded case. *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC'00)*, pp. 745–754, 2000.

[CLRS01] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms.* 2nd edition, the MIT Press, Cambridge, MA, 2001.

[CRS04] A. Czumaj, C. Riley, and C. Scheideler. Perfectly balanced allocation. *Proceedings of the 7th International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM'03)*, pp. 240–251, 2003.

[Leo??] A. Leon-Garcia. *Probability and Random Processes.*

[MRS01] M. Mitzenmacher, A. W. Richa, and R. Sitaraman. The power of two random choices: A survey of techniques and results. In *Handbook of Randomized Computing*, Rajasekaran et al., eds., Volume I, pp. 255-312, Kluwer Academic Press, 2001.

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

[RS98] M. Raab and A. Steger. "Balls into bins" — a simple and tight analysis. *Proceedings of the 2nd International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM'98)*, pp. 159–170, 1998.

[SEK00] P. Sanders, S. Egner, and J. Korst. Fast concurrent access to parallel disks. *Algorithmica*, 35(1): 21–55, 2003.

[Vöc99] B. Vöcking. How asymetry helps load balancing. *Journal of the ACM*, 50(4):568–589, 2003

Appendix

A Basic probability distributions

A.1 Bernoulli distribution

Bernoulli probability distribution is used to describe 0-1 random variables. For example, it might describe the random variable corresponding to coin tossing: either we will obtain HEAD or TAIL.

Let X be a random variable. Let p be the probability that $X = 1$ and let the probability of getting a 0 be $1 - p$. The expected value and the variance of X are

$$\begin{aligned} \mathbf{E}[X] &= p \cdot 1 + (1-p) \cdot 0 = p \ , \\ \mathbf{Var}[X] &= p \cdot (1-p) \ . \end{aligned}$$

A.2 Binomial distribution

Binomial probability distribution is used to describe the outcome of multiple Bernoulli trials. Let $n \in \mathbb{N}$ and let p be $0 \le p \le 1$. A random variable X that describes the number of successes (getting 1s) in n independent Bernoulli trials with success probability p each is called to have *binomial distribution with parameters n and p.*

For example, if we toss a coin n times, and in each trial we have the probability that we will get a HEAD equal to p and we will get TAIL with probability $1-p$, then the random variable describing the number of HEADs has binomial distribution with parameters n and p.

It is easy to see that

$$\mathbf{Pr}[X = k] \ = \ \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \qquad \text{for all } k = 0, 1, \dots \ .$$

It is also known that

$$\begin{aligned} \mathbf{E}[X] &= n\,p \ , \\ \mathbf{Var}[X] &= n\,p\,(1-p) \ . \end{aligned}$$

A.3 Geometric distribution

Geometric probability distribution corresponds to the "waiting time" for the first success. For example, if we toss a coin that in each trial will have the probability of getting a HEAD equal to p and of getting a TAIL $1 - p$, then the random variable describing the number of tosses needed to obtain the first HEAD has geometric distribution with parameter p.

For any p, $0 < p \le 1$, we say a random variable X has *geometric distribution with parameter p*, if

$$\mathbf{Pr}[X = k] \ = \ p \cdot (1-p)^{k-1} \qquad \text{for all } k = 0, 1, \dots \ .$$

It is also known that

$$\mathbf{E}[X] \ = \ \frac{1}{p} \qquad \text{and} \qquad \mathbf{Var}[X] \ = \ \frac{1-p}{p^2} \ .$$

To see the first claim, that $\mathbf{E}[X] = \frac{1}{p}$, it is enough to do some simple algebraic transformations:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{k=1}^{\infty} k \cdot \mathbf{Pr}[X = k] = \sum_{k=1}^{\infty} k \cdot p \cdot (1-p)^{k-1} = \sum_{k=1}^{\infty} \sum_{i=1}^{k} p \cdot (1-p)^{k-1} \\
&= \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} p \cdot (1-p)^{k-1} = \sum_{i=1}^{\infty} p \cdot (1-p)^{i-1} \sum_{s=0}^{\infty} (1-p)^{s} \\
&= \sum_{i=1}^{\infty} p \cdot (1-p)^{i-1} \cdot \frac{1}{1-(1-p)} = \sum_{i=1}^{\infty} (1-p)^{i-1} = \sum_{j=0}^{\infty} (1-p)^{j} = \frac{1}{1-(1-p)} \\
&= \frac{1}{p} \; .
\end{aligned}$$