.1. *Limit of the Binomial Distribution*

have shown that, when throwing m balls randomly into b bins, the probability p_r a bin has r balls is approximately the Poisson distribution with mean m/b. In gen-, the Poisson distribution is the limit distribution of the binomial distribution with ameters n and p, when n is large and p is small. More precisely, we have the fol-ing limit result.

eorem 5.5: *Let X_n be a binomial random variable with parameters n and p, where a function of n and $\lim_{n\to\infty} np = \lambda$ is a constant that is independent of n. Then, any fixed k,*

$$\lim_{n\to\infty} \Pr(X_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

s theorem directly applies to the balls-and-bins scenario. Consider the situation ere there are m balls and n bins, where m is a function of n and $\lim_{m\to\infty} m/n = \lambda$. X_m be the number of balls in a specific bin. Then X_m is a binomial random variable h parameters m and $1/n$. Theorem 5.5 thus applies and says that

$$\lim_{m\to\infty} \Pr(X_m = r) = \frac{e^{-m/n}(m/n)^r}{r!},$$

tching the approximation of Eqn. (5.2).

Before proving Theorem 5.5, we describe some of its applications. Distributions of type arise frequently and are often modeled by Poisson distributions. For exam-consider the number of spelling or grammatical mistakes in a book, including this k. One model for such mistakes is that each word is likely to have an error with ne very small probability p. The number of errors is then a binomial random vari-e with large n and small p that can therefore be treated as a Poisson random variable. another example, consider the number of chocolate chips inside a chocolate chip kie. One possible model is to split the volume of the cookie into a large number of ll disjoint compartments, so that a chip lands in each compartment with some prob-ity p. With this model, the number of chips in a cookie roughly follows a Poisson ribution. We will see similar applications of the Poisson distribution in continuous ings in Chapter 8.

of of Theorem 5.5: We can write

$$\Pr(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$$\Pr(X_n = k) \le \frac{}{k!} p^k \frac{}{(1-p)^k}$$
$$\le \frac{(np)^k}{k!} \frac{e^{-pn}}{1-pk}$$
$$= \frac{e^{-pn}(np)^k}{k!} \frac{1}{1-pk}.$$

The second line follows from the first by Eqn. (5.3) and the fact that $(1-p)^k \ge 1-pk$ for $k \ge 0$. Also,

$$\Pr(X_n = k) \ge \frac{(n-k+1)^k}{k!} p^k (1-p)^n$$
$$\ge \frac{((n-k+1)p)^k}{k!} e^{-pn}(1-p^2)^n$$
$$\ge \frac{e^{-pn}((n-k+1)p)^k}{k!}(1-p^2 n),$$

where in the second inequality we applied Eqn. (5.3) with $x = -p$.

Combining, we have

$$\frac{e^{-pn}(np)^k}{k!} \frac{1}{1-pk} \ge \Pr(X_n = k) \ge \frac{e^{-pn}((n-k+1)p)^k}{k!}(1-p^2 n).$$

In the limit, as n approaches infinity, p approaches zero because the limiting value of pn is the constant λ. Hence $1/(1-pk)$ approaches 1, $1-p^2 n$ approaches 1, and the difference between $(n-k+1)p$ and np approaches 0. It follows that

$$\lim_{n\to\infty} \frac{e^{-pn}(np)^k}{k!} \frac{1}{1-pk} = \frac{e^{-\lambda}\lambda^k}{k!}$$

and

$$\lim_{n\to\infty} \frac{e^{-pn}((n-k+1)p)^k}{k!}(1-p^2 n) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Since $\lim_{n\to\infty} \Pr(X_n = k)$ lies between these two values, the theorem follows. ■

5.4. The Poisson Approximation

The main difficulty in analyzing balls-and-bins problems is handling the dependencies that naturally arise in such systems. For example, if we throw m balls into n bins and find that bin 1 is empty, then it is less likely that bin 2 is empty because we know that the m balls must now be distributed among $n-1$ bins. More concretely: if we know

the number of balls in the first $n-1$ bins, then the number of balls in the last bin is completely determined. The loads of the various bins are not independent, and independent random variables are generally much easier to analyze, since we can apply Chernoff bounds. It is therefore useful to have a general way to circumvent these sorts of dependencies.

We have already shown that, after throwing m balls independently and uniformly at random into n bins, the distribution of the number of balls in a given bin is approximately Poisson with mean m/n. We would like to say that the joint distribution of the number of balls in *all* the bins is well approximated by assuming the load at *each* bin is an *independent* Poisson random variable with mean m/n. This would allow us to treat bin loads as independent random variables. We show here that we can do this when we are concerned with sufficiently rare events. Specifically, we show in Corollary 5.9 that taking the probability of an event using this Poisson approximation for all of the bins and multiplying it by $e\sqrt{m}$ gives an upper bound for the probability of the event when m balls are thrown into n bins. For rare events, this extra $e\sqrt{m}$ factor will not be significant. To achieve this result, we now introduce some technical machinery.

Suppose that m balls are thrown into n bins independently and uniformly at random, and let $X_i^{(m)}$ be the number of balls in the ith bin, where $1 \le i \le n$. Let $Y_1^{(m)}, \ldots, Y_n^{(m)}$ be independent Poisson random variables with mean m/n. We derive a useful relationship between these two sets of random variables. Tighter bounds for specific problems can often be obtained with more detailed analysis, but this approach is quite general and easy to apply.

The difference between throwing m balls randomly and assigning each bin a number of balls that is Poisson distributed with mean m/n is that, in the first case, we know there are m balls in total, whereas in the second case we know only that m is the expected number of balls in all of the bins. But suppose when we use the Poisson distribution we end up with m balls. In this case, we do indeed have that the distribution is the same as if we threw m balls into n bins randomly.

Theorem 5.6: *The distribution of $(Y_1^{(m)}, \ldots, Y_n^{(m)})$ conditioned on $\sum_i Y_i^{(m)} = k$ is the same as $(X_1^{(k)}, \ldots, X_n^{(k)})$, regardless of the value of m.*

Proof: When throwing k balls into n bins, the probability that $(X_1^{(k)}, \ldots, X_n^{(k)}) = (k_1, \ldots, k_n)$ for any $k_1, \ldots, k_n$ satisfying $\sum_i k_i = k$ is given by

$$\frac{\binom{k}{k_1; k_2; \ldots; k_n}}{n^k} = \frac{k!}{(k_1!)(k_2!)\cdots(k_n!)n^k}.$$

Now, for any $k_1, \ldots, k_n$ with $\sum_i k_i = k$, consider the probability that

$$(Y_1^{(m)}, \ldots, Y_n^{(m)}) = (k_1, \ldots, k_n)$$

conditioned on $(Y_1^{(m)}, \ldots, Y_n^{(m)})$ satisfying $\sum_i Y_i^{(m)} = k$:

$$\Pr\left((Y_1^{(m)}, \ldots, Y_n^{(m)}) = (k_1, \ldots, k_n) \mid \sum_{i=1}^{n} Y_i^{(m)} = k\right)$$

$$= \frac{\Pr((Y_1^{(m)} = k_1) \cap (Y_1^{(m)} = k_2) \cap \cdots \cap (Y_n^{(m)} = k_n))}{\Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = k\right)}.$$

The probability that $Y_i^{(m)} = k_i$ is $e^{-m/n}(m/n)^{k_i}/k_i!$, since the $Y_i^{(m)}$ are independent Poisson random variables with mean m/n. Also, by Lemma 5.2, the sum of the $Y_i^{(m)}$ is itself a Poisson random variable with mean m. Hence

$$\frac{\Pr((Y_1^{(m)} = k_1) \cap (Y_1^{(m)} = k_2) \cap \cdots \cap (Y_n^{(m)} = k_n))}{\Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = k\right)} = \frac{\prod_{i=1}^{n} e^{-m/n}(m/n)^{k_i}/k_i!}{e^{-m}m^k/k!}$$

$$= \frac{k!}{(k_1!)(k_2!)\cdots(k_n!)n^k},$$

proving the theorem. ∎

With this relationship between the two distributions, we can prove strong results about *any* function on the loads of the bins.

Theorem 5.7: *Let $f(x_1, \ldots, x_n)$ be a nonnegative function. Then*

$$\mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})] \le e\sqrt{m}\,\mathbf{E}[f(Y_1^{(m)}, \ldots, Y_n^{(m)})]. \tag{5.4}$$

Proof: We have that

$$\mathbf{E}[f(Y_1^{(m)}, \ldots, Y_n^{(m)})] = \sum_{k=0}^{\infty} \mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \mid \sum_{i=1}^{n} Y_i^{(m)} = k\right] \Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = k\right)$$

$$\ge \mathbf{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \mid \sum_{i=1}^{n} Y_i^{(m)} = m\right] \Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = m\right)$$

$$= \mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})] \Pr\left(\sum Y_i^{(m)} = m\right),$$

where the last equality follows from the fact that the joint distribution of the $Y_i^{(m)}$ given $\sum_{i=1}^{n} Y_i^{(m)} = m$ is exactly that of the $X_i^{(m)}$, as shown in Theorem 5.6. Since $\sum_{i=1}^{n} Y_i^{(m)}$ is Poisson distributed with mean m, we now have

$$\mathbf{E}[f(Y_1^{(m)}, \ldots, Y_n^{(m)})] \ge \mathbf{E}[f(X_1^{(m)}, \ldots, X_n^{(m)})] \frac{m^m e^{-m}}{m!}.$$

We use the following loose bound on $m!$, which we prove as Lemma 5.8:

$$m! < e\sqrt{m}\left(\frac{m}{e}\right)^m.$$

This yields

$$\mathbf{E}[f(Y_1^{(m)},\ldots,Y_n^{(m)})] \geq \mathbf{E}[f(X_1^{(m)},\ldots,X_n^{(m)})]\frac{1}{\mathrm{e}\sqrt{m}},$$

and the theorem is proven. ■

We prove the upper bound we used for factorials, which closely matches the loose lower bound we used in Lemma 5.1.

Lemma 5.8:

$$n! \leq \mathrm{e}\sqrt{n}\left(\frac{n}{\mathrm{e}}\right)^n. \tag{5.5}$$

Proof: We use the fact that

$$\ln(n!) = \sum_{i=1}^{n} \ln i.$$

We first claim that, for $i \geq 2$,

$$\int_{i-1}^{i} \ln x\, dx \geq \frac{\ln(i-1)+\ln i}{2}.$$

This follows from the fact that $\ln x$ is concave, since its second derivative is $-1/x^2$, which is always negative. Therefore,

$$\int_1^n \ln x\, dx \geq \sum_{i=1}^{n} \ln i - \frac{\ln n}{2}$$

or, equivalently,

$$n\ln n - n + 1 \geq \ln(n!) - \frac{\ln n}{2}.$$

The result now follows simply by exponentiating. ■

Theorem 5.7 holds for any nonnegative function on the number of balls in the bins. In particular, if the function is the indicator function that is 1 if some event occurs and 0 otherwise, then the theorem gives bounds on the probability of events. Let us call the scenario in which the number of balls in the bins are taken to be independent Poisson random variables with mean m/n the *Poisson case,* and the scenario where m balls are thrown into n bins independently and uniformly at random the *exact case*.

Corollary 5.9: *Any event that takes place with probability p in the Poisson case takes place with probability at most* $p\mathrm{e}\sqrt{m}$ *in the exact case.*

Proof: Let f be the indicator function of the event. In this case, $\mathbf{E}[f]$ is just the probability that the event occurs, and the result follows immediately from Theorem 5.7. ■

This is a quite powerful result. It says that any event that happens with small probability in the Poisson case also happens with small probability in the exact case, where balls are thrown into bins. Since in the analysis of algorithms we often want to show that certain events happen with small probability, this result says that we can utilize an analysis of the Poisson approximation to obtain a bound for the exact case. The Poisson approximation is easier to analyze because the numbers of balls in each bin are independent random variables.[1]

We can actually do even a little bit better in many natural cases. Part of the proof of the following theorem is outlined in Exercises 5.13 and 5.14.

Theorem 5.10: *Let* $f(x_1,\ldots,x_n)$ *be a nonnegative function such that* $\mathbf{E}[f(X_1^{(m)},\ldots,X_n^{(m)})]$ *is either monotonically increasing or monotonically decreasing in m. Then*

$$\mathbf{E}[f(X_1^{(m)},\ldots,X_n^{(m)})] \leq 2\mathbf{E}[f(Y_1^{(m)},\ldots,Y_n^{(m)})]. \tag{5.6}$$

The following corollary is immediate.

Corollary 5.11: *Let* $\mathcal{E}$ *be an event whose probability is either monotonically increasing or monotonically decreasing in the number of balls. If* $\mathcal{E}$ *has probability p in the Poisson case, then* $\mathcal{E}$ *has probability at most* $2p$ *in the exact case.*

To demonstrate the utility of this corollary, we again consider the maximum load problem for the case $m = n$. We have shown via a union bound argument that the maximum load is at most $3\ln n/\ln\ln n$ with high probability. Using the Poisson approximation, we prove the following almost-matching lower bound on the maximum load.

Lemma 5.12: *When n balls are thrown independently and uniformly at random into n bins, the maximum load is at least* $\ln n/\ln\ln n$ *with probability at least* $1 - 1/n$ *for n sufficiently large.*

Proof: In the Poisson case, the probability that bin 1 has load at least $M = \ln n/\ln\ln n$ is at least $1/\mathrm{e}M!$, which is the probability it has load exactly M. In the Poisson case, all bins are independent, so the probability that no bin has load at least M is at most

$$\left(1 - \frac{1}{\mathrm{e}M!}\right)^n \leq \mathrm{e}^{-n/(\mathrm{e}M!)}.$$

We now need to choose M so that $\mathrm{e}^{-n/(\mathrm{e}M!)} \leq n^{-2}$, for then (by Theorem 5.7) we will have that the probability that the maximum load is not at least M in the exact case is at most $\mathrm{e}\sqrt{n}/n^2 < 1/n$. This will give the lemma. Because the maximum load is clearly monotonically increasing in the number of balls, we could also apply the slightly better Theorem 5.10, but this would not affect the argument substantially.

It therefore suffices to show that $M! \leq n/2\mathrm{e}\ln n$, or equivalently that $\ln M! \leq \ln n - \ln\ln n - \ln(2\mathrm{e})$. From our bound (5.5), it follows that

$$M! \leq \mathrm{e}\sqrt{M}\left(\frac{M}{\mathrm{e}}\right)^M \leq M\left(\frac{M}{\mathrm{e}}\right)^M$$

[1] There are other ways to handle the dependencies in the balls-and-bins model. In Chapter 12 we describe a more general way to deal with dependencies (using martingales) that applies here. Also, there is a theory of negative dependence that applies to balls-and-bins problems that also allows these dependencies to be dealt with nicely.