Probability and Computing

Randomization and Probabilistic Techniques in Algorithms and Data Analysis

Second Edition

Michael Mitzenmacher Eli Upfal

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi - 110002, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107154889
10.1017/9781316651124

First published 2017

Printed in the United Kingdom by Clays, St Ives plc

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data
Names: Mitzenmacher, Michael, 1969– author. | Upfal, Eli, 1954– author.
Title: Probability and computing / Michael Mitzenmacher Eli Upfal.
Description: Second edition. | Cambridge, United Kingdom ;
New York, NY, USA : Cambridge University Press, [2017] |
Includes bibliographical references and index.
Identifiers: LCCN 2016041654 | ISBN 9781107154889
Subjects: LCSH: Algorithms. | Probabilities. | Stochastic analysis.
Classification: LCC QA274.M574 2017 | DDC 518/.1 – dc23
LC record available at https://lccn.loc.gov/2016041654

ISBN 978-1-107-15488-9 Hardback

Additional resources for this publication at www.cambridge.org/Mitzenmacher.

CHAPTER FOUR

Chernoff and Hoeffding Bounds

This chapter introduces large deviation bounds commonly called Chernoff and Hoeffding bounds. These bounds are extremely powerful, giving exponentially decreasing bounds on the tail distribution. These bounds are derived by applying Markov's inequality to the moment generating function of a random variable. We start this chapter by defining and discussing the properties of the moment generating function. We then derive Chernoff bounds for the binomial distribution and other related distributions, using a set balancing problem as an example, and the Hoeffding bound for sums of bounded random variables. To demonstrate the power of Chernoff bounds, we apply them to the analysis of randomized packet routing schemes on the hypercube and butterfly networks.

4.1. Moment Generating Functions

Before developing Chernoff bounds, we discuss the special role of the moment generating function $\mathbf{E}[e^{tX}]$.

Definition 4.1: *The* moment generating function *of a random variable X is*

$$M_X(t) = \mathbf{E}[e^{tX}].$$

We are mainly interested in the existence and properties of this function in the neighborhood of zero.

The function $M_X(t)$ captures all of the moments of X.

Theorem 4.1: *Let X be a random variable with moment generating function $M_X(t)$. Under the assumption that exchanging the expectation and differentiation operands is legitimate, for all $n > 1$ we then have*

$$\mathbf{E}[X^n] = M_X^{(n)}(0),$$

where $M_X^{(n)}(0)$ is the nth *derivative of $M_X(t)$ evaluated at $t = 0$.*

Proof: Assuming that we can exchange the expectation and differentiation operands, then

$$M_X^{(n)}(t) = \mathbf{E}[X^n e^{tX}].$$

Computed at $t = 0$, this expression yields

$$M_X^{(n)}(0) = \mathbf{E}[X^n].$$ ■

The assumption that expectation and differentiation operands can be exchanged holds whenever the moment generating function exists in a neighborhood of zero, which will be the case for all distributions considered in this book.

As a specific example, consider a geometric random variable X with parameter p, as in Definition 2.8. Then, for $t < -\ln(1-p)$,

$$\begin{aligned} M_X(t) &= \mathbf{E}[e^{tX}] \\ &= \sum_{k=1}^{\infty} (1-p)^{k-1} p e^{tk} \\ &= \frac{p}{1-p} \sum_{k=1}^{\infty} (1-p)^k e^{tk} \\ &= \frac{p}{1-p} ((1-(1-p)e^t)^{-1} - 1). \end{aligned}$$

It follows that

$$M_X^{(1)}(t) = p(1-(1-p)e^t)^{-2} e^t \quad \text{and}$$
$$M_X^{(2)}(t) = 2p(1-p)(1-(1-p)e^t)^{-3} e^{2t} + p(1-(1-p)e^t)^{-2} e^t.$$

Evaluating these derivatives at $t = 0$ and using Theorem 4.1 gives $\mathbf{E}[X] = 1/p$ and $\mathbf{E}[X^2] = (2-p)/p^2$, matching our previous calculations from Section 2.4 and Section 3.3.1.

Another useful property is that the moment generating function of a random variable (or, equivalently, all of the moments of the variable) uniquely defines its distribution. However, the proof of the following theorem is beyond the scope of this book.

Theorem 4.2: *Let X and Y be two random variables. If*

$$M_X(t) = M_Y(t)$$

for all $t \in (-\delta, \delta)$ for some $\delta > 0$, then X and Y have the same distribution.

One application of Theorem 4.2 is in determining the distribution of a sum of independent random variables.

Theorem 4.3: *If X and Y are independent random variables, then*

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Proof:

$$M_{X+Y}(t) = \mathbf{E}[e^{t(X+Y)}] = \mathbf{E}[e^{tX} e^{tY}] = \mathbf{E}[e^{tX}]\mathbf{E}[e^{tY}] = M_X(t) M_Y(t).$$

Here we have used that X and Y are independent – and hence e^{tX} and e^{tY} are independent – to conclude that $\mathbf{E}[e^{tX}e^{tY}] = \mathbf{E}[e^{tX}]\mathbf{E}[e^{tY}]$. ■

Thus, if we know $M_X(t)$ and $M_Y(t)$ and if we recognize the function $M_X(t)M_Y(t)$ as the moment generating function of a known distribution, then that must be the distribution of $X + Y$ when Theorem 4.2 applies. We will see examples of this in subsequent sections and in the exercises.

4.2. Deriving and Applying Chernoff Bounds

The Chernoff bound for a random variable X is obtained by applying Markov's inequality to e^{tX} for some well-chosen value t. From Markov's inequality, we can derive the following useful inequality: for any $t > 0$,

$$\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

In particular,

$$\Pr(X \geq a) \leq \min_{t>0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

Similarly, for any $t < 0$,

$$\Pr(X \leq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

Hence

$$\Pr(X \leq a) \leq \min_{t<0} \frac{\mathbf{E}[e^{tX}]}{e^{ta}}.$$

Bounds for specific distributions are obtained by choosing appropriate values for t. While the value of t that minimizes $\mathbf{E}[e^{tX}]/e^{ta}$ gives the best possible bounds, often one chooses a value of t that gives a convenient form. Bounds derived from this approach are generally referred to collectively as *Chernoff bounds*. When we speak of a Chernoff bound for a random variable, it could actually be one of many bounds derived in this fashion.

4.2.1. *Chernoff Bounds for the Sum of Poisson Trials*

We now develop the most commonly used version of the Chernoff bound: for the tail distribution of a sum of independent 0–1 random variables, which are also known as *Poisson trials*. (Poisson trials differ from Poisson random variables, which will be discussed in Section 5.3.) The distributions of the random variables in Poisson trials are not necessarily identical. *Bernoulli trials* are a special case of Poisson trials where the independent 0–1 random variables have the same distribution; in other words, all trials are Poisson trials that take on the value 1 with the same probability. Also recall that the binomial distribution gives the number of successes in n independent Bernoulli

trials. Our Chernoff bound will hold for the binomial distribution and also for the more general setting of the sum of Poisson trials.

Let $X_1, \ldots, X_n$ be a sequence of independent Poisson trials with $\Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$, and let

$$\mu = \mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} p_i.$$

For a given $\delta > 0$, we are interested in bounds on $\Pr(X \geq (1+\delta)\mu)$ and $\Pr(X \leq (1-\delta)\mu)$ – that is, the probability that X deviates from its expectation μ by $\delta\mu$ or more. To develop a Chernoff bound we need to compute the moment generating function of X. We start with the moment generating function of each X_i:

$$\begin{aligned} M_{X_i}(t) &= \mathbf{E}[e^{tX_i}] \\ &= p_i e^t + (1 - p_i) \\ &= 1 + p_i(e^t - 1) \\ &\leq e^{p_i(e^t - 1)}, \end{aligned}$$

where in the last inequality we have used the fact that, for any y, $1 + y \leq e^y$. Applying Theorem 4.3, we take the product of the n generating functions to obtain

$$\begin{aligned} M_X(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &\leq \prod_{i=1}^{n} e^{p_i(e^t - 1)} \\ &= \exp\left\{\sum_{i=1}^{n} p_i(e^t - 1)\right\} \\ &= e^{(e^t - 1)\mu}. \end{aligned}$$

Now that we have determined a bound on the moment generating function, we are ready to develop concrete versions of the Chernoff bound for a sum of Poisson trials. We start with bounds on the deviation above the mean.

Theorem 4.4: *Let $X_1, \ldots, X_n$ be independent Poisson trials such that $\Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbf{E}[X]$. Then the following Chernoff bounds hold:*

1. *for any $\delta > 0$,*

$$\Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^\mu; \tag{4.1}$$

2. *for $0 < \delta \leq 1$,*

$$\Pr(X \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/3}; \tag{4.2}$$

3. *for $R \geq 6\mu$,*

$$\Pr(X \geq R) \leq 2^{-R}. \tag{4.3}$$

The first bound of the theorem is the strongest, and it is from this bound that we derive the other two bounds, which have the advantage of being easier to state and compute with in many situations.

Proof: Applying Markov's inequality, for any $t > 0$ we have

$$\begin{aligned}\Pr(X \geq (1+\delta)\mu) &= \Pr(e^{tX} \geq e^{t(1+\delta)\mu}) \\ &\leq \frac{\mathbf{E}[e^{tX}]}{e^{t(1+\delta)\mu}} \\ &\leq \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}}.\end{aligned}$$

For any $\delta > 0$, we can set $t = \ln(1+\delta) > 0$ to get Eqn. (4.1):

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

To obtain Eqn. (4.2) we need to show that, for $0 < \delta \leq 1$,

$$\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \leq e^{-\delta^2/3}.$$

Taking the logarithm of both sides, we obtain the equivalent condition

$$f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3} \leq 0.$$

Computing the derivatives of $f(\delta)$, we have:

$$\begin{aligned}f'(\delta) &= 1 - \frac{1+\delta}{1+\delta} - \ln(1+\delta) + \frac{2}{3}\delta \\ &= -\ln(1+\delta) + \frac{2}{3}\delta; \\ f''(\delta) &= -\frac{1}{1+\delta} + \frac{2}{3}.\end{aligned}$$

We see that $f''(\delta) < 0$ for $0 \leq \delta < 1/2$ and that $f''(\delta) > 0$ for $\delta > 1/2$. Hence $f'(\delta)$ first decreases and then increases over the interval $[0, 1]$. Since $f'(0) = 0$ and $f'(1) < 0$, we can conclude that $f'(\delta) \leq 0$ in the interval $[0, 1]$. Since $f(0) = 0$, it follows that $f(\delta) \leq 0$ in that interval, proving Eqn. (4.2).

To prove Eqn. (4.3), let $R = (1+\delta)\mu$. Then, for $R \geq 6\mu$, $\delta = R/\mu - 1 \geq 5$. Hence, using Eqn. (4.1),

$$\begin{aligned}\Pr(X \geq (1+\delta)\mu) &\leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \\ &\leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu} \\ &\leq \left(\frac{e}{6}\right)^{R} \\ &\leq 2^{-R}.\end{aligned}$$

■

We obtain similar results bounding the deviation below the mean.

Theorem 4.5: *Let $X_1, \ldots, X_n$ be independent Poisson trials such that* $\Pr(X_i = 1) = p_i$. *Let* $X = \sum_{i=1}^{n} X_i$ *and* $\mu = \mathbf{E}[X]$. *Then, for* $0 < \delta < 1$:

1. $$\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}; \tag{4.4}$$

2. $$\Pr(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}. \tag{4.5}$$

Again, the bound of Eqn. (4.4) is stronger than Eqn. (4.5), but the latter is generally easier to use and sufficient in most applications.

Proof: Using Markov's inequality, for any $t < 0$ we have

$$\begin{aligned}\Pr(X \leq (1-\delta)\mu) &= \Pr(e^{tX} \geq e^{t(1-\delta)\mu}) \\ &\leq \frac{\mathbf{E}[e^{tX}]}{e^{t(1-\delta)\mu}} \\ &\leq \frac{e^{(e^t-1)\mu}}{e^{t(1-\delta)\mu}}.\end{aligned}$$

For $0 < \delta < 1$, we set $t = \ln(1-\delta) < 0$ to get Eqn. (4.4):

$$\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}.$$

To prove Eqn. (4.5) we must show that, for $0 < \delta < 1$,

$$\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}} \leq e^{-\delta^2/2}.$$

Taking the logarithm of both sides, we obtain the equivalent condition

$$f(\delta) = -\delta - (1-\delta)\ln(1-\delta) + \frac{\delta^2}{2} \leq 0$$

for $0 < \delta < 1$.

Differentiating $f(\delta)$ yields

$$\begin{aligned}f'(\delta) &= \ln(1-\delta) + \delta, \\ f''(\delta) &= -\frac{1}{1-\delta} + 1.\end{aligned}$$

Since $f''(\delta) < 0$ in the range $(0, 1)$ and since $f'(0) = 0$, we have $f'(\delta) \leq 0$ in the range $[0, 1)$. Therefore, $f(\delta)$ is nonincreasing in that interval. Since $f(0) = 0$, it follows that $f(\delta) \leq 0$ when $0 < \delta < 1$, as required. ■

Often the following form of the Chernoff bound, which is derived immediately from Eqn. (4.2) and Eqn. (4.4), is used.

Corollary 4.6: *Let $X_1, \ldots, X_n$ be independent Poisson trials such that* $\Pr(X_i = 1) = p_i$. *Let* $X = \sum_{i=1}^{n} X_i$ *and* $\mu = \mathbf{E}[X]$. *For* $0 < \delta < 1$,

$$\Pr(|X - \mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}. \tag{4.6}$$

In practice we often do not have the exact value of $\mathbf{E}[X]$. Instead we can use $\mu \geq \mathbf{E}[X]$ in Theorem 4.4 and $\mu \leq \mathbf{E}[X]$ in Theorem 4.5 (see Exercise 4.7).

4.2.2. *Example: Coin Flips*

Let X be the number of heads in a sequence of n independent fair coin flips. Applying the Chernoff bound of Eqn. (4.6), we have

$$\begin{aligned}\Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{1}{2}\sqrt{6n\ln n}\right) &\leq 2\exp\left\{-\frac{1}{3}\frac{n}{2}\frac{6\ln n}{n}\right\} \\ &= \frac{2}{n}.\end{aligned}$$

This demonstrates that the concentration of the number of heads around the mean $n/2$ is very tight; most of the time, the deviations from the mean are on the order of $O(\sqrt{n\ln n})$.

To compare the power of this bound to Chebyshev's bound, consider the probability of having no more than $n/4$ heads or no fewer than $3n/4$ heads in a sequence of n independent fair coin flips. In the previous chapter, we used Chebyshev's inequality to show that

$$\Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{4}{n}.$$

Already, this bound is worse than the Chernoff bound just calculated for a significantly larger event! Using the Chernoff bound in this case, we find that

$$\begin{aligned}\Pr\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) &\leq 2\exp\left\{-\frac{1}{3}\frac{n}{2}\frac{1}{4}\right\} \\ &\leq 2e^{-n/24}.\end{aligned}$$

Thus, Chernoff's technique gives a bound that is exponentially smaller than the bound obtained using Chebyshev's inequality.

4.2.3. *Application: Estimating a Parameter*

Suppose that we are interested in evaluating the probability that a particular gene mutation occurs in the population. Given a DNA sample, a lab test can determine if it carries the mutation. However, the test is expensive and we would like to obtain a relatively reliable estimate from a small number of samples.

Let p be the unknown value that we are trying to estimate. Assume that we have n samples and that $X = \tilde{p}n$ of these samples have the mutation. Given a sufficiently large number of samples, we expect the value p to be close to the sampled value $\tilde{p}$. We express this intuition using the concept of a confidence interval.

Definition 4.2: *A* $1-\gamma$ confidence interval *for a parameter* p *is an interval* $[\tilde{p} - \delta, \tilde{p} + \delta]$ *such that*

$$\Pr(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) \geq 1 - \gamma.$$

Notice that, instead of predicting a single value for the parameter, we give an interval that is likely to contain the parameter. If p can take on any real value, it may not make sense to try to pin down its exact value from a finite sample, but it does make sense to estimate it within some small range.

Naturally we want both the interval size 2δ and the error probability γ to be as small as possible. We derive a trade-off between these two parameters and the number of samples n. In particular, given that among n samples (chosen uniformly at random from the entire population) we find the mutation in exactly $X = \tilde{p}n$ samples, we need to find values of δ and γ for which

$$\Pr(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) = \Pr(np \in [n(\tilde{p} - \delta), n(\tilde{p} + \delta)]) \geq 1 - \gamma.$$

Now $X = n\tilde{p}$ has a binomial distribution with parameters n and p, so $\mathbf{E}[X] = np$. If $p \notin [\tilde{p} - \delta, \tilde{p} + \delta]$ then we have one of the following two events:

1. if $p < \tilde{p} - \delta$, then $X = n\tilde{p} > n(p + \delta) = \mathbf{E}[X](1 + \delta/p)$;
2. if $p > \tilde{p} + \delta$, then $n\tilde{p} < n(p - \delta) = \mathbf{E}[X](1 - \delta/p)$.

We can apply the Chernoff bounds in Eqns. (4.2) and (4.5) to compute

$$\Pr(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) = \Pr\left(X < np\left(1 - \frac{\delta}{p}\right)\right) + \Pr\left(X > np\left(1 + \frac{\delta}{p}\right)\right) \tag{4.7}$$

$$< e^{-np(\delta/p)^2/2} + e^{-np(\delta/p)^2/3} \tag{4.8}$$

$$= e^{-n\delta^2/2p} + e^{-n\delta^2/3p}. \tag{4.9}$$

The bound given in Eqn. (4.9) is not useful because the value of p is unknown. A simple solution is to use the fact that $p \leq 1$, yielding

$$\Pr(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) < e^{-n\delta^2/2} + e^{-n\delta^2/3}.$$

Setting $\gamma = e^{-n\delta^2/2} + e^{-n\delta^2/3}$, we obtain a trade-off between δ, n, and the error probability γ.

We can apply other Chernoff bounds, such as those in Exercises 4.13 and 4.16, to obtain better bounds. We return to the subject of parameter estimation when we discuss the Monte Carlo method in Chapter 11.

4.3. Better Bounds for Some Special Cases

We can obtain stronger bounds using a simpler proof technique for some special cases of symmetric random variables.

We consider first the sum of independent random variables when each variable assumes the value 1 or -1 with equal probability.

Theorem 4.7: *Let $X_1, \ldots, X_n$ be independent random variables with*

$$\Pr(X_i = 1) = \Pr(X_i = -1) = \frac{1}{2}.$$

Let $X = \sum_{i=1}^{n} X_i$. *For any* $a > 0$,

$$\Pr(X \geq a) \leq e^{-a^2/2n}.$$

Proof: For any $t > 0$,

$$\mathbf{E}[e^{tX_i}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t}.$$

To estimate $\mathbf{E}[e^{tX_i}]$, we observe that

$$e^t = 1 + t + \frac{t^2}{2!} + \cdots + \frac{t^i}{i!} + \cdots$$

and

$$e^{-t} = 1 - t + \frac{t^2}{2!} + \cdots + (-1)^i \frac{t^i}{i!} + \cdots,$$

using the Taylor series expansion for e^t. Thus,

$$\begin{aligned}
\mathbf{E}[e^{tX_i}] &= \frac{1}{2}e^t + \frac{1}{2}e^{-t} \\
&= \sum_{i \geq 0} \frac{t^{2i}}{(2i)!} \\
&\leq \sum_{i \geq 0} \frac{(t^2/2)^i}{i!} \\
&= e^{t^2/2}.
\end{aligned}$$

Using this estimate yields

$$\mathbf{E}[e^{tX}] = \prod_{i=1}^{n} \mathbf{E}[e^{tX_i}] \leq e^{t^2 n/2}$$

and

$$\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbf{E}[e^{tX}]}{e^{ta}} \leq e^{t^2 n/2 - ta}.$$

Setting $t = a/n$, we obtain

$$\Pr(X \geq a) \leq e^{-a^2/2n}.$$ ■

By symmetry we also have

$$\Pr(X \leq -a) \leq e^{-a^2/2n}.$$

Combining the two results yields our next corollary.

Corollary 4.8: *Let* $X_1, \ldots, X_n$ *be independent random variables with*

$$\Pr(X_i = 1) = \Pr(X_i = -1) = \frac{1}{2}.$$

Let $X = \sum_{i=1}^{n} X_i$. *Then, for any* $a > 0$,

$$\Pr(|X| \geq a) \leq 2e^{-a^2/2n}.$$

Applying the transformation $Y_i = (X_i + 1)/2$ allows us to prove the following.

Corollary 4.9: *Let* $Y_1, \ldots, Y_n$ *be independent random variables with*

$$\Pr(Y_i = 1) = \Pr(Y_i = 0) = \frac{1}{2}.$$

Let $Y = \sum_{i=1}^{n} Y_i$ *and* $\mu = \mathbf{E}[Y] = n/2$.

1. For any $a > 0$,

$$\Pr(Y \geq \mu + a) \leq e^{-2a^2/n}.$$

2. For any $\delta > 0$,

$$\Pr(Y \geq (1 + \delta)\mu) \leq e^{-\delta^2 \mu}. \tag{4.10}$$

Proof: Using the notation of Theorem 4.7, we have

$$Y = \sum_{i=1}^{n} Y_i = \frac{1}{2}\left(\sum_{i=1}^{n} X_i\right) + \frac{n}{2} = \frac{1}{2}X + \mu.$$

Applying Theorem 4.7 yields

$$\Pr(Y \geq \mu + a) = \Pr(X \geq 2a) \leq e^{-4a^2/2n},$$

proving the first part of the corollary. The second part follows from setting $a = \delta\mu = \delta n/2$. Again applying Theorem 4.7, we have

$$\Pr(Y \geq (1 + \delta)\mu) = \Pr(X \geq 2\delta\mu) \leq e^{-2\delta^2\mu^2/n} = e^{-\delta^2\mu}. \qquad \blacksquare$$

Note that the constant in the exponent of the bound of Eqn. (4.10) is 1 instead of the 1/3 in the bound of Eqn. (4.2).

Similarly, we have the following result.

Corollary 4.10: *Let* $Y_1, \ldots, Y_n$ *be independent random variables with*

$$\Pr(Y_i = 1) = \Pr(Y_i = 0) = \frac{1}{2}.$$

Let $Y = \sum_{i=1}^{n} Y_i$ *and* $\mu = \mathbf{E}[Y] = n/2$.

1. For any $0 < a < \mu$,

$$\Pr(Y \leq \mu - a) \leq e^{-2a^2/n}.$$

2. For any $0 < \delta < 1$,

$$\Pr(Y \leq (1 - \delta)\mu) \leq e^{-\delta^2 \mu}. \tag{4.11}$$

4.4. Application: Set Balancing

Given an $n \times m$ matrix $\mathbf{A}$ with entries in $\{0, 1\}$, let

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}.$$

Suppose that we are looking for a vector $\bar{b}$ with entries in $\{-1, 1\}$ that minimizes

$$\|\mathbf{A}\bar{b}\|_\infty = \max_{i=1,\ldots,n} |c_i|.$$

This problem arises in designing statistical experiments. Each column of the matrix $\mathbf{A}$ represents a subject in the experiment and each row represents a feature. The vector $\bar{b}$ partitions the subjects into two disjoint groups, so that each feature is roughly as balanced as possible between the two groups. One of the groups serves as a control group for an experiment that is run on the other group.

Our randomized algorithm for computing a vector $\bar{b}$ is extremely simple. We randomly choose the entries of $\bar{b}$, with $\Pr(b_i = 1) = \Pr(b_i = -1) = 1/2$. The choices for different entries are independent. Surprisingly, although this algorithm ignores the entries of the matrix $\mathbf{A}$, the following theorem shows that $\|\mathbf{A}\bar{b}\|_\infty$ is likely to be only $O(\sqrt{m \ln n})$. This bound is fairly tight. In Exercise 4.15 you are asked to show that, when $m = n$, there exists a matrix $\mathbf{A}$ for which $\|\mathbf{A}\bar{b}\|_\infty$ is $\Omega(\sqrt{n})$ for any choice of $\bar{b}$.

Theorem 4.11: *For a random vector $\bar{b}$ with entries chosen independently and with equal probability from the set $\{-1, 1\}$,*

$$\Pr\left(\|\mathbf{A}\bar{b}\|_\infty \geq \sqrt{4m \ln n}\right) \leq \frac{2}{n}.$$

Proof: Consider the ith row $\bar{a}_i = a_{i,1}, \ldots, a_{i,m}$, and let k be the number of 1s in that row. If $k \leq \sqrt{4m \ln n}$, then clearly $|\bar{a}_i \cdot \bar{b}| = |c_i| \leq \sqrt{4m \ln n}$. On the other hand, if $k > \sqrt{4m \ln n}$ then we note that the k nonzero terms in the sum

$$Z_i = \sum_{j=1}^{m} a_{i,j} b_j$$

are independent random variables, each with probability $1/2$ of being either $+1$ or -1.

Now using the Chernoff bound of Corollary 4.8 and the fact that $m \geq k$,

$$\Pr\left(|Z_i| > \sqrt{4m \ln n}\right) \leq 2e^{-4m \ln n/2k} \leq \frac{2}{n^2}.$$

By the union bound, the probability that the bound fails for any row is at most $2/n$. ∎

4.5. The Hoeffding Bound

Hoeffding's bound extends the Chernoff bound technique to general random variables with a bounded range.

Theorem 4.12 [Hoeffding Bound]: *Let $X_1, \ldots, X_n$ be independent random variables such that for all $1 \leq i \leq n$, $\mathbf{E}[X_i] = \mu$ and $\Pr(a \leq X_i \leq b) = 1$. Then*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}.$$

Proof: The proof relies on the following bound for the moment generating function, which we prove first.

Lemma 4.13 [Hoeffding's Lemma]: *Let X be a random variable such that $\Pr(X \in [a, b]) = 1$ and $\mathbf{E}[X] = 0$. Then for every $\lambda > 0$,*

$$\mathbf{E}[e^{\lambda X}] \leq e^{\lambda^2(b-a)^2/8}.$$

Proof: Before beginning, note that since $\mathbf{E}[X] = 0$, if $a = 0$ then $b = 0$ and the statement is trivial. Hence we may assume $a < 0$ and $b > 0$.

Since $f(x) = e^{\lambda x}$ is a convex function, for any $\alpha \in (0, 1)$,

$$f(\alpha a + (1-\alpha)b) \leq \alpha e^{\lambda a} + (1-\alpha)e^{\lambda b}.$$

For $x \in [a, b]$, let $\alpha = \frac{b-x}{b-a}$; then $x = \alpha a + (1-\alpha)b$ and we have

$$e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}.$$

We consider $e^{\lambda X}$ and take expectations. Using the fact that $\mathbf{E}[X] = 0$, we have

$$\begin{aligned}
\mathbf{E}[e^{\lambda X}] &\leq \mathbf{E}\left[\frac{b-X}{b-a}e^{\lambda a}\right] + \mathbf{E}\left[\frac{X-a}{b-a}e^{\lambda b}\right] \\
&= \frac{b}{b-a}e^{\lambda a} - \frac{\mathbf{E}[X]}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} + \frac{\mathbf{E}[X]}{b-a}e^{\lambda b} \\
&= \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}.
\end{aligned}$$

We now require some manipulation of this final expression. Let $\phi(t) = -\theta t + \ln(1 - \theta + \theta e^t)$, for $\theta = \frac{-a}{b-a} > 0$. Then

$$\begin{aligned}
e^{\phi(\lambda(b-a))} &= e^{-\theta\lambda(b-a)}(1 - \theta + \theta e^{\lambda(b-a)}) \\
&= e^{\lambda a}(1 - \theta + \theta e^{\lambda(b-a)}) \\
&= e^{\lambda a}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{\lambda(b-a)}\right) \\
&= \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b},
\end{aligned}$$

which equals the upper bound we derived for $\mathbf{E}[e^{\lambda X}]$. It is not hard to verify that $\phi(0) = \phi'(0) = 0$, and $\phi''(t) \leq 1/4$ for all t. By Taylor's theorem, for any $t > 0$ there is a

$t' \in [0, t]$ such that

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{1}{2}t^2\phi''(t') \leq \frac{1}{8}t^2.$$

Thus, for $t = \lambda(b-a)$, we have

$$\phi(\lambda(b-a)) \leq \frac{\lambda^2(b-a)^2}{8}.$$

It follows that

$$\mathbf{E}[e^{\lambda X}] \leq e^{\phi(\lambda(b-a))} \leq e^{\lambda^2(b-a)^2/8}. \quad \blacksquare$$

We now return to the proof of Theorem 4.12. Let $Z_i = X_i - \mathbf{E}[X_i]$ and $Z = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

For any $\lambda > 0$, by Markov's inequality,

$$\Pr(Z \geq \epsilon) = \Pr(e^{\lambda Z} \geq e^{\lambda\epsilon}) \leq e^{-\lambda\epsilon}\mathbf{E}[e^{\lambda Z}] \leq e^{-\lambda\epsilon}\prod_{i=1}^{n}\mathbf{E}[e^{\lambda Z_i/n}]$$

$$\leq e^{-\lambda\epsilon}\prod_{i=1}^{n} e^{\lambda^2(b-a)^2/n^2} \leq e^{-\lambda\epsilon+\lambda^2(b-a)^2/8n},$$

where for the key second to last inequality we have used Hoeffding's Lemma with the fact that Z_i/n is bounded between $(a-\mu)/n$ and $(b-\mu)/n$. Setting $\lambda = \frac{4n\epsilon}{(b-a)^2}$ gives

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \epsilon\right) = \Pr(Z \geq \epsilon) \leq e^{-2n\epsilon^2/(b-a)^2}.$$

Applying the same argument for $\Pr(Z \leq -\epsilon)$ with $\lambda = -\frac{4n\epsilon}{(b-a)^2}$ gives

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leq -\epsilon\right) = \Pr(Z \leq -\epsilon) \leq e^{-2n\epsilon^2/(b-a)^2}.$$

Applying a union bound on the two cases gives the theorem. $\blacksquare$

The proof of the following more general version of the bound is left as an exercise (Exercise 4.20).

Theorem 4.14: *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbf{E}[X_i] = \mu_i$ and $\Pr(a_i \leq X_i \leq b_i) = 1$ for constants a_i and b_i. Then*

$$\Pr\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mu_i\right| \geq \epsilon\right) \leq 2e^{-2\epsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}.$$

Note that Theorem 4.12 bounds the deviation of the average of the n random variables while Theorem 4.14 bounds the deviation of the sum of the variables.

Examples:

1. Consider n independent random variables $X_1, \ldots, X_n$ such that X_i is uniformly distributed in $\{0, \ldots, \ell\}$. For all i, $\mu = \mathbf{E}[X_i] = \ell/2$, and

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\ell}{2}\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/\ell^2}.$$

   In particular,

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \delta\mu\right) \leq 2e^{-n\delta^2/2}.$$

2. Consider n independent random variables $Y_1, \ldots, Y_n$ such that Y_i is uniformly distributed in $\{0, i\}$. Let $Y = \sum_{i=1}^{n} Y_i$. Then $\mathbf{E}[Y_i] = i/2$, and $\mu = \mathbf{E}[Y] = \sum_{i=1}^{n} i/2 = n(n+1)/4$. Applying Theorem 4.14 with $c_i = i$ we have

$$\Pr\left(\left|Y - \frac{n(n+1)}{4}\right| \geq \epsilon\right) \leq 2e^{-2\epsilon^2/\sum_{i=1}^{n} c_i^2} = 2e^{-2\epsilon^2/(n(n+1)(2n+1)/6)}$$
$$= 2e^{-12\epsilon^2/(n(n+1)(2n+1))}.$$

   We can conclude

$$\Pr(|Y - \mu| \geq \delta\mu) \leq 2e^{-12\delta^2 n^2(n+1)^2/(16n(n+1)(2n+1))} \leq 2e^{-3n\delta^2/8}.$$

4.6.* Application: Packet Routing in Sparse Networks

A fundamental problem in parallel computing is how to communicate efficiently over sparse communication networks. We model a communication network by a directed graph on N nodes. Each node is a routing switch. A directed edge models a communication channel, which connects two adjacent routing switches. We consider a synchronous computing model in which (a) an edge can carry one packet in each time step and (b) a packet can traverse no more than one edge per step. We assume that switches have buffers or queues to store packets waiting for transmission through each of the switch's outgoing edges.

Given a network topology, a *routing algorithm* specifies, for each pair of nodes, a route – or a sequence of edges – connecting the pair in the network. The algorithm may also specify a queuing policy for ordering packets in the switches' queues. For example, the First In First Out (FIFO) policy orders packets by their order of arrival. The Furthest To Go (FTG) policy orders packets in decreasing order of the number of edges they must still cross in the network.

Our measure of the performance of a routing algorithm on a given network topology is the maximum time – measured as the number of parallel steps – required to route an arbitrary *permutation routing* problem, where each node sends exactly one packet and each node is the address of exactly one packet.

Of course, routing a permutation can be done in just one parallel step if the network is a complete graph connecting all of the nodes to each other. Practical considerations, however, dictate that a network for a large-scale parallel machine must be sparse.

Each node can be connected directly to only a few neighbors, and most packets must traverse intermediate nodes en route to their final destination. Since an edge may be on the path of more than one packet and since each edge can process only one packet per step, parallel packet routing on sparse networks may lead to congestion and bottlenecks. The practical problem of designing an efficient communication scheme for parallel computers leads to an interesting combinatorial and algorithmic problem: designing a family of sparse networks connecting any number of processors, together with a routing algorithm that routes an arbitrary permutation request in a small number of parallel steps.

We discuss here a simple and elegant randomized routing technique and then use Chernoff bounds to analyze its performance on the hypercube network and the butterfly network. We first analyze the case of routing a permutation on a hypercube, a network with N-processors and $O(N \log N)$ edges. We then present a tighter argument for the butterfly network, which has N nodes and only $O(N)$ edges.

4.6.1. *Permutation Routing on the Hypercube*

Let $\mathcal{N} = \{0 \leq i \leq N-1\}$ be the set of processors in our parallel machine and assume that $N = 2^n$ for some integer n. Let $\bar{x} = (x_1, \ldots, x_n)$ be the binary representation of the number $0 \leq x \leq N-1$.

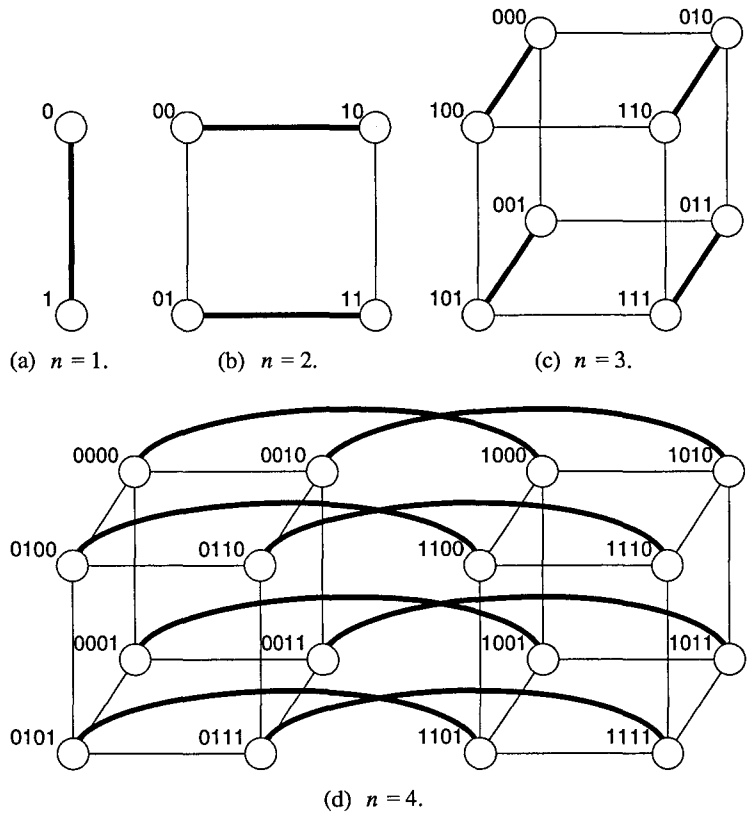
Definition 4.3: *The n-dimensional* hypercube *(or n-*cube*) is a network with $N = 2^n$ nodes such that node x has a direct connection to node y if and only if $\bar{x}$ and $\bar{y}$ differ in exactly one bit.*

See Figure 4.1. Note that the total number of directed edges in the n-cube is nN, since each node is adjacent to n outgoing and n ingoing edges. Also, the diameter of the network is n; that is, there is a directed path of length up to n connecting any two nodes in the network, and there are pairs of nodes that are not connected by any shorter path.

The topology of the hypercube allows for a simple bit-fixing routing mechanism, as shown in Algorithm 4.1. When determining which edge to cross next, the algorithm simply considers each bit in order and crosses the edge if necessary.

Although it seems quite natural, using only the bit-fixing routes can lead to high levels of congestion and poor performance, as shown in Exercise 4.22. There are certain permutations on which the bit-fixing routes behave poorly. It turns out, as we will show, that these routes perform well if each packet is being sent from a source to a destination chosen uniformly at random. This motivates the following approach: first route each packet to a randomly chosen intermediate point, and then route it from this intermediate point to its final destination.

It may seem unusual to first route packets to a random intermediate point. In some sense, this is similar in spirit to our analysis of Quicksort in Section 2.5. We found there that for a list already sorted in reverse order, Quicksort would take $\Omega(n^2)$ comparisons, whereas the expected number of comparisons for a randomly chosen permutation is only $O(n \log n)$. Randomizing the data can lead to a better running time for Quicksort.



(a) $n = 1$. (b) $n = 2$. (c) $n = 3$.

(d) $n = 4$.

Figure 4.1: Hypercubes of dimensions 1, 2, 3, and 4.

n-Cube Bit-Fixing Routing Algorithm:

1. Let $\bar{a}$ and $\bar{b}$ be the origin and the destination of the packet.
2. For $i = 1$ to n, do:
   (a) If $a_i \neq b_i$ then traverse the edge $(b_1, \ldots, b_{i-1}, a_i, \ldots, a_n) \rightarrow (b_1, \ldots, b_{i-1}, b_i, a_{i+1}, \ldots, a_n)$.

Algorithm 4.1: n-Cube bit-fixing routing algorithm.

Here, too, randomizing the routes that packets take – by routing them through a random intermediate point – avoids bad initial permutations and leads to good expected performance.

The two-phase routing algorithm (Algorithm 4.2) is executed in parallel by all the packets. The random choices are made independently for each packet. Our analysis holds for any queueing policy that obeys the following natural requirement: if a queue is not empty at the beginning of a time step, some packet is sent along the edge associated with that queue during that time step. We prove that this routing strategy achieves asymptotically optimal parallel time.

Two-Phase Routing Algorithm:

Phase I – Route the packet to a randomly chosen node in the network using the bit-fixing route.

Phase II – Route the packet from its random location to its final destination using the bit-fixing route.

Algorithm 4.2: Two-phase routing algorithm.

Theorem 4.15: *Given an arbitrary permutation routing problem, with probability* $1 - O(N^{-1})$ *the two-phase routing scheme of Algorithm 4.2 routes all packets to their destinations on the n-cube in* $O(n) = O(\log N)$ *parallel steps.*

Proof: We first analyze the run-time of Phase I. To simplify the analysis we assume that no packet starts the execution of Phase II before all packets have finished the execution of Phase I. We show later that this assumption can be removed.

We emphasize a fact that we use implicitly throughout. If a packet is routed to a randomly chosen node $\bar{x}$ in the network, we can think of $\bar{x} = (x_1, \ldots, x_n)$ as being generated by setting each x_i independently to be 0 with probability 1/2 and 1 with probability 1/2.

For a given packet M, let $T_1(M)$ be the number of steps for M to finish Phase I. For a given edge e, let $X_1(e)$ denote the total number of packets that traverse edge e during Phase I.

In each step of executing Phase I, packet M is either traversing an edge or waiting in a queue while some other packet traverses an edge on M's route. This simple observation relates the routing time of M to the total number of packet transitions through edges on the path of M, as follows.

Lemma 4.16: *Let* $e_1, \ldots, e_m$ *be the* $m \leq n$ *edges traversed by a packet M in Phase I. Then*

$$T_1(M) \leq \sum_{i=1}^{m} X_1(e_i).$$

Let us call any path $P = (e_1, e_2, \ldots, e_m)$ of $m \leq n$ edges that follows the bit-fixing algorithm a *possible packet path.* We denote the corresponding nodes by $v_0, v_1, \ldots, v_m$ with $e_i = (v_{i-1}, v_i)$. Following the definition of $T_1(M)$, for any possible packet path P we let

$$T_1(P) = \sum_{i=1}^{m} X_1(e_i).$$

By Lemma 4.16, the probability that Phase I takes more than T steps is bounded by the probability that, for some possible packet path P, $T_1(P) \geq T$. Note that there are at most $2^n \cdot 2^n = 2^{2n}$ possible packet paths, since there are 2^n possible origins and 2^n possible destinations.

To prove the theorem, we need a high-probability bound on $T_1(P)$. Since $T_1(P)$ equals the summation $\sum_{i=1}^{n} X_1(e_i)$, it would be natural to try to use a Chernoff bound. The difficulty here is that the $X_1(e_i)$ are not independent random variables, since a packet that traverses an edge is likely to traverse one of its adjacent edges. To circumvent this difficulty, we first use a Chernoff bound to prove that, with high probability, no more than $6n$ different packets cross any edge of P. We then condition on this event to derive a high-probability bound on the total number of transitions these packets make through edges of the path P, again using a Chernoff bound.[1]

Let us now fix a specific possible packet path P with m edges. To obtain a high-probability bound on the number of packets that cross an edge of P, let us call a packet *active* at a node v_{i-1} on the path P if it reaches v_{i-1} and has the possibility of crossing edge e_i to v_i. That is, if v_{i-1} and v_i differ in the jth bit then – in order for a packet to be active at v_{i-1} – its jth bit cannot have been fixed by the bit-fixing algorithm when it reaches v_{i-1}. We may also call a packet active if it is active at some vertex on the path P. We bound the total number of active packets.

For $k = 1, \ldots, N$, let H_k be a 0–1 random variable such that $H_k = 1$ if the packet starting at node k is active and $H_k = 0$ otherwise. Notice that the H_k are independent because (a) each H_k depends only on the choice of the intermediate destination of the packet starting at node k and (b) these choices are independent for all packets. Let $H = \sum_{k=1}^{N} H_k$ be the total number of active packets.

We first bound $\mathrm{E}[H]$. Consider all the active packets at v_{i-1}. Assume that $v_{i-1} = (b_1, \ldots, b_{j-1}, a_j, a_{j+1}, \ldots, a_n)$ and $v_i = (b_1, \ldots, b_{j-1}, b_j, a_{j+1}, \ldots, a_n)$. Then only packets that start at one of the addresses $(*, \ldots, *, a_j, \ldots, a_n)$, where $*$ stands for either a 0 or a 1, can reach v_{i-1} before the jth bit is fixed. Similarly, each of these packets actually reaches v_{i-1} only if its random destination is one of the addresses $(b_1, \ldots, b_{j-1}, *, \ldots, *)$. Thus, there are no more than 2^{j-1} possible active packets at v_{i-1}, and the probability that each of these packets is actually active at v_{i-1} is $2^{-(j-1)}$. Hence the expected number of active packets per vertex is 1 and, since we need only consider the m vertices $v_0, \ldots, v_{m-1}$, it follows by linearity of expectations that

$$\mathrm{E}[H] \leq m \cdot 1 \leq n.$$

Since H is the sum of independent 0–1 random variables, we can apply the Chernoff bound (we use the bound of Eqn. (4.3)) to prove

$$\Pr(H \geq 6n \geq 6\mathrm{E}[H]) \leq 2^{-6n}.$$

The high-probability bound for H can help us obtain a bound for $T_1(P)$ as follows. Using

$$\begin{aligned} \Pr(A) &= \Pr(A \mid B)\ \Pr(B) + \Pr(A \mid \bar{B})\ \Pr(\bar{B}) \\ &\leq \Pr(B) + \Pr(A \mid \bar{B}), \end{aligned}$$

[1] This approach overestimates the time to finish a phase. In fact, there is a deterministic argument showing that, in this setting, the delay of a packet on a path is bounded by the number of different packets that traverse edges of the path, and hence there is no need to bound the total number of traversals of these packets on the path. However, in the spirit of this book we prefer to present the probabilistic argument.

we find for a given possible packet path P that

$$\begin{aligned}\Pr(T_1(P) \geq 30n) &\leq \Pr(H \geq 6n) + \Pr(T_1(P) \geq 30n \mid H < 6n) \\ &\leq 2^{-6n} + \Pr(T_1(P) \geq 30n \mid H < 6n).\end{aligned}$$

Hence if we show

$$\Pr(T_1(P) \geq 30n \mid H < 6n) \leq 2^{-3n-1},$$

we then have

$$\Pr(T_1(P) \geq 30n) \leq 2^{-3n},$$

which proves sufficient for our purposes.

We therefore need to bound the conditional probability $\Pr(T_1(P) \geq 30n \mid H \leq 6n)$. In other words, conditioning on having no more than $6n$ active packets that might use edges of P, we need a bound on the total number of transitions that these packets take through edges of P.

We first observe that, if a packet leaves the path, it cannot return to that path in this phase of the routing algorithm. Indeed, assume that the active packet was at v_i and that it moved to $w \neq v_{i+1}$. The smallest index bit in which v_{i+1} and w differ cannot be fixed later in this phase, so the route of the packet and the path P cannot meet again in this phase.

Now suppose we have an active packet on our path P at node v_i. What is the probability that the packet crosses e_i? Let us think of our packet as fixing the bits in the binary representation of its destination one at a time by independent random coin flips. The nodes of the edge e_i differ in one bit (say, the jth bit) in this representation. It is therefore clear that the probability of the packet crossing edge e_i is at most $1/2$, since to cross this edge it must choose the appropriate value for the jth bit. (In fact, the probability might be less than $1/2$; the packet might cross some other edge before choosing the value of the jth bit.)

To obtain our bound, let us view as a *trial* each point in the algorithm where an active packet at a node v_i on the path P might cross edge e_i. The trial is successful if the packet leaves the path but a failure if the packet stays on the path. Since the packet leaves the path on a successful trial, if there are at most $6n$ active packets then there can be at most $6n$ successes. Each trial is successful, independently, with probability at least $1/2$. The number of trials is itself a random variable, which we use in our bound of $T_1(P)$.

We claim that the probability that the active packets cross edges of P more than $30n$ times is less than the probability that a fair coin flipped $36n$ times comes up heads fewer than $6n$ times. To see this, think of a coin being flipped for each trial, with heads corresponding to a success. The coin is biased to come up heads with the proper probability for each trial, but this probability is always at least $1/2$ and the coins are independent for each trial. Each failure (tails) corresponds to an active packet crossing an edge, but once there have been $6n$ successes we know there are no more active packets left that can cross an edge of the path. Using a fair coin instead of a coin possibly biased in favor of success can only lessen the probability that the active packets cross edges of

P more than $30n$ times, as can be shown easily by induction (on the number of biased coins).

Letting Z be the number of heads in $36n$ fair coin flips, we now apply the Chernoff bound of Eqn. (4.5) to prove:

$$\Pr(T_1(P) \geq 30n \mid H \leq 6n) \leq \Pr(Z \leq 6n) \leq e^{-18n(2/3)^2/2} = e^{-4n} \leq 2^{-3n-1}.$$

It follows that

$$\Pr(T_1(P) \geq 30n) \leq \Pr(H \geq 6n) + \Pr(T_1(P) \geq 30n \mid H \leq 6n) \leq 2^{-3n},$$

as we wanted to show. Because there are at most 2^{2n} possible packet paths in the hypercube, the probability that there is *any* possible packet path for which $T_1(P) \geq 30n$ is bounded by

$$2^{2n}2^{-3n} = 2^{-n} = O(N^{-1}).$$

This completes the analysis of Phase I. Consider now the execution of Phase II, assuming that all packets completed their Phase I route. In this case, Phase II can be viewed as running Phase I backwards: instead of packets starting at a given origin and going to a random destination, they start at a random origin and end at a given destination. Hence no packet spends more than $30n$ steps in Phase II with probability $1 - O(N^{-1})$.

In fact, we can remove the assumption that packets begin Phase II only after Phase I has completed. The foregoing argument allows us to conclude that the total number of packet traversals across the edges of any packet path during Phase I and Phase II together is bounded by $60n$ with probability $1 - O(N^{-1})$. Since a packet can be delayed only by another packet traversing that edge, we find that every packet completes both Phase I and Phase II after $60n$ steps with probability $1 - O(N^{-1})$ regardless of how the phases interact, concluding the proof of Theorem 4.15 ■

Note that the run-time of the routing algorithm is optimal up to a constant factor, since the diameter of the hypercube is n. However, the network is not fully utilized because $2nN$ directed edges are used to route just N packets. At any give time, at most $1/2n$ of the edges are actually being used. This issue is addressed in the next section.

4.6.2. *Permutation Routing on the Butterfly*

In this section we adapt the result for permutation routing on the hypercube networks to routing on butterfly networks, yielding a significant improvement in network utilization. Specifically, our goal in this section is to route a permutation on a network with N nodes and $O(N)$ edges in $O(\log N)$ parallel time steps. Recall that the hypercube network had N nodes but $\Omega(N \log N)$ edges. Although the argument will be similar in spirit to that for the hypercube network, there is some additional complexity to the argument for the butterfly network.

We work on the wrapped butterfly network, defined as follows.

Definition 4.4: *The* wrapped butterfly network *has $N = n2^n$ nodes. The nodes are arranged in n columns and 2^n rows. A node's address is a pair (x, r), where $1 \leq x \leq 2^n$*

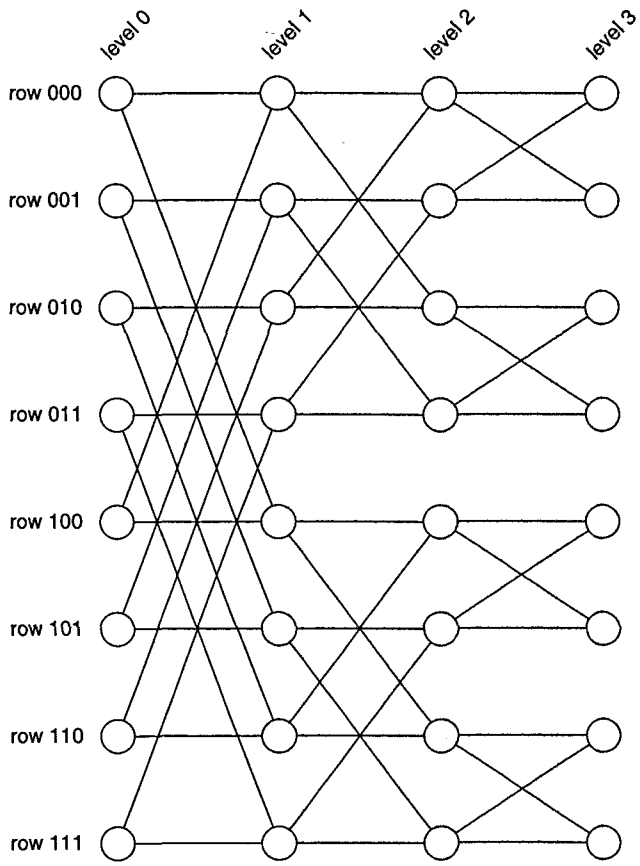


Figure 4.2: The butterfly network. In the wrapped butterfly, levels 0 and 3 are collapsed into one level.

is the row number and $0 \leq r \leq n-1$ *is the column number of the node. Node* (x, r) *is connected to node* (y, s) *if and only if* $s = r + 1 \bmod n$ *and either:*

1. $x = y$ *(the "direct" edge); or*
2. *x and y differ in precisely the* sth *bit in their binary representation (the "flip" edge).*

See Figure 4.2. To see the relation between the wrapped butterfly and the hypercube, observe that by collapsing the n nodes in each row of the wrapped butterfly into one "super node" we obtain an n-cube network. Using this correspondence, one can easily verify that there is a unique directed path of length n connecting node (x, r) to any other node (w, r) in the same column. This path is obtained by bit fixing: first fixing bits $r+1$ to n, then bits 1 to r. See Algorithm 4.3. Our randomized permutation routing algorithm on the butterfly consists of three phases, as shown in Algorithm 4.4.

Unlike our analysis of the hypercube, our analysis here cannot simply bound the number of active packets that possibly traverse edges of a path. Given the path of a packet, the expected number of other packets that share edges with this path when

Wrapped Butterfly Bit-Fixing Routing Algorithm:

1. Let (x, r) and (y, r) be the origin and the destination of a packet.
2. For $i = 0$ to $n - 1$, do:
   (a) $j = ((i + r) \bmod n) + 1$;
   (b) if $a_j = b_j$ then traverse the direct edge to column $j \bmod n$, else traverse the flip edge to column $j \bmod n$.

Algorithm 4.3: Wrapped butterfly bit-fixing routing algorithm.

Three-Phase Routing Algorithm:

For a packet sent from node (x, r) to node (y, s):

Phase I – Choose a random $w \in [1, \ldots, 2^n]$. Route the packet from node (x, r) to node (w, r) using the bit-fixing route.
Phase II – Route the packet to node (w, s) using direct edges.
Phase III – Route the packet from node (w, s) to node (y, s) using the bit-fixing route.

Algorithm 4.4: Three-phase routing algorithm.

routing a random permutation on the butterfly network is $\Omega(n^2)$ and not $O(n)$ as in the n-cube. To obtain an $O(n)$ routing time, we need a more refined analysis technique that takes into account the order in which packets traverse edges.

Because of this, we need to consider the priority policy that the queues use when there are several packets waiting to use the edge. A variety of priority policies would work here; we assume the following rules.

1. The priority of a packet traversing an edge is $(i - 1)n + t$, where i is the current phase of the packet and t is the number of edge traversals the packet has already executed in this phase.
2. If at any step more than one packet is available to traverse an edge, the packet with the smallest priority number is sent first.

Theorem 4.17: *Given an arbitrary permutation routing problem on the wrapped butterfly with $N = n2^n$ nodes, with probability $1 - O(N^{-1})$ the three-phase routing scheme of Algorithm 4.4 routes all packets to their destinations in $O(n) = O(\log N)$ parallel steps.*

Proof: The priority rule in the edge queues guarantees that packets in a phase cannot delay packets in earlier phases. Because of this, in our forthcoming analysis we can consider the time for each phase to complete separately and then add these times to bound the total time for the three-phase routing scheme to complete.

We begin by considering the second phase. We first argue that with high probability each row transmits at most $4n$ packets in the second phase. To see this, let X_w be the

number of packets whose intermediate row choice is w in the three-phase routing algorithm. Then X_w is the sum of 0–1 independent random variables, one for each packet, and $\mathbf{E}[X_w] = n$. Hence, we can directly apply the Chernoff bound of Eqn. (4.1) to find

$$\Pr(X_w \geq 4n) \leq \left(\frac{e^3}{4^4}\right)^n \leq 3^{-2n}.$$

There are 2^n possible rows w. By the union bound, the probability that any row has more than $4n$ packets is only $2^n \cdot 3^{-2n} = O(N^{-1})$.

We now argue that, if each row has at most $4n$ packets for the second phase, then the second phase takes at most $5n$ steps to complete. Combined with our previous observations, this means the second phase takes at most $5n$ steps with probability $1 - O(N^{-1})$. To see this, note that in the second phase the routing has a special structure: each packet moves from edge to edge along its row. Because of the priority rule, each packet can be delayed only by packets already in a queue when it arrives. Therefore, to place an upper bound on the number of packets that delay a packet p, we can bound the total number of packets found in each queue when p arrives at the queue. But in Phase II, the number of other packets that an arriving packet finds in a queue cannot increase in size over time, since at each step a queue sends a packet and receives at most one packet. (It is worth considering the special case when a queue becomes empty at some point in Phase II; this queue can receive another packet at some later step, but the number of packets an arriving packet will find in the queue after that point is always zero.) Since there are at most $4n$ packets total in the row to begin with, p finds at most $4n$ packets that delay it as it moves from queue to queue. Since each packet moves at most n times in the second phase, the total time for the phase is $5n$ steps.

We now consider the other phases. The first and third phases are again the same by symmetry, so we consider just the first phase. Our analysis will use a delay sequence argument. ■

Definition 4.5: *A* delay sequence *for an execution of Phase I is a sequence of n edges $e_1, \ldots, e_n$ such that either $e_i = e_{i+1}$ or e_{i+1} is an outgoing edge from the end vertex of e_i. The sequence $e_1, \ldots, e_n$ has the further property that e_i is (one of) the last edges to transmit packets with priority up to i among e_{i+1} and the two incoming edges of e_{i+1}.*

The relation between the delay sequence and the time for Phase I to complete is given by the following lemma.

Lemma 4.18: *For a given execution of Phase I and delay sequence $e_1, \ldots, e_n$, let t_i be the number of packets with priority i sent through edge e_i. Let T_i be the time that edge e_i finishes sending all packets with priority number up to i, so that T_n is the earliest time at which all packets passing through e_n during Phase I have passed through it. Then:*

1. $T_n \leq \sum_{i=1}^{n} t_i$.
2. *If the execution of Phase I takes T steps, then there is a delay sequence for this execution for which $\sum_{i=1}^{n} t_i \geq T$.*

Proof: By the design of the delay sequence, at time T_i the queue of e_{i+1} already holds all of the packets that it will need to subsequently transmit with priority $i + 1$, and at

that time it has already finished transmitting all packets with priority numbers up to i. Thus,

$$T_{i+1} \leq T_i + t_{i+1}.$$

Since $T_1 = t_1$, we have

$$\begin{aligned} T_n &\leq T_{n-1} + t_n \\ &\leq T_{n-2} + t_{n-1} + t_n \\ &\leq \sum_{i=1}^{n} t_i, \end{aligned}$$

proving the first part of the lemma.

For the second part, assume that Phase I took T steps and let e be an edge that transmitted a packet at time T. We can construct a delay sequence with $e_n = e$ by choosing e_{n-1} to be the edge among e and its two incoming edges that last transmits packets of priority $n-1$, and similarly choosing e_{n-2} down to e_1. By the first part of the lemma, $\sum_{i=1}^{n} t_i \geq T$. ∎

Returning to the proof of Theorem 4.17, we now show that the probability of a delay sequence with $T \geq 40n$ is only $O(N^{-1})$. We call any sequence of edges $e_1, \ldots, e_n$ such that either $e_i = e_{i+1}$ or e_{i+1} is an outgoing edge from the end vertex of e_i a *possible delay sequence.* For a given execution and a possible delay sequence, let t_i be the number of packets with priority i sent through e_i. Let $T = \sum_{i=1}^{n} t_i$. We first bound $\mathbf{E}[T]$. Consider the edge $e_i = v \rightarrow v'$. Packets with priority i pass through this edge only if their source is at distance $i-1$ from v. There are precisely 2^{i-1} nodes that are connected to v by a directed path of length $i-1$. Since packets are sent in Phase I to random destinations, the probability that each of these nodes sends a packet that traverses edge e_i is 2^{-i}, giving

$$\mathbf{E}[t_i] = 2^{i-1}2^{-i} = \frac{1}{2} \quad \text{and} \quad \mathbf{E}[T] = \frac{n}{2}.$$

The motivation for using the delay sequence argument should now be clear. Each possible delay sequence defines a random variable T, where $\mathbf{E}[T] = n/2$. The maximum of T over all delay sequences bounds the run-time of the phase. So we need a bound on T that holds with sufficiently high probability to cover all possible delay sequences. A high-probability bound on T can now be obtained using an argument similar to the one used in the proof of Theorem 4.15. We first bound the number of different packets that contribute to edge traversals counted in T.

For $j = 1, \ldots, N$, let $H_j = 1$ if any traversal of the packet sent by node j is counted in T; otherwise, $H_j = 0$. Clearly, $H = \sum_{j=1}^{N} H_j \leq T$ and $\mathbf{E}[H] \leq \mathbf{E}[T] = n/2$, where the H_j are independent random variables. Applying the Chernoff bound of Eqn. (4.3) therefore yields

$$\Pr(H \geq 5n) \leq 2^{-5n}.$$

Conditioning on the event $H \leq 5n$, we now proceed to prove a bound on T, following the same line as in the proof of Theorem 4.15. Given a packet u with at least one

traversal counted in T, we consider how many additional traversals of u are counted in T. Specifically, if u is counted in t_i then we consider the probability that it is counted in t_{i+1}. We distinguish between two cases as follows.

1. If $e_{i+1} = e_i$ then u cannot be counted in t_{i+1}, since its traversal with priority $i+1$ is in the next column. Similarly, it cannot be counted in any t_j, $j > i$.
2. If $e_{i+1} \neq e_i$, then the probability that u continues through e_{i+1} (and is counted in t_{i+1}) is at most $1/2$. If it does not continue through e_{i+1}, then it cannot intersect with the delay sequence in any further traversals in this phase.

As in the proof of Theorem 4.15, the probability that $T \geq 40n$ is less than the probability that a fair coin flipped $40n$ times comes up heads fewer than $5n$ times. (Keep in mind that, in this case, the first traversal by each packet in H must be counted as contributing to T.) Letting Z be the number of heads in $40n$ fair coin flips, we now apply the Chernoff bound (4.5) to prove

$$\Pr(T \geq 40n \mid H \leq 5n) \leq \Pr(Z \leq 5n) \leq e^{-20n(3/4)^2/2} \leq 2^{-5n}.$$

We conclude that

$$\Pr(T \geq 40n) \leq \Pr(T \geq 40n \mid H \leq 5n) + \Pr(H \geq 5n) \leq 2^{-5n+1}.$$

There are no more than $2N3^{n-1} \leq n2^n3^n$ possible delay sequences, since a sequence can start in any one of the $2N$ edges of the network, and by Definition 4.5, if e_i is the ith edge in the sequence, there are only three possible assignments for e_{i+1}. Thus, the probability that, in the execution of Phase I, there is a delay sequence with $T \geq 40n$ is bounded above (using the union bound) by

$$n2^n3^n2^{-5n+1} \leq O(N^{-1}).$$

Since Phase III is entirely similar to Phase I and since Phase II also finishes in $O(n)$ steps with probability $1 - O(N^{-1})$, we have that the three-phase routing algorithm finishes in $O(n)$ steps with probability $1 - O(N^{-1})$.

4.7. Exercises

Exercise 4.1: Alice and Bob play checkers often. Alice is a better player, so the probability that she wins any given game is 0.6, independent of all other games. They decide to play a tournament of n games. Bound the probability that Alice loses the tournament using a Chernoff bound.

Exercise 4.2: We have a standard six-sided die. Let X be the number of times that a 6 occurs over n throws of the die. Let p be the probability of the event $X \geq n/4$. Compare the best upper bounds on p that you can obtain using Markov's inequality, Chebyshev's inequality, and Chernoff bounds.

Exercise 4.3: (a) Determine the moment generating function for the binomial random variable $B(n, p)$.

(b) Let X be a $B(n, p)$ random variable and Y a $B(m, p)$ random variable, where X and Y are independent. Use part (a) to determine the moment generating function of $X + Y$.

(c) What can we conclude from the form of the moment generating function of $X + Y$?

Exercise 4.4: Determine the probability of obtaining 55 or more heads when flipping a fair coin 100 times by an explicit calculation, and compare this with the Chernoff bound. Do the same for 550 or more heads in 1000 flips.

Exercise 4.5: We plan to conduct an opinion poll to find out the percentage of people in a community who want its president impeached. Assume that every person answers either yes or no. If the actual fraction of people who want the president impeached is p, we want to find an estimate X of p such that

$$\Pr(|X - p| \leq \varepsilon p) > 1 - \delta$$

for a given ε and δ, with $0 < \varepsilon, \delta < 1$.

We query N people chosen independently and uniformly at random from the community and output the fraction of them who want the president impeached. How large should N be for our result to be a suitable estimator of p? Use Chernoff bounds, and express N in terms of p, ε, and δ. Calculate the value of N from your bound if $\varepsilon = 0.1$ and $\delta = 0.05$ and if you know that p is between 0.2 and 0.8.

Exercise 4.6: **(a)** In an election with two candidates using paper ballots, each vote is independently misrecorded with probability $p = 0.02$. Use a Chernoff bound to give an upper bound on the probability that more than 4% of the votes are misrecorded in an election of 1,000,000 ballots.

(b) Assume that a misrecorded ballot always counts as a vote for the other candidate. Suppose that candidate A received 510,000 votes and that candidate B received 490,000 votes. Use Chernoff bounds to upper bound the probability that candidate B wins the election owing to misrecorded ballots. Specifically, let X be the number of votes for candidate A that are misrecorded and let Y be the number of votes for candidate B that are misrecorded. Bound $\Pr((X > k) \cup (Y < \ell))$ for suitable choices of k and ℓ.

Exercise 4.7: Throughout the chapter we implicitly assumed the following extension of the Chernoff bound. Prove that it is true.

Let $X = \sum_{i=1}^{n} X_i$, where the X_i are independent 0–1 random variables. Let $\mu = \mathbf{E}[X]$. Choose any μ_L and μ_H such that $\mu_L \leq \mu \leq \mu_H$. Then, for any $\delta > 0$,

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

Similarly, for any $0 < \delta < 1$,

$$\Pr(X \leq (1 - \delta)\mu_L) \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{(1-\delta)}}\right)^{\mu_L}.$$

Exercise 4.8: We show how to construct a random permutation π on $[1, n]$, given a black box that outputs numbers independently and uniformly at random from $[1, k]$ where $k \geq n$. If we compute a function $f\,[1, n] \rightarrow [1, k]$ with $f(i) \neq f(j)$ for $i \neq j$, this yields a permutation; simply output the numbers $[1, n]$ according to the order of the $f(i)$ values. To construct such a function f, do the following for $j = 1, \ldots, n$: choose $f(j)$ by repeatedly obtaining numbers from the black box and setting $f(j)$ to the first number found such that $f(j) \neq f(i)$ for $i < j$.

Prove that this approach gives a permutation chosen uniformly at random from all permutations. Find the expected number of calls to the black box that are needed when $k = n$ and $k = 2n$. For the case $k = 2n$, argue that the probability that each call to the black box assigns a value of $f(j)$ to some j is at least $1/2$. Based on this, use a Chernoff bound to bound the probability that the number of calls to the black box is at least $4n$.

Exercise 4.9: Suppose that we can obtain independent samples $X_1, X_2, \ldots$ of a random variable X and that we want to use these samples to estimate $\mathbf{E}[X]$. Using t samples, we use $\left(\sum_{i=1}^{t} X_i\right)/t$ for our estimate of $\mathbf{E}[X]$. We want the estimate to be within $\varepsilon\mathbf{E}[X]$ from the true value of $\mathbf{E}[X]$ with probability at least $1 - \delta$. We may not be able to use Chernoff's bound directly to bound how good our estimate is if X is not a 0–1 random variable, and we do not know its moment generating function. We develop an alternative approach that requires only having a bound on the variance of X. Let $r = \sqrt{\mathbf{Var}[X]}/\mathbf{E}[X]$.

(a) Show using Chebyshev's inequality that $O(r^2/\varepsilon^2\delta)$ samples are sufficient to solve the problem.

(b) Suppose that we need only a weak estimate that is within $\varepsilon\mathbf{E}[X]$ of $\mathbf{E}[X]$ with probability at least $3/4$. Argue that $O(r^2/\varepsilon^2)$ samples are enough for this weak estimate.

(c) Show that, by taking the median of $O(\log(1/\delta))$ weak estimates, we can obtain an estimate within $\varepsilon\mathbf{E}[X]$ of $\mathbf{E}[X]$ with probability at least $1 - \delta$. Conclude that we need only $O((r^2 \log(1/\delta))/\varepsilon^2)$ samples.

Exercise 4.10: A casino is testing a new class of simple slot machines. Each game, the player puts in \$1, and the slot machine is supposed to return either \$3 to the player with probability $4/25$, \$100 with probability $1/200$, or nothing with all remaining probability. Each game is supposed to be independent of other games.

The casino has been surprised to find in testing that the machines have lost \$10,000 over the first million games. Derive a Chernoff bound for the probability of this event. You may want to use a calculator or program to help you choose appropriate values as you derive your bound.

Exercise 4.11: Consider a collection $X_1, \ldots, X_n$ of n independent integers chosen uniformly from the set $\{0, 1, 2\}$. Let $X = \sum_{i=1}^{n} X_i$ and $0 < \delta < 1$. Derive a Chernoff bound for $\Pr(X \geq (1 + \delta)n)$ and $\Pr(X \leq (1 - \delta)n)$.

Exercise 4.12: Consider a collection $X_1, \ldots, X_n$ of n independent geometrically distributed random variables with mean 2. Let $X = \sum_{i=1}^{n} X_i$ and $\delta > 0$.

(a) Derive a bound on $\Pr(X \geq (1+\delta)(2n))$ by applying the Chernoff bound to a sequence of $(1+\delta)(2n)$ fair coin tosses.
(b) Directly derive a Chernoff bound on $\Pr(X \geq (1+\delta)(2n))$ using the moment generating function for geometric random variables.
(c) Which bound is better?

Exercise 4.13: Let $X_1, \ldots, X_n$ be independent Poisson trials such that $\Pr(X_i = 1) = p$. Let $X = \sum_{i=1}^{n} X_i$, so that $\mathbf{E}[X] = pn$. Let

$$F(x, p) = x \ln(x/p) + (1-x) \ln((1-x)/(1-p)).$$

(a) Show that, for $1 \geq x > p$,

$$\Pr(X \geq xn) \leq \mathrm{e}^{-nF(x,p)}.$$

(b) Show that, when $0 < x, p < 1$, we have $F(x, p) - 2(x-p)^2 \geq 0$. (*Hint:* Take the second derivative of $F(x, p) - 2(x-p)^2$ with respect to x.)
(c) Using parts (a) and (b), argue that

$$\Pr(X \geq (p+\varepsilon)n) \leq \mathrm{e}^{-2n\varepsilon^2}.$$

(d) Use symmetry to argue that

$$\Pr(X \leq (p-\varepsilon)n) \leq \mathrm{e}^{-2n\varepsilon^2},$$

and conclude that

$$\Pr(|X - pn| \geq \varepsilon n) \leq 2\mathrm{e}^{-2n\varepsilon^2}.$$

Exercise 4.14: Modify the proof of Theorem 4.4 to show the following bound for a weighted sum of Poisson trials. Let $X_1, \ldots, X_n$ be independent Poisson trials such that $\Pr(X_i) = p_i$ and let $a_1, \ldots, a_n$ be real numbers in $[0, 1]$. Let $X = \sum_{i=1}^{n} a_i X_i$ and $\mu = \mathbf{E}[X]$. Then the following Chernoff bound holds: for any $\delta > 0$,

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

Prove a similar bound for the probability that $X \leq (1-\delta)\mu$ for any $0 < \delta < 1$.

Exercise 4.15: Let $X_1, \ldots, X_n$ be independent random variables such that

$$\Pr(X_i = 1 - p_i) = p_i \quad \text{and} \quad \Pr(X_i = -p_i) = 1 - p_i.$$

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

$$\Pr(|X| \geq a) \leq 2\mathrm{e}^{-2a^2/n}.$$

Hint: You may need to assume the inequality

$$p_i \mathrm{e}^{\lambda(1-p_i)} + (1-p_i)\mathrm{e}^{-\lambda p_i} \leq \mathrm{e}^{\lambda^2/8}.$$

This inequality is difficult to prove directly.

Exercise 4.16: Let $X_1, \ldots, X_n$ be independent Poisson trials such that $\Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} a_i X_i$ and $\mu = \mathrm{E}[X]$. Use the result of Exercise 4.15 to prove that if $|a_i| \leq 1$ for all $1 \leq i \leq n$, then for any $0 < \delta < 1$,

$$\Pr(|X - \mu| \geq \delta\mu) \leq 2e^{-2\delta^2\mu^2/n}.$$

Exercise 4.17: Suppose that we have n jobs to distribute among m processors. For simplicity, we assume that m divides n. A job takes 1 step with probability p and $k > 1$ steps with probability $1 - p$. Use Chernoff bounds to determine upper and lower bounds (that hold with high probability) on when all jobs will be completed if we randomly assign exactly n/m jobs to each processor.

Exercise 4.18: In many wireless communication systems, each receiver listens on a specific frequency. The bit $b(t)$ sent at time t is represented by a 1 or -1. Unfortunately, noise from other nearby communications can affect the receiver's signal. A simplified model of this noise is as follows. There are n other senders, and the ith has strength $p_i \leq 1$. At any time t, the ith sender is also trying to send a bit $b_i(t)$ that is represented by 1 or -1. The receiver obtains the signal $s(t)$ given by

$$s(t) = b(t) + \sum_{i=1}^{n} p_i b_i(t).$$

If $s(t)$ is closer to 1 than -1, the receiver assumes that the bit sent at time t was a 1; otherwise, the receiver assumes that it was a -1.

Assume that all the bits $b_i(t)$ can be considered independent, uniform random variables. Give a Chernoff bound to estimate the probability that the receiver makes an error in determining $b(t)$.

Exercise 4.19: Recall that a function f is said to be *convex* if, for any x_1, x_2 and for $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

(a) Let Z be a random variable that takes on a (finite) set of values in the interval $[0, 1]$, and let $p = \mathrm{E}[Z]$. Define the Bernoulli random variable X by $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$. Show that $\mathrm{E}[f(Z)] \leq \mathrm{E}[f(X)]$ for any convex function f.

(b) Use the fact that $f(x) = e^{tx}$ is convex for any $t \geq 0$ to obtain a Chernoff bound for the sum of n independent random variables with distribution Z as in part (a), based on a Chernoff bound for independent Poisson trials.

Exercise 4.20: Prove Theorem 4.14.

Exercise 4.21: We prove that the Randomized Quicksort algorithm sorts a set of n numbers in time $O(n \log n)$ with high probability. Consider the following view of Randomized Quicksort. Every point in the algorithm where it decides on a pivot element is called a *node.* Suppose the size of the set to be sorted at a particular node is s. The node is called *good* if the pivot element divides the set into two parts, each of size not exceeding $2s/3$. Otherwise the node is called *bad.* The nodes can be thought of as

forming a tree in which the root node has the whole set to be sorted and its children have the two sets formed after the first pivot step and so on.

(a) Show that the number of good nodes in any path from the root to a leaf in this tree is not greater than $c \log_2 n$, where c is some positive constant.
(b) Show that, with high probability (greater than $1 - 1/n^2$), the number of nodes in a given root to leaf path of the tree is not greater than $c' \log_2 n$, where c' is another constant.
(c) Show that, with high probability (greater than $1 - 1/n$), the number of nodes in the longest root to leaf path is not greater than $c' \log_2 n$. (*Hint:* How many nodes are there in the tree?)
(d) Use your answers to show that the running time of Quicksort is $O(n \log n)$ with probability at least $1 - 1/n$.

Exercise 4.22: Consider the bit-fixing routing algorithm for routing a permutation on the n-cube. Suppose that n is even. Write each source node s as the concatenation of two binary strings a_s and b_s each of length $n/2$. Let the destination of s's packet be the concatenation of b_s and a_s. Show that this permutation causes the bit-fixing routing algorithm to take $\Omega(\sqrt{N})$ steps.

Exercise 4.23: Consider the following modification to the bit-fixing routing algorithm for routing a permutation on the n-cube. Suppose that, instead of fixing the bits in order from 1 to n, each packet chooses a random order (independent of other packets' choices) and fixes the bits in that order. Show that there is a permutation for which this algorithm requires $2^{\Omega(n)}$ steps with high probability.

Exercise 4.24: Assume that we use the randomized routing algorithm for the n-cube network (Algorithm 4.2) to route a total of up to $p2^n$ packets, where each node is the source of no more than p packets and each node is the destination of no more than p packets.

(a) Give a high-probability bound on the run-time of the algorithm.
(b) Give a high-probability bound on the maximum number of packets at any node at any step of the execution of the routing algorithm.

Exercise 4.25: Show that the expected number of packets that traverse any edge on the path of a given packet when routing a random permutation on the wrapped butterfly network of $N = n2^n$ nodes is $\Omega(n^2)$.

Exercise 4.26: In this exercise, we design a randomized algorithm for the following packet routing problem. We are given a network that is an undirected connected graph G, where nodes represent processors and the edges between the nodes represent wires. We are also given a set of N packets to route. For each packet we are given a source node, a destination node, and the exact route (path in the graph) that the packet should take from the source to its destination. (We may assume that there are no loops in the

path.) In each time step, at most one packet can traverse an edge. A packet can wait at any node during any time step, and we assume unbounded queue sizes at each node.

A *schedule* for a set of packets specifies the timing for the movement of packets along their respective routes. That is, it specifies which packet should move and which should wait at each time step. Our goal is to produce a schedule for the packets that tries to minimize the total time and the maximum queue size needed to route all the packets to their destinations.

(a) The dilation d is the maximum distance traveled by any packet. The congestion c is the maximum number of packets that must traverse a single edge during the entire course of the routing. Argue that the time required for any schedule should be at least $\Omega(c + d)$.

(b) Consider the following unconstrained schedule, where many packets may traverse an edge during a single time step. Assign each packet an integral delay chosen randomly, independently, and uniformly from the interval $[1, \lceil \alpha c / \log(Nd) \rceil]$, where α is a constant. A packet that is assigned a delay of x waits in its source node for x time steps; then it moves on to its final destination through its specified route without ever stopping. Give an upper bound on the probability that more than $O(\log(Nd))$ packets use a particular edge e at a particular time step t.

(c) Again using the unconstrained schedule of part (b), show that the probability that more than $O(\log(Nd))$ packets pass through any edge at any time step is at most $1/(Nd)$ for a sufficiently large α.

(d) Use the unconstrained schedule to devise a simple randomized algorithm that, with high probability, produces a schedule of length $O(c + d \log(Nd))$ using queues of size $O(\log(Nd))$ and following the constraint that at most one packet crosses an edge per time step.