Chapter 4

The First and Second Moment Methods

When analyzing randomized algorithms or data structures with random inputs, we are usually interested in time complexities, storage requirements or the value of a particular parameter characterizing the algorithm. These quantities are *random* (non-deterministic) since either the input is assumed to vary or the algorithm itself makes random decisions. We often are satisfied with average values, however for many applications this might be too simplistic. In many instances these quantities of interest, though random, behave in a very deterministic manner for large inputs. The *first and second moment methods* are the most popular probabilistic tools used to derive such relationships. We discuss them in this chapter together with some interesting applications.

IMAGINE that X_n is a random variable representing a quantity of interest (e.g., the number of pattern occurrences in a random text of length n). Usually, it is not difficult to compute the average $\mathbf{E}[X_n]$ of X_n but it is equally easy to come up with examples that show how poor such a measure of variability of X_n could be. But, often we can discover a refined information about X_n; e.g., that with high probability $X_n \sim \mathbf{E}[X_n]$ as $n \to \infty$. In this case, the random variable X_n converges (in probability or almost surely) to a *deterministic* value $a_n = \mathbf{E}[X_n]$ when n is larger and larger (cf. Example 2).

Consider the following example: Let $M_n = \max\{C_1, \ldots, C_n\}$ where C_i are dependent random variables (e.g., M_n may represent the longest path in a digital tree). Again, M_n varies in a random fashion but for large n we shall find out that $M_n \sim a_n$ where a_n is a (deterministic) sequence (cf. Example 1). More precisely, for every $\varepsilon > 0$ the probability of $(1-\varepsilon)a_n \le M_n \le (1+\varepsilon)a_n$ becomes closer and closer to 1, thus according to the definition from Section 2.2.2 we say that $M_n \stackrel{pr}{\to} a_n$ or $M_n \to a_n$ (pr.).

These and other problems can be tackled and solved successfully by two simple probabilistic methods called the **first and second moment methods**. They resemble Boole's and Bonferroni's inequalities, but instead of bounding the probability of a union of events by marginal probabilities they use the first and the second moments. We also briefly discuss the **fourth moment method** recently proposed by B. Berger [38].

Here is the plan for the chapter: We first discuss some theoretical underpinning of the methods which we illustrate through a handful examples. Then, we present several applications such as a Markov approximation of a stationary distribution, computing the number of primes dividing a given number, and estimating the height in tries, PATRICIA tries, digital search trees and suffix trees.

4.1 The Methods

Let us start by recalling Markov's inequality (2.6) from Chapter 2: For a nonnegative random variable X, we can bound the probability that X exceeds $t > 0$ as follows: $\Pr\{X \geq t\} \leq \mathbf{E}[X]/t$. If in addition, X is integer-valued random variable, after setting $t = 1$ we obtain

$$\Pr\{X > 0\} \leq \mathbf{E}[X], \tag{4.1}$$

which is the *first moment method*. We can derive it in another manner (without restoring to the Markov inequality). Observe that

$$\Pr\{X > 0\} = \sum_{k=1}^{\infty} \Pr\{X = k\} \leq \sum_{k=0}^{\infty} k\Pr\{X = k\} = \mathbf{E}[X].$$

The above implies Boole's inequality (3.6). Indeed, let A_i $(i = 1, \ldots, n)$ be events, and set $X = I(A_1) + \cdots + I(A_n)$ where, as before, $I(A) = 1$ if A occurs, and zero otherwise. Inequality (3.6) follows.

In a typical application of (4.1), we expect to show that $\mathbf{E}[X] \to 0$, and hence that $X = 0$ occurs **with high probability** (**whp**). Throughout, we shall often write **whp** instead a more formal "convergence in probability". We illustrate the first moment method on a simple example.

Example 4.1: *Maximum of Dependent Random Variables*

Let $M_n = \max\{X_1, \ldots, X_n\}$ where X_i are dependent but identically distributed random variables with the distribution function $F(\cdot)$. We assume that the distribution function is such that for all $c > 1$

$$1 - F(cy) = \delta(y)(1 - F(y)),$$

where $\delta(y) \to 0$ as $y \to \infty$. We shall prove that **whp** $M_n/a_n \leq 1$ where a_n is the smallest solution of the following equation:

$$n(1 - F(a_n)) = 1,$$

provided $a_n \to \infty$ as $n \to \infty$. Indeed, observe that $\{M_n > x\} = \bigcup_{i=1}^{n} A_i$ where $A_i = \{X_i > x\}$. Thus, by the first moment method,

$$\Pr\{M_n > x\} \leq n\mathbf{E}[I(A_1)] = n(1 - F(x)).$$

Let us set now $x = (1 + \varepsilon)a_n$ for any $\varepsilon > 0$, where a_n is defined as above. Then

$$\Pr\{M_n > (1 + \varepsilon)a_n\} \leq n(1 - F((1 + \varepsilon)a_n)) = \delta(a_n)n(1 - F(a_n)) = \delta(a_n) \to 0,$$

since $a_n \to \infty$. Thus $M_n \leq (1 + \varepsilon)a_n$ **whp**, as desired (cf. also Chapter 3.3.2). □

The first moment method uses only the knowledge of the average value of X to bound the probability. A better estimate of the probability can be obtained if we know the variance of X. Indeed, Chebyshev already noticed that (cf. (2.7))

$$\Pr\{|X - \mathbf{E}[X]| \geq t\} \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Setting in the Chebyshev inequality $t = |\mathbf{E}[X]|$ we arrive at

$$\Pr\{X = 0\} \leq \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2}. \tag{4.2}$$

Indeed, we have

$$\Pr\{X = 0\} \leq \Pr\{X\,(X - 2\mathbf{E}[X]) \geq 0\} = \Pr\{|X - \mathbf{E}[X]| \geq |\mathbf{E}[X]|\} \leq \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2}$$

Inequality (4.2) is known as (Chebyshev's version of) the *second moment method.*

We now derive a refinements of (4.2). Shepp [374] proposed to apply Schwarz's inequality (2.8) to obtain the following:

$$\mathbf{E}[X]^2 = \mathbf{E}[I(X \neq 0)X]^2 \leq \mathbf{E}[I(X \neq 0)]\mathbf{E}[X^2] = \Pr\{I(X \neq 0)\}\mathbf{E}[X^2],$$

which leads to a refinement of the second moment inequality

$$\Pr\{X = 0\} \leq \frac{\mathbf{Var}[X]}{\mathbf{E}[X^2]}. \tag{4.3}$$

Actually, another formulation of Shepp's inequality due to Chung and Erdős is quite useful. Consider $X_n = I(A_1) + \cdots + I(A_n)$ for a sequence of events $A_1, \ldots, A_n$, and observe that $\{X_n > 0\} = \bigcup_{i=1}^n A_i$. Rewriting (4.3) as

$$\Pr\{X_n > 0\} \geq \frac{\mathbf{E}[X_n]^2}{\mathbf{E}[X_n^2]},$$

we obtain after some simple algebra

$$\Pr\{\bigcup_{i=1}^n A_i\} \geq \frac{(\sum_{i=1}^n \Pr\{A_i\})^2}{\sum_{i=1}^n \Pr\{A_i\} + \sum_{i \neq j} \Pr\{A_i \cap A_j\}}\,. \tag{4.4}$$

In a typical application, if we are able to prove that $\mathbf{Var}[X_n]/\mathbf{E}[X_n^2] \to 0$, then $X_n \xrightarrow{pr} \mathbf{E}[X_n]$ as $n \to \infty$, that is, for any $\varepsilon > 0$

$$\lim_{n \to \infty} \Pr\{|X_n - \mathbf{E}[X_n]| \geq \varepsilon \mathbf{E}[X_n]\} = 0.$$

We present below two simple examples delaying a discussion of a more sophisticated problem till the application Section 4.2.

Example 4.2: *Given Pattern Occurrences in a Random Text*

Let H be a given string (pattern) of length m, and T be a random string of length n. Both strings are over a finite alphabet, and we assume that $m = o(n)$. Our goal is to find a typical (in a probabilistic sense) behavior of the number of the pattern occurrences in T. We denote this by O_n. Is it true that $O_n \sim \mathbf{E}[O_n]$ **whp**? We compute the mean $\mathbf{E}[O_n]$ and the variance $\mathbf{Var}[O_n]$. Let I_i be a random variable equal to 1 if the pattern H occurs at position i, and 0 otherwise. Clearly $O_n = I_1 + I_2 + \cdots + I_{n-m+1}$, and hence

$$\mathbf{E}[O_n] = \sum_{i=1}^{n-m+1} \mathbf{E}[I_i] = (n-m+1)P(H),$$

where $P(H) = \mathsf{Pr}\{T_i^{i+m-1} = H\}$ for some $1 \le i \le n-m+1$, that is, the probability that the pattern H occurs at position i of the text T. The variance is also easy to compute since

$$\mathbf{Var}[O_n] = \sum_{i=1}^{n-m+1} \mathbf{E}[I_i^2] + \sum_{1 \le i < j \le n-m+1} \mathbf{Cov}[I_i I_j] = \sum_{i=1}^{n-m+1} \mathbf{E}[I_i] + \sum_{1 \le i < j \le n-m+1} \mathbf{Cov}[I_i I_j],$$

where

$$\mathbf{Cov}[I_i I_j] = \mathbf{E}[I_i I_j] - \mathbf{E}[I_i]\mathbf{E}[I_j].$$

Let us compute the covariance $\mathbf{Cov}[I_i I_j]$. Observe that $\mathbf{Cov}[I_i I_j] = 0$ for $|j-i| > m$, and otherwise $\mathbf{Cov}[I_i I_j] \le \mathbf{E}[I_i I_j] \le \mathbf{E}[I_i] = P(H)$. Thus

$$\mathbf{Var}[O_n] \le (n-m+1)P(H) + \sum_{|j-i| \le m} \mathbf{E}[I_i] \le (n-m+1)P(H) + 2m^2 P(H).$$

In view of the above, and Chebyshev's inequality we obtain for any $\varepsilon > 0$

$$\begin{aligned} \mathsf{Pr}\{|O_n/\mathbf{E}[O_n] - 1| > \varepsilon\} &\le \frac{(n-m+1+2m^2)P(H)}{\varepsilon^2 (n-m+1)^2 P^2(H)} \\ &\le \frac{1}{\varepsilon^2 (n-m+1)P(H)} + \frac{2m^2}{\varepsilon^2 (n-m+1)^2 P(H)} \to 0, \end{aligned}$$

provided $m = o(n)$ (we remind the reader that H is *given*). This proves that $O_n/\mathbf{E}[O_n] \to 1$ **whp**. A fuller discussion of this problem can be found in Section 7.6.2. □

In the next example, we consider a *deterministic* problem and use randomized technique to solve it. It is adopted from Alon and Spencer [9].

Example 4.3: *Erdős Distinct Sum Problem*

Consider a set of positive integers $\{x_1, \ldots, x_k\} \subset \{1, \ldots, n\}$. Let $f(n)$ denote the maximal k such that there exists a set $\{x_1, \ldots, x_k\}$ with *distinct* sums. The simplest set with distinct sums is $\{2^i : i \le \log_2 n\}$, which also shows that $f(n) \ge 1 + \lfloor \log_2 n \rfloor$. Erdős asked to prove that

$$f(n) \le \log_2 n + C$$

for some constant C. We shall show that

$$f(n) \le \log_2 n + \log_2 \log_2 n + O(1).$$

Indeed, let $\{x_1, \ldots, x_k\}$ be a distinct sum set, and let B_i be a random variable taking values 0 and 1 with equal probability $\frac{1}{2}$. Consider

$$X = B_1 x_1 + \cdots + B_k x_k.$$

Certainly,

$$\begin{aligned} \mathbf{E}[X] &= \frac{x_1 + \cdots + x_k}{2}, \\ \mathbf{Var}[X] &= \frac{x_1^2 + \cdots + x_k^2}{4} \le \frac{n^2 k}{4}. \end{aligned}$$

Using Chebyshev's inequality with $t = \lambda\sqrt{\mathbf{Var}[X]}$ we obtain, after reversing the inequality,

$$1 - \frac{1}{\lambda^2} \le \Pr\{|X - \mathbf{E}[X]| \le \lambda n \sqrt{k}/2\}$$

for some $\lambda > 1$. On the other hand, we should observe that due to the assumption that $\{x_1, \ldots, x_k\}$ is a distinct sum set, the probability that X has a particular value that is equal either to 0 or 2^{-k}. Since there are $\lambda n \sqrt{k}$ values within $|X - \mathbf{E}[X]| \le \lambda n \sqrt{k}/2$, we obtain

$$\Pr\{|X - \mathbf{E}[X]| \le \lambda n \sqrt{k}/2\} \le 2^{-k} \lambda n \sqrt{k}.$$

Comparing the above two inequalities leads to

$$n \ge \frac{2^k}{\sqrt{k}} \frac{1 - \lambda^{-2}}{\lambda}.$$

After setting the optimal $\lambda = \sqrt{3}$ we arrive at

$$f(n) \le \log_2 n + \log_2 \log_2 n + O(1),$$

which completes the proof of the claim. □

Finally, we discuss the *fourth moment method* due to Berger [38]. We shall follow the presentation of Coffman and Lueker [63]. We start with a simple (deterministic) inequality. Consider for $x \geq 0$ the function

$$f(x) = \frac{qx - x^3}{q^{3/2}},$$

where q is a parameter. It is easy to check that the function reaches its maximum value $f_{\max} = 2/(3\sqrt{3})$ at $x = \sqrt{q/3}$. Thus we can write

$$x \geq \frac{3\sqrt{3}}{2\sqrt{q}} \left(x^2 - x^4/q\right).$$

If we now replace x by a random variable, say Y, take the expectation of both sides of the above, then we arrive at

$$\mathbf{E}[|Y|] \geq \frac{3\sqrt{3}}{2\sqrt{q}} \left(\mathbf{E}[Y^2] - \frac{\mathbf{E}[Y^4]}{q}\right).$$

After setting $q = 3\mathbf{E}[Y^4]/\mathbf{E}[Y^2]$, we finally obtain Berger's fourth moment inequality

$$\mathbf{E}[|Y|] \geq \frac{\mathbf{E}[Y^2]^{3/2}}{\mathbf{E}[Y^4]^{1/2}}. \tag{4.5}$$

We illustrate the fourth moment method on a simple example, while the reader is encouraged to read Berger [38] for considerably deeper analysis.

Example 4.4: *Discrepancy of a Sum of Random Variables*

As in [63], consider $Y = X_1 + \cdots + X_n$ where $X_i \in \{-1, 0, 1\}$ are i.i.d. and $\Pr\{X_i = -1\} = \Pr\{X_i = +1\} = p$ while $\Pr\{X_i = 0\} = 1 - 2p$. Actually, we allow p to vary with n provided $np \geq 1$. We wish to find a lower bound for $\mathbf{E}[|Y|]$. Observe that $\mathbf{E}[Y^2] = np$ and

$$\mathbf{E}[Y^4] = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \mathbf{E}[X_i X_j X_k X_l].$$

When all indices i, j, k, l are different, then clearly $\mathbf{E}[X_i X_j X_k X_l] = 0$. In fact, $\mathbf{E}[X_i X_j X_k X_l]$ is not vanishing *only if* all four indices are the same or two of them are equal to one value and the other two equal to another value. But there is only $O(n^2)$ such indices, thus $\mathbf{E}[Y^4] = \Theta(n^2 p^2)$, and by the fourth moment inequality we prove that

$$\mathbf{E}[|Y|] = \Omega(\sqrt{np}).$$

In fact, one can prove that $\mathbf{E}[|Y|] = \Theta(\sqrt{np})$. □

We summarize this section by presenting all the results derived so far.

Theorem 4.1 *Let X be a random variable.*

[**First Moment Method**] *If X is nonnegative, integer-valued, then*

$$\Pr\{X > 0\} \leq \mathbf{E}[X].$$

[**Chebyshev's Second Moment Method**]

$$\Pr\{X = 0\} \leq \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2}.$$

[**Shepp's Second Moment Method**]

$$\Pr\{X = 0\} \leq \frac{\mathbf{Var}[X]}{\mathbf{E}[X^2]}.$$

[**Chung and Erdős' Second Moment Method**] *For any set of events $A_1, \ldots, A_n$*

$$\Pr\{\bigcup_{i=1}^{n} A_i\} \geq \frac{\left(\sum_{i=1}^{n} \Pr\{A_i\}\right)^2}{\sum_{i=1}^{n} \Pr\{A_i\} + \sum_{i \neq j} \Pr\{A_i \cap A_j\}}.$$

[**Fourth Moment Method**] *For an arbitrary random variable Y*

$$\mathbf{E}[|Y|] \geq \frac{\mathbf{E}[Y^2]^{3/2}}{\mathbf{E}[Y^4]^{1/2}},$$

provided it possess the first four moments.

4.2 Applications

It is time to apply what we have learned so far to some interesting problems. We start this section with a kth order Markov approximation of a stationary distribution (cf. Section 4.2.1). Then we shall present the Hardy and Ramanujan result [181] concerning the number of primes dividing a large integer n (cf. Section 4.2.2), and finally we consider the height in digital trees and suffix trees (cf. Sections 4.2.3–4.2.6). In Chapter 6 we shall discuss other applications of information theoretical flavor in which the methods of this chapter are used.

4.2.1 Markov Approximation of a Stationary Distribution

In information theory and other applications, one often approximates a stationary process by a kth order Markov chain. Such an approximation can be easily derived through the Markov inequality, hence the first moment method. We discuss it next. The presentation below is adopted from Cover and Thomas [69].

Let $\{X_k\}_{k=-\infty}^{\infty}$ be a two sided stationary process with the underlying distribution $P(x_1^n) = \Pr\{X_1^n = x_1^n\}$. We consider the *random* variable $P(X_0^{n-1})$ and estimate its probabilistic behavior through a kth order Markov approximation. The latter is defined as follows:

$$P^k(X_0^{n-1}) := P(X_0^{k-1}) \prod_{i=k}^{n-1} P(X_i|X_{i-k}^{i-1})$$

for fixed k.

Let us evaluate the ratio $P^k(X_0^{n-1})/P(X_0^{n-1})$ by the first moment method. First of all, we have

$$\begin{aligned}\mathbf{E}\left[\frac{P^k(X_0^{n-1})}{P(X_0^{n-1})}\right] &= \sum_{x_0^{n-1}} P(x_0^{n-1})\frac{P^k(x_0^{n-1})}{P(x_0^{n-1})}\\ &= \sum_{x_0^{n-1}} P^k(x_0^{n-1}) \le 1.\end{aligned}$$

Now, by Markov's inequality, for any $a_n \to \infty$

$$\Pr\left\{\frac{1}{n}\log\frac{P^k(X_0^{n-1})}{P(X_0^{n-1})} \ge \frac{1}{n}\log a_n\right\} = \Pr\left\{\frac{P^k(X_0^{n-1})}{P(X_0^{n-1})} \ge a_n\right\} \le \frac{1}{a_n}.$$

Taking $a_n = n^2$, we conclude by the Borel-Cantelli lemma that the event

$$\frac{1}{n}\log\frac{P^k(X_0^{n-1})}{P(X_0^{n-1})} \ge \frac{1}{n}\log a_n$$

occurs only finitely often with probability one. Hence,

$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{P^k(X_0^{n-1})}{P(X_0^{n-1})} \le 0 \qquad \text{(a.s.)}.$$

Actually, we can also find an upper bound on $P(X_0^{n-1})$. Let us consider $P(X_0^{n-1}|X_{-\infty}^{-1})$ as an approximation of $P(X_0^{n-1})$. We shall use Markov's inequality, so we must first evaluate

$$\begin{aligned}\mathbf{E}\left[\frac{P(X_0^{n-1})}{P(X_0^{n-1}|X_{-\infty}^{-1})}\right] &= \mathbf{E}_{X_{-\infty}^{-1}}\left[\mathbf{E}_{X_0^{n-1}}\left[\frac{P(X_0^{n-1})}{P(X_0^{n-1}|X_{-\infty}^{-1})}\,\Big|X_{-\infty}^{-1}\right]\right]\\ &= \mathbf{E}_{X_{-\infty}^{-1}}\left[\sum_{x_0^{n-1}}\frac{P(x_0^{n-1})}{P(x_0^{n-1}|X_{-\infty}^{-1})}P(x_0^{n-1}|X_{-\infty}^{-1})\right] \le 1.\end{aligned}$$

By the same argument as above we conclude that

$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{P(X_0^{n-1})}{P(X_0^{n-1}|X_{-\infty}^{-1})} \le 0 \qquad \text{(a.s.)}.$$

In summary, we formulate the following lemma (cf. [8, 69]).

Lemma 4.2 (Algoet and Cover 1988 Sandwich Lemma) *Let P be a stationary measure and P^k its kth order Markov approximation, as defined above. Then*

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{P^k(X_0^{n-1})}{P(X_0^{n-1})} \leq 0 \quad \text{(a.s.)}, \tag{4.6}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{P(X_0^{n-1})}{P(X_0^{n-1}|X_{-\infty}^{-1})} \leq 0 \quad \text{(a.s.)} \tag{4.7}$$

for large n.

In fact, the proof of the above lemma allows us to conclude the following: Given $\varepsilon > 0$, there exists N_ε such that for $n > N_\varepsilon$ with probability at least $1 - \varepsilon$ we have

$$\frac{1}{n^2} P^k(X_0^{n-1}) \leq P(X_0^{n-1}) \leq n^2 P(X_0^{n-1}|X_{-\infty}^{-1}).$$

The "sandwich lemma" was proved by Algoet and Cover [8] who used it to establish the existence of the entropy rate for general stationary processes. We shall return to it in Chapter 6.

4.2.2 Excursion into Number Theory

We present below Turán's proof of the Hardy and Ramanujan result concerning the number of primes dividing a number chosen randomly between 1 and n. We denote the number by $\nu(n)$, and we prove that n has close to $\ln \ln n$ prime factors. We use the second moment method, thus we again solve a deterministic problem by a probabilistic tool. We shall follow the exposition of Alon and Spencer [9].

Let throughout this section p denote a prime number. We need two well know results from number theory (cf. [181, 413]), namely,

$$\sum_{p \leq x} \frac{1}{p} = \ln \ln x + C + o(1), \tag{4.8}$$

$$\pi(x) = \frac{x}{\ln x}(1 + o(1)), \tag{4.9}$$

where C is a constant $C = 0.26149\ldots$, and $\pi(x)$ denotes the number of primes smaller than x.

We now choose x randomly from the set $\{1, \ldots, n\}$. For prime p we set $X_p = 1$ if $p|x$, and zero otherwise. We ask for the number of prime factors $\nu(x)$ of x. Observe that

$$X := \sum_{p \leq n} X_p = \nu(x).$$

Since x can be chosen on n different ways, and in $\lfloor n/p \rfloor$ cases it will be divisible by p, we easily find that

$$\mathbf{E}[X_p] = \frac{\lfloor n/p \rfloor}{n} = \frac{1}{p} + O\left(\frac{1}{n}\right),$$

and by (4.8) we also have

$$\mathbf{E}[X] = \sum_{p \le n} \left(\frac{1}{p} + O\left(\frac{1}{n}\right)\right) = \ln\ln n + O(1).$$

Now we bound the variance

$$\mathbf{Var}[X] = \sum_{p \le n} \mathbf{Var}[X_p] + \sum_{p \ne q \le n} \mathbf{Cov}[X_p X_q] \le \mathbf{E}[X] + \sum_{p \ne q \le n} \mathbf{Cov}[X_p X_q],$$

since $\mathbf{Var}[X_p] \le \mathbf{E}[X_p]$. In the above p and q are two distinct primes. Observe that $X_p X_q = 1$ if and only if $p|x$ and $q|x$ which further implies that $pq|x$. In view of this we have

$$\begin{aligned}
\mathbf{Cov}[X_p X_q] &= \mathbf{E}[X_p X_q] - \mathbf{E}[X_p]\mathbf{E}[X_q] = \frac{\lfloor n/(pq) \rfloor}{n} - \frac{\lfloor n/p \rfloor}{n}\frac{\lfloor n/q \rfloor}{n} \\
&\le \frac{1}{pq} - \left(\frac{1}{p} - \frac{1}{n}\right)\left(\frac{1}{q} - \frac{1}{n}\right) \\
&\le \frac{1}{n}\left(\frac{1}{p} + \frac{1}{q}\right).
\end{aligned}$$

Then by (4.9)

$$\sum_{p \ne q} \mathbf{Cov}[X_p X_q] \le \frac{2\pi(n)}{n} \sum_{p \le n} \frac{1}{p} \le \frac{2n(\ln\ln n + O(1))}{n \ln n} = O\left(\frac{\ln\ln n}{\ln}\right) \to 0.$$

Finally, by Chebyshev's inequality with $t = \omega(n)\sqrt{\ln\ln n}$, where $\omega(n) \to \infty$ arbitrarily slowly, we arrive at the main result

$$\Pr\{|\nu(x) - \ln\ln n| > \omega(n)\sqrt{\ln\ln n}\} \le \frac{c}{\omega(n)},$$

where c is a constant. This is summarized in the theorem below.

Theorem 4.3 (Hardy and Ramanujan, 1920; Turán 1934) *For $\omega(n) \to \infty$ arbitrarily slowly, the number of $x \in \{1, \ldots, n\}$ such that*

$$|\nu(x) - \ln\ln n| > \omega(n)\sqrt{\ln\ln n}$$

is $O(n/\omega(n)) = o(n)$.

4.2.3 Height in Independent Tries

In the next three subsections, we shall analyze three digital trees, namely: tries, PATRICIA tries and digital search trees (in short: DST), and their generalizations such as b-tries. As we discussed in Chapter 1.1, these digital trees are built over a set $\mathcal{X} = \{X^1, \ldots, X^n\}$ of n independent strings of (possibly) infinite length. Each string X^i is generated by either a memoryless source of a Markovian source. We concentrate on estimating the typical length of the height in digital trees. We recall that the height is the longest path in a tree.

Here, we consider a trie built over n strings generated independently by a probabilistic source. The reader may recall that in Theorem 1.3 we express the height H_n in terms of the alignments C_{ij} as

$$H_n = \max_{1 \le i < j \le n} \{C_{ij}\} + 1, \tag{4.10}$$

where C_{ij} is defined as the length of the longest common prefix of strings X^i and X^j. We shall use this relation to assess the typical probabilistic behavior of the height H_n for memoryless and Markovian sources. Throughout this section, we apply the first moment method to bound the height from above, and the Chung-Erdős second moment method to establish a lower bound.

The results of the next three subsections are mostly adopted from Devroye [88, 93, 95], Pittel [326] and Szpankowski [399, 400, 401]. However, the reader is advised to study also a series of papers by Arratia and Waterman [20, 21, 22, 24] on similar topics.

Memoryless Source.

Let us first assume that the strings from $\mathcal{X}$ are generated by a memoryless source. To simplify the analysis, we only consider a binary alphabet $\mathcal{A} = \{0, 1\}$ and assume that a memoryless source outputs an independent sequence of 0's and 1's with probabilities p and $q = 1 - p$, respectively. Observe that $P_2 = p^2 + q^2$ is the probability that two independently generated strings agree on a given position (i.e., either both symbols are 0 or both are 1). We also write $Q_2 = 1/P_2$ to simplify our exposition. Clearly, the alignment C_{ij} has a geometric distribution, that is, for all $i, j \in \{1, \ldots, n\}$ and $k \ge 0$

$$\Pr\{C_{ij} = k\} = P_2^k (1 - P_2)$$

and hence $\Pr\{C_{ij} \ge k\} = P_2^k$.

We now derive a typical behavior of H_n. We start with an upper bound. By the first moment method (or equivalently the Boole inequality), for any nonnegative integer k

$$\Pr\{H_n > k\} = \Pr\{\max_{1 \le i < j \le n} \{C_{ij}\} \ge k\} \le n^2 \Pr\{C_{ij} \ge k\},$$

since the number of pairs (i,j) is bounded by n^2. Set now $k = \lfloor 2(1+\varepsilon)\log_{Q_2} n \rfloor$ for any $\varepsilon > 0$. Then

$$\Pr\{H_n > 2(1+\varepsilon)\log_{Q_2} n\} \le \frac{n^2}{n^{2(1+\varepsilon)}} = \frac{1}{n^{2\varepsilon}} \to 0.$$

Thus $H_n/(2\log_{Q_2} n) \le 1$ (pr.).

We now match the above upper bound and prove that $H_n/(2\log_{Q_2} n) = 1$ (pr.) by showing that $\Pr\{H_n > 2(1-\varepsilon)\log_{Q_2} n\} \to 1$ for any $\varepsilon > 0$. We set $k = \lfloor 2(1-\varepsilon)\log_{Q_2} n \rfloor$. We use the Chung-Erdős formulation of the second moment method with $A_{ij} = \{C_{ij} \ge k\}$. Observe that the sums in (4.4) are over pairs (i,j). Let us define

$$\begin{aligned} S_1 &= \sum_{1 \le i < j \le n} \Pr\{A_{ij}\}, \\ S_2 &= \sum_{(i,j) \ne (l,m)} \Pr\{A_{ij} \cap A_{lm}\}, \end{aligned}$$

where the summation in S_2 is over all pairs $1 \le i,j \le n$, $1 \le l,m \le n$ such that $(i,j) \ne (l,m)$. Then

$$\Pr\{H_n \ge k\} \ge \frac{S_1^2}{S_1 + S_2}.$$

The following is obvious for $n \ge 2$

$$S_1 = \sum_{1 \le i < j \le n} \Pr\{A_{ij}\} = \frac{1}{2} n(n-1) P_2^k \ge \frac{n^2}{4} P_2^k.$$

The sum S_2 a little harder to deal with. We must consider two cases: (i) all indices i,j,l,m are different, (ii) either $i = l$ (i.e., we have (i,j) and (i,m)) or $j = m$ (so we have (i,j) and (l,j)). Let us split $S_2 = S_2' + S_2''$ such that S_2' is the summation over all different indices (as in (i) above), and S_2'' covers the latter case. Notice that C_{ij} and C_{lm} are independent when the indices are different, hence

$$S_2' \le \frac{n^4}{4} P_2^{2k}.$$

To evaluate S_2'' we must compute the probability $\Pr\{C_{ij} \ge k, C_{im} \ge k\}$. But, as in Chapter 3.3.1, we easily see that

$$\Pr\{C_{ij} \ge k, C_{i,m} \ge k\} = (p^3 + q^3)^k = P_3^k,$$

since the probability of having the same symbol at a given position for three strings X_i, X_j and X_m is equal to $P_3 = p^3 + q^3$. But, there are no more than $n^3/6$ pairs with one common index, thus

$$S_2'' \le n^3 (p^3 + q^3)^k.$$

In summary,

$$S_2 = \sum_{(i,j),(l,m)} \mathsf{Pr}\{C_{ij} \geq k,\ C_{lm} \geq k\} \leq \frac{n^4}{4} P_2^{2k} + n^3 (p^3 + q^3)^k \ .$$

To proceed further, we need to bound the sum S_2''. To accomplish this, we need a useful inequality that we discuss next.

Lemma 4.4 *For all $s \geq t > 0$ the following holds*

$$(p^s + q^s)^{1/s} \leq (p^t + q^t)^{1/t}, \tag{4.11}$$

where $0 < q = 1 - p < 1$.

Proof. Let $f(x) = (p^x + q^x)^{1/x}$ for $x > 0$, and then

$$f'(x) = \frac{f(x)}{x} \left(\frac{p^x \ln p + q^x \ln q}{p^x + q^x} - \frac{\ln f(x)}{x} \right) .$$

For $p \geq q$ define $a = q/p$. For $x \geq 1$ we proceed as follows:

$$\begin{aligned} \frac{p^x \ln p + q^x \ln q}{p^x + q^x} - \frac{\ln f(x)}{x} &\leq \frac{1}{1 + a^x} \ln p + \frac{1}{1 + a^{-x}} \ln q - \frac{1}{x} \ln p \\ &\leq \frac{1}{1 + a^{-x}} (\ln q - \ln p) + \left(1 - \frac{1}{x}\right) \ln p \\ &\leq 0. \end{aligned}$$

Thus $f'(x) < 0$ for $x \geq 1$ and $f(x)$ is a decreasing function. This can be easily extended to $0 \leq x < 1$ so the lemma is proved for all $x \geq 0$. In Figure 4.1 we plot the function $f(s) = (p^s + q^s)^{1/s}$ for $s = 2$ and $s = 3$. ■

Now, the rest is easy: An application of the Chung–Erdős formula (4.4) leads to

$$\begin{aligned} \mathsf{Pr}\{H_n > k\} &\geq \frac{1}{1/S_1 + S_2'/S_1^2 + S_2''/S_1^2} \\ &\geq \frac{1}{4n^{-2} P_2^{-k} + 1 + 16(p^3 + q^3)^k / (n P_2^{2k})} \\ &\geq \frac{1}{1 + 4n^{-2\varepsilon} + 16 n^{-1} P_2^{-k/2}} \geq \frac{1}{1 + 4n^{-2\varepsilon} + 16 n^{-\varepsilon}} = 1 - O(1/n^{\varepsilon}) \to 1, \end{aligned}$$

where the third inequality follows after setting $k = 2(1 - \varepsilon) \log_{Q_2} n$, and the last is a consequence of (4.11). Thus, we have shown that $H_n / (2 \log_{Q_2} n) \geq 1$ (pr.), which completes our proof of

$$\mathsf{Pr}\{2(1 - \varepsilon) \log_{Q_2} n \leq H_n \leq 2(1 + \varepsilon) \log_{Q_2} n\} = 1 - O(n^{-\varepsilon}) \tag{4.12}$$

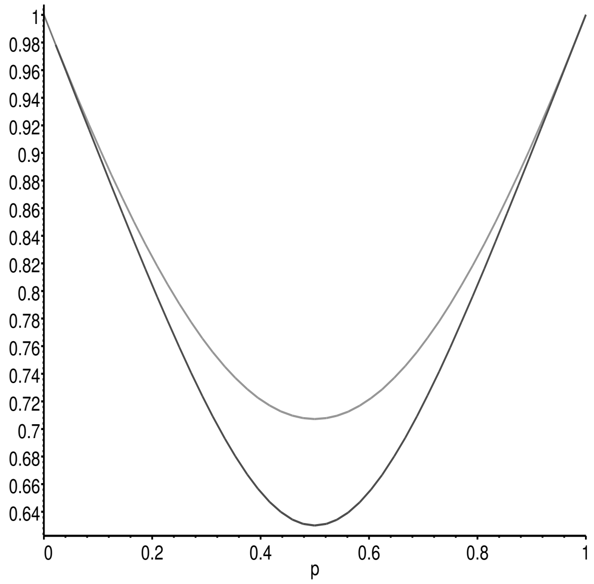

Figure 4.1: Function $(p^3+q^3)^{\frac{1}{3}}$ (lower curve) plotted against $(p^2+q^2)^{\frac{1}{2}}$.

for any $\varepsilon > 0$, where $Q_2^{-1} = P_2 = p^2+q^2$. In passing we should observe that the above implies the convergence in probability of $H_n/\log_{Q_2} n \to 1$, but the rate of convergence is too slow to use the Borel-Cantelli lemma and to conclude almost sure convergence of $H_n/\log_{Q_2} n$. Nevertheless, at the end of this section we provide an argument (after Kingman [231] and Pittel [326]) that allows us to justify such a convergence (cf. Theorem 4.7 for a complete statement of the result).

Memoryless Source for b-Tries.

We now generalize the above to **b-tries** in which every (external) node is capable to store up to b strings. As in Theorem 1.3 we express the height H_n in terms of the alignments $C_{i_1,\ldots,i_{b+1}}$ as

$$H_n = \max_{1\le i_1<\cdots<i_{b+1}\le n}\{C_{i_1,\ldots,i_{b+1}}\}+1,$$

where $C_{i_1,\ldots,i_{b+1}}$ is the length of the longest common prefix of strings $X^{i_1},\ldots,X^{i_{b+1}}$. We follow the same route as for $b=1$, so we only outline the proof. First of all, observe that

$$\Pr\{C_{i_1,\ldots,i_{b+1}} \ge k\} = P_{b+1}^k,$$

where $P_{b+1} = p^{b+1}+q^{b+1}$ represents the probability of a match in a *given* position of $b+1$ strings $X^{i_1},\ldots,X^{i_{b+1}}$. As before, we write $Q_{b+1} = 1/P_{b+1}$. Our goal is to prove that

$H_n \sim (b+1)\log_{Q_{b+1}} n$ (pr.). To simplify the analysis, we let $\mathcal{D}_{\mathbf{i}} = \{\mathbf{i} = (i_1, \ldots, i_{b+1}) : i_k \neq i_l$ whenever $k \neq l\}$ and $A_{\mathbf{i}} = \{C_{i_1,\ldots,i_{b+1}} \geq k\}$.

The upper bound is an easy application of the first moment method. We have for $k = \lfloor (1+\varepsilon)(b+1)\log_{Q_{b+1}} n \rfloor$

$$
\begin{aligned}
\Pr\{H_n > k\} &= \Pr\{\max_{\mathcal{D}_{\mathbf{i}}}\{C_{i_1,\ldots,i_{b+1}}\} \geq k\} \\
&\leq n^{b+1}\Pr\{C_{i_1,\ldots,i_{b+1}} \geq (1+\varepsilon)(b+1)\log_{Q_{b+1}} n\} \\
&= \frac{1}{n^{(b+1)\varepsilon}} \to 0,
\end{aligned}
$$

which is the desired result.

The lower bound is slightly more intricate due to complicated indexing. But, in principle it is nothing else than the second moment method as in the case $b = 1$. Let us denote by S_1 and S_2 the two sums appearing in the Chung–Erdős formula. The first sum can be estimated as follows:

$$
S_1 = \sum_{\mathcal{D}_{\mathbf{i}}} \Pr\{A_{\mathbf{i}}\} = \binom{n}{b+1} P_{b+1}^k \geq \frac{(n-b)^{b+1}}{(b+1)!} P_{b+1}^k.
$$

The second sum S_2 we upper bound by

$$
S_2 = \sum_{\mathcal{D}_{\mathbf{i}} \neq \mathcal{D}_{\mathbf{j}}} \Pr\{A_{\mathbf{i}} \cap A_{\mathbf{j}}\} \leq \frac{1}{[(b+1)!]^2} n^{2(b+1)} P_{b+1}^{2k} + \sum_{i=1}^{b} \binom{b+1}{i} n^{2(b+1)-i} P_{2(b+1)-i}^k.
$$

Using the Chung-Erdős formula we obtain

$$
\begin{aligned}
\Pr\{H_n > k\} &\geq \frac{1}{1/S_1 + S_2/S_1^2} \\
&\geq \frac{1}{\frac{(b+1)!}{(n-b)^{b+1} P_{b+1}^k} + \left(1 + \frac{b}{n-b}\right)^{2(b+1)} + \sum_{i=1}^{b} \binom{b+1}{i} [(b+1)!]^2 \frac{n^{2(b+1)-i}}{(n-b)^{2(b+1)}} \frac{P_{2(b+1)-i}^k}{P_{b+1}^{2k}}}.
\end{aligned}
$$

Setting now $k = \lfloor (1-\varepsilon)(b+1)\log_{Q_{b+1}} n \rfloor$ and using inequality (4.11) we finally arrive at

$$
\begin{aligned}
\Pr\{H_n > (1-\varepsilon)(b+1)\log_{Q_{b+1}} n\} &\geq \frac{1}{c_1 n^{(b+1)\varepsilon} + 1 + \sum_{i=1}^{b} c_2(b) n^{i\varepsilon}} \\
&\geq \frac{1}{1 + c(b) n^{\varepsilon}} = 1 - O(n^{-\varepsilon}),
\end{aligned}
$$

where $c_1, c_2(b)$ and $c(b)$ are constants. In summary, as in the case $b = 1$, we conclude that $H_n \sim (b+1)\log_{Q_{b+1}} n$ (pr.), or more precisely for any $\varepsilon > 0$,

$$
\Pr\{(1-\varepsilon)(b+1)\log_{Q_{b+1}} n \leq H_n \leq (1+\varepsilon)(b+1)\log_{Q_{b+1}} n\} = 1 - O\left(\frac{1}{n^{\varepsilon}}\right), \tag{4.13}
$$

where $Q_{b+1}^{-1} = P_{b+1} = p^{b+1} + q^{b+1}$ (cf. Theorem 4.7 for a precise formulation of this result).

Markovian Source.

Our last generalization deals with the Markovian model. We now assume that all strings $X^1, \ldots, X^n$ are generated independently by a *stationary Markovian source* over a finite alphabet $\mathcal{A}$ of size V. More precisely, the underlying Markov chain is stationary with the transition matrix P and the stationary distribution $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi} = \boldsymbol{\pi}\mathsf{P}$. For simplicity, we also assume $b = 1$. We shall follow the footsteps of the previous analysis once we find formulas on $\mathsf{Pr}\{C_{ij} \geq k\}$ and $\mathsf{Pr}\{C_{ij} \geq k, C_{lm} \geq k\}$. In principle, it is not that difficult. We need, however, to adopt a new approach which Arratia and Waterman [21] called the "analysis by pattern", Jacquet and Szpankowski [207] named it the "string-ruler approach", and it was already used in Pittel [326]. The idea is to choose a *given* pattern $w \in \mathcal{A}^k$ of length k, and measure relationships between strings $X^1, \ldots, X^n$ by comparing them to w. In particular, it is easy to see that

$$C_{ij} \geq k \quad \Rightarrow \quad \exists_{w \in \mathcal{A}^k} \;\; X_1^i X_2^i \ldots X_k^i = X_1^j X_2^j \ldots X_k^j = w,$$

that is, if $C_{ij} \geq k$, then there exists a word w such that prefixes of length k of X^i and X^j are the same and equal to w. Thus

$$\mathsf{Pr}\{C_{ij} \geq k\} = \sum_{w \in A^k} [\mathsf{Pr}\{w\}]^2, \tag{4.14}$$

$$\mathsf{Pr}\{C_{ij} \geq k, C_{lm} \geq k\} = \sum_{w \in A^k} \sum_{u \in \mathcal{A}^k} [\mathsf{Pr}\{w\}]^2 [\mathsf{Pr}\{u\}]^2 \qquad i \neq l, \; j \neq m, \tag{4.15}$$

$$\mathsf{Pr}\{C_{ij} \geq k, C_{im} \geq k\} = \sum_{w \in A^k} [\mathsf{Pr}\{w\}]^3 \qquad i = l, \; j \neq m. \tag{4.16}$$

To complete our analysis, we must compute the probabilities $\mathsf{Pr}\{C_{ij} \geq k\}$ and $\mathsf{Pr}\{C_{ij} \geq k, C_{lm} \geq k\}$ as $k \to \infty$. As shown in (4.14)–(4.16) these probabilities depend on $\mathsf{Pr}\{w\}$ where $w \in \mathcal{A}^k$ is a word of length k. Let $w = w_{j_1} w_{j_2} \ldots w_{j_k}$. Then, in the Markovian model

$$\mathsf{Pr}\{w\} = \pi_{j_1} p_{j_1, j_2} p_{j_2, j_3} \cdots p_{j_{k-1}, j_k}.$$

Thus, for any r we have

$$\sum_{w \in \mathcal{A}^k} [\mathsf{Pr}\{w\}]^r = \sum_{1 \leq j_1, j_2, \ldots, j_k \leq V} \left(\pi_{j_1} p_{j_1, j_2} \cdots p_{j_{k-1}, j_k}\right)^r.$$

We can rewrite succinctly the above using the Schur's product of matrices. Let $\mathsf{P} \circ \mathsf{P} \circ \cdots \circ \mathsf{P} := \mathsf{P}_{[r]}$ be the r-th Schur product (i.e., element-wise product) of P. We also write $\boldsymbol{\psi} = (1, 1, \ldots, 1)$ for the unit vector, and $\boldsymbol{\pi}_{[r]} = (\pi_1^r, \ldots, \pi_V^r)$ for the the r-th power of the stationary vector. The above formula becomes in terms of matrices just introduced

$$\sum_{w \in \mathcal{A}^k} [\mathsf{Pr}\{w\}]^r = \langle \boldsymbol{\pi}_{[r]}, \mathsf{P}_{[r]}^{k-1} \boldsymbol{\psi} \rangle,$$

Table 4.1: SPECTRAL PROPERTIES OF NON-NEGATIVE MATRICES

Let $\mathbf{r}$ and $\mathbf{l}$ be, respectively, the right eigenvector and the left eigenvector of a matrix A associated with the eigenvalue λ, that is,

$$\mathbf{l}\mathsf{A} = \lambda\mathbf{l} \quad , \quad \mathsf{A}\mathbf{r} = \lambda\mathbf{r}.$$

To avoid heavy notation, we do not specify whether vectors are column or row vectors since this should be clear from the context. Consider now a nonnegative matrix A (all elements of A are nonnegative). We also assume it is irreducible (cf. [191, 316]; the reader may think of A as a transition matrix of an irreducible Markov chain). Let $\lambda_1, \ldots, \lambda_m$ are eigenvalues of A associated with the eigenvectors $\mathbf{r}_1, \ldots, \mathbf{r}_m$. We assume that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_m|$.

Theorem 4.5 (Perron–Frobenius) *Let A be a $V \times V$ irreducible nonnegative matrix. Then*

- A *has a positive real eigenvalue λ_1 of the largest value, that is $\lambda_1 \geq |\lambda_{i\neq 1}|$.*
- *The eigenvector $\mathbf{r}_1$ associated with λ_1 has all positive coordinates.*
- *λ_1 is of multiplicity one.*
- *$\lambda_1 > |\lambda_{i\neq 1}|$ if there exists k such that all entries of A^k are strictly positive or the main-diagonal entries are strictly positive.*

Assume first that all eigenvalues are of multiplicity one. Then the left eigenvectors $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_V$ are orthogonal with respect to the right eigenvectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_V$, that is $\langle \mathbf{l}_i, \mathbf{r}_j \rangle = 0$ for $i \neq j$ where $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the inner (scalar) product of $\mathbf{x}$ and $\mathbf{y}$. (Indeed, $\langle \mathbf{l}_i, \mathbf{r}_j \rangle = \frac{\lambda_j}{\lambda_i} \langle \mathbf{l}_i, \mathbf{r}_j \rangle = 0$ since $\lambda_i \neq \lambda_j$.) Setting $\langle \mathbf{l}_i, \mathbf{r}_i \rangle = 1$ for all $1 \leq i \leq V$ we can write for any vector $\mathbf{x} = \langle \mathbf{l}_1, \mathbf{x} \rangle \mathbf{r}_1 + \sum_{i=2}^{V} \langle \mathbf{l}_i, \mathbf{x} \rangle \mathbf{r}_i$ which yields

$$\mathsf{A}\mathbf{x} = \langle \mathbf{l}_1, \mathbf{x} \rangle \lambda_1 \mathbf{r}_1 + \sum_{i=2}^{V} \langle \mathbf{l}_i, \mathbf{x} \rangle \lambda_i \mathbf{r}_i.$$

Since A^k has eigenvalues $\lambda_1^k, \lambda_2^k, \ldots, \lambda_V^k$, then — dropping the assumption about eigenvalues $\lambda_2, \ldots, \lambda_V$ being simple — we arrive at

$$\mathsf{A}^k\mathbf{x} = \langle \mathbf{l}_1, \mathbf{x} \rangle \mathbf{r}_1 \lambda^k + \sum_{i=2}^{V} q_i(k) \langle \mathbf{l}_i, \mathbf{x} \rangle \mathbf{r}_i \lambda_i^k$$

where $q_i(k)$ is a polynomial in k ($q_i(k) \equiv 1$ when the eigenvalues $\mu_2, \ldots, \mu_V$ are simple). In particular, for irreducible non-negative matrices by the above and the Perron-Frobenius theorem $\mathsf{A}^k\mathbf{x} = \beta\lambda^k(1 + O(\rho^k))$ for some β and $\rho < 1$.

where $\langle \mathbf{x}, \mathbf{y} \rangle$ is the scalar product of vectors $\mathbf{x}$ and $\mathbf{y}$. Assuming that the underlying Markov chain is aperiodic and irreducible, by the Perron-Frobenius Theorem 4.5 (cf. Table 4.1) the above can be represented as for some $\rho < 1$

$$\langle \boldsymbol{\pi}_{[r]}, \mathsf{A}_{[r]}^{k-1} \boldsymbol{\psi} \rangle = \lambda_{[r]}^{k-1} \langle \boldsymbol{\pi}_{[r],\mathbf{r}} \rangle \langle \mathbf{l}, \boldsymbol{\psi} \rangle (1 + O(\rho^k)),$$

where $\mathbf{l}$ and $\mathbf{r}$ are the left and the right principal eigenvectors of $\mathsf{P}_{[r]}$. In summary, there exists a constant β such that

$$\sum_{w \in \mathcal{A}^k} [\mathsf{Pr}\{w\}]^r = \beta \lambda_{[r]}^k (1 + O(\rho^k)), \tag{4.17}$$

hence

$$\begin{aligned} \mathsf{Pr}\{C_{ij} \geq k\} &= \beta_1 \lambda_{[2]}^k (1 + O(\rho^k)), && && (4.18) \\ \mathsf{Pr}\{C_{ij} \geq k, C_{lm} \geq k\} &= \beta_1^2 \lambda_{[2]}^{2k} (1 + O(\rho^k)) && i \neq l,\ j \neq m, && (4.19) \\ \mathsf{Pr}\{C_{ij} \geq k, C_{im} \geq k\} &= \beta_2 \lambda_{[3]}^k (1 + O(\rho^k)) && i = l,\ j \neq m && (4.20) \end{aligned}$$

for some constants β_1 and β_2.

We need the following inequality on $\lambda_{[r]}^{1/r}$ that extends Lemma 4.4 to Markov models and is interesting on its own right.

Lemma 4.6 (Karlin and Ost, 1985) *Let $\mathsf{P}_{[r]}$ the rth order Schur product of a transition matrix P of an aperiodic irreducible Markov chain. The function*

$$F(r) = \lambda_{[r]}^{1/r}$$

is decreasing for $r > 0$, where $\lambda_{[r]}$ is the largest eigenvalue of $\mathsf{P}_{[r]}$.

Proof. We follow the arguments of Karlin and Ost [219]. From (4.17) we observe that

$$\lim_{k \to \infty} \left(\sum_{w \in \mathcal{A}^k} [\mathsf{Pr}\{w\}]^r \right)^{\frac{1}{k}} = \lambda_{[r]} \leq 1.$$

But (either by a probabilistic argument or an algebraic one)

$$\sum_{w \in \mathcal{A}^k} [\mathsf{Pr}\{w\}]^{r+s} \leq \sum_{w \in \mathcal{A}^k} [\mathsf{Pr}\{w\}]^r \sum_{w \in \mathcal{A}^k} [\mathsf{Pr}\{w\}]^s,$$

hence together with the above we conclude that

$$\lambda_{[r+s]} \leq \lambda_{[r]} \lambda_{[s]}. \tag{4.21}$$

Furthermore, it is easy to see that $\log \lambda_{[r]}$ is convex as a function of r. Let now $f(s) = \log \lambda_{[s]}$ and set for $r < s < 2r$

$$s = \frac{2r-s}{r} r + \frac{s-r}{r} 2r.$$

By convexity

$$f(s) \le \frac{2r-s}{r} f(r) + \frac{s-r}{r} f(2r).$$

But (4.21) implies $f(2r) \le 2f(r)$, hence the above yields

$$f(s) \le \frac{s}{r} f(r)$$

and therefore $f(r)/r$ is decreasing. This proves the lemma. ■

The rest is easy, and we imitate our derivation from the above. The reader is asked in Exercise 2 to complete the proof of the following for any $\varepsilon > 0$

$$\Pr\{2(1-\varepsilon)\log_{Q_2} n \le H_n \le 2(1+\varepsilon)\log_{Q_2} n\} = 1 - O(n^{-\varepsilon}), \tag{4.22}$$

where $Q_2^{-1} = \lambda_{[2]}$ is the largest eigenvalue of $\mathsf{P} \circ \mathsf{P} = \mathsf{P}_{[2]}$. Thus $H_n \sim 2\log_{Q_2} n$ (pr.).

Almost Sure Convergence.

For simplicity of derivations, we again assume here $b = 1$. In (4.12), (4.13) and (4.22) we proved that **whp** the height H_n of a trie is asymptotically equal to $2\log_{Q_2} n$ with the rate of convergence $O(n^{-\varepsilon})$. This rate does not yet justify an application of the Borel-Cantelli Lemma in order to improve the result to almost sure convergence. Nevertheless, we shall show in this section that $H_n/(2\log_{Q_2} n) \to 1$ (a.s.) thanks to the fact that H_n is a nondecreasing sequence. We apply here a method suggested by H. Kesten and reported in Kingman [231].

First of all, observe that for any n

$$H_n \le H_{n+1},$$

that is, H_n – though random – is a nondecreasing sequence. Furthermore, the rate of convergence $O(n^{-\varepsilon})$ and the Borel-Cantelli Lemma justify almost sure convergence of $H_n/(2\log_{Q_2} n)$ along the exponential skeleton $n = s2^r$ for some integers s and r. Indeed, we have

$$\sum_{r=0}^{\infty} \Pr\left\{ \left| \frac{H_{s2^r}}{2\log_{Q_2}(s2^r)} - 1 \right| \ge \varepsilon \right\} < \infty.$$

We must extend the above to *every* n. Fix s. For every n we find such r that

$$s2^r \le n \le (s+1)2^r.$$

Since logarithm is a slowly varying function and H_n is a nondecreasing sequence, we have

$$\limsup_{n\to\infty} \frac{H_n}{2\log_{Q_2} n} \le \limsup_{r\to\infty} \frac{H_{(s+1)2^r}}{2\log_{Q_2}(s+1)2^r} \frac{2\log_{Q_2}(s+1)2^r}{2\log_{Q_2} s2^r} = 1 \qquad \text{(a.s.)}.$$

In a similar manner, we prove

$$\liminf_{n\to\infty} \frac{H_n}{2\log_{Q_2} n} \ge \liminf_{r\to\infty} \frac{H_{s2^r}}{2\log_{Q_2}(s+1)2^r} = 1 \quad \text{(a.s.)}.$$

We summarize the main findings of this section in the following theorem.

Theorem 4.7 (Pittel, 1985; Szpankowski, 1991) *Consider a b-trie built over n independent strings generated according to a stationary Markov chain with the transition matrix* P. *Then*

$$\lim_{n\to\infty} \frac{H_n}{\ln n} = \frac{(b+1)}{\ln \lambda_{[b+1]}^{-1}} \qquad \text{(a.s.)},$$

where $\lambda_{[b+1]}$ is the largest eigenvalue of the $(b+1)$-st order Schur product of P. *In particular, in the Bernoulli model $\lambda_{[b+1]} = \sum_{i=1}^{V} p_i^{b+1}$ where p_i is the probability of generating the ith symbol from the alphabet $\mathcal{A} = \{1, \ldots, V\}$.*

4.2.4 Height in Independent PATRICIA Tries

In Chapter 1.1 we described how to obtain PATRICIA trie from a regular trie: In PATRICIA we compress a path from the root to a terminal node by avoiding (compressing) unary nodes (cf. Figure 1.1). In this section, we look at the height H_n of PATRICIA and derive its typical probabilistic behavior. To simplify our analysis, we again assume a binary alphabet $\mathcal{A} = \{0, 1\}$ with $p \ge q = 1-p$. Actually, we shall write $p_{\max} = \max\{p, q\}$ and $Q_{\max} = 1/p_{\max}$. We again assume the memoryless model.

We start with an upper bound. Following Pittel [326] we argue that for fixed k and b the event $H_n \ge k + b - 1$ implies that there exist b strings, say $X^{i_1}, \ldots, X^{i_b}$ such that their common prefix is of length at least k. In other words,

$$H_n \ge k + b - 1 \qquad \Rightarrow \qquad \exists_{i_1,\ldots,i_b} \;\; C_{i_1,i_2,\ldots,i_b} \ge k.$$

This is true since in PATRICIA there is no unary nodes. Hence, if there is a path of length at least $k + b - 1$, then there must be b strings sharing the common prefix of length k. But, as in the analysis of b tries, we know that

$$\mathsf{Pr}\{C_{i_1,i_2,\ldots,i_b} \ge k\} = P_b^k,$$

where $P_b = p^b + q^b$. Thus, for fixed b and $k(b) = b(1+\varepsilon)\log_{P_b^{-1}} n$

$$\mathsf{Pr}\{H_n \ge k + b - 1\} \le n^b \mathsf{Pr}\{C_{i_1,i_2,\ldots,i_b} \ge k\} = O(n^{-\varepsilon}).$$

The above is true for *all* values of b, hence we allow now b to increase arbitrary slowly to infinity. Observe then

$$\begin{aligned}\lim_{b\to\infty} k(b) &= (1+\varepsilon)\lim_{b\to\infty}\frac{b\log n}{\log P_b^{-1}} = (1+\varepsilon)\lim_{b\to\infty}\frac{\log n}{\log P_b^{-1/b}} \\ &\le (1+\varepsilon)\frac{\log n}{\log p_{\max}^{-1}} = (1+\varepsilon)\log_{Q_{\max}} n,\end{aligned}$$

since $\lim_{b\to\infty} P_b^{1/b} \le \lim_{b\to\infty} p_{\max}^{\frac{b-1}{b}} = p_{\max}$. In summary, for any $\varepsilon > 0$ we just proved that

$$\Pr\{H_n \ge (1+\varepsilon)\log_{Q_{\max}} n\} = O(n^{-\varepsilon}),$$

which is the desired upper bound.

Not surprisingly, we shall use the second moment method to prove a lower bound. Let "1" occurs with the probability $p_{\max}$. We show that with high probability the following k strings: $X^{i_1} = 10\ldots$, $X^{i_2} = 110\ldots,\cdots,X^{i_k} = 111\cdots 10\ldots$ appear among all n strings $X^1,\ldots X^n$, where string X^{i_j} has the first j symbols equal to 1 following by a 0. Observe that if this is true, then $H_n \ge k+1$ **whp**, and we select $k = (1-\varepsilon)\log_{Q_{\max}} n$ to complete the proof.

Let now $\mathbf{i} = (i_1, i_2, \ldots, i_k)$, and define $Z_{\mathbf{i}} = 1$ if $X^{i_1} = 10\ldots$, $X^{i_2} = 110\ldots,\cdots,X^{i_k} = 111\cdots 10\ldots$, and zero otherwise. Let also

$$Z = \sum_{\mathbf{i}\in\mathcal{D}_{\mathbf{i}}} Z_{\mathbf{i}},$$

where $\mathcal{D}_{\mathbf{i}} = \{\mathbf{i} = (i_1,\ldots,i_k) :\ i_j \ne i_l \text{ whenever } j \ne l\}$. Observe that

$$\Pr\{Z > 0\} \le \Pr\{H_n > k\}.$$

We estimate the left-hand side of the above, by the second moment method. First of all

$$\mathbf{E}[Z] = \binom{n}{k} p_{\max}^{\frac{k(k+1)}{2}}(1-p_{\max})^k,$$

since the probability of X^i is $p_{\max}^i(1-p_{\max})$, and the strings are independent. In a similar manner we compute the variance

$$\mathbf{Var}[Z] \le \mathbf{E}[Z] + \sum_{\mathcal{D}_{\mathbf{i}}\ne\mathcal{D}_{\mathbf{j}}} \mathbf{Cov}[Z_{\mathbf{i}}Z_{\mathbf{j}}].$$

The covariance $\mathbf{Cov}[Z_{\mathbf{i}}Z_{\mathbf{j}}]$ depends on how many indices are the same in $\mathbf{i}$ and $\mathbf{j}$. If, say $0 < l < k$, indices are the same, then

$$\mathbf{Cov}[Z_{\mathbf{i}}Z_{\mathbf{j}}] \le \mathbf{E}[Z_{\mathbf{i}}Z_{\mathbf{j}}] \le \binom{n}{2k-l} p_{\max}^{k(k+1)-l(l+1)/2}(1-p_{\max})^{2k-l}.$$

Thus, by the second moment method

$$\begin{aligned}
\Pr\{Z=0\} &\le \frac{\mathbf{Var}[Z]}{\mathbf{E}[Z]^2} \\
&\le \frac{1}{\mathbf{E}[Z]} + \frac{1}{\mathbf{E}[Z]^2}\sum_{l=1}^{k-1}\binom{n}{2k-l} p_{\max}^{k(k+1)-l(l+1)/2}(1-p_{\max})^{2k-l}. \qquad (4.23)
\end{aligned}$$

Let us now estimate the terms of the above. First of all, for $k=(1-\varepsilon)\ln n/\ln p_{\max}^{-1}$ we obtain for some constant c

$$\begin{aligned}
\mathbf{E}[Z] &\ge c\frac{n^k}{k!}p_{\max}^{k^2/2}(1-p_{\max})^k \\
&\ge \exp\left((1-\varepsilon)\frac{\ln^2 n}{\ln p_{\max}^{-1}} - \frac{(1-\varepsilon)^2}{2}\frac{\ln^2 n}{\ln p_{\max}^{-1}} - O(\ln n \ln\ln n)\right) \\
&= \exp\left(\frac{1}{2}(1-\varepsilon)\frac{\ln^2 n}{\ln p_{\max}^{-1}}\right) \to \infty,
\end{aligned}$$

where ε is an arbitrary small positive number. The lth term of (4.23) can be estimated as follows

$$\begin{aligned}
&\frac{\binom{n}{2k-l}p_{\max}^{k(k+1)-l(l+1)/2}(1-p_{\max})^{2k-l}}{\mathbf{E}[Z]^2} = \frac{\binom{n}{2k-l}p_{\max}^{-l(l+1)/2}(1-p_{\max})^{-l}}{\binom{n}{l}^2} \\
&= O\left(n^{-l}p_{\max}^{-l^2/2}(1-p_{\max})^{-l}\right) \to 0,
\end{aligned}$$

where the convergence to zero is true for $l \le k = (1-\varepsilon)\ln n/\ln p_{\max}^{-1}$ by the same arguments as above (indeed, the function $f(l) = n^l p_{\max}^{l^2/2}$ is nondecreasing with respect to l for $l \le k = \ln n/\ln p_{\max}^{-1}$). Thus

$$\Pr\{H_n \ge (1-\varepsilon)\ln n/\ln p_{\max}^{-1}\} \ge 1 - O\left((\exp(-\alpha\ln^2 n + \ln\ln n)\right)$$

for some $\alpha > 0$.

In summary, we prove the following result (the almost sure convergence below is established in the same fashion as in the previous section).

Theorem 4.8 (Pittel, 1985) *Consider a PATRICIA trie built over n independent strings generated by a memoryless source (Bernoulli model) with p_i being the probability of generating the ith symbol from the alphabet $\mathcal{A} = \{1,\ldots,V\}$. Then*

$$\lim_{n\to\infty}\frac{H_n}{\log n} = \frac{1}{\log p_{\max}^{-1}} \qquad \text{(a.s.)},$$

where $p_{\max} = \max\{p_1, p_2, \ldots, p_V\}$.

4.2.5 Height in Independent Digital Search Trees

To complete our discussion of digital trees, we shall analyze the height of a digital search tree and prove a result similar to the one presented in Theorem 4.8.

We start with a lower bound, and use an approach similar to the one discussed for PATRICIA tries. Actually, the lower bound from Section 4.2.4 works fine for digital search trees. Nevertheless, we provide another derivation so the reader can again see the second moment method at work. Set $Z_0 = n$, and let Z_i for $i \geq 1$ be the number of strings whose prefix of length i consists of 1's. We recall that 1 occurs with the probability $p_{\max}$. Let also $k = (1-\varepsilon)\ln n/\ln p_{\max}^{-1}$. Observe that

$$Z_k > 0 \quad \Rightarrow \quad H_n > k,$$

where H_n is the height in a digital search tree. We prove that $\Pr\{Z_k > 0\}$ **whp** when $k = (1-\varepsilon)\ln n/\ln p_{\max}^{-1}$ using the second moment method. Thus, we must find $\mathbf{E}[Z_k]$ and $\mathbf{Var}[Z_k]$.

All strings $X^1, \ldots, X^n$ are independent, hence Z_1 has the binomial distribution with parameters n and $p_{\max}$, that is, $Z_1 \sim \text{binomial}(n, p_{\max})$. In general, $Z_i \sim \text{binomial}(Z_{i-1}, p_{\max})$. Using conditional expectation we easily prove that (the reader is asked to provide details in Ex. 8)

$$\begin{aligned} \mathbf{E}[Z_k] &= np_{\max}^k, & (4.24)\\ \mathbf{Var}[Z_k] &= np_{\max}^k(1-p_{\max}^k). & (4.25) \end{aligned}$$

Then, for $Q_{\max} = 1/p_{\max}$

$$\Pr\{H_n \leq (1-\varepsilon)\log_{Q_{\max}} n\} \leq \Pr\{Z_k = 0\} \leq \frac{\mathbf{Var}[Z_k]}{\mathbf{E}[Z_k]^2} \leq \frac{1}{np_{\max}^k} = n^{-\varepsilon}.$$

This proves that $H_n/\ln n \geq 1/\ln p_{\max}^{-1}$ **whp**.

The upper bound is slightly more complicated. Let us consider a *given* word w (of possibly infinite length) whose prefix of length k is denoted as w_k. Define $T_n(w)$ to be the length of a path in a digital search tree that follows the symbols of w until it reaches the last node of the tree on its path. For example, referring to Figure 1.1 we have $T_4(00000\ldots) = 2$. The following relationship is quite obvious:

$$\Pr\{T_n(w) \geq k\} = \Pr\{T_{n-1}(w) \geq k\} + \Pr\{T_{n-1}(w) = k-1\}\Pr\{w_k\},$$

since when inserting a new string to the tree we either do not follow the path of w (the first term above) or we follow the path leading to the second term of the recurrence. Observe now that for given w

$$T_n(w) \geq T_{n-1}(w),$$

so that $T_n(w)$ is a nondecreasing sequence with respect to n. Since, in addition, $\Pr\{w_k\} \leq p_{\max}^k$ and $\Pr\{T_{n-1}(w) = k-1\} \leq \Pr\{T_{n-1}(w) \geq k-1\}$, we can iterate the above recurrence to get

$$\Pr\{T_n(w) \geq k\} \leq \Pr\{T_{n-2} \geq k\} + 2p_{\max}^k \Pr\{T_{n-1}(w) \geq k-1\},$$

so that after iterations with respect to n the above leads to

$$\Pr\{T_n(w) \geq k\} \leq np_{\max}^k \Pr\{T_{n-1}(w) \geq k-1\}.$$

Iterating now with respect to k, yields

$$\Pr\{T_n(w) \geq k\} \leq n^k p_{\max}^{k^2}.$$

Set now $k = (1+\varepsilon)\log n / \log p_{\max}^{-1}$ to obtain

$$\begin{aligned}
\Pr\{H_n > (1+\varepsilon)\log_{p_{\max}} n^{-1}\} &\leq \sum_{w \in \mathcal{A}^k} \Pr\{T_n(w) \geq k\} \\
&\leq 2^k n^k p_{\max}^{k^2} \\
&\leq \exp\left(k \ln 2 + k \ln n - k^2 \ln p_{\max}^{-1}\right) \\
&\leq \exp\left(-k(1+\varepsilon)\ln n + k \ln n + k \ln 2\right) \\
&\leq \exp(-\alpha \ln^2 n) \to 0
\end{aligned}$$

for some $\alpha > 0$ as long as $\varepsilon > \ln 2 / \ln n$.

Theorem 4.9 (Pittel, 1985) *Consider a digital search tree built over n independent strings generated by a memoryless source (Bernoulli model) with p_i being the probability of generating the ith symbol from the alphabet $\mathcal{A} = \{1, \ldots, V\}$. Then*

$$\lim_{n\to\infty} \frac{H_n}{\log n} = \frac{1}{\log p_{\max}^{-1}} \qquad \text{(a.s.)},$$

where $p_{\max} = \max\{p_1, p_2, \ldots, p_V\}$.

4.2.6 Height in a Suffix Tree

Finally, we consider the height of a suffix tree built from the first n suffixes of a (possibly infinite) string X_1^∞ generated by a memoryless source. As before, we assume a binary alphabet with p and $q = 1-p$ being the probability of generating the symbols. We denote by $X(i) = X_i^\infty$ the ith suffix, where $1 \leq i \leq n$. (We recall that a suffix tree is a trie built from dependent strings, namely, $X(1), \ldots, X(n)$.) We restrict the analysis to the memoryless model. The reader is asked in Exercise 7 to extend this analysis to the mixing model.

The analysis of the height follows the same footsteps as in the case of tries (i.e., we use the first moment method to prove an upper bound and the second moment method to prove a lower bound) except that we must find the probability law governing the self-alignment C_{ij} that represents the length of the longest prefix common to $X(i)$ and $X(j)$. We recall that the height H_n is related to C_{ij} by $H_n = \max_{i \neq j}\{C_{ij}\} + 1$.

We need to know the probability of $A_{ij} = \{C_{ij} \geq k\}$. Let $d = |i - j|$ and consider the suffix $X(i)$ and the suffix $X(i + d)$. Below, we assume that $j = i + d$. When $d \geq k$, then $X(i)$ and $X(i + d)$ are independent in the memoryless model, hence as in tries we easily find that

$$\Pr\{C_{i,i+d} \geq k\} = P_2^k \qquad d \geq k,$$

where $P_2 = p^2 + q^2$. The problem arises when $d < k$. We need to identify conditions under which k symbols starting at position i are the same as k symbols following position $i + d$, that is,

$$X_i^{i+k-1} = X_{i+d}^{i+d+k-1},$$

where X_i^{i+k-1} and $X_{i+d}^{i+d+k-1}$ are k-length prefixes of suffixes $X(i)$ and $X(i+d)$, respectively. The following simple combinatorial lemma provides a solution.

Lemma 4.10 (Lothaire, 1982) *Let X_1^∞ be a string whose ith and jth suffixes are $X(i) = X_i^\infty$ and $X(j) = X_j^\infty$, respectively, where $j - i = d > 0$. Let also Z be the longest common prefix of length $k \geq d$ of $X(i)$ and $X(j)$. Then there exists a word $w \in \mathcal{A}^d$ of length d such that*

$$Z = w^{\lfloor k/d \rfloor} \bar{w}, \tag{4.26}$$

where $\bar{w}$ is a prefix of w, and w^l is the string resulting from the concatenation of $l = \lfloor k/d \rfloor$ copies of the word w.

We leave the formal proof of the above lemma to the reader (cf. Exercise 9). Its meaning, however, should be quite clear. To illustrate Lemma 4.10, let is assume that $X_1^{k+d} = X_1^{3d+2} = x_1 x_2 \ldots x_d x_1 x_2 \ldots x_d \ldots x_1 x_2 \ldots x_d x_1 x_2$, where $k = 2d+2$ and $d > 2$. One easily identifies $w = x_1 x_2 \ldots x_d$ and $\bar{w} = x_1 x_2$. The common prefix Z of X_1^{k+d-1} and X_{d+1}^{k+2d} can be represented as $Z_1^{2d+2} = w^2 \bar{w}$, as stated in Lemma 4.10.

In view of the above, we can now derive a formula on the probability $\Pr\{C_{i,i+d} \geq k\}$ which is pivotal to our analysis of the height. An application of Lemma 4.10 leads to

$$\begin{aligned} \Pr\{C_{i,i+d} \geq k\} &= \sum_{w \in \mathcal{A}^{k \wedge d}} \Pr\{w^{\lfloor \frac{k}{d} \rfloor + 1} \bar{w}\} & (4.27) \\ &= \left(p^{\lfloor \frac{k}{d} \rfloor + 1} + q^{\lfloor \frac{k}{d} \rfloor + 1}\right)^{d-r} \left(p^{\lfloor \frac{k}{d} \rfloor + 2} + q^{\lfloor \frac{k}{d} \rfloor + 2}\right)^{r}, & (4.28) \end{aligned}$$

where $r = k - d\lfloor k/d \rfloor$ and $k \wedge d = \min\{d, k\}$ (so word w has length $k \wedge d$). Observe that by restricting the length of w to $k \wedge d$, we cover also the case $k < d$ in which case we have $\Pr\{C_{i,i+d} \geq k\} = (p^2 + q^2)^k$.

To see why (4.27) implies (4.28), say for $d < k$, we observe that $d - r$ symbols of w must be repeated $l = \lfloor k/d \rfloor$ times in w^l, while the last r symbols of w must occur $l + 1$ times. The probability of having the same symbol on a given position in l strings is $p^{l+1} + q^{l+1}$. Thus (4.28) follows.

Before we proceed, we need an inequality on (4.28) which we formulate in the form of a lemma.

Lemma 4.11 *For $d < k$ we have*

$$\Pr\{C_{i,i+d} \ge k\} \le (p^2 + q^2)^{k/2} = P_2^{\frac{k+d}{2}} \tag{4.29}$$

Proof. The result follows from Lemma 4.4 (cf. (4.11)) and (4.28). Indeed, let $P_l = (p^l + q^l)$. Then

$$\Pr\{C_{i,i+d} \ge k\} = P_{l+1}^{d-r} P_{l+2}^{r} \le P_2^{\frac{1}{2}[(d-r)(l+1)+r(l+2)]} = P_2^{\frac{k+d}{2}},$$

since $k = dl + r$. ∎

Now, we are ready to prove an upper bound for the height. We use the Boole inequality and the above to obtain

$$\begin{aligned}
\Pr\{\max_{i,d}\{C_{i,i+d}\} \ge k\} &\le n\left(\sum_{d=1}^{k} \Pr\{C_{i,i+d} \ge k\} + \sum_{d=k+1}^{n} \Pr\{C_{i,i+d} \ge k\}\right) \\
&\le n\left(\sum_{d=1}^{k} P_2^{\frac{k+d}{2}} + \sum_{d=k+1}^{n} P_2^{k}\right) \\
&\le n\left(P_2^{\frac{k+1}{2}}(1 - \sqrt{P_2})^{-1} + nP_2^{k}\right),
\end{aligned}$$

where the second inequality is a consequence of Lemma 4.11. From the above we conclude that

$$\Pr\{H_n > 2(1+\varepsilon)\log_{Q_2} n\} = \Pr\{\max_{i,d}\{C_{i,i+d}\} \ge 2(1+\varepsilon)\log_{Q_2} n\} \le \frac{c}{n^{\varepsilon}},$$

where (as before) $Q_2 = P_2^{-1}$ and c is a constant. Thus **whp** we have $H_n \le 2(1+\varepsilon)\log_{Q_2} n$.

For the lower bound, we use the Chung and Erdős second moment method, as one can guess. We need, however, to overcome some difficulties. Let $\mathcal{D} = \{(i,j) : |i-j| > k\}$. Observe that

$$\Pr\{H_n > k\} = \Pr\{\bigcup_{i \ne j} A_{ij}\} \ge \Pr\{\bigcup_{i,j \in \mathcal{D}} A_{ij}\} \ge \frac{S_1^2}{S_1 + S_2},$$

where

$$\begin{aligned}
S_1 &= \sum_{i,j \in \mathcal{D}} \Pr\{A_{ij}\}, \\
S_2 &= \sum_{(i,j) \ne (l,m) \in \mathcal{D}} \Pr\{A_{ij} \cap A_{lm}\}.
\end{aligned}$$

The first sum is easy to compute from what we have learned so far. Since $|i-j|>k$ in $\mathcal{D}$ we immediately find that

$$S_1 \geq (n^2-(2k+1)n)P_2^k.$$

To compute S_2 we split it into three terms, namely S_{21}, S_{22} and S_{23} such that the summation in S_{2i} $(i=1,2,3)$ is over the set $\mathcal{D}_i$ defined as follows:

$$\begin{aligned}
\mathcal{D}_1 &= \{(i,j),(l,m)\in\mathcal{D}: \ \min\{|l-i|,|l-j|) \geq k \text{ and } \min\{|m-i|,|m-j|) \geq k\} \\
\mathcal{D}_2 &= \{(i,j),(l,m)\in\mathcal{D}: \ \min\{|l-i|,|l-j|) \geq k \text{ and } \min\{|m-i|,|m-j|) < k \\
&\qquad \text{ or } \min\{|l-i|,|l-j|) < k \text{ and } \min\{|m-i|,|m-j|) \geq k\} \\
\mathcal{D}_3 &= \{(i,j),(l,m)\in\mathcal{D}: \ \min\{|l-i|,,|l-j|) < k \text{ and } \min\{|m-i|,|m-j|) < k\}
\end{aligned}$$

Using the above and similar arguments as in the case of tries, we obtain

$$\begin{aligned}
S_{21} &= \sum_{(i,j),(l,m)\in\mathcal{D}_1} \Pr\{A_{ij}\cap A_{lm}\} \leq n^4 P_2^{2k}, \\
S_{22} &= \sum_{(i,j),(l,m)\in\mathcal{D}_2} \Pr\{A_{ij}\cap A_{lm}\} \leq \sum_{(i,j),(i,m)\in\mathcal{D}_2} \Pr\{A_{ij}\cap A_{im}\} \leq 8kn^3P_3^k \leq 8kn^2 P_2^{\frac{3}{2}k} \\
S_{23} &= \sum_{(i,j),(l,m)\in\mathcal{D}_3} \Pr\{A_{ij}\cap A_{lm}\} \leq \sum_{(i,j),(i,j)\in\mathcal{D}_3} \Pr\{A_{ij}\cap A_{ij}\} \leq 16k^2n^2P_2^k,
\end{aligned}$$

where, as always, $P_3=p^3+q^3$. The last inequality for S_{22} follows from Lemma 4.4. The rest is a matter of algebra. Proceeding as in Section 4.2.3 we arrive at

$$\Pr\{H_n > 2(1-\varepsilon)\log_{Q_2} n\} \geq \frac{1}{1+c_1n^{-\varepsilon}\log n + c_2 n^{-2\varepsilon}} = 1 - O\left(\frac{\log n}{n^{\varepsilon}}\right),$$

for some constants c_1, c_2. As before we can extend this convergence in probability to almost sure convergence by considering an exponential skeleton $n=s2^r$. Thus, we just proved the following interesting result.

Theorem 4.12 (Devroye, Rais and Szpankowski, 1992) *Consider a suffix tree built from n suffixes of a string generated by a memoryless source with p_i being the probability of generating the ith symbol from the alphabet $\mathcal{A}=\{1,\ldots,V\}$. Then*

$$\lim_{n\to\infty}\frac{H_n}{\log n} = \frac{1}{\log P_2^{-1}} \qquad \text{(a.s.)},$$

where $P_2=\sum_{i=1}^V p_i^2$.

In passing we point out that results of Sections 4.2.3–4.2.6 can be extended to the mixing probabilistic model (cf. [326, 400, 401]). The reader is asked to try to prove such extensions in Exercises 6–7. Additional details can be also found in Chapter 6.

4.3 Extensions and Exercises

4.1 Prove the following alternative formulation of the second moment method. *Let* $\{A_i\}_{i=1}^n$ *be a set of identically distributed events such that* $n\mathsf{Pr}\{A_i\} \to \infty$. *If*

$$\sum_{i \neq j} \frac{\mathsf{Pr}\{A_i | A_j\}}{n\mathsf{Pr}\{A_i\}} \to 1$$

as $n \to \infty$, *then* $\mathsf{Pr}\{\bigcup_{i=1}^n A_i\} \to 1$.

4.2 Consider the height of a b-trie in a Markovian model and prove

$$\mathsf{Pr}\{(b+1)(1-\varepsilon)\log_{Q_{b+1}} n \leq H_n \leq (b+1)(1+\varepsilon)\log_{Q_{b+1}}\} = 1 - O(n^{-\varepsilon}),$$

where $Q_{b+1}^{-1} = \lambda_{[b+1]}$ is the largest eigenvalue of $\mathsf{P}_{b+1]} = \mathsf{P} \circ \cdots \mathsf{P}$.

4.3 Find another proof for the lower bound in the derivation of the height in PATRICIA tries.

4.4 ⚠ Extend the analysis PATRICIA trie to a Markovian source. What is the equivalence of $p_{\max}$ in the Markovian model? (cf. Chapter 6.)

4.5 ⚠ Extend the analysis of digital search trees to a Markovian source.

4.6 ⚠ (Pittel 1985) Establish typical behaviors of heights in all three digital trees in the mixing model (cf. Chapter 6).

4.7 ⚠ (Szpankowski 1993) Extend the analysis of the height in a suffix tree (cf. Section 4.2.6) to the mixing model.

4.8 Prove (4.24) and (4.25).

4.9 Prove Lemma 4.10.

4.10 ⚠ (Pittel 1985) Consider the fill-up level F_n (cf. Section 1.1) in tries, PATRICIA tries and digital search trees. Prove the following theorem:

Theorem 4.13 (Pittel, 1985) *Consider tries, PATRICIA tries and digital search trees built over* n *independent strings generated by a memoryless source with* p_i *being the probability of generating the* ith *symbol from the alphabet* $\mathcal{A} = \{1, \ldots, V\}$. *Then*

$$\lim_{n\to\infty} \frac{F_n}{\log n} = \frac{1}{\log p_{\min}^{-1}} \qquad \text{(a.s.)},$$

where $p_{\min} = \min\{p_1, p_2, \ldots, p_V\}$.

Can this theorem be extended to suffix trees?

4.11 Consider again digital trees as above, including suffix trees. Using the same notation as in Exercise 10 prove or disprove the following result.

Theorem 4.14 (Pittel, 1985) *Consider digital trees built over n independent strings generated by a memoryless source with p_i being the probability of generating the ith symbol from the alphabet $\mathcal{A} = \{1, \dots, V\}$. Then*

$$\lim_{n\to\infty} \frac{s_n}{\log n} = \frac{1}{\log p_{\min}^{-1}} \qquad \text{(a.s.)},$$

where $p_{\min} = \min\{p_1, p_2, \dots, p_V\}$.

4.12 Consider the following "comparison" of two random strings. Let X_1^n and Y_1^m of length n and $m < n$, respectively, be generated by a binary memoryless source with probabilities p and $q = 1-p$ of the two symbols. Define C_i to be the number matches between X_i^{i+m-1} and Y, that is,

$$C_i = \sum_{j=1}^{n-m+1} equal(X_{i+j-1}, Y_j),$$

where $equal(x, y)$ is one when $x = y$ and zero otherwise (i.e., the Hamming distance). Define $M_{mn} = \max_{1\le i\le n-m+1}\{C_i\}$. Prove that *if* $\log n = o(m)$ *then*

$$\lim_{n\to\infty} \frac{M_{m,n}}{m} = P_2 \qquad \text{(a.s.)},$$

where $P_2 = p^2 + q^2$.

4.13 (Atallah, Jacquet and Szpankowski 1993) Consider the same problem as in Exercise 12 above. Let $n = \Theta(m^\alpha)$ and $P_3 = p^3 + q^3$a Define

$$\beta = \frac{(P_2 - P - 3)(P_2 - 3P_2^2 + 2P_3)}{6(P_3 - p_2^2)}.$$

Prove that if $P_2 - P_3 \le (1-\alpha)\beta$ then $M_{m,n} - mP \sim \sqrt{2m(P_2 - P_3)\log n}$ (pr.) as $n, m \to \infty$.

4.14 Consider an n-dimensional unit cube $I_n = \{0,1\}^n$, i.e., a binary sequence $(x_1, \dots, x_n)$ is regarded as a vertex of the unit cube in n dimensions. Let T_n be the cover time of a random walk, that is, time required by a simple walk to visit all 2^n vertices of the cube. Prove that $T_n \sim (1+n)2^n \log 2$ **whp**.