Chapter 5

Subadditive Ergodic Theorem and Large Deviations

The celebrated *ergodic theorem* of Birkhoff finds many applications in computer science. However, for probabilistic analysis of algorithms a generalization of it due to Kingman, called *subadditive ergodic theorem* is even more important. Kingman's result proves the existence of a limit, but neither says what the limit is nor what is the rate of convergence. Azuma's inequality tries to remedy the latter problem. Its easiest and general derivation is based on Hoeffding's inequality for *martingale differences.* We also review *martingales* since they are useful probabilistic tools for analysis of algorithms.

Birkhoff's ergodic theorem asserts that a *sample mean* of a stationary ergodic process converges almost surely and in moments to the actual mean of the process. In analysis of algorithms we often encounter a situation in which a quantity of interest, say X_0^n, satisfies either a *subadditivity* property (i.e., $X_0^n \leq X_0^m + X_m^n$ for $m \leq n$) or a *superadditivity* property (i.e., $X_0^n \geq X_0^m + X_m^n$). Does it suffice to conclude that X_0^n/n has a limit as $n \to \infty$? In 1976 Kingman [145] proposed conditions under which the answer is in the affirmative. We shall discuss in this chapter a generalization of Kingman's *subadditive ergodic theorem* and its applications to the analysis of algorithms.

The subadditive ergodic theorem has its "cavity". It says that a limit exists but does not say what it is. Actually, finding this limit might be quite troublesome. But, we can do the next best thing, namely, show that the process X_0^n/n is well concentrated around its limit. This can be accomplished through a powerful lemma initiated by Azuma, and further developed by others (cf. McDiarmid [184], Talagrand [239, 240]). We present here a proof of Azuma's inequality based on *martingale differences.* Martingales find many applications in probabilistic analysis of algorithms, and we shall review here some of their properties. In recent years, Talagrand [239] developed a novel isoperimetric theory of concentration inequalities. We will not discuss it here. The reader is either refer to the original paper of Talagrand [239] or a more readable account of Steele [225], Talagrand [239] or McDiarmid [113].

Finally, we observe that Azuma's inequality is an example of a large deviations result. We discuss here in depth large deviations theory of i.i.d. random variables. We shall return to large deviations in Chapter 8. In the application section of this chapter we first discuss two pattern matching problems, namely the edit distance problem and the Knuth–Morris-Pratt algorithm. The last application deals with a certain problem that arose in the analysis of the hashing with lazy deletion in which martingales play a crucial role.

5.1 Subadditive Sequence

Let us start with a deterministic situation. Assume a deterministic sequence $\{x_n\}_{n=0}^{\infty}$ satisfies the so called *subadditivity property*, that is,

$$x_{m+n} \le x_n + x_m$$

for all integers $m, n \ge 0$. We will prove that

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\ge 1} \frac{x_m}{m} = \alpha$$

for some $\alpha \in [-\infty, \infty)$. Indeed, it suffices to fix $m \ge 0$, write $n = km + l$ for some $0 \le l \le m$, and observe that by a consecutive applications of the subadditivity property we arrive at

$$x_n \le kx_m + x_l \ .$$

Dividing by n, and considering $n \to \infty$ with $n/k \to m$, we finally arrive at

$$\limsup_{n\to\infty} \frac{x_n}{n} \le \inf_{m\ge 1} \frac{x_m}{m} \le \alpha \ ,$$

where the inequality follows from arbitrariness of m. This completes the derivation since

$$\liminf_{n\to\infty} \frac{x_n}{n} = \alpha$$

is automatic from the definition of $\liminf$ (cf. Table 5.1). We just derived the following theorem of Fekete.

Theorem 5.1 (Fekete 1923) *If a sequence of real numbers $\{x_n\}$ satisfies the subadditive condition*

$$x_{m+n} \le x_n + x_m \tag{5.1}$$

for all integers $m, n \ge 0$, then

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\ge 1} \frac{x_m}{m}. \tag{5.2}$$

If the subadditivity (5.1) is replaced by the superadditivity property

$$x_{m+n} \ge x_n + x_m \tag{5.3}$$

for all integers $m, n \ge 0$, then

$$\lim_{n\to\infty} \frac{x_n}{n} = \sup_{m\ge 1} \frac{x_m}{m}. \tag{5.4}$$

Example 5.1: *Longest Common Subsequence*

In Chapter 1.5 we discussed the *longest common subsequence* (LCS) problem which is a special case of the edit distance problem. To recall: two ergodic stationary sequences $X = X_1, X_2, \ldots, X_n$ and $Y = Y_1, Y_2, \ldots, Y_n$ are given. Let

$$\begin{aligned} L_n = \max\{K :\ X_{i_k} = Y_{j_k} \text{ for } 1 \le k \le K, \quad &\text{where} \quad 1 \le i_1 < i_2 < \cdots < i_K \le n, \\ &\text{and} \quad 1 \le j_1 < j_2 < \cdots < j_K \le n\} \end{aligned}$$

be the length of the longest common subsequence. Observe that

$$L_{1,n} \ge L_{1,m} + L_{m,n}$$

where $L_{m,n}$ is the longest common subsequence of X_m^n and Y_m^n. The above inequality follows from a simple observation that the longest subsequence may cross the boundary of X_1^m, Y_1^m and X_m^n, Y_m^n, hence it may be bigger then the sum of LCS in each subregion $(1, m)$ and (m, n). Thus, $a_n = \mathbf{E}[L_{1,n}]$ is superadditive. By Theorem 5.1

$$\lim_{n\to\infty} \frac{a_n}{n} = \alpha = \sup_{m\ge 1} \frac{\mathbf{E}[L_m]}{m}.$$

Interestingly enough, we still do not know the value of α even for the i.i.d case (memoryless source). Steele in 1982 conjectured that $\alpha = 2/(1+\sqrt{2}) \approx 0.8284$, and some simulation studies support this guess. The best bounds were are due to Deken [54] who proved that $0.7615 \le \alpha \le 0.8376$ till 1995 when Dančik and Paterson [48] improved them slightly.

Example 5.2: *Rate of Convergence in Theorem 5.1 Can be Arbitrary Slow*

Let $x_n = f(n) \ge 0$ with $f(n)/n$ decreasing where f is an arbitrary function. Then

$$\begin{aligned} x_{n+m} &= f(n+m) = m\frac{f(n+m)}{n+m} + n\frac{f(n+m)}{n+m} \le m\frac{f(m)}{m} + n\frac{f(n)}{n} \\ &= x_m + x_n. \end{aligned}$$

Thus

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\ge 1} \frac{f(m)}{m}$$

and due to the arbitrariness of f the convergence can be as slow as desired. □

It should not be a surprise that the subadditivity property can be relaxed while still keeping unchanged the main thesis of Fekete's theorem. For example, if $x_{m+n} \le x_n + x_m + c$, where c is a constant, then $y_n = x_n + c$ satisfies the subadditivity property, hence x_n/n has a limit that is equal to $\inf x_m/m$. Can we replace the constant c be a sequence $c_n = o(n)$ so that the thesis of Fekete's theorem still holds? The next result solves this problem, while exercises provide further generalizations. We shall follow the presentation of Steele [225].

Table 5.1: Basic properties of limsup and liminf.

We start with a formal definition.

Definition 5.2 *Let a_n be a sequence of real numbers. Limit superior and limit inferior are defined as*

$$\limsup_{n\to\infty} a_n = \inf_{n\geq 0}\{\sup_{k\geq n}\{a_k\}\} = \lim_{n\to\infty}\left(\sup_{m\geq n} a_m\right),$$
$$\liminf_{n\to\infty} a_n = \sup_{n\geq 0}\{\inf_{k\geq n}\{a_k\}\} = \lim_{n\to\infty}\left(\inf_{m\geq n} a_m\right).$$

The following theorem is useful.

Theorem 5.3 *We say that $L<\infty$ is the limit superior of a_n if the following two conditions hold:*
(i) *For every $\varepsilon>0$ there exists an integer N such that $n>N$ implies*

$$a_n < L+\varepsilon,$$

that is, the above holds for all but finitely many n;
(ii) *For every $\varepsilon>0$ and given m, there exists an integer $n>m$ such that*

$$a_n > L-\varepsilon,$$

that is, infinitely many a_n satisfy the above condition.
A similar statement is true for limit inferior.

In words, we can characterize the above limits as follows: Let us consider $\liminf a_n$. Take all possible convergent subsequences a_{n_k} of a_n, and let E be the set of all limits of such subsequences. Then $\liminf a_n = \inf E$. Finally, observe that $\liminf a_n$ and $\limsup a_n$ always exist, and $\lim a_n = a$ if and only if $\liminf a_n = a = \limsup a_n$.
Finally, in applications we often use the following simple result.

Theorem 5.4 *Assume that $a_n \leq b_n$ for all $n=1,2,\ldots$ Then we have*

$$\liminf_{n\to\infty} a_n \leq \liminf_{n\to\infty} b_n, \qquad \limsup_{n\to\infty} a_n \leq \limsup_{n\to\infty} b_n.$$

Theorem 5.5 (DeBruijn and Erdős 1952) *Let c_n be a positive and nondecreasing sequence fulfilling*

$$\sum_{k=1}^{\infty} \frac{c_k}{k^2} < \infty. \tag{5.5}$$

If x_n satisfies

$$x_{n+m} \leq x_n + x_m + c_{n+m}, \tag{5.6}$$

then

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\geq 1} \frac{x_m}{m}.$$

Proof. The idea is to find a new sequence $b_n = o(n)$ such that $a_n = x_n + b_n$ is subadditive. Let

$$b_n = n \sum_{k=n}^{\infty} \frac{c_k}{k^2}$$

which is $o(n)$ due to (5.5). Observe that due to monotonicity of c_n

$$\sum_{k=\ell}^{h} \frac{c_k}{k^2} \geq c_\ell \int_{\ell}^{h+1} t^{-2} dt = c_\ell \left(\frac{1}{\ell} - \frac{1}{h+1} \right).$$

Then $a_n = x_n + b_n$ satisfies

$$\begin{aligned}
a_{m+n} - a_n - a_m &= x_{n+m} - x_n - x_m - m \sum_{k=m}^{m+n} \frac{c_k}{k^2} - n \sum_{k=n}^{m+n} \frac{c_k}{k^2} \\
&\leq c_{n+m} - m c_m \left(\frac{1}{m} - \frac{1}{n+m+1} \right) - n c_n \left(\frac{1}{n} - \frac{1}{n+m+1} \right) \\
&\leq c_{n+m} \left(1 - m \left(\frac{1}{m} - \frac{1}{n+m+1} \right) - n \left(\frac{1}{n} - \frac{1}{n+m+1} \right) \right) \\
&\leq 0
\end{aligned}$$

where the last inequality follows again from the monotonicity of c_n. Thus, we prove that $x_n + b_n$ is subadditive, hence it has the limit equal to $\inf_{m\geq 1}(x_m/m + b_m/m)$. But, by (5.5) $b_m/m = o(m)$, which completes the proof. ∎

As the reader may expect, there are many interesting embellishments of the subadditive theorem. We discuss some of them in the exercise section, but two stand out above all and we present a short account on them here following Steele's presentation from [225]. The first one deals with the rate of convergence (for a generalization see Exercise 2).

Theorem 5.6 (Pólya and Szegő 1924) *If a real sequence x_n satisfies*

$$x_n + x_m - c \leq x_{n+m} \leq x_n + x_m + c \tag{5.7}$$

for all $n, m \geq 1$, then there is a finite constant α such that

$$\left|\frac{x_n}{n} - \alpha\right| \leq \frac{c}{n} \tag{5.8}$$

for all $n \geq 1$.

Proof. First of all, observe that setting $m = n$ in (5.7) we obtain

$$\left|\frac{x_{2m}}{2m} - \frac{x_m}{m}\right| \leq \frac{c}{2m}$$

for all $m \geq 1$. Also, since $x_n + c$ is subadditive, by Fekete's theorem $x_n/n \to \inf x_m/m = \alpha$. Using this and the above we obtain

$$\left|\alpha - \frac{x_n}{n}\right| < \left|\frac{x_{2m}}{2m} - \frac{x_m}{m}\right| + \left|\frac{x_{4m}}{4m} - \frac{x_{2m}}{2m}\right| + \cdots < \sum_{k\geq 1} \frac{c}{2^k m} = \frac{c}{m},$$

as desired. ∎

Finally, the next theorem due to DeBruijn and Erdős shows that we can relax the requirement that the subadditivity holds for all m and n. We ask in Exercise 3 to prove it (cf. [225]).

Theorem 5.7 (DeBruijn and Erdős 1952) *If a real sequence x_n satisfies the subadditivity condition for the restricted range of m and n*

$$x_{m+n} \leq x_n + x_m \qquad \text{for} \quad \frac{1}{2}n \leq m \leq 2n,$$

then

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\geq 1} \frac{x_m}{m} = \alpha$$

where $-\infty \leq \alpha < \infty$.

5.2 Subadditive Ergodic Theorem

In the early seventies there was a resurgence of interest in generalizing Fekete's deterministic subadditivity result to a sequence of random variables. Such an extension would have an impact on many research problems of those days. For example, Chvatal and Sankoff [39] used ingenious tricks to establish the probabilistic behavior of the *Longest Common Superstring* problem (cf. Example 1) while we know that it is a trivial consequence of a stochastic extension of the subadditivity result. In 1976 Kingman [145] presented the first proof of what later will be known as the *Subadditivity Ergodic Theorem.* Below, we present an extension of Kingman's result due to Liggett [172] and Derriennic [55].

To formulate it we must consider a sequence of doubly-indexed random variables $X_m^n = (X_m, X_{m+1}, \dots, X_n)$. However, traditionally, when dealing with subadditivity theorems, one denotes such a substring as $X_{m,n} = X_m^n$, and we adopt it here. Our goal is to say something about probabilistic behavior of $X_{0,n}$ as $n \to \infty$. One expects that $X_{0,n} \sim n\alpha$ in a probabilistic sense, where α is a constant. This turns out to be true under certain assumptions specified below in the *Subadditive Ergodic Theorem.*

Theorem 5.8 (Kingman 1976, Liggett 1985) *Let $X_{m,n}$ $(m < n)$ be a sequence of random variables satisfying the following properties*

(i) $X_{0,n} \le X_{0,m} + X_{m,n}$ *(subadditivity);*

(ii) *For every k, $\{X_{nk,(n+1)k},\ n \ge 1\}$ is a stationary sequence.*

(iii) *The distribution of $\{X_{m,m+k},\ k \ge 1\}$ does not depend on m.*

(iv) $\mathbf{E}[X_{0,1}] < \infty$ *and for each n, $\mathbf{E}[X_{0,n}] \ge c_0 n$ where $c_0 > -\infty$.*

Then,

$$\begin{aligned} \lim_{n\to\infty} \frac{\mathbf{E}[X_{0,n}]}{n} &= \inf_m \frac{\mathbf{E}[X_{0,m}]}{m} := \alpha, && (5.9) \\ \lim_{n\to\infty} \frac{X_{0,n}}{n} &= X \quad \text{a.s. and in } L^1, && (5.10) \end{aligned}$$

such that $\mathbf{E}[X] = \alpha$. Finally, if all stationary sequences in (ii) are ergodic, then

$$\lim_{n\to\infty} \frac{X_{0,n}}{n} = \alpha \quad \text{(a.s.)} \tag{5.11}$$

Clearly, (5.9) follows directly from Fekete's Theorem 5.1 while the proof of almost sure convergence and in mean (5.10) is beyond the scope of this book. An accessible proof can be found in Durrett [69]. Also, the original proof of Kingman [145] is readable. In fact, Kingman proved a weaker version of the theorem in which (ii) and (iii) are replaced by

(ii') $X_{m,n}$ is stationary (i.e., the joint distributions of $X_{m,n}$ are the same as $X_{m+1,n+1}$) and ergodic.

There are examples in which (ii') does not hold while Liggett's version works fine (cf. [69]).

Before, we proceed with examples (see also Section 5.5) we mention a generalization of the subadditive ergodic theorem that is known as *Almost Subadditive Ergodic Theorem.*

Theorem 5.9 (Derriennic 1983) *Let (ii') and (iv) above be fulfilled and (i) be replaced by*

$$X_{0,n} \le X_{0,m} + X_{m,n} + A_n \tag{5.12}$$

such that $\lim_{n\to\infty} \mathbf{E}[A_n/n] = 0$, *then (5.11) holds, too.*

We must point out, however, that the above findings establish only the existence of a constant α such that (5.11) holds. They say *nothing* how to compute the limit, and in fact many ingenious methods have been devised in the past to bound the constant.

Example 5.3 *The First Birth Problem For an Age-Dependent Branching Process*
Let us consider an *age-dependent branching process* in which each individual i lives for time t_i with distribution F before producing k offsprings with probability p_k. The process starts with one individual in generation 0 who is born at time 0, and when dying its offsprings start independent copies of the original process. Our interest lies in estimating the birth time B_n of the first member of generation n (cf. [28, 29, 146]). To estimate B_n, let $B_{0,m}$ be the birth of the first individual in generation m, and $B_{m,n}$ be the time needed for this individual to have an offspring in generation $n > m$. Only little reflection is needed to find out that $B_{0,n} \le B_{0,m} + B_{m,n}$, however, it is *not* true that for $l > 0$ we also have $B_{l,n} \le B_{l,m} + B_{m,n}$. Since $B_n = B_{0,n}$ is a nonnegative process, hence all assumptions of Theorem 5.8 are fulfilled, and therefore there exists a constant α such that

$$\frac{B_n}{n} \to \alpha \qquad \text{(a.s.)}.$$

Interestingly enough, in this case we can compute the constant α. In Exercise 16 we guide the reader through Biggins' derivation of the constant (cf. [28]).

The first birth problem is of prime importance to the analysis of algorithms since in 1986 Devroye [60] used it to prove the long standing conjecture concerning the height of a binary search tree.

Example 5.4 *Increasing Sequences in Random Permutations*
Let π be a permutation of $\{1, 2, \ldots, n\}$ and let $l_n(\pi)$ (or shortly l_n) be the longest increasing sequence in π, that is, the largest k for which there are $i_1 < i_2 < \cdots < i_k$ such that $\pi(i_1) < \pi(i_2) < \cdots < \pi(i_k)$. This example is interesting since it is easy to see that $\ell(\pi)$ is *not* linear in n, hence the subadditive ergodic theorem cannot be applied. We have the following estimate by Boole's inequality

$$\Pr\{l_n(\pi) \ge 2e\sqrt{n}\} \le \exp(-2e\sqrt{n}).$$

Indeed, there are $\binom{n}{k}$ subsequences of length k and each has probability $1/k!$, so that

$$\Pr\{l_n(\pi) \ge k\} \le \binom{n}{k}\frac{1}{k!}.$$

Using Stirling's approximation with $k = 2e\sqrt{n}$, we prove the estimate. But, for any k

$$\begin{aligned} \mathbf{E}[l_n(\pi)] &= \sum_{0\le i\le k} i\mathsf{Pr}\{l_n = i\} + \sum_{k+1\le i\le n} i\mathsf{Pr}\{l_n = i\} \\ &\le k + n\mathsf{Pr}\{l_n(\pi) \ge k\}, \end{aligned}$$

hence for sufficiently large n we have $\mathbf{E}[\ell_n(\pi)] \le c\sqrt{n}$ for some $c > e$. The longest increasing sequence cannot grow linearly!

To circumvent this problem, we follow Hammersley and *poissonize*, that is, embed the process l_n into a Poisson stream of rate 1 in the plane. Poissonization is a powerful probabilistic technique to which we devote Chapter 10 of this book. We denote by $l(t,s)$ $(t < s)$ the length of the longest increasing sequence in the rectangle with corners $(s,s),(s,t),(t,t)$ and (t,s). By the longest sequence in the plane we mean the largest k for which there are points (x_i, y_i) in the Poisson process with $s < x_1 < \cdots < x_k < t$ and $s < y_1 < \cdots < y_k < t$. Now, we are in the superadditive world, that is,

$$l(0, s+t) \ge l(0,s) + l(s, s+t)$$

since the longest increasing sequence can go through two other rectangles not included in $l(0,s)$ and $l(s,s+t)$. Then, by the superadditive ergodic theorem

$$\lim_{t\to\infty} \frac{l(t)}{t} = \alpha \qquad \text{(a.s.)}$$

The ease at which we obtain this result has its price. We must *depoissonize* it to recover the probabilistic behavior of l_n. Chapter 10 of this book is devoted to *analytic depoissonization*. Here, we apply a theorem from that chapter adopted to our situation. Let us concentrate on estimating the average $\mathbf{E}[l_n]$ which we denote for simplicity as g_n. The idea is to choose such t that on a square of the area t^2 there are exactly n points. In other words, we shall condition on the Poisson process having n points in the square $(0,t)^2$. But we just proved that

$$\mathbf{E}[l(t)] = \sum_{k=0}^{\infty} g_k \frac{(t^2)^k}{k!} e^{-t^2} \sim \alpha t, \tag{5.13}$$

for large t. We need to recover $g_n = \mathbf{E}[l_n]$ for large n from the above. This is exactly where the analytic depoissonization can help. Setting $t^2 = z$ in (5.13), we observe that the so called Poisson transform $\widetilde{G}(z) = \sum_{k=0}^{\infty} g_k \frac{z^k}{k!} e^{-z}$ of $g_n = \mathbf{E}[l_n]$ grows like $\alpha\sqrt{z}$, hence by the depoissonization Theorem 10.3 of Chapter 10 we conclude that $\mathbf{E}[l_n] = g_n = \alpha\sqrt{n} + o(\sqrt{n})$. This completes the derivations. □

5.3 Martingales and Azuma's Inequality

The subadditive ergodic theorem is a very powerful tool that allows to predict the growth rate of a quantity of interest. As we have seen before (cf. Example 4), the technique is not restricted to linear growth, but then one must use also other approaches. The drawback of the subadditivity method lies in its inability to evaluate the constant α. But we can try to do the next best thing, namely, assess the deviation from the most likely value αn. This is known as the *concentration of measure (probability)*. Surprisingly enough, there is a general technique for the concentration of measure even without any knowledge about α. Such results are known in literature under the name *Azuma's inequality* or the *method of bounded differences* (cf. [184]).

Azuma's inequality is an outcome of Hoeffding's large deviations lemma applied to martingales. Before we plunge into technical details, we would like to make two comments. Large deviations is discussed in depth in the next section of this chapter and we return to it in Chapter 8. Azuma's inequality could be viewed as a special case of large deviations, but we are of the opinion that its applicability and usefulness deserve a special attention. In fact, at the heart of Azuma's inequality are *martingales*. Martingale is another very useful tool in the probabilistic analysis of algorithms, so we recall their basic properties in Table 5.3. The application Section 5.5.3 provides a more sophisticated example of martingales applications. In passing, we mention here that Régnier [203] (cf. Mahmoud [182]) used the *martingales convergence theorem* to establish the limiting distribution of the path length in a binary search tree or the quicksort algorithm.

Martingales are standard probabilistic tools, and the interested reader is referred to Durrett [69] for a more detailed discussion. Of supreme importance for martingales is its convergence theorem (cf. Theorem 5.12). However, in the analysis of algorithms a bound on the degree of fluctuation of a martingale is even more interesting. This bound will lead directly to Azuma's inequality which finds an abundance of applications.

Let then $\{Y_n\}_{n=0}^{\infty}$ be a martingale with respect to the filtration $\mathcal{F}$, or simply Y_n is a martingale with respect to $X_0, X_1, \ldots$, as discussed in Definition 5.10 and Definition 5.11 of Table 5.3. We now define the *martingale difference* sequence $\{D_n\}_{n=1}^{\infty}$ as

$$D_n = Y_n - Y_{n-1}, \tag{5.14}$$

so that

$$Y_n = Y_0 + \sum_{i=1}^{n} D_i. \tag{5.15}$$

Clearly, D_n depends only on the knowledge contained in $X_0, \ldots, X_n$, or simply is $\mathcal{F}_n$-measurable where $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ is the smallest σ-field generated by $(X_0, \ldots, X_n)$. Also, by the definition of martingales $\mathbf{E}[|D_n|] < \infty$. In fact, for all n

$$\begin{aligned} \mathbf{E}[D_{n+1}|\ \mathcal{F}_n] &= \mathbf{E}[Y_{n+1}|\ \mathcal{F}_n] - \mathbf{E}[Y_n|\mathcal{F}_n] = Y_n - Y_n \\ &= 0. \end{aligned} \tag{5.16}$$

Thus, also $\mathbf{E}[D_n] = 0$. Actually, let us consider the product $D_{n+1}D_n$. Observe that

$$\mathbf{E}[D_{n+1}D_n|\ \mathcal{F}_n] = D_n\mathbf{E}[D_{n+1}|\ \mathcal{F}_n] = 0.$$

Table 5.2: BASIC PROPERTIES OF MARTINGALES.

Martingales, besides Markov processes and stationary processes, are another broad class of processes that often arise in the analysis of algorithms.

Definition 5.10 *A sequence $Y_n = f(X_1, \ldots, X_n)$ $(n \geq 0)$ is a martingale with respect to the sequence $X_0, X_1, \ldots$ if for all $n \geq 0$*

(i) $\mathbf{E}[|Y_n|] < \infty$,

(ii) $\mathbf{E}[Y_{n+1}|\ X_0, X_1, \ldots, X_n] = Y_n$.

We should observe that $\mathbf{E}[Y_{n+1}|\ X_0, X_1, \ldots, X_n]$ defines a *random variable* depending on $X_0, \ldots, X_n$ (or more precisely, defined on the σ-field with respect to each $X_0, X_1, \ldots, X_n$). The best known and simplest example of a martingale is constructed as follows: define $S_n = X_1 + \cdots + X_n$ with X_i being i.i.d. with mean $\mu = \mathbf{E}[X_i]$. Then, $Y_n = S_n - n\mu$ is a martingale since

$$\mathbf{E}[S_{n+1}|\ X_0, X_1, \ldots, X_n] = \mathbf{E}[S_n + X_{n+1}|\ X_0, X_1, \ldots, X_n] = S_n + \mu.$$

Above we use $\mathbf{E}[S_n|\ X_1, \ldots, X_n] = S_n$ since S_n is completely defined on $X_1, \ldots, X_n$.
From the above, we may conclude that for the definition of martingale we only require the *knowledge* (i.e., information) obtained from $X_0, X_1, \ldots, X_n$. In fact, this vague idea can be made rigorous, if we agree to enter the realm of σ-fields. We do, since the advantage of such approach will be immediately visible (e.g., shortening our notation). In general, let $\mathcal{F} = \{\mathcal{F}_0, \mathcal{F}_1, \ldots\}$ be a sequence of sub-σ-fields of $\mathcal{F}$. We call $\mathcal{F}$ a *filtration* if $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all n (i.e., $\mathcal{F}_n$ is more coarser than $\mathcal{F}_{n+1}$). A sequence Y_n $(n \geq 0)$ is said to be *adapted* to the filtration $\mathcal{F}$, if Y_n is $\mathcal{F}_n$ measurable for all n (i.e., can be properly defined on $\mathcal{F}_n$). With this in mind, we can generalize the definition of martingales to the following one.

Definition 5.11 *We say that the sequence Y_n is a martingale with respect to the filtration $\mathcal{F}$ if for all $n \geq 0$ (i) above holds together with*

(ii) $\mathbf{E}[Y_{n+1}|\ \mathcal{F}_n] = Y_n$.

We recover the previous definition by choosing $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$, i.e., the smallest σ-fields with respect to which the variables X_i are defined. Finally, we discuss one result known as the *martingale convergence theorem* which we present here in a simpler version.

Theorem 5.12 *If Y_n is a martingale with $\mathbf{E}[Y_n^2] \leq M < \infty$ for some M, then there exists a random variable Y such that Y_n converges to Y almost surely and in mean square.*

In general, we can conclude that for any $1 \le i_1 < i_2 < \cdots < i_k \le n$

$$\mathbf{E}[D_{i_1} D_{i_2} \cdots D_{i_k}] = 0. \tag{5.17}$$

This property, as pointed out by Steele [225], is the only one required to prove the theorem below (but we formulate it in a more traditional way and ask the reader to generalize it in Exercise 4).

Theorem 5.13 (Hoeffding's Inequality) *Let $\{Y_n\}_{n=0}^{\infty}$ be a martingale with respect to the filtration $\mathcal{F}$, and let there exist constant c_n such that for all $n \ge 0$*

$$|Y_n - Y_{n-1}| \le c_n. \tag{5.18}$$

Then for $x > 0$

$$\Pr\{|Y_n - Y_0| \ge x\} = \Pr\left\{ |\sum_{i=1}^{n} D_i| \ge x \right\} \le 2 \exp\left(-\frac{x^2}{2\sum_{i=1}^{n} c_i^2} \right) \tag{5.19}$$

Proof. We start with a simple inequality for the exponential function. Observe that for $\beta > 0$ the exponential function $e^{\beta d}$ is convex for any d, and hence on the interval $d \in (-1, 1)$ we obtain

$$e^{\beta d} \le \frac{1}{2}(1-d)e^{-\beta} + \frac{1}{2}(1+d)e^{\beta}.$$

If now d is replaced by a random variable D with mean $\mathbf{E}[D] = 0$, then the above becomes

$$\mathbf{E}[e^{\beta D}] \le \frac{1}{2}\left(e^{-\beta} + e^{\beta}\right) < e^{\frac{1}{2}\beta^2} \tag{5.20}$$

where the last inequality follows immediately from the Taylor expansions of both sides and a comparison of the coefficients at β^{2k}.

We use the above inequality established for D with mean zero to prove the claim. We only derive (5.19) for $Y_n - Y_0 \ge x$ since our arguments will be valid for the other side of the inequality. Observe now that by Markov's inequality (cf. (2.5)) we have

$$\Pr\{Y_n - Y_0 \ge x\} = \Pr\{e^{\theta(Y_n - Y_0)} \ge e^{\theta x}\} \le e^{-\theta x}\mathbf{E}[e^{\theta(Y_n - Y_0)}] \tag{5.21}$$

for some $\theta > 0$. Since $D_n = Y_n - Y_{n-1}$, we can write $Y_n - Y_0 = Y_{n-1} - Y_0 + D_n$. Conditioning now on $\mathcal{F}_{n-1}$ (the reader can think of conditioning on all knowledge extracted from $Y_0, Y_1, \ldots, Y_{n-1}$) we obtain

$$\begin{aligned} \mathbf{E}[e^{\theta(Y_n - Y_0)} | \mathcal{F}_{n-1}] &= e^{\theta(Y_{n-1} - Y_0)}\mathbf{E}[e^{\theta D_n} | \mathcal{F}_{n-1}] \\ &\le e^{\theta(Y_{n-1} - Y_0)} e^{\frac{1}{2}\theta^2 c_n^2} \end{aligned}$$

where the last inequality follows from (5.20) applied to D_n/c_n which is smaller than one by (5.18). Now, taking expectation of the above with respect to $\mathcal{F}_n$ we have

$$\mathbf{E}[e^{\theta(Y_n-Y_0)}] \le e^{\frac{1}{2}\theta^2 c_n^2}\mathbf{E}[e^{\theta(Y_{n-1}-Y_0)}],$$

and after n iterations we arrive at

$$\mathbf{E}[e^{\theta(Y_n-Y_0)}] \le \exp\left(\frac{1}{2}\theta^2\sum_{i=1}^{n} c_i^2\right).$$

Hence, by (5.21) we prove that

$$\Pr\{Y_n - Y_0 \ge x\} \le \exp\left(-\theta x + \frac{1}{2}\theta^2\sum_{i=1}^{n} c_i^2\right)$$

which becomes (5.19) after the substitution $\theta = x\left(\sum_{i=1}^{n} c_i^2\right)^{-1}$ (which maximizes the exponent of the above). This completes the proof. ∎

Our derivation of (5.19) is quite simple but we did not get the optimal constant in the exponent. The reader is asked in Exercise 5 to prove that (5.19) can be improved to

$$\Pr\{|Y_n - Y_0| \ge x\} = \Pr\left\{|\sum_{i=1}^{n} D_i| \ge x\right\} \le 2\exp\left(-\frac{2x^2}{\sum_{i=1}^{n} c_i^2}\right). \tag{5.22}$$

Before we derive the Azuma concentration of measure (cf. Theorem 5.15), we illustrate it on an example. The reader should study very carefully this example since it explains the main idea behind the Azuma inequality and a construction of martingale differences in combinatorial optimization and analysis of algorithms.

Example 5.5 *Bin Packing*

Consider n objects with random sizes $X_1, X_2, \ldots, X_n$ having a common distribution on $[0,1]$, and unlimited collection of bins each of size 1. Let $B_n := B_n(X_1, \ldots, X_n)$ be the minimum number of bins required to pack the objects. First of all, it is not difficult to see that B_n is subadditive, that is, $B_{0,n} \le B_{0,m} + B_{m,n}$, hence by Theorem 5.8 there exists a constant α such that

$$\begin{aligned}\lim_{n\to\infty}\frac{\mathbf{E}[B_n]}{n} &= \alpha,\\ \lim_{n\to\infty}\frac{B_n}{n} &= \alpha \qquad \text{(a.s.)}\end{aligned}$$

The evaluation of α is quite complicated, if possible at all, but we can answer the next interesting question, namely, how close is B_n to its mean value $\mathbf{E}[B_n] \sim \alpha n$. As expected, we shall use Hoeffding's inequality. Thus, we must define a martingale and martingale differences. Here it comes. Let

$$\begin{aligned} Y_i &= \mathbf{E}[B_n(X_1,\ldots,X_n)|\ X_1,\ldots,X_i] \\ &= \mathbf{E}[B_n|\ \mathcal{F}_i], \end{aligned}$$

where throughout we assume that $\mathcal{F}_i = \sigma(X_1,\ldots,X_i)$. In other words, Y_i is the average of the minimum number of bins required to pack n items, provided someone reveals to us sizes $X_1,\ldots,X_i$ of the first i objects. Is it a martingale? Yes, it is and we prove it below.

Our task is to show that $\{Y_i\}_{i=0}^n$ is a finite martingale, that is, for all $0 \le i \le n-1$

$$\mathbf{E}[Y_{i+1}|\ \mathcal{F}_i] = \mathbf{E}[\mathbf{E}[B_n|\ \mathcal{F}_{i+1}]|\ \mathcal{F}_i] = \mathbf{E}[B_n|\ \mathcal{F}_i] = Y_i \tag{5.23}$$

The middle equality, which we shall prove in its whole generality in Lemma 5.14, is at the heart of many martingale applications. In this example, instead of being general and abstract we present a simple and boring proof that reveals all the details. Even more, we simply prove the above for $n=3$, that is, we show

$$\mathbf{E}[Y_2|\ X_1] = \mathbf{E}[\mathbf{E}[B_n(X_1,X_2,X_3)|\ X_1,X_2]|\ X_1] = \mathbf{E}[B_n(X_1,X_2,X_3)|\ X_1] = Y_1. \tag{5.24}$$

Indeed, let us compute the above averages. Observe that the right-hand side of the above becomes

$$\mathbf{E}[B_n(X_1,X_2,X_3)|\ X_1] = \int_{x_2}\int_{x_3} B(X_1,x_2,x_3) f(X_1,x_2,x_3|X_1)dx_2dx_3$$

where $f(X_1,x_2,x_3|X_1) = f(X_1,x_2,x_3)/f(X_1)$ is the conditional density of (X_1,x_2,x_3) given X_1. Needless to say, the above average is a random variable (since X_1 is a random variable). Let us now compute the left-hand side of (5.24). Let $Z(X_1,X_2) = \mathbf{E}[B_n(X_1,X_2,X_3)|\ X_1,X_2]$, that is, $Z(X_1,X_2)$ is a random function of X_1,X_2 whose value is

$$Z(X_1,X_2) = \mathbf{E}[B_n(X_1,X_2,X_3)|\ X_1,X_2] = \int_{x_3} B(X_1,X_2,x_3)\frac{f(X_1,X_2,x_3)}{f(X_1,X_2)}dx_3.$$

Its distribution is $f(x_1,x_2)$ when $X_1 = x_1$ and $X_2 = x_2$. Then,

$$\begin{aligned} \mathbf{E}[Z(X_1,X_2)|\ X_1] &= \int_{x_2} Z(X_1,x_2)\frac{f(X_1,x_2)}{f(X_1)}dx_2 \\ &= \int_{x_2}\int_{x_3} B(X_1,x_2,x_3)\frac{f(X_1,x_2,x_3)}{f(X_1)}dx_2dx_3, \end{aligned}$$

which is equal to $\mathbf{E}[B_n(X_1,X_2,X_3)|\ X_1]$ as computed above. This proves (5.24). The same analysis shows that for any $1 \le i \le n$ we have $\mathbf{E}[Y_{i+1}|\ \mathcal{F}_i] = Y_i$.

Knowing that Y_i is a martingale, we define the martingale difference as $D_i = Y_i - Y_{i-1}$, and verify that

$$\begin{aligned} \mathbf{E}[D_i|\mathcal{F}_{i-1}] &= \mathbf{E}[\mathbf{E}[B_n|\ \mathcal{F}_i]|\ \mathcal{F}_{i-1}] - \mathbf{E}[B_n|\mathcal{F}_{i-1}] \\ &= \mathbf{E}[B_n|\ \mathcal{F}_{i-1}] - \mathbf{E}[B_n|\mathcal{F}_{i-1}] = 0, \end{aligned} \tag{5.25}$$

as desired. In fact, the equality (5.25) follows from (5.23) (cf. also Lemma 5.14 below).

It remains now to observe that $Y_n = \mathbf{E}[B_n|X_1, \ldots, X_n] = B_n$ and $Y_0 = \mathbf{E}[B_n]$, hence we are almost ready to apply Hoeffding's inequality, if we can prove that $D_i = Y_i - Y_{i-1}$ is bounded. The only difference between Y_i and Y_{i-1} is that the former knows one more size, namely X_i. But, in the worst case this cannot improve the packing by more than one bin! Thus, $|D_i| = |Y_i - Y_{i-1}| \leq 1$. In view of this, Theorem 5.13 allows us to write

$$\Pr\{|B_n - \mathbf{E}[B_n]| \geq x\} \leq 2\exp\left(-\frac{x^2}{2n}\right).$$

We do not know exactly $\mathbf{E}[B_n]$, except that it is $O(n)$. Let then $x = \varepsilon n$ for any $\varepsilon > 0$. Thus

$$\Pr\{|B_n - \alpha n| \geq \varepsilon n\} \leq 2\exp\left(-\frac{1}{2}\varepsilon^2 n(1+o(1))\right)$$

which is a useful bound as long as $n\varepsilon^2 \to \infty$. □

In the above example, we used the following result that is crucial for any martingale application; in particular, for the Azuma inequality.

Lemma 5.14 *If $\mathcal{F}_1 \subset \mathcal{F}_2$, then*

$$\mathbf{E}[\mathbf{E}[Y|\mathcal{F}_1]|\mathcal{F}_2] = \mathbf{E}[Y|\mathcal{F}_1], \tag{5.26}$$

$$\mathbf{E}[\mathbf{E}[Y|\mathcal{F}_2]|\mathcal{F}_1] = \mathbf{E}[Y|\mathcal{F}_1]. \tag{5.27}$$

In words, the smaller σ-field always wins. In particular,

$$\mathbf{E}[\mathbf{E}[Y|\ X_1, X_2]|\ X_1] = \mathbf{E}[Y|\ X_1] = \mathbf{E}[\mathbf{E}[Y|\ X_1]|\ X_1, X_2].$$

Proof. The proof of (5.26) is quite simple. Observe that $\mathbf{E}[f(X)|\ X] = f(X)$ since $f(X) \in \sigma(X)$ for a measurable function f. But $\mathbf{E}[Y|\mathcal{F}_1] \in \mathcal{F}_2$ since $\mathcal{F}_1 \subset \mathcal{F}_2$, hence (5.26) follows. The second part is a little more intricate, but we have already given a sketch of the proof above in Example 5. Here is another derivation: Notice that $\mathbf{E}[Y|\mathcal{F}_1] \in \mathcal{F}_1$ and for $A \in \mathcal{F}_1 \subset \mathcal{F}_2$ we have

$$\int_A \mathbf{E}[Y|\mathcal{F}_1]dP = \int_A Y dP = \int_A \mathbf{E}[Y|\mathcal{F}_2]dP$$

which completes the proof. ■

We are now ready to derive the Azuma inequality. Let

$$Y_n = F_n(X_1, \ldots, X_n)$$

be a martingale with respect to $X_1, \ldots, X_n$. Define the martingale difference as

$$\begin{aligned} D_i &= \mathbf{E}[Y_n|\mathcal{F}_i] - \mathbf{E}[Y_n|\mathcal{F}_{i-1}] \\ &= \mathbf{E}[Y_n|\ X_1, \ldots, X_i] - \mathbf{E}[Y_n|\ X_1, \ldots, X_{i-1}]. \end{aligned}$$

Indeed, the above is a martingale difference since by Lemma 5.14

$$\mathbf{E}[D_i|\mathcal{F}_{i-1}] = \mathbf{E}[\mathbf{E}[Y_n|\mathcal{F}_i|\mathcal{F}_{i-1}] - \mathbf{E}[Y_n|\mathcal{F}_{i-1}] = \mathbf{E}[Y_n|\mathcal{F}_{i-1}] - \mathbf{E}[Y_n|\mathcal{F}_{i-1}] = 0.$$

Also, it is easy to see that

$$\mathbf{E}[Y_n|\mathcal{F}_n] = Y_n, \qquad \text{and} \qquad \mathbf{E}[Y_n|\mathcal{F}_0] = \mathbf{E}[Y_n]$$

since $\mathcal{F}_0$ does not have any information about $X_1, \ldots, X_n$ while $\mathcal{F}_n$ completely describes Y_n.

In view of the above, we can apply Hoeffding's inequality to D_i provided they are bounded, say $|D_i| \le c_i$. However, the real trick of Azuma is to explore this generic representation to find a simple way of bounding D_i. In order to accomplish this, let $\hat{X}_i$ be a new independent random variable with the same distribution as X_i. The point to observe is that $\mathcal{F}_i$ has no information about this new random variable, so that

$$\mathbf{E}[F_n(X_1, \ldots, X_i, \ldots, X_n)|\mathcal{F}_{i-1}] = \mathbf{E}[F_n(X_1, \ldots, \hat{X}_i, \ldots, X_n)|\mathcal{F}_i].$$

Hence

$$\begin{aligned} D_i &= \mathbf{E}[F_n(X_1, \ldots, X_i, \ldots, X_n)|\mathcal{F}_i] - \mathbf{E}[F_n(X_1, \ldots, X_i, \ldots, X_n)|\mathcal{F}_{i-1}] \\ &= \mathbf{E}[F_n(X_1, \ldots, X_i, \ldots, X_n)|\mathcal{F}_i] - \mathbf{E}[F_n(X_1, \ldots, \hat{X}_i, \ldots, X_n)|\mathcal{F}_i]. \end{aligned}$$

In view of this, if we postulate for every $1 \le i \le n$

$$|F_n(X_1, \ldots, X_i, \ldots, X_n)|\mathcal{F}_i] - F_n(X_1, \ldots, \hat{X}_i, \ldots, X_n)|\mathcal{F}_i]| \le c_i$$

then $|D_i| \le c_i$, and Hoeffding's inequality implies immediately the following Azuma's result.

Theorem 5.15 (Azuma's type Inequality) *Let $Y_n = F_n(\cdot, \ldots, \cdot)$ be a martingale with respect to $X_1, \ldots, X_n$ such that for every $1 \le i \le n$ there exist constant c_i*

$$|F_n(X_1, \ldots, X_i, \ldots, X_n) - F_n(X_1, \ldots, \hat{X}_i, \ldots, X_n)| \le c_i \ , \tag{5.28}$$

where $\hat{X}_i$ is independent of X_i and has the same distribution as X_i. Then

$$\begin{aligned} \Pr\{|Y_n - \mathbf{E}[Y_n]| \ge x\} &= \Pr\{|F_n(X_1, \ldots, X_n) - \mathbf{E}[F_n(X_1, \ldots, X_n)]| \ge x\} \\ &\le 2\exp\left(-\frac{x^2}{2\sum_{i=1}^n c_i^2}\right) \end{aligned} \tag{5.29}$$

for some $x > 0$.

In the application section (cf. Section 5.5) we discuss some non-trivial applications of Azuma's inequality. The reader can also consult McDiarmid [184] for many interesting applications in graph theory. In passing, we point out that the concentration of mean for the bin packing Example 5 can be obtained directly from the Azuma inequality.

5.4 Large Deviations

Hoeffding's inequality presented in the previous section is an example of a large deviations bound when applied to the martingale $S_n - n\mathbf{E}[X_1]$ where $S_n = X_1 + \cdots + X_n$. We devote this section to a detailed discussion of large deviations of S_n. Before we plunge into technical details, we first offer an explanation why one must investigate large deviations; why the central limit theorem studied so extensively in probability theory is not sufficient.

Let us consider a sequence $X_1, \ldots, X_n$ of i.i.d. random variables, and let $S_n = X_1 + \cdots + X_n$. Define $\mu := \mathbf{E}[X_1]$ and $\sigma^2 := \mathbf{Var}[X_1]$. To study limiting laws of S_n one must normalize it appropriately. For example, for the central limit theorem the following normalization is useful

$$s_n := \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Let $F_n(x) := \Pr\{s_n \leq x\}$ and $\Phi(x)$ be the distribution function of the *standard normal distribution*, that is,

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt \ .$$

The *Central Limit Theorem* asserts that $F_n(x) \to \Phi(x)$ for continuity points of $\Phi_n(\cdot)$, provided $\sigma < \infty$ (cf. [69, 73]). A stronger version is due to Berry-Essén (cf. Lemma 8.28 of Chapter 8) who proved that

$$|F_n(x) - \Phi(x)| \leq \frac{2\rho}{\sigma^2\sqrt{n}} \tag{5.30}$$

where $\rho = \mathbf{E}[|X-\mu|^3] < \infty$. Finally, Feller [73] showed that if centralized moments $\mu_2, \ldots, \mu_r$ exist, then

$$F_n(x) = \Phi(x) - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \sum_{k=3}^{r} n^{-\frac{1}{2}k+1} R_k(x) + O\left(n^{-\frac{1}{2}r+\frac{1}{2}}\right) \tag{5.31}$$

uniformly in x, where $R_k(x)$ is a polynomial of degree k depending only on $\mu_1, \ldots, \mu_r$ but not on n and r. We shall return to these problems in Chapter 8 where we study some analytic tools useful for such investigations.

One should notice, however, the weakness of the central limit result that is able only to assess the probability of *small deviations* from the mean. Observe that (5.31) is true only for $x = O(1)$ (i.e., for $S_n \in (\mu n - O(\sqrt{n}), \mu n + O(\sqrt{n}))$ due to a polynomial rate of convergence as shown in (5.30). To see this even more explicitly we quote here a result of Greene and Knuth [101] that we prove in Section 8.4.3, namely

$$\Pr\{S_n = \mu n + r\} = \frac{1}{\sigma\sqrt{2\pi n}} \exp\left(\frac{-r^2}{2\sigma^2 n}\right)\left(1 - \frac{\kappa_3}{2\sigma^4}\left(\frac{r}{n}\right) + \frac{\kappa_3}{6\sigma^6}\left(\frac{r^3}{n^2}\right)\right) + O\left(n^{-\frac{3}{2}}\right) \quad (5.32)$$

where κ_3 is the third cumulant of X_1. Now, it should be clear that when $r = O(\sqrt{n})$ (i.e., $x = O(1)$ in our previous formulas) the error term is of the same order as the leading term of the asymptotic expression, thus the whole estimate becomes useless.

From the above discussion, one should conclude that the central limit theorem has limited range of application, and one should expect another law for *large deviations* from the mean, that is, when $x \to \infty$. The most interesting from the application point of view is the case when $x = O(\sqrt{n})$ (or $r = O(n)$), that is, for $\Pr\{S_n = n(\mu + \delta)\}$ for $\delta \neq 0$. Hereafter, we shall discuss this large deviations behavior hereafter.

Actually, we have already touched large deviations results when derived Hoeffding's inequality. To see this, let us define $Y_n = S_n - n\mu$ where $X_1, \ldots, X_n$ are i.i.d. random variables. Notice that Y_n is martingale with respect to $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ since

$$\mathbf{E}[Y_{n+1}|\mathcal{F}_n] = \mathbf{E}[S_n + X_{n+1} - (n+1)\mu|\mathcal{F}_n] = S_n - (n+1)\mu - \mu = Y_n.$$

Defining the martingale difference $D_n = Y_n - Y_{n-1}$ and recognizing that $\mathbf{E}[Y_n|\mathcal{F}_n] = Y_n$ as well as $\mathbf{E}[Y_n|\mathcal{F}_0] = \mathbf{E}[Y_n]$, we obtain from (5.22) (several generalization and extensions of this results are discussed in the exercise section).

Theorem 5.16 (Hoeffding) *Let $X_1, \ldots, X_n$ be i.i.d. random variables such that $|X_i| \leq 1$ for every i. Then*

$$\Pr\{S_n \geq n(\mu + \delta)\} \leq \exp\left(-\frac{1}{2}n\delta^2\right) \quad (5.33)$$

for $\delta > 0$.

In passing we observe that directly from Theorem 5.13 we can obtain a weaker bound, namely, $\Pr\{S_n \geq n(\mu + \delta)\} \leq \exp\left(-\frac{1}{8}n\delta^2\right)$.

Let us now explore more the large deviation behavior of $S_n = X_1 + \cdots X_n$ for i.i.d. (not necessary bounded) random variables. We start with evaluating $\Pr\{S_n \geq an\}$ for $a > \mu$ as $n \to \infty$. Observe that

$$\Pr\{S_{n+m} \geq (n+m)a\} \geq \Pr\{S_m \geq ma,\ S_{n+m} - S_m \geq na\} = \Pr\{S_n \geq na\}\Pr\{S_m \geq ma\}$$

since S_m and $S_{n+m} - S_m$ are independent. Taking logarithm of both sides, and recognizing that $\log \Pr\{S_n \geq an\}$ is a superadditive sequence, by Fekete's Theorem 5.1 we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{S_n \geq na\} = -I(a)$$

where $I(a) \geq 0$. Thus S_n may decay *exponentially* (provided $I(a) > 0$) when far away from its mean, as we have already seen it for bounded random variables. Unfortunately, we obtain the above result from the subadditive property that allows us to conclude the existence of the above limit, but says nothing about $I(a)$. In particular, we must discover when $I(a) = 0$, that is, the decay of S_n is *subexponential.* We discuss it next.

We want to know when S_n has subexponential or *heavy* tail which will lead to $I(a) = 0$. Intuitively, when the tail of the distribution of a random variable X is not exponential, then the so called *moment generating function* defined as

$$M(\lambda) = \mathbf{E}[e^{\lambda X}]$$

cannot exist for $\lambda > 0$. (We shall discuss in depth generating functions in Chapter 7 for complex λ, but here we assume that λ is real and positive.) To make this claim precise, let us consider a sequence $X_1, \ldots, X_n$ of i.i.d. random variables such that for all $\lambda > 0$

$$M(\lambda) = \mathbf{E}[e^{\lambda X_1}] = \infty. \tag{5.34}$$

Then we claim

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathsf{Pr}\{S_n \geq an\} = 0 \tag{5.35}$$

where, as always, $S_n = X_1 + \cdots + X_n$. It suffices to prove that

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{Pr}\{t \leq S_n \leq t+1\} = 0$$

and set $t = an$. We prove now that (5.34) implies the above. Let us assume contrary and postulate the existence of $\delta > 0$ such that $\mathsf{Pr}\{t-1 \leq S_n \leq t\} \leq e^{-t\delta}$. Then

$$\begin{aligned}
\mathbf{E}[e^{\lambda S_n}] &= \int_{-\infty}^{\infty} e^{\lambda s} dP(s) \leq 1 + \int_0^{\infty} e^{\lambda s} dP(s) \\
&= 1 + \sum_{t=0}^{\infty} \int_t^{t+1} e^{\lambda s} dP(s) \\
&\leq 1 + \sum_{t=0}^{\infty} e^{\lambda t} \mathsf{Pr}\{t \leq S_n \leq t+1\} \\
&< \infty \qquad \text{for} \qquad \lambda < \delta.
\end{aligned}$$

Thus the moment generating functions of S_n and X_1 exist at least for $\lambda < \delta$. This contradicts our basic assumption (5.34) and proves (5.35).

Hereafter, we adopt the hypothesis that the moment generating function of X_1 exists, that is, we assume that

$$M(\lambda) = \mathbf{E}[e^{\lambda X_1}] < \infty \;\; \text{for some } \lambda > 0. \tag{5.36}$$

In addition, we define the *cumulant function* of X_1 as

$$\kappa(\lambda) = \log M(\lambda).$$

Certainly, $\kappa'(0) = \mathbf{E}[X_1] = \mu$ and $\kappa''(0) = \mathbf{Var}[X_1]$. In Exercise 12 we ask the reader to prove that the cumulant function $\kappa(\lambda)$ is convex.

For a moment, we now drop the assumption about independence of $X_1, \ldots, X_n$. Then by Markov's inequality

$$e^{\lambda na}\mathsf{Pr}\{S_n \geq na\} = e^{\lambda na}\mathsf{Pr}\{e^{\lambda S_n} \geq e^{\lambda na}\} \leq \mathbf{E}[e^{\lambda S_n}]$$

for any $a > 0$. Actually, due to arbitrariness of λ, subject to $\lambda > 0$, we just derived the so called *Chernoff's bound.*

Lemma 5.17 (Chernoff's Bound) *Let $X_1, \ldots, X_n$ be random variables for which the moment generating functions $M_i(\lambda)$ exist for some $\lambda > 0$. Then*

$$\mathsf{Pr}\{S_n \geq na\} \leq \min_{\lambda>0} \left\{ e^{-\lambda na}\mathbf{E}[e^{\lambda S_n}] \right\} . \tag{5.37}$$

for any $a > 0$.

Example 5.6 *Bounds on the Binomial Distribution*

Let X_i be Bernoulli distributed with $\mathsf{Pr}\{X_1 = 1\} = p$ and $q = 1 - p$. Hence, $S_n = X_1 + \cdots + X_n$ has the *binomial*(n, p) distribution. By Chernoff's bound we have

$$\mathsf{Pr}\{S_n \geq na\} \leq \exp\left(-na\lambda + n\ln(pe^{\lambda} + q)\right).$$

Set now $a = p(1+\varepsilon)$. Then for large n

$$\begin{aligned}\mathsf{Pr}\{S_n \geq na\} &\leq \exp\left(-np\varepsilon\lambda - np\lambda + np\lambda + np\lambda^2/2 + O(n\lambda^3)\right)\\ &\leq \exp\left(-np\varepsilon\lambda + np(1+\delta)\lambda^2/2\right),\end{aligned}$$

for some $\delta > 0$. Setting $\delta = 1/2$ and $\lambda = 2\varepsilon/3$, we finally obtain

$$\mathsf{Pr}\{S_n \geq np(1+\varepsilon)\} \leq \exp(-np\varepsilon^2/3) \tag{5.38}$$

which is a useful inequality that we shall use throughout the book. □

In the i.i.d. case the bound (5.37) becomes

$$\mathsf{Pr}\{S_n \geq na\} \leq \min_{\lambda>0} \{\exp(-n(a\lambda - \kappa(\lambda))\}$$

since

$$\mathbf{E}[e^{\lambda(X_1+\cdots+X_n)}] = \mathbf{E}[\prod_{i=1}^{n} e^{\lambda X_i}] = \prod_{i=1}^{n} \mathbf{E}[e^{\lambda X_i}] = e^{n\kappa(\lambda)}$$

where the second equality follows from independence. We shall write the above in a different manner by introducing the so called Fenchel–Legendre transform

$$I(a) = \sup_{\lambda>0}\{a\lambda - \kappa(\lambda)\}.$$

Then

$$\Pr\{S_n \geq na\} \leq e^{-nI(a)}$$

and we shall prove below a matching lower bound.

We first need to study some properties of the Fenchel–Legendre transform. Let now $L(\lambda) = a\lambda - \kappa(\lambda)$ and $I(a) = \sup_{\lambda>0} L(\lambda)$. To find an explicit formula for $I(a)$ we must solve a simple optimization problem, that is, to find such λ_a, if exists, that $L(\lambda_a) = I(a)$. We shall look at λ_a among solutions of $L'(\lambda_a) = 0$, that is, such that

$$a = \kappa'(\lambda_a) = \frac{M'(\lambda_a)}{M(\lambda_a)}. \tag{5.39}$$

To study the existence of λ_a and to prove a lower bound for the large deviations result, we introduce the so called *twisted distribution* known also as *shift of mean method* (cf. Section 8.4.3) and *exponential change of measure*. Let X^θ be a random variable with the distribution

$$F_\theta(x) = \frac{1}{M(\theta)} \int_{-\infty}^{x} e^{\theta y} dF(y).$$

It is easy to check that $F_\theta(x)$ is a distribution function, and

$$\mathbf{E}[X^\theta] = \kappa'(\theta) = \frac{M'(\theta)}{M(\theta)}.$$

The exponential change of measure plays an important role in probability since the mean of X^θ can be shifted to any position by selecting proper θ (e.g., by choosing $\theta = \lambda_a$ the mean of X^θ becomes a). Then, we may apply several useful probability laws that are true only around the mean (e.g., weak law of numbers and central limit theorem) Moreover, since

$$\kappa''(\theta) = \mathbf{Var}[X^\theta] \geq 0,$$

$\kappa'(\theta) = \mathbf{E}[X^\theta]$ is strictly increasing as long as *the distribution F is not a point mass at μ.* In view of this, and the fact $\kappa'(0) = \mu$, we conclude that for $a > \mu$ equation (5.39) has always a solution $\lambda_a \geq 0$ such that it maximizes the concave function $L(\lambda)$.

We now know that λ_a maximizes $L(\lambda)$, however, to use this fact for the large deviations, we must relate it to the assumption (5.36) since it might happen that $M(\lambda)$ is defined only for $\lambda < \lambda_a$. To handle this properly, we introduce

$$\begin{aligned} x_0 &= \sup\{x : F(x) < 1\}, \\ \lambda_{\max} &= \sup\{\lambda : M(\lambda) < \infty\}. \end{aligned}$$

We are ready now to formulate the basic large deviations result.

Theorem 5.18 (Large Deviations I) *Assume $X_1, \ldots, X_n$ are i.i.d. Let $M(\lambda) = \mathbf{E}[e^{\lambda X_1}] < \infty$ for some $0 < \lambda < \lambda_{\max}$, and the distribution of X_i is not a point mass at μ. Set $\mu < a < x_0$ and assume there is $\lambda_a \in (0, \lambda_{\max})$ such that*

$$a = \frac{M'(\lambda_a)}{M(\lambda_a)}.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{S_n \geq na\} = -(a\lambda_a - \log M(\lambda_a)) = -I(a).$$

Proof. The upper bound follows immediately from the Chernoff bound and the discussion above. Following Durrett [69], for the lower bound we introduce the twisted distribution F_λ such that $\lambda \in (\lambda_a, \lambda_{\max})$, and consider $X_1^\lambda, \ldots, X_n^\lambda$ with distribution F_λ. Observe that $S_n^\lambda = X_1^\lambda + \cdots + X_n^\lambda$ has mean $\mathbf{E}[S_n^\lambda] = n\kappa'(\lambda) > na$ where the inequality follows from the monotonicity of $\kappa'(\lambda)$ and $\lambda > \lambda_a$. It is also easy to compute the distribution F_n^λ of S_n^λ in terms of the distribution F_n of S_n. Indeed, since $M^n(\lambda)$ is the moment generating function of S_n, we immediately realize that

$$\frac{dF_n(x)}{dF_n^\lambda(x)} = e^{-\lambda x} M^n(\lambda),$$

where we wrote the twisted distributed in the derivative form (i.e., Radon-Nikodym derivative) rather than the integral form. Now, we obtain

$$\begin{aligned} \Pr\{S_n > an\} &= \int_{na}^{\infty} dF_n(x) = \int_{na}^{\infty} e^{-\lambda x} M^n(\lambda) dF_n^\lambda(x) \\ &\geq M^n(\lambda) e^{-\lambda na}(1 - \Pr\{S_n^\lambda \leq na\}) \\ &\to M^n(\lambda) e^{-\lambda na} \end{aligned}$$

where the last line follows from the weak law of large numbers, since as shown above, $\mathbf{E}[S_n^\lambda] > na$. Thus,

$$\liminf_{n\to\infty} \frac{1}{n} \log \Pr\{S_n > na\} \geq -(a\lambda - \log M(\lambda)),$$

and the theorem follows from arbitrariness of $\lambda > \lambda_a$. ∎

In the theorem above, we assumed that $\lambda_a < \lambda_{\max}$, however, there is an abundance of examples (see below) when this condition is violated. In this case, we should replace the above theorem by the result below which we ask the reader to prove in Exercise 14.

Theorem 5.19 (Large Deviations II) *Suppose that*

$$\lim_{\lambda \to \lambda_{\max}} \kappa'(\lambda) = x_1 \leq a.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{S_n \geq na\} = -(a\lambda_{\max} - \log M(\lambda_{\max}))$$

for $a \geq x_1$.

Example 5.7 *Large Deviations of Some Known Distributions*

We apply here the large deviation results to some distributions.

1. Standard Normal Distribution: In this case it is easy to check that

$$M(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda y} e^{-\frac{1}{2}y^2} dy = e^{\frac{1}{2}\lambda^2}$$

by completing the square in the exponent. Then, $\kappa(\lambda) = \frac{1}{2}\lambda^2$, $\kappa'(\lambda) = \lambda$, and hence $\lambda_a = a$ which implies $I(a) = \frac{1}{2}a^2$ so that for $a > 0$

$$\log \Pr\{S_n \geq an\} \sim -\frac{n}{2}a^2$$

which can be verified with "bare hands".

2. Poisson Distribution: Consider the Poisson distribution with mean θ. Then easy calculations show

$$M(\lambda) = \exp\left(\theta(e^{\lambda} - 1)\right),$$

and $\lambda_a = \log(a/\theta)$ so that

$$I(a) = a\left(\log(a/\theta) - 1\right) + \theta.$$

Theorem 5.18 implies for $a > \theta$

$$\Pr\{S_n \geq an\} = \left(\frac{a}{\theta}\right)^{-na} e^{-n(\theta - a) + o(n)}.$$

3. Perverted Exponential Distribution: Define the density f as follows: it is zero for $x < 1$ and for $x \geq 1$ we set $f(x) = Cx^{-3}e^{-x}$ with C chosen so that f is a probability density (cf. [69]). Observe that $M(\lambda)$ can exist only for $\lambda \leq 1$ and

$$\kappa'(\lambda) \leq \kappa'(1) = 2.$$

Thus, $\lambda_{\max} = 1$ and $\kappa'(\lambda) = a$ does not have a solution for $a > 2$. In other words, as $\lambda \to \lambda_{\max}$ we only have $\kappa'(\lambda) \to 2 = x_1$. By Theorem 5.19 for $a > 2$ we only can write

$$n^{-1} \log \Pr\{S_n \geq an\} \to -a$$

in this case. □

A major strengthening of the large deviation theorem just discussed is due to Gärtner and Ellis [67] who extended it to weakly dependent random variables. Let us consider S_n as a sequence of random variables (e.g., $S_n = X_1 + \ldots + X_n$), and let $M_n(\lambda) = \mathbf{E}[e^{\lambda S_n}]$. The following is quite useful. Its proof can be found in Dembo and Zeitouni [56].

Theorem 5.20 (Gärtner-Ellis 1984) *For dependent random variable $X_1, \ldots, X_n$, let*

$$\lim_{n\to\infty} \frac{\log M_n(\lambda)}{n} = c(\lambda)$$

exist and is finite in a subinterval of the real axis. If there exists λ_a such that $c'(\lambda_a)$ is finite and $c'(\lambda_a) = a$, then

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{S_n \geq na\} = -(a\lambda_a - c(\lambda_a)) \ .$$

We return to large deviations of weakly dependent random variables in Chapter 8 where we analyze the so called *quasi-power* case for which $M_n(s) \sim a(s)c^n(s)$ where $a(s)$ and $c(s)$ are functions of complex s.

5.5 Applications

This is a long chapter and we need a few good applications of the subadditive ergodic theorem, Azuma's inequality, martingales, Chernoff's bound and large deviations to illustrate these important techniques. We hope this section presents some interesting problems. We shall continue using these techniques throughout the book.

Hereafter, we discuss three applications. First, we deal with the edit distance problem already described in Section 1.5 (cf. [177]). Then, we turn our attention to the Knuth–Morris–Pratt pattern matching algorithm and prove that its complexity grows linearly with the size of the text (cf. [205]). For this application, unlike other examples discussed so far, the hardest part is to establish the subadditive property. Finally, we investigate an interesting problem on a random sequence that arose in the analysis of hashing with lazy deletions (cf. [4]).

5.5.1 Edit Distance

We consider the edit distance problem presented in Section 1.5 and in Example 1. We want to estimate the minimum cost $C_{\min}$ or the maximum cost $C_{\max}$ of transforming one sequence into another. As mentioned before, this problem can be reduced to finding the longest (shortest) path in a special grid graph (cf. Figure 1.8). Let us assume that the weights of insertion W_I, deletion W_D and substitution W_Q are independently distributed. We first show that for some constant $\alpha > 0$

$$\lim_{n\to\infty} \frac{C_{\max}}{n} = \lim_{n\to\infty} \frac{\mathbf{E}[C_{\max}]}{n} = \alpha \quad \text{(a.s.)} ,$$

provided ℓ/s has a limit as $n \to \infty$. Indeed, let us consider the $\ell \times s$ grid of Figure 1.8 with the starting point O and the ending point E. Call it Grid(O,E). We also choose an arbitrary point, say A, inside the grid so that we can consider two grids, namely Grid(O,A) and Grid(A,E). Actually, point A splits the edit distance problem into two subproblems with objective functions $C_{\max}(O,A)$ and $C_{\max}(A,E)$. Clearly, $C_{\max}(O,E) \geq C_{\max}(O,A) + C_{\max}(A,E)$. Thus, under our assumption regarding weights, the objective function $C_{\max}$ is superadditive, and a direct application of Theorem 5.8 proves the result.

As expected, the above calculation does not indicate how to compute α. From Example 1, we know that even in the simplest case of the Longest Common Superstring problem, the constant α is unknown. But we can prove that $C_{\max}$ is concentrated around αn.

Theorem 5.21 (Louchard and Szpankowski 1995) (i) *If all weights are bounded random variables, say* $\max\{W_I, W_D, W_Q\} \leq 1$, *then for arbitrary* $\varepsilon > 0$ *and large* n *the random variable*

$$\Pr\{|C_{\max} - \mathbf{E}[C_{\max}]| > \varepsilon \mathbf{E}[C_{\max}]\} \leq 2\exp(-\frac{1}{2}\varepsilon^2 \alpha n). \tag{5.40}$$

(ii) *If the weights are unbounded but such that for large* n, $W_{max} = \max\{W_I, W_D, W_Q\}$ *satisfies the following*

$$n\Pr\{W_{max} \geq n^{1/2-\delta}\} \leq U(n) \tag{5.41}$$

for some $\delta > 0$ *and a function* $U(n) \to 0$ *as* $n \to \infty$, *then*

$$\Pr\{|C_{\max} - EC_{\max}| > \varepsilon EC_{\max}\} \leq 2\exp(-\beta n^{\delta}) + U(n) \tag{5.42}$$

for any $\varepsilon > 0$ *and some* $\beta > 0$.

Proof: Part (i) is a direct consequence of Azuma's inequality (cf. Theorem 5.15). It suffices to assume in (5.28) that X_i is equal to the ith symbol of the first string X for $1 \leq i \leq \ell$, and for $\ell + 1 \leq i \leq n = s + \ell$, X_i is equal to the $(i - \ell)$th symbol of the second string Y. Under the Bernoulli model, the X_i are i.i.d. and hence the Azuma condition (5.28) holds, with $F_n(\cdot) = C_{\max}$. More precisely,

$$|C_{\max}(X_1, \ldots, X_i, \ldots, X_n) - C_{\max}(X_1, \ldots, \hat{X}_i, \ldots, X_n)| \leq \max_{1\leq i\leq n} \{W_{max}(i)\} \tag{5.43}$$

where $W_{max}(i)$ is the ith independent version of W_{max} defined in the theorem. Clearly, we can apply (5.29) of Theorem 5.15 with $c_i = 1$. Inequality (5.40) follows from the above and $t = \varepsilon \mathbf{E}[C_{\max}] = O(n)$.

To prove part (ii), we proceed as follows for some c

$$\begin{aligned}
\Pr\{|C_{\max} - \mathbf{E}[C_{\max}]| \geq t\} &= \Pr\{|C_{\max} - \mathbf{E}[C_{\max}]| \geq t\ ,\ \max_{1\leq i\leq n}\{W_{max}(i)\} \leq c\} \\
&+ \Pr\{|C_{\max} - \mathbf{E}[C_{\max}]| \geq t\ ,\ \max_{1\leq i\leq n}\{W_{max}(i)\} > c\} \\
&\leq 2\exp(-\frac{1}{2}t^2/nc^2) + n\Pr\{W_{max} > c\}\ .
\end{aligned}$$

Set now $t = \varepsilon \mathbf{E}[C_{\max}] = O(n)$ and $c = O(n^{1/2-\delta})$, then

$$\Pr\{|C_{\max} - \mathbf{E}[C_{\max}]| \geq \varepsilon E C_{\max}\} \leq 2\exp(-\beta n^{\delta}) + n\Pr\{W_{max} > n^{1/2-\delta}\}\ ,$$

for some constant $\beta > 0$, and this implies (5.42) provided (5.41) holds. ■

5.5.2 Knuth–Morris–Pratt Pattern Matching Algorithms

Here, we shall analyze the number of comparisons C_n performed by a Knuth–Morris–Pratt like algorithm (described in details in Section 1.3) when it searches for occurrences of a *given* pattern H of length m in a *random* text T of size n. In the derivation we shall use the subadditive ergodic Theorem 5.8. Unlike all other examples in this section, the main difficulty here lies in establishing the subadditivity property for C_n. We follow here the presentation of Régnier and Szpankowski [205] but we also discuss some new results.

We recall that Knuth–Morris–Pratt algorithms (in short: KMP) scan the text from left to right. The main idea can be described as follows: When searching for pattern occurrences the algorithm finds a mismatch at, say, position l of the text and position k of the pattern. Then the next comparison is made at position l of the text and position i of the pattern such that i is the largest integer satisfying

$$T_{l-i+1}^{l-1} = H_1^{i-1},$$

that is, the largest suffix of the already searched pattern that is a prefix of the pattern. The pattern is shifted by $k - i$ positions to be aligned at $l - i + 1$ position of the text (cf. Section 1.3). We shall write $M(l, k) = 1$ if the the algorithm makes a comparison of position l of the text with position k of the pattern; otherwise $M(l, k) = 0$.

Our goal is to prove the following result.

Theorem 5.22 (Regnier and Szpankowski 1995) *Consider the Knuth-Morris-Pratt algorithm for pattern matching. Let H be a given pattern of length m.*

(i) *If the text* T_1^n *is given, then*

$$\lim_{n\to\infty} \frac{\max_T C_n}{n} = \alpha_1(H) \tag{5.44}$$

where $\alpha_1(H) \geq 1$ *is a constant.*

(ii) *If the text is generated by a stationary and ergodic source, then*

$$\begin{aligned}\lim_{n\to\infty} \frac{C_n}{n} &= \alpha_2(H) \qquad \text{a.s.}\\ \lim_{n\to\infty} \frac{\mathbf{E}[C_n]}{n} &= \alpha_2(H)\end{aligned}$$

where $\alpha_2(H) \geq 1$ *is a constant.*

We prove this theorem by showing that C_n is subadditive. For this, we denote by $C_{r,m}$ the number of comparisons performed on T_r^m ($r \leq m$). Actually, we prove a stronger result, namely, that the subadditivity holds for the so called ℓ-convergent pattern matching algorithms that are defined below together with some other useful notations.

Definition 5.23 (i) *Given a string searching algorithm, a text* T *and a pattern* H*, a position* AP *in the text* T *satisfying for some* k $(1 \leq k \leq m)$

$$M[AP + (k-1), k] = 1$$

is said to be an **alignment position** *(i.e., the first position of the pattern is aligned with position* AP *of the text).*

(ii) *For a given pattern* H*, a position* i *in the text* T_1^n *is an* **unavoidable alignment position** *for an algorithm if for any* r, l *such that* $r \leq i$ *and* $l \geq i + m$*, the position* i *is an alignment position when the algorithm is run on* T_r^l.

(iii) *An algorithm is said to be* ℓ**-convergent** *if, for any text* T *and pattern* H*, there exists an increasing sequence* $\{U_i\}_{i=1}^n$ *of unavoidable alignment positions satisfying* $U_{i+1} - U_i \leq \ell$ *where* $U_0 = 0$ *and* $n - \max_i U_i \leq \ell$.

Our main effort will concentrate on proving the following lemma.

Lemma 5.24 *An* ℓ*-convergent pattern matching algorithm satisfies the following inequality for all* r *such that* $1 \leq r \leq n$*:*

$$|C_{1,n} - (C_{1,r} + C_{r,n})| \leq m^2 + \ell m \ , \tag{5.45}$$

provided any comparison is done only once.

Proof. Let U_r be the smallest unavoidable position greater than r. We evaluate in turn $C_{1,n} - (C_{1,r} + C_{U_r,n})$ and $C_{r,n} - C_{U_r,n}$ (cf. Figure 5.1) . We start our analysis by considering $C_{1,n} - (C_{1,r} + C_{U_r,n})$. This part involves the following contributions:

- Those comparisons that are performed after position r but with alignment positions before r. We call this contribution S_1. Observe that those comparisons contribute to $C_{1,n}$ but not to $C_{1,r}$. To avoid counting the last character r twice, we must subtract one comparison. Thus

$$S_1 = \sum_{AP<r} \sum_{i \geq r} M(i, i - AP + 1) - 1.$$

- The next contribution called S_2 accounts for alignments AP satisfying $r \leq AP \leq U_r$ that only contribute to $C_{1,n}$, that is,

$$S_2 = \sum_{AP=r}^{U_r - 1} \sum_{i \leq m} M(AP + (i-1), i) \ .$$

- Finally, since the alignment positions after U_r on the text $T_{U_r}^n$ and T_1^n are the same, the only difference in contribution may come from the amount of information saved from previous comparisons done on T_1^r. This is clearly bound by

$$|C_{1,n} - (C_{1,r} + C_{U_r,n} + S_1 + S_2)| \leq m \ .$$

Now, we evaluate $C_{r,n} - C_{U_r,n}$ (see second part of Figure 5.1). We assume that the algorithm runs on T_r^n and let AP be any alignment position satisfying $r \leq AP < U_r$. The following contributions must be considered:

- The contribution S_3

$$S_3 = \sum_{AP=r}^{U_r - 1} \sum_{i} M(AP + (i-1), i)$$

  counts for the number of comparisons associated positions $r \leq AP < U_r$. This sum is the same as S_2 but the associated alignment positions and searched text positions $AP + k - 1$ may be different.

- Additional contribution may come from the alignment at position U_r. But, no more than m comparisons can be saved from previous comparisons, hence

$$|C_{r,n} - C_{U_r,n} - S_3| \leq m.$$

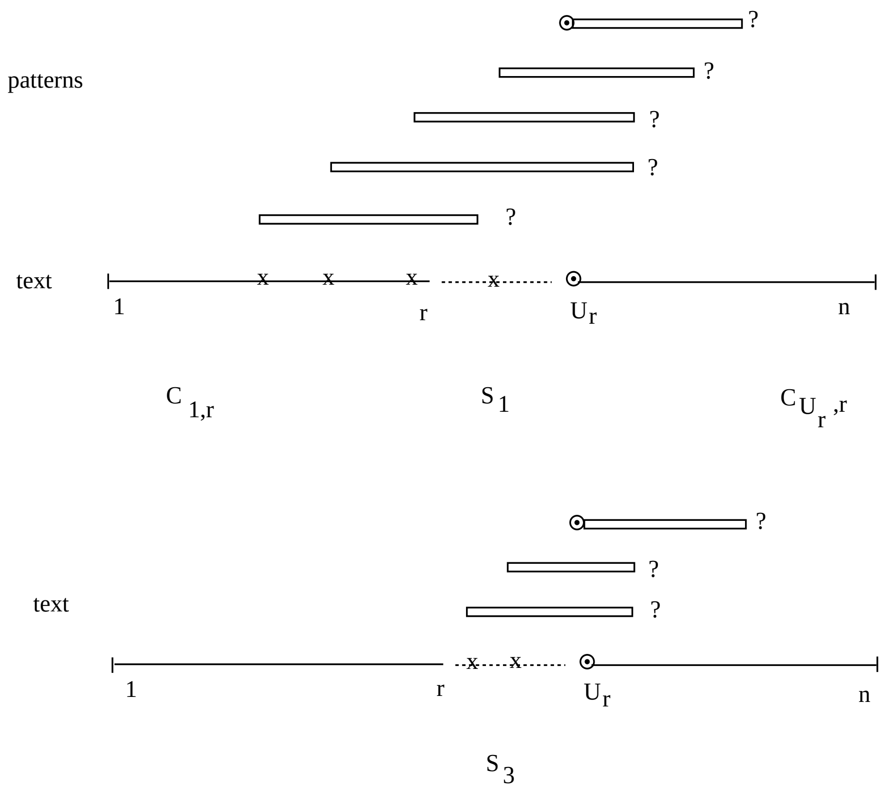

Figure 5.1: Illustration to the proof of Lemma 5.24

To complete the proof, it remains to find upper bounds on S_1, S_2, and S_3. For $\ell \geq U_r - r$ we easily see that S_2 and S_3 are smaller than ℓm. So are their difference. With respect to S_1, for a given alignment position AP, we have $|i - AP| \leq m$, hence $|r - AP| \leq m$, and for any AP the index i has at most m different values. Thus $S_1 \leq m^2$, and this proves the lemma. ■

To prove Theorem 5.22, it is enough to show that KMP is $\ell = m$-convergent algorithm. Before we establish this fact, we prove an interesting property of KMP-like algorithms.

Lemma 5.25 *Given a pattern H and a text T, KMP-like algorithms have the same set of unavoidable alignment positions $\mathcal{U} = \bigcup_{l=1}^{n}\{U_l\}$, where*

$$U_l = \min\{\min_{1\leq k\leq l}\{T_k^l \preceq H\},\ l+1\}$$

and $T_k^l \preceq H$ means that the substring T_k^l is a prefix of the pattern H.

Proof. Let l be a text position such that $1 \leq l \leq n$, and r be any text position satisfying $r \leq U_l$. Let $\{A_i\}$ be the set of alignment positions defined by a KMP–like algorithm that

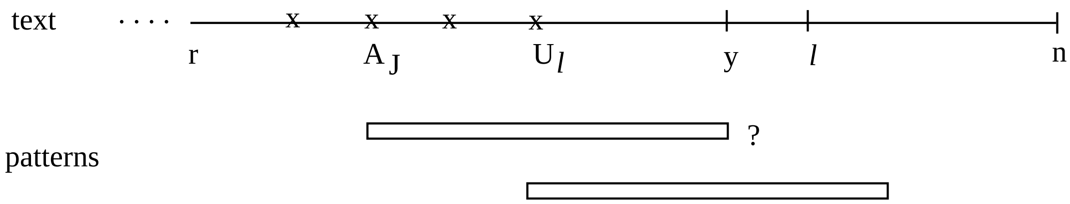

Figure 5.2: Illustration to the proof of Lemma 5.25.

runs on T_r^n. As it contains r, we may define (cf. Figure 5.2).

$$A_J = \max\{A_i :\ A_i < U_l\}$$

Hence, we have $A_{J+1} \geq U_l$. Using an adversary argument, we shall prove that $A_{J+1} > U_l$ cannot be true, thus showing that $A_{J+1} = U_l$. Let $y = \max\{k : M(A_J + (k-1), k) = 1\}$, that is, y is the rightest position we can do a comparison starting from A_J We observe that we have $y \leq l$. Otherwise, according to the KMP algorithm rule, we would have $T_{A_J}^l \preceq H$, which contradicts the definition of U_l. Also, since KMP–like algorithms are sequential (i.e., the text–pattern comparisons define non-decreasing sequences), then $A_{J+1} \leq y + 1 \leq l + 1$. Hence $U_l = l + 1$ contradicts the assumption $A_{J+1} > U_l$ and we may assume $U_l \leq l$. In that case, $H_{U_l}^l \preceq H$ and an occurrence of H at position U_l is consistent with the available information. Let the adversary assumes that H does occur. As sequence $\{A_i\}$ is non-decreasing and A_{J+1} has been chosen greater than U_l so that this occurrence will not be detected by the algorithm. This leads to the desired contradiction. Thus $A_{J+1} = U_l$, as needed. ■

To complete the proof of Theorem 5.22, it is necessary to show that KMP is m-convergent. But this is easy. Let AP be an alignment position and define $l = AP + m$. As $|H| = m$, one has $l - (m-1) \leq U_l \leq l$. Hence, $U_l - AP \leq m$ which establishes the m-convergence of the KMP algorithm, and the main Theorem 5.22.

As in other applications, since we obtained the result through the subadditive ergodic theorem, the constants α_1 and α_2 are unknown. In [43] some bounds for α_1 were found. But, even if we cannot find the value of α_2, we still are able to show that C_n is concentrated around $\alpha_2 n$. As expected, we shall use Azuma's inequality. Let us now assume that T is generated by a memoryless source. It suffices to observe that C_n, as a function of random text $T = T_1, \ldots, T_n$, satisfies

$$|C_n(T_1, \ldots, T_i, \ldots, T_n) - C_n(T_1, \ldots, \hat{T}_i, \ldots, T_n)| \leq 2m^2$$

where $\hat{T}_i$ is an independent copy of the ith symbol T_i of the text.. In view of this and Theorem 5.15 we prove the following.

Theorem 5.26 *Let T be generated by a memoryless source while the pattern H of length m is given. The number of comparisons C_n made by the Knuth–Morris–Pratt algorithm is* *concentrated around its mean* $\mathbf{E}[C_n] = \alpha_2 n(1+o(1))$. *More precisely,*

$$\Pr\{|C_n - \alpha_2 n| \ge \varepsilon n\} \le 2\exp\left(-\frac{1}{4}\varepsilon^2 \frac{n}{m^4}(1+o(1))\right)$$

for any $\varepsilon > 0$.

5.5.3 Large Deviations of a Random Sequence

We consider here a large deviation result proved in Aldous *et al.* [4] who studied hashing with lazy deletion. We reformulate the problem as a coding problem.

Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of independent random variables with uniform distribution over $\mathcal{A} = \{1,\ldots,V\}$. Let also $f : \{1,2,3,\ldots\} \to \{1,2,3,\ldots\}$ be a $1-1$ function such that $f(i) > i$ for all $i \ge 1$. Our goal is to estimate the probability that

$$Y_{f(i)} \quad \text{is different than all of} \quad Y_1,\ldots,Y_i.$$

Let A_i be the event that the above holds, and let $B = \sum_{i=1}^{\infty} I(A_i)$. We want to estimate $\Pr\{B \ge b\}$ for large b.

A direct application of Chernoff's bound leads to

$$\Pr\{B \ge b\} \le \min_{z>1}\left\{z^{-b}\mathbf{E}[z^B]\right\}$$

where we set $z = e^{\lambda}$ in Theorem 5.17. The problem is that we do not know how to assess the generating function of B since $I(A_i)$ are dependent. To overcome this difficulty, we adopt a different approach based on martingales. Let the events A_i be adapted to the increasing σ-fields $\mathcal{F}_i$ (i.e., $\mathcal{F}_i$ contains all the information about $A_1,\ldots,A_i$). We first prove that for some $w \ge 1$

$$M_n = \frac{w^{\sum_{i=1}^n I(A_i)}}{\prod_{i=1}^n \mathbf{E}[w^{I(A_i)}|\mathcal{F}_{i-1}]}$$

is a positive martingale. Indeed,

$$\begin{aligned}
\mathbf{E}[M_{n+1}|\mathcal{F}_n] &= \frac{\mathbf{E}[w^{\sum_{i=1}^n I(A_i)} w^{I(A_{n+1})}|\mathcal{F}_n]}{\prod_{i=1}^n \mathbf{E}[w^{I(A_i)}|\mathcal{F}_{i-1}]\mathbf{E}[w^{I(A_{n+1})}|\mathcal{F}_n]} \\
&= \frac{w^{\sum_{i=1}^n I(A_i)}}{\prod_{i=1}^n \mathbf{E}[w^{I(A_i)}|\mathcal{F}_{i-1}]} \cdot \frac{\mathbf{E}[w^{I(A_{n+1})}|\mathcal{F}_n]}{\mathbf{E}[w^{I(A_{n+1})}|\mathcal{F}_n]} \\
&= M_n.
\end{aligned}$$

Observe that the above martingale is the "multiplicative" analogue of the additive martingale we have been working so far. But by *Martingale Convergence Theorem* (cf. Theorem 5.12) M_n converges almost surely and in mean to the random variable

$$M_\infty = \frac{w^B}{D_w}$$

where

$$D_w = \prod_{i\geq 1} \mathbf{E}[w^{I(A_i)}|\mathcal{F}_{i-1}] = \prod_{i\geq 1}(1+(w-1)\mathsf{Pr}\{A_i|\mathcal{F}_{i-1}\}).$$

In addition, we know that $\mathbf{E}[M_\infty] \leq \mathbf{E}[M_0] = 1$, hence

$$\mathbf{E}[w^B/D_w] \leq 1. \tag{5.46}$$

Now, we proceed as follows. Instead of using Chernoff's bound directly, we write for some $a > 0$

$$\begin{aligned}
\mathsf{Pr}\{B \geq b\} &\leq \mathsf{Pr}\{D_w > a\} + \mathsf{Pr}\{B \geq b,\ D_w \leq a\} \\
&\leq \mathbf{E}[D_w/a] + \mathsf{Pr}\{w^B/D_w \geq w^b/a\} \\
&\leq \mathbf{E}[D_w/a] + a/w^b \\
&\leq 2\sqrt{w^{-b}\mathbf{E}[D_w]}, \qquad a = \sqrt{w^b\mathbf{E}[D_w]}
\end{aligned}$$

where the second inequality follows from Markov's inequality, the third one is a consequence of (5.46), and in the last one we set $a = \sqrt{w^b\mathbf{E}[D_w]}$. Since $w \geq 1$ is arbitrary, hence we finally arrive at

$$\mathsf{Pr}\{B \geq b\} \leq 2\sqrt{\inf_{w>1} w^{-b}\mathbf{E}[D_w]}. \tag{5.47}$$

We evaluate now the required infimum in (5.47). First of all, observe that

$$\mathsf{Pr}\{A_i|\mathcal{F}_{i-1}\} = L_i/V$$

where L_i is the number of values *not* taken by $Y_1, \ldots, Y_i$. Thus, for $w > 1$ the quantity D_w is

$$D_w = \prod_{i\geq 0}(1+(w-1)L_i/V) \leq \exp\left(\frac{w-1}{V}\sum_{i\geq 0} L_i\right). \tag{5.48}$$

Now we can re-write

$$\sum_{i\geq 0} L_i = \sum_{k=1}^{V} kW_k, \tag{5.49}$$

where W_k is the waiting time for the process L_i to go from k to $k-1$ (e.g., $L_0 = V$, $L_1 = V-1$). A little thought reveals that the random variables W_k are independent with (different) geometric distributions

$$\mathsf{Pr}\{W_k = \ell\} = \left(1-\frac{k}{V}\right)^{\ell-1}\frac{k}{V}, \qquad \ell \geq 1.$$

(This is just the elementary argument for the classical coupon collector's problem with equally-likely coupons.) Hence, the associated generating function $\mathbf{E}[z^{W_k}]$ may be written as

$$\mathbf{E}[z^{W_k}] = \frac{zk/V}{1 - z(1-k/V)} = \frac{1}{1 - V/k(1-z^{-1}))} \tag{5.50}$$

for $|z| < V/(V-k)$. Combining (5.48)–(5.50) gives

$$\begin{aligned}
\mathbf{E}[D_w] &= \mathbf{E}\left[\prod_{i\geq 1}\left(1 - \frac{w-1}{V}L_i\right)\right] \\
&\leq \mathbf{E}\left[\exp\left(\frac{w-1}{V}\sum_{i\geq 1} L_i\right)\right] \leq \mathbf{E}\left[\exp\left(\frac{w-1}{V}\sum_{k=1}^{V} kW_k\right)\right] \\
&= \prod_{k=1}^{V} \mathbf{E}\left[\exp\left(\frac{w-1}{V}kW_k\right)\right] = \prod_{k=1}^{V}\left(1 - \frac{V}{k}\left(1 - e^{-k(w-1)/V}\right)\right)^{-1} \\
&\leq \prod_{k=1}^{V}(1-(w-1))^{-1} = (2-w)^{-V},
\end{aligned}$$

where the last inequality follows from $1 - y^{-1}(1-e^{-ay}) \geq 1 - a$ for $a, y > 0$. By (5.47) we obtain

$$\Pr\{B \geq b\} \leq 2\sqrt{\inf_{1<w<2} w^{-b}(2-w)^{-V}}.$$

Elementary calculus gives the exact infimum at $w = 2b/(b+V)$. Since throughout the proof we require $w > 1$, our result holds only for $b > V$.

In summary, we prove the following lemma.

Lemma 5.27 (Aldous, Hofri and Szpankowski 1992) *Let* $f : \{1,2,3,\ldots\} \to \{1,2,3,\ldots\}$ *be a* $1-1$ *function with* $f(i) > i$. *Let* $(Y_i; i \geq 1)$ *be independent, uniform on* $\{1,\ldots,V\}$. *Define* A_i *to be the event*

$$Y_{f(i)} \quad \text{is different than all of} \quad Y_1,\ldots,Y_i.$$

Let B *be the counting random variable* $B = \sum_{i\geq 1} I(A_i)$. *Then*

$$\Pr\{B \geq b\} \leq 2\left(\frac{V+b}{2b}\right)^{b/2}\left(\frac{V+b}{2V}\right)^{V/2}$$

for $b > V$.

5.6 Extensions and Exercises

5.1 Prove the following generalization of the DeBruijn and Erdős result:

Theorem 5.28 *Let x_n and c_n be two sequences of real numbers such that*

$$x_{m+n} \leq x_m + x_n + c_n$$

for all $m, n \geq 1$. Then $lim_{n\to\infty} x_n/n$ exists provided $\lim_{n\to\infty} c_n/n = 0$, however, the limit may be greater than $\inf x_m/m$.

5.2 Prove or disprove the following generalization of the Pólya and Szegő result:

Theorem 5.29 *If a real sequence x_n satisfies*

$$x_m + x_n - b(n+m) \leq x_{n+m} \leq x_m + x_n + b(n+m)$$

for all $m, n \geq 1$, where $b(n)$ is a positive sequence, then there is a finite constant ω such that

$$\left| \frac{x_n}{n} - \omega \right| \leq \frac{\beta_n}{n} \to 0$$

provided that for every $m \geq 1$

$$\beta_m = \sum_{k \geq 1} b(2^k m) 2^{-k} < \infty$$

and $\beta_m = o(m)$.

5.3 Prove the DeBruijn and Erdős Theorem 5.7. Extend all other subadditive theorems so that the subadditivity property holds only in the range $\frac{1}{2}n \leq m \leq 2n$.

5.4 Prove Hoeffding's Inequality of Theorem 5.13 using only (5.17) (cf. [225]).

5.5 Prove a refined Theorem 5.13 in which (5.19) is replaced by the optimal

$$\Pr\{|Y_n - Y_0| \geq x\} = \Pr\left\{ |\sum_{i=1}^{n} D_i| \geq x \right\} \leq 2\exp\left(-\frac{2x^2}{\sum_{i=1}^{n} c_i^2} \right).$$

5.6 *Doob–Levy's Martingale.* Let Y have a finite second moment, and let $X_1, \ldots, X_n$ be a sequence of variables. Prove that $Y_n = \mathbf{E}[Y|X_1, \ldots, X_n]$ is a martingale with respect to $X_1, \ldots, X_n$.

5.7 Construct a counter-example to the following statement

$$\mathbf{E}[\mathbf{E}[Y|\mathcal{F}_1]|\mathcal{F}_2] = \mathbf{E}[\mathbf{E}[Y|\mathcal{F}_2]|\mathcal{F}_1].$$

5.8 **Knapsack Problem**. The objective of the knapsack problem is to pack a knapsack of a finite capacity with objects of maximum wealth. More precisely, let $V_1, \ldots, V_n$ and $W_1, \ldots, W_n$ be independent non-negative random variables with finite means and $W_i \le M$ for all i and some fixed M. Assume that the knapsack has finite capacity C and your goal is to pack as many objects as you can that maximizes the wealth of the packed objects. In other words, we search for a zero–one vector $(z_1, z_2, \ldots, z_n)$ such that

$$\sum_{i=1}^{n} z_i V_i \le C,$$

and

$$Z_{\max} = \max\{\sum_{i=1}^{n} z_i W_i\}$$

is maximized. Prove that

$$\Pr\{|Z_{\max} - \mathbf{E}[Z_{\max}]| \ge x\} \le 2\exp\left(-\frac{x^2}{2nM^2}\right)$$

for $x > 0$.

5.9 Prove the following generalization of the Efron-Stein inequality due to Steele [224]: Let $F(x_1, \ldots, x_n)$ be any function of n variables, and X_i $\hat{X}_i$ for $1 \le i \le n$ be $2n$ random variables. Define $\hat{S}_i = S(X_1, \ldots, X_{i-1}, \hat{X}_i, X_{i+1}, \ldots, X_n)$. Then

$$\mathbf{Var}[S(X_1, \ldots, X_n)] \le \frac{1}{2}\mathbf{E}\left[\sum_{i=1}^{n}(S(X_1, \ldots, X_n) - \hat{S}_i)^2\right].$$

5.10 Prove that the following are equivalent: (i) $I(a) = \infty$, (ii) $\Pr\{X_1 \ge a\} = 0$, and (iii) $\Pr\{S_n \ge na\} = 0$ for all n.

5.11 Prove the following large deviation result due to Bernstein: *Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2, and suppose that $|X_i - \mathbf{E}[X_i]| \le M < \infty$. Then for all $x > 0$*

$$\Pr\{S_n \ge n\mu + x\sqrt{n}\} \le \exp\left(-\frac{x^2}{2\sigma^2 + \frac{2Mx}{3\sqrt{n}}}\right)$$

where, as usual, $S_n = X_1 + \cdots + X_n$.

5.12 Prove that the cumulant function $\kappa(\lambda) = \log \mathbf{E}[e^{\lambda X}]$, if exists, is a convex function.

Hint: Use Hölder inequality.

5.13 Prove that the Fenchel–Legendre transform $I(a)$ attains its minimum at $a = \mathbf{E}[X] = \mu$ and $I(\mu) = 0$.

5.14 Prove Theorem 5.19.

5.15 (Feller 1971) Let $F_n(x) = \Pr\{S_n \le x\}$ where $S_n = X_1 + \cdots X_n$ and $X_1, \dots X_n$ are i.i.d. Prove the following large deviations result for $x = o(\sqrt{n})$

$$\frac{1 - F_n(x)}{1 - \Phi(x)} = \left(1 + O\left(\frac{x}{\sqrt{n}}\right)\right) \exp\left(\frac{\mu_3}{6\sigma^3} \frac{x^3}{\sqrt{n}}\right)$$

where $\Phi(x)$ is the standard normal distribution function.

5.16 Consider the age-dependent branching process from Ex. 3. Let t_i be the life time for the i-th individual, and let $T_n = t_1 + \cdots + t_n$. We recall that each individual generating k offspring with probability p_k when dying. Let $\mu = \sum_{k \ge 0} k p_k$. From large deviations theory, we know that for $a < \mathbf{E}[t_i]$

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{T_n \le an\} = -I(a)$$

where $I(a) = a\lambda_a - \log \mathbf{E}[e^{\lambda_a t_1}]$ and λ_a is a solution of $a\mathbf{E}[e^{\lambda t_1}] = \frac{d}{d\lambda}\mathbf{E}[e^{\lambda t_1}]$. Prove that the constant α such that $b_n \sim \alpha n$ (a.s.) is given by

$$\alpha = \inf\{a : \ \log \mu - I(a) > 0\}.$$

Hint. Follow Biggins' footsteps by first considering a new random variable $Z_n(an)$ which denotes the number of individuals in generation n born by time an. Show:

(i) $\mathbf{E}[Z_n(an)] = \mu^n \Pr\{T_n \le an\}$;
(ii) $\Pr\{Z_n(an) > 0\} > 0$ if $\mathbf{E}[Z_n(an)] > 1$, hence $B_n \le an$ and $\alpha \le a$.

5.17 (Szpankowski 1995) Consider the general combinatorial optimization problem described in Section 1.6. To recall, we define $Z_{\max}$ and $Z_{\min}$ as

$$Z_{\max}(\mathcal{S}_n) = \max_{\alpha \in \mathcal{B}_n} \left\{ \sum_{i \in \mathcal{S}_n(\alpha)} w_i(\alpha) \right\} \quad , \quad Z_{\min}(\mathcal{S}_n) = \min_{\alpha \in \mathcal{B}_n} \left\{ \sum_{i \in \mathcal{S}_n(\alpha)} w_i(\alpha) \right\} ,$$

where $\mathcal{B}_n$ is a set of all feasible solutions, $\mathcal{S}_n(\alpha)$ is a set of objects from $\mathcal{S}_n$ belonging to the α-th feasible solution, and $w_i(\alpha)$ is the weight assigned to the i-th object in the α-th feasible solution. Let us adopt the following assumptions:

(A) The cardinality $|\mathcal{B}_n|$ of $\mathcal{B}_n$ is fixed and equal to m. The cardinality $|\mathcal{S}_n(\alpha)|$ of the set $\mathcal{S}_n(\alpha)$ does *not* depend on $\alpha \in \mathcal{B}_n$ and for all α it is equal to N, i.e., $|\mathcal{S}_n(\alpha)| = N$.

(B) For all $\alpha \in \mathcal{B}_n$ and $i \in \mathcal{S}_n(\alpha)$ the weights $w_i(\alpha)$ are identically and independently distributed (i.i.d.) random variables with common distribution function $F(\cdot)$, and the mean value μ, the variance σ^2, and the third moment μ_3 are finite.

Then prove the following result: *Under assumptions (A)-(B), as* $N, m \to \infty$ *with* $n \to \infty$

$$Z_{\min} = N\mu - o(N) \qquad \text{(pr.)} \qquad Z_{\max} = N\mu + o(N)$$

provided $\log m = o(N)$.