Techniques of Average Case Analysis of Algorithms

Wojciech Szpankowski*
Department of Computer Science, Purdue University
W. Lafayette, IN 47907, U.S.A.
spa@cs.purdue.edu

1 Introduction

An *algorithm* is a finite set of instructions for a treatment of *data* to meet some desired objectives. The most obvious reason for *analysing* algorithms and data structures (associated with them) is to discover their characteristics in order to evaluate their suitability for various applications or to compare them with other algorithms for the same application. Needless to say, we are interested in *efficient* algorithms in order to use efficiently scarce resources such as computer space and time.

Most often algorithm designs are finalized to the optimization of the asymptotic *worst case* performance, as popularized by Aho, Hopcroft and Ullman [2]. Insightful, elegant and generally useful constructions have been set up in this endeavor. Along these lines, however, the design of an algorithm is usually targeted at coping efficiently with unrealistic, even pathological inputs and the possibility is neglected that a simpler algorithm that works fast "on average" might perform just as well, or even better in practice. This alternative solution, called also a probabilistic approach, became an important issue two decades ago when it became clear that the prospects for showing the existence of polynomial time algorithms for NP-hard problems, were very dim. This fact, and apparently high success rate of heuristic approaches to solving certain difficult problems, led Richard Karp [57] to undertake a more serious investigation of probabilistic approximation algorithms. (But, one must realize that there are problems which are also hard "on average" as shown by Levin [68].) In the last decade we have witnessed an increasing interest in the probabilistic, also called *average case* analysis and design of algorithms, possibly due to the high success rate of randomized algorithms for computational geometry, scientific visualization, molecular biology, etc. (e.g., see [14, 43, 75, 97]).

The *average case* analysis of algorithms can be roughly divided into categories, namely: *analytical* (also called *precise*) and *probabilistic* analysis of algorithms. The former was popularized by Knuth's monumental three volumes *The Art of Computer Programming* [64, 65, 66] whose prime goal was to accurately predict the performance characteristics of an algorithm. Such an analysis more than often sheds light on properties of computer programs and provides useful

*This research was partially supported by NSF Grants NCR-9206315, NCR-9415491 and CCR-9201078, and NATO Collaborative Grant CRG.950060.

insights of combinatorial behaviors of such programs. Probabilistic methods were introduced by Erdös and Rényi and popularized by Erdös and Spencer in their book [23] (cf. also [5]). In general, nicely structured problems are amiable to an analytical approach that usually gives much more precise information about the algorithm under consideration. On the other hand, structurally complex algorithms are more likely to be first solved by a probabilistic tool that later could be further enhanced by a more precise analytical approach. The average case analysis of algorithms, as a discipline, uses a number of branches of mathematics: combinatorics, probability theory, graph theory, real and complex analysis, and occasionally algebra, geometry, number theory, operations research, and so forth.

In this chapter, we choose one facet of the theory of algorithms, namely that of algorithms and data structures on words (strings) and present a brief exposition on certain analytical and probabilistic methods that have become popular in such an endouver. Our choice of the area stems from a fact that there has been a resurgence of interest in *string algorithms* due to several novel applications, most notably in computational molecular biology and data compression. Our choice of methods covered here is aimed at closing a gap between analytical and probabilistic methods. There are excellent books on analytical methods (cf. Knuth's three volumes [64, 65, 66], Sedgewick and Flajolet [84]) and probabilistic methods (cf. Alon and Spencer [5], Coffman and Lueker [17], and Motwani and Raghavan [75]), however, remarkably very few books have been dedicated to both analytical and probabilistic analysis of algorithms (with possible exceptions of Hofri [46] and Mahmoud [73]). Finally, before we launch our journey through probabilistic and analytical methods, we should add that in recent years several useful surveys on analysis of algorithms have been published. We mentioned here: Frieze and McDiarmid [36], Karp [58], Vitter and Flajolet [96], and Flajolet [28].

This chapter is organized as follows: In the next section we describe some algorithms and data structures on words (e.g., digital trees, suffix trees, edit distance, Lempel-Ziv data compression algorithm, etc.) that we use throughout to illustrate our ideas and methods of analysis. Then, we present probabilistic models for algorithms and data structures on words together with a short review from probability and complex analysis. Section 4 is devoted to probabilistic methods and we discuss the sieve method, first and second moment methods, subadditive ergodic theorem, techniques of information theory (e.g., entropy and its applications), and large deviations (i.e., Chernoff's bound) and Azuma's type inequality. Finally, in the last section we concentrate on analytical techniques that we define as such in which complex analysis plays an important role. We plan to touch here analytical techniques for recurrences and asymptotics (i.e., Rice's formula, singularity analysis, etc.), Mellin transform and its applications, and poissonization and depoissonization.

2 Data Structures and Algorithms on Words

As mentioned above, in this survey we choose one facet of the theory of algorithms, namely that of data structures and algorithms on words (strings) to illustrate several probabilistic and analytical techniques of the analysis of algorithms. In this section, we briefly recall to the reader certain data structures and algorithms on words that we use extensively throughout this chapter.

Algorithms on words have experienced a new wave of interest due to a number of novel applications in computer science, telecommunications, and biology. Among others, these include dynamic hashing, partial match retrieval of multidimensional data, conflict resolution algorithms for broadcast communications, pattern matching, data compression, and searching and sorting. To satisfy these diversified demands various data structures were proposed for these algorithms. Undoubtly, the most popular data structures in algorithms on words are digital trees [66, 73] (e.g., tries, PATRICIA, digital search trees), and in particular suffix trees [2, 6, 19, 83, 84, 91]. We discuss them briefly below, together with general *edit distance* problem [8, 10, 16, 19, 67, 71, 76, 83, 97], and the *shortest common superstring* [13, 37, 67, 95] problem which recently became very popular due to possible application to the DNA sequencing problem.

2.1 Digital Trees

We start our discussion with a brief review of the **digital trees**. The most basic digital tree known as a **trie** (the name comes from re*trie*val) is defined first, and then other digital trees are described in terms of the trie.

The primary purpose of a trie is to store a set S of strings (words, keys), say $S = \{X_1, \ldots, X_n\}$. Each word $X = x_1x_2x_3\ldots$ is a finite or infinite string of symbols taken from a finite alphabet $\Sigma = \{\omega_1, \ldots, \omega_V\}$ of size $V = |\Sigma|$. A string will be stored in a leaf of the trie. The trie over S is built recursively as follows: For $|S| = 0$, the trie is, of course, empty. For $|S| = 1$, $trie(S)$ is a single node. If $|S| > 1$, S is split into V subsets $S_1, S_2, \ldots, S_V$ so that a string is in S_j if its first symbol is ω_j. The tries $trie(S_1), trie(S_2), \ldots, trie(S_V)$ are constructed in the same way except that at the k-th step, the splitting of sets is based on the k-th symbol. They are then connected from their respective roots to a single node to create $trie(S)$. Figure 2.1 illustrates such a construction.

There are many possible variations of the trie. One such variation is the b-*trie* in which a leaf is allowed to hold as many as b strings (cf. [31, 73, 91]). The b-trie is particularly useful in algorithms for extendible hashing in which the capacity of a page or other storage unit is b. A second variation of the trie, the **PATRICIA trie**, eliminates the waste of space caused

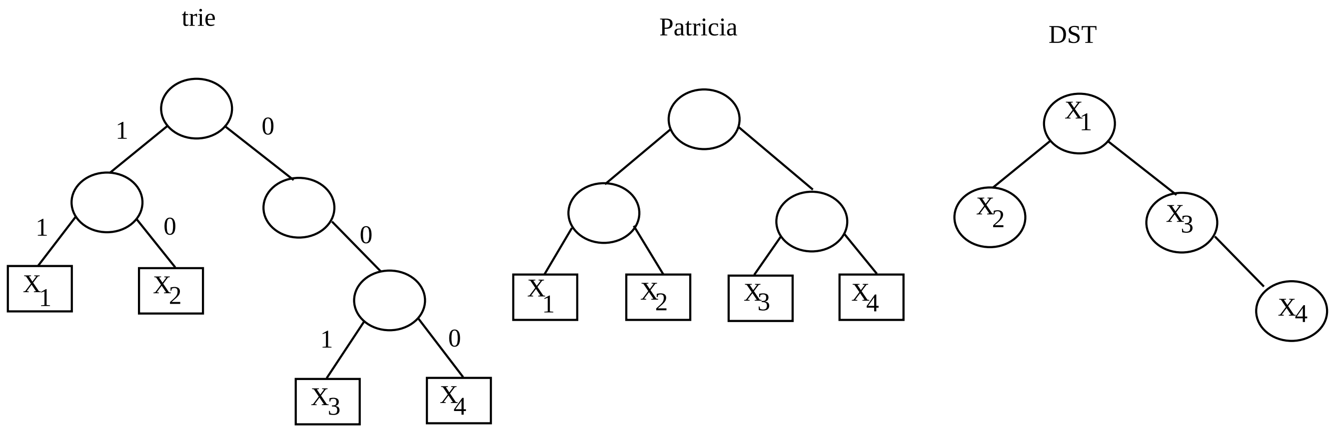

Figure 1: A trie, Patricia trie and a digital search tree (DST) built from the following four strings $X_1 = 11100\ldots$, $X_2 = 10111\ldots$, $X_3 = 00110\ldots$, and $X_4 = 00001\ldots$.

by nodes having only one branch. This is done by collapsing one-way branches into a single node. In a **digital search tree** keys (strings) are directly stored in nodes, and hence external nodes are eliminated. The branching policy is the same as in tries. Figure 2.1 illustrates these definitions.

The **suffix tree** and the *compact suffix tree* are similar to the trie and PATRICIA trie, but differ in the structure of the words that are being stored. In suffix trees and compact suffix trees, the words are suffixes of a given string X; that is, the word $X_j = x_j x_{j+1} x_{j+2} \ldots$ is the suffix of X which begins at the j-th position of X. Thus a suffix tree is a trie and a compact suffix tree is a PATRICIA trie in which the words are all suffixes of a given string.

Certain characteristics of tries and suffix trees are of primary importance. Hereafter, we assume that a digital tree is built from n strings or a suffix tree is constructed from a string of length n. The *m-depth* $D_n(m)$ of the m-th leaf in a trie is the number of internal nodes on the path from the root to the leaf. The (typical) depth of the trie D_n then, is the average depth over all its leaves, that is,

$$\Pr\{D_n \le k\} = \frac{1}{n} \sum_{m=1}^{n} \Pr\{D_n(m) \le k\} .$$

The *path length* L_n is the sum of all depths, that is,

$$L_n = \sum_{m=1}^{n} D_n(m) .$$

Closely related to the depth of a trie is the *depth of insertion*, which gives the depth of the $(n+1)$-st key inserted into a trie of n keys. The *height* H_n of the trie is the maximum depth of a leaf in the trie and can also be defined as the length of the longest path from the root to a leaf, that is,

$$H_n^{(b)} = \max_{1 \le m \le n} \{D_n(m)\} .$$

The *shortest path* s_n of the trie is the length of the shortest such path. Finally, the *size* S_n of the trie is given by the number of internal nodes in the trie. These characteristics are very useful in determining the expected size and shape of the data structures involved in algorithms on words. We study some of them in this chapter.

2.2 String Editing Problem

The string editing problem arises in many applications, notably in text editing, speech recognition, machine vision and, last but not least, molecular sequence comparison (cf. [97]). Algorithmic aspects of this problem have been studied rather extensively in the past (cf. [10, 76, 83, 97]). In fact, many important problems on words are special cases of string editing, including the *longest common subsequence* problem (cf. [19, 16, 83]) and the problem of *approximate pattern matching* (cf. [19]). In the following, we review the string editing problem and its relationship to the longest path problem in a special grid graph.

Let Y be a string consisting of ℓ symbols on some alphabet Σ of size V. There are three operations that can be performed on a string, namely *deletion* of a symbol, *insertion* of a symbol, and *substitution* of one symbol for another symbol in Σ. With each operation is associated a *weight* function. We denote by $W_I(y_i)$, $W_D(y_i)$ and $W_Q(x_i, y_j)$ the weight of insertion and deletion of the symbol $y_i \in \Sigma$, and substitution of x_i by $y_j \in \Sigma$, respectively. An *edit script* on Y is any sequence of edit operations, and the total weight of it is the sum of weights of the edit operations.

The **string editing** problem deals with two strings, say Y of length ℓ (for ℓong) and X of length s (for $s$hort), and consists of finding an edit script of minimum (maximum) total weight that transforms X into Y. The maximum (minimum) weight is called the *edit distance from X to Y*, and its is also known as the Levenshtein distance. In molecular biology, the Levenshtein distance is used to measure similarity (homogeneity) of two molecular sequences, say DNA sequences (cf. [83]).

The string edit problem can be solved by the standard dynamic programming method. Let $C_{\max}(i,j)$ denote the maximum weight of transforming the prefix of Y of size i into the prefix of X of size j. Then, (cf. [10, 76, 97])

$$C_{\max}(i,j) = \max\{C_{\max}(i-1,j-1)+W_Q(x_i,y_j), C_{\max}(i-1,j)+W_D(x_i), C_{\max}(i,j-1)+W_I(y_j)\}$$

for all $1 \le i \le \ell$ and $1 \le j \le s$. We compute $C_{\max}(i,j)$ row by row to obtain finally the total cost $C_{\max} = C_{\max}(\ell, s)$ of the maximum edit script. A similar procedure works for the minimum edit distance.

The key observation for us is to note that interdependency among the partial optimal weights $C_{\max}(i,j)$ induce an $\ell \times s$ grid-like directed acyclic graph, called further a *grid graph*. In such a

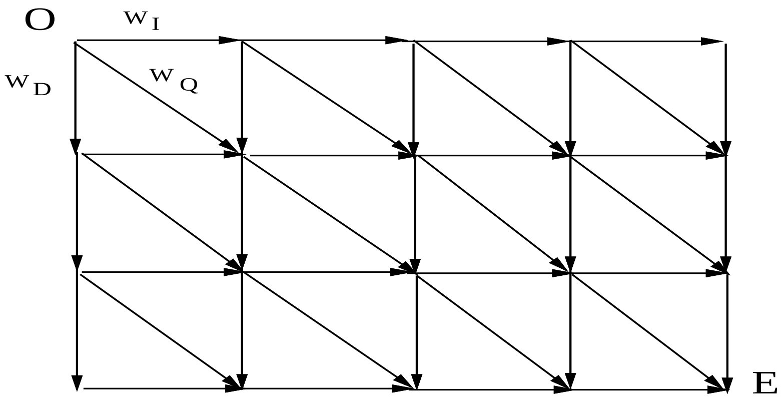

Figure 2: Example of a grid graph of size $\ell = 4$ and $s = 3$.

graph vertices are points in the grid and edges go only from (i, j) point to neighboring points, namely $(i, j+1)$, $(i+1, j)$ and $(i+1, j+1)$. A horizontal edge from $(i-1, j)$ to (i, j) carries the weight $W_I(y_j)$; a vertical edge from $(i, j-1)$ to (i, j) has weight $W_D(x_i)$; and finally a diagonal edge from $(i-1, j-1)$ to (i, j) is weighted according to $W_Q(x_i, y_j)$. Figure 2 shows an example of such an edit graph. The edit distance is the longest (shortest) path from the point $O = (0, 0)$ to $E = (\ell, s)$.

Finally, we should mention that by selecting properly the distributions of W_I, W_D and W_Q we can model several variations of the string editing problem. For example, in the standard setting the deletion and insertion weights are identical, and usually constant, while the substitution weight takes two values, one (high) when matching between a letter of X and a letter of Y occurs, and another value (low) in the case of a mismatch (e.g., in the *Longest Common Substring* problem [16, 83], one sets $W_I = W_D = 0$, and $W_Q = 1$ when a matching occurs, and $W_Q = -\infty$ in the other case).

2.3 Shortest Common Superstring

Various versions of the **shortest common superstring** (in short: SCS) problem play important roles in data compression and DNA sequencing. In fact, in laboratories DNA sequencing (cf. [67, 97]) is routinely done by sequencing large numbers of relatively short fragments, and then heuristically finding a short common superstring. The problem can be formulated as follows: given a collection of strings, say $X_1, X_2, \ldots, X_n$ over an alphabet Σ, find the shortest string Z such that each of X_i appears as a substring (a consecutive block) of Z.

It is known that computing the shortest common superstring is NP-hard. Thus constructing a good approximation to SCS is of prime interest. It has been shown recently, that a greedy algorithm can compute in $O(n \log n)$ time a superstring that in the worst case is only β times (where $2 \leq \beta \leq 4$) longer than the shortest common superstring [13, 95]. Often, one is interested in maximizing total overlap of SCS using a greedy heuristic and to show that such a heuristic produces an overlap $O_n^{\rm gr}$ that approximates well the optimal overlap $O_n^{\rm opt}$ where n is the number of strings.

More precisely, suppose $X = x_1x_2\ldots x_r$ and $Y = y_1y_2\ldots y_s$ are strings over the same finite alphabet Σ. We also write $|X|$ for the length of X. We define their *overlap* $o(X,Y)$ by

$$o(X,Y) = \max\{j:\ y_i = x_{r-i+1},\ 1 \le i \le j\}.$$

If $X \neq Y$ and $k = o(X,Y)$, then

$$X \oplus Y = x_1x_2\ldots x_r y_{k+1}y_{k+2}\ldots y_s.$$

Let $\mathcal{S}$ be a set of all superstrings built over the strings $X_1,\ldots,X_n$. Then,

$$O_n^{\rm opt} = \sum_{i=1}^{n} |X_i| - \min_{Z\in\mathcal{S}} |Z|.$$

A generic greedy algorithm for the SCS problem can be described as follows: Its input is the n strings $X_1, X_2, \ldots, X_n$ over Σ. It outputs a string Z which is a superstring of the input.

`Generic greedy algorithm`

1. $I \leftarrow \{X_1, X_2, X_3, \ldots, X_n\}$; $O_n^{\rm gr} \leftarrow 0$;
2. **repeat**
3. choose $X, Y \in I$; $Z = X \oplus Y$;
4. $I \leftarrow (I \setminus \{X,Y\}) \cup \{Z\}$;
5. $O_n^{\rm gr} \leftarrow O_n^{\rm gr} + o(X,Y)$;
6. **until** $|I| = 1$

Different variants of the above generic algorithm can be envisioned by interpreting appropriately the "choose" statement in Step 3 above. We shall discuss some probabilistic aspects of it in sections below.

3 Probabilistic Models

In this section, we first discuss a few probabilistic models of randomly generated strings. Then, we briefly review some basic facts from probability theory (e.g., types of stochastic convergence), and finally we provide some elements of complex analysis that we shall use in this chapter.

3.1 Probabilistic Models of Strings

As expected, random shape of data structures on words depends on the underlying probabilistic assumptions concerning the strings involved. Below, we discuss a few basic probabilistic models that one often encounters in the analysis of problems on words.

We start with the most elementary model, namely the **Bernoulli model** that is defined as follows:

(B) BERNOULLI MODEL

Symbols of the alphabet $\Sigma = \{\omega_1, \ldots, \omega_V\}$ occur independently of one another; thus, a key $X = x_1x_2x_3\ldots$ can be described as the outcome of an infinite sequence of Bernoulli trials in which $\Pr\{x_j = \omega_i\} = p_i$ and $\sum_{i=1}^{V} p_i = 1$. If $p_1 = p_2 = \ldots = p_V = 1/V$, then the model is called *symmetric*; otherwise, it is *asymmetric*. Throughout the paper we only consider *binary alphabet* $\Sigma = \{0, 1\}$ with p being the probability of "0" and $q = 1 - p$ the probability of "1".

In general, when one deals with many strings (e.g., when building a digital tree) additional assumption is made concerning the independence of the strings involved.

In many cases, assumption (B) is not very realistic. For instance, if the strings are words from the English language, then there certainly is a dependence among the symbols of the alphabet. As an example, h is much more likely to follow an s than a b. When this is the case, assumption (B) can be replaced by

(M) MARKOVIAN MODEL

There is a Markovian dependency between consecutive symbols in a key; that is, the probability $p_{ij} = \Pr\{X_{k+1} = \omega_j | X_k = \omega_i\}$ describes the conditional probability of sampling symbol ω_j immediately after symbol ω_i.

There are two further generalizations of the Markovian model, namely *mixing model* and the *stationary model* that are very useful in practice, especially when dealing with problems of data compression or molecular biology when one expects long dependency among symbols of a string.

(MX) MIXING MODEL

Let $\mathcal{F}_m^n$ be a σ-field generated by $\{X_k\}_{k=m}^n$ for $m \le n$. There exists a function $\alpha(\cdot)$ of g such that: (i) $\lim_{g\to\infty} \alpha(g) = 0$, (ii) $\alpha(1) < 1$, and (iii) for any m, and two events $A \in \mathcal{F}_{-\infty}^m$ and $B \in \mathcal{F}_{m+g}^{\infty}$ the following holds

$$(1 - \alpha(g))\Pr\{A\}\Pr\{B\} \le \Pr\{AB\} \le (1 + \alpha(g))\Pr\{A\}\Pr\{B\} .$$

In words, model (MX) says that the dependency between $\{X_k\}_{k=1}^m$ and $\{X_k\}_{k=m+g}^\infty$ is getting weaker and weaker as g becomes larger (note that when the sequence $\{X_k\}$ is i.i.d., then $\Pr\{AB\} = \Pr\{A\}\Pr\{B\}$). The "quantity" of dependency is characterized by $\alpha(g)$ (cf. [11, 24]).

The most general probabilistic model that can provide some useful results, is the stationary model.

(S) Stationary Model

The sequence $\{X_k\}_{k=1}^\infty$ of letters from a finite alphabet is a *stationary and ergodic* sequence of random variables.

To explain how the stationary model works, we need to introduce some notations. Let $X_m^n = (X_m, ..., X_n)$ for $m < n$, and let for every $n \geq 1$ the nth order probability distribution for $\{X_k\}$ be $P(X_1^n) = \Pr\{X_k = x_k,\ 1 \leq k \leq n,\ x_k \in \mathcal{A}\}$. In the stationary model, this probability does not depend on time-shift, that is, if τ is an integer , then for every n and τ the following holds $P(X_{1+\tau}^{n+\tau}) = P(X_1^n)$ (cf. [11, 24]).

3.2 Quick Review from Probability: Types of Stochastic Convergence

We begin with some elementary definitions from probability theory. The reader is refered to [24, 25, 26, 85] for more detailed discussions. Let the random variable X_n denote the value of a parameter of interest depending on n (e.g., depth in a suffix tree and/or trie built over n strings). The expected value $\mathbf{E}[X_n]$ or mean and the variance $\mathbf{Var}[X_n]$ can be computed as $\mathbf{E}[X_n] = \sum_{k=0}^\infty k\Pr\{X_n = k\}$ and $\mathbf{Var}[X_n] = \sum_{k=0}^\infty (k - \mathbf{E}[X_n])^2 \Pr\{X_n = k\}$.

Convergence of Random Variables. It is important to note the different ways in which random variables are said to converge. To examine the different methods of convergence, let X_n be a sequence of random variables, and let their distribution functions be $F_n(x)$, respectively. A good and easy to read account on various types of convergence can be found in Shiryayev [85].

The first notion of convergence of a sequence of random variables is known as **convergence in probability**. The sequence X_n converges to a random variable X *in probability*, denoted $X_n \to X$ (pr.) or $X_n \xrightarrow{p} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\{|X_n - X| < \epsilon\} = 1 \ .$$

Note that this does not say that the difference between X_n and X becomes very small. What converges here is the *probability* that the difference between X_n and X becomes very small. It is, therefore, possible, although unlikely, for X_n and X to differ by a significant amount and for such differences to occur infinitely often.

A stronger kind of convergence which does not allow such behavior is called **almost sure convergence** or **strong convergence**. A sequence of random variables X_n converges to a random variable X *almost surely*, denoted $X_n \to X$ (a.s.) or $X_n \stackrel{(a.s.)}{\to} X$, if for any $\epsilon > 0$,

$$\lim_{N\to\infty} \Pr\{\sup_{n\geq N} |X_n - X| < \epsilon\} = 1 \ .$$

From this formulation of almost sure convergence, it is clear that if $X_n \to X$ (a.s.), the probability of infinitely many large differences between X_n and X is zero. The sequence X_n in this case is said to satisfy the strong law of large numbers. As the term strong implies, almost sure convergence implies convergence in probability.

A simple criterion for almost sure convergence can be inferred from the *Borel-Cantelli lemma*. We give it in the following corollary.

Lemma 1 (Borel-Cantelli) *Let $\epsilon > 0$. If $\sum_{n=0}^{\infty} \Pr\{|X_n - X| > \epsilon\} < \infty$, then $X_n \to X$ (a.s.).*

Proof. It follows directly from the following chain of inequalities (the reader is referred to Section 4.1 for more explanations on these inequalities):

$$\Pr\{\sup_{n\geq N} |X_n - X| \geq \epsilon\} = \Pr\{\bigcup_{n\geq N} (|X_n - X| \geq \epsilon)\} \leq \sum_{n\geq N} \Pr\{|X_n - X| \geq \epsilon\} \to 0 \ .$$

The last convergence is a consequence of our assumption that $\sum_{n=0}^{\infty} \Pr\{|X_n - X| > \epsilon\} < \infty$. ■

A third type of convergence is defined on the distribution functions $F_n(x)$. The sequence of random variables X_n **converges in distribution** or **converges in law** to the random variable X, denoted $X_n \stackrel{d}{\to} X$ if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for each point of continuity of $F(x)$. Almost sure convergence implies convergence in distribution.

Finally, the convergence *in mean of order p* implies that $\mathbf{E}[|X_n - X|^p] \to 0$ as $n \to \infty$. It is well known that almost sure convergence and convergence in mean imply the convergence in probability. On the other hand, the convergence in probability leads to the convergence in distribution. If the limiting random variable X is a constant, then the convergence in distribution also implies the convergence in probability (cf. [11, 24]).

Generating Functions. The distribution of a random variable can also be described using generating functions. The *ordinary generating function* $G_n(u)$, and a bivariate *exponential generating function* $g(z, u)$ are defined as

$$\begin{aligned} G_n(u) &= \mathbf{E}[u^{X_n}] = \sum_{k=0}^{\infty} \Pr\{X_n = k\} u^k \ , \\ g(z,u) &= \sum_{n=0}^{\infty} G_n(u) \frac{z^n}{n!} \ . \end{aligned}$$

These functions are well-defined for any complex numbers z and u such that $|u| < 1$. Observe that

$$\begin{aligned} \mathbf{E}[X_n] &= G_n'(1) , \\ \mathbf{Var}[X_n] &= G_n''(1) + G_n'(1) - [G_n'(1)]^2 . \end{aligned}$$

Levy's Continuity Theorem. Our next step is to relate convergence in distribution to convergence of generating functions. The following results, known as *Levy's continuity theorem* is an archi-fact for most distributional analysis. For our purpose we formulate it in terms of the Laplace transform of X_n, namely $G_n(e^{-t}) = \mathbf{E}[e^{-tX_n}]$ for real t.

Theorem 1 (Continuity Theorem) *Let X_n and X be random variables with Laplace transforms $G_n(e^{-t})$ and $G(e^{-t})$, respectively. A necessary and sufficient condition for $X_n \xrightarrow{d} X$ is that $G_n(e^{-t}) \to G(e^{-t})$ for all $t \geq 0$.*

The above theorem holds if we set $t = i\nu$ for $-\infty < \nu < \infty$ (i.e., we consider characteristic functions). Moreover, if the above holds for t complex number, then we automatically derive convergence in moments due to the fact that an analytical function possesses all its derivatives.

Finally, a key result used in establishing central limit theorem (i.e., convergence to a normal distribution) is a theorem by Goncharov (cf. [64], Chap 1.2.10, Ex. 13). This theorem states that a sequence of random variables X_n with mean $\mathbf{E}[X_n] = \mu_n$ and standard deviation $\sigma_n = \sqrt{\mathbf{Var}[X_n]}$ approaches a normal distribution if the following holds:

$$\lim_{n\to\infty} e^{-\tau\mu_n/\sigma_n} P_n(e^{\tau/\sigma_n}) = e^{\tau^2/2}$$

for all $\tau = i\nu$ and $-\infty < \nu < \infty$, and X_n converges in moments if τ is a complex number.

3.3 Review from Complex Analysis

Much of the necessary complex analysis involves the use of Cauchy's integral formula and Cauchy's residue theorem. We briefly recall a few facts from analytical functions, and then discuss the above two theorems. For precise definitions and formulations the reader is referred to [45, 82]. We shall follow here Flajolet and Sedgewick [35].

A function $f(z)$ of complex variable z is analytical at point $z = a$ if it is differentiable in a neighbourhood of $z = a$ or equivalently it has a convergent series representation around $z = a$. Let us concentrate our discussion only on *meromorphic functions* that are analytical with an exception of a finite number of points called poles. More formally, a meromorphic function $f(z)$ can be represented in a neighbourhood of $z = a$ with $z \neq a$ by Laurent series as follows:

$$f(z) = \sum_{n \geq -M} f_n (z-a)^n ,$$

for some integer M. If the above holds with $f_{-M} \neq 0$, then it is said that $f(z)$ has a *pole* of order M at $z = a$. **Cauchy's Integral Theorem** says that the coefficients f_n of an analytical function in a disk can be computed as

$$f_n := [z^n]f(z) = \frac{1}{2\pi i} \oint f(z) \frac{dz}{z^{n+1}}$$

and the circle is traversed counterclockwise.

An important tool frequently used in the analytical analysis of algorithms is *residue theory.* The residue of $f(z)$ at a point a is the coefficient of $(z-a)^{-1}$ in the expansion of $f(z)$ around a, and it is denoted as

$$\mathrm{Res}[f(z);\ z = a] = f_{-1}\ .$$

There are many simple rules to evaluate residues and the reader can find them in any standard book on complex analysis (e.g., [45, 82]). Actually, the easiest way to compute a residue of a function is to use the `series` commend in MAPLE that produces a series development of a function. The residue is simply the coefficient at $(z-a)^{-1}$. For example, the following session of MAPLE computes series of $f(z) = \Gamma(z)/(1-2^z)$ at $z = 0$:

```
series(GAMMA(z)/(1-2^z), z=0, 4);
```

$$-\frac{1}{\ln(2)} z^{-2} - \frac{-\gamma - \frac{1}{2}\ln(2)}{\ln(2)} z^{-1} - \frac{-\frac{1}{6}\ln(2)^2 + \frac{1}{12}\pi^2 + \frac{1}{2}\gamma^2 + \frac{1}{4}(2\gamma + \ln(2))\ln(2)}{\ln(2)} + \mathrm{O}(z)$$

From the above we see that $\mathrm{Res}[f(z);\ z = 0] = \frac{\gamma}{\log 2} + \frac{1}{2}$.

Residues are very important in evaluating contour integrals. In fact, a well-known theorem in complex analysis, that is, **Cauchy's residue theorem** states that if $f(z)$ is analytic within and on the boundary of C except at a finite number of poles $a_1, a_2, \ldots, a_N$ inside of C having residues $\mathrm{Res}[f(z);\ z = a_1], \ldots, \mathrm{Res}[f(z);\ z = a_N]$, then

$$\oint_C f(z)dz = 2\pi i \sum_{j=1}^{N} \mathrm{Res}[f(z);\ z = a_j]\ ,$$

where the curve C is traversed counterclockwise.

4 Probabilistic Techniques

In this section we discuss several probabilistic techniques that have been successfully applied to the average case analysis of algorithms. We start with some elementary **inclusion-exclusion** principle known also as **sieve methods**. Then, we present very useful **first** and **second moment methods**. We continue with the **subadditive ergodic theorem** that is quite popular for deducing some properties of problems on words. Next, we turn our attention to some probabilistic methods of information theory, and in particular we discuss **entropy** and **asymptotic equipartition property**. Finally, we look at some **large deviations** results and **Azuma's type inequality**. In this section, as well in the next one where analytical techniques are discussed, we adopt the following scheme of presentation: First, we describe the method and give a short intuitive derivation. Then, we illustrate it on some non-trivial examples taken from the problems on words discussed in Section 2.

4.1 Sieve Method and Its Variations

The *inclusion-exclusion principle* is one of the oldest tools in combinatorics, number theory (where this principle is known as *sieve method*), discrete mathematics, and probabilistic analysis. It provides a tool to estimate probability of a union of *not* disjoint events, say $\bigcup_{i=1}^{n} A_i$ where A_i are events for $i = 1, \ldots, n$. However, before we plunge into our discussion, let us first show a few examples of problems on words for which an estimation of the probability of a union of events is required.

Example 1: *Depth and Height in a Trie*

In Section 2.1 we discussed tries built over n binary strings $X_1, \ldots, X_n$. We assume that those strings are generated according to the Bernoulli model with one symbol, say "0", occurring with probability p and the other, say "1", with probability $q = 1 - p$. Let C_{ij}, known as *alignment* between ith and jth strings, be defined as the length of the longest string that is a prefix of X_i and X_j. Then, it is easy to see that the mth depth $D_n(m)$ (i.e., length of a path in trie from the root to the external node containing X_m), and the height H_n (i.e., the length of the longest path in a trie) can be expressed as follows:

$$D_n(m) = \max_{1 \le i \ne m \le n} \{C_{i,m}\} + 1 \ , \qquad (1)$$

$$H_n = \max_{1 \le i < j \le n} \{C_{ij}\} + 1 \ . \qquad (2)$$

Certainly, the alignments C_{ij} are dependent random variables even for the Bernoulli model. The above equations expressed the depth and the height as an *order statistic* (i.e., maximum of the sequence $C_{i,j}$ for $i, j = 1, \ldots, n$). We can estimate some probabilities associated with the depth and the height as a union of properly defined events. Indeed, let $A_{ij} = \{C_{ij} > k\}$ for some k. Then, one finds

$$\Pr\{D_n(m) > k\} = \Pr\{\bigcup_{i=1, \ne m}^{n} A_{i,m}\} \ , \qquad (3)$$

$$\mathsf{Pr}\{H_n > k\} \quad = \quad \mathsf{Pr}\{\bigcup_{i,j=1}^{n} A_{i,j}\} \ . \tag{4}$$

In passing, we should point out that for the Shortest Common Superstring Problem (cf. Section 2.3) we need to estimate a quantity $M_n(m)$ which is similar to $D_n(m)$ except that C_{im} is defined as the length of the longest string that is a prefix of X_i and suffix of X_m for fixed m. One easily observes that $M_n(m) \stackrel{d}{=} D_n(m)$, that is, these two quantities are equal *in distribution*. $\square$

We have just seen that often we need to estimate a probability of union of events. The following formula is known as *inclusion-exclusion formula* (cf. [12])

$$\mathsf{Pr}\{\bigcup_{i=1}^{n} A_i\} = \sum_{r=1}^{n}(-1)^{r+1} \sum_{|J|=r} \mathsf{Pr}\{\bigcap_{j\in J} A_j\} \ . \tag{5}$$

The next example illustrates it on the depth on a trie.

Example 2: *Generating Function of the Depth in a Trie*

Let us compute the generating function of the depth $D_n := D_n(1)$ for the first string X_1. We start with (3), and after some algebraic manipulation (5) leads to (cf. [52, 53])

$$\begin{aligned}
\mathsf{Pr}\{D_n \ge k\} &= \mathsf{Pr}\{\bigcup_{i=2}^{n}[C_{i,1} \ge k]\} \\
&= \frac{1}{n}\sum_{r=2}^{n}(-1)^r \binom{n}{r} \mathsf{Pr}\{C_{2,1} \ge k, \ldots, C_{r,1} \ge k\}
\end{aligned}$$

since the probability $\mathsf{Pr}\{C_{2,1} \ge k, \ldots, C_{r,1} \ge k\}$ does not depend on the choice of strings (i.e., it is the same for any r-tuple of strings selected). Moreover, it can be easily explicitly computed. Indeed, we obtain

$$\mathsf{Pr}\{C_{2,1} \ge k, \ldots, C_{r,1} \ge k\} = (p^r + q^r)^k$$

since r independent strings must agree on the first k symbols. Thus, the generating function $D_n(u) = \mathbf{E}[u^{D_n}]$ becomes

$$\mathbf{E}[z^{D_n}] = 1 - \frac{1-z}{n}\sum_{r=2}^{n}(-1)^r \binom{n}{r} r \frac{1}{1 - z(p^r + q^r)} \ .$$

The last formula is a simple consequence of the above, and the following well known fact from the generating function $\mathbf{E}[u^X]$ for a random variable X:

$$\mathbf{E}[u^X] = \frac{1}{1-u}\sum_{k=0}^{\infty} \mathsf{Pr}\{X \le k\} u^k$$

for $|u| < 1$. $\square$

In many real computations, however, one cannot explicitly compute the probability of the events union. Often, one must retreat to inequalities that actually are enough to reach one's goal. The most simple yet still very powerful is the following inequality

$$\Pr\{\bigcup_{i=1}^{n} A_i\} \leq \sum_{i=1}^{n} \Pr\{A_i\} \ . \tag{6}$$

The latter is an example of a series of inequalities due to Bonferroni which can be formulated as follows: For every even integer $e \geq 0$ we have

$$\begin{aligned}\sum_{j=1}^{m}(-1)^{j-1} \sum_{1\leq t_1\leq \cdots \leq t_j \leq n} \Pr\{A_{t_1} \cap \cdots \cap A_{t_j}\} &\leq \Pr\{\bigcup_{i=1}^{n} A_i\} \\ &\leq \sum_{j=1}^{m+1}(-1)^{j-1} \sum_{1\leq t_1\leq \cdots \leq t_j \leq n} \Pr\{A_{t_1} \cap \cdots \cap A_{t_j}\} \ .\end{aligned}$$

In combinatorics (e.g., enumeration problems) and probability the so called *inclusion-exclusion principle* is very popular and had many successes. We formulate it in a form of a theorem whose proof can be found in Bollobás [12].

Theorem 2 (Inclusion-Exclusion Principle) *Let $A_1, \ldots, A_n$ be events in a probability space, and let p_k be the probability of exactly k of them to occur. Then:*

$$p_k = \sum_{r=k}^{n}(-1)^{r+k}\binom{r}{k} \sum_{|J|=r} \Pr\{\bigcap_{j\in J} A_j\} \ .$$

Example 3: *Computing a Distribution Through Its Moments*

Let X be a random variable defined on $\{0, 1, \ldots, n\}$, and let $\mathbf{E}_r[X] = \mathbf{E}X(X-1)\cdots(X-r+1)$ be the rth factorial moment of X. Then:

$$\Pr\{X = k\} = \frac{1}{k!}\sum_{r=k}^{n}(-1)^{r+k}\frac{\mathbf{E}_r[X]}{(r-k)!} \ .$$

Indeed, it suffices to set $A_i = \{X \geq i\}$ for all $i = 1, \ldots, n$, and observe that $\sum_{|J|=r} \Pr\{\bigcap_{j\in J} A_j\} = \mathbf{E}_r[X]/r!$. Since the event $\{X = k\}$ is equivalent to the event that exactly k of A_i occur, a simple application of Theorem 2 proves the announced result. □

Finally, we say a few words about the so called *Lovász Local Lemma* (cf. [5, 12, 75]) which can be simply stated as follows: *If there are n mutually independent (bad) events B_i each of probability strictly smaller than one, then there is a positive probability that none of the bad events happens.* We shall formulate this in a more precise manner below. It should be clear, however, from the above that this lemma can be used to prove the existence of some complicated structure by showing that it must occur with a positive probability.

Let us start with a simple statement. *If* $\sum_{i=1}^n \mathsf{Pr}\{B_i\} < 1$, *then* $\mathsf{Pr}\{\bigcap_{i=1}^n \bar{B}_i\} > 0$. Indeed,

$$\mathsf{Pr}\{\bigcap_{i=1}^n \bar{B}_i\} = 1 - \mathsf{Pr}\{\bigcup_{i=1}^n A_i\} \geq 1 - \sum_{i=1}^n \mathsf{Pr}\{B_i\} > 0$$

where the first inequality follows from Bonferroni's inequality (6).

A stronger statement is in fact true. First of all, let us notice that *if* $\mathsf{Pr}\{B_i\} < 1$ *and all* n *events are mutually independent, then* $\mathsf{Pr}\{\bigcap_{i=1}^n \bar{B}_i\} > 0$. Indeed, it suffices to observe that $\mathsf{Pr}\{\bigcap_{i=1}^n \bar{B}_i\} = \prod_{i=1}^n (1 - \mathsf{Pr}\{B_i\}) > 0$. Erdös and Lovász (cf. [12]) proved that the above conclusion is true even if there is some dependency among the events. For example: *let every* d *events be dependent and let* $\mathsf{Pr}\{B_i\} \leq P$ *for all* $i = 1, \ldots, n$ *such that* $4dP < 1$. *Then,* $\mathsf{Pr}\{\bigcap_{i=1}^n \bar{B}_i\} > 0$. The reader is referred to [5, 12] for a more general formulation of this lemma, and for interesting applications.

4.2 Inequalities: First and Second Moment Methods

In this subsection, we review some inequalities that play a considerable role in probabilistic analysis of algorithms. In particular, we discuss *first* and *second moment methods* that are "bread-and-butter" of a typical probabilistic analysis.

We start with a few standard inequalities (cf. [24, 85]):

Markov Inequality: For a nonnegative random variable X and $\varepsilon > 0$ the following holds:

$$\mathsf{Pr}\{X \geq \varepsilon\} \leq \frac{\mathbf{E}[X]}{\varepsilon} .$$

Indeed: let $I(A)$ be the indicator function of A (i.e., $I(A) = 1$ if A occurs, and zero otherwise). Then,

$$\mathbf{E}[X] \geq \mathbf{E}[XI(X \geq \varepsilon)] \geq \varepsilon \mathbf{E}[I(X \geq \varepsilon)] = \varepsilon \mathsf{Pr}\{X \geq \varepsilon\} .$$

Chebyshev's Inequality: If one replaces X by $|X - E[X]|$ in the Markov inequality, then

$$\mathsf{Pr}\{|X - \mathbf{E}[X]| > \varepsilon\} \leq \frac{\mathbf{Var}[X]}{\varepsilon^2} .$$

Schwarz's Inequality (also called Cauchy-Schwarz): Let X and Y be such that $\mathbf{E}[X^2] < \infty$ and $E[Y^2] < \infty$. Then:

$$\mathbf{E}[|XY|]^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2] ,$$

where $\mathbf{E}[X]^2 := (\mathbf{E}[X])^2$.

Jensen's Inequality: Let $f(\cdot)$ be a downward convex function, that is, for $\lambda \in (0,1)$

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda) y) .$$

Then:

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)] \ .$$

The remainder part of this subsection is devoted to the first and the second moment methods that we illustrate on several examples arising in the analysis of digital trees. The *first moment method* for a nonnegative random variable X boils down to

$$\mathsf{Pr}\{X > 0\} \leq \mathbf{E}[X] \ . \tag{7}$$

This follows directly from Markov's inequality after setting $\varepsilon = 1$. The above inequality implies also the basic Bonfferroni inequality (6). Indeed, let A_i ($i = 1, \ldots, n$) be events, and set $X = I(A_1) + \cdots + I(A_n)$. Inequality (6) follows.

In a typical usage of (7), we expect to show that $\mathbf{E}[X] \to 0$, just $X = 0$ occurs almost always or *with high probability* (whp). We illustrate it in the next example.

Example 4: *Upper Bound on the Height in a Trie*

In Example 1 we showed that the height H_n of a trie is given by (2) or (4). Thus, using the first moment method we have

$$\mathsf{Pr}\{H_n \geq k+1\} \leq \mathsf{Pr}\{\max_{1 \leq i < j \leq n}\{C_{ij}\} \geq k\} \leq n^2 \mathsf{Pr}\{C_{ij} \geq k\}$$

for any integer k. From Example 2 we know that $\mathsf{Pr}\{C_{ij} \geq k\} = (p^2 + q^2)^k$. Let $P = p^2 + q^2$, $Q = P^{-1}$, and set $k = 2(1+\varepsilon)\log_Q n$ for any $\varepsilon > 0$. Then, the above implies

$$\mathsf{Pr}\{H_n \geq 2\log_Q n + 1\} \leq \frac{n^2}{n^{2(1+\varepsilon)}} = \frac{1}{n^{2\varepsilon}} \to 0 \ ,$$

thus $H_n/(2\log_Q n) \leq 1$ (pr.). In the example below, we will actually prove that $H_n/(2\log_Q n) = 1$ (pr.) by establishing a lower bound. □

Let us look now at the *second moment method.* Setting in the Chebyshev inequality $\varepsilon = \mathbf{E}[X]$ we easily prove that

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

But, one can do better (cf. [5, 17]). Using Schwarz's inequality for a random variable X we obtain the following chain of inequalities

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

which finally implies the second moment inequality

$$\mathsf{Pr}\{X > 0\} \geq \frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]} \ . \tag{8}$$

Actually, another formulation of this inequality due to Chung and Erdös is quite popular. To derive it, set in (8) $X = I(A_1) + \cdots + I(A_n)$ for a sequence of events $A_1, \ldots, A_n$. Noting that $\{X > 0\} = \bigcup_{i=1}^n A_i$, we obtain after some algebra

$$\Pr\{\bigcup_{i=1}^n A_i\} \geq \frac{(\sum_{i=1}^n \Pr\{A_i\})^2}{\sum_{i=1}^n \Pr\{A_i\} + \sum_{i\neq j} \Pr\{A_i \cap A_j\}} \ . \tag{9}$$

In a typical application, we are able to prove that $\mathbf{Var}[X]/\mathbf{E}[X^2] \to 0$, thus showing that $\{X > 0\}$ almost always. The next example – which is a continuation of Example 4 – illustrates this point.

Example 5: *Lower Bound for the Height in a Trie.*

We now prove that $\Pr\{H_n \geq 2(1-\varepsilon)\log_Q n\} \to 1$, just completing the proof that $H_n/(2\log_Q n) \to 1$ (pr.). We use the Chung-Erdös formulation, and set $A_{ij} = \{C_{ij} \geq k\}$. Throughout this example, we assume $k = 2(1-\varepsilon)\log_Q n$. Observe that now in (9) we must replace the single summation index i by a double summation index (i,j). The following is obvious

$$\sum_{1\leq i<j\leq n} \Pr\{A_{ij}\} = \frac{1}{2}n(n-1)P^k \ ,$$

where $P = p^2 + q^2$. The other sum in (9) is a little harder to deal with. We must sum over $(i,j),(l,m)$, and we consider two cases: (i) all indices are different, (ii) $i = l$ (i.e., we have $(i,j),\ (i,m)$). In the second case we must consider the probability $\Pr\{C_{ij} \geq k, C_{i,m} \geq k\}$. But, as in Example 2, we obtain $\Pr\{C_{ij} \geq k, C_{i,m} \geq k\} = (p^3+q^3)^k$ since once you choose a symbol in the string X_i you must have the same symbol on the same position in X_j, X_m. In summary,

$$\sum_{(ij),(l,m)} \Pr\{C_{ij} \geq k,\ C_{lm} \geq k\} \leq \frac{1}{4}n^4 P^{2k} + n^3(p^3+q^3)^k \ .$$

We need a simple inequality that can be easily verified (see [56, 7] for a more elaborate extension), namely:

$$(p^3+q^3)^{\frac{1}{3}} \leq P^{\frac{1}{2}} \ .$$

Then, (9) becomes

$$\begin{aligned}\Pr\{H_n \geq k\} = \Pr\{\bigcup_{i=1}^n A_i\} &\geq \frac{1}{n^{-2}P^{-k} + 1 + 4(p^3+q^3)^k/(nP^{2k})} \\ &\geq \frac{1}{1+n^{-2\varepsilon} + 4/(nP^{k/2})} \geq \frac{1}{1+n^{-2\varepsilon}+4n^{-\varepsilon}} \to 1 \ .\end{aligned}$$

Thus, we have shown that $H_n/(2\log_Q n) \geq 1$ (pr.), which completes our proof of

$$\lim_{n\to\infty} \Pr\{2(1-\varepsilon)\log_Q n \leq H_n \leq 2(1+\varepsilon)\log_Q n\} = 1$$

for any $\varepsilon > 0$. □

We complete this subsection with two results concerning order statistics that find plenty of applications in average case analysis of algorithms. These results are direct consequences of the methods discussed here, but for completeness we give a short derivation (cf. [4, 39]).

Lemma 2 *Let $Y_1,\ Y_2,\dots,Y_m$ be a sequence of random variables with distribution functions $F_1(y),\ F_2(y)$,..., $F_m(y)$, respectively. Let $R_i(y) = \Pr\{Y_i \geq y\}$ be the complement function of the distribution function $F_i(y)$. Define $M_m = \max_{1\leq i\leq m}\{Y_i\}$.*

(i) *If a_m is the smallest solution of*

$$\sum_{k=1}^{m} R_k(a_m) = 1 \ , \tag{10}$$

then

$$\mathbf{E}[M_m] \leq a_m + \sum_{k=1}^{m}\sum_{j=a_m}^{\infty} R_k(j) \ .$$

(ii) *If distribution functions $F_i(y)$ of $Y_1,\ Y_2,\dots,Y_m$ satisfying for all $1 \leq i \leq m$ the following two conditions*

$$F_i(y) < 1 \qquad \text{for all} \quad y < \infty \ ,$$

and

$$\lim_{y\to\infty} \sup_i \frac{1 - F_i(cy)}{1 - F_i(y)} = 0 \qquad \text{for} \quad c > 1 \ , \tag{11}$$

then $M_m/a_m \leq 1$ in probability (pr.), that is, for any $\varepsilon > 0$

$$\lim_{m\to\infty} \mathsf{Pr}\{M_m \geq (1+\varepsilon)a_m\} = 0 \ ,$$

where a_m solves asymptotically (10).

(iii) *If $Y_1,\dots,Y_m$ are independently and identically distributed (i.e., i.i.d.) with common distribution function $F(\cdot)$, then $M_m \sim a_m$ (pr.), that is, for any $\varepsilon > 0$*

$$\lim_{m\to\infty} \mathsf{Pr}\{(1-\varepsilon)a_m \leq M_m \leq (1+\varepsilon)a_m\} = 1 \ ,$$

where a_m is a solution of (10) which in this case becomes $mR(a_m) = 1$.

Proof. For (i) observe that, for any a,

$$M_m \leq a + \sum_{k=1}^{m} [Y_k - a]^+$$

where t^+ denotes $\max\{0,\ t\}$. Since $[Y_k - a]^+$ is a nonnegative random variable, then $\mathbf{E}[Y_k - a]^+ = \int_a^\infty R_k(y)dy$, so that (assuming for simplicity that Y_i is a continuous random variable)

$$\mathbf{E}[M_m] \leq a + \sum_{k=1}^{m} \int_a^\infty R_k(x)dx \ .$$

Minimizing the right-hand side of the above with respect to a yields part (i). For part (ii) we apply inequality (6) to get

$$\Pr\{M_m > x\} = \Pr\{Y_1 > x, \text{ or }, ..., \text{ or }, Y_m > x\} \le \sum_{i=1}^{m} R_i(x) \ .$$

Let $x = (1+\varepsilon)a_m$ where $\varepsilon > 0$ and a_m be defined by (10). Then, (11) with $c = 1+\varepsilon$ implies $R((1+\varepsilon)a_m) = o(1)R(a_m)$, hence

$$\Pr\{M_m \ge (1+\varepsilon)a_m\} = o(1)\sum_{i=1}^{m} R(a_m) = o(1) \ .$$

Finally, part (iii) can be proved using the additional assumption about the independence of $Y_1, \ldots, Y_m$. ■

In certain applications, we are interested not only in the maximum value but the r-th maximum value of the sequence $Y_1, \ldots, Y_m$. Let $\min_{1\le i\le m}\{Y_i\} = Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(m)} = \max_{1\le i\le m}$, and we call $Y_{(r)}$ the rth order statistics of $Y_1, \ldots, Y_m$. Below, we present a simple asymptotic result concerning the behavior of $Y_{(r)}$ for the so called *exchangeable* random variables. A sequence $\{Y_i\}_{i=1}^m$ is exchangeable if for any k-tuple $\{j_1, \ldots, j_k\}$ of the index set $\{1, \ldots, m\}$ the following holds: $\Pr\{Y_{j_1} < y_{j_1}, \ldots, Y_{j_k} < y_{j_k}\} = \Pr\{Y_1 < y_1, \ldots, Y_k < y_k\}$, that is, the joint distribution depends only on the *number* of variables (cf. [39]).

Lemma 3 *Let $\{Y_i\}_{i=1}^m$ be exchangeable random variables, and let $R^{(r)}(x) = \Pr\{Y_1 > x, \ldots, Y_r > x\}$ satisfies (11). Define $a_m^{(r)}$ as the smallest solution of*

$$\binom{m}{r-1} R^{(m-r+1)}(a_m^{(r)}) = 1 \ .$$

Then, $M_{(r)} \le a_m^{(r)}$ (pr.). If, in addition, Y_i are i.i.d., then $M_{(r)} \sim a_m^{(r)}$ (pr.) for $m \to \infty$.

Proof. One should apply inequality (6) to $\Pr\{M_m^{(r)} > x\} = \Pr\{\bigcup_{j_1,..,j_{m-r+1}} \bigcap_{i=1}^{m-r+1}(Y_{j_i} > y)\}$ for all distinct $j_1, ..., j_{m-r+1} \in \{1, ..., m\}$. ■

4.3 Subadditive Ergodic Theorem

The celebrated *ergodic theorem* of Birkhoff [25, 26] found many useful applications in computer science. It is used habitually during a computer simulation run or whenever one must perform experiments and collect data. However, for probabilistic analysis of algorithms a generalization of this result due to Kingman [59] is more important. We briefly review it here and illustrate on a few examples.

Let us start with the following well known fact: Assume a (deterministic) sequence $\{x_n\}_{n=0}^{\infty}$ satisfies the so called *subadditivity property*, that is,

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

for all integers $m, n \geq 0$. It is easy to see that then (cf. [24])

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

for some $\alpha \in [-\infty, \infty)$. Indeed, it suffices to fix $m \geq 0$, write $n = km + l$ for some $0 \leq l \leq m$, and observe that by the above subadditivity property

$$x_n \leq kx_m + x_l \ .$$

Taking $n \to \infty$ with $n/k \to m$ we finally arrive at

$$\limsup_{n\to\infty} \frac{x_n}{n} \leq \frac{x_m}{m} \leq \alpha \ ,$$

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

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

is automatic. One can also see that replacing "$\leq$" in the subadditivity property by "$\geq$" (thus, *superadditivity property*) will not change our conclusion except that $\inf_{m\geq 1} \frac{x_m}{m}$ should be replaced by $\sup_{m\geq 1} \frac{x_m}{m}$.

In the early seventies people started asking whether the above deterministic subadditivity result could be extended 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 [16] used ingenious tricks to establish the probabilistic behavior of the *Longest Common Superstring* problem (cf. Section 2.2 and below) while we show below that it is a trivial consequence of a stochastic extension of the above subadditivity result. In 1976 Kingman [59] presented the first proof of what later will be called *Subadditivity Ergodic Theorem.* Below, we present an extension of Kingman's result.

To formulate it properly we must consider a sequence of of doubly-indexed random variables $X_{m,n}$ with $m \leq n$. One can think of it as $X_{m,n} = (X_m, X_{m+1}, \ldots, X_n)$, that is, as a substring of a single-indexed sequence X_n.

Theorem 3 (Subadditive Ergodic Theorem) (i) *Let $X_{m,n}$ $(m < n)$ be a sequence of non-negative random variables satisfying the following three properties*

(a) $X_{0,n} \leq X_{0,m} + X_{m,n}$ *(subadditivity);*

(b) *$X_{m,n}$ is stationary (i.e., the joint distributions of $X_{m,n}$ are the same as $X_{m+1,n+1}$) and ergodic (cf. [11]);*

(c) $\mathbf{E}[X_{0,1}] < \infty$.

Then,

$$\lim_{n\to\infty} \frac{\mathbf{E}[X_{0,n}]}{n} = \gamma \qquad \text{and} \qquad \lim_{n\to\infty} \frac{X_{0,n}}{n} = \gamma \quad \text{(a.s.)} \tag{12}$$

for some constant γ.

(ii) (**Almost Subadditive Ergodic Theorem [21]**) *If the subadditivity inequality is replaced by*

$$X_{0,n} \le X_{0,m} + X_{m,n} + A_n \tag{13}$$

such that $\lim_{n\to\infty} \mathbf{E}[A_n/n] = 0$, then (12) holds, too.

We must point out, however, that the above result proves only the existence of a constant γ such that (12) holds. It says *nothing* how to compute it, and in fact many ingenious methods have been devised in the past to bound this constant. We discuss it in a more detailed way in the examples below.

Example 6: *String Editing Problem*

Let us consider the string editing problem of Section 2.2. To recall, one is interested in estimating the minimum $C_{\min}$ or the maximum cost $C_{\max}$ of transforming one sequence into another. In a particular case (i.e., *Longest Common Superstring* problem) one selects the longest common superstring of two given strings. As mentioned in Section 2.2 this problem can be reduced to finding the longest (shortest) path in a special grid graph (cf. Figure 2). Let us assume that the weights W_I, W_D and W_Q are independently distributed, thus we adopt the Bernoulli model (B) of Section 3.1. Then, using the subadditive ergodic theorem it is easy to prove that for some constant $\alpha > 0$

$$\lim_{n\to\infty} \frac{C_{\max}}{n} = \lim_{n\to\infty} \frac{EC_{\max}}{n} = \alpha \qquad \text{(a.s.)} \ ,$$

provided ℓ/s has a limit as $n \to \infty$. Indeed, let us consider the $\ell \times s$ grid with starting point O and ending point E (cf. Figure 2). 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, $C_{\max}(O,E) \ge C_{\max}(O,A) + C_{\max}(A,E)$. Thus, under our assumption regarding weights, the objective function $C_{\max}$ is superadditive, and direct application of the *Superadditive Ergodic Theorem* proves the result.

We should observe, however, that the above result does not tell us how to compute α. In fact, even in the simplest case of the Longest Common Superstring problem the constant α is unknown, but there are some bounds on α (cf. [16, 83]). □

4.4 Entropy and Its Applications

Entropy and *mutual information* was introduced by Shannon in 1948, and over a night a new field of *information theory* was born. Over the last fifty years information theory underwent many changes, and remarkable progress was achieved. These days entropy and the **Shannon-McMillan-Breiman Theorem** are standard tools of the average case analysis of algorithms. In this subsection, we review some elements of information theory and illustrate its usage to the analysis of algorithms.

Let us start with a simple observation: Consider a binary sequence of symbols of length n, say $(X_1, \ldots, X_n)$, with p denoting the probability of one symbol and $q = 1 - p$ the probability of the other symbol. When $p = q = 1/2$, then $\Pr\{X_1, \ldots, X_n\} = 2^{-n}$ and it does not matter what are the actual values of $X_1, \ldots, X_n$. In general, $\Pr\{X_1, \ldots, X_n\}$ is not the same for all possible values of $X_1, \ldots, X_n$, however, we shall show that a **typical** sequences $(X_1, \ldots, X_n)$ have "asymptotically" the same probability. Indeed, consider $p \neq q$ in the example above. Then, a typical sequence has the probability of its occurrence (we use here the central limit theorem for i.i.d. sequences):

$$p^{np+O(\sqrt{n})}q^{nq+O(\sqrt{n})} = e^{-n(-p\log p - q\log q)+O(\sqrt{n})} \sim e^{-nh}$$

where $h = -p\log p - q\log q$ is the *entropy* of the underlying Bernoulli model. Thus, a typical sequence X_1^n has asymptotically the same probability equal to e^{-nh}.

To be more precise, let us consider a *stationary and ergodic sequence* $\{X_k\}_{k=1}^{\infty}$ (cf. Section 3.1), and define $X_m^n = (X_m, X_{m+1}, \ldots, X_n)$ for $m \leq n$ as a substring of $\{X_k\}_{k=1}^{\infty}$. The entropy h of $\{X_k\}_{k=1}^{\infty}$ is defined as (cf. [11, 18, 24])

$$h := -\lim_{n\to\infty} \frac{\mathbf{E}[\log \Pr\{X_1^n\}]}{n} , \tag{14}$$

where one can prove the limt above exists. We must point out that $\Pr\{X_1^n\}$ is a *random variable* since X_1^n is a random sequence!

We show now how to derive the Shannon-McMillan-Breiman theorem in the case of the Bernoulli model and the mixing model, and later we formulate the theorem in its full generality. Consider first the Bernoulli model, and let $\{X_k\}$ be generated by a Bernoulli source. Thus

$$\begin{aligned} -\frac{\log \Pr\{X_1^n\}}{n} &= -\frac{1}{n}\sum_{i=1}^{n} \log \Pr\{X_i\} \\ &\to \mathbf{E}[-\log \Pr\{X_1\}] = h \qquad \text{(a.s.)} , \end{aligned}$$

where the last implication follows from the *Strong Law of Large Numbers* (cf. [11]) applied to the sequence $(-\log \Pr\{X_1\}, \ldots, -\log \Pr\{X_n\})$. One *should* notice a difference between the definition of the entropy (14) and the result above. In (14) we take the *average* of $\log \Pr\{X_1^n\}$

while in the above we proved that *almost surely* for all but finitely sequences the probability $\Pr\{X_1^n\}$ can be closely approximated by e^{-nh}. For the Bernoulli model, we have already seen it above, but we are aiming at showing that the above conclusion is true for much more general probabilistic models.

As the next step, let us consider the mixing model (MX) (that includes as a special case the Markovian model (M)). For the mixing model the following is true:

$$\Pr\{X_1^{n+m}\} \leq c\Pr\{X_1^n\}\Pr\{X_{n+1}^{n+m}\}$$

for some constant $c > 0$ and any integers $n, m \geq 0$. Taking logarithm we obtain

$$\log \Pr\{X_1^{n+m}\} \leq \log \Pr\{X_1^n\} + \log \Pr\{X_{n+1}^{n+m}\} + \log c$$

which satisfies the subadditivity property (13) of the *Subadditive Ergodic Theorem* discussed in Subsection 4.3. Thus, by (12) we have

$$h = -\lim_{n\to\infty} \frac{\log \Pr\{X_1^n\}}{n} \qquad \text{(a.s.)} \ .$$

Again, the reader should notice the difference between this result and the definition of the entropy.

We are finally ready to state the Shannon-McMillan-Breiman in its full generality (cf. [11, 24]).

Theorem 4 (Shannon-McMillan-Breiman) *For a stationary and ergodic sequence* $\{X_k\}_{k=-\infty}^{\infty}$ *the following holds*

$$h = -\lim_{n\to\infty} \frac{\log \Pr\{X_1^n\}}{n} \qquad \text{(a.s.)} \ .$$

where h *is the entropy of the process* $\{X_k\}$.

An important conclusion of this result is the so called **Asymptotic Equipartition Property** (AEP) which basically asserts that asymptotically all sequences have the same probability approximately equal to e^{-nh}. More precisely:

> *For a stationary and ergodic sequence* X_1^n, *the state space* Σ^n *can be partitioned into two subsets* $\mathcal{B}_n^\varepsilon$ *("bad set") and* $\mathcal{G}_n^\varepsilon$ *("good set") such that for given* $\varepsilon > 0$ *there is* N_ε *so that for* $n \geq N_\varepsilon$ *we have* $\Pr\{\mathcal{B}_n^\varepsilon\} \leq \varepsilon$, *and* $e^{-nh(1+\varepsilon)} \leq \Pr\{x_1^n\} \leq e^{-nh(1-\varepsilon)}$ *for all* $x_1^n \in \mathcal{G}_n^\varepsilon$.

Example 7: *Shortest Common Superstring or Depth in a Trie/Suffix Tree*

For concreteness let us consider the Shortest Common Superstring discussed in Section 2.3, but the same arguments as below can be used to derive the depth in a trie (cf. [79]) or a suffix tree (cf. [91]). Define C_{ij} as the length of the longest suffix of X_i that is equal to the prefix of X_j. Let

$$M_n(i) = \max_{1 \le j \le n, j \ne i} \{C_{ij}\} .$$

We write M_n for a generic random variable distributed as $M_n(i)$ (observe that $M_n \stackrel{d}{\to} M_n(i)$ for all i, where $\stackrel{d}{\to}$ means "equal in distribution"). We would like to prove that in the mixing model, for any $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr\left\{(1-\varepsilon)\frac{1}{h}\log n \le M_n \le (1+\varepsilon)\frac{1}{h}\log n\right\} = 1 - O(1/n^{\varepsilon})$$

provided $\alpha(g) \to 0$ as $g \to \infty$, that is, $M_n/\log n \to h$ (pr.). To prove an upper bound, we take *any fixed* typical sequence $w_k \in \mathcal{G}_k^{\varepsilon}$ as defined in AEP above, and observe that

$$\Pr\{M_n \ge k\} \le n\Pr\{w_k\} + \Pr\{\mathcal{B}_k\}.$$

The result follows immediately after substituting $k = (1+\varepsilon)h^{-1}\log n$ and noting that $\Pr\{w_k\} \le e^{nh(1-\varepsilon)}$. For a lower bound, let $w_k \in \mathcal{G}_k^{\varepsilon}$ be any fixed typical sequence with $k = \frac{1}{h}(1-\varepsilon)\log n$. Define Z_k as the number of strings $j \ne i$ such that a prefix of length k is equal to w_k *and* a suffix of length k of the ith string is equal to $w_k \in \mathcal{G}_k$. Since w_k is fixed, the random variables C_{ij} are independent, and hence by the *second moment method* (cf. Section 4.2)

$$\Pr\{M_n < k\} = \Pr\{Z_k = 0\} \le \frac{\mathbf{Var} Z_k}{(\mathbf{E} Z_k)^2} \le \frac{1}{n\Pr\{w_k\}} = O(n^{-\varepsilon^2}) ,$$

since $\mathbf{Var} Z_k \le nP(w_k)$, and this completes the derivation. □

In many problems on words another kind of entropy is widely used (cf. [7, 8, 9, 91]). It is called **Rényi entropy** and defined as follows: For $-\infty \le b \le \infty$, the bth *order Rényi entropy* is

$$h_b = \lim_{n\to\infty} \frac{-\log(\mathbf{E}[\Pr\{X_1^n\}^{b-1}])}{bn} = \lim_{n\to\infty} \frac{-\log\left(\sum_{w\in\Sigma^n} (\Pr\{w\})^b\right)^{-1/b}}{n} , \tag{15}$$

provided the above limit exists. In particular: by *inequalities on means* we obtain

$$\begin{aligned} h_0 &= h , \\ h_{-\infty} &= \lim_{n\to\infty} \frac{\max\{-\log \Pr\{X_1^n\} , \Pr\{X_1^n\} > 0\}}{n} , \\ h_{\infty} &= \lim_{n\to\infty} \frac{\min\{-\log \Pr\{X_1^n\} , \Pr\{X_1^n\} > 0\}}{n} . \end{aligned}$$

For example, the entropy $h_{-\infty}$ appears in the formulation of the shortest path in digital trees (cf. [79, 91]), the entropy h_∞ is responsible for the height in PATRICIA tries (cf. [79, 91]), while h_2 determines the height in a trie. Indeed, we claim that in a mixing model the height H_n in a trie behaves probabilistically as $H_n/\log n \to 2/h_2$. Consider first the Bernoulli model as in Examples 4 and 5. Using our definition of h_2 in the Bernoulli model one immediately proves that $h_2 = \log P^{-1} = \log Q$ where $P = p^2 + q^2$ as defined in Example 4. This confirms our observation. An extension to a mixing model follows the footsteps of our proof from Examples 4 and 5 and can be found in [79, 90, 91].

4.5 Central Limit and Large Deviations Results

Convergence of a sum of independent, identically distributed (i.i.d.) random variables is central to probability theory. In the analysis of algorithms, we mostly deal with *weakly dependent* random variables, but often results from the i.i.d. case can be extended to this new situation by some clever tricks. A more systematic treatment of such cases is usually done through generating functions and complex analysis techniques (cf. [32, 49, 50, 52, 53, 54, 73]) which we briefly discuss it in the next section. Hereafter, we concentrate on the i.i.d. case.

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 $\sigma^2 := \mathbf{Var}[X_1]$. We pay particular interest to another random variable, namely

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

which distribution function we denote as $F_n(x) = \Pr\{s_n \le x\}$. Let also $\Phi(x)$ 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 $F_n(\cdot)$, provided $\sigma < \infty$ (cf. [24, 26]). A stronger version is due to Berry-Esséen who proved that

$$|F_n(x) - \Phi(x)| \le \frac{2\rho}{\sigma^2\sqrt{n}} \tag{16}$$

where $\rho = \mathbf{E}[|X - \mu|^3] < \infty$. Finally, Feller [26] has shown that if centralized moments $\mu_2, \ldots, \mu_r$ exist, then

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

uniformly in x, where $R_k(x)$ is a polynomial depending only on $\mu_1, \ldots, \mu_r$ but not on n and r.

One should notice from the above, in particular from (16), the weakness of central limit results which are able only to assess the probability of *small deviations* from the mean. Indeed, the results above are true for $x = O(1)$ (i.e., for $S_n \in (\mu n - O(\sqrt{n}), \mu n + O(\sqrt{n}))$ due to only a polynomial rate of convergence as shown in (16). To see it more clearly, we quote a result from Greene and Knuth [41] who estimated

$$\Pr\{S_n = \mu n + r\} = \frac{1}{\sigma\sqrt{2\pi n}} \exp\left(\frac{-r^2}{2\sigma^2 n}\right) \left(1 - \frac{\kappa_3}{2\sigma^4}\left(\frac{r}{n}\right) + \frac{\kappa_3}{6\sigma^6}\left(\frac{r^3}{n^2}\right)\right) + O\left(n^{-\frac{3}{2}}\right) \tag{17}$$

where κ_3 is the third cumulant of X_1. Observe now that when $r = O(\sqrt{n})$ (which is equivalent to $x = O(1)$ in our previous formulæ) the error term *dominates* the leading term of the above asymptotic, thus the estimate is quite 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 $x_n \to \infty$ in the above formulæ. 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 $\delta \neq 0$. We shall discuss this large deviations behavior next.

Let us first try to "guess" a large deviation behavior of $S_n = X_1 + \cdots X_n$ for i.i.d. random variables. We estimate $\mathsf{Pr}\{S_n \geq an\}$ for $a > 1$ as $n \to \infty$. Observe that (cf. [24]

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

since S_m and $S_{n+m} - S_m$ are independent. Taking logarithm of the above, and recognizing that $\log \mathsf{Pr}\{S_n \geq an\}$ is a superadditive sequence (cf. Subsection 4.3), we obtain

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

where $I(a) \geq 0$. Thus, S_n decays *exponentially* when far away from its mean, not in a Gaussian way as the central limit theorem would predict! Unfortunately, we obtain the above result from the subadditive property which allowed us to conclude the existence of the above limit, but says nothing about $I(a)$.

In order to take a full advantage of the above derivation, we should say something about $I(a)$ and, more importantly, to show that $I(a) > 0$ under some mild conditions. For the latter. let us first assume that the *moment generating function*

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

Let also $\kappa(\lambda) = \log M(\lambda)$ be the *cumulant function* of X_1. Then, by Markov's inequality (cf. Subsection 4.2)

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

Actually, due to arbitrariness of λ subject to $\lambda > 0$, we finally arrive at the so called **Chernoff bound**, that is,

$$\mathsf{Pr}\{S_n \geq na\} \leq \min_{\lambda>0} \left\{ e^{-\lambda na}\mathbf{E}[e^{\lambda S_n}] \right\} . \tag{18}$$

We should emphasize that the above bound is true for *dependent* random variables since we only used Markov's inequality applied to S_n.

Returning to the i.i.d. case, we can rewrite the above as

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

But, under mild conditions the above minimization problem is easy to solve. One finds that the minimum is attended at λ_a which satisfies $a = M'(\lambda_a)/M(\lambda_a)$. Thus, we proved that $I(a) \geq a\lambda_a - \log M(\lambda_a)$. However, a careful evaluation of the above leads to the following classical large deviations result (cf. [24])

Theorem 5 *Assume $X_1, \ldots, X_n$ are i.i.d. Let $M(\lambda) = \mathbf{E}[e^{\lambda X_1}] < \infty$ for some $\lambda > 0$, the distribution of X_i is not a point mass at μ, and there exists $\lambda_a > 0$ in the domain of the definition of $M(\lambda)$ such that*

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

Then:

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

for $a > \mu$.

A major strengthening of this theorem is due to Gärtner and Ellis 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 known (cf. [22]):

Theorem 6 (Gärtner-Ellis) *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

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

Let us return again to the i.i.d. case and see if we can strengthen Theorem 5 which in its present form gives only a logarithmic limit. We explain our approach on a simple example, following Greene and Knuth [41]. Let us assume that $X_1, \ldots, X_n$ are discrete i.i.d. with common generating function $G(z) = \mathbf{E}[z^X]$. We recall that $[z^m]G(z)$ denote the coefficient at z^m of $G(z)$. In (17) we show how to compute such a coefficient at $m = \mu n + O(\sqrt{n})$ of $G^n(z) = Ez^{S_n}$. We observed also that (17) cannot be used for large deviations since the error term was dominating the leading term in such a case. But, one may shift the mean of S_n to a new value such that (17) is valid again. Thus, let us define a new random variable $\widetilde{X}$ whose generating function is

$$\widetilde{G}(z) = \frac{G(z\alpha)}{G(\alpha)}$$

where α is a constant that is to be determined. Observe that $\mathbf{E}[\widetilde{X}] = \widetilde{G}'(1) = \alpha G'(\alpha)/G(\alpha)$. Assume one needs large deviations result around $m = n(\mu + \delta)$ where $\delta > 0$. Clearly, (17) cannot be applied directly. Now, a proper choice of α can help. Let us select α such that the new $\widetilde{S}_n = \widetilde{X}_1 + \cdots + \widetilde{X}_n$ has mean $m = n(\mu + \delta)$. This results in setting α to be a solution of

$$\frac{\alpha G'(\alpha)}{G(\alpha)} = \frac{m}{n} = \mu + \delta \ .$$

In addition, we have the following obvious identity

$$[z^m]G^n(z) = \frac{G^n(\alpha)}{\alpha^m}[z^m]\left(\frac{G(\alpha z)}{G(\alpha)}\right)^n \ . \tag{19}$$

But, now we can use (17) to the right-hand side of the above since the new random variable $\widetilde{S}_n$ has mean around m.

To illustrate the above technique that is called **shift of mean** we present an example.

Example 8: *Large Deviations by "Shift of Mean"* (cf. [41]).

Let S_n be be binomially distributed with parameter 1/2, that is, $G^n(z) = ((1+z)/2)^n$. We want to estimate the probability $\Pr\{S_n = n/3\}$ which is far away from its mean ($ES_n = n/2$) and central limit result (17) cannot be applied. We apply shift of mean, and compute α as

$$\frac{\alpha G'(\alpha)}{G(\alpha)} = \frac{\alpha}{1+\alpha} = \frac{1}{3} \ ,$$

thus, $\alpha = 1/2$. Using (17) we obtain

$$[z^{n/3}]\left(\frac{2}{3} + \frac{1}{3}z\right)^n = \frac{3}{2\sqrt{\pi n}}\left(1 - \frac{7}{24n}\right) + O(n^{-5/2}) \ .$$

To obtain the result we want (i.e., coefficient at $z^{n/3}$ of $(z/2 + 1/2)^n$), one must apply (18). This finally leads to

$$[z^{n/3}](z/2 + 1/2)^n = \left(\frac{3 \cdot 2^{1/3}}{4}\right)^n \frac{3}{2\sqrt{\pi n}}\left(1 - \frac{7}{24n} + O(n^{-2})\right)$$

which is a large deviations result (the reader should observe the exponential decay of the above probability). □

The last example showed that one may expect a stronger large deviation result than the one presented in Theorem 5. Indeed, under proper mild conditions it can be proved that Theorem 5 extends to (cf. [22])

$$\Pr\{S_n \geq na\} \sim \frac{1}{\sqrt{2\pi n}\sigma_a \lambda_a} \exp(-nI(a))$$

for a constant σ_a, and λ_a and $I(a) = a\lambda_a - \log M(\lambda_a)$ defined as in Theorem 5.

Finally, we deal with an interesting extension of the above large deviations results initiated by Azuma, and recently significantly extended by Talagrand [93]. These results are known in the literature under the name **Azuma's type inequality** or **method of bounded differences** (cf. [74]). It can be formulated as follows:

Theorem 7 (Azuma's type Inequality) *Let X_i be i.i.d. random variables such that for some function $f(\cdot, \ldots, \cdot)$ the following is true*

$$|f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X_i', \ldots, X_n)| \leq c_i \ , \tag{20}$$

where $c_i < \infty$ are constants, and X_i' has the same distribution as X_i. Then,

$$\Pr\{|f(X_1, \ldots, X_n) - Ef(X_1, \ldots, X_n)| \geq t\} \leq 2\exp(-2t^2/\sum_{i=1}^{n} c_i^2) \tag{21}$$

for some $t > 0$.

We finish this long subsection, and the whole Section 4, with an application of the Azuma inequality (cf. [71]):

Example 9: *Concentration of Mean for the Editing Problem*

Let us consider again the editing problem from Section 2.2. The following is true:

$$\Pr\{|C_{\max} - EC_{\max}| > \varepsilon EC_{\max}\} \leq 2\exp(-\varepsilon^2 \alpha n) \ .$$

provided all weights are bounded random variables, say $\max\{W_I, W_D, W_Q\} \leq 1$. Indeed, under the Bernoulli model, the X_i are i.i.d. (where X_i, $1 \leq i \leq n = \ell + s$, represents symbols of the two underlying sequences), and therefore (20) holds with $f(\cdot) = C_{\max}$. More precisely,

$$|C_{\max}(X_1, \ldots, X_i, \ldots, X_n) - C_{\max}(X_1, \ldots, X_i', \ldots, X_n)| \leq \max_{1 \leq i \leq n}\{W_{\max}(i)\} \ .$$

where $W_{max}(i) = \max\{W_I(i), W_D(i), W_Q(i)\}$. Setting $c_i = 1$ and $t = \varepsilon EC_{\max} = O(n)$ in the Azuma inequality we obtain the desired result. $\square$

5 Analytical Techniques

Analytical (or precise) analysis of algorithms was initiated by Knuth almost thirty years ago in his *magnum opus* [64, 65, 66] who treated many aspects of fundamental algorithms, seminumerical algorithms, or sorting and searching. A modern introduction to analytical methods can be found in a marvelous book [84] by Sedgewick and Flajolet, while advanced analytical techniques are covered in a forthcoming book *Analytical Combinatorics* by Flajolet and Sedgewick. In this section, we only touch "a tip of an iceberg" and briefly discuss functional equations arising in the analysis of digital trees, complex asymptotics techniques, Mellin transform, and analytical depoissonization.

5.1 Recurrences and Functional Equations

Recurrences and functional equations are widely used in computer science. For example, the divide-and-conquer recurrence equations (cf. Chapter 1) appear in the analysis of searching and sorting algorithms (cf. [66]). Hereafter, we concentrate on recurrences and functional equations that arise in the analysis of digital trees and problems on words.

However, to introduce the reader into the main subject we first consider two well known functional equations that should be in a "knapsack" of every computer scientist. Let us enumerate the number of *unlabeled binary trees* built over n vertices. Call this number b_n, and let $B(z) = \sum_{n=0}^{\infty} b_n z^n$ be its *ordinary generating function.* Since each such tree is constructed in a recursive manner with left and right subtrees being unlabeled binary trees, we immediately arrive at the following recurrence for $n \geq 1$

$$b_n = b_0 b_{n-1} + \cdots + b_{n-1} b_0$$

with $b_0 = 1$ by definition. Multiplying by z^n and summing from $n = 1$ to infinity, we obtain $B(z) - 1 = zB^2(z)$ which is a simple functional equation that can be solved to find

$$B(z) = \frac{1 - \sqrt{1-4z}}{2z} .$$

To derive the above functional equation, we used a simple fact that the generating function $C(z)$ of the convolution c_n of two sequences, say a_n and b_n (i.e., $c_n = a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0$), is the product of $A(z)$ and $B(z)$, that is, $C(z) = A(z)B(z)$.

The above functional equation and its solution can be used to obtain an explicit formula on b_n. Indeed, we first recall that $[z^n]B(z)$ denotes the coefficient at z^n of $B(z)$ (i.e., b_n). A standard analysis leads to (cf. [64, 73])

$$b_n = [z^n]B(z) = \frac{1}{n+1}\binom{2n}{n} ,$$

which is the famous *Catalan number.*

Let us now consider a more challenging example, namely, enumeration of *rooted labeled trees.* Let t_n the number of rooted labeled trees, and $t(z) = \sum_{n=0}^{\infty} \frac{t_n}{n!} z^n$ its *exponential generating function.* It is known that $t(z)$ satisfies the following functional equation (cf. [46, 84, 98])

$$t(z) = ze^{t(z)} .$$

The easiest way of finding t_n, which is the coefficient at z^n, is by *Lagrange's Inversion Formula.* Let $\Phi(u)$ be a formal power series with $[u^0]\Phi(u) \neq 0$, and let $X(z)$ be a solution of $X = z\Phi(X)$. The coefficients of $X(z)$ or in general $\Psi(X(z))$ where Ψ is an arbitrary series can be found by

$$\begin{aligned} [z^n]X(z) &= \frac{1}{n}[u^{n-1}]\,(\Phi(u))^n , \\ [z^n]\Psi(X(z)) &= \frac{1}{n}[u^{n-1}]\,(\Phi(u))^n\,\Psi'(u) . \end{aligned}$$

In particular, an application of the above to $t(z)$ leads to $t_n = n^{n-1}$, and to an interesting formula (which we encounter again in Example 14)

$$t(z) = \sum_{n=1}^{\infty} \frac{n^{n-1}}{n!} z^n \tag{22}$$

where $T(z) = ze^{T(z)}$.

After these introductory remarks, we can now concentrate on certain recurrences that arise in problems on words; in particular in digital trees and shortest common superstring problems. Let x_n be a generic notation for a quantity of interest (e.g., depth, size or path length in a digital tree built over n strings). Given x_0 and x_1, the following three recurrences originate from problems on tries, PATRICIA tries and digital search trees, respectively (cf. [28, 31, 34, 46, 51, 52, 60, 61, 62, 66, 72, 73, 84, 86, 86, 87, 88, 89]):

$$x_n = a_n + \beta \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (x_k + x_{n-k}) , \qquad n \geq 2 \tag{23}$$

$$x_n = a_n + \beta \sum_{k=1}^{n-1} \binom{n}{k} p^k q^{n-k} (x_k + x_{n-k}) - \alpha(p^n + q^n) x_n , \qquad n \geq 2 \tag{24}$$

$$x_{n+1} = a_n + \beta \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (x_k + x_{n-k}) \qquad n \geq 0 \tag{25}$$

where a_n is a known sequence (also called additive term), α and β are some constants, and finally $p + q = 1$.

To solve this recurrences and to obtain explicit or asymptotic expression for x_n we apply exponential generating functions. We need to know the following two obvious facts: Let a_n and b_n be sequences with $a(z) = \sum_{n=0}^{\infty} \frac{a_n}{n!} z^n$ and $b(z)$ as their exponential generating functions. (Hereafter, we consequently use lower-case letters for exponential generating functions, like $a(z)$, and upper-case letters for ordinary generating functions, like $A(z)$). Then:

- For any integer $h \geq 0$
$$\frac{d^h}{dz^h} a(z) = \sum_{n=0}^{\infty} \frac{a_{n+h}}{n!} z^n .$$

- If $c_n = \sum_{r=0}^{n} \binom{n}{r} a_r b_{n-r}$, then the exponential generating function $c(z)$ of c_n becomes
$$c(z) = a(z) b(z) .$$

Now, we are ready to attack the above recurrences and show how they can be solved. Let us start with the simplest one, namely (23). Multiplying it by z^n, summing up, and taking into account the initial conditions we obtain

$$x(z) = a(z) + \beta x(zp) e^{zq} + \beta x(zq) e^{zp} + d(z) \tag{26}$$

where $d(z) = d_0 + d_1 z$ and d_0 and d_1 depend on the initial condition for $n = 0, 1$. The trick is to introduce the so called **Poisson transform** $\widetilde{X}(z) = x(z)e^{-z}$ which reduces the above functional equation to

$$\widetilde{X}(z) = \widetilde{A}(z) + \beta\widetilde{X}(zp) + \beta\widetilde{X}(z) + d(z)e^{-z} . \tag{27}$$

Observe that $\widetilde{x}_n$ and x_n are related by $x_n = \sum_{k=0}^{n} \binom{n}{k}\widetilde{x}_k$. Using this, and comparing coefficients of $\widetilde{X}(z)$ at z^n we finally obtain

$$x_n = x_0 + n(x_1 - x_0) + \sum_{k=2}^{n}(-1)^k \binom{n}{k} \frac{\hat{a}_k + kd_1 - d_0}{1 - \beta(p^k + q^k)} , \tag{28}$$

where $n![z^n]\widetilde{A}(z) = \widetilde{a}_n := (-1)^n \hat{a}_n$. In fact, $\hat{a}_n$ and a_n form the so called *binomial inverse relations*, and

$$\hat{a}_n = \sum_{k=0}^{n} \binom{n}{k}(-1)^k a_k , \qquad a_n = \sum_{k=0}^{n} \binom{n}{k}(-1)^k \hat{a}_k ,$$

that is, $\hat{\hat{a}}_n = a_n$ (cf. [66]).

Example 10: *Average Path Length in a Trie*

Let us consider a trie in the Bernoulli model, and estimate the average ℓ_n of the external path length, i.e., $\ell_n = \mathbf{E}[L_n]$ (cf. Section 2.1). Clearly, $\ell_0 = \ell_1 = 0$ and for $n \geq 2$

$$\ell_n = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}(\ell_k + \ell_{n-k}) .$$

Thus, by (28)

$$\ell_n = \sum_{k=2}^{n}(-1)^k \binom{n}{k} \frac{k}{1 - p^k - q^k} .$$

Below, we shall discuss the asymptotics of ℓ_n (cf. Example 15). □

Let us now consider recurrence (24) which is much more intricate. It has an exact solution only for some special cases (cf. [66, 86, 88]) that we discuss below. We first consider a simplied version of (24), namely

$$x_n(2^n - 2) = 2^n a_n + \sum_{k=1}^{n-1} \binom{n}{k} x_k$$

with $x_0 = x_1 = 0$ (for a more general recurrence of this type see [86]). After multiplying by z^n and summing up we arrive at

$$x(z) = (e^{z/2} + 1)x(z/2) + a(z) - a_0 \tag{29}$$

where $x(z)$ and $a(z)$ are exponential generating functions of x_n and a_n. To solve this recurrence we must observe that after multiplying both sides by $z/(e^z - 1)$ and defining

$$\check{X}(z) = x(z)\frac{z}{e^z - 1} \tag{30}$$

we obtain a new functional equation that is easy to solve, namely:

$$\check{X}(z) = \check{X}(z/2) + \check{A}(z)$$

where in the above we assume for simplicity $a_0 = 0$. This function equation is of similar type to $\widetilde{X}(x)$ considered above, and the coefficient $\check{x}_n$ at z^n of $\check{X}(z)$ can be easily extracted. One must, however, translate coefficients $\check{x}_n$ into the original sequence x_n. In order to accomplish this, let us introduce the Bernoulli polynomials $B_n(x)$ and Bernoulli numbers $B_n = B_n(0)$, that is, $B_n(x)$ are defined as

$$\frac{ze^{tz}}{e^z - 1} = \sum_{k=0}^{\infty} B_k(t) \frac{z^k}{k!} \ .$$

Furthermore, we introduce *Bernoulli inverse relations* for a sequence a_n as

$$\check{a}_n = \sum_{k=0}^{n} \binom{n}{k} B_k a_{n-k} \qquad \Longleftrightarrow \qquad a_n = \sum_{k=0}^{n} \binom{n}{k} \frac{\check{a}_k}{k+1} \ .$$

One should know that (cf. [66])

$$a_n = \binom{n}{r} q^n \qquad \Longleftrightarrow \qquad \check{a}_n = \binom{n}{r} q^r B_{n-r}(q)$$

for $0 < q < 1$. For example, for such a choice of a_n as above (i.e., $a_n = \binom{n}{r} q^n$) the above recurrence has a particular simply solution, namely:

$$x_n = \sum_{k=1}^{n} (-1)^k \binom{n}{k} \frac{B_{k+1}(1-q)}{k+1} \frac{1}{2^{k+1} - 1} \ .$$

A general solution to the above recurrence can be found in [86], and it involves $\check{a}_n$.

Example 11: *Unsuccessful Search in PATRICIA*

Let us consider the number of trials u_n in an unsuccessful search of a string in a PATRICIA trie constructed over the symmetric Bernoulli model (i.e., $p = q = 1/2$). As in Knuth [66] (cf. [88])

$$u_n(2^n - 2) = 2^n(1 - 2^{1-n}) + \sum_{k=1}^{n-1} \binom{n}{k} u_k$$

and $u_0 = u_1 = 0$. A simple application of the above derivation leads, after some algebra, to

$$u_n = 2 - \frac{4}{n+1} + 2\delta_{n0} + \frac{2}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k}{2^{k-1} - 1}$$

where $\delta_{n,k}$ is the Kronecker delta, that is, $\delta_{n,k} = 1$ for $n = k$ and zero otherwise. $\square$

We were able to solve the functional equations (26) and (29) exactly since we reduce them to a simple functional equation of the form (27). In particular, equation (29) became (27) since luckily $e^z - 1 = (e^{z/2} - 1)(e^{z/2} + 1)$, as already pointed out by Knuth [66], but one cannot expect that much luck with other functional equations. Let us consider a general functional equation

$$F(z) = a(z) + b(z)F(\sigma(z)) \tag{31}$$

where $a(z), b(z), \sigma(z)$ are known function. Formally, iterating this equation we obtain its solution as

$$F(z) = \sum_{k=0}^{\infty} a(\sigma^{(k)}(z)) \prod_{j=0}^{k-1} b(\sigma^{(j)}(z))$$

where $\sigma^{(k)}(z)$ is the kth iterate of $\sigma(\cdot)$. When applying the above to solve real problems, one must assure the existence of the infinite series involved (cf. [29]). In some cases (cf. [33, 63]), we can provide asymptotic solutions to such complicated formulæ by appealing to the Mellin transform which we discuss below in Subsection 5.3.

Finally, we deal with the recurrence (25). Multiplying by $z^n/n!$ and using the above discussed properties of exponential generating functions we obtain for $x(z) = \sum_{n \geq 0} x_n \frac{z^n}{n!}$

$$x'(z) = a(z) + x(zp)e^{zq} + x(zq)e^{zp} \ ,$$

which becomes after substitution $\widetilde{X}(z) = x(z)e^{-z}$

$$\widetilde{X}'(z) + \widetilde{X}(z) = \widetilde{A}(z) + \widetilde{X}(zp) + \widetilde{X}(zq) \ . \tag{32}$$

The above is a differential-functional equation that we did not discuss so far. It can be solved since a direct translation of coefficients gives: $\widetilde{x}_{n+1} + \widetilde{x}_n = \widetilde{a}_n + \widetilde{x}_n(p^n + q^n)$. Fortunately, this is a simple linear recurrence that has an explicit solution. Taking into account $x_n = \sum_{k=0}^{n} \binom{n}{k} \widetilde{x}_k$, we finally obtain

$$x_n = x_0 - \sum_{k=1}^{n} (-1)^k \binom{n}{k} \sum_{i=1}^{k-1} \hat{a}_i \prod_{j=i+1}^{k-1} (1 - p^j - q^j) = x_0 - \sum_{k=1}^{n} (-1)^k \binom{n}{k} \sum_{i=1}^{k-1} \hat{a}_i \frac{Q_k}{Q_i} \ , \tag{33}$$

where $Q_k = \prod_{j=2}^{k} (1 - p^j - q^j)$, and $\hat{a}_n$ is the binomial inverse of a_n as defined above. In passing, we should observe that solutions of the recurrences (23)-(25) have a form of an alternating sum, that is, $x_n = \sum_{k=1}^{n} (-1)^k \binom{n}{k} f_n$ where f_n has an explicit formula. In subsection 5.3, we discuss how to obtain asymptotics of such an alternating sum.

Example 12: *Expected Path Length in a Digital Search Tree*

Let ℓ_n be the expected path length in a digital search tree. Then (cf. [35, 66, 90]) for all $n \geq 0$

$$\ell_{n+1} = n + \sum_{k=1}^{n} \binom{n}{k} p^k q^{n-k} (\ell_k + \ell_{n-k})$$

with $\ell_0 = 0$. By (33) it has the following solution

$$\ell_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} Q_{k-1}$$

where Q_k is defined above. $\square$

We were quite lucky when solving the above differential-functional equation since we could reduce it to a linear recurrence of *first* order. However, this is not any longer true when we consider the so called b-digital search trees in which one assumes that a node of such a tree can store up to b strings. Then, the general recurrence (25) becomes

$$x_{n+b} = a_n + \beta \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (x_k + x_{n-k}) \qquad n \geq 0$$

provided $x_0, \ldots, x_{b-1}$ are given. Our previous approach would lead to a linear recurrence of order b that does not possess a nice explicit solution. The "culprit" lies in the fact that the exponential generating function of a sequence $\{x_{n+b}\}_{n=0}^{\infty}$ is the b-th derivative of the exponential generating function $x(z)$ of $\{x_n\}_{n=0}^{\infty}$. On the other hand, if one consider *ordinary* generating function $X(z) = \sum_{n \geq 0} x_n z^n$, then the sequence $\{x_{n+b}\}_{n=0}^{\infty}$ translates into $z^{-b}(X(z) - x_0 - \cdots - x_{b-1} z^{b-1})$. This observation led Flajolet and Richmond [31] to reconsider the standard approach to the above binomial recurrences, and to introduce ordinary generating function into the play. A careful reader observes, however, that then one must translate into ordinary generating functions sequences such as $s_n = \sum_{k=0}^{n} \binom{n}{k} a_k$ (which were easy under exponential generating functions since they become $a(z)e^z$). But, it is not difficult to see that

$$s_n = \sum_{k=0}^{n} \binom{n}{k} a_k \qquad \Longrightarrow \qquad S(z) = \frac{1}{1-z} A\left(\frac{z}{1-z}\right) \ .$$

Indeed,

$$\begin{aligned} \frac{1}{1-z} A\left(\frac{z}{1-z}\right) &= \sum_{m=0}^{\infty} a_m z^m \frac{1}{(1-z)^{m+1}} = \sum_{m=0}^{\infty} a_m z_m \sum_{j=0}^{\infty} \binom{m+j}{j} z^j \\ &= \sum_{n=0}^{\infty} z^n \sum_{k=0}^{\infty} \binom{n}{k} a_k \ . \end{aligned}$$

Thus, the above recurrence for $p = q = 1/2$ and any $b \geq 1$ can be translated into ordinary generating functions as

$$\begin{aligned} X(z) &= \frac{1}{1-z} G\left(\frac{z}{1-z}\right) \\ G(z)(1+z)^b &= 2z^b G(z/2) + P(z) \end{aligned}$$

$P(z)$ is a function of a_n and initial conditions. But, the latter functional equation falls under (31) and its solution is given above.

Finally(!) an interested reader may ask how to translate from exponential generating function back to ordinary generating functions. Let a_n be a sequence such that its ordinary generating function $A(z)$ exists, say in a unit disk. Let also $a(z)$ denote its exponential generating function. Then, **Borel transform** (cf. [94, 99]) asserts that

$$A(z) = \int_0^\infty e^{-t} a(zt) dt$$

at least for $|z| \le 1$. The above is rather easy to understand. Formally, it suffices to develop $a(zt)$ into Taylor's series, and integrate term by term (noting that $\int_0^\infty e^{-t} t^n dt = n!$).

5.2 Complex Asymptotics

When analyzing an algorithm we often aim at predicting its rate of growth of time or space complexity for large inputs, n. Precise analysis of algorithms launched by Knuth [64, 65, 66] aims at obtaining precise asymptotics of some performance measure of an algorithm. For example, in the previous subsection we studied some parameters of tries (e.g., path length ℓ_n, unsuccessful search u_n, etc.) that depend on input of size n. We observed that these quantities are expressed by some complicated alternating sums (cf. Examples 10-12). One might be interested in precise rate of growth of these quantities. More precisely, if x_n represents a quantity of interest with input size n, we may look for a simple explicit function a_n (e.g., $a_n = \log n$ or $a_n = \sqrt{n}$) such that $x_n \sim a_n$ (i.e., $\lim_{n\to\infty} x_n/a_n = 1$) or we may be aiming at a very precise asymptotic expansion such as $x_n = a_n^1 + a_n^2 + \cdots + o(a_n^k)$ where for each $1 \le i \le k$ we have $a_n^{i+1} = o(a_n^i)$.

The reader is referred to an excellent recent survey by Odlyzko [78] on asymptotic methods. In this subsection, we briefly discuss some elementary facts of asymptotic evaluation, and describe a few useful methods.

It is well recognized that complex analysis through generating functions provides the most powerful approach to deal with asymptotic evaluation of a sequence $\{a_n\}_{n=0}^\infty$. Let $A(z) = \sum_{n=0}^\infty a_n z^n$ be its generating function. In the previous subsection, we look at $A(z)$ as a *formal power series*. Now, we ask whether $A(z)$ converges, and what is its region of convergence. It turns out that the radius of convergence for $A(z)$ is responsible for the asymptotic behavior of a_n for large n. Indeed, by *Hadamard's Theorem* [45, 94] we know that radius R of convergence of $A(z)$ (where z is a complex variable) is given by

$$R = \frac{1}{\limsup_{n\to\infty} |a_n|^{1/n}} .$$

In other words, for every $\varepsilon > 0$ there exists N such that for $n > N$ we have

$$|a_n| \le (R^{-1} + \varepsilon)^n ;$$

and for infinitely many n we have

$$|a_n| \geq (R^{-1} - \varepsilon)^n \ .$$

Informally saying, $\frac{1}{n} \log |a_n| \sim 1/R$; or even less formally the exponential growth of a_n is determined by $(1/R)^n$. In summary, singularities of $A(z)$ determine asymptotic behavior of its coefficients for large n. In fact, formally from *Cauchy's Integral Theorem* (cf. Section 3.3) we know that

$$|a_n| \leq \frac{M(r)}{r^n}$$

where $M(r)$ is the maximum value of $|A(z)|$ for circle $r < R$.

Our goal now is to make a little more formal our discussion above, and deal with multiple singularities. We restrict ourselves to *meromorphic* functions $A(z)$, i.e., ones that are analytical with the exception of a finite number of *poles*. To make our discussion more concrete we study the following function (cf. [98])

$$A(z) = \sum_{j=1}^{r} \frac{a_{-j}}{(z-\rho)^j} + \sum_{j=0}^{\infty} a_j (z-\rho)^j \ .$$

More precisely, we assume that $A(z)$ has the above *Laurent expansion* around a pole ρ of multiplicity r. Let us further assume that the pole ρ is the closest to the origin, that is, $R = |\rho|$ (and there are no more poles on the circle of convergence). In other words, the sum of $A(z)$ which we denote for simplicity as $A_1(z)$, is analytical in the circle $|z| \leq |\rho|$, and its possible radius of convergence $R' > |\rho|$. Thus, coefficients a'_n of $A_1(n)$ are bounded by $|a'_n| = O((1/R' + \varepsilon)^n)$ for any $\varepsilon > 0$. Let us now deal with the first part of $A(z)$. Using the fact that $[z^n](1-z)^{-r} = \binom{n+r-1}{r-1}$ for r a positive integer, we obtain:

$$\begin{aligned}
\sum_{j=1}^{r} \frac{a_j}{(z-\rho)^j} &= \sum_{j=1}^{r} \frac{a_j(-1)^j}{\rho^j(1-z/\rho)^j} \\
&= \sum_{j=1}^{r} (-1)^j a_j \rho^{-j} \sum_{n=0}^{\infty} \binom{n+j-1}{n} \left(\frac{z}{\rho}\right)^n \\
&= \sum_{n=1}^{\infty} z^n \sum_{j=1}^{r} (-1)^j a_j \binom{n}{j-1} \rho^{-(n+j)} \ .
\end{aligned}$$

In summary, we prove that

$$[z^n]A(z) = \sum_{j=1}^{r} (-1)^j a_j \binom{n}{j-1} \rho^{-(n+j)} + O((1/R' + \varepsilon)^n)$$

for $R' > \rho$ and any $\varepsilon > 0$.

Example 13: *Frequency of a Given Pattern Occurrence*

Let H be a *given* pattern of size m, and consider a random text of length n generated according to the Bernoulli model. An old and well studied problem of pattern matching (cf. [26]) asks for an estimation of the number O_n of pattern H occurrences in the text. Let $T_r(z) = \sum_{n=0}^{\infty} \Pr\{O_n = r\} z^n$ denote the generating function of $\Pr\{O_n = r\}$ for $|z| \le 1$. It can be proved (cf. [38, 44]) that

$$T_r(z) = \frac{z^m P(H)(D(z) + z - 1)^{r-1}}{D^{r+1}(z)} .$$

where $D(z) = P(H)z^m + (1 - z)A_H(z)$ and $A(z)$ is the so called *autocorrelation polynomial* (a polynomial of degree m). It is also easy to see that there exists smallest $\rho > 1$ such that $D(\rho) = 0$. Then, an easy application of the above analysis leads to

$$\Pr\{O_n(H) = r\} = \sum_{j=1}^{r+1} (-1)^j a_j \binom{n}{j-1} \rho^{-(n+j)} + O(\rho_1^{-n})$$

where $\rho_1 > \rho$ and $a_{r+1} = \rho^m P(H) (\rho - 1)^{r-1} (D'(\rho))^{-r-1}$. □

The method just described can be called the method of *subtracted singularities*, and its general description follows: Imagine that we are interested in the asymptotic formula for coefficients a_n of a function $A(z)$ whose circle of convergence is R. Let us also assume that we can find a simpler function, say $\bar{A}(z)$ that has the same singularities as $A(z)$ (e.g., in the example above $\bar{A}(z) = \sum_{j=1}^{r} \frac{a_{-j}}{(z-\rho)^j}$). Then, $A_1(z) = A(z) - \bar{A}(z)$ is analytical in a larger disk, of radius $R' > R$, say, and its coefficients are not dominant in an asymptotic sense. To apply this method successfully, we need to develop asymptotics of some known functions (e.g., $(1 - z)^\alpha$ for any real α) and establish the so called *transfer theorems* (cf. [30]). This leads us to the so called *singularity analysis* of Flajolet and Odlyzko [30] which we discuss next.

We start with the observation that

$$[z^n]A(z) = \rho^n [z^n]A(z/\rho) ,$$

that is, we need only to study singularities at, say, $z = 1$. The next observation deals with asymptotics of $(1 - z)^{-\alpha}$. Above, we show how to obtain coefficients at z^n of this function when α is a natural number. Then, the function $(1 - z)^{-\alpha}$ has a pole of order α at $z = 1$. However, when $\alpha \ne 1, 2 \ldots$, then the function has an *algebraic singularity* (in fact, it is then a multi-valued function). Luckily enough, we can proceed formally as follows:

$$\begin{aligned} [z^n](1 - z)^{-\alpha} &= \binom{n + \alpha - 1}{n} = \frac{\Gamma(n + \alpha)}{\Gamma(\alpha)\Gamma(n + 1)} \\ &= \frac{n^{\alpha-1}}{\Gamma(\alpha)} \left(1 + \frac{\alpha(\alpha - 1)}{2n} + O\left(\frac{1}{n^2}\right)\right) \end{aligned}$$

provided $\alpha \notin \{0, -1, -2, \ldots\}$. In the above, $\Gamma(x) = \int_0^\infty e^{-t}x^{t-1}dx$ is the Euler Gamma function (cf. [5, 45]), and the latter asymptotic expansion follows from the Stirling formula. Even more generally, let

$$A(z) = (1-z)^{-\alpha}\left(\frac{1}{z}\log\frac{1}{1-z}\right)^{\beta} .$$

Then, as shown by Flajolet and Odlyzko [30]

$$a_n = [z^n]A(z) = \frac{n^{\alpha-1}}{\Gamma(\alpha)}\left(1 + C_1\frac{\beta}{\log n} + C_2\frac{\beta(\beta-1)}{2\log^2 n} + O\left(\frac{1}{\log^3 n}\right)\right) \qquad (34)$$

provided $\alpha \notin \{0, -1, -2, \ldots\}$, and C_1 and C_2 are constants that can be calculated explicitly.

The most important aspect of the singularity theory comes next: In many instances we do *not* have an explicit expression for the generating function $A(z)$ but only an expansion of $A(z)$ around a singularity. For example: let $A(z) = (1-z)^{-\alpha} + O(B(z))$. In order to pass to coefficients of a_n we need a "transfer theorem" that will allow us to pass to coefficients of $B(z)$ under the "Big Oh" notation. These transfer theorems are jewels of Flajolet and Odlyzko theory [30], and we discuss them below.

We need a definition of Δ-analyticity around the singularity $z = 1$:

$$\Delta = \{z : \ |z| < R, \ z \neq 1, \ |\arg(z-1)| > \phi\}$$

for some $R > 1$ and $0 < \phi < \pi/2$ (i.e., the domain Δ is an extended disk around $z = 1$ with a circular part rooted at $z = 1$ deleted). Then:

Theorem 8 (Flajolet and Odlyzko 1990) *Let $A(z)$ be Δ-analytical that satisfies in a neighbourhood of $z = 1$ either*

$$A(z) = O\left((1-z)^{-\alpha}\log^{\beta}(1-z)^{-1}\right)$$

or

$$A(z) = o\left((1-z)^{-\alpha}\log^{\beta}(1-z)^{-1}\right) .$$

Then, either

$$[z^n] = O\left(n^{\alpha-1}\log^{\beta} n\right)$$

or

$$[z^n] = o\left(n^{\alpha-1}\log^{\beta} n\right) ,$$

respectively.

A classical example of singularity analysis is the Flajolet and Odlyzko analysis of the height of binary trees (cf. [30]), however, we finish this subsection with a simpler application that quite well illustrates the theory.

Example 14: *Certain Sums from Coding Theory*

In coding theory the following sum is of some interest:

$$S_n = \sum_{i=0}^{n} \binom{n}{i} (i/n)^i (1-i/n)^{n-i}$$

Let $s_n = n^n S_n$. If $s(z)$ denotes the exponential generating function of s_n, then by a simple application of convolution principle of exponential generating functions we obtain $s(z) = (b(z))^2$ where $b(z) = (1-t(z))^{-1}$ and $t(z)$ is the "tree function" defined in Subsection 5.1 (cf. (22)). In fact, we already know that this function also satisfies the functional equation $t(z) = ze^{t(z)}$. One observes that $z = e^{-1}$ is the singularity point of $t(z)$, and (cf. [92])

$$\begin{aligned}
t(z) - 1 &= \sqrt{2(1-ez)} + \frac{2}{3}(1-ez) + \frac{11\sqrt{2}}{36}(1-ez)^{3/2} + \frac{43}{135}(1-ez)^2 + O((1-ez)^{5/2}) \ , \\
s(z) &= \frac{1}{2h(z)\left(1 + \frac{\sqrt{2}}{3}\sqrt{h(z)} + \frac{11}{36}h(z) + O(h^{3/2}(z))\right)^2} \\
&= \frac{1}{2(1-ez)} + \frac{\sqrt{2}}{3\sqrt{(1-ez)}} + \frac{1}{36} + \frac{\sqrt{2}}{540}\sqrt{1-ez} + O(1-ez) \ .
\end{aligned}$$

Thus, an application of the singularity analysis leads finally to the following asymptotic expansion

$$S_n = \sqrt{\frac{n\pi}{2}} + \frac{2}{3} + \frac{\sqrt{2\pi}}{24}\frac{1}{\sqrt{n}} - \frac{4}{135}\frac{1}{n} + O(1/n^{3/2}) \ .$$

For more sophisticated examples the reader is referred to [30, 35, 92]. □

5.3 Mellin Transform and Asymptotics

In previous sections, we study functional equations such as (27) or more generally (32). They can be summarized by the following general functional equation:

$$f^{(b)}(z) = a(z) + \alpha f(zp) + \beta f(zq) \tag{35}$$

where $f^{(b)}(z)$ denotes the bth derivative of $f(z)$, α, β are constants, and $a(z)$ is a known function. An *important* point to observe is that in the applications described so far the unknown function $f(z)$ was usually a Poisson transform, that is, $\widetilde{f}(z) = \sum_{n\geq 0} f_n \frac{z^n}{n!} e^{-z}$. We briefly discuss consequences of this point at the end of this subsection where some elementary *depoissonization* results will be presented. An effective approach to solve asymptotically (either for $z \to 0$ or $z \to \infty$) the above function equation is by the so called **Mellin transform** which we discuss next. D.E. Knuth [66], together with De Bruijn, is responsible for introducing the Mellin transform in the "orbit" of the average case analysis of algorithms, however, it was popularized by Flajolet and his school who applied Mellin transforms to "countably" many problems of analysis of algorithms and analytical combinatorics. We base this subsection mostly on a beautiful survey of Flajolet *et al.* [33].

For a function $f(x)$ defined on $x \in [0, \infty)$ we define the Mellin transform as

$$\mathcal{M}(f, s) = f^*(s) = \int_0^\infty f(x) x^{s-1} dx$$

where s is a complex number. For example, observe that from the definition of the Euler gamma function, we have $\Gamma(s) = \mathcal{M}(e^x, s)$. The Mellin transform is a special case of the Laplace transform (set $x = e^t$) or the Fourier transform (set $x = e^{i\omega}$). Therefore, using the inverse Fourier transform, one establishes the inverse Mellin transform as (cf. [20, 45]), namely:

$$f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} f^*(s) x^{-s} ds$$

provided $f(x)$ is continuous. In the above, the integration is along a vertical line $\Im(s) = c$, and c must belong to the so called fundamental strip where the Mellin transform exists (see properly (P1) below).

The usefulness of the Mellin transform to the analysis of algorithms is a consequence of a few properties that we discuss in the sequel.

(P1) Fundamental Strip

Let $f(x)$ be a piecewise continuous function on the interval [0,∞) such that

$$f(x) = \begin{cases} O(x^\alpha) & x \to 0 \\ O(x^\beta) & x \to \infty \; . \end{cases}$$

Then the Mellin transform of $f(x)$ exists for any complex number s in the **fundamental strip** $-\alpha < \Re(s) < -\beta$, which we will denote $\langle -\alpha; -\beta \rangle$.

(P2) Smallness of Mellin transforms

Let $s = \sigma + it$. By the Riemann-Lebesgue lemma

$$f^*(\sigma + it) = o(|t|^{-r}) \qquad \text{as} \;\; t \to \pm\infty$$

provided $f \in \mathcal{C}^r$ where $\mathcal{C}^r$ is the set of functions having continuous r derivatives. More formally:

(P3) Basic Functional Properties

The following holds in appropriate strips:

$$\begin{array}{rcll} f(\mu x) & \Leftrightarrow & \mu^{-s} f^*(s) & (\mu > 0) \\ f(x^\rho) & \Leftrightarrow & \dfrac{1}{\rho} f^*(s/\rho) & (\rho > 0) \\ \dfrac{d}{dx} f(x) & \Leftrightarrow & -(s-1) f^*(s) & \\ \displaystyle\int_0^x f(t) dt & \Leftrightarrow & -\dfrac{1}{s} f^*(s+1) & \\ f(x) = \displaystyle\sum_{k \geq 0} \lambda_k g(\mu_k x) & \Leftrightarrow & f^*(s) = g^*(s) \displaystyle\sum_{k \geq 0} \lambda_k \mu_k^{-s} & \text{(Harmonic Sum Rule)} \end{array}$$

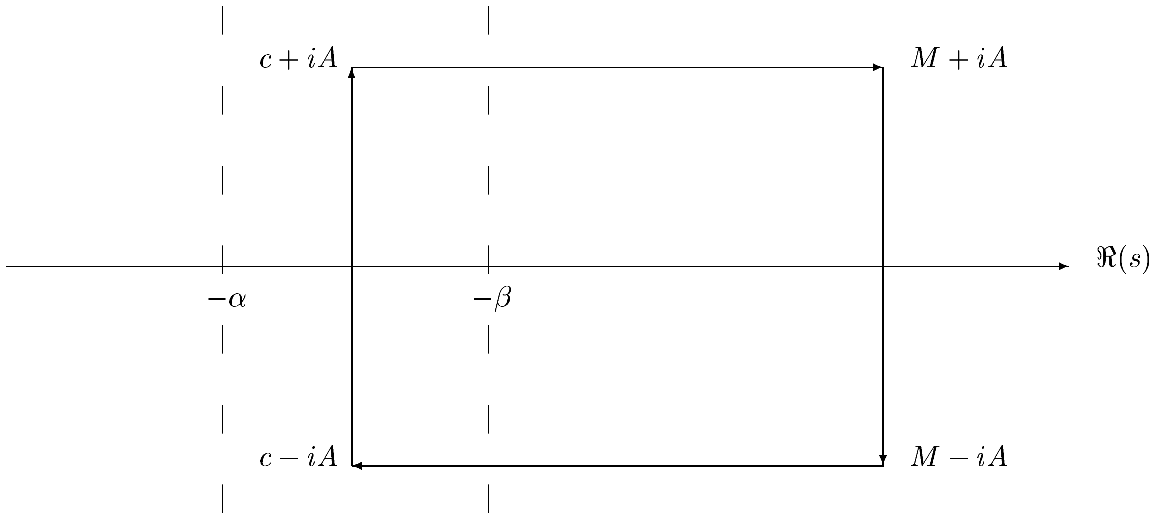

Figure 3: The fundamental strip of $f^*(s)$ and the integration contour

(P4) Asymptotics for $x \to 0$ and $x \to \infty$

Let the fundamental strip of $f^*(s)$ be the set of all s such that $-\alpha < \Re(s) < -\beta$ and assume that for $s = \sigma + i\tau$, $f^*(s) = O(|s|^r)$ with $r > 1$ as $|s| \to \infty$. If $f^*(s)$ can be analytically continued to a meromorphic function for $-\beta \le \Re(s) \le M$ with finitely many poles λ_k such that $\Re(\lambda_k) < M$, then as $x \to \infty$,

$$F(x) = -\sum_{\lambda_k \in \mathcal{H}} \mathrm{Res}\{F^*(s)x^{-s}, s = \lambda_k\} + O(x^{-M}) \qquad x \to \infty$$

where M is as large as we want. (In a similar fashion one can continue the function $f^*(s)$ to the left to get an asymptotic formula for $x \to 0$.) This property is so important that we provide here a sketch of a proof. Consider the rectangle R given in Figure 3 with the corners as illustrated. Choose A so that the sides of R do not pass through any singularities of $F^*(s)x^{-s}$. When evaluating

$$\lim_{A\to\infty} \int_R = \lim_{A\to\infty} \left(\int_{c-iA}^{c+iA} + \int_{c+iA}^{M+iA} + \int_{M+iA}^{M-iA} + \int_{M-iA}^{c-iA} \right),$$

the second and fourth integrals contribute very little since $F^*(s)$ is small for s with a large imaginary part by property (P2). The contribution of the fourth integral is computed as follows:

$$\left| \int_{M+i\infty}^{M-i\infty} F^*(s)x^{-s}ds \right| = \left| \int_{\infty}^{-\infty} F^*(M+it)x^{-M-it}dt \right| \le |x^{-M}| \int_{\infty}^{-\infty} |F^*(M+it)||x^{-it}|dt$$

But the last integrand decreases exponentially as $|t| \to \infty$, thus giving a contribution of $O(x^{-M})$. Finally, using Cauchy's residue theorem and taking into account the negative direction of R, we have

$$-\sum_{\lambda_k \in \mathcal{H}} Res\{F^*(s)x^{-s}, s = \lambda_k\} = \frac{1}{2i\pi}\int_{c-i\infty}^{c+i\infty} F^*(s)x^{-s}ds + O(x^{-M}),$$

which proves the desired result.

Specifically, the above implies that if the above smallness condition on $f^*(s)$ is satisfied for $-\beta < \Re(s) \le M$, $(M > 0)$, then

$$f^*(s) = \sum_{k=0}^{K} \frac{d_k}{(s-b)^{k+1}}, \tag{36}$$

implies

$$f(x) = -\sum_{k=0}^{K} \frac{d_k}{k!} x^{-b}(-\log x)^k + O(x^{-M}) \qquad x \to \infty . \tag{37}$$

In a similar fashion, if for $-M < \Re(s) < -\alpha$ the smallness condition of $f^*(s)$ holds and

$$f^*(s) = \sum_{k=0}^{K} \frac{d_k}{(s-b)^{k+1}}, \tag{38}$$

then

$$f(x) = \sum_{k=0}^{K} \frac{d_k}{k!} x^{-b}(-\log x)^k + O(x^{M}) \qquad x \to 0 . \tag{39}$$

Mellin Transform in The Complex Plane (cf. [20, 33, 55])
If $f(z)$ is analytic in a cone $\theta_1 \le \arg(z) \le \theta_2$ with $\theta_1 < 0 < \theta_2$, then the Mellin transform $f^*(s)$ can be defined by replacing the path of integration $[0, \infty[$ by any curve starting at $z = 0$ and going to ∞ inside the cone, and it is identical with the real transform $f^*(s)$ of $f(z) = F(z)\Big|_{z\in\mathbb{R}}$. In particular, if $f^*(s)$ fulfills an asymptotic expansion as (36) or (38), then (37) or (39) for $f(z)$ holds in $z \to \infty$ and $z \to 0$ in the cone, respectively.

Let us now apply Mellin transforms to some problems studies above. For example, consider a trie for which the functional equation (27) becomes

$$\widetilde{X}(z) = \widetilde{A}(z) + \widetilde{X}(zp) + \widetilde{X}(zq)$$

where $p + q = 1$ and $\widetilde{A}(z)$ is the Poisson transform of a known function. Thanks to property (P3) the Mellin transform translates the above functional equation to an algebraic one which can be immediately solved resulting in

$$X^*(s) = \frac{A^*(s)}{1 - p^{-s} - q^{-s}}$$

provided there exists a fundamental strip for $X^*(s)$ where also $A^*(s)$ is well defined. Now, thanks to property (P4) we can easily compute asymptotics of $\widetilde{X}(z)$ as $z \to \infty$ in a cone. More formally, we obtain asymptotics for z real, say x, and then either analytically continue our results or apply property (P5) which basically says that there is a cone in which the asymptotic results for real x can be extended to a complex z. Examples of usage of this technique can be found in [27, 35, 46, 49, 50, 51, 54, 55, 63, 66, 73].

This is a good plan to attack the problem, however, one must translate asymptotics of the Poisson transform $\widetilde{X}(z)$ into the original sequence, say x_n. One would like to have $x_n \sim \widetilde{X}(n)$, but this is not true in general (e.g., take $x_n = (-1)^n$). To assure the above asymptotic equivalence, we enter another area of research called **depoissonization** that was recently actively pursued [49, 50, 51, 54, 55, 81]. Due to lack of space, we cite below only one result that found many applications in the analysis of algorithms:

Theorem 9 (Jacquet and Szpankowski 1995, 1996) *Let $\widetilde{X}(z)$ be the Poisson transform of a sequence x_n that is assumed to be an entire function of z. We postulate that in a cone S_θ ($\theta < \pi/2$) the following two conditions simultaneously hold for some real numbers $A, B, R > 0$, β, and $\alpha < 1$:*

(I) *For $z \in S_\theta$*

$$|z| > R \quad \Rightarrow \quad |\widetilde{X}(z)| \leq B|z|^\beta \ ,$$

(O) *For $z \notin S_\theta$*

$$|z| > R \quad \Rightarrow \quad |\widetilde{X}(z)e^z| \leq A\exp(\alpha|z|) \ .$$

Then,

$$x_n = \widetilde{X}(n) + O(n^{\beta-1}) \tag{40}$$

for large n.

The verification of conditions (I) and (O) is usually not too difficult, and can be accomplished directly on the functional equation at hand through the so called *increasing domains* method discussed in [54].

Finally, we should say that there is an easier (however, not that powerful) approach to deal with a majority of functional equations of type (27). As we pointed out, such equations possess solutions that can be represented as some alternating sums (cf. (28) and Examples 10-12). Let us consider a general alternating sum

$$S_n = \sum_{k=m}^{n} (-1)^k \binom{n}{k} f_k$$

where f_k is a known, but otherwise, general sequence. The following two equivalent approaches (cf. [34, 66, 87]) use complex integration (the second one is actually a Mellin-like approach) to simplify the computations of asymptotics of S_n for $n \to \infty$ (usually through residue calculus).

Theorem 10 (Rice's Formula) (i) *Let $f(s)$ be an analytical continuation of $f(k) = f_k$ that contains the half line $[m, \infty)$. Then,*

$$S_n := \sum_{k=m}^{n} (-1)^k \binom{n}{k} f_k = \frac{(-1)^n}{2\pi i} \int_C f(s) \frac{n!}{s(s-1)\cdots(s-n)} ds$$

where C is a positively enclosed curve that encircles $[m, n]$ and does not include any of the integers $0, 1, \ldots, m-1$.

(ii) (**Szpankowski 1988**) *Let $f(s)$ be analytical left to the vertical line $(\frac{1}{2} - m - i\infty, \frac{1}{2} - m + i\infty)$ and it does not grow too fast at infinity, then*

$$\begin{aligned} S_n &= \frac{1}{2\pi i} \int_{\frac{1}{2}-m-i\infty}^{\frac{1}{2}-m+i\infty} f(-z) B(N+1, z) dz \\ &= \frac{1}{2\pi i} \int_{\frac{1}{2}-m-i\infty}^{\frac{1}{2}-m+i\infty} f(-z) n^{-z} \Gamma(z) \left(1 - \frac{z(z+1)}{2n} + \frac{z(1+z)}{24n^2}(3(1+z)^2 + z - 1) + O(n^{-3})\right) dz \end{aligned}$$

where $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the Beta function.

The precise growth condition for $f(z)$ of part (ii) can be found in [87].

Example 15: *Asymptotics of Some Alternating Sums*

In Examples 10-12 we deal with alternating sums of the following general type:

$$S_n(r) = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \binom{k}{r} \frac{1}{p^{-k} - q^{-k}}$$

where $p + q = 1$. We now use Theorem 10 to obtain asymptotics of S_n as n becomes large and r is fixed. To simplify our computation we use part(ii) of the above theorem which leads to

$$S_n(r) = \frac{1}{2\pi i} \frac{(-1)^n}{r!} \int_{\frac{1}{2}-[2-r]^+-i\infty}^{\frac{1}{2}-[2-r]^++i\infty} n^{r-z} \Gamma(z) \frac{1}{1 - p^{r-z} - q^{r-z}} dz + e_n$$

$x^+ = \max\{0, x\}$, where e_n is an error term that we discuss later. The above integral should remind the reader of the integral appearing in the inverse Mellin transform. Thus, we can estimate it using a similar approach. First of all, we observe that the function under the integral has infinitely many poles at

$$1 = p^{r-z} + q^{r-z} \; .$$

It can be proved (cf. [54]) that these poles, say z_k for $k = 0, \pm 1, \ldots$, lie on a line $\Re(z) = r - 1$ provided $\log p / \log q$ is rational, which we assume to hold. Thus, we can write $z_k = r - 1 + iy_k$ where $y_0 = 0$ and otherwise a real number for $k \neq 0$. Observe also that the line at $\Re(z) = r - 1$ lies *right* to the line of integration $(\frac{1}{2} - [2-r]^+ - i\infty, \frac{1}{2} - [2-r]^+ + i\infty)$. To take advantages of the Cauchy residue theorem, as in Figure 3, we consider a big rectangle with left side being the line of integration, the right size position at $\Re(z) = M$ (where M is a large number), and bottom and top side position at $\Im(z) = \pm A$, say. We further observe that the right side contributes only $O(n^{r-M})$ due to the factor n^{r-M} in the integral. Both, bottom and top sides, contributes negligible, too, since the gamma function decays exponentially fast with the increase of imaginary part (i,e., when $A \to \infty$). In summary, the integral is equal to a circular integral (around the rectangle) plus a negligible part $O(n^{r-M})$. But, then by Cuachy's residue theorem the latter integral is equal to minus the sum of all residues at z_k, that is,

$$S_n(r) = -\sum_{k-\infty}^{\infty} \text{Res}\left(\frac{n^{r-z}\Gamma(z)}{1-p^{r-z}-q^{r-z}},\ z = z_k\right) + O(n^{r-M}) \ .$$

We can compute the residues using MAPLE (as shown in Section 3.3). Equivalently, for $k = 0$ (the main contribution to the asymptotics comes from $z_0 = r - 1$) we can use the following expansions around $w = z - z - z_0$

$$\begin{aligned} n^{r-z} &= n(1 - w\ln n + O(w^2)) \ , \\ (1-p^{r-z}-q^{r-z})^{-1} &= -w^{-1}h^{-1} + \frac{1}{2}h_2h^{-2} + O(w), \\ \Gamma(z) &= (-1)^{r+1}(w^{-1} - \gamma + \delta_{r,0}) + O(w) \qquad r = 0,1 \end{aligned}$$

where $h = -p\ln p - q\ln q$, $h_2 = p\ln^2 p + q\ln^2 q$, and $\gamma = 0.577215\ldots$ is the Euler constant. Considering in addition the residues coming from z_k for $k \neq 0$ we finally arrive at

$$S_n(r) = \begin{cases} \frac{1}{h}n(\ln n + \gamma - \delta_{r,0} + \frac{1}{2}h_2) + (-1)^r nP_r(n) + e_n & r = 0,1 \\ n\frac{(-1)^r}{r(r-1)h} + (-1)^r nP_r(n) + e_n & r \geq 2 \end{cases}$$

where the error term can be computed easily to be $e_n = O(1)$ (using the arguments as above and observing that the error term has a similar integral representation but with term n^{-1} in front of it). In the above $P_r(n)$ is a contribution from z_k for $k \neq 0$, and it is a fluctuating function with small amplitude. For example, when $p = q = 1/2$, then

$$P_r(n) = \frac{1}{\ln 2}\sum_{k\neq 0}\Gamma(r + 2\pi ik/\log 2)\exp(-2\pi ik\log_2 n)$$

is a periodic function of $\log x$ with period 1, mean 0 and amplitude $\leq 10^{-6}$ for $r = 0,1$. $\Box$

ACKNOWLEDGEMENT

The author thanks his colleagues P. Jacquet, G. Louchard, H. Prodinger and K. Park for reading earlier versions of this chapter and comments that led to improvements of the presentation.

References

[1] M. Abramowitz, and I. Stegun, *Handbook of Mathematical Functions*, Dover, New York 1964.

[2] A. Aho, J. Hopcroft, and J. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading 1974.

[3] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*, Springer Verlag, New York 1989.

[4] D. Aldous, M. Hofri, and W. Szpankowski, Maximum Size of a Dynamic Data Structure: Hashing with Lazy Deletion Revisited, *SIAM J. Computing*, 21, 713-732, 1992.

[5] N. Alon and J. Spencer, *The Probabilistic Method*, John Wiley & Sons, New York 1992.

[6] A. Apostolico, The Myriad Virtues of Suffix Trees, *Combinatorial Algorithms on Words*, 85-96, Springer-Verlag, ASI F12 (1985).

[7] R. Arratia and M. Waterman, The Erdös-Rényi Strong Law for Pattern Matching with Given Proportion of Mismatches, *Annals of Probability*, 17, 1152-1169, 1989.

[8] R. Arratia and M. Waterman, A Phase Transition for the Score in Matching Random Sequences Allowing Deletions, *Annals of Applied Probability*, 4, 200-225, 1994.

[9] R. Arratia, L. Gordon, and M. Waterman, The Erdös-Rényi Law in Distribution for Coin Tossing and Sequence Matching, *Annals of Statistics*, 18, 539-570, 1990.

[10] A. Apostolico, M. Atallah, L. Larmore, and S. McFaddin, Efficient Parallel Algorithms for String Editing and Related Problems, *SIAM J. Comput.*, 19, 968-988, 1990.

[11] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York 1968.

[12] B. Bollobás, *Random Graphs*, Academic Press, London 1985.

[13] A. Blum, T. Jiang, M. Li, J. Tromp, M. Yannakakis, Linear Approximation of Shortest Superstring, *J. the ACM*, 41, 630-647, 1994

[14] G. Brassard and P. Bratley, *Algorithmics. Theory and Practice*, Prentice Hall, Englewood Cliffs, 1988.

[15] S-N. Choi and M. Golin, Lopsided trees: Algorithms, Analyses and Applications, *Proc. the 23rd International Colloquium on Automata Languages and Programming* (ICALP '96), July 1996.

[16] V. Chvatal and D. Sankoff, Longest Common Subsequence of Two Random Sequences, *J. Appl. Prob.*, 12, 306-315, 1975.

[17] E. Coffman and G. Lueker, *Probabilistic Analysis of Packing and Partitioning Algorithms*, John Wiley & Sons, New York 1991.

[18] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York (1991).

[19] M. Crochemore and W. Rytter, *Text Algorithms*, Oxford University Press, New York (1995).

[20] B. Davies, *Integral Transforms and Their Applications*, Springer-Verlag, New York 1978.

[21] Y. Derriennic, Un Théorème Ergodique Presque Sous Additif, *Ann. Probab.*, 11, 669-677, 1983.

[22] A. Dembo and O. Zeitouni, *Large Deviations Techniques*, Jones and Bartlett Publishers, Boston 1993.

[23] P. Erdös and J. Spencer, *Probabilistic Methods in Combinatorics*, Academic Press, New York 1974.

[24] Durrett, R., *Probability: Theory and Examples*, Wadsworth, Belmont CA 1991.

[25] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol.I, John Wiley & Sons, 1970

[26] W. Feller *An Introduction to Probability Theory and its Applications*, Vol.II, John Wiley & Sons, 1971

[27] Fill, J. A., Mahmoud, H. M., and Szpankowski, W. On the distribution for the duration of a randomized leader election algorithm. *Ann. Appl. Probab.*, 1996.

[28] P. Flajolet, Analytic Analysis of Algorithms, *Lectures Notes in Computer Science*, Vol. 623, Ed. W. Kuich, 186-210, Springer-Verlag 1992.

[29] P. Flajolet, M. Régnier and D. Sotteau, Algebraic Methods for Trie Statistics, *Annals of Discrete Mathematics*, 25, 145-188, 1985.

[30] P. Flajolet and A. Odlyzko, Singularity Analysis of Generating Functions, *SIAM J. Disc. Methods*, 3, 216-240, 1990.

[31] P. Flajolet and B. Richmond, Generalized Digital Trees and Their Difference-Differential Equations, *Random Structures and Algorithms*, 3, 305-320, 1992.

[32] P. Flajolet and M. Soria, General Combinatorial Schemas: Gaussian Limit Distributions and Exponential Tails, *Discrete Mathematics*, 114, 159-180 (1993).

[33] P. Flajolet, X. Gourdon, P. Dumas, Mellin Transforms and Asymptotics: Harmonic sums, *Theoretical Computer Science*, 144, 3-58, 1995.

[34] P. Flajolet, and R. Sedgewick, Mellin Transforms and Asymptotics: Finite Differences and Rice's Integrals. *Theoretical Computer Science*, 144, 101–124, 1995.

[35] P. Flajolet, and R. Sedgewick, *Analytical Combinatorics*, in preparation; see also INRIA TR-1888 1993, TR-2026 1993 and TR-2376 1994.

[36] A. Frieze and C. McDiarmid, Algorithmic Theory of Random Graphs, *Random Structures & Algorithms*, 10, 1997.

[37] A. Frieze and W. Szpankowski, Greedy Algorithms for the Shortest Common Superstring That Are Asymptotically Optimal, *Proc. European Symposium on Algorithms*, Barcelona (1996).

[38] I. Fudos, E. Pitoura and W. Szpankowski, On Pattern Occurrences in a Random Text, *Information Processing Letters*, 57, 307-312, 1996.

[39] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, Robert E. Krieger Publishing Company, Malabar, Florida 1987.

[40] Z. Galil and R. Giancarlo, Data Structures and Algorithms for Approximate String Matching, *J. Complexity*, 4, 33-72, (1988).

[41] D.H. Greene and D.E. Knuth, *Mathematics for the Analysis of Algorithms*, Birkhauser, 1981

[42] M. Golin, Limit Theorems for Minimum-Weight Triangulations, Other Euclidean Functionals and Probabilistic Recurrence Relations, *Seventh Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA96), 252-260, 1996

[43] G.H. Gonnet and R. Baeza-Yates, *Handbook of Algorithms and Data Structures*, Addison-Wesley, Workingham (1991).

[44] L. Guibas and A. M. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *J. Combin.Theory Ser. A*, 30, 183-208, 1981.

[45] P. Henrici, *Applied and Computational Complex Analysis*, Vols. 1–3, John Wiley & Sons 1977.

[46] M. Hofri, *Analysis of Algorithms. Computational Methods and Mathematical Tools*, Oxford University Press, New York 1995.

[47] H-K. Hwang, Large Deviations for Combinatorial Distributions I: Central Limit Theorems, *Ann. Appl. Probab.*, 6, 297-319, 1996.

[48] H-K. Hwang, Limit Theorems for Mergesort, *Random Structures and Algorithms*, 8, 319-336, 1996.

[49] P. Jacquet and M. Régnier, Limiting Distributions for Trie Parameters, *Lecture Notes in Computer Science*, 214, 196-210, 1986.

[50] P. Jacquet and M. Régnier, Normal Limiting Distribution of the Size of Tries, *Proc. Performance'87*, 209-223, North Holland, Amsterdam 1987

[51] P. Jacquet and W. Szpankowski, Ultimate Characterizations of the Burst Response of an Interval Searching Algorithm: A Study of a Functional Equation, *SIAM J. Computing*, 18, 777-791, 1989.

[52] P. Jacquet and W. Szpankowski, Analysis of Digital Tries with Markovian Dependency, *IEEE Trans. Information Theory*, 37, 1470-1475, 1991.

[53] P. Jacquet and W. Szpankowski, Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combin.Theory Ser. A*, 66, 237-269, 1994.

[54] P. Jacquet and W. Szpankowski, Asymptotic Behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161-197, 1995.

[55] P. Jacquet and W. Szpankowski, Analytical Depoissonization and Its Applications, preprint.

[56] S. Karlin and F. Ost, Counts of Long Aligned Word Matches Among Random Letter Sequences, *Adv. Appl. Prob.*, 19, 293-351, 1987.

[57] R. Karp, The Probabilistic Analysis of Some Combinatorial Search Algorithms. In *Algorithms and Complexity*, ed. J.F. Traub, Academic Press, New York 1976.

[58] R. Karp, An Introduction to Randomized Algorithms, *Discrete Applied Mathematics*, 34, 165-201, 1991.

[59] J.F.C. Kingman, *Subadditive Processes*, in Ecole d'Eté de Probabilités de Saint-Flour V-1975, Lecture Notes in Mathematics, 539, Springer-Verlag, Berlin (1976).

[60] P. Kirschenhofer and H. Prodinger, On Some Applications of Formulæ of Ramanujan in the Analysis of Algorithms, *Mathematika*, 38, 14-33, 1991.

[61] P. Kirschenhofer, H. Prodinger and W. Szpankowski, On the Variance of the External Path in a Symmetric Digital Trie *Discrete Applied Mathematics*, 25, 129–143, 1989.

[62] P. Kirschenhofer, H. Prodinger and W. Szpankowski, Digital Search Trees Again Revisited: The Internal Path Length Perspective, *SIAM J. Computing*, 23, 598-616, 1994.

[63] P. Kirschenhofer, H. Prodinger and W. Szpankowski, Analysis of a Splitting Process Arising in Probabilistic Counting and Other Related Algorithms, *Random Structures & Algorithms*, to appear.

[64] D. E. Knuth, *The Art of Computer Programming. Fundamental Algorithms,* Vol. 1. Addison-Wesley, Reading, Mass. 1973.

[65] D.E. Knuth, *The Art of Computer Programming. Seminumerical Algorithms.* Vol. II. Addison Wesley, Reading, Mass. 1981.

[66] D.E. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3., Addison-Wesley, Reading, MA 1973.

[67] A. Lesek (Ed.), *Computational Molecular Biology, Sources and Methods for Sequence Analysis*, Oxford University Press, 1988.

[68] L. Levin, Average Case Complete Problems, *SIAM J. Computing*, 15, 285-286, 1986.

[69] G. Louchard, Random Walks, Gaussian Processes and List Structures, *Theor. Comp. Sci.*, 53, 99-124, 1987.

[70] G. Louchard, R. Schott, Probabilistic Analysis of Some Distributed Algorithms, *Random Structures & Algorithms*, 2, 151-186, 1991.

[71] G. Louchard and W. Szpankowski, A Probabilistic Analysis of a String Editing Problem and its Variations, *Combinatorics, Probability and Computing*, 4, 143-166, 1994.

[72] G. Louchard and W. Szpankowski, Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm, *IEEE Trans. Information Theory*, 41, 478-488, 1995.

[73] H. Mahmoud, *Evolution of Random Search Trees*. Wiley, New York 1992.

[74] C. McDiarmid, On the Method of Bounded Differences, in *Surveys in Combinatorics*, J. Siemons (Ed.), vol 141, pp. 148-188, London Mathematical Society Lecture Notes Series, Cambridge University Press, 1989.

[75] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, Cambridge 1995.

[76] E. Myeres, An $O(ND)$ Difference Algorithm and Its Variations, *Algorithmica*, 1, 251-266, 1986.

[77] C. Newman, Chain Lengths in Certain Random Directed Graphs, *Random Structures & Algorithms*, 3, 243-254, 1992.

[78] A. Odlyzko, Asymptotic Enumeration, in *Handbook of Combinatorics*, Vol. II, (Eds. R. Graham, M. Götschel and L. Lovász), Elsevier Science, 1063-1229, 1995.

[79] B. Pittel, Asymptotic Growth of a Class of Random Trees, *Annals of Probability* , 18, 414 - 427 (1985).

[80] B. Pittel, Paths in a Random Digital Tree: Limiting Distributions, *Adv. Appl. Prob.*, 18, 139-155 (1986).

[81] B. Rais, P. Jacquet, and W. Szpankowski, Limiting Distribution for the Depth in Patricia Tries, *SIAM J. Discrete Mathematics*, 6, 197-213, 1993.

[82] R. Remmert, *Theory of Complex Functions*, Springer Verlag, New York 1991.

[83] D. Sankoff and J. Kruskal (Eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, Reading, Mass., 1983.

[84] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley Publishing Company, Reading Mass., 1995.

[85] A. N. Shiryayev, *Probability*, Springer-Verlag, New York 1984.

[86] W. Szpankowski, Solution of a Linear Recurrence Equation Arising in the Analysis of Some Algorithms, *SIAM J. Alg. Disc. Methods*, 8, 233-250, 1987.

[87] W. Szpankowski, The Evaluation of an Alternating Sum with Applications to the Analysis of Some Data Structures, *Information Processing Letters*, 28, 13-19, 1988.

[88] W. Szpankowski, Patricia Tries Again Revisited, *JACM*, 37, 691-711, 1990.

[89] W. Szpankowski, A Characterization of Digital Search Trees From the Successful Search Viewpoint, *Theoretical Computer Science*, 85, 117-134, 1991.

[90] W. Szpankowski, On the Height of Digital Trees and Related Problems, *Algorithmica*, 6, 256-277, 1991.

[91] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176-1198, 1993.

[92] W. Szpankowski, On Asymptotics of Certain Sums Arising in Coding Theory, *IEEE Trans. Information Theory*, 41, 2087-2090, 1995.

[93] M. Talagrand, A New look at Independence, *Ann. Appl. Probab.*, 6, 1-34, 1996.

[94] E. C. Titchmarsh, *The Theory of Functions*, Oxford University Press, Oxford 1944.

[95] E. Ukkonen, A Linear-Time Algorithm for Finding Approximate Shortest Common Superstrings, *Algorithmica*, 5, 313-323, 1990.

[96] J. Vitter and P. Flajolet, Average-Case Analysis of Algorithms and Data Structures, In *Handbook of Theoretical Computer Science*, Ed. J. van Leewen. 433-524, Elsevier Science Publishers, 1990.

[97] M. Waterman, *Introduction to Computational Biology*, Chapman & Hall, London 1995.

[98] H. Wilf, *generatingfunctionology*, Academic Press, Boston 1990.

[99] E. Whittaker and G. Watson, *A Course of Modern Analysis*, Cambridge University Press, Cambridge 1927.