Deviations from normality in random strings

Philippe Flajolet, Robert F. Tichy, Peter Kirschenhofer

▶ **To cite this version:**

Philippe Flajolet, Robert F. Tichy, Peter Kirschenhofer. Deviations from normality in random strings. [Research Report] RR-0719, INRIA. 1987. inria-00075833

HAL Id: inria-00075833
https://inria.hal.science/inria-00075833

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 719

DEVIATIONS FROM NORMALITY IN RANDOM STRINGS

Philippe FLAJOLET
Peter KIRSCHENHOFER
Robert TICHY

AOUT 1987

DEVIATIONS FROM NORMALITY IN RANDOM STRINGS

Philippe Flajolet

Peter Kirschenhofer

Robert Tichy

Abstract: *We show that almost all binary strings of length n contain all blocks of size $(1-\epsilon)\log_2 n$ a close to normal number of times. From this, we derive tight bounds on the discrepancy of random infinite strings. Our results are obtained through explicit generating function expressions and contour integration estimates*

DEVIATIONS PAR RAPPORT A LA NORME DANS LES SUITES ALEATOIRES

Résumé: Nous montrons que presque toutes les suites binaires de longueur n contiennent tous les blocks de longueur $(1-\epsilon)\log_2 n$ un nombre de fois proche de la normale. De là nous déduisons des bornes précises concernant la discrépance des suites aléatoires infinies. Ces résultats sont obtenus à partir de formes explicites de séries génératrices et d'estimations d'intégrales de contour.

PAPIER RÉCUPÉRÉ ET RECYCLÉ

DEVIATIONS FROM NORMALITY IN RANDOM STRINGS

Philippe Flajolet†
INRIA, Rocquencourt
78150 Le Chesnay (France)

Peter Kirschenhofer and *Robert Tichy*
Technische Universität Wien
Wiedner Hauptstraße 8–10
A–1040 Vienna (Austria)

Abstract: *We show that almost all binary strings of length n contain all blocks of size $(1-\epsilon)\log_2 n$ a close to normal number of times. From this, we derive tight bounds on the discrepancy of random infinite strings. Our results are obtained through explicit generating function expressions and contour integration estimates*

1. Introduction.

Emile Borel introduced in 1908 the notion of *normal numbers* characterized by the property that, in their binary representation, each block pattern of zeros and ones occurs with its natural probability (namely $1/2^k$ with k the length of the block). He then proved that almost all real $[0,1]$ numbers are normal, and later in his life conducted various experiments on digits of particular numbers like e, π or $\sqrt{2}$.

Our purpose in this paper is to provide statistical estimates for the occurrences of blocks in random binary strings of either finite or infinite length, and in particular try to determine quantitatively which deviations from the "norm" are to be expected in a random string.

To take a particular example, if one computes (say, with the Gauss-Salamin method) 10,000 bits of π and if one looks, for various values of k, at the least frequent block and most frequent block of size k, one finds:

Length (k)	Least Frequent	Most Frequent
2	$(11)^{2480}$	$(00)^{2509}$
3	$(111)^{1226}$	$(000)^{1275}$
4	$(1100)^{603}$	$(0000)^{652}$
5	$(11100)^{296}$	$(00000)^{341}$
6	$(100111)^{133}$	$(101101)^{176}$
7	$(0101100)^{57}$	$(0110110)^{97}$

† Research of the three authors was supported by the French Austrian scientific cooperation programme.

It would obviously be of interest to determine whether such deviations from the norm point out to specific "non-random" properties of the decimals of π.

Much in the same vein, tests on occurences of blocks in bits produced by (pseudo) *random number generators* are often employed and the reader may refer to Knuth's encyclopedic treatment on this subject (see especially Section 3.5 of [Kn69]).

The present paper concerns itself with *extremal statistics* regarding occurrences of blocks in random strings. The basic concept that formalizes our previous numerical observations, is that of *discrepancy* and it is often used in [0,1] distributions problems (see *e.g.* Hlawka's or Kuipers and Niederreiter's books [Hl79], [KN74]). Let $b = b_1 b_2 \dots b_n$ be a (finite) binary string. Then its k–discrepancy is defined as:

$$D_k(b) = \max_{|u|=k} \left| \frac{\Omega(b,u)}{n} - \frac{1}{2^k} \right|$$

where $\Omega(b,u)$ is the number of occurrences of a pattern (block) $u = u_1 u_2 \dots u_k$ in u:

$$\Omega(b,u) = \mathrm{card}\{j \mid b_j b_{j+1} \dots b_{j+k-1} = u_1 u_2 \dots u_k\}.$$

In the definition of discrepancy, $\Omega(b,u)/n$ represents the observed frequency of block b in u and $1/2^k$ is the probability of occurrence of block b at at any position in a random string. Thus, the discrepancy does represent deviations from normality observed in string b. Stated informally, a string b will pass a randomness test with block length k if the discrepancy is "much smaller" than $\frac{1}{2^k}$. It is therefore of interest to determine for what range of values of k (as a function of n) this test should be meaningful, as well as to determine what is an "acceptable" deviation from the norm.

Our main result for infinite strings is contained in Theorem 1 below. In essence, a "finite version" of this theorem (Theorem 2) states that *almost all* binary strings of length n contain *all* patterns of length k a (close to) normal number of times as long as $k < (1-\epsilon)\log_2 n$. Notice that much stronger results cannot be expected to hold since a string of length n has only n bit positions so that, when $k > \log_2 n$, some patterns are certain not to occur while others will tend to occur only once.

To state Theorem 1 precisely, we first need to introduce the notion of discrepancy for infinite strings.

Definition. Let $\omega = b_1 b_2 b_3 \cdots$ be an infinite string. Then, for integers k and n, the discrepancy $D_k(\omega, n)$ is $D_k(b_1 b_2 \cdots b_n)$. Let $k(n)$ be a non–decreasing sequence of positive integers. Then, the string ω is said to be $k(n)$–uniformly distributed if

$$\lim_{n\to\infty} 2^{k(n)} D_{k(n)}(\omega, n) = 0.$$

In previous works, we have established several properties of $k(n)$–uniformly distributed sequences. Earlier relevant results also appear in [FKT86].

Our Theorem 1 states that almost all infinite strings are fairly uniformly distributed. Here, our measures on finite and infinite strings are the usual product measures, with individual 0–1 bits being equally likely. The notation $\lg n$ is the binary logarithm $\lg n = \log_2 n$.

Theorem 1. *Let $k(n) \leq \lg n - \lg\lg n - 2\lg\lg\lg n$ be a non-decreasing sequence of positive integers. Then almost all infinite binary strings ω are $k(n)$ uniformly distributed.*

As a direct consequence, we get

Corollary 1. *Almost all infinite strings ω are $k(n)$ uniformly distributed for $k(n) = \lfloor(1-\epsilon)\lg n\rfloor$, with $\epsilon > 0$.*

Notice that from earlier research [KT85], the uniform distribution property was only known to hold for $k(n) \sim \frac{1}{3}\lg n$. As already said, our result is in essence the best possible.

To attain our goal, we mostly study distribution problems on finite strings. The transfer to infinite strings is then easy by the Borel-Cantelli lemma.

In Section 2, we introduce a particular Markov chain (with 2^k states) that records information about the simultaneous occurrences of all k–blocks in a random string. Interestingly enough, the graph of this Markov chain is nothing but a De Bruijn graph (see e.g. [Kn68, p.379]) used classically to construct minimal sequences that contain all possible k–blocks once and only once. The Markov chain is equivalent to a probabilistic traversal of this graph, while the construction of the minimal De Bruijn sequences corresponds to a particular deterministic traversal. Consideration of this Markov chain shows *a priori* that rational generating functions are to be expected in this range of problems. It also provides useful probabilistic intuitions and could lead to numerical approximations for parameters of interest when k is kept fixed.

We then proceed in Section 3 with the computation of the distribution of the number of occurrences of a fixed pattern in a random string of length n. This is achieved via generating functions. Here, the situation is greatly helped by the fact that closely related generating functions have earlier been computed by Guibas and Odlyzko [GO81a], [GO81b]. In particular, it turns out that, although the number of occurrences of a pattern of length k in an n-string has average:

$$\frac{n-k+1}{2^k},$$

the corresponding variance depends deeply on the overlap structure present in the pattern block. The *correlation polynomials* of Guibas and Odlyzko are essential to our treatment.

Section 4 uses crude saddle point estimates that suffice to obtain exponential tail results for occurrence probabilities. Such results are needed if we want to let k vary with n and approach $\lg n$. In Section 5, these estimates are used to derive rather directly the proof of our main result.

Notice the difference with two previous approaches [KT85], [FKT86]. Firstly, tail estimates based on Tchebycheff's inequality are too weak to lead to Theorem 1. Secondly, another approach based on W. Philipp's law of iterated logarithm leads to accurate probability distribution estimates for discrepancies, but is limited to slowly growing sequences $k(n)$. What renders our proof possible is the clear relation (from Guibas and Odlyzko's works) between generating functions and pattern structures.

2. A Universal Markov Chain.

We introduce here a Markov chain that is in a sense "universal" for counting pattern occurrences. It takes into account the simultaneous occurrence of all k–blocks in a random string of size n. Fix k, the block length. The Markov chain $\mathcal{M}^{(k)}$ has $\ell = 2^k$ states; state i means: "the block of $\{0,1\}$ which corresponds to the binary representation of integer i with length k has just occurred". Thus if a new element $\alpha \in \{0,1\}$ of a random string is added, the new state is $j = (2i+\alpha) \bmod 2^k$. Whence

Definition: The Markov chain $\mathcal{M}^{(k)}$ has 2^k states. Its transition matrix $M^{(k)}$ is given by

$$M^{(k)}_{ij} = \frac{1}{2} \qquad \text{if} \qquad (j = 2i \bmod 2^k) \text{ or } (j = 2i+1 \bmod 2^k),$$

all other entries being equal to 0.

It is of interest to note that the graph $\Gamma^{(k)}$ associated to $\mathcal{M}^{(k)}$, whose adjacency matrix is $2M^{(k)}$, is nothing but a classical De Bruijn graph used in combinatorics [Kn68,p.379]: The fact that this graph has a Eulerian circuit (all its nodes are of even degree) entails the existence of a (minimal) string of length $2^k + k - 1$ which contains every k–block once and only once.

Let V be the diagonal matrix with elements $(v_0, v_1, \ldots, v_{\ell-1})$. Then from the standard matrix theory of Markov chains results that the Taylor coefficient of $[v_0^{n_0} v_1^{n_1} \ldots v_{\ell-1}^{n_{\ell-1}}]$ in the quantity:

$$(1,1,\ldots,1)\,(1 - VM^{(k)})^{-1}\,(\frac{1}{2^k}, \frac{1}{2^k}, \ldots, \frac{1}{2^k})^t \tag{1}$$

represents the probability that a random string of length $n = n_0 + n_1 + \cdots + n_{\ell-1} + k - 1$ has n_j occurrences of block with number j, for all j.

Let $(N_0, N_1, \ldots, N_{\ell-1})$ denote the random vector where N_j represents the (random) number of times state j is reached in a sequence of n transitions of the Markov chain $\mathcal{M}^{(k)}$. The expectation of each N_j is asymptotically, for large n, $\sim n/2^k$: The matrix being doubly stochastic, the stationary probability of each state is $1/2^k$. In other words, a random n–string tends to contain each k block about $n/2^k$ times.

Stronger normality results follow if we appeal to the standard theory of limit theorems for Markov chains. We then find that, in the limit, vector $(N_0, N_1, \ldots, N_{\ell-1})$ obeys a limiting multivariate Gaussian distribution. This strongly suggests that deviation from expected values for occurrences of *any* "small" block should be of small amplitude.

The above observations are useful when k stays fixed while n varies and they may be used to derive aproximate numerical estimations in this case. However, for the purpose of proving Theorem 1, we must let k vary and approach $\lg n$ so that we need uniform error terms in n and k, for the class of Markov chains $\mathcal{M}^{(k)}$. We shall therefore need to continue in Section 3 with another route, less probabilistic and more analytic.

As a first result here, notice that restrictions of the "universal" generating function (1) give almost all conceivable generating functions of interest, when counting

occurrences of blocks in words. In particular, we expect such generating functions to be rational. Let $\pi_{u,n}^{(r)}$ be the probability that a random string of length n contains the pattern u exactly r times. The associated bivariate generating function:

$$P_u(z,v) = \sum_{n,r\geq 0} \pi_{u,n}^{(r)} v^r z^n \tag{2}$$

is obtained from Eq. (1) by the substitution $v_j \mapsto zv$ and $v_i \mapsto z$ for $i \neq j$ with j being the number whose binary representation (with length k) coincides with the string u. Thus $P_u(z,v)$ is a linear fractional transformation of v with coefficients that are rational in z.

Proposition 1. *The bivariate generating function for the probabilities of occurrence of pattern u is of the form*

$$P_u(z,v) = \frac{A_u(z) + vB_u(z)}{C_u(z) + vD_u(z)},$$

for some rational functions $A_u(z)$, $B_u(z)$, $C_u(z)$ and $D_u(z)$.

In particular, for $r \geq 1$, we find that the generating functions

$$P_u^{(r)}(z) = \sum_{n\geq 0} \pi_{u,n}^{(r)} z^n \tag{3}$$

are given by

$$P_u^{(r)}(z) = \alpha(z)\big(\beta(z)\big)^r, \tag{4}$$

for some rational functions $\alpha(z)$ and $\beta(z)$ that depend on pattern u.

The purpose of the next section is to make explicit the dependency of those functions with respect to the structure of the pattern using the correlation polynomials of Guibas and Odlyzko.

3. Generating Functions for Pattern Occurrences

This section relies heavily on explicit expressions for generating functions related to occurrences of patterns in strings. These were derived by Guibas and Odlyzko [GO78], [GO81a], [GO81b] and later surveyed by Odlyzko [Od84]. Our notations follow Odlyzko's survey, except that the variable in our generating function is z while he uses z^{-1}. Thus our generating functions are the usual ones, and they are analytic at the origin while Guibas and Odlyzko's are analytic at ∞.

Let $u = u_1 u_2 \cdots u_k$ be a binary string of length k. The primary notion is that of the *correlation polynomial* associated to u. This is a polynomial $p(z) \equiv p_u(z)$ of degree $k-1$, such that $p(0) = 1$; the correlation polynomial has 0–1 coefficients given by†

$$[z^\ell]p(z) = 1 \qquad \text{if} \qquad u_1 u_2 \cdots u_{k-\ell} = u_{\ell+1} u_{\ell+2} \cdots u_k \tag{5}$$

† We let as usual $[z^n]f(z)$ denote the coefficient of z^n in the Taylor expansion of $f(z)$ at the origin.

and $[z^\ell]p(z) = 0$ if the condition in (5) is not satisfied. In other words, the correlation polynomial describes the way the pattern "matches" slided versions of itself. For instance, the correlation polynomial of u ='00100100' is $p(z) = 1 + z^3 + z^6 + z^7$. Given a a string u, we define the following sets of binary strings:

1. The set $\mathcal{F}_u$ is the set of binary strings that end with u and contain only a single occurrence of u.
2. The set $\mathcal{G}_u$ is the set of strings that start with u, end with u and contain exactly two occurrences of u. Note that the two occurrences of u are allowed to overlap.
3. The set $\mathcal{H}_u$ is the set of strings that start with u and contain only one occurrence of u.

If $\mathcal{L}$ is a set of strings, we let $L(z)$ denote the generating function of $\mathcal{L}$, in the usual sense of combinatorial analysis. Thus $[z^n]L(z)$ is the number of strings in set $\mathcal{L}$. Observe that, since there are 2^n binary strings of size n, $[z^n]L(\frac{z}{2})$ is also the *probability* that a random string of length n belongs to $\mathcal{L}$.

Guibas and Odlyzko have provided expressions for the generating functions of set $\mathcal{F}_u$ and $\mathcal{G}_u$, which in our notations read as

$$F_u(z) = \frac{z^k}{z^k + (1-2z)p(z)} \qquad \text{and} \qquad G_u(z) = z^k \frac{z^k + (1-2z)(p(z)-1)}{z^k + (1-2z)p(z)} \tag{6}$$

with still $p(z) \equiv p_u(z)$ the correlation polynomial of u. These are equations (4.5) and (4.10) in [Od84].

Now comes an easy combinatorial argument. First, let $\tilde{u}$ denote the mirror image of u (elements of u are taken in reverse order). There is a clear bijection between $\mathcal{H}_u$ and $\mathcal{F}_{\tilde{u}}$. Also from the definition of the correlation polynomial, it immediately results that $p_{\tilde{u}}(z) = p_u(z)$. Thus, $F_u(z)$ is also the generating function of the set $\mathcal{H}_u$.

Next observe that there is a direct bijection between the following two sets: (i) the set $\mathcal{O}_u^r$ of strings containing r possibly overlapping occurrences of pattern u; (ii) the set of $r+2$ tuples $\mathcal{F}_u \times (\mathcal{G}_u)^r \times \mathcal{H}_u$. Furthermore under this bijection, there corresponds to a string of length n in $\mathcal{O}_u^r$ a tuple of strings with total length $n + kr$. Thus, from standard combinatorial analysis (products of sets correspond to products of generating functions etc.), see *e.g.* [GJ83], we find

$$O_u^{(r)}(z) = z^{-kr}(F_u(z))^2(G_u(z))^r. \tag{7}$$

Equations (6) and (7) thus provide for the explicit form of the generating function of probabilities $\pi_{u,r}^{(r)}$ since, from a previous observation, $P^{(r)}(z) = O_u^{(r)}(\frac{z}{2})$.

Proposition 2. *The generating function $P^{(r)}(z)$ for probabilities of a pattern u occurring k times is given by*

$$P_u^{(r)}(z) = 2^k z^k \frac{[z^k + (1-2z)(p(z)-1)]^{r-1}}{[z^k + (1-2z)p(z)]^{r+1}} \qquad \text{for } r \geq 1 \tag{8a}$$

where $p(z)$ is the correlation polynomial of string u.

The generating function $P_u^{(0)}(z)$ is also found from [Od84] to be

$$P_u^{(0)}(z) = \frac{p(z)}{z^k + (1-2z)p(z)}. \tag{8b}$$

4. Saddle Point Estimates

We now have at our disposal the explicit form of Proposition 2 , Eq. (8) for generating functions of probabilities. One can return to the probabilities themselves by means of Cauchy's theorem,

$$\pi_{u,n}^{(r)} \equiv [z^n] P_u^{(r)}(z) = \frac{1}{2i\pi} \int_{O+} P_u^{(r)}(z) \frac{dz}{z^{n+1}}. \tag{9}$$

We shall get bounds on the probabilities, when r is far from the mean – namely $n/2^k$ –, by estimating the integral in (9) along a circle $|z| = R$, where R is chosen so as to traverse an approximate saddle point of the integrand. We shall find that, in our range of values of r, k and n, it is sufficient to take $R = 1 \pm \epsilon$ (with adequate $\epsilon \to 0$), and use trivial bounds on the integral. This leads to uniform exponential tail results for the probabilities: These are summarised by Equations (18) and (23) below. In the next section, we shall see how to derive discrepancy estimates from there.

In the sequel, n is large and tends to infinity. The block lengths we consider are $k = k(n)$ with

$$k(n) = \lfloor \lg n - \lg\lg n - 2\lg\lg n \rfloor.$$

A pattern† u (block) of length k in a random string of length n has an expected number of occurrences that is asymptotic to $j_0(n) = n/2^k$. We are interested in the probabilities that the random variable J_n representing this number of occurrences deviates from the mean. Set

$$j = \frac{n}{2^k} + \delta \frac{n}{2^k}, \tag{10}$$

where $\delta \in [-1, +1]$. We need estimates on the probabilities that $J_n < j$ when $\delta < 0$ (lower tail) and $J_n > j$ when $\delta > 0$ (upper tail). Thus, we need to estimate

$$L_n(\delta) = \Pr\{J_n < j\} \quad (\delta < 0), \qquad \text{and} \qquad U_n(\delta) = \Pr\{J_n > j\} \quad (\delta > 0).$$

These quantities are sums of the $\pi_{u,n}^r$ probabilities defined earlier:

$$L_n(\delta) = \sum_{r<j} \pi_{u,n}^{(r)} \qquad U_n(\delta) = \sum_{r>j} \pi_{u,n}^{(r)}. \tag{11}$$

We shall use integral representation (9) to evaluate these sums.

For $k = k(n)$ and δ in the fixed interval, all subsequent estimates are uniform in n and δ, and implied constants in $O(.)$ notations are absolute constants. Now comes a batch of notations. We set $\mu = 2^k$ so that

$$\mu = 2^k = \theta_0 \frac{n}{\lg n (\lg\lg n)^2} \qquad \text{with } \frac{1}{2} \le \theta_0 \le 2.$$

We can rewrite Eq. (8) as

$$P^{(r)}(z) \equiv P_u^{(r)}(z) = 2^k a(z) b(z)^{r-1},$$

† Throughout this section we omit all subscripts u in formulæ for readability.

with

$$a(z) = \frac{z^k}{Q^2(z)}, \qquad b(z) = 1 + \frac{2^k(z-1)}{Q(z)}, \tag{12a}$$

where

$$Q(z) = z^k + 2^k(1-z)p(\frac{z}{2}). \tag{12b}$$

We have obviously

$$b(1) = a(1) = 1, \qquad \text{and} \qquad b'(1) = \mu = 2^k. \tag{12c}$$

Upper Tail. There δ is strictly positive, accordingly $j > n/2^k$, and from (9), (11), (12), we find

$$U_n(\delta) = \frac{1}{2i\pi} \int_{O+} 2^k a(z) \frac{b^j(z)}{1-b(z)} \frac{dz}{z^{n+1}}. \tag{13}$$

We propose to evaluate the integral in (13) using the contour

$$|z| = 1 - \epsilon, \qquad \epsilon = \epsilon(n) = \frac{\lg n \lg\lg n}{n} \tag{14}$$

whose choice is dictated by a saddle point heuristic. Provided we check that the integrand in (13) is analytic for $|z| < 1$, trivial bounds on the integral lead to

$$U_n(\delta) \leq 2^k \frac{a(1-\epsilon)}{1-b(1-\epsilon)} b^j(1-\epsilon)(1-\epsilon)^{-n}. \tag{15}$$

The analyticity condition that justifies (15) is given by the following lemma.

Lemma 1. *For large enough patterns ($k \geq k_0$), the polynomial $Q(z)$ has no zeroes in the domain $|z| < 1 + \frac{1}{2k}$.*

Proof. Using the substitution $z/2 = 1/w$, the equation $Q(z) = 0$ is equivalent to $1 + (w-2)p(1/w) = 0$. Following again Guibas and Odlyzko [GO78, Lemma 3], we find that this equation has only one zero in the domain $|w| \geq 1.7$. Applying Lemma 4 in [GO78], we obtain for the zero w (and k large enough) $|w| \leq 2 - (1/k)$. Hence, for $k \geq k_0$, the equation $Q(z)$ has no solution satisfying $|z| \leq 1 + \frac{1}{2k} < \frac{2}{2-(1/k)}$. ■

Remark that $\mu\epsilon = \theta_0/(\lg\lg n)$ tends to 0 as $n \to \infty$. Estimates that follow are stated for values of functions $a(z)$ and $b(z)$ at $1 \pm \epsilon$ since they will be used later.

$$\begin{cases} a(1 \pm \epsilon) = 1 + O(\mu\epsilon) \\ b(1 \pm \epsilon) = 1 \pm \mu\epsilon + O(\mu^2\epsilon^2) \\ b^j(1 \pm \epsilon) = \exp\left(\pm \epsilon\mu j + O(j\mu^2\epsilon^2)\right) \\ (1+\epsilon)^{-n} = \exp\left(\mp n\epsilon + O(n\epsilon^2)\right). \end{cases} \tag{16}$$

These bounds all follow by inspection from the explicit forms (12), and the observation that

$$Q(1 \pm \epsilon) = (1 \pm \epsilon)^k \mp \mu\epsilon p(\frac{1 \pm \epsilon}{2}) = 1 + O(\mu\epsilon), \tag{17}$$

since the correlation polynomial $p(z)$ has 0–1 coefficients. Applying estimates (16) to bound (14), we get an upper bound on $U_n(\delta)$ in the form

$$U_n(\delta) = \frac{2^k}{\epsilon\mu}\frac{1+O(\mu\epsilon)}{1+O(\mu\epsilon)}\exp\big(-\delta n\epsilon + O(\frac{n^2}{\lg n(\lg\lg n)^2}\epsilon^2)\big).$$

This gives our main upper bound for the upper tail:

$$U_n(\delta) \le \exp\big(-\delta\lg n\lg\lg n + c_1\lg n\big) \tag{18}$$

for some absolute constant c_1 and n large enough.

Lower Tail. Now δ is strictly negative, and accordingly $j < n/2^k$. From (9), (11), (12), we find

$$L_n(\delta) = \frac{1}{2i\pi}\int_{O+} 2^k a(z)\frac{1-b^j(z)}{1-b(z)}\frac{dz}{z^{n+1}}. \tag{19}$$

We evaluate the integral in (19) using the contour

$$|z| = 1+\epsilon, \qquad \epsilon = \epsilon(n) = \frac{\lg n\lg\lg n}{n}. \tag{20}$$

If we know that $b(1+\epsilon) > 1$ and that $a(z)$, $b(z)$ have no poles in the domain $|z| \le 1+\epsilon$, then (19) is upper bounded by

$$L_n(\delta) \le j2^k a(1+\epsilon)b^j(1+\epsilon)(1+\epsilon)^{-n}. \tag{21}$$

The transition from (20) to (21) is obtained via Lemma 1: Note that $b(1+\epsilon) > 1$ is equivalent to $Q(1+\epsilon) > 0$, which follows from Lemma 1, $Q(1) > 0$ and $1+\epsilon < 1+\frac{1}{2k}$. The conclusion for the lower tail comes directly from (21) and (16), so that

$$L_n(\delta) = j2^k(1+O(\mu\epsilon))\exp(\epsilon\mu j + O(j\mu^2\epsilon^2) - n\epsilon + O(n\epsilon^2)). \tag{22}$$

But $j = O(n)$ and $2^k = O(n^2)$, so that finally

$$L_n(\delta) = \exp(\delta n\epsilon + c_2\lg n) \tag{23}$$

for some absolute constant c_2 and large enough n.

5. Discrepancies of Finite and Infinite Strings

From Equations (18) and (23), we have exponential tail estimates for the probability distribution of occurrences of a single pattern u with length k when $k = k(n)$. Returning to discrepancies is easy: if δ is > 0, and b represents a random string of length n, we have

$$\begin{aligned}\Pr\{2^k D_k(b) > \delta\} &\le \sum_{|u|=k}\Pr\{|\Omega(b,u) - n| > \delta\} \\ &\le 2^k(L_n(-\delta) + U_n(+\delta))\end{aligned} \tag{24}$$

Thus, by (4) and estimates (18), (23), we obtain

Theorem 2. *Let $k = k(n) \equiv \lfloor \lg n - \lg\lg n - 2\lg\lg\lg n \rfloor$ and δ be such that $0 < \delta < 1$. Then the probability distribution of dicrepancy over the set of strings b of length n satisfies*

$$\Pr\{2^k D_k(b) > \delta\} < n^{-\delta \lg\lg n + c}. \tag{25}$$

where c denotes an absolute constant.

In essence, the result can be extended to smaller values of k using the following lemma from [KT85] (Proposition 2.1 there):

Lemma 2. *Let s and t be two integers such that $s \leq t \leq \lg n$. Then, for an arbitrary string b of length n, we have*

$$2^s D_s(b) \leq 2^t D_t(b) + 2^s \frac{t}{n}. \tag{26}$$

Applying Lemma 2 to Theorem 1 shows that bound (25) actually holds not only for $k = k(n)$ but also for all $k < k(n)$. Notice that a slight improvement of our result is possible, and the $2\lg\lg\lg n$ coud be replaced by $(1+\eta)\lg\lg\lg n$ for any $\eta > 0$.

We are now ready to complete the proof of Theorem 1. Choose δ a function of n:

$$\delta = \delta(n) = (\lg\lg n)^{-1/2} \qquad (n \geq 4)$$

and observe that

$$\sum_n n^{-\delta(n)\lg\lg n + c} < \infty.$$

Hence, by the Borel–Cantelli lemma (cf [Fe68, p.201]), we obtain Theorem 1.

References

[Fe68] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd Edition, J. Wiley (1968).

[FKT86] Ph. Flajolet, P. Kirschenhofer and R. Tichy, "Discrepancy of Sequences in Infinite Strings", preprint 1986.

[GJ83] I. Goulden and D. Jackson, *Combinatorial Enumerations*, J. Wiley (1983).

[GO78] L. J. Guibas and A. M. Odlyzko, Maximal prefix-synchronized codes, *SIAM J. Appl. Math.* **35** (1978), 401-418.

[GO81a] L. J. Guibas and A. M. Odlyzko, Periods in strings, *J. Comb Th. (A)* **30** (1981), 19-42.

[GO81b] L. J. Guibas and A. M. Odlyzko, Strings overlaps, pattern matching and nontransitive games, *J. Comb Th. (A)* **30** (1981) 183-208.

[Hl79] E. Hlawka, *Theorie der Gleichverteilung*, Bibl. Inst., Mannheim-Wien-Zürich (1969).

[Kn68] D. E. Knuth, *The Art of Computer Programming, Vol 1: Fundamental Algorithms*, Addison-Wesley (1968).

[Kn69] D. E. Knuth, *The Art of Computer Programming, Vol 2: Semi-Numerical Algorithms*, Addison-Wesley (1969).

[KN74] L. Kuipers and H. Niderreiter, *Uniform Distribution of Sequences*, J. Wiley (1974).

[KT85] P. Kirschenhofer and R. Tichy, Some distribution properties of 0-1 sequences, *Manuscripta Mathematica* **54** (1985), 205-219.

[Od84] A. M. Odlyzko, Enumeration of Strings, in *Combinatorial Algorithms on Words*, A. Apostolicao and Z. Galil Eds, NATO ASI Series F12, Springer (1984).

[Re80] P. Révész, Strong Theorems in Coin Tossing, Proc. 1978 Int. Congress of Mathematicians, Helsinki (1980).

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique