The Computational Complexity Column

Eric Allender
Rutgers University, Department of Computer Science
Piscataway, NJ 08855 USA
allender@cs.rutgers.edu

With this issue of the Bulletin, my tenure as editor of the Computational Complexity Column comes to an end. Lance Fortnow will edit this column beginning with the June issue, and we can look forward to a long series of informative columns under his direction. It has been a pleasure, and I am glad that the column will be in good hands.

Although columns appearing in this space focus on complexity classes and lower bounds, it is important to realize that algorithmic techniques (upper bounds) frequently find application in proving lower bounds, and in proving new relationships among complexity classes. This is especially evident in the exciting developments in the field of derandomization. We are fortunate, in this issue of the Bulletin, to have an overview of some important algorithmic and techniques by a leading figure in the field of derandomization.

Low-discrepancy sets for high-dimensional rectangles: a survey

Aravind Srinivasan [1]

Abstract

A sub-area of *discrepancy theory* that has received much attention in computer science recently, is that of explicit constructions of low-discrepancy point sets for various types of rectangle families in high dimension. This research has led to interesting applications in error-control coding, distributed protocols, Web document filtering, derandomization, and other areas. We give a short survey of this area here.

1 Introduction

One major approach in the general area of derandomization is that of *explicit* or *efficient constructions.* This is the problem of giving efficient deterministic constructions of various discrete structures (e.g., error-correcting codes, hash function families) whose existence has been shown (typically via probabilistic arguments). Our notion of efficiency throughout will be that of time polynomial in the size of some natural description of such a structure. A particular class of such structures that has received much attention in the last decade, is that of *pseudorandom generators* or *low discrepancy sets* for rectangle families. (The reader is referred to [9, 20] for investigations into discrepancy theory.) These studies have led to interesting applications in coding theory, derandomization, fault-tolerant leader election protocols, and testing Web documents for similarity. We give a short survey of this area here. This survey certainly does not cover all the interesting research in this field; our purpose is to give a sample of the key ideas here, so that the interested reader can delve into the relevant literature for further information.

[1] Bell Laboratories, Lucent Technologies, 600-700 Mountain Ave., Murray Hill, NJ 07974-0636, USA. srin@research.bell-labs.com

Given positive integers n and m, let $\mathcal{C}_m^n$ be the following family of n-dimensional *combinatorial rectangles*:

$$\mathcal{C}_m^n = \{S_1 \times \cdots \times S_n : \forall i,\ S_i \subseteq \{0, 1, \ldots, m-1\}\}.$$

Given a finite (multi-)set A, let $U(A)$ denote the uniform distribution on A. A finite multiset $S \subseteq \{0, 1, \ldots, m-1\}^n$ is defined to be an *ϵ-approximation* for $\mathcal{C}_m^n$, if for $\vec{X}$ sampled from $U(S)$, we have for all $R = S_1 \times \cdots \times S_n \in \mathcal{C}_m^n$ that

$$|\Pr(\vec{X} \in R) - (\prod_i |S_i|)/m^n| \leq \epsilon. \tag{1}$$

In other words, a random sample from S looks approximately like a random element of $\{0, 1, \ldots, m-1\}^n$, for each rectangle in $\mathcal{C}_m^n$: the *discrepancy of S w.r.t. $\mathcal{C}_m^n$*,

$$\max_{R=S_1 \times \cdots \times S_n \in \mathcal{C}_m^n} |\Pr(\vec{X} \in R) - (\prod_i |S_i|)/m^n|, \tag{2}$$

is at most ϵ. Furthermore, we would like a low-discrepancy (multi-)set S to be *constructible* in the following way. Suppose the ith element of S is the vector $s_{i,1}s_{i,2}\cdots s_{i,n} \in \{0, 1, \ldots, m-1\}^n$. We desire a deterministic algorithm, which, given any i and j, can construct $s_{i,j}$ in time poly$(\log|S| + \log n + \log m)$; note that representing i and j in the natural way needs $\log|S| + \log n$ bits. For all the constructions of low-discrepancy sets surveyed here, such a constructibility property will be true.

A natural subclass of $\mathcal{C}_m^n$ is the following family of *geometric rectangles*:

$$\mathcal{G}_m^n = \{[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] :\ \forall i, ((a_i \leq b_i) \wedge (a_i, b_i \in \{0, 1, \ldots, m-1\}))\}.$$

(This family is sometimes also referred to as *axis-parallel rectangles*.) We can analogously define low-discrepancy sets S for $\mathcal{G}_m^n$. This survey will consider the efficient construction of "small" low-discrepancy sets S as defined above. We shall consider variants of this basic problem, and applications thereof.

It is not hard to show that if C is a sufficiently large constant and if we choose Cmn/ϵ^2 points independently from $U(\{0, 1, \ldots, m-1\}^n)$, then this random multi-set Z of points is an ϵ-approximation for $\mathcal{C}_m^n$ with high probability. Briefly, consider any fixed combinatorial rectangle $R = S_1 \times \cdots \times S_n \in \mathcal{C}_m^n$. A Chernoff-type argument [12, 14] shows that for any desired constant $C' > 0$, there is a large enough value for C such that (1) fails with probability at most $2^{-C'mn}$ for Z. Now, $\mathcal{C}_m^n$ has 2^{mn} elements; so, the probability that Z is not an ϵ-approximation for $\mathcal{C}_m^n$ is at most $2^{(2-C')mn}$, a tiny quantity if $C' > 2$.

Thus, there certainly *exists* an ϵ-approximation for $\mathcal{C}_m^n$ with cardinality $O(mn/\epsilon^2)$. The primary question is: is there an efficient deterministic construction of such a multiset? (In fact, a construction of such a multiset with cardinality bounded by any fixed polynomial of m, n, and ϵ^{-1} would be a major breakthrough.) "Deterministic construction" here, and throughout this survey, refers to the notion of constructibility presented above after (2).

We next present two useful families of rectangles that are closely related to $\mathcal{C}_m^n$ and $\mathcal{G}_m^n$ respectively. Let us say that $R = S_1 \times S_2 \times \cdots \times S_n \in \mathcal{C}_m^n$ is *trivial* in dimension i iff $S_i = \{0, 1, \ldots, m-1\}$; R is nontrivial in dimension i otherwise. Let $\mathcal{C}_{m,k}^n$ denote the subset of $\mathcal{C}_m^n$ containing the rectangles that are nontrivial in at most k dimensions. In a natural way, one can also define $\mathcal{G}_{m,k}^n \subseteq \mathcal{G}_m^n$, and ϵ-approximations for $\mathcal{C}_{m,k}^n$ and $\mathcal{G}_{m,k}^n$. A probabilistic construction shows that there exists an ϵ-approximation of cardinality

$$O\left(\left(mk + \log\left(\binom{n}{k}\right)\right)/\epsilon^2\right) \leq O((mk + k\log n)/\epsilon^2)$$

for $\mathcal{C}_{m,k}^n$ (and hence for $\mathcal{G}_{m,k}^n$). The major open question here is to construct an ϵ-approximation of cardinality poly$(m + k + \log n + \epsilon^{-1})$ for $\mathcal{C}_{m,k}^n$, or at least for $\mathcal{G}_{m,k}^n$ as a first step. We will illustrate the utility of $\mathcal{C}_{m,k}^n$ and $\mathcal{G}_{m,k}^n$ later through some applications.

We now start by considering the basic but very important case where $m = 2$, in §2. The general cases of geometric rectangles and combinatorial rectangles are then studied in §3 and §4 respectively. Sample applications are presented throughout. We conclude with some open problems in §5.

2 The case $m = 2$

Note that $\mathcal{C}_m^n = \mathcal{G}_m^n$ if $m = 2$. Breakthrough work on low-discrepancy sets for $\mathcal{C}_2^n$ was presented in [21] through a notion of *small-bias sample spaces*, which will be introduced shortly. In particular, the problem of constructing an ϵ-approximation for $\mathcal{C}_{2,k}^n$ of cardinality poly$(k + \log n + \epsilon^{-1})$ was settled in their work. Let us briefly survey some of the key ideas and applications of their and related results.

2.1 From bias to approximation

Let $[\ell]$ denote the set $\{1, 2, \ldots, \ell\}$; $\log x$ will denote $\log_2 x$. To construct an ϵ-approximation for $\mathcal{C}_{2,k}^n$, we want an efficient construction of a "small" multiset $S \subseteq \{0,1\}^n$ such that for a vector $(X_1, X_2, \ldots, X_n)$ sampled according to $U(S)$,

$$\forall j \in [k]\ \forall \{i_1, i_2, \ldots, i_j\} \subseteq [n]\ \forall (b_1, b_2, \ldots, b_j) \in \{0,1\}^j,\ |\Pr(\bigwedge_{\ell=1}^{j} (X_{i_\ell} = b_\ell)) - 2^{-j}| \leq \epsilon. \quad (3)$$

(Setting $k = n$ will yield an ϵ-approximation for $\mathcal{C}_2^n$.)

A key idea in [21] is to approach this through *small-bias sample spaces*. Given a distribution D on $\{0,1\}^n$ and any $T \subseteq [n]$, define

$$\text{bias}_D(T) = |\Pr((\bigoplus_{i \in T} X_i) = 1) - \Pr((\bigoplus_{i \in T} X_i) = 0)|;$$

$(X_1, X_2, \ldots, X_n) \in \{0,1\}^n$ is a random vector chosen according to D, and "$\bigoplus$" is the usual XOR operation. Define D to be *k-wise ϵ-biased* if $\text{bias}_D(T) \leq \epsilon$ for all T such that $|T| \leq k$; D is simply called *ϵ-biased* if it is n-wise ϵ-biased. Via Fourier analysis, it is shown in [21] that if D is k-wise ϵ-biased, then our desired property (3) also holds. (Thus, in particular, we have for all distinct indices $i_1, i_2, \ldots, i_k$ that

$$\left(\sum_{(b_1, b_2, \ldots, b_k) \in \{0,1\}^k} |\Pr(\bigwedge_{\ell=1}^{k} (X_{i_\ell} = b_\ell)) - 2^{-j}|\right) \leq 2^k \epsilon. \quad (4)$$

It is shown in [3, 21] that the left-hand-side of (4) can in fact be upper-bounded by $2^{k/2}\epsilon$.)

Another interesting result of [21] (using a construction of [1]) is the following reduction. Suppose we have an efficient construction of a multiset $S \subseteq \{0,1\}^n$ with $|S| \leq F(n, \epsilon)$, such that $U(S)$ is ϵ-biased. Then, for any $k \in [n]$, there is also an efficient construction of an $S' \subseteq \{0,1\}^n$ with $|S'| \leq F(O(k \log n), \epsilon)$, such that $U(S')$ is k-wise ϵ-biased.

2.2 Constructing small-bias spaces

Having seen that small-bias spaces suffice to solve our ϵ-approximation problem for $m = 2$, we now present a construction of small-bias spaces due to [3]. We will also survey another approach to small-bias spaces from [21], and show a connection to error-correcting codes.

Suppose $m = \lceil \log(n/\epsilon) \rceil$. The following construction of a multiset $S \subseteq \{0,1\}^n$ such that $U(S)$ is ϵ-biased, is due to [3]. We will describe S by showing how to generate an element according to $U(S)$; it will be immediate then that $|S| = 2^{2m}$, i.e., that $|S| = O(n^2/\epsilon^2)$. Choose elements x, y of the finite field $GF[2^m]$ independently and uniformly at random. Next generate a vector $X = (X_1, X_2, \ldots, X_n) \in \{0,1\}^n$ by defining $X_i = (x^i \cdot y) \bmod 2$ for each i. Here, x^i denotes raising x to the ith power in the field $GF[2^m]$; $x^i \cdot y$ denotes interpreting x^i and y as m-bit strings and taking their dot product. This vector X represents a vector chosen according to $U(S)$.

As mentioned above, it is immediate that $|S| = 2^{2m}$. Let us quickly see why the above distribution on $(X_1, X_2, \ldots, X_n)$ is ϵ-biased; we need to show that $\text{bias}_D(T) \leq \epsilon$ for all $T \subseteq [n]$. Fix any $T = \{i_1, i_2, \ldots, i_j\} \subseteq [n]$. By elementary properties of sums over the field $GF[2^m]$, we have

$$\bigoplus_{\ell \in T} X_\ell = (x^{i_1} + x^{i_2} + \cdots + x_{i_j}) \cdot y \bmod 2, \tag{5}$$

where the sum $t(x) = x^{i_1} + x^{i_2} + \cdots + x_{i_j}$ is taken in $GF[2^m]$. Now if $t(x)$ is nonzero for our random x, it is easy to check that the right-hand-side of (5) is equally likely to be 0 or 1. On the other hand, if $t(x)$ is zero, the right-hand-side of (5) is 0. Since $t(z)$ is a polynomial of degree at most n, it has at most n roots; thus, $\Pr(t(x) = 0) \leq n/2^m \leq \epsilon$. This, combined with the above observations, helps show that $\text{bias}_D(T) \leq \epsilon$.

Thus we have an efficient construction of a multiset $S \subseteq \{0,1\}^n$ with $|S| \leq O(n^2/\epsilon^2)$ such that $U(S)$ is ϵ-biased; a different such construction with $|S| \leq O(n/\epsilon^3)$ is presented in [2]. For k-wise ϵ-bias, the reduction of [21] mentioned at the end of §2.1 yields constructions of size

$$O(\min\{((k \log n)/\epsilon)^2, (k \log n)/\epsilon^3\}). \tag{6}$$

We now sketch a different approach to small-bias spaces due to [21], which has connections to coding theory. An efficient construction of a multi-set $A \subseteq \{0,1\}^n$ with $|A| = \text{poly}(n)$ is shown in [21], such that for a certain constant $\beta > 0$,

$$\text{for all nonempty } T \subseteq [n], \ \Pr((\bigoplus_{i \in T} X_i) = 1) \geq \beta; \tag{7}$$

$(X_1, X_2, \ldots, X_n)$ is a vector chosen according to $U(A)$. (As expected, a random construction helps show the *existence* of such an A; the challenge is in efficient construction.) This construction is used as a key building block in [21] in designing small-bias sample spaces.

As shown in [21], such a set A has a close connection to *linear codes.* Recall that a linear code over a field $\mathbf{F}$ is basically given by a *generator matrix* $G \in \mathbf{F}^{n \times m}$, for some n and m with $m \geq n$.

Given a *message* $a \in \mathbf{F}^n$, it is encoded by the *codeword* aG. Clearly, the set of all codewords form a subspace; hence, the *minimum distance* of the coding represented by G (the maximum, over all distinct pairs of codewords x, y, of the Hamming distance between x and y) equals the *minimum weight* (the minimum, over all nonzero codewords x, of the number of nonzero symbols in x) of the code. The minimum distance is a fundamental parameter of a code. Let us specialize the discussion to the case where $\mathbf{F} = GF[2]$. Suppose we have an efficient construction of a matrix $G \in (GF[2])^{n \times m}$, with $m \leq \text{poly}(n)$ and with minimum distance at least βm, for some constant $\beta \in (0, 1)$. If we consider the set A of m column-vectors of G, a moment's reflection shows that (7) holds. Indeed, matrices G with these properties with, in fact, $m = O(n)$, are efficiently constructible [17]. Conversely, construction of an A satisfying (7) yields a linear code. Thus we see a connection to coding theory. These connections are further explored in [2].

2.3 Some applications

Suppose independent random bits $Y_1, Y_2, \ldots, Y_n$ with $\Pr(Y_i = 1) = 1/2$ for each i, are used by some randomized algorithm. Now suppose that random bits X_i are "almost k-wise independent" in the sense of (3). Then, one may expect that if we take k and ϵ^{-1} sufficiently large based on the analysis of the given randomized algorithm, the analysis will work out if we use $(X_1, X_2, \ldots, X_n)$ in place of $(Y_1, Y_2, \ldots, Y_n)$. In such cases, one may choose the X_i from a small sample space (of cardinality such as (6)), and deterministic exhaustive search of this space will lead us to a derandomization. This has proved to be a powerful idea for derandomization over the last several years. As a quick example, suppose we are given an n-vertex, m-edge undirected graph $G = (V, E)$. Finding a *maximum cut* in G, i.e., a partition of V into two subsets such that a maximum number of edges "cross" this cut, is a well-known NP-hard problem. Suppose we just wish to find a "large" cut in G; a classical algorithm finds a cut with at least $m/2$ edges crossing the cut. Here is an efficient NC algorithm that finds a cut with at least $m(1-\delta)/2$ edges crossing the cut, for a given constant $\delta > 0$. Suppose a random bit-vector $X = (X_1, X_2, \ldots, X_n)$ satisfies (3), with $k = 2$ and $\epsilon = \delta/4$. Consider a randomized algorithm that samples such a vector X and constructs a cut in G by setting $V_\ell = \{i \in V : X_i = \ell\}$, for $\ell = 0, 1$. The probability that a given edge $\{i, j\}$ is cut is given by

$$\Pr(X_i = 1,\ X_j = 0) + \Pr(X_i = 0,\ X_j = 1) \geq (1/4 - \epsilon) + (1/4 - \epsilon) = (1 - \delta)/2.$$

Thus, by linearity of expectation, the expected number of edges cut is at least $m(1-\delta)/2$. So, as in (6), we can construct a sample space of size $O((\log n)/\epsilon^3) = O(\log n)$ such that at least one sample from this space will lead to a cut in which at least $m(1-\delta)/2$ edges cross the cut. Exhaustive search (e.g., using $(n+m)\log n$ processors and $O(\log n)$ time in total) of such a space leads to an efficient deterministic NC algorithm for our problem. See [21, 11] for extensions to the more general problem of finding heavy codewords in linear codes, and for more efficient algorithms for smaller values of δ.

The above application shows how small-bias spaces, combined with the fact that randomized algorithms are typically robust to small changes in the underlying probabilities, leads to fruitful derandomization schemes. Furthermore, the work of [4] presents a nice *approximate method of conditional probabilities*, which works well with small-bias sample spaces to lead to efficient derandomization approaches. The reader is referred to [4, 6, 11] for a study of this useful approach.

Finally, explicit constructions of small-bias spaces have had an impact on constructing various combinatorial objects; one of these that has seen many applications (e.g., in computational learning theory and hardness-of-approximation results) is the following. Given integers k, n with $k \leq n$, a family $V(n,k)$ of subsets of $[n]$ is called (n,k)-universal if for any $T \subseteq [n]$ with $|T| = k$ and any $T' \subseteq T$, there is a $T'' \in V(n,k)$ such that for all $t \in T$, $t \in T''$ iff $t \in T'$. A simple random construction shows the existence of such a family $V(n,k)$ of cardinality poly$(2^k + \log n)$. Suppose a family $W(n,k)$ has the property that for some $\epsilon < 2^{-k}$, (3) holds for $(X_1, X_2, \ldots, X_n)$ sampled according to $U(W(n,k))$. (Subsets of $[n]$ are identified here with their characteristic vectors.) Then, the reader may verify that $W(n,k)$ is (n,k)-universal. So, from (6), we see that there are explicit constructions of (n,k)-universal sets with cardinality poly$(2^k + \log n)$; see [22] for an improved bound on the polynomial in this "poly$(2^k + \log n)$" construction.

The reader is referred to [21] for some further applications of small-bias spaces.

3 Geometric rectangles

Suppose we have independent random variables $Y_1, Y_2, \ldots, Y_n$ such that $\Pr(Y_i = 0) = \Pr(Y_i = 1) = 1/2$ for each i. Then, a random vector $(X_1, X_2, \ldots, X_n)$ satisfying (3) can be considered an "almost k-wise independent" analog of $\vec{Y} = (Y_1, Y_2, \ldots, Y_n)$. As seen in §2, such approximations are very useful in derandomizing, say, randomized algorithms for which $\vec{Y}$ is the vector of random bits used. However, a randomized algorithm could use a vector of independent random variables $\vec{Y}$ where each Y_i has an arbitrary distribution over a domain of size larger than 2. Formally, our "almost k-wise independence" requirement now is as follows. Suppose $Y_1, Y_2, \ldots, Y_n \in \{0, 1, \ldots, \ell - 1\}$ are *independent* random variables with arbitrary marginal (i.e., individual) distributions and joint distribution D, for an arbitrary ℓ. A multi-set $S \subseteq \{0, 1, \ldots, \ell - 1\}^n$ is called a (k, ϵ)-*approximation* for D if, for $\vec{X} = (X_1, X_2, \ldots, X_n)$ sampled from $U(S)$,

$$\forall I \subseteq [n] \ (|I| \leq k) \ \forall a_1, \ldots, a_{|I|} \in \{0, 1, \ldots, \ell - 1\}, \ |\Pr(\bigwedge_{i \in I} (X_i = a_i)) - \prod_{i \in I} Pr(Y_i = a_i)| \leq \epsilon. \quad (8)$$

Thus, $\vec{X}$ is an "almost k-wise independent" analog of $\vec{Y}$; if $|S|$ is "small", we may hope for reasonable derandomization approaches for randomized algorithms that use $\vec{Y}$ as their sequence of random variables.

As pointed out in [13], there is a single approach to all possible distributions D above, as follows. Suppose $m \geq 2$ is an integer. We can approximate the random variables Y_i by new random variables Y'_i, as follows. Let $p_{i,j} = \Pr(Y_i = j)$. Given the vector $(p_{i,0}, p_{i,1}, \ldots, p_{i,\ell-1})$, the reader can verify that we can efficiently construct a vector of non-negative reals $(p'_{i,0}, p'_{i,1}, \ldots, p'_{i,\ell-1})$ such that: **(P1)** $\sum_j p'_{i,j} = 1$; **(P2)** each $p'_{i,j}$ is an integer multiple of $1/m$, and **(P3)** for any indices $0 \leq j' \leq j \leq \ell - 1$, $|\sum_{t=j'}^{j} p'_{i,t} - \sum_{t=j'}^{j} p_{i,t}| \leq 2/m$. Now consider independent random variables $Y'_1, Y'_2, \ldots, Y'_n \in \{0, 1, \ldots, \ell - 1\}$, with $\Pr(Y'_i = j) = p'_{i,j}$. Using (P3), we can verify that for all $I \subseteq [n]$ with $|I| \leq k$ and for all $a_1, b_1, a_2, b_2, \ldots, a_{|I|}, b_{|I|} \in \{0, 1, \ldots, \ell - 1\}$ such that $a_i \leq b_i$,

$$|\Pr(\bigwedge_{i \in I} (Y_i \in [a_i, b_i])) - \Pr(\bigwedge_{i \in I} (Y'_i \in [a_i, b_i]))| \leq 2k/m. \quad (9)$$

Moreover, (P2) suggests an obvious way of generating Y'_i via a sample Z_i from $U(\{0, 1, \ldots, m-1\})$: $Y'_i = j$ iff

$$Z_i \in \Phi_j \doteq [\sum_{t=0}^{j-1} m p'_{i,t}, \ (\sum_{t=0}^{j} m p'_{i,t}) - 1].$$

(For example, suppose $p'_{i,0} = p'_{i,1} = 1/8$, $p'_{i,2} = 3/4$, and $m = 8$. Choose a sample Z_i from $U(\{0, 1, \ldots, 7\})$. If $Z_i = 0$, define $Y'_i = 0$; else if $Z_i = 1$, define $Y'_i = 1$; else define $Y'_i = 2$.) Suppose $m \geq 4k/\epsilon$; using (9) and the fact that Φ_j is not just an arbitrary subset of $\{0, 1, \ldots, m-1\}$ but is an *interval*, it is not hard to verify the following. Suppose a multi-set S is an $(\epsilon/2)$-approximation for $\mathcal{G}^n_{m,k}$. Choose a random sample $(Z'_1, Z'_2, \ldots, Z'_n)$ from $U(S)$, and define a vector $\vec{X} = (X_1, X_2, \ldots, X_n) \in \{0, 1, \ldots, \ell - 1\}^n$ by "$X_i = j$ iff $Z'_i \in \Phi_j$". Then, for all $I \subseteq [n]$ with $|I| \leq k$ and for all $a_1, b_1, a_2, b_2, \ldots, a_{|I|}, b_{|I|} \in \{0, 1, \ldots, \ell - 1\}$ such that $a_i \leq b_i$,

$$|\Pr(\bigwedge_{i \in I}(X_i \in [a_i, b_i])) - \Pr(\bigwedge_{i \in I}(Y_i \in [a_i, b_i]))| \leq \epsilon. \tag{10}$$

Note that this is stronger than (8).

Thus, ϵ-approximations for $\mathcal{G}^n_{m,k}$ have direct applications to approximating arbitrary independent random variables. Three constructions S_1, S_2, S_3 of such ϵ-approximations were presented in [13], with

$$|S_1| = (\log n + 2^k + 1/\epsilon)^{O(1)},\ |S_2| = (n/\epsilon)^{O(\log(1/\epsilon))},\ \text{and } |S_3| = (n/\epsilon)^{O(\log n)}. \tag{11}$$

S_2 and S_3 are also ϵ-approximations for $\mathcal{G}^n_m$. (The parameter m does not figure in (11) since, as discussed above, we may assume that $m \leq O(k/\epsilon)$.)

By utilizing and extending the results of [13], these bounds were improved to $\text{poly}(\log n + 1/\epsilon + (\lceil k/\log(1/\epsilon)\rceil)^{\log(1/\epsilon)})$ in [11]. Briefly, by extending some ideas behind the construction of S_2, it is shown in [11] that for certain values $k' = O(\log(1/\epsilon))$ and $\epsilon' = 1/\text{poly}(1/\epsilon + (\lceil k/\log(1/\epsilon)\rceil)^{\log(1/\epsilon)})$, any ϵ'-approximation for $\mathcal{G}^n_{m,k'}$ is also an ϵ-approximation for $\mathcal{G}^n_{m,k}$. Using S_1 now gives the desired construction. These constructions have been further improved in [19] to one of cardinality

$$\text{poly}(\log n + 1/\epsilon + (1/\epsilon)^{\sqrt{\log(\lceil k/\log(1/\epsilon)\rceil)}}). \tag{12}$$

Two of the applications of explicit low-discrepancy sets for geometric rectangles are as follows. First, the strong property (10) is used in [23] to develop a derandomized NC version of a hypergraph coloring algorithm presented in [23]. Second, a useful notion of pseudorandom *permutation* families, that of *approximate min-wise independent permutation families*, has been introduced in [10]. Efficient construction of such families has applications to a quantitative notion of similarity of Web documents [10]. Certain constructions of such families have been developed in [16]; via a connection to low-discrepancy sets for geometric rectangles, different constructions that have smaller size for certain ranges of the parameters of interest, have been shown in [25].

4 Combinatorial rectangles and hitting sets

This short section considers low-discrepancy sets, and a related notion of *hitting sets*, for combinatorial rectangles. Explicit ϵ-approximations for $\mathcal{C}^n_m$ with cardinality $\text{poly}(m + n + (1/\epsilon)^{O(\log(1/\epsilon))})$ were presented in [7], and were improved to $\text{poly}(m+n+(1/\epsilon)^{O(\sqrt{\log(1/\epsilon)})})$ in [19]. These approaches introduce certain new ideas and also build on [15, 18]. An ϵ-approximation for $\mathcal{C}^n_{m,k}$ with cardinality

$$\text{poly}(\log n + m^{O(\log(1/\epsilon))} + (1/\epsilon)^{\log(\lceil k/\log(1/\epsilon)\rceil)})$$

is given in [8], by extending an idea of [11].

The following reductions are also shown in [8, 19]. Suppose that for all (n, m, k, ϵ), an ϵ-approximation S' for $\mathcal{C}_{m,k}^n$ with $\log |S'| = O(\log\log n + k + \log m + \log(1/\epsilon))$ can be efficiently constructed. Then, for all (n, m, k, ϵ), the following constructions of ϵ-approximations S'' for $\mathcal{C}_{m,k}^n$ are possible, as shown respectively in [8] and [19]:

- one with $\log |S''| = O(\log\log n + \log k + \log m + \log(1/\epsilon) + \log(1/\epsilon)\log(\lceil k/\log(1/\epsilon)\rceil))$, and
- one with $\log |S''| = O(\log\log n + \log k + \log m + \log(1/\epsilon) + \log(1/\epsilon)\log\log(1/\epsilon))$.

Note that $\log |S'|$ is allowed to be quite high as a function of k. Also, constructions of such S' are indeed known for geometric rectangles: see S_1 in (11). Thus, these reductions could be potentially useful approaches to get improved bounds for ϵ-approximations for $\mathcal{C}_{m,k}^n$.

Finally, we consider the concept of *hitting sets* for combinatorial rectangles. Suppose S is an ϵ-approximation for $\mathcal{C}_m^n$, and that a random vector $\vec{X}$ is sampled from $U(S)$. It is immediate from (1) that

$$\forall R = S_1 \times \cdots \times S_n \in \mathcal{C}_m^n \text{ such that } v(R) \doteq (\prod_i |S_i|)/m^n > \epsilon, \;\; \Pr(\vec{X} \in R) > 0. \qquad (13)$$

Now suppose that we want a multi-set $T \subseteq \{0, 1, \ldots, m-1\}^n$ such that a random vector $\vec{X}$ sampled from $U(T)$ satisfies (13)—but not necessarily (1). Such a T is called an ϵ-hitting set for $\mathcal{C}_m^n$, since it "hits" (i.e., has a nonempty intersection with) every combinatorial rectangle $R \in \mathcal{C}_m^n$ for which $v(R) > \epsilon$. Every ϵ-approximation is an ϵ-hitting set, but not vice versa. Work of [18] solves the hitting set problem, by presenting an ϵ-hitting set of size $\text{poly}(m + \log n + 1/\epsilon)$ constructible in $\text{poly}(m + n + 1/\epsilon)$ time. This construction has had an interesting application to efficient protocols for leader election in a model of fault-tolerant asynchronous distributed computing [24].

5 Open Questions

One of the main open questions in our context is the construction of ϵ-approximations of cardinality $\text{poly}(m + n + \epsilon^{-1})$ for $\mathcal{C}_m^n$. An even more general question is to construct ϵ-approximations of cardinality $\text{poly}(m + k + \log n + \epsilon^{-1})$ for $\mathcal{C}_{m,k}^n$. A first step may be to approach the corresponding problems for *geometric* rectangles. Can the two reductions mentioned in §4 be of help? As mentioned in the journal version of [13], one can also ask the following easier versions of these issues of efficient construction. Instead of a deterministic construction of suitably small multi-sets S that are ϵ-approximations for the above-mentioned rectangle families, we could first aim for randomized *Las Vegas* constructions of such multi-sets: randomized constructions that are guaranteed to be correct, and whose expected running times are suitably small. Note that the randomized construction sketched, e.g., in §1 is *Monte Carlo.* This "Las Vegas" issue is of complexity that lies between the Monte Carlo and deterministic notions of constructibility.

References

[1] N. Alon, L. Babai, and A. Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of Algorithms*, 7:567–583, 1986.

[2] N. Alon, J. Bruck, J. Naor, M. Naor, and R. Roth. Construction of asymptotically good, low-rate error-correcting codes through pseudo-random graphs. *IEEE Trans. Info. Theory*, 38:509–516, 1992.

[3] N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple constructions of almost k–wise independent random variables. *Random Structures & Algorithms*, 3(3):289–303, 1992.

[4] N. Alon and M. Naor. Derandomization, witnesses for Boolean matrix multiplication and construction of perfect hash functions. *Algorithmica*, 16:434–449, 1996.

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

[6] N. Alon and A. Srinivasan. Improved parallel approximation of a class of integer programming problems. *Algorithmica*, 17:449–462, 1997.

[7] R. Armoni, M. Saks, A. Wigderson, and S. Zhou. Discrepancy sets and pseudorandom generators for combinatorial rectangles. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 412–421, 1996.

[8] P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: learning and pseudorandom sets. *Journal of Computer and System Sciences*, 57:376–388, 1998.

[9] J. Beck and V. T. Sós. Discrepancy theory. In *Handbook of combinatorics, Volume II*, chapter 26, pages 1405–1446. Elsevier Science B.V. and the MIT Press, 1995.

[10] A. Z. Broder, M. Charikar, A. Frieze and M. Mitzenmacher. Min-wise independent permutations. In *Proc. ACM Symposium on Theory of Computing*, pages 327–336, 1998.

[11] S. Chari, P. Rohatgi, and A. Srinivasan. Improved algorithms via approximations of probability distributions. To appear in *Journal of Computer and System Sciences.*

[12] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[13] G. Even, O. Goldreich, M. Luby, N. Nisan, and B. Veličković. Approximations of general independent distributions. In *Proc. ACM Symposium on the Theory of Computing*, pages 10–16, 1992. Also: Efficient approximations for product distributions, *Random Structures & Algorithms*, 13: 1–16, 1998.

[14] W. Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

[15] R. Impagliazzo, N. Nisan, and A. Wigderson. Pseudorandomness for network algorithms. In *Proc. ACM Symposium on Theory of Computing*, pages 356–364, 1994.

[16] P. Indyk. A small approximately min-wise independent family of hash functions. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 454–456, 1999.

[17] J. Justesen. A class of asymptotically good algebraic codes. *IEEE Trans. Info. Theory*, 18:652–656, 1972.

[18] N. Linial, M. Luby, M. Saks, and D. Zuckerman. Efficient construction of a small hitting set for combinatorial rectangles in high dimension. *Combinatorica*, 17:215–234, 1997.

[19] C.-J. Lu. Improved pseudorandom generators for combinatorial rectangles. In *Proc. International Conference on Automata, Languages and Programming*, pages 223–234, 1998.

[20] J. Matoušek. **Geometric Discrepancy: An Illustrated Guide**. *Algorithms and Combinatorics*, Eds.: R.L. Graham, B. Korte, L. Lovász, A. Wigderson and G.M. Ziegler, Vol. 18, Springer-Verlag, 1999.

[21] J. Naor and M. Naor. Small–bias probability spaces: efficient constructions and applications. *SIAM J. Comput.*, 22(4):838–856, 1993.

[22] M. Naor, L. J. Schulman, and A. Srinivasan. Splitters and near-optimal derandomization. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 182–191, 1995.

[23] J. Radhakrishnan and A. Srinivasan. Improved bounds and algorithms for hypergraph two-coloring. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 684–693, 1998.

[24] A. Russell and D. Zuckerman. Perfect information leader election in $\log^* n + O(1)$ rounds. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 576–583, 1998.

[25] M. Saks, A. Srinivasan, S. Zhou and D. Zuckerman. Low discrepancy sets yield approximate min-wise independent permutation families. In *Proc. International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 11–15, 1999.