COMBINATORICA
Bolyai Society – Springer-Verlag

COMBINATORICA 17 (2) (1997) 215–234

EFFICIENT CONSTRUCTION OF A SMALL HITTING SET FOR COMBINATORIAL RECTANGLES IN HIGH DIMENSION*

NATHAN LINIAL[†], MICHAEL LUBY[‡], MICHAEL SAKS[§] and DAVID ZUCKERMAN[¶]

Received March 22, 1995

We describe a deterministic algorithm which, on input integers d, m and real number $\epsilon \in (0,1)$, produces a subset S of $[m]^d = \{1,2,3,\ldots,m\}^d$ that hits every combinatorial rectangle in $[m]^d$ of volume at least ϵ, i.e., every subset of $[m]^d$ the form $R_1 \times R_2 \times \ldots \times R_d$ of size at least ϵm^d. The cardinality of S is polynomial in $m(\log d)/\epsilon$, and the time to construct it is polynomial in md/ϵ. The construction of such sets has applications in derandomization methods based on small sample spaces for general multivalued random variables.

1. Introduction

This paper is motivated by the *witness finding* problem: design an efficient algorithm that on input a positive integer n and a real $\epsilon > 0$ produces a list S of elements in $\{0,1\}^n$ such that, for any *witness set* $\mathrm{R} \subseteq \{0,1\}^n$ where $|\mathrm{R}|/2^n \geq \epsilon$, $\mathrm{S} \cap \mathrm{R} \neq \emptyset$. The running time of the algorithm should be polynomial in n and $1/\epsilon$.

* A preliminary version of this paper appeared in Proceedings of the 25th Annual ACM Symposium on Theory of Computing, 1993.

[†] Research partially done while visiting the International Computer Science Institute. Research supported in part by a grant from the Israel-USA Binational Science Foundation.

[‡] A large portion of this research was done while still at the International Computer Science Institute in Berkeley, California. Research supported in part by National Science Foundation operating grants CCR-9304722 and NCR-9416101, and United States-Israel Binational Science Foundation grant No. 92-00226.

[§] Supported in part by NSF under grants CCR-8911388 and CCR-9215293 and by AFOSR grants AFOSR-89-0512 AFOSR-90-0008, and by DIMACS, which is supported by NSF grant STC-91-19999 and by the New Jersey Commission on Science and Technology. Research partially done while visiting the International Computer Science Institute.

[¶] Partially supported by NSF NYI Grant No. CCR-9457799. Most of this research was done while the author was at MIT, partially supported by an NSF Postdoctoral Fellowship. Research partially done while visiting the International Computer Science Institute.

Mathematics Subject Classification (1991): 068Q25, 05B40, 68R05

This is a fundamental problem in complexity theory where R is usually the witness set for some language under consideration.

The witness finding problem is easy to solve using a randomized algorithm: Just sample independently at random $O(1/\epsilon)$ strings from the uniform distribution on $\{0,1\}^n$. For any fixed R with $|\mathrm{R}|/2^n \geq \epsilon$, it is easy to see that with probability $\geq 1/2$ at least one of the sampled strings is in R. (Note the order of quantification: it is clearly not true that with probability $\geq 1/2$ the set of sampled strings contains an element from all witness sets.) Overall, this algorithm uses $O(n/\epsilon)$ random bits.

A solution to the witness finding problem is a key component in many known efficient randomized algorithms. In these applications, it is typical for $1/\epsilon$ to be polynomial in n. Over the past few years, research has centered on designing efficient algorithms for the witness finding problem that use fewer random bits. A randomized algorithm for witness finding that uses $O(n)$ random bits and solves the problem for $\epsilon = 1/poly(n)$ was introduced in [9] and [16], and subsequently [4] found a substantially simpler randomized algorithm using only $2n$ random bits (in fact, a simple modification of their procedure reduces this to n bits).

Unfortunately, there is no deterministic algorithm for the general witness finding problem. If the algorithm deterministically produces a list of length k, it completely misses the complementary set of size $2^n - k$. This impossibility results, of course, from the fact that no restriction is imposed on the witness sets, i.e., R is allowed to be an arbitrary subset of $\{0,1\}^n$. In many applications it is possible to derive some structural properties of the witness set, even though the set itself remains unknown. Can we find, then, natural and interesting classes of exponential size witness sets for which the witness finding problem is solvable deterministically in polynomial time? This is exactly what the present paper is about.

The class of witness sets considered here is this: Let d and m be positive integers, and let $U = [m]^d$, i.e., the universe U consists of all d-dimensional lattice points with all coordinates in $[m]$. Of course, $|U| = m^d$. Witness sets are all *combinatorial rectangles* within U, i.e., sets of the form $\mathrm{R} = \mathrm{R}_1 \times \cdots \times \mathrm{R}_d$, with $\mathrm{R}_i \subseteq [m]$ for all $i \in [d]$. Note that there are 2^{md} combinatorial rectangles, while there are 2^{m^d} subsets of $\{0,1\}^U$, so the restriction to rectangles should be helpful. The *volume of* R is defined as

$$\mathrm{vol}(\mathrm{R}) = |\mathrm{R}|/|U| = \left(\prod_{i \in [d]} |\mathrm{R}_i| \right) / m^d,$$

i.e., the volume of R is the fraction of points in U that lie in R. Our algorithm produces an (m,d,ϵ)-hitting set $\mathrm{S} \subseteq U$ i.e., for any combinatorial rectangle R, if $\mathrm{vol}(\mathrm{R}) \geq \epsilon$ then $\mathrm{S} \cap \mathrm{R} \neq \emptyset$. The cardinality $|\mathrm{S}|$ is polynomial in $m \log(d)/\epsilon$. It takes time polynomial in $m \cdot d/\epsilon$ to construct S. In Section 3 we show that $|\mathrm{S}|$ is optimal to within a polynomial factor by giving an $\Omega(m/\varepsilon + \log d)$ lower bound.

As note above, there are only 2^{md} rectangles. It follows that a set S of $O(md/\epsilon)$ points drawn uniformly at random from U almost surely hits all rectangles of volume at least ϵ. With a little more effort, the same can be shown for a random set S of size polynomial in $m\log(d)/\epsilon$. (In Section 3 we point out that any (m,d,ϵ)-hitting set must have size at least $\Omega(m+\log(d)+1/\epsilon)$ (assuming $1/\epsilon \leq m^d$)). However, this does *not* provide a solution to the problem we consider: this is only an existence proof, while we are looking for efficient constructions.

This work was motivated by the problem of finding efficient constructions of small sample spaces that approximate the independent uniform distribution on many multivalued random variables (these can easily be used to simulate non-uniformly distributed random variables, see e.g., [5]). Let $X = \langle X_1,\ldots,X_d\rangle$ be a sequence of d random variables taking on values in $[m]$. The set S described above can be viewed as a sample space for the sequence X with the following properties. For each $i \in [d]$, we can view $\mathrm{R}_i \subseteq [m]$ as a set of possible values for X_i. Given a combinatorial rectangle $\mathrm{R} = \mathrm{R}_1 \times \cdots \times \mathrm{R}_d$, we can view $X \in \mathrm{R}$ as the global event that simultaneously, for all $i \in [d]$, $X_i \in \mathrm{R}_i$. Thus, $\mathrm{vol}(\mathrm{R})$ is the probability that $X \in \mathrm{R}$ if the random variables $X_1, X_2, \ldots, X_d$ are uniformly distributed on $[m]$ and independent. The set S is spread uniformly in the following sense: for any rectangle R for which $Pr(X \in \mathrm{R}) \geq \epsilon$ where X is the vector random variable $(X_1, X_2, \ldots, X_d)$ and the sampling is done uniformly from $[m]^d$, then under uniform sampling from S, $Pr(X \in \mathrm{R}) > 0$.

Our construction can also be viewed as an efficient deterministic solution for the d-dimensional version of the battleship game. A "battleship" corresponds to a combinatorial rectangle R, and S defines a deterministic, efficiently constructible, short probe sequence that hits all battleships of size at least ϵ.

Another aspect of our work is that we provide an easily constructible and small ϵ-net for combinatorial rectangles in d dimensions. Recall that an ϵ-net for a class of objects is a set of points S such that any object of size at least ϵ is hit by at least one point in S.

A *geometric* rectangle is $\mathrm{R} = [a_1,b_1) \times \cdots \times [a_d,b_d) \subseteq [0,1)^d$, and $\mathrm{vol}(\mathrm{R}) = \prod_{i\in[d]}(b_i - a_i)$. It is a natural problem to find a set S of points in $[0,1)^d$ which meets every geometric rectangle R with $\mathrm{vol}(\mathrm{R}) \geq \epsilon$. As noted in [5], this geometric question easily reduces to the combinatorial version we consider here, where $m = O(d/\epsilon)$. Solving the combinatorial problem turned out to require many more ideas than a solution to the geometric version.

Two constructions given in [5] are comparable to the work described here. The first is a set S of size $(md)^{O(\log(1/\epsilon))}$ and in the second one S has size $(md/\epsilon)^{O(\log(d))}$. There is one aspect in which the constructions of [5] is stronger than those described here: for each combinatorial rectangle R, the fraction of points in S that belong to R is within an additive factor ϵ of $\mathrm{vol}(\mathrm{R})$ (so consequently S hits each R with $\mathrm{vol}(\mathrm{R}) \geq \epsilon$ at least once). On the other hand, the construction here improves over the constructions of [5] in terms of $|\mathrm{S}|$, and this improvement is more substantial than might first appear. For interesting cases of d, m and $1/\epsilon$, i.e., when all parameter are polynomial in n, we give the first explicit constructions of size polynomial in n. In contrast, the constructions in [5] are of size $n^{O(\log(n))}$. One of the constructions in [5] is based on Nisan's pseudorandom generator that maps $O(\log^2(n))$ bits to $n^{O(1)}(n)$ bits and fools any logspace machine [14]. Combinatorial rectangles may be viewed as a special case of nonuniform logspace tests; hence trying all seeds of Nisan's generator gives the construction in [5]. A similar idea gives the $n^{O(\log(n))}$ size universal traversal sequences found in [14], which is the best explicit construction to date. It is important to identify interesting special cases when this $n^{O(\log(n))}$ barrier can be broken, and the size brought down to polynomial. This paper provides such a case.

2. Some preliminaries

We use log to denote the logarithm to base 2.

2.1. Rectangles and hitting sets

For integers $m,d \geq 1$, a *rectangle* R in $[m]^d$ is a subset of the form $\mathrm{R}_1 \times \mathrm{R}_2 \times \ldots \times \mathrm{R}_d$. The *volume* of the rectangle R, denoted $\mathrm{vol}(\mathrm{R})$, is defined to be $\Pi_{i=1}^{d}(|R_i|/m)$. Observe that if R is a rectangle in $[m]^d$ then we can also view it as a rectangle of $[m']^d$ for $m' > m$; however the volume of R changes by a factor $(m/m')^d$ if we do this.

For rectangle R and $J \subseteq [d]$, we define $\mathrm{R}_J = \cap_{j \in J} \mathrm{R}_j$, abbreviating $R_{\{i,j\}}$ as $R_{i,j}$. A rectangle R is said to have *pairwise independent projections* if, for all $i \neq j$, $|\mathrm{R}_{i,j}|/m = |\mathrm{R}_i|/m \times |\mathrm{R}_j|/m$. Such a rectangle is called a *PIP-rectangle.*

Throughout the paper, ϵ denotes a parameter in the range $(0,1)$ and $k = k(\epsilon)$ denotes $\ln(1/\epsilon)$. A rectangle R in $[m]^d$ of volume at least ϵ is said to be an (m,d,ϵ)-rectangle. A subset S of $[m]^d$ that has a non-empty intersection with all (m,d,ϵ) rectangles is an (m,d,ϵ)*-hitting set.* A subset S of $[m]^d$ that has a non-empty intersection with all (m,d,ϵ) PIP-rectangles is an (m,d,ϵ) PIP-hitting set. Trivially an (m,d,ϵ)-hitting set is an (m,d,ϵ) PIP-hitting set, but the reverse is not typically true.

2.2. Directed bipartite graphs

We denote a directed bipartite graph with parts X,Y and all edges directed from X to Y by $G = (X,Y,E)$. For $(x,y) \in E$, we say y is an *out-neighbor* of x and x is an *in-neighbor* of y. For $x \in X$, the set of out-neighbors of x is denoted $G^+(x)$ and the size of $G^+(x)$, called the out-degree of x, is denoted $deg^+(x) = deg_G^+(x)$.

Similarly, For $y \in Y$, the set of in-neighbors of y is denoted $G^{-}(y)$ and the size of $G^{-}(y)$, called the in-degree of y, is denoted $deg^{-}(y) = deg_G^{-}(y)$. $\Delta^{+}(G)$ denotes the maximum out-degree of any $x \in X$ and $\Delta^{-}(G)$ is the maximum in-degree of any $y \in Y$. For $W \subseteq X$, $G^{+}(W)$ is the union of $G^{+}(x)$ for $x \in W$ and and for $Z \subseteq Y$, $G^{-}(Z)$ is the union of $G^{-}(y)$ for $y \in Z$.

2.3. Universal families of hash functions

Our construction makes use of two standard tools of derandomization: universal families of hash functions and expanders. In the next two subsections, we review the definitions and relevant properties.

Definition. A family of functions H mapping $[r]$ to $[s]$ is a *universal hash function family*, if for all $i \neq i' \in [r]$, and for all $j, j' \in [s]$, the fraction of functions $h \in H$ that map i to j and i' to j' is exactly $1/s^2$. In other words, if we consider H as a probability space with uniform probability function, then the random variables $h(1), h(2), \ldots, h(r)$ are each uniformly distributed over $[s]$ and are pairwise independent.

There are various explicit constructions known for universal families of hash functions. For our purposes we will need the following well-known fact ([3]):

Lemma 1. *Let r, s be integers with s a power of 2. Then there is an explicitly constructible family $H_{r,s}$ of universal hash functions of size at most $s \cdot \max(2r, s)$. The time to construct the family is polynomial in its size.*

The upper bound on size is actually a little better than this, but this form of the bound is convenient for our purposes.

2.4. Expanders

For $\alpha \in (0,1)$ and positive integers $n, \Delta > 0$, an undirected graph G is an (n, Δ, α)-expander if G has n vertices, maximum degree Δ, and for any subset A of vertices, the fraction of vertices in $V(G) - A$ that have a neighbor in A is at least $\alpha|A|/n$. We will need:

Lemma 2. [13], [7] *For each integer n that is a perfect square, there is an explicitly constructible $(n, 8, \alpha)$ expander for $\alpha = (2 - \sqrt{3})/4$.*

If a, b are vertices, $\mathrm{dist}_G(a,b)$ is, as usual, the length (number of edges) of the shortest path from a to b. For vertex subsets A and B, $\mathrm{dist}_G(A,B)$ is the minimum over $a \in A$, $b \in B$ of $\mathrm{dist}_G(a,b)$. The following well-known property of expanders is the key property we need. For completeness we give a proof.

Lemma 3. *Let G be an (n,Δ,α) expander, with $\alpha\in(0,1)$. If A,B are subsets of $V(G)$ then:*

$$\mathrm{dist}_G(A,B) \le \frac{2}{\alpha}(\log\frac{n}{|A|} + \log\frac{n}{|B|}).$$

Proof. For $A\subseteq V(G)$ and $i\ge 0$, let $N_i(A)$ be the set of vertices of distance at most i from A. The expansion property implies that for any A, $|N_1(A)|\ge|A|(1+\alpha(n-|A|)/n)$. Noting that $N_i(A)=N_1(N_{i-1}(A))$, we have that for positive integers s:

$$|N_s(A)| \ge |A|\prod_{t=0}^{s-1}(1+\alpha(n-|N_t(A)|/n)) \ge |A|(1+\alpha(n-|N_{s-1}(A)|)/n)^s.$$

Similarly,

$$|N_s(B)| \ge |B|(1+\alpha(n-|N_{s-1}(B)|)/n)^s.$$

Let i be the largest index such that $|N_i(A)| \le n/2$ and j be the largest index such that $|N_j(B)|\le n/2$. Then $N_{i+1}(A)\cap N_{j+1}(B)$ is nonempty, and so $\mathrm{dist}_G(A,B)\le i+1+j+1$. Now $n/2\ge|N_i(A)|\ge|A|(1+\alpha/2)^i\ge|A|2^{(\alpha/2)i}$, since $1+x\ge 2^x$ for $x\in[0,1]$. Thus $i\le\frac{2}{\alpha}\log\frac{n}{2|A|}\le\frac{2}{\alpha}(\log\frac{n}{|A|}-1)\le\frac{2}{\alpha}\log\frac{n}{|A|}-1$. Similarly $j\le\frac{2}{\alpha}\log\frac{n}{|B|}-1$. ∎

3. A lower bound

Before giving our construction, we give a lower bound.

Proposition 4. *For $m^{-d}\le\varepsilon\le 2/9$, any (m,d,ε)-hitting set has size $\Omega(m+1/\varepsilon+\log d)$.*

We remark that some upper bound on ε is necessary, since for $\varepsilon > 1/4$ and any d, $\{0^d,1^d\}$ is a $(2,d,\varepsilon)$-hitting set. The lower bound on ϵ is also necessary since the set $[m]^d$ is trivially an (m,d,ϵ)-hitting set for all ϵ.

Proof. A lower bound of $m(1-\epsilon)$ follows by noting that if S is such a hitting set and $R_1\subseteq[m]$ is the set of values that don't appear as the first coordinate of a point in S, then $R_1\times[m]^{d-1}$ is a rectangle of volume at least $1-|S|/m$ that is not hit by S.

A lower bound of $1/2\epsilon$ follows from a simple probabilistic argument: choose positive integers $T_1,T_2,\dots,T_d$ all at most m such that $2\epsilon m^d\ge T_1T_2\dots T_m\ge\epsilon m^d$. Select subsets $R_1,R_2,\dots,R_d$ of $[m]$ where R_i is chosen uniformly at random from among all subsets of size T_i. Then for any fixed point in $[m]^d$ the probability that it is in $R=R_1\times\dots\times R_d$ is $T_1T_2\dots T_d/m^d\le 2\epsilon$ and so the expected size of $R\cap S$ is at most $2|S|\epsilon$. If S is a hitting set this expectation is at least 1 and so $|S|>1/2\epsilon$.

Next, we prove a $\log_2 d$ lower bound, and we start with the case $m=2$. The proof reduces to the easy and well-known fact that the edge-set of K_d, the complete graph on d vertices cannot be covered with fewer than $\log_2 d$ complete bipartite subgraphs. Let S be a $(2,d,\varepsilon)$-hitting set. With every point $(x_1,\dots,x_d) \in S$, we associate the subgraph of K_d consisting of the edges $[p,q] \in E(K_d)$ where $x_p \neq x_q$. If this collection of complete bipartite subgraphs fails to cover the edge $[i,j] \in E(K_d)$, then there is no point $(x_1,\dots,x_d) \in S$ with $x_i = 1$ and $x_j = 2$. But then S misses the rectangle with $\{1\}$ in the ith coordinate, $\{2\}$ in the jth coordinate, and $\{1,2\}$ everywhere else, although this rectangle has volume $1/4 > \varepsilon$.

To deal with $m \geq 3$, let $\phi : [m]^d \to [2]^d$ be defined via $\phi(x_1,\dots,x_d) = (\lceil 2x_1/m \rceil, \dots, \lceil 2x_d/m \rceil)$. Let S be an (m,d,ε)-hitting set, and consider its image $\phi(S) \subset [2]^d$. As we observed, if $|\phi(S)| < \log_2 d$, then it misses a rectangle R of the form $\{1\} \times \{2\} \times \{1,2\}$ (in some order of coordinates). But then S misses $\phi^{-1}(\mathrm{R})$ whose volume is $\lfloor m/2 \rfloor \cdot \lceil m/2 \rceil / m^2 \geq 2/9$, a contradiction. Consequently, $|S| \geq |\phi(S)| \geq \log_2 d$, as needed. ∎

4. Overview of the hitting set construction

Our goal is to give an explicit construction of a "small" set that meets all combinatorial rectangles of volume ϵ from $[m]^d$. We want the size of the set to be polynomial in $m, 1/\epsilon$ and $\log d$.

Our construction has two main parts. The first part is a construction, based on expander graphs, of an (m,d,ϵ)-hitting set whose size is polynomial in m, $1/\epsilon$, and 2^d. This construction generalizes one of [17], which is closely related to a previous construction of [1]. This construction is described in Section 5.

The inadequacy of this construction for our problem is that the dependence on d is exponential and we want the dependence on d to be logarithmic. Note that in the case that d is small, on the order of $\log(1/\epsilon)$, the size of the hitting set size is polynomial in m and $1/\epsilon$. The second part of our construction is a "reduction" of the general problem of building an (m,d,ϵ)-hitting set to that of building an (m^*,d^*,ϵ^*)-hitting set where m^* is bounded above by a polynomial in m, $1/\varepsilon$, and $\log d$, ϵ^* is bounded below by a polynomial in ϵ, and d^* is $O(k^*) = O(\ln(1/\epsilon^*)) = O(\ln(1/\epsilon))$. We then use the expander-based construction to get an (m^*,d^*,ϵ^*)-hitting set. The reduction specifies how to transform this hitting set into an (m,d,ϵ)-hitting set whose size is polynomial in $m, 1/\epsilon$ and $\log d$.

This second part of the construction, the reduction, is the composition of a sequence of reductions. Each reduction has the same general form which, for clarity, we first describe very generally. Suppose that $\mathcal{A}$ is a family of subsets of a set X for which we wish to construct a small hitting set H. Suppose that $\mathcal{B}$ is a family of subsets of some set Y. Then the problem of finding a hitting set for $\mathcal{A}$ can be reduced to the problem of finding a hitting set for $\mathcal{B}$ as follows. Suppose that $G = (X,Y,E)$ is any directed bipartite graph that satisfies: if $A \in \mathcal{A}$, then $G^+(A)$

contains some $B \in \mathscr{B}$. Then it is easy to see that if $H' \subset Y$ is a hitting set for $\mathscr{B}$ then $G^-(H')$ is a hitting set for $\mathscr{A}$. Note that the size of the hitting set for $\mathscr{A}$ is at most $|H'|\Delta^-(G)$. We call such a bipartite graph a *reduction* from the hitting set problem for $\mathscr{A}$ to the problem for $\mathscr{B}$, and $\Delta^-(G)$ is called the *cost* of the reduction.

Our sequence of reductions will take us through a sequence of "simpler" hitting set problems, ending with the (m^*,d^*,ϵ^*)-hitting set problem. The cost of each reduction will always be bounded by a polynomial in m, $\log d$ and $1/\epsilon$. Using the hitting set construction based on expanders to build an (m^*,d^*,ϵ^*)-hitting set, and applying the reductions we obtain an (m,d,ϵ)-hitting set of the desired size.

We will need a sequence of reductions to accomplish our aim. We employ two types of reductions: *dimension reductions* and *PIP-reductions.* As its name suggests, a dimension reduction reduces a hitting set problem for rectangles for dimension d to one for rectangles of some lower dimension. A PIP-reduction reduces the (m,d,ϵ)-hitting set problem to an (m',d,ϵ) PIP-hitting set problem, for some m' that is bounded by a polynomial in m and d. The cost of dimension reductions depends on the details of the reduction, while a PIP-reduction always has cost 1.

Our reduction sequence is divided into three main reductions. The second and third reductions each consist of two subreductions—a PIP-reduction followed by a dimension reduction. We will denote by m_0,d_0,ϵ_0 and k_0, the parameters corresponding to the hitting set problem we wish to solve and for $1 \le i \le 3$, we use m_i,d_i,ϵ_i and k_i to denote the parameters of the hitting set problem after the i^{th} main reduction, so that (m_3,d_3,ϵ_3) corresponds to (m^*,d^*,ϵ^*) above. In what follows, it is helpful to keep in mind that the values ϵ_1,ϵ_2 and ϵ_3 are each polynomial in ϵ_0, and thus $k_i = \Theta(k_j)$ for any $i,j \in \{0,1,2,3\}$. We now summarize the sequence of reductions.

Reduction 1. Reduce the (m_0,d_0,ϵ_0)-hitting set problem to the (m_1,d_1,ϵ_1)-hitting set problem, where $m_1 = m_0$, $d_1 = O(k_0^3(\log d_0)^2/\epsilon_0)$, $\epsilon_1 = \epsilon_0/2$. This is accomplished by a dimension reduction of cost $\lceil 2k_0^2(\log d)/\epsilon_0 \rceil$.

Reduction 2. We reduce the (m_1,d_1,ϵ_1)-hitting set problem to the (m_2,d_2,ϵ_2)-hitting set problem, where $m_2 = O(m_1^2 d_1^2)$, $d_2 = O(k_1^2)$ and $\epsilon_2 = (\epsilon_1/4)^2$. This is accomplished by a sequence of two sub-reductions.

Reduction 2a. This is a PIP-reduction that reduces the (m_1,d_1,ϵ_1)-hitting set problem to the $(m_2,d_1,\epsilon_1/4)$ PIP-hitting set problem. This reduction has cost 1.

Reduction 2b. This is a dimension reduction that reduces the $(m_2,d_1,\epsilon_1/4)$ PIP-hitting set problem to the (m_2,d_2,ϵ_2)-hitting set problem. The cost of this reduction is $O(d_1^2)$.

Reduction 3. We reduce the (m_2,d_2,ϵ_2)-hitting set problem to the (m_3,d_3,ϵ_3)-hitting set problem, where $m_3 = O(m_2^2 d_2^2)$, $d_3 = O(k_2) = O(k_0)$ and $\epsilon_3 = (\epsilon_2/4)^2$. Like reduction 2, this is accomplished by a sequence of two sub-reductions.

Reduction 3a. This is a PIP-reduction that reduces the (m_2,d_2,ϵ_2)-hitting set problem to the $(m_3,d_2,\epsilon_2/4)$ PIP-hitting set problem. This reduction has cost 1.

Reduction 3b. This is a dimension reduction that reduces the $(m_3,d_2,\epsilon_2/4)$ PIP-hitting set problem to the (m_3,d_3,ϵ_3)-hitting set problem. The cost of this reduction is $d_2^{O(\log k_2)} \cdot 2^{O(k_2)}$, which is bounded by a polynomial in $1/\epsilon_0$.

Before proceeding to the details of the constructions, let us clarify what each of these reductions accomplishes, and why we need them all. Recall that we need to reduce the dimension of the problem to $O(k_0)$, at a cost at most polynomial in $m, \log d$ and $1/\epsilon$. Since the cost of each reduction has this bound, applying them in succession accomplishes our goal.

Why do we need all three? Note that reduction 3 could be applied directly to any (m,d,ϵ)-hitting set to reduce the dimension to $O(\log(1/\epsilon))$; the problem is that its cost is polynomial in $1/\varepsilon$ times $d^{O(\log\log 1/\varepsilon)}$. This cost would be acceptable if d is bounded, say, by a polynomial in $\log(1/\epsilon)$. So, given reduction 3, it suffices to reduce the general (m_0,d_0,ϵ_0)-hitting set problem to the case that the dimension is polynomial in k_0. Now, reduction 2 accomplishes this, getting the dimension down to $O(k_0^2)$, at a cost that is polynomial in the dimension. This cost is still too high for us to apply reduction 2 to the initial problem, since we want the dependence on d_0 to be only polylogarithmic. Thus reduction 1 is applied first; this gets the dimension down to polynomial in $1/\epsilon_0$ and $\log d_0$, at an acceptable cost.

In the rest of this paper, we describe the construction. First we give the expander-based construction of an (m,d,ϵ)-hitting set whose size is polynomial in $m, 1/\epsilon$ and 2^d. Then, in preparation for describing the sequence of three reductions, we describe the two types of reductions, dimension reductions and PIP-reductions, in more detail. Finally we describe each of the three reductions.

5. A hitting set construction for low dimension

Here we present a construction which for any positive integers m and d and $\epsilon \in (0,1)$, produces an (m,d,ϵ)-hitting set of size polynomial in $m, 1/\epsilon$ and 2^d. This generalizes a construction in [17]. To describe the construction we need a few definitions. If G is a graph, a *walk* of length s in G is a sequence $v_0, v_1, v_2, \ldots, v_s$ where for each $i \geq 1$, either v_i is equal to or adjacent to v_{i-1}. Let $W_s(G)$ denote the set of all walks of G of length s. For $1 \leq t \leq s$, let $W_{s,t}$ be the set of all sequences $v_1, v_2, \ldots, v_t$ which are (not necessarily consecutive) subsequences of some walk of length s.

Lemma 5. *Let m,d be positive integers and* R *be a rectangle in $[m]^d$. Suppose G is an (m,Δ,α)-expander with $1/2 > \alpha > 0$. If $s = 1 + \frac{4}{\alpha}(d + \log(1/\mathrm{vol}(\mathrm{R})))$, then $W_{s,d}$ contains a point from* R.

Before proving the lemma, note that it implies that if $s = 1 + \frac{4}{\alpha}(d + \log(1/\epsilon))$, then $W_{s,d}$ hits all (m,d,ϵ) rectangles. Also, the size of $W_{s,d}$ is trivially bounded above by $2^s|W_s|$ and $|W_s|$ is bounded above by $m(\Delta(G)+1)^s$ where $\Delta(G)$ is the maximum degree of G. Using the explicit bounded degree expanders mentioned in Section 2.4, the size of $W_{s,d}$ in this case is polynomial in $m, 1/\epsilon$, and 2^d as required.

One technical point needs to be made. The construction of expanders we use requires that m be a perfect square. If m is not a perfect square, we want to round m up to the next perfect square $\hat{m}$. Every rectangle R in $[m]^d$ is also a rectangle in $[\hat{m}]^d$, but the volume of R relative to the larger space is reduced by a factor $(m/\hat{m})^d$. Since $m/\hat{m} \geq 1/2$, it would suffice to find a $(\hat{m}, d, \hat{\epsilon})$-hitting set, where $\hat{\epsilon} = \epsilon/2^d$, and thus since our construction is polynomial in $1/\hat{\epsilon}$, $\hat{m}$ and 2^d, it is also polynomial in $1/\epsilon, m$ and 2^d.

Proof of Lemma 5. We say that the walk $v_0, v_1, v_2, \dots, v_s$ in W_s *traverses* the sequence of sets $\mathrm{R}_1, \mathrm{R}_2, \dots, \mathrm{R}_d$ if there are indices $1 \leq i_1 < i_2 < \dots < i_d \leq s$ such that $v_{i_j} \in \mathrm{R}_j$ for each j between 1 and d. We want to show that for any (m,d,ϵ)-rectangle there is a walk in W_s that traverses it. For $j \geq 2$, define $t_j = \frac{2}{\alpha}(2+\log\frac{m}{|R_{j-1}|}+\log\frac{m}{|R_j|})$. Observe that $1 + \sum_{j=2}^{d} t_j \leq s$ and thus Lemma 5 follows immediately from:

Claim. *Let* $\mathrm{R}_1, \mathrm{R}_2, \dots, \mathrm{R}_d$ *be subsets of* $[m]$. *Then there is a subset* Q_d *of* R_d *of size at least* $|\mathrm{R}_d|/2$ *such that for each* $w \in Q_d$ *there is a walk of length at most* $1 + \sum_{j=2}^{d} t_j$ *that traverses* $\mathrm{R}_1, \mathrm{R}_2, \dots, \mathrm{R}_d$ *and ends at* w.

We prove the claim by induction on d. For $d = 1$, the claim is trivial since we can take $Q_1 = \mathrm{R}_1$, and for each $w \in Q_1$, the trivial walk w satisfies the claim.

Now suppose $d > 1$ and that the claim holds for $d-1$. Then there is a subset Q_{d-1} of size at least $|R_{d-1}|/2$ satisfying the conclusion of the claim. Let Y be the set of vertices whose distance from Q_{d-1} exceeds t_d. Then $|Y| \leq |R_d|/2$ since by the definition of t_d and by Lemma 3 with $A = Q_{d-1}$ and $B = Y$ we have:

$$\frac{2}{\alpha}(2 + \log\frac{m}{|R_{d-1}|} + \log\frac{m}{|R_d|}) < \mathrm{dist}\,(Q_{d-1}, Y)$$

Now, by lemma 3,

$$\begin{aligned}\mathrm{dist}\,(Q_{d-1}, Y) &\leq \frac{2}{\alpha}\left(\log\frac{m}{|Q_{d-1}|} + \log\frac{m}{|Y|}\right)\\ &\leq \frac{2}{\alpha}\left(2 + \log\frac{m}{|R_{d-1}|} + \log\frac{m}{2|Y|}\right).\end{aligned}$$

This implies that $|Y| \leq |R_d|/2$. So the set $Q_d = R_d - Y$ has size at least $|R_d|/2$ and for every $i \in Q_d$, $\mathrm{dist}\,(Q_{d-1}, i) \leq t_d$. With the induction hypothesis, this implies that Q_d satisfies the Claim for d. ▮

6. Reductions

We now give a detailed description of the two types of reductions, PIP-reductions and dimension reductions.

6.1. PIP-reductions

The purpose of this reduction is to reduce the (m,d,ϵ)-hitting set problem to the (m',d,ϵ') PIP-hitting set problem with $m'=O(m^2d^2)$ and $\epsilon'\geq\epsilon/4$.

First consider the case that m is a power of 2. Let T be a universal family of hashing functions that map $[d]$ to $[m]$ of size $m'=m\times\max\{m,2d\}$ as in Lemma 1. Identify the set T with the set $[m']$ and identify T^d with $[m']^d$. Now define a map from T^d to $[m]^d$ as follows: A point $f=(f_1,\dots,f_d)\in T^d$ (i.e, each f_i is a function from the family T) is mapped to $x_f=(f_1(1),\dots,f_d(d))$. Define the bipartite graph G on vertex sets $[m]^d$ and T^d with edge set $\{(x_f,f):f\in T^d\}$.

Now if $\mathrm{R}=\mathrm{R}_1\times\mathrm{R}_2\times\dots\times\mathrm{R}_d$ is any rectangle in $[m]^d$, then $G^+(\mathrm{R})$ is a rectangle $F_{\mathrm{R}}=F_1\times\dots\times F_d\subseteq[m']^d$ in T^d, with $F_i=\{f\in T:f(i)\in\mathrm{R}_i\}$. The definition of a universal family of hash functions easily implies that F_{R} is a PIP-rectangle of the same volume as R. Thus if H' is an (m',d,ϵ) PIP-hitting set then $G^-(H)$ is a (m,d,ϵ)-hitting set. Clearly, every $f\in T$ has in-degree 1 in G, so the cost of the reduction is 1.

Now, suppose that m is not a power of 2. We'd like just to replace m by $\hat{m}$, the least power of 2 greater than m, and view rectangles in $[m]^d$ as rectangles in $[\hat{m}]^d$; which is similar to what we did in the previous section. However, this is not adequate here, because, when we increase m to $\hat{m}$, the volume of each rectangle is reduced by a factor $(m/\hat{m})^d$, which can be close to $(1/2)^d$. We can't afford such a drastic reduction in volume, since, unlike in the previous section, we are not willing to have an exponential dependence of the cost on d. Instead, we perform two simple reductions to reduce to the power of 2 case. For the first reduction, we reduce from the (m,d,ϵ)-hitting set to the (cm,d,ϵ)-hitting set where c is the least integer greater than or equal to d such that the interval $[cm,(c+1)m]$ contains a power of 2. (Note c is between d and $2d$). This reduction is obtained by defining the bipartite graph G from $[m]^d$ to $[cm]^d$ which connects point $(p_1,p_2,\dots,p_d)$ in $[m]^d$ to $(q_1,q_2,\dots,q_d)$ in $[cm]^d$ if $q_i\equiv p_i \bmod m$ for i between 1 and d. The cost of this reduction is clearly 1. For the second preliminary reduction, let $\hat{m}$ be the power of 2 in the interval $[cm,(c+1)m]$. Note that $(cm/\hat{m})^d\geq(c/(c+1))^d\geq(1-1/d)^d\geq1/4$ and thus every (cm,d,ϵ)-rectangle contains a $(\hat{m},d,\epsilon/4)$-rectangle. Thus we may reduce the (cm,d,ϵ)-hitting set problem to the $(\hat{m},d,\epsilon/4)$-hitting set problem, where $\hat{m}=\Theta(md)$ is a power of 2. After doing these two preliminary reductions we can use the above reduction for the case that m is a power of 2 to reduce the

$(\hat{m},d,\epsilon/4)$-hitting set problem to the (m',d,ϵ')-hitting set problem where $m'=\hat{m}\times \max(\hat{m},2d)=\hat{m}^2=O(m^2d^2)$ and $\epsilon'=\epsilon/4$.

7. Dimension reductions

The three dimension reductions used in the construction are not all the same, but they have a common structure. We want to reduce the hitting set problem for a collection $\mathcal{A}$ of (possibly restricted) (m,d,ϵ)-rectangles to the hitting set problem for the set $\mathcal{B}$ of all (m,p,γ)-rectangles where p is some number less than d, and γ is not much smaller than ϵ. So we need to define a bipartite graph between $[m]^d$ and $[m]^p$. This bipartite graph will be completely specified by a family F of functions from $[d]$ to $[p]$. The graph $G=G_F$ associated to F is defined as follows. For each point $(x_1,x_2,\dots,x_p)$ of $[m]^p$ and each function $f\in F$, there will be an edge from $(x_{f(1)},x_{f(2)},\dots,x_{f(d)})$ to $(x_1,x_2,\dots,x_p)$. Thus each vertex in $[m]^p$ will have in-degree $|F|$ (counting possibly multiple edges).

Let us now formulate sufficient conditions on the set F of functions so that G_F is a reduction. We need that if $\mathrm{R}=\mathrm{R}_1\times\mathrm{R}_2\times\dots\times\mathrm{R}_d$ is a rectangle in $\mathcal{A}$ then $G^+(\mathrm{R})$ contains some rectangle in $\mathcal{B}$. Fix some function $f\in F$, and examine the edges of G that are defined by f. Note that f^{-1} defines an ordered partition $f^{-1}(1),f^{-1}(2),\dots,f^{-1}(p)$ of $[d]$ into p (possibly empty) parts. Define the rectangle $\mathrm{R}_f=\mathrm{R}_{f^{-1}(1)}\times\mathrm{R}_{f^{-1}(2)}\times\dots\times\mathrm{R}_{f^{-1}(p)}$ in $[m]^p$, where we recall that for $J\subseteq[m]$, R_J is defined to be $\cap_{j\in J}R_j$. If $(x_1,x_2,\dots,x_p)$ is an arbitrary point in the rectangle R_f then it is joined by an edge labeled by the function f to the point $(y_1,y_2,\dots,y_d)$ with $y_i=x_{f(i)}$ for each $i\in[d]$. Since this point is in R, we conclude that $G^-(\mathrm{R})$ contains R_f. We say that a function f is γ-good for a rectangle R if $\mathrm{vol}(\mathrm{R}_f)$ is at least γ. We now have a sufficient condition on the family F of functions such that G_F is a reduction from $\mathcal{A}$ to $\mathcal{B}$ is given by:

Condition D. For each rectangle R in $\mathcal{A}$, there is a function $f\in F$ that is γ-good for R.

Below, we present two lemmas. The first gives a sufficient condition for a function f to be $\epsilon/2$-good for a fixed rectangle of volume ϵ. The second gives a (much weaker) sufficient condition for a function f to be ϵ^2-good for a fixed PIP-rectangle of volume ϵ. We will use these sufficient conditions to guide the choices of the specific reductions used in the construction.

We first need some notation. For the rest of the paper, we will typically consider a fixed function f from $[d]$ to $[p]$ and a fixed rectangle R in $[m]^d$. In this situation we define the following real parameters associated with R and f. For $i\in[d]$, define $\beta_i=|\mathrm{R}_i|/m$ and $\delta_i=1-\beta_i$. For any subset $S\subseteq[d]$, define

$$\beta_S=|\mathrm{R}_S|/m,$$

$$\pi(S) = \prod_{j\in S} \beta_j,$$
$$\nu(S) = \sum_{j\in S} \delta_j,$$
$$\mu(S) = \sum_{i,j\in S, i\neq j} \delta_i\delta_j.$$

Let us compare the volume of R_f to the volume of R. $\mathrm{vol}(\mathrm{R}_f)$ is equal to $\prod_{t=1}^{p}\beta_{f^{-1}(t)}$. We can write $\mathrm{vol}(\mathrm{R})=\prod_{t=1}^{p}\pi(f^{-1}(t))$. The term $\pi(f^{-1}(t))$ can be thought of as an "estimate" of $\beta_{f^{-1}(t)}$ that would be valid, had the sets R_j for $j\in f^{-1}(t)$ been mutually independent. The problem is that in general, for a subset S of $[d]$, β_S can be much smaller than $\pi(S)$. Thus for f to be γ-good, we want that for each t, $\beta_{f^{-1}(t)}$ should be not much less than $\pi(f^{-1}(t))$.

It is useful to think about the problem of choosing a function f that is γ-good for R in the following way. Think of elements of the set $[d]$ as "items", and think of δ_i as the "weight" of item i. The sequence $\delta=\delta(\mathrm{R})=(\delta_1,\ldots,\delta_d)$ is called the *weight sequence* of the rectangle R. Recall that $k=k(\epsilon)=\ln(1/\epsilon)$. Noting that $\mathrm{vol}(\mathrm{R})=\prod_{i=1}^{d}(1-\delta_i)\leq e^{-\nu([d])}$, we have:

Proposition 6. *For any rectangle* R *of volume at least ϵ, $\nu([d])$, the sum of the weights of all items associated with* R*, is at most k.*

Think of the elements of $[p]$ as "bins" into which f places the items. In the two lemmas below, the sufficient conditions for f to be γ-good for R depend only on the "weight distribution" of the items in the bin; roughly they require that bins not be too crowded. Intuitively, such a condition ensures that for each t, $\beta_{f^{-1}(t)}$ is not too much smaller than $\pi(f^{-1}(t))$.

To formulate the conditions on the weight distribution, we define $\rho=\rho(f,\delta)$ to be $\sum_{t=1}^{p}\mu(f^{-1}(t))$. Finally, for each $t\in[p]$, let $f^*(t)$ denote the singleton set containing a fixed $i\in f^{-1}(t)$ such that δ_i is maximum, provided that $f^{-1}(t)$ is non-empty; otherwise $f^*(t)=\emptyset$. Let $f^{\#}(t)$ denote the set $f^{-1}(t)-f^*(t)$. In words, $f^{\#}(t)$ is the set that remains after removing a heaviest item from $f^{-1}(t)$.

We can now state the two sufficient conditions for γ-goodness.

Lemma 7. *Suppose* R *is a rectangle with weight sequence δ, and that f maps $[d]$ to $[p]$. Then*

$$\mathrm{vol}(\mathrm{R}_f) \geq \mathrm{vol}(\mathrm{R}) - \rho(f,\delta).$$

In particular, if R *has volume ϵ and $\rho(f,\delta)\leq\epsilon/2$ then f is $\epsilon/2$-good for* R.

Lemma 8. *Let* R *be any PIP-rectangle and f maps $[d]$ to $[p]$. Let δ be the weight sequence for* R. *If for each $t \in [p]$, $\nu(f^{\#}(t)) \le 1/2$ then $\mathrm{vol}(\mathrm{R}_f) \ge (\mathrm{vol}(R))^2$. In particular, if* R *has volume at least ϵ then f is ϵ^2-good for* R.

In the first lemma it is required that the sum of the pairwise products of weights of elements that are mapped to the same bin should be small. The condition in the second lemma says that for each bin, the sum of the weights of the items assigned to the bin, excluding the heaviest item in the bin, is at most 1/2.

Proof of Lemma 7. As noted above, $\mathrm{vol}(\mathrm{R}) = \prod_{t=1}^{p} \pi(f^{-1}(t))$. Also, $\mathrm{vol}(\mathrm{R}_f) = \prod_{t=1}^{p} \beta_{f^{-1}(t)}$. For q between 1 and p, let $G(q)$ denote the inequality:

$$\prod_{t=1}^{q} \beta_{f^{-1}(t)} \ge \prod_{t=1}^{q} \pi(f^{-1}(t)) - \sum_{t=1}^{q} \mu(f^{-1}(t)).$$

Note that $G(p)$ is the conclusion of the lemma; we will prove that $G(q)$ holds for each $q \le p$, by induction. It is easily proved by induction on $|S|$ that for any $S \subseteq [d]$, $\pi(S) \le 1 - \nu(S) + \mu(S)$. Thus:

$$\beta_S \ge 1 - \nu(S) \ge \pi(S) - \mu(S). \tag{1}$$

Applying this with $S = f^{-1}(1)$, we obtain the base case $G(1)$.

For the induction step $q > 1$, we assume $G(q-1)$. If the right hand side of $G(q-1)$ is negative, then so is the right hand side of $G(q)$ (since the first term on the left can only decrease and the second term can only increase), and so the relation follows. So assume the right hand side of $G(q-1)$ is positive. Multiply both sides by $\beta_{f^{-1}(q)}$. We obtain:

$$\begin{aligned}
\prod_{t=1}^{q} \beta_{f^{-1}(t)} &\ge \beta_{f^{-1}(q)} \left(\prod_{t=1}^{q-1} \pi(f^{-1}(t)) - \sum_{t=1}^{q-1} \mu(f^{-1}(t)) \right) \\
&\ge \beta_{f^{-1}(q)} \prod_{t=1}^{q-1} \pi(f^{-1}(t)) - \sum_{t=1}^{q-1} \mu(f^{-1}(t)) \\
&\ge (\pi(f^{-1}(q)) - \mu(f^{-1}(q))) \prod_{t=1}^{q-1} \pi(f^{-1}(t)) - \sum_{t=1}^{q-1} \mu(f^{-1}(t)) \\
&\ge \prod_{t=1}^{q} \pi(f^{-1}(t)) - \sum_{t=1}^{q} \mu(f^{-1}(t)),
\end{aligned}$$

as required to prove the lemma. (Here the second inequality follows from $\beta_S \le 1$ for any S and the third inequality follows from inequality (1)). ∎

Proof of Lemma 8. Since $\mathrm{vol}(\mathrm{R}) = \prod_{t=1}^{p} \pi(f^{-1}(t))$ and $\mathrm{vol}(\mathrm{R}_f) = \prod_{t=1}^{p} \beta_{f^{-1}(t)}$ the lemma follows from:

Proposition 9. *For any weight sequence δ and $t \in [p]$, if $\nu(f^{\#}(t)) \leq 1/2$ then $\beta_{f^{-1}(t)} \geq \pi(f^{-1}(t))^2$.*

So, let us prove the proposition. If $f^{-1}(t)$ is empty, both sides of the inequality are 1. So suppose $f^{-1}(t)$ is nonempty. Let $S = f^{-1}(t)$, $\{s^*\} = f^*(t)$. We have $\mathrm{R}_S = \cap_{i \in S} R_i = \mathrm{R}_{s^*} - (\cup_{i \in S - \{s^*\}} (R_{s^*} \cap ([m] - R_i))$. Thus:

$$|R_S| \geq |\mathrm{R}_{s^*}| - \sum_{i \in S - \{s^*\}} |\mathrm{R}_{s^*} \cap ([m] - R_i)|,$$

and therefore

$$\begin{aligned} \beta_S &= \beta_{s^*} - \sum_{i \in S - \{s^*\}} \beta_{s^*} \delta_i \\ &= \beta_{s^*}(1 - \nu(S - \{s^*\})), \end{aligned}$$

where the first equality comes from the fact that R is a PIP-rectangle.

Now $\pi(S)^2 = \pi(S - \{s^*\})^2 \beta_{s^*}^2 \leq \pi(S - \{s^*\})^2 \beta_{s^*}$. Furthermore, $\pi(S - \{s^*\}) = \prod_{i \in S - \{s^*\}} (1 - \delta_i) \leq e^{-\nu(S - \{s^*\})}$, so $\pi(S)^2 \leq \beta_{s^*} e^{-2\nu(S - \{s^*\})}$. Now, if $x \leq 1$, $e^{-x} \leq 1 - x/2$ and so since $\nu(S - \{s^*\}) \leq 1/2$ by hypothesis, we get $\pi(S)^2 \leq \beta_{s^*}(1 - \nu(S - \{s^*\})) \leq \beta_S$, as required to prove the proposition and the lemma. ∎

Armed with Lemmas 7 and 8 we are ready to present the three reductions needed to accomplish the construction.

8. The reduction sequence

8.1. Reduction 1.

We will choose a family of mappings F_1 from $[d_0]$ to $[d_1]$ that satisfies Condition D, and for this we will make use of Lemma 6. It will be useful to introduce some additional notation. For $i, j \in [d_0]$ and function f with domain $[d_0]$, define the indicator function $\chi_{i,j}(f)$ to be 1 if $f(i) = f(j)$ and 0 otherwise. With this notation $\rho(f, \delta) = \sum_{i,j \in [d_0], i \neq j} \chi_{i,j}(f) \delta_i \delta_j$. Suppose that F is a family of mappings and let $C(F)$ be the maximum over all $i, j \in [d_0]$, $i \neq j$, of $\sum_{f \in F} \chi_{i,j}(f)$. Then for any rectangle R, the average of $\rho(f, \delta)$ over all $f \in F$ is at most:

$$\begin{aligned} \frac{\sum\limits_{f \in F} \rho(f, \delta)}{|F|} &= \frac{\sum\limits_{f \in F} \sum\limits_{i,j \in [d_0], i \neq j} \chi_{i,j}(f) \delta_i \delta_j}{|F|} \\ &= \frac{\sum\limits_{i,j \in [d_0], i \neq j} \delta_i \delta_j \sum\limits_{f \in F} \chi_{i,j}(f)}{|F|} \end{aligned}$$

$$\leq \frac{C(F)}{|F|}\mu([d_0])$$

$$\leq \frac{C(F)}{|F|}\left(\sum_{i=1}^{d_0}\delta_i\right)^2$$

$$\leq \frac{C(F)}{|F|}k_0^2.$$

So if we can choose F^1 such that $C(F^1)/|F^1| \leq \epsilon_0/(2k_0^2)$, then for each R there must be an $f \in F^1$, such that $\rho(f,\delta) \leq \epsilon_0/2$. We can then apply Lemma 7, to conclude that the volume of R_f is at least $\epsilon_0/2$. Thus Condition D is satisfied.

So it remains to construct F^1 such that $C(F^1)/|F^1|$ is at most $\epsilon_0/2k_0^2$. For each positive integer a, let ϕ_a denote the a^{th} prime, and define the function f_a by $f_a(i) = i \bmod \phi_a{+}1$, so that f_a maps every integer to $[\phi_a]$. Note that $f_a(i) = f_a(j)$ if and only if a divides $i-j$. Thus the number of functions f_a for which $f_a(i) = f_a(j)$ is the number of prime divisors of $i-j$ which is at most $\log|i-j| \leq \log d_0$, for $1 \leq i \neq j \leq d_0$, i.e., $C(F^1) \leq \log d_0$. Define the parameter t to be $\lceil 2k_0^2(\log d_0)/\epsilon_0 \rceil$, and define $F^1 = \{f_a | 1 \leq a \leq t\}$. Then $|F^1| = t \geq 2k_0^2(\log d_0)/\varepsilon_0$, so $C(F^1)/|F^1| \leq \epsilon_0/2k_0^2$. Furthermore, if we define $d_1 = \phi_t$ then $d_1 = O(t\log t)$ and all of the functions in F^1 map $[d_0]$ to $[d_1]$. Thus F^1 achieves the desired dimension reduction, and the cost of this reduction is at most $t = \lceil 2k_0^2(\log d_0)/\epsilon_0 \rceil$.

8.2. Reduction 2.

For reduction 2a, we first apply a PIP-reduction, to reduce to the problem of finding a $(m_2, d_1, \epsilon_1/4)$ PIP-hitting set.

For reduction 2b, we will choose F^2 to be a family of universal hashing functions from $[d_1]$ to $[d_2]$. We want to show that for each $(m_2, d_1, \epsilon_1/4)$ PIP-rectangle R, there is an f satisfying the hypothesis of Lemma 8, i.e., for each $t \in [d_2]$, $\nu(f^{\#}(t)) \leq 1/2$.

Fix such a rectangle R and consider a map f from $[d_1]$ to $[d_2]$. Let us say that bin $t \in [d_2]$ is *bad* for the map f if $\nu(f^{\#}(t)) > 1/2$. Let $B(f)$ be the set of $i \in [d_1]$ such that $f(i)$ is bad for f. Then the hypothesis of Lemma 8 is equivalent to $B(f) = \emptyset$. Notice that $\nu(B(f))$ is the sum of the weights of items that are placed into bad bins and if $B(f)$ is nonempty then $\nu(B(f))$ must be greater than 1/2. Thus a sufficient condition for f to satisfy the hypothesis of Lemma 8 is $\nu(B(f)) \leq 1/2$. The following fact about universal families of hashing functions is key:

Lemma 10. *Suppose that* $\mathrm{R} \subseteq [z]^q$ *is a PIP-rectangle. Let* F *be a universal family of hashing functions from* $[q]$ *to* $[p]$. *For each* $i \in [q]$, *if* f *is chosen uniformly from* F *then the probability that* i *belongs to* $B(f)$ *does not exceed* $2\nu([q])/p$. *Thus the expectation of* $\nu(B(f))$ *is at most* $2\nu([q])^2/p$.

Proof. Note first that the expectation $\mathrm{E}[\nu(B(f))] = \sum_{i=1}^{q} \delta_i \mathrm{Prob}[i \in B(f)]$. For $i \in [q]$, define the random variable $X_i = \sum_{j \neq i} \delta_j \chi_{i,j}(f)$, i.e., the sum of the weights of the elements other than i mapped to the same location as i. The event $i \in B(f)$ implies that $X_i \geq 1/2$, so $\mathrm{E}[\nu(B(f))] \leq \sum_{i=1}^{q} \delta_i \mathrm{Prob}[X_i \geq 1/2]$. By Markov's inequality, $\mathrm{Prob}[X_i \geq 1/2] \leq 2\mathrm{E}[X_i]$, and this is at most $2\nu([q])/p$, hence $\mathrm{E}[\nu(B(f))] \leq \sum_{i=1}^{q} \delta_i (2\nu([q])/p = 2\nu([q])^2/p$. ∎

Now, taking q to be d_1 and p to be d_2 in the above lemma, we have by Proposition 6 that $\nu([d_1]) \leq k_1$ and so if we choose d_2 to be the least power of 2 that is at least $4k_1^2$, we conclude from the lemma that for any rectangle R, there is an $f \in F^2$ that satisfies the hypothesis of Lemma 8. Note also that the family F^2 has size $O(d_1 \max(d_2, 2d_1))$. We may assume $d_2 < d_1$, since otherwise we may skip reduction 2, and thus $|F^2| = O(d_1^2)$.

8.3. Reduction 3.

For reduction 3a, we first apply a PIP-reduction, to reduce to the problem of finding a $(m_3, d_2, \epsilon_2/4)$ PIP-hitting set, as in reduction 2a.

To define reduction 3b, we begin by following the argument in reduction 2b. As in that reduction, it suffices to define a family F^3 of functions from $[d_2]$ to $[d_3]$ such that for each $(m_3, d_2, \epsilon_2/4)$ PIP-rectangle R there is an $f \in F^3$, such that for each $t \in [d_3]$, $B(f)$ is empty. In reduction 2b, a universal hash function family was shown to be sufficient to achieve this goal; provided that we did not go below dimension $O(k_0^2)$. What we want is to get the dimension down to $O(k_0)$ (which is also $O(k_2)$). We will need a more complicated family of functions, one that is similar to a family previously used in [15].

Look again at Lemma 10. Notice that if we take $p \geq 4\nu([q])$ then this implies that there is an $f \in F$ for which $\nu(B(f)) \leq \nu([q])/2$, i.e., the weight of the elements that are mapped to bad locations is at most half the total weight. What we want to do is to "collect" the elements that are mapped to bad locations and remap them to new locations.

This idea leads to the following iterative mapping scheme. Partition the set of positive integers into the consecutive intervals $I_0 = \{1\}, I_1 = \{2,3\}, I_2 = \{4,5,6,7\}, \ldots$ (i.e., $I_j = \{2^j, \ldots, 2^{j+1} - 1\}$). Let F_j be a universal family of hashing functions from $[q]$ to I_j. Define a *remapping sequence* of order r to be a sequence $\sigma = (f_r, J_r, f_{r-1}, J_{r-1}, \ldots, f_1, J_1, f_0)$ where each f_j is in F_j and each J_j is a subset of I_j.

Such a sequence defines a map g^σ from $[q]$ to $I_0 \cup I_1 \cup \ldots \cup I_r$ as follows: each element i is mapped to $f_j(i)$ where j is the first index (starting from r downwards) such that $f_j(i) \notin J_j$. An alternative, more algorithmic description of g^σ is this: tentatively map each element of $[q]$ to I_r using f_r. For those elements i that are not mapped to J_r, set $g^\sigma(i) = f_r(i)$. The elements that are mapped to J_r are remapped according to f_{r-1}; again for those elements i that are not mapped to J_{r-1}, $g^\sigma(i) = f_{r-1}(i)$ and the rest are remapped.

Lemma 11. *Let* $\mathrm{R} \subseteq [z]^q$ *be a PIP-rectangle, and let* ℓ *be the least nonnegative integer such that* $2^\ell \geq 4\nu([q])$. *Then there is a remapping sequence* σ *of order at most* ℓ *such that the associated map* g^σ *has* $B(g^\sigma) = \emptyset$.

The proof of this lemma is an easy induction on ℓ. If $\ell = 0$ then $\nu([q]) \leq 1/4$ and so if we choose our remapping sequence to be (f_0) where f_0 is the unique function in F_0, then $g^\sigma = f_0$, and $B(g^\sigma)$ is trivially empty. For $\ell > 0$, by Lemma 10 we can choose $f_\ell \in F_\ell$ such that the total weight of the elements mapped to bad bins is at most half the total weight. Then choose J_ℓ to be the set of bad locations for f_ℓ. The set of unmapped elements is $B(f_\ell)$, and we can apply induction (with q replaced by $|B(F_\ell)|$) to get a sequence of order $\ell - 1$ that maps these elements with no bad bins. ∎

This lemma now allows us to construct our family F^3 of functions. Let ℓ be $\lceil 2 + \log k_2 \rceil$, define d_3 to be $2^{\ell+1}$, and let F^3 be the set of all maps of the form g^σ where σ is a remapping sequence of length ℓ. The previous lemma implies that for any (m, d, ϵ) PIP-rectangle there is an $f \in F^3$ which satisfies the hypothesis of Lemma 8, as required.

Finally, it remains to observe that the size of the family F^3 is at most the product of the sizes of the F_r for r between 0 and ℓ times 2^{2d_3} (for the choices of the sets J_r) which is $2^{O(k_2)} = 2^{O(k_0)}$ and is thus bounded by a polynomial in $1/\epsilon$.

9. Open problems

The problem considered in this paper was directly motivated by the *discrepancy* problem stated in [5]. The discrepancy problem is the stronger version of the hitting problem, where, instead of hitting each rectangle R with $\mathrm{vol}(\mathrm{R}) \geq \epsilon$ at least once, each rectangle is hit a fraction of times that is within ϵ of its volume. An explicit construction for a sample space of polynomial size that solves the discrepancy problem is still not known. Besides the application to an explicit construction of a small sample space, a solution to the discrepancy problem has a number of other applications, including applications to numerical integration.

There are further natural questions that generalize the discrepancy question. For example: Is there a polynomial time algorithm that on input d, m, ϵ, and k, produces a set S of size at most polynomial in d, m, $1/\epsilon$ and k with the following property: for every set of at most k rectangles, the volume of the set of points that are contained in at least one of the k rectangles has discrepancy at most ϵ with respect to S. (When $k = 1$ this is the discrepancy problem.) A solution to this problem would immediately yield a polynomial time deterministic approximation algorithm for the DNF counting problem. Building on ideas of the present paper, [2], made some progress on this problem. As described in [11] progress on this problem has also been made using a different approach. See also [8].

A somewhat less natural generalization of the discrepancy problem is motivated by the *GF[2]* counting problem considered in [10] and [12]: Is there a polynomial time algorithm that on input d, m, ϵ, and k, produces a set S of size at most polynomial in d, m, $1/\epsilon$ and k with the following property: for every set of at most k rectangles, the volume of the set of points that are contained in an odd number of the k rectangles has discrepancy at most ϵ with respect to S. (When $k=1$ this is again the discrepancy problem.) A solution to this problem would immediately yield a polynomial time deterministic approximation algorithm for the above-mentioned *GF[2]* counting problem. As described in [12], some progress on this problem has been made using a different approach.

Acknowledgment. The authors are very grateful to Boban Veličković for discovering a crucial error in the original paper, and to the anonymous referee for a careful reading and very helpful suggestions.

References

[1] M. Ajtai, J. Komlós, and E. Szemerédi: Deterministic Simulation in LOGSPACE, in: *Proc. of 19th ACM Symposium on Theory of Computing,* 1987, 132–140.

[2] R. Armoni, M. Saks, A. Wigderson, and S. Zhou: Discrepancy sets and pseudorandom generators for combinatorial rectangles, *Proc. 37th IEEE Symposium on Foundations of Computer Science,* 1996.

[3] J. L. Carter, M. N. Wegman: Universal classes of hash functions, *J. Comput. System Sci.*, **18** (1979), 143–154.

[4] B. Chor, O. Goldreich: On the Power of Two–Point Based Sampling, *Journal of Complexity*, **5** (1989), 96–106.

[5] G. Even, O. Goldreich, M. Luby, N. Nisan, and B. Boban Veličković: Approximations of General Independent Distributions, in: *Proc. of the 24th ACM Symposium on Theory of Computing*, 1992.

[6] J. Friedman: Constructing $O(n\log(n))$ size monotone formulae for the kth threshold function of n boolean variables, *SIAM J. on Computing*, **15** (1986).

[7] O. Gabber, Z. Galil: Explicit constructions of linear-sized superconcentrators, *Journal of Computer System Science,* **22** (1981), 407–420.

[8] J. Kahn, N. Linial, and A. Samorodnitsky: Inclusion–Exclusion: Exact and approximate, *Combinatorica*, **16** (1996), 465–477.

[9] R. Karp, M. Pippenger, and M. Sipser: *Time-Randomness Tradeoff,* presented at the AMS conference on probabilistic computational complexity, Durham, New Hampshire, 1982.

[10] M. Karpinski, M. Luby: Approximating the Number of Solutions to a *GF[2]* Formula, *Journal of Algorithms*, **14** (1993), 280–287.

[11] M. Luby, B. Veličković: On Deterministic Approximation of DNF, in: *Proceedings of 23rd ACM Symposium on Theory of Computing,* 1991, 430–438, *Algorithmica* (special issue devoted to randomized algorithms), Vol. 16, No. 4/5, October/November 1996, 415–433.

[12] M. Luby, A. Wigderson, and B. Veličković: Deterministic Approximate Counting of Depth-2 Circuits, in: *Proceedings of the Second Israeli Symposium on Theory of Computing and Systems*, 1993, 18–24.

[13] G. A. Margulis: Explicit construction of concentrators, *Problemy Peredači Informacii* 9, (1973) 71–80. (English translation in *Problems Inform. Transmission,* 1975).

[14] N. Nisan: Pseudorandom Generators for Space-Bounded Computation, *Combinatorica,* **12** (1992), 449–461.

[15] J. Schmidt, A. Siegel: The Spatial Complexity of oblivious k-probe hash functions, *SIAM Journal on Computing,* **19** (1990), 775–786.

[16] M. Sipser: Expanders, Randomness, or Time vs. Space, *Journal of Computer and System Sciences*, **36** (1988), 379–383.

[17] D. Zuckerman: Simulating BPP Using a General Weak Random Source, *Algorithmica,* **16** (1996), 367–391.

Nathan Linial

Hebrew University,
Computer Science Department,
Jerusalem, Israel
nati@cs.huji.ac.il

Michael Saks

Department of Mathematics,
Rutgers University,
New Brunswick, NJ 08854, USA
saks@lagrange.rutgers.edu

Michael Luby

DEC/SRC, 130 Lytton Avenue,
Palo Alto,
California 94301
luby@pa.dec.com

David Zuckerman

Department of Computer Sciences,
The University of Texas at Austin,
Austin, Texas, 78713, USA
diz@cs.utexas.edu