Approximations of General Independent Distributions[1]

Guy Even[2] Oded Goldreich[3] Michael Luby[4] Noam Nisan[5]
Boban Veličković[6]

(Reproduced from files dating Nov. 1991)

[1]Preliminary version has appeared in the *Proceedings of the 24th ACM Symp. on Theory of Computing (STOC)*, pages 10–16, 1992. Research partially supported by NSF operating grant CCR-9016468 and by grant No. 89-00312 from the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel.

[2]Computer Science Department, Technion, Haifa, Israel.

[3]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.

[4]International Computer Science Institute, 1947 Center Street, Berkeley, California 94704, USA.

[5]Department of Computer Science, Hebrew University, Jerusalem, Israel

[6]Department of Mathematics, U.C. Berkeley.

Abstract

This report describes efficient constructions of small probability spaces that approximate the joint distribution for general random variables. These results yield efficient constructions of small sets with low discrepancy in high dimensional space.

Chapter 1 contains the text (more or less) as it has appeared in the *STOC92* proceedings. In that version the problem of approximating the product distribution of general random variables is reduced to the construction of small discrepancy sets, and all constructions are presented in terms of the latter problem.

Chapter 2 contains contains an ealier version in which only the first construction is presented. The exposition is in terms of the original problem (i.e., of approximating the product distribution of general random variables).

In Chapter 3 we provide details for a construction of a small bias sample space over $GF(p)$, for $p > 2$. References to this construction are made in Section 2.3.3 (and at the end of Section 1.3.1) as well as in subsequent literature, however so far details have only appeared in the first author's M.Sc. Thesis (in Hebrew).

Contents

Chapter 1

The STOC92 Version: Approximation and Discrepancies

Abstract

We describe efficient constructions of small probability spaces that approximate the joint distribution for general random variables. Previous work on efficient constructions concentrate on approximations of the joint distribution for the special case of uniform boolean-valued random variables. Our results yield efficient constructions of small sets with low discrepancy in high dimensional space and have applications to derandomizing randomized algorithms.

1.1 Introduction

The problem of constructing small sample spaces that "approximate" the independent distribution on n random variables has received considerable attention recently (cf. [7, Chor Goldreich] [9, Karp Wigderson], [12, Luby], [1, Alon Babai Itai], [15, Naor Naor], [2, Alon Goldreich Håstad Peralta], [3, Azar Motwani Naor]). The primary motivation for this line of research is that random variables that are "approximately" independent suffices for the analysis of many interesting randomized algorithm and hence constructing a *small probability space that "approximates" the independent distribution* yields a way to "derandomize" these algorithms, i.e. convert them to deterministic algorithms of reasonable complexity by using the deterministically constructed sample space in place of the "internal coin tosses" of the algorithm. The culmination of previous works are constructions of small sample spaces that approximate a constant amount of independence for general random variables (see for example [12, Luby] or [1, Alon Babai Itai]) or that approximate complete independence for identically and uniformly distributed boolean-valued random variables [15, Naor Naor], [2, Alon Goldreich Håstad Peralta], [3, Azar Motwani Naor]). Although previous results are sufficient for some applications to derandomizing algorithms, in many applications what is needed is a small sample space that approximates more than a constant amount of independence for general random variables. In this paper we present constructions of small efficiently constructible sample spaces for this more general case.

1.1.1 Definitions of Approximation

The probability distribution on n general m-valued random variables is described by a n by m *probability matrix* $\mathcal{P}_{n,m} = \{p_{i,v} : i \in \{1,\ldots,n\}, v \in \{0,\ldots,m-1\}\}$, which is a matrix of non-negative entries such that the sum of the entries in each row is equal to 1. The (i,v)-entry $p_{i,v}$ specifies the probability that the i^{th} random variable should take on value v. For all values of

$l \in \{1,\ldots,n\}$, for all $I =< i_1,\ldots,i_l >$, where $1 \le i_1 < \ldots < i_l \le n$, and for all $V =< v_1,\ldots,v_l > \in \{0,\ldots,m-1\}^l$, let $p_{I,V} = \prod_{j=1}^{l} p_{i_j,v_j}$ be the probability that the subsequence of random variables indexed by I should take on value V if the random variables were truly independent.

From $\mathcal{P}_{n,m}$ we want to produce a finite set S that induces a distribution on n random variables $x_1,\ldots,x_n$ which approximates the independent distribution for $\mathcal{P}_{n,m}$. All constructions for S given in this paper are efficient in the sense that there is a deterministic algorithm which produces S in time polynomial in the length of the description of $\mathcal{P}_{n,m}$ and in the length of the description of S. The description of each point $s \in S$ consists of, for each index i, a value $x_i(s) \in \{0,\ldots,m-1\}$ for the i^{th} random variable. We view S as a sample space that induces a distribution on $x_1,\ldots,x_n$ defined by choosing a point randomly and uniformly from S. We let x_I be the subsequence of random variables indexed by I, we let $x_I = V$ denote the event that the subsequence x_I takes on the value V, and we let $\mathrm{P}_S[x_I = V]$ be the probability that event $x_I = V$ occurs in the distribution induced by S. We say that S is *independent* for $\mathcal{P}_{n,m}$ if it induces a distribution on $x_1,\ldots,x_n$ such that, for all l, $I =< i_1,\ldots,i_l >$ and $V =< v_1,\ldots,v_l >$, $\mathrm{P}_S[x_I = V] = p_{I,V}$. (Hereafter, the quantification "for $\mathcal{P}_{n,m}$" is omitted for brevity whenever $\mathcal{P}_{n,m}$ is clear from the context.) We say S is a (k,ϵ)-*approximation* if for any subsequence I of size $l \le k$ and for any set of possible values $V \in \{0,\ldots,m-1\}^l$, $|\mathrm{P}_S[x_I = V] - p_{I,V}| \le \epsilon$. S is an ϵ-*approximation* if this statement is true with no restriction on the size of I and S is a k-*wise independent approximation* if this statement is true with $\epsilon = 0$.

1.1.2 Previous Work on Approximation

Let $\mathcal{U}_{n,m}$ be the probability matrix with all entries equal to $1/m$ that describes the special case of n identically and uniformly distributed m-valued random variables. Thus, $\mathcal{U}_{n,2}$ is the important subcase where all entries are $1/2$ that describes n identically and uniformly distributed boolean-valued random variables. It is fairly easy to prove that S has to be of size at least 2^n to be independent even for $\mathcal{U}_{n,2}$. Constructions of sample spaces that are k-wise independent approximations for $\mathcal{U}_{n,m}$ of size $\max\{n,m\}^k$, and that are (k,ϵ)-approximations for general $\mathcal{P}_{n,m}$ with size $(\max\{n,k/\epsilon\})^k$, are implicit in many works; a brief survey of some of these constructions can be found in either [12, Luby] or [1, Alon Babai Itai]. For constant k and $1/\epsilon$ polynomial in n, this yields a sample space of size polynomial in n.

It has been recognized that in many other examples, what is needed is a sample space that has more than a constant amount of independence between the n random variables; typically logarithmic in n independence suffices. On the other hand, it has been shown that the sample space has to be of size at least $n^{k/2}$ in order to be a k-wise independent approximation for $\mathcal{U}_{n,2}$ [6, Chor Freidmann Goldreich Håstad Rudich Smolensky], and for non-constant k this is not polynomial in n. [15, Naor Naor] introduced the idea of allowing the error parameter ϵ and gave an ingenious construction of a sample space that is an ϵ-approximation for $\mathcal{U}_{n,2}$ where the size of the sample space is $O(n\log(n)/\epsilon^4)$. Simpler constructions with a sample space of size $O((n\log(n))^2/\epsilon^2)$ for $\mathcal{U}_{n,2}$ were subsequently presented in [2, Alon Goldreich Håstad Peralta]. These constructions can be extended to $\mathcal{U}_{n,m}$, basically using the same ideas, but in a slightly more complicated way (cf. [2, Alon Goldreich Håstad Peralta], [3, Azar Motwani Naor], [8, Even]), where the size of the resulting sample space is $O((n\log(n))^2/\epsilon^2)$.

1.1.3 New Results on Approximation

For some applications the constructions described in the previous subsection are quite useful. For example, in the analysis of some of the randomized algorithms for graph problems presented in [12, Luby] and [1, Alon Babai Itai], approximate pairwise independence of the random variables suffices. Thus, the construction of a sample space of polynomial size that is a pairwise independent approximation for general $\mathcal{P}_{n,m}$ can be used to convert these randomized algorithms into deterministic algorithms. In other applications (see [15, Naor Naor]), approximations of identically and uniformly distributed boolean-valued random variables suffice. However, in the more typical application the random variables are general and more than a constant amount of independence is required in the analysis, and thus it is of primary importance to develop constructions for these cases.

In this paper, we describe three constructions of small sample spaces that are approximations of the independent distribution for general $\mathcal{P}_{n,m}$; the first two constructions are new and the third is a construction based on a theorem in [16, Nisan]. The first construction yields a sample space that is a (k, ϵ)-approximation, where the size of the sample space is polynomial in $\log(n)$, 2^k and $1/\epsilon$. Previous results that achieve the same kind of approximation result in a sample space of size polynomial in $\log(n)$ and $(k/\epsilon)^k$. In contrast to previous results, when $k = O(\log(n))$ and $1/\epsilon$ is polynomial in n the size of the sample space in our construction is polynomial in n. This case is important to some applications, and in particular this construction improves the running time of some of the algorithms presented in [13, Luby Veličković].

The second and third constructions for n general random variables yield samples spaces that are ϵ-approximations for general $\mathcal{P}_{n,m}$, where the size of the sample space is polynomial in $(n/\epsilon)^{\log(1/\epsilon)}$ for the second construction and polynomial in $(n/\epsilon)^{\log(n)}$ for the third. In contrast, the previous bound on the sample space size, implicit in the classical work on discrepancy theory (see e.g. [4, Beck Chen] or [17, Niederreiter]), is polynomial in n^n/ϵ. For interesting cases of n and ϵ, i.e. when $1/\epsilon$ is polynomial in n, the results presented here are dramatic improvements.

1.1.4 Discrepancies

Let $[0,1)^n$ be the n dimensional unit cube, let $\mathcal{R}_n$ be the set of all axis parallel rectangles within $[0,1)^n$, and for each $R \in \mathcal{R}_n$, let $\text{vol}(R)$ be the volume of R. For any finite set of points S in $[0,1)^n$ and for any $R \in \mathcal{R}_n$, define the *discrepancy of S on R* as $\text{disc}_S(R) = |\text{vol}(R) - |S \cap R|/|S||$. This quantity is the absolute value of the difference between the probability that a randomly chosen point from $[0,1)^n$ falls in R and the probability that a randomly chosen point from S falls in R. For any $\mathcal{K}_n \subseteq \mathcal{R}_n$, the *discrepancy of S on $\mathcal{K}_n$* is defined as $\Delta_S(\mathcal{K}_n) = \max_{R \in \mathcal{K}_n} \text{disc}_S(R)$. Finding explicit constructions of sets with small discrepancy have a variety of applications, including applications to numerical integration. The discrepancy problem can be stated as follows: given n and ϵ, construct a set S in $[0,1)^n$ with $\Delta_S(\mathcal{R}_n) \leq \epsilon$.

As we describe (and as also has been describe before by others, e.g. [17, Niederreiter]), there is an close connection between discrepancy and approximating independent distributions of n general random variables. Our primary interest in sets with small discrepancy is that they are *universal* for the problem of constructing sample spaces that are good approximations for general distributions. For example, a set S in $[0,1)^n$ for which $\Delta_S(\mathcal{R}_n) \leq \epsilon$ can be viewed as *universal* in the following sense: There is a simple efficient algorithm that given S and *any* $\mathcal{P}_{n,m}$ computes a sample space that is an ϵ-approximation for $\mathcal{P}_{n,m}$. To be of interest in the problem of approximating random variables, it is crucial that the size of S be small in terms of both parameters ϵ and n.

Classical work on the discrepancy problem concentrates on minimizing the size of S primarily as a function of $1/\epsilon$ and then secondarily as a function of n [4, Beck Chen], [17, Niederreiter], i.e. the dimension n is thought of as arbitrary but fixed and the goal is to find a set S with size as small as possible as a function of $1/\epsilon$. Although classical work shows that there are explicit constructions of S with size smaller than that implied by a random construction for fixed n, the bounds are exponential in n and say nothing non-trivial for values of n and ϵ interesting for the case of approximating general distributions. i.e. when n and $1/\epsilon$ are comparable.

The constructions presented here give new results for the discrepancy problem. For any constant $\beta < 1$ let $\mathcal{R}_n^{[0,\beta)} \subset \mathcal{R}_n$ be the set of rectangles R such that in each dimension i the length of R is either 1 or else it is in the range $[0,\beta)$. The first construction yields, for any constant $\beta < 1$, a set S in $[0,1)^n$ of size polynomial in both n and $1/\epsilon$ such that $\Delta_S(\mathcal{R}_n^{[0,\beta)}) \leq \epsilon$. The second construction yields a set S in $[0,1)^n$ of size polynomial in $(n/\epsilon)^{\log(1/\epsilon)}$ such that $\Delta_S(\mathcal{R}_n) \leq \epsilon$, and the third construction yields a set S with the same properties of size $(n/\epsilon)^{\log(n)}$. In contrast to these new results, the previous known bounds from classical discrepancy theory on the size of an explicitly constructible set S in $[0,1)^n$ with small discrepancy are exponential in n [4, Beck Chen], [17, Niederreiter].

It is easy to see that a random set of points S in $[0,1)^n$ of size $cn\log(n/\epsilon)/\epsilon^2$ for some constant $c > 1$ has the property that $\Delta_S(\mathcal{R}_n) \leq \epsilon$ with high probability. The crucial property missing from this proof of existence is efficient constructibility. We leave this as an open question, i.e. the problem of finding an explicit construction of a set S in $[0,1)^n$ with $\Delta_S(\mathcal{R}_n) \leq \epsilon$ and with $|S|$ polynomial in both n and $1/\epsilon$. As stated above, a solution to this problem would yield a universal set S of size polynomial in both n and $1/\epsilon$ that for all $\mathcal{P}_{n,m}$ can be interpreted as a sample space that is an ϵ-approximation for $\mathcal{P}_{n,m}$.

1.2 Linking discrepancy and approximation

In this section we provide the (straightforward) link between sets S with small discrepancy and sample spaces that approximate the independent distribution on n random variables.

Definition (classes of rectangles): The n dimensional unit cube is $[0,1)^n$. Let $R = \prod_{i\in\{1,\ldots,n\}} r_i$ be an axis-parallel rectangle in $[0,1)^n$, where each $r_i = [a_i, b_i)$ is a subinterval of $[0,1)$. We say R is *trivial* in dimension i if $r_i = [0,1)$. Without loss of generality, we restrict attention to those rectangles for which there is no $i \in \{1,\ldots,n\}$ with $a_i = b_i$. The volume of R is $\mathrm{vol}(R) = \prod_{i\in\{1,\ldots,n\}} b_i - a_i$.

- Define $\mathcal{R}_n$ to be the set of all axis parallel rectangles in $[0,1)^n$.
- For constant $\beta < 1$, define $\mathcal{R}_n^{[0,\beta)}$ to be the subset of $\mathcal{R}_n$ consisting of all rectangles $R = \prod_{i\in\{1,\ldots,n\}} r_i$ such that for each i either R is trivial in dimension i or else $r_i = [a_i, b_i)$ with $b_i - a_i \leq \beta$.
- For positive integer k, define $\mathcal{R}_n^k$ to be the subset of $\mathcal{R}_n$ consisting of all rectangles R such that R is trivial for all but at most k dimensions.
- For any $\mathcal{K}_n \subseteq \mathcal{R}_n$ and any positive integer $m \geq 2$, define $\mathcal{K}_{n,m}$ as the set of rectangles $R = \prod_{i\in\{1,\ldots,n\}} r_i \in \mathcal{K}_n$ such that each r_i is of the form $[a_i/m, b_i/m)$ for integers a_i and b_i satisfying $0 \leq a_i < b_i \leq m$. For example, $\mathcal{R}_{n,2}$ is the subset of $\mathcal{R}_n$ consisting of all rectangles $R = \prod_{i\in\{1,\ldots,n\}} r_i$ such that for each $i \in \{1,\ldots,n\}$, $r_i = [0,1/2)$ or $r_i = [1/2,1)$ or $r_i = [0,1)$.

Definition (projection sample space): Let S be a finite subset of points from $[0,1)^n$ and let $\mathcal{P}_{n,m}$ be a probability matrix. S can be viewed as the *projection sample space for* $\mathcal{P}_{n,m}$, inducing a distribution on random variables $x_1,\ldots,x_n$ as follows. For all $i \in \{1,\ldots,n\}$ let interval $r_{i,0} = [0, p_{i,0})$ and for all $v \in \{1,\ldots,m\}$ let interval $r_{i,v} = [a_{i,v}, b_{i,v})$, where $a_{i,v} = \sum_{0\leq w<v} p_{i,w}$ and $b_{i,v} = a_{i,v} + p_{i,v}$. Random variable x_i at a point $s =< s_1,\ldots,s_n >\in S$ takes on the unique value v that satisfies $s_i \in r_{i,v}$.

A set $S \subset [0,1)^n$ as just described is *universal* in the sense that it can be interpreted in a straightforward way as a sample space for *any* $\mathcal{P}_{n,m}$. The interpretation has the property that it is *coordinate independent* in the sense that the value given to x_i at sample point $s \in S$ depends only on the i^{th} coordinate of s and on the i^{th} row of $\mathcal{P}_{n,m}$.

The crucial links between discrepancies and approximations are the following observations.

1. If $\Delta_S(\mathcal{R}_n) \leq \epsilon$ then for any $\mathcal{P}_{n,m}$ the projection sample space of S for $\mathcal{P}_{n,m}$ is an ϵ-approximation.

2. If $\Delta_S(\mathcal{R}_n^k) \leq \epsilon$ then for any $\mathcal{P}_{n,m}$ the projection sample space of S for $\mathcal{P}_{n,m}$ is a (k,ϵ)-approximation.

3. If $\Delta_S(\mathcal{R}_{n,2}) \leq \epsilon$ then the projection sample space of S for $\mathcal{U}_{n,2}$ is an ϵ-approximation.

From this discussion it is clear that in order to produce sample spaces which approximate the independent distribution for n general random variables it suffices to produce small finite sets $S \subset [0,1)^n$ with small discrepancy.

Definition (the natural mapping to $[0,1)^n$): We can view a sample space S that induces a distribution on $x_1,\ldots,x_n$ for $\mathcal{U}_{n,m}$ in a natural way as a finite set of points in $[0,1)^n$, where the i^{th} coordinate of $s \in S$ is $x_i(s)/m$.

From this the converse of observation 3 follows, i.e. it is not hard to verify that if S is an ϵ-approximation for $\mathcal{U}_{n,2}$ then $\Delta_S(\mathcal{R}_{n,2}) \leq \epsilon$ when sample points in S are mapped to $[0,1)^n$ in the natural way. The converses of observations 1 and 2 are not so obvious. For example, it is true that if S is a sample space that is a (k,ϵ)-approximation for $\mathcal{U}_{n,m}$ then $\Delta_S(\mathcal{R}_{n,m}^k) \leq \epsilon m^k$ when points in S are mapped to $[0,1)^n$ in the natural way, but this is too weak of a bound for most purposes.

Some further useful observations are:

4. $\Delta_S(\mathcal{R}_{n,4n/\epsilon}) \leq \epsilon/2$ implies that $\Delta_S(\mathcal{R}_n) \leq \epsilon$. This is because for any rectangle $R \in \mathcal{R}_n$ there are rectangles $R^-, R^+ \in \mathcal{R}_{n,4n/\epsilon}$ such that $R^- \subseteq R \subseteq R^+$ and such that $\mathrm{vol}(R^+) - \mathrm{vol}(R^-) \leq \epsilon/2$.

5. By similar reasoning to that used in observation 4, $\Delta_S(\mathcal{R}_{n,4k/\epsilon}^k) \leq \epsilon/2$ implies that $\Delta_S(\mathcal{R}_n^k) \leq \epsilon$.

6. $\Delta_S(\mathcal{R}_n^k) \leq \epsilon$ implies that $\Delta_S(\mathcal{R}_n^{[0,\beta)}) \leq \epsilon + \beta^k$. This follows because $\mathrm{vol}(R) \leq \beta^k$ for any rectangle $R \in \mathcal{R}_n^{[0,\beta)}$ which is non-trivial in more than k dimensions. This shows for constant β and for $k = O(\log(1/\epsilon))$ that $\Delta_S(\mathcal{R}_n^k) \leq \epsilon/2$ implies that $\Delta_S(\mathcal{R}_n^{[0,\beta)}) \leq \epsilon$.

1.3 The Constructions

All of the results are stated in terms of constructions of sets with small discrepancy. The constructions of small sample spaces that approximate the independent distribution follow from the observations of the previous section.

1.3.1 Construction based on reducing from boolean to general

Theorem 1 *There is an explicitly constructible finite set $S \subset [0,1)^n$ with $\Delta_S(\mathcal{R}_n^k) \leq \epsilon$ such that $|S|$ is polynomial in $\log(n)$, 2^k and $1/\epsilon$.*

PROOF:

Let $x_1^1, \ldots, x_1^l, \ldots, x_n^1, \ldots, x_n^l$ be n blocks of l boolean-valued random variables each, where l is a positive integer whose value is determined later. For each $i \in \{1, \ldots, n\}$ we let random variable $x_i = .x_i^1 \cdots x_i^l$ be a binary fraction where x_i^j is the j^{th} most significant bit. We show, the event $< x_1, \ldots, x_n > \in R$ occurs with probability within ϵ of vol(R) for every rectangle $R \in \mathcal{R}_n^k$ when these random variables have the properties we develop below.

Without loss of generality, fix a rectangle $R = \prod_{i \in \{1,\ldots,n\}} [a_i, b_i) \in \mathcal{R}_n^k$ such that the first k dimensions are the non-trivial ones, i.e. for all $i = k+1, \ldots, n$, $[a_i, b_i) = [0, 1)$. For simplicity of presentation, for all $i = 1, \ldots, k$, we restrict the i^{th} interval to be of the special form $[0, b_i)$. (The analysis for the case when the interval is of the general form $[a_i, b_i)$ is no more difficult technically, just not as clean.) To determine if the event $< x_1, \ldots, x_n > \in R$ occurs, it is enough to determine if the k subevents $x_1 \in [0, b_1), \ldots, x_k \in [0, b_k)$ all occur simultaneously.

We think of determining the outcomes of the k subevents starting with subevent $x_1 \in [0, b_1)$ and ending with $x_k \in [0, b_k)$. Let b_1^j be the j^{th} bit in the binary expansion of b_1. We compare the bits $x_1^1, \ldots, x_1^l$ with $b_1^1, \ldots, b_1^l$ one at a time, starting with the most significant bit and working down, stopping as soon as $x_1 \in [0, b_1)$ or $x_1 \notin [0, b_1)$ has been determined. Note that if $x_1^1 \neq b_1^1$ then the outcome of the first subevent is determined one way or the other, i.e. if $x_1^1 = 0$ and $b_1^1 = 1$ then $x_1 \in [0, b_1)$, whereas if $x_1^1 = 1$ and $b_1^1 = 0$ then $x_1 \notin [0, b_1)$. In this case, we move on to determine the outcome of the second subevent. On the other hand, if $x_1^1 = b_1^1$ then the outcome of the first subevent hasn't been determined and we next compare x_1^2 with b_1^2, etc.

Determining the outcomes of the k subevents can be viewed as a complete binary tree labeled with the boolean-valued random variables. The root of the tree is labeled with x_1^1, the left edge out of the root corresponds to $x_1^1 = b_1^1$ and the right edge corresponds to $x_1^1 \neq b_1^1$. At each subsequent node of the tree, the node is labeled with the boolean-valued random variable that is considered next; e.g. the left child of the root is labeled x_1^2 and the right child label is x_2^1.

Suppose for now that $x_1^1, \ldots, x_1^l, \ldots, x_n^1, \ldots, x_n^l$ are independently and uniformly distributed. A random setting of the variables defines a random path down the tree, and it is easy to see that if a random path is taken down this tree (and l is infinite) then the probability that the k subevents all simultaneously occur is exactly vol(R). Furthermore, on average the values of two boolean-valued random variables are examined to determine the outcome of each subevent, and thus on average we examine $2k$ boolean-valued random variables to determine the outcomes of all k subevents. Consider the probability that the outcomes of all k subevents are not determined by the time the first k' boolean-valued random variables are examined. This probability is exactly the same as the probability that there are less than k "heads" in k' tosses of a fair coin. By a standard analysis, when k' is set to a value that is $O(k + \log(1/\epsilon))$ it can be easily shown that this probability is at most $\epsilon/2$. This shows the probability a random path down the tree to depth k' doesn't determine the outcomes of all k subevents is at most $\epsilon/2$. Consequently if $x_1^1, \ldots, x_1^l, \ldots, x_n^1, \ldots, x_n^l$ are k'-wise independent and uniformly distributed then the probability that all k subevents occur simultaneously is within $\epsilon/2$ of vol(R).

There are only $2^{k'}$ paths down to depth k' in the tree. Thus, if, for every path down to depth k' in the tree, the actual probability of the path is within $\epsilon/2^{k'+1}$ of $1/2^{k'}$ then the analysis shows that the probability all k subevents occur simultaneously is within ϵ of vol(R). From this it follows that any distribution on $x_1^1, \ldots, x_1^l, \ldots, x_n^1, \ldots, x_n^l$ that is a (k', ϵ')-approximation for $\mathcal{U}_{nl,2}$ (with $k' = O(k + \log(1/\epsilon))$ and $\epsilon' = \epsilon/2^{k'+1}$ and $l = k'$) has the property that the probability that all k subevents occur simultaneously is within ϵ of vol(R) for all $R \in \mathcal{R}_n^k$. For these values of l, k' and ϵ' we can use [15, Naor Naor] or [2, Alon Goldreich Håstad Peralta] to construct a sample space S which induces a distribution on $x_1^1, \ldots, x_1^l, \ldots, x_n^1, \ldots, x_n^l$ that is a (k', ϵ')-approximation for $\mathcal{U}_{nl,2}$

with $|S|$ polynomial in $\log(n)$, 2^k and $1/\epsilon$.

It should be noted that this analysis uses components of analysis for "Discrete Distribution Generating tree" described in [10, Knuth Yao] and also component of an analysis presented in [13, Luby Veličković].

Setting $k = O(\log(1/\epsilon))$ and using observation 6 from section 1.2 shows that for any constant $\beta < 1$ there is a set S of size polynomial in $1/\epsilon$ and $\log(n)$ with $\Delta_S(\mathcal{R}_n^{[0,\beta)}) \leq \epsilon$.

Alternative construction for special case

For the special case when p is a small prime (e.g., $p = 3$) there is a smaller sample space that is a (k, ϵ)-approximation for $\mathcal{U}_{n,p}$; the construction is a generalization of the construction for $\mathcal{U}_{n,2}$, and can be found in [2, Alon Goldreich Håstad Peralta], [3, Azar Motwani Naor] and [8, Even].

1.3.2 Construction based on the inclusion-exclusion formula

Let $k = O(\log(1/\epsilon))$ and let S be a sample space that is a k-wise independent approximation for $\mathcal{U}_{n,4n/\epsilon}$. In this section, we show this implies $\Delta_S(\mathcal{R}_n) \leq \epsilon$ when points in S are mapped in the natural way to $[0,1)^n$. Using standard constructions (see for example [12, Luby] or [1, Alon Babai Itai]), there is a constructible set S with these properties of size polynomial in $(n/\epsilon)^{\log(1/\epsilon)}$.

The first easy observation is that $\Delta_S(\mathcal{R}^k_{n,4n/\epsilon}) = 0$ when points in S are mapped in the natural way to $[0,1)^n$. Then, we use the theorem described below to show that $\Delta_S(\mathcal{R}^k_{n,4n/\epsilon}) = 0$ and $k = O(\log(1/\epsilon))$ implies that $\Delta_S(\mathcal{R}_{n,4n/\epsilon}) \leq \epsilon/2$. Finally, observation 4 from section 1.2 shows that $\Delta_S(\mathcal{R}_n) \leq \epsilon$.

Theorem 2 *Let $\mathcal{P}_{n,2}$ be a general probability matrix for n boolean-valued random variables $x_1, \ldots, x_n$. Then,*

$$|\mathrm{P}_D[\bigwedge_{i\in\{1,\ldots,n\}} x_i = 0] - \prod_{i\in\{1,\ldots,n\}} p_{i,0}| \leq 2^{-\Omega(k)}$$

for any probability space D that induces a distribution on $x_1, \ldots, x_n$ that is a k-wise independent approximation for $\mathcal{P}_{n,2}$. (Note that if $x_1, \ldots, x_n$ are independently distributed for $\mathcal{P}_{n,2}$, then the event $\bigwedge_{i\in\{1,\ldots,n\}} x_i = 0$ has probability exactly $\prod_{i\in\{1,\ldots,n\}} p_{i,0}$.)

PROOF: The idea is to use the inclusion-exclusion formula. Fix D to be any space that induces a distribution on $x_1, \ldots, x_n$ that is a k-wise independent approximation for $\mathcal{P}_{n,2}$. Define $T_0 = 1$ and for all $j = 1, \ldots, k$ define

$$T_j = \sum_{I=<i_1,\ldots,i_j>} \prod_{l=1,\ldots,j} p_{i_j,1},$$

i.e. T_j is the j^{th} term of the inclusion-exclusion formula. Then, for all even values of j, $\sum_{l=0,\ldots,j}(-1)^j T_j$ is an upper bound on $\mathrm{P}_D[\bigwedge_{i\in\{1,\ldots,n\}} x_i = 0]$, this quantity is a lower bound for all odd values of j and T_k is an upper bound on $|\mathrm{P}_D[\bigwedge_{i\in\{1,\ldots,n\}} x_i = 0] - \prod_{i\in\{1,\ldots,n\}} p_{i,0}|$.

Define $\alpha = \sum_{i\in\{1,\ldots,n\}} p_{i,1}$. There are two cases to the proof, depending on whether $\alpha \leq \frac{k}{2e}$ or $\alpha > \frac{k}{2e}$. (Where $e = 2.718\ldots$.) Suppose that $\alpha \leq \frac{k}{2e}$. We show this implies $T_k \leq 2^{-k}$, which finishes the proof for the first case. This inequality holds because, subject to the restriction that $\sum_{i\in\{1,\ldots,n\}} p_{i,1} = \alpha$, T_k is maximized when, for all $i \in \{1,\ldots,n\}$, $p_{i,1} = \alpha/n$. Thus, $T_k \approx (\frac{\alpha}{ek})^k \leq 2^{-k}$. Now suppose that $\alpha > \frac{k}{2e}$. Consider the first $n' < n$ random variables such that $\frac{k}{2e} - 1 < \sum_{i\in\{1,\ldots,n'\}} p_{i,1} \leq \frac{k}{2e}$ and let $\alpha' = \sum_{i\in\{1,\ldots,n'\}} p_{i,1} \approx \frac{k}{2e}$. We first show this implies $\prod_{i\in\{1,\ldots,n'\}} p_{i,0} \leq 2^{-\Omega(k)}$ and then we show how to finish the proof from this for the second case. Subject to the restriction that $\sum_{i\in\{1,\ldots,n\}} p_{i,1} = \alpha'$, $\prod_{i\in\{1,\ldots,n'\}} p_{i,0}$ is maximized when, for all $i \in \{1,\ldots,n'\}$,

$p_{i,1} = \alpha'/n'$. Thus, because $\alpha' \approx \frac{k}{2e}$, $\prod_{i\in\{1,\ldots,n'\}} p_{i,0} \leq (1-\alpha'/n')^{n'} = 2^{-\Omega(k)}$. From the same proof as used in the first case, noting that $\alpha' \leq \frac{k}{2e}$, $\mathrm{P}_D[\bigwedge_{i\in\{1,\ldots,n'\}} x_i = 0] \leq \prod_{i\in\{1,\ldots,n'\}} p_{i,0} + 2^{-k}$. Because $\prod_{i\in\{1,\ldots,n'\}} p_{i,0} \leq 2^{-\Omega(k)}$ and because $\mathrm{P}_D[\bigwedge_{i\in\{1,\ldots,n\}} x_i = 0] \leq \mathrm{P}_D[\bigwedge_{i\in\{1,\ldots,n'\}} x_i = 0]$, this implies that $0 \leq \mathrm{P}_D[\bigwedge_{i\in\{1,\ldots,n\}} x_i = 0] \leq 2^{-\Omega(k)}$. This and $0 \leq \prod_{i\in\{1,\ldots,n\}} p_{i,0} \leq \prod_{i\in\{1,\ldots,n'\}} p_{i,0}$ finishes the proof of the second case.

The obvious corollary to this theorem we use to prove the result stated at the beginning of this section is that $\Delta_S(\mathcal{R}^k_{n,m}) = 0$ implies that $\Delta_S(\mathcal{R}_{n,m}) \leq 2^{-\Omega(k)}$.

This theorem should be contrasted with the main theorem of [11, Linial Nisan]. Loosely stated, the above theorem says that if the leading $O(\log(1/\epsilon))$ terms of the inclusion-exclusion formula are exactly the same as they are for the independent distribution then the probability of the union of n events is completely determined to within an error ϵ. Loosely stated, one direction of the main theorem in [11, Linial Nisan] says that an arbitrary specification of less than $\sqrt{n}$ of the leading terms doesn't even determine the probability of the union of the n events to within a constant amount.

One application of the result is to deterministic approximation of the number of satisfying truth assignments to a disjunctive normal form boolean formula [13, Luby Veličković]. A more philosophical application is that the result says that the probability of unions of events that are somewhat independent and the probability of unions of events that are totally independent are not very different. This gives some partial justification for modeling "real world" events, which are somewhat independent but not totally so, by events that are totally independent, without drastically affecting the probability of their union.

1.3.3 Construction based on hashing

The third construction uses the results given in section 5 of [16, Nisan] and observation 4 of section 1.2. The result is that there is an efficiently constructible set S of size polynomial in $(n/\epsilon)^{\log(n)}$ such that $\Delta_S(\mathcal{R}_n) \leq \epsilon$.

1.4 Open Problems

One open problem motivated by this work can be found at the end of subsection 1.1.4. An even harder problem, which was motivation for this work, is the following generalization of that problem: Find an efficiently constructible set S of size polynomial in n, m and $1/\epsilon$ such that for any union of at most m rectangles in n dimensional space, the fraction of points in S that fall in their union is within ϵ of the volume of their union. A positive solution to this problem would provide an efficient deterministic approximation algorithm for the DNF counting problem.

1.5 Acknowledgments

We thank Emo Welzl for discussions which led to the understanding of the connections between approximations of distributions and discrepancy theory. We thank Nati Linial and Avi Wigderson for a number of helpful technical discussions. We thank Josef Beck for sharing with us his enthusiasm for this work and his knowledge about discrepancies.

Chapter 2

Eariler Version: Approximation per se

Abstract

Recently, the problem of contructing small sample spaces, inducing k-wise independent and almost k-wise independent random variables, has received considerable attention. However, the positive results obtained so far refer to the special case of *identically distributed* random variables each *uniformly* distributed over the same finite set (typically $\{0,1\}$).

In this paper, we deal with the general problem: given a specification of n independent distributions, we show how to construct a small sample space defining a sequence of n random variables such that the joint distribution of every k variables is statistically close to the corresponding joint distribution specified.

Our construction reduces the general problem of (k-wise) approximating an arbitrary product distribution to the extensively studied special case of (k-wise) approximating uniform distribution over $\{0,1\}^n$.

2.1 Introduction

In recent years, much research effort has been invested in constructing small sample spaces for k-wise independent and almost k-wise independent random variables (cf. [7, 1, 15, 2, 3]). The motivation for this line of research has been the belief that *limited stochastic independence* suffices for the analysis of many interesting randomized algorithm and hence constructing *small probability spaces implementing limited independence* yields a way to "derandomize" these algorithms (i.e., convert them to deterministic algorithms of reasonable complexity) [1]. A typical example of the use of this methodology has been provided by Luby in his work on the maximal indepedent set problem [12]. Surprisingly, it is often ignored that the random variables used in that work are *neither identically distributed nor uniformly distributed* over some sets, and furthermore that this is likely to be the case in many applications. In contrast, all constructions (for limited independence), presented so far, apply to random variables *uniformly distributed over the same set* (in most cases the two-element set $\{0,1\}$). Hence, it is of primary importance to investigate the extent to which these constructions can be generalized to deal with the "k-wise approximation" of arbitrarly stochastically indepedent events.

[1] By enumerating determinstically all elements in the sample space, and running the algorithm using each of them as the "outcome of the internal coin tosses" of the algorithm.

2.1.1 What are k-wise ϵ-approximations

Throughout the paper we consider the approximation of product distributions; namely, distributions which are the product of many (say n) independent distributions. In other words, we consider random variables of the form $X = X_1 \cdots X_n$, where the X_i's are indepedent random variables. These X_i's are **not** necessarily *identically* distributed or *uniformly* distributed over some finite sets. The support of the product variable X, provided none of the X_i is trivial, has cardinality exponential in n. Our aim is to approximate such an X by a random variable $Y = Y_1 \cdots Y_n$ which has much smaller support. It follows that the Y_i's cannot be indepedent of each other. Hence, our aim is to approximate product distributions by distributions of smaller support (which necessarily are not product distributions themselves).

By a k-wise ϵ-approximation of a product (random variable) $X = X_1 \cdots X_n$, we mean a random variable $Y = Y_1 \cdots Y_n$ (where the Y_i are not necessarily independent) so that every k-subproduct of the X_i's is "approximated with error ϵ" by joint distribution of the corresponding Y_i's. In the sequel, the phrase "approximated with error ϵ" means that the variation distance[2], between the resulting random variables, is bounded by ϵ.

2.1.2 Known and new results about k-wise ϵ-approximations

All previous works deal with the approximation of *identical* random variables which are *uniformly* distributed over a finite set. In particaulr, Naor and Naor [15] presented an efficient k-wise ϵ-approximation of identical random variables each uniformly distributed over $\{0,1\}$. The support of their approximation has cardinality $O((k \log n) \cdot 2^{2k} \cdot \frac{1}{\epsilon^4})$. Simpler constructions using a support of size $O((k \log n)^2 \cdot 2^k \cdot \frac{1}{\epsilon^2})$, were presented in [2]. These constructions can be easily extended to approximate identical random variables, each uniformly distributed over a finite field (cf. [2, 3, 8]). The size of the support, for a finite field of cardinality q, is $O((k \log n)^2 \cdot q^k \cdot \frac{1}{\epsilon^2})$. Hence, the support size is polynomial in $\frac{q^k}{\epsilon}$.

The obvious way to get k-wise ϵ-approximation of arbitrary (n-fold product) distributions from the above is to use a $k \cdot \log_2(2/\epsilon)$-wise $\epsilon/2$-approximation of product of $n \cdot \log_2(2/\epsilon)$ independent and uniformly distributed 0-1 random variables[3]. This yields a support size which is polynomial in $(\frac{2}{\epsilon})^k$.

In this paper we present a simple method for constructing k-wise ϵ-approximations, of arbitrary (n-fold product) distributions, using a much smaller support. Let s be a bound on the number of elements in the support of a single distribution in the n-fold product distribution, and suppose $\epsilon < 1/s$ (which is the natural case). Then, loosely speaking, the k-wise ϵ-approximation presented in this paper has support of size polynomial in $\frac{s^k}{\epsilon}$. Hence, whenever $s \ll \frac{1}{\epsilon}$, our improvement is meaningful. Let us consider two typical examples:

Example 1 *Suppose we wish to approximate a product of n indepedently distributed 0-1 random variables, each assigned 1 with probability $\frac{1}{2}+\epsilon$ and 0 otherwise.* Using previously known techniques, a k-wise $\frac{\epsilon}{2}$-approximation of this n-fold random variable would have required using a sample space

[2]The variation distance between the random variables U and V is

$$\sum_{\alpha} |\mathrm{P}(U=\alpha) - \mathrm{P}(V=\alpha)|$$

.

[3]These $n \cdot \log_2(2/\epsilon)$ 0-1 variables are partitioned into blocks of length $\log_2(2/\epsilon)$, each encoding elements of the corresponding set in the obvious manner.

of size $\epsilon^{\Theta(k)}$, whereas an n-fold of uniformly and indentically distributed 0-1 random variables could be k-wise $\frac{\epsilon}{2}$-approximated using a sample space of size $\mathrm{poly}(\frac{1}{\epsilon}, 2^k, \log n)$. Using our results the first (i.e., "non-uniform") n-fold can be k-wise ϵ-approximated at essentially the same "cost" as the "uniform" one (i.e., using a sample space of size $\mathrm{poly}(\frac{1}{\epsilon}, 2^k, \log n)$).

Example 2 *Suppose we wish to approximate an n-fold product of indepedently distributed random variables, where the i^{th} random variable is uniformly distributed over the set $\{1, 2, ..., s_i\}$. Let $s \stackrel{\mathrm{def}}{=} \max_i\{s_i\}$ and $L \stackrel{\mathrm{def}}{=} \mathrm{lcm}\{s_i : i \leq n\}$.* Using previously known techniques a k-wise ϵ-approximation of this n-fold random variable would have required using a sample space of size $\min\{\frac{L^{\Theta(k)}}{\epsilon}, \frac{1}{\epsilon^{\Theta(k)}}\}$, whereas the n-fold consisting of indepedent random variables each uniformly distributed over $\{1, 2, ..., s\}$ could be k-wise ϵ-approximated using a sample space of size $\mathrm{poly}(\frac{1}{\epsilon}, s^k, \log n)$. Using our results the first n-fold (in which variables are not identical) can be k-wise ϵ-approximated at essentially the same "cost" as the indentical case (i.e., using a sample space of size $\mathrm{poly}(\frac{1}{\epsilon}, s^k, \log n)$).

2.1.3 An overview of our construction

Our construction is quite simple and is described below. For simplicity, we consider here the special case of approximating n-folds of 0-1 distributions. Namely, each random variable X_i satisfies $X_i \in \{0, 1\}$. Let $p_i = \mathrm{P}(X_i = 0)$. To construct a k-wise ϵ-approximation of $X = X_1 \cdots X_n$, we use a $O(k + \log(2/\epsilon))$-wise $\mathrm{poly}(\epsilon/2^k)$-approximation of the uniform distribution over $\{0, 1\}^{nl}$, where $l \stackrel{\mathrm{def}}{=} O(\log(2/\epsilon)$. The approximation to X, denoted $Y = Y_1 \cdots Y_n$, is determined by letting $Y_i = 0$ if the $B_i < p_i \cdot 2^l$, where B_i is the integer encoded in the i^{th} (l-bit long) block of the nl-bit long sample string. The crucial point is that we are using a $O(k + \log(2/\epsilon))$-wise $\mathrm{poly}(\epsilon/2^k)$-approximation of the uniform distribution over $\{0, 1\}^{nl}$, rather than using a $(k \cdot \log(2/\epsilon))$-wise $(\epsilon/2)$-approximation of it. This requires a more careful analysis.

The analysis of the approximation Y uses in an essential way the fact that each Y_i is determined by specific *fixed* locations in the binary string produced by the approximation to the uniform distribution.

We end the introduction by presenting an *alternative construction of unknown quality.* The problem of constructing k-wise approximations to arbitrary product distributions, is remenisence of the classic problem of generating arbitrary probability distributions by using a uniform probability distribution over binary strings (or in other words by using an unbiased coin). In particular, Knuth and Yao have extensively analyzed the expected number of coin tosses required in such schemes [10]. A natural suggestion is to use one of these schemes (termed "Discrete Distribution Generating tree") to produce an a k-wise approximation to the n-fold distribution by using as input a $O(k)$-wise approximation to the uniform binary distribution. We do not know whether this alternative approach works and our conjecture is that it does not.

2.2 Formal Setting

2.2.1 Preliminaries

Convention: Throughout the rest of this paper we consider only random variables ranging over finite sets. Without loss of generality, each finite set, say of cardinality s, is associated with the set of the first s non-negative integers.

We recall two standard definitions. The first definition will be used in the definition of approximation, whereas the second definition is given merely for methodological purposes.

Definition 1 (distance between distributions): *Let X and Y be two random variables ranging over some finite set S.*

- (max-norm): *The distance in max-norm (L_∞ norm) between X and Y is defined as* $\max_{e \in S} |\mathrm{P}(X = e) - \mathrm{P}(Y = e)|$.
- (variation distance): *The variation distance (L_1* norm distance*) between X and Y is defined as* $\sum_{e \in S} |\mathrm{P}(X = e) - \mathrm{P}(Y = e)|$.

Definition 2 (k-wise independence): *A sequence of random variables $Z = Z_1, ..., Z_n$ is* k-wise independent *if for any k positions $i_1 < i_2 < \cdots < i_k$, the random variables $Z_{i_1}, Z_{i_2}, \ldots, Z_{i_k}$ are totally independent. Namely, for every k-long sequence of integers, $\alpha = \alpha_1, \alpha_2, ..., \alpha_k$, we have*

$$Pr[Z_{i_1} Z_{i_2} \cdots Z_{i_k} = \alpha] = Pr[Z_{i_1} = \alpha_1] \cdot Pr[Z_{i_2} = \alpha_2] \cdots Pr[Z_{i_k} = \alpha_k]$$

2.2.2 Definition of k-wise approximation

The following definition is central to the current paper.

Definition 3 (k-wise approximation): *Let $X = X_1 \cdots X_n$ be a product of independent random variables, and $Y = Y_1 \cdots Y_n$ be an arbitrary sequence of (not necessarily independent) random variables.*

- (max-norm approximation): *We say that Y is a* k-wise ϵ-approximation of X in max-norm *if for any $l \le k$ positions, $i_1 < i_2 < \cdots < i_l$, the max-norm distance between $X_{i_1} X_{i_2} \cdots X_{i_l}$ and $Y_{i_1} Y_{i_2} \cdots Y_{i_l}$ is bounded above by ϵ.*
- (L_1 approximation): *We say that Y is a* k-wise ϵ-approximation of X in norm L_1 *if for any k positions, $i_1 < i_2 < \cdots < i_k$, the L_1-norm distance between $X_{i_1} X_{i_2} \cdots X_{i_k}$ and $Y_{i_1} Y_{i_2} \cdots Y_{i_k}$ is bounded above by ϵ.*

The above two measures of approximation seem to be most useful in applications. Other notions of approximation are discussed in [5].

When constructing a k-wise approximation to a product variable X, we get as input a "specification" of X. A *specification of X* is an n-by-s matrix, $P = \{p_{i,j}\}$, satisfying $\mathrm{P}(X_i = a) = p_{i,a}$ (for every pair i, a) and $\sum_{a=0}^{s-1} p_{i,a} = 1$ (for every i). (We stress that we consider k-wise approximations only to products of independent random variables.)

Definition 4 (k-wise approximators): *Let A be an algorithm that on input a specification of a product variable* $\mathrm{spec}(X)$*, an integer k a rational ϵ, and integers $i \le n$, and j, outputs an element (in the support of X_i). Algorithm A is called a* k-wise L_∞ (resp. L_1) ϵ-approximator of X *if, for $M = A(\mathrm{spec}(X), k, \epsilon)$, the random variable $Y = Y_1, ..., Y_n$ defined by selecting j uniformly in $\{1, ..., M\}$ and setting $Y_i = A(\mathrm{spec}(X), k, \epsilon, i, j)$, for each $i \le n$, constitues a k-wise ϵ-approximation of X in max norm (resp. L_1 norm). ($M = A(\mathrm{spec}(X), k, \epsilon)$ is called the* size of A's sample space.*) Algorithm A is called a* product approximator *if for every X, k, ϵ as above, A constitutes a k-wise ϵ-approxmator of X.*

2.2.3 Our Result

Theorem 1 (efficient product approximator): *There exists a polynomial-time product approximator in L_1 norm (resp., L_∞ norm) satisfying, for every X, k, ϵ, the size of A's sample space is bounded above by*

$$(O(s))^k \cdot (\frac{1}{\epsilon})^{10} \cdot F^2$$

$$(\text{resp., } 2^{16k}(\frac{1}{\epsilon})^{10} \cdot F^2 \text{ })$$

where s is a bound on the support of individual X_i, and $F \stackrel{\text{def}}{=} O((k+\log(1/\epsilon))\cdot(\log n+\log\log(k/\epsilon)))$. In fact, the algorithm can be implemented in NC.

A better result is possible for the special case of $s = 2$.

2.3 Constructions

We first present our construction for the special (yet interesting) case of approximating Boolean-valued random variables. We later generalize the construction to handle random variables ranging over arbitrary sets.

2.3.1 Special Case: Boolean-valued random variables

Assume we are give a specification of a random variable $X = X_1, ..., X_n$, consisting of a sequence of n independent Boolean random variables. Clearly, it suffices to specify the probability that each of these variables is assigned 0. Let $p_i \stackrel{\text{def}}{=} \text{P}(X_i = 0)$, for every $i \leq n$, and denote by $p_i(1), p_i(2), ...$ the bits in the binary expansion of p_i (i.e., $p_i = \sum_{j\geq 1} p_i(j)\cdot 2^{-j}$). We construct a k-wise ϵ-approximation of X as follows.

Let l and t be integers to be determined latter ($l = 2 + \log_2(k/\epsilon)$ and $t = 5(k + \log(1/\epsilon))$ will do). In our construction we use an arbitrary t-wise $(\epsilon/2^{t+1})$-approximation in max-norm of the uniform distribution over $\{0,1\}^{ln}$. Let us denote the 0-1 random variable in this approximation by $Z_1(1), ..., Z_1(l), ..., Z_n(1), ..., Z_n(l)$.

Construction 1 *Let $Z_1(1), ..., Z_1(l), ..., Z_n(1), ..., Z_n(l)$ be a t-wise $(\epsilon/2^{t+1})$-approximation in L_∞ norm of the uniform distribution over $\{0,1\}^{ln}$. For every i, if the string $Z_i(1)\cdots Z_i(l)$ is smaller than (in lexicographic order) the string $p_i(1)\cdots p_i(l)$ then set $Y_i = 0$ otherwise set $Y_i = 1$.*

Our analysis of the above construction is somewhat analogous to the proof of Theorem 3 in [13]. We fix k variables in Y, without loss of generality $Y_1, ..., Y_k$, and a k-bit string, α, and evaluate the difference $\text{P}(X_1 \cdots X_k = \alpha) - \text{P}(Y_1 \cdots Y_k = \alpha)$.

Consider a mental experiment in which the Y_i's are determined by a random walk on an infinite labelled binary tree as follows. The edges in the tree are labelled by $\{0,1\}$, so that each node has one 0-child and one 1-child. Pictorially, one may visualize the 0-child as the left-child and the 1-child as the right-child. The nodes in the tree are labelled by pairs of the form (i, σ), where $i \in \{1, .., k\}$ and $\sigma \in \{0, 1, *\}$. The root is labelled $(1, *)$. There is a unique path going down from the root with all nodes on it labelled $(1, *)$. This is the path corresponding to the binary expansion of p_1. All the nodes reached by following this path upto some node and then leaving it to the left (assuming the path continues to the right) are labelled $(1, 0)$. Intuitively, reaching such a node via a random walk down the tree results in setting Y_1 to 0. Likewise, the nodes reached by following the "p_1-expansion path" upto some node and then leaving it to the right are labelled $(1,1)$. From each node labelled (i,σ), $i<k$ and $\sigma\in\{0,1\}$, there is a unique path going down labelled $(i+1,*)$. This is the path corresponding to the binary expansion of p_{i+1}. The nodes reached from a node labelled (i,σ) by following the "p_{i+1}-expansion path" upto some node and then leaving it to the left are labelled $(i+1,0)$, and reaching them via a random walk results in setting Y_{i+1} to 0. Finaly, both children of nodes labelled (k,σ), with $\sigma\in\{0,1\}$, are labelled (k,σ) too. These nodes are called *complete*. Intuitively, reaching them via a random path from the root means that all Y_i's were given values.

The following claims are easily verified

Claim 1 *Consider a random infinite path going down the tree and set $Y_i=\sigma_i$ if and only if the path goes through a node labelled (i,σ_i). Then $X_1\cdots X_k$ and $Y_1\cdots Y_k$ are identically distrubted.*

Claim 2 *The number of nodes at level t which are not complete is $\sum_{i=0}^{k-1}\binom{t}{i}\ll 2^{\frac{3}{4}t+k}$.*

Claim 3 *Consider a random path of length t going down the tree and set $Y_i=\sigma_i$ if the path goes through a node labelled (i,σ_i). In case the path does not go through any node labelled (i,σ) (with $\sigma\in\{0,1\}$), set Y_i arbitrarily. Then the variation distance between $X_1\cdots X_k$ and $Y_1\cdots Y_k$ is bounded by $2^{-(\frac{t}{4}-k)}$.*

Clearly, the mental experiment described in Claim 3 corresponds to the setting of the Y_i in Construction 1, provided that the $Z_i(j)$'s are t-wise independent and that $l\geq k$. Waiving these requirements (namely, allowing the $Z_i(j)$'s to constitute a t-wise $(\epsilon/2^{t+1})$-approximation in max-norm of the uniform 0-1 distribution and l be arbitrary) adds error terms bounded by $\frac{\epsilon}{2}$ and $k\cdot 2^{-l}$, respectively. Hence, we get

Proposition 1 *Let $t=4(k+\log_2(4/\epsilon))$ and $l=\log_2(4k/\epsilon)$. Then Y_i's presented in Construction 1 constitute a k-wise ϵ-approximation (in L_1 norm) of the n-fold X.*

2.3.2 The General Case

The construction for the general case extends Construction 1 in the obvious manner. Let $X=X_1\cdots X_n$ be an n-fold random variable, s a bound on the support of each X_i, and $P=\{p_{i,j}:1\leq i\leq n, 0\leq j\leq s-1\}$ be a specification of the n-fold X. For every i,j, let $q_{i,j}\stackrel{\rm def}{=}\sum_{h=1}^{j}p_{i,h}$. Denote by $q_{i,j}(1),q_{i,j}(2),\ldots$ the bits in the binary expansion of $q_{i,j}$. Let l and t be integers to be determined latter ($l=2+\log_2(k/\epsilon)$ and $t=9(k+\log(1/\epsilon))$ will do).

Construction 2 *Let $Z_1(1),\ldots,Z_1(l),\ldots,Z_n(1),\ldots,Z_n(l)$ be a t-wise $(\epsilon/2^{t+1})$-approximation in L_∞ norm of the uniform distribution over $\{0,1\}^{ln}$. For every i, if the string $Z_i(1)\cdots Z_i(l)$ is between (in lexicographic order) the string $q_{i,j}(1)\cdots q_{i,j}(l)$ and the string $q_{i,j+1}(1)\cdots q_{i,j+1}(l)$ then set $Y_i=j$.*

Extending the argument used in the previous subsection we can easily evaluate the quality of Construction 2 as a max-norm approximator. The statement and proof of Claim 2 are slightly changed: the number of nodes at level t which are not complete is now bounded by $\sum_{i=0}^{k-1}\binom{t}{i}2^i\ll 2^{\frac{3}{4}t+2k}$. We get

Proposition 2 *Let $t=4\cdot(2k+\log_2(4/\epsilon))$ and $l=3+\log_2(k/\epsilon)$. Then Y_i's presented in Construction 2 constitute a k-wise ϵ-approximation in L_∞ norm of the n-fold X.*

Approximation in L_1 norm follows immediately by bounding the L_1 approximation error by s^k times the L_∞ approximation error. Using the known results on t-wise approximation of the uniform distribution over $\{0,1\}^{ln}$, Theorem 1 follows[4].

2.3.3 Alternative construction for special case

In the special case where the specification of the n-fold variable X can be expressed by a matrix in which all entries are rationals of the form $\frac{i}{p}$, for some small prime p (e.g., $p = 3$), much better k-wise approximation schemes can be constructed. In this case, a k-wise ϵ-approximation of X is constructed using a k-wise ϵ-approximation of the uniform distribution over $GF(p)^n$, in the obvious manner[5].

The construction can be extended to the case that where the specification of the n-fold variable X can be well approximated by a matrix in which all entries are rationals of the form $\frac{i}{p}$, for some small prime p (e.g., $p = 3$). By well approximation we mean that the absolute difference between an entry in the specification matrix of X and the corresponding entry in the approximation matrix should not exceed the approximation error in the desired construction (i.e., the parameter ϵ). Hence, this approach is applicable only if the specification matrix has good approximation by a rational matrix with relatively small common denominator.

[4] Recall that we need a t-wise $2^{-(t+1)}\epsilon$-approximation in max-norm of the uniform distribution over $\{0,1\}^{ln}$. By results of [2], such approximations can be efficiently constructed having sample space of size $(\frac{t \cdot \log_2(ln)}{2^{-t-1}\epsilon})^2 = (2t\log_2(ln))^2 \cdot (\frac{2^t}{\epsilon})^2$. Substituting the values of t and l, we get $F^2 \cdot (\frac{2^{16k}}{\epsilon^{10}})$, where $F = O((k + \log(1/\epsilon)) \cdot (\log_2 n + \log\log(k/\epsilon)))$.

[5] Recall that k-wise ϵ-approximation (in max norm) of the uniform distribution over $GF(p)^n$ can be constructed using support of the same cardinality as in the construction of such approximations for the uniform binary distribution [2, 3, 8].

Chapter 3

Some Probablity Spaces over $GF(p)$

In the first section we define what we mean by a small-bias probability space over $GF(p)^n$, for prime $p \geq 2$, and provide a construction. Our definition generalizes the one commonly used for $p = 2$, and our constructions of ϵ-biased spaces over $GF(p)^n$ maintain the size of known constructions for $GF(2)^n$, independent of p. In the second section we relate such small-bias spaces to more standard notions of approximation, which refer to the pointwise difference between probability spaces.

3.1 Small Bias Spaces over $GF(p)$, for $p > 2$

In this section we present a construction of small biased probability spaces over the prime field $GF(p)$. Such spaces consists of n-long sequences over $GF(p)$, where n and ϵ are parameters so that for every t-long sequence $(c_1, ..., c_t)$ of elements in $GF(p)$, so that not all c_i's are zero, and every $v \in GF(p)$

$$|\mathrm{P}(\sum_{i=1}^{n} c_i r_i = v) - \frac{1}{p}| < \epsilon$$

where the probability is taken uniformly over all possible sequences, $(r_1, ..., r_t)$, in the sample space.

The sample space we construct has size $O(n/\epsilon)^2$. It generalizes the first construction (i.e., the LFSR Construction) of [2] (which was presented there for $p = 2$). We point out that other two constructions are known, generalizing the second and third constructions of [2] (cf., [3] and [2], respectively).

A point in the sample space is specified by two sequences of length $m \stackrel{\text{def}}{=} \log_p(n/\epsilon)$ over $GF(p)$, denoted $f_0 \cdots f_{m-1}$ and $s_0 \cdots s_{m-1}$, where $f_0 = 1$ and $t^m + \sum_{i=0}^{m-1} f_i \cdot t^i$ is an irreducible polynomial. The n-bit sample string, denoted $r_0 \cdots r_{n-1}$ is determined by $r_i = s_i$ for $i < m$ and $r_i = \sum_{j=0}^{m-1} f_j \cdot r_{i-m+j}$ for $i \geq m$.

3.1.1 Formal Setting

The following definition of small-bias sample spaces implies the informal definition presented above. Both definitions are legitimate generalizations of the definition of small-biased sample spaces for the binary case (and indeed they are equivalent for $p = 2$).

Definition 5 *Let n be an integer, p be a prime and ω be a p^{th} root of unity (in the complex field). A set $S \subseteq GF(p)^n$ is said to have ϵ-*bias (sample space for $GF(p)^n$) *if, for every n-long sequence*

$(a_1, ..., a_n)$ *of elements in* $GF(p)$*, so that not all* a_i*'s are zero, the expectation of (the magnitude of)* $\omega^{\sum_{i=1}^{n} a_i r_i}$*, taken over all* $(r_1, ..., r_n) \in S$ *with uniform distribution, is bounded above by* ϵ*. That is,*

$$\left\| \mathrm{E}_{(r_1,...,r_n)\in S} \left(\omega^{\sum_{i=1}^{n} a_i r_i} \right) \right\| \le \epsilon \tag{3.1}$$

Theorem 2 *For every integer* n*, prime* p *and* $\epsilon > 0$*, there exists an efficiently constructible* ϵ*-bias sample space for* $GF(p)^n$ *of size* $(2n/\epsilon)^2$*.*

3.1.2 The Construction

Our construction is based on linear feedback shift register (LFSR) sequences over $GF(p)$. We stress that the arithmetics in the LFSR is that of $GF(p)$.

Definition 6 (linear feedback shift register sequences): *Given two sequences* $\overline{s} = s_0, s_1, \ldots s_{m-1}$ *and* $\overline{f} = f_0, f_1, \ldots f_{m-1}$ *over* $GF(p)$*, the* shift register sequence generated *by the* feedback rule $\overline{f}$ *and the* start sequence $\overline{s}$ *is* $r_0, r_1, \ldots r_{n-1}$ *where* $r_i = s_i$ *for* $i < m$ *and* $r_i = \sum_{j=0}^{m-1} f_j \cdot r_{i-m+j}$ *for* $i \ge m$*.*

Our sample space will consist of all shift register sequences generated by "non-degenerate" feedback rules and any starting sequence. A feedback rule $f_0, f_1, \ldots f_{m-1}$ is called non-degenerate if $f(t) \stackrel{\text{def}}{=} t^m + \sum_{j=0}^{m-1} f_j \cdot t^j$ is an irreducible polynomial over $GF(p)$.

Construction 1 (Sample Space $S_n^m(p)$): *The sample space* $S_n^m(p)$ *is the set of all shift register sequences generated by a non-degenerate feedback rule. Namely,* $S_n^m(p)$ *contains all sequences* $\overline{r} = r_0 r_1 \cdots r_{n-1}$ *such that there exists a non-degenerate feedback rule,* $\overline{f}$*, and a start sequence,* $\overline{s}$*, generating* $\overline{r}$*.*

For the rest of this section we consider polynomials over $GF(p)$. The number of irreducible monic polynomials of degree m is (cf., [14, Chap. 4, Thm. 15])

$$\frac{1}{m} \sum_{d|m} \mu(m/d) \cdot p^d$$

where μ is the ordinary Möbius function (i.e. $\mu(x) = (-1)^s$ where s is the number of primes that divide x if x is squarefree and $\mu(x) = 0$ otherwise). Since also $\mu(1) = 1$ the above expression is $(1 + O(p^{-m/2}))\frac{p^m}{m}$. For the rest of this section we will, for notational simplicity, treat the number of irreducible monic polynomials of degree m as if it is exactly $\frac{p^m}{m}$. (The error introduced is absorbed in the error term.) Hence, with this convention we say that the size of $S_n^m(p)$ is $\frac{p^{2m}}{m}$. Thus, setting $m \stackrel{\text{def}}{=} \log_p(1/\epsilon)$ and proving the proposition below, Theorem 2 follows.

Proposition 3 : *The sample space* $S_n^m(p)$ *is* $\frac{n-1}{p^m}(1 + O(2^{-m/2}))$*-biased.*

Proof: We fix an arbitrary (not all-zero) sequence, $\alpha \stackrel{\text{def}}{=} (a_0, ..., a_{n-1})$, and consider the value of the expression in the l.h.s. of Eq. (3.1). Furthermore, we fix a feedback rule, $\overline{f}$, and consider

$$\left\| \mathrm{E}_{(s_0,...,s_{m-1})\in GF(p)^m} \left(\omega^{\sum_{i=0}^{n-1} a_i r_i} \right) \right\| \le \epsilon \tag{3.2}$$

where $r_0, ..., r_{n-1}$ is the shift register sequence generated by the feedback rule $\overline{f}$ and the start sequence $\overline{s} \stackrel{\text{def}}{=} (s_0, ..., s_{m-1})$.

Towards evaluating Eq. (3.2), we consider the distribution of $(\alpha, r)_p \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} a_i r_i$ (when we only vary the starting vector $\overline{s}$). A key observation is that the r_i's are a linear combination of the s_j's (which are the only indeterminates as the f_i's were fixed). It is useful (and standard practice) to notice that in $GF(p)$, the reduction of t^j modulo $f(t)$ $(= t^m + \sum_{i=0}^{m-1} f_i \cdot t^i)$ is a linear combination of $t^0, t^1, \ldots t^{m-1}$ and that this linear combination is identical to the coefficients in the expression of r_i as a linear combination of the s_j's. Hence, a linear combination of the r_i's (which is exactly what $(\alpha, r)_p$ is) corresponds to a linear combination of the corresponding powers of t^i. This linear combination can be either identically zero or not. The first case means that the polynomial $f(t)$ divides the polynomial $a(t) \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} \alpha_i \cdot t^i$; whereas in the second case $(\alpha, r)_p$ being a non-constant combination of the s_i's is unbiased when the s_i's are uniformly selected. Thus, Eq. (3.2) is 1 if the polynomial $f(t)$ divides the polynomial $a(t)$ and is 0 otherwise (since in that case the expectation equals $\frac{1}{p} \sum_{j=0}^{p-1} \omega^j = 0$).

Thus, the value of Eq. (3.1) equals the probability that the polynomial $f(t)$ divides the polynomial $a(t)$. The latter probability is bounded by the fraction of irreducible monic polynomials of degree m which divide a specific polynomial of degree $n-1$. There are at most $\frac{n-1}{m}$ irreducible monic polynomials of degree m which divide a polynomial of degree $n-1$. Dividing by the number of irreducible monic polynomials of degree m (i.e., $\frac{p^m}{m}$), the proposition follows. ■

3.2 Notions of approximations versus various norms

Notation. One may view probability spaces over a finite Abelian group G as elements in the vector space $\mathcal{H}_G$ consisting of functions that map G to the complex numbers.

An inner product in the vector space $\mathcal{H}_G$ is defined by

$$\langle f, g \rangle = \sum_{x \in G} f(x) \cdot \overline{g(x)}$$

We are interested in two orthonormal bases: the character functions $\{\pi_g\}_{g \in G}$ and Kronecker functions $\{\delta_g\}_{g \in G}$. The Character function for a cyclic group G generated by g is defined by

$$\pi_{g^i}(g^j) = \frac{1}{\sqrt{|G|}} \cdot \omega^{ij}$$

where $\omega = e^{2\pi i/|G|}$ is the complex root of unity of order $|G|$. For $G = GF(p)^n$, the characters are of the form

$$\pi_{f_1, f_2, \ldots, f_n}(x_1, x_2, \ldots, x_n) = p^{-n/2} \omega^{\sum_{i=0}^{n} f_i x_i}$$

where $\omega = e^{2\pi i/p}$. The Kronecker function δ_g (for $g \in G$) is defined by

$$\delta_g(g') = \begin{cases} 1 & \text{if } g = g' \\ 0 & \text{otherwise} \end{cases}$$

Norms over the vector space $\mathcal{H}_G$ are defined by considering the Fourier series. Namely, the series of coefficients in the representation of functions by orthonormal bases. Given an orthonormal base $B = \{b_1, b_2, \ldots, b_{|G|}\}$, $r \geq 1$, and a function $f \in \mathcal{H}_G$, The norm $N_{B,r}(f)$ is defined by

$$N_{B,r}(f) = \left(\sum_{i=1}^{|G|} |\langle f, b_i \rangle|^r \right)^{1/r}$$

When $r = \infty$, $N_{B,\infty}(f)$ equals $\max_{i=1}^{|G|} |\langle f, b_i \rangle|$.

Relations between definitions. Let C denote the orthonormal basis of characters, and let K denote the orthonormal basis of Kronecker functions defined above. Let S denote a probability space over G, and let U denote the uniform probability space over G. The following relations hold:

1. The probability space S is ε-biased iff

$$N_{C,\infty}(S-U) \leq \frac{\varepsilon}{\sqrt{|G|}}$$

This follows from the following observation. Since the group of characters of G is isomorphic to G, we denote by π_g the character that corresponds by the isomorphism to $g \in G$. By definition,

$$N_{C,\infty}(S-U) = \max_{g \in G} |\langle S-U, \pi_g \rangle|$$

Note that $\langle S-U, \pi_I \rangle = 0$, for the identity $I \in G$ (since $\pi_I(g) = 1$, for every $g \in G$). Morover, $\langle U, \pi_g \rangle = 0$, for $g \in G - I$. Therefore,

$$N_{C,\infty}(S-U) = \frac{1}{\sqrt{|G|}} \cdot \max_{g \in G-I} |E_{j\sim S}[\pi_g(j)]|$$

For $G = GF(p)^n$, the latter coincides with Eq. (3.1), divided by $\sqrt{|G|}$.

2. The probability space S is an ε-approximation iff

$$N_{K,\infty}(S-U) \leq \varepsilon$$

This follows from $N_{K,\infty}(S-U) = \max_{g \in G} |S(g) - U(g)|$.

3. The probability space S is an ε-L1-approximation iff

$$N_{K,1}(S-U) \leq 2\varepsilon$$

This follows from $N_{K,1}(S-U) = \sum_{g \in G} |S(g) - U(g)|$.

Relations between approximations. The previous paragraph shows that one may view different types of approximations of the uniform probability space as bounding the norms over the approperiate basis. We now bound the norms when bases are changed.

Claim 3 *Let B_1 and B_2 denote any two orthonormal bases. Then,*

$$N_{B_1,\infty}(f) \leq \sqrt{|G|} \cdot N_{B_2,\infty}(f) \tag{3.3}$$

$$N_{B_1,1}(f) \leq |G| \cdot N_{B_2,\infty}(f) \tag{3.4}$$

Let $B_1 = K$ and $B_2 = C$. Then Eq. (3.3) implies that an ε-biased space is also an ε-approximation. Eq. (3.4) implies that an ε-biased space is also an $\varepsilon \cdot (\sqrt{|G|}/2)$-L1-approximation.

Proof: The first part is proved by the following transitions. The first and third transitions are trivial, and the second trasition follows from Parseval's formula.

$$\begin{aligned} N_{B_1,\infty}(f) &\leq N_{B_1,2}(f) \\ &= N_{B_2,2}(f) \\ &\leq \sqrt{|G|} \cdot N_{B_2,\infty}(f) \end{aligned}$$

For the second part we use the following three equations, justified by Cauchy-Schwartz inequality, Parsevals' formula, and a trivial substitution, respectively.

$$\begin{aligned} N_{B_{1},1}(f) &\leq \sqrt{|G|} \cdot N_{B_{1},2}(f) \\ N_{B_{1},2}(f) &= N_{B_{2},2}(f) \\ N_{B_{2},2}(f) &\leq \sqrt{|G|} \cdot N_{B_{2},\infty}(f) \end{aligned}$$

The second part follows. $\square$

Bibliography

[1] Alon, N., Babai, L., Itai, A., "A Fast and Simple Randomized Parallel Algorithm for the Maximal Independent Set Problem", *Journal of Algorithms*, 7, pp. 567–583, 1986.

[2] Alon, N., Goldreich, O., Håstad, J., Peralta, R., "Simple Constructions of Almost k-wise Independent Random Variables", *Proc. 31st FOCS*, 1990.

[3] Azar, Y., Motwani, R., Naor, J., "An efficient construction of a multiple value small bias probability space", to appear.

[4] Beck, J., Chen, W., "Irregularities of distribution", Cambridge University Press, 1987.

[5] Ben-Natan, R., "On Dependent Random Variables Over Small Sample Spaces", M.Sc. Thesis, Computer Science Dept., Hebrew University, Jerusalem, Israel, Feb. 1990.

[6] B. Chor, J. Freidmann, O. Goldreich, J. Hastad, S. Rudich, and R. Smolensky, "The bit extraction problem and t-resilient functions", *Proc. 26th FOCS*, 1985, pp. 396–407

[7] Chor, B., Goldreich, O., "On the Power of Two–Point Based Sampling," *Jour. of Complexity*, Vol 5, 1989, pp. 96–106.

[8] Even, G., "Construction of Small Probabilistic Spaces for Deterministic Simulation", M. Sc. (in Computer Science) thesis, submitted to the Senate of the Technion (Israel Institute of Technology) in Aug. 1991. (In Hebrew, abstract in English).

[9] Karp, R., Wigderson, A., "A Fast Parallel Algorithm for the Maximal Independent Set Problem", proceedings of 16^{th} ACM Symposium on Theory of Computing, 1984.

[10] Knuth, D., Yao, A., "The complexity of non uniform random number generation", in Algorithms and Complexity, Ed. J. Traub, AC Press, New York, pp. 357-428, 1976.

[11] Linial, N., Nisan, N., "Approximate Inclusion-Exclusion", 22^{nd} *STOC*, 1990.

[12] Luby, M., "A Simple Parallel Algorithm for the Maximal Independent Set Problem," 17^{th} *STOC*, May 6-8 1985, pp. 1-10, *SIAM J. on Computing*, November 1986, Volume 15, No. 4, pp. 1036-1053

[13] Luby, M., Veličković, B., "On Deterministic Approximation of DNF", *Proc. 23rd STOC*, 1991, pp. 430–438.

[14] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North Holland, Amsterdam, 1977.

[15] Naor, J., Naor, M., "Small-bias Probability Spaces: Efficient Constructions and Applications", *22nd STOC*, 1990, pp. 213–223.

[16] Nisan, N., "Pseudo-random Generators for Space-Bounded Computation", 22^{nd} *STOC*, May 14-16 1990, pp. 204-212.

[17] Niederreiter, H., "Constructions of Low-Discrepancy Point Sets and Sequences", get the correct reference here, probably several.