TECHNION - Israel Institute of Technology
Computer Science Department

A SIMPLE CONSTRUCTION OF ALMOST k-Wise INDEPENDENT RANDOM VARIABLES

by

O. Goldreich and J. Hastad

Technical Report #638

June 1990

A Simple Construction of Almost k-wise Independent Random Variables

Oded Goldreich*
Computer Science Dept.
Technion
Haifa, Israel

Johan Håstad
Royal Institute of Technology
Stockholm, Sweden

June 27, 1990

Abstract

We present a simple construction of a small probability space on n bits for which any k bits are almost independent. The number of bits used to specify a point in the sample space is $O(\log\log n + k + \log\frac{1}{\epsilon})$, where ϵ is the statistical difference between of the distribution induced on any k bit locations and the uniform distribution. This is asymptotically comparable to the construction recently presented by Naor and Naor. An additional advantage of our construction is its simplicity. Loosely speaking, the sample space consists of the set of sequences obtained from a linear feedback shift register on various short start and feedback sequences.

*Supported by grant No. 86-00301 from the United States - Israel Binational Science Foundation (BSF), Jerusalem, Israel.

1 Introduction

In recent years, randomization has played a central role in the development of efficient algorithms. Notable examples are the massive use of randomness in computational number theory (e.g., primality testing [16, 17, 10, 1]) and in parallel algorithms (e.g. [12, 14]).

A randomized algorithm can be viewed as a two-stage procedure in which first a "sample point" is chosen at random and next a deterministic procedure is applied to the sample point. In the generic case the sample point is an arbitrary string of specific length (say n), the sample space consists of the set of all 2^n strings, and "choosing a sample at random" amounts to taking the outcome of n consequative unbiased coin tosses. However, as observed by Luby [12], in many cases the algorithm "behaves as well" when the sample is chosen from a much smaller sample space. If points in the smaller sample space can be compactly represented and generated (i.e. reconstructed to their full length from the compact representation) then this yields a saving in the number of coin tosses required for the procedure. In some cases the required number of coin tosses gets so small that one can deterministically scan all possible outcomes (e.g. [12]).

To summerize, the construction of small sample spaces which have some randomness properties is of major theoretical and practical importance. A typical property is that the probability distribution, induced on every k bit locations in a string randomly selected in the sample space, should be uniform. Such a sample space is called *k-wise independent.*

Alon, Babai and Itai [3] presented an efficient construction of k-wise independent sample spaces of size $n^{k/2}$, where n is (as above) the length of the strings in the sample space. This result is the best possible, in view of the matching lower bound of Chor. et. al. [5]. Hence, k-wise independent sample spaces of size polynomial in n are only possible for constant k. This fact led Naor and Naor to introduce the notion of *almost k-wise independent* sample spaces. Loosely speaking, the probability distribution induced on every k bit locations in the sample string is *"statistically close"* to uniform. Clearly, if an algorithm "behaves well" on points chosen from a k-wise independent sample space then it will "behave essentially as well" on points chosen from an almost k-wise independent sample space.

Naor and Naor presented an efficient construction of an almost k-wise independent sample space [15]. Points in their sample space are specified by $O(\log\log n + k + \log\frac{1}{\epsilon})$ bits, where ϵ is a bound on the statistical difference between the distribution induced on k bit locations and the uniform one.

The heart of their construction is a sample space of size $(\frac{n}{\epsilon})^{O(1)}$ for which the exclusive-or of any fixed bit locations, in the sample point, induces a 0-1 random variable with bias bounded by ϵ (i.e. the exclusive-or of these bits is 1 with probability $\frac{1}{2} \pm \epsilon$). The constant in the exponent depends, among other things, on the constants involved in an explicit construction of an expander (namely the degree and second eigenvalue of the expander). Using the best known expanders [13] this constant is around 10.

We present a construction of a sample space of size $(\frac{n}{\epsilon})^2$ for which the exclusive-or of any fixed bit locations, in the sample point, induces a 0-1 random variable with bias bounded by ϵ. Our construction is so simple that it can be described the the rest of this paragraph. A point in our sample space is specified by two bit strings of length $m \stackrel{\text{def}}{=} \log n/\epsilon$ each, 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 $r_i = \sum_{j=0}^{m-1} f_j \cdot r_{i-m-1+j}$ for $i \geq m$.

2 Formal Setting

We will consider probability distributions on binary strings of length n. In particular, we will construct probability distributions which are uniform over some set $S \subseteq \{0,1\}^n$. The parameter that will be of interest to us is the "size of the probability space"; namely, the number of strings in the support (i.e. $|S|$). The aim is to construct "small" probability spaces which have "good" randomness properties. In particular we will be interested in k-wise independence.

2.1 Almost k-wise Independence

Definition 1 (k-wise independence): *A probability-space S is k-wise independent if when $X = x_1 \cdots x_n$ is chosen uniformly from S then for any k positions $i_1 < i_2 < \cdots < i_k$ and any k-bit string α, we have*

$$Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] = 2^{-k}.$$

For all practical purposes it is sufficient that a set of bits is "almost" k-wise independent. There are several standard ways of quantifying this condition (i.e. interpreting the phrase "almost"): cf. [4]. We use two very natural ways corresponding to the L_∞ and L_1 norms:

Definition 2 (almost k-wise independence): *Let S be probability-space and $X = x_1 \cdots x_n$ be chosen uniformly from S.*

- (max-norm): *S is* (ϵ, k)-independent *if for any k positions $i_1 < i_2 < \cdots < i_k$ and any k-bit string α, we have*

$$|Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-k}| \leq \epsilon.$$

- (statistical closeness): *S is* ϵ-away from k-independence *if for any k positions $i_1 < i_2 < \cdots < i_k$ we have*

$$\sum_{\alpha \in \{0,1\}^k} |Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-k}| \leq \epsilon.$$

Clearly, if S is (ϵ, k)-independent then it is at most $2^k\epsilon$-away from k-independence, whereas if S is ϵ-away from k-independence then it is (ϵ, k)-independent. The first relation seems more typical.

2.2 The Basic Construction

The heart of our construction is a sample space which is very close to random with respect to "linear Boolean tests" (i.e., tests which take the exclusive-or of the bits in some fixed locations in the string). Following Naor and Naor [15], this sample space can be used in various ways to achieve almost k-wise independence. Our construction is based on feedback shift register sequences.

Definition 3 (linear feedback shift register sequences): *Let $\overline{s} = s_0, s_1, \ldots s_{m-1}$ and $\overline{f} = f_0, f_1, \ldots f_{m-1}$ be two sequences of m bits each. 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 $r_i = \sum_{j=0}^{m-1} f_j \cdot r_{i-m-1+j}$ for $i \geq m$.*

Our sample space will consist of all shift register sequences generated by "non-degenerate" feedback rules and any starting sequence.

Construction 1 (The Sample Space S_n^m): *The sample space S_n^m is the set of all shift register sequences generated by a feedback rule $\overline{f} = f_0 f_1 \cdots f_{m-1}$ with $f_0 = 1$ and $f(t) \stackrel{\text{def}}{=} t^m + \sum_{j=0}^{m-1} f_j \cdot t^j$ being an irreducible polynomial (such a feedback rule is called* non-degenerate*). Namely, S_n^m 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 $\overline{r}$.*

Hence, the size of the sample space S_n^m is at most 2^{2m} (actually, it is $\approx \frac{2^{2m}}{m}$). As stated before, we start by evaluating the quality of this sample space with respect to "linear Boolean tests" (i.e. the exclusive-or of a specific subset of the bits).

Definition 4 :

- *Let $(\alpha,\beta)_2$ denote the inner-product mod 2 of the binary vectors α and β (i.e. $(\alpha_1 \cdots \alpha_n, \beta_1 \cdots \beta_n)_2 = \sum_{i=1}^{n} \alpha_i\beta_i \bmod 2$).*
- *A 0-1 random variable X is called ϵ-biased if*

$$|Pr[X = 0] - Pr[X = 1]| \leq \epsilon.$$

Proposition 1 : *For any nonzero α the random variable $(\alpha, r)_2$ is $n2^{-m}$-biased when r is selected uniformly in S_n^m.*

Setting $m = k + \log n + \log \frac{1}{\epsilon}$, the sample space S_n^m is ϵ-away from k-wise independence. The proof of Proposition 1 is given in Section 3. Using the XOR-Lemma of Vazirani [18] we immediately get

Theorem 1 : *For any $k \leq n$, the sample space S_n^m is $(n2^{-m}, k)$-independent.*

A sample space is called *linear* if its elements are obtained by a linear transformation of their succinct representation (equivalently, the sample space is a linear subspace). Note that the construction of a k-wise independent sample space presented by Alon, Babai and Itai [3] is linear. Naor and Naor observed that a sample space with is almost unbiased with respect to linear Boolean tests can be used to sample points in a linear k-wise independent sample space while only moderately increasing the bias with respect to linear Boolean tests. Hence, we can efficiently construct a sample space R_N^m having the same size as S_n^m but containing much longer strings. For $N < 2^{\frac{n}{k}}$, the new space R_N^m has the same guarantee for almost independence. Namely,

Theorem 2 *For any $k \leq n$, the sample space R_N^m (containing $\approx \frac{2^{2m}}{m}$ strings each of length N) is $(k\lceil \log N \rceil 2^{-m}, k)$-independent.*

Setting $m = k + \log k + \log\log N + \log \frac{1}{\epsilon}$, the sample space R_N^m is ϵ-away from k-wise independence.

3 Proof of Proposition 1

For the rest of the paper we consider only polynomials over $GF(2)$. The number of irreducible monic polynomial of degree m is

$$\frac{1}{m}\sum_{d|m}\mu(\frac{m}{d})2^d$$

This expression is well approximated by $\frac{2^m}{m}$. For the rest of this abstract we will, for notational simplicity, treat the number of irreducible monic polynomials of degree m as if it is exactly $\frac{2^m}{m}$. (The tiny error introduced by the approximation is negligable anyhow.) Hence, the size of A_n^m is $\frac{2^{2m}}{m}$. We now turn to the proof of Proposition 1.

Fix the feedback rule and consider the distribution of $(\alpha, r)_2$ when we only vary the starting vector. 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 $GF(2)$, the reduction of t^i 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 expression of r_i as a function of the s_j's. Hence, a linear combination of the r_i's (which is exactly what $(\alpha, r)_2$ 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 $g(t) \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} \alpha_i \cdot t^i$; whereas in the second case $(\alpha, r)_2$ being a non constant combination of the s_i's is unbiased when the s_i's are uniformly selected.

Hence we get the following expression for the bias of $(\alpha, r)_2$ when r is uniformly selected in S_n^m:

$$\begin{aligned} 2 \cdot \mathrm{E}_r|(\alpha, r)_2 - \frac{1}{2}| &\leq \mathrm{E}_{\overline{f}}|\mathrm{E}_{\overline{s}} 2 \cdot (\alpha, r)_2 - 1| \\ &= \frac{\sum_{f(t)|g(t)} |\mathrm{E}_{\overline{s}} 2 \cdot (\alpha, r)_2 - 1| + \sum_{f(t)\not|g(t)} |\mathrm{E}_{\overline{s}} 2 \cdot (\alpha, r)_2 - 1|}{2^m/m} \end{aligned}$$

The first term can be bounded by the number of irreducible monic polynomials of degree m which divide a polynomial of degree $n-1$: there are at most $\frac{n-1}{m}$ such polynomials (out of the $\frac{2^m}{m}$ irreducible monic polynomials of degree m). The second term is identically zero. The proposition follows. ■

4 Using the Sample Space

As is clear from the above, the points in the sample space (S_n^m as well as the points in R_N^m) are specified by irreducible monic polynomials of degree m (used to specify a non-degenerated feedback rule) and another m-bit string (specifying the start sequence). However, the reader may wonder whether problems are not encountered once we wish to generate sample points. As will be clear from this section, the answer to this worry depends on the application: either there is no difficulty or the difficulties can be easily resolved.

In some applications we are allowed to use a preprocessing stage of complexity 2^m. Two notable examples follow

- The sample space is used for *deterministic simulation* of a randomized algorithm. In such a case the overall complexity will be a factor of $\frac{2^{2m}}{m}$ anyhow, so we might as well go through a preprocessing stage which costs less...
- The sample space contains strings of length comparable to 2^m. This is the case, for example, when m is selected such that the sample space is ϵ-away from $\log n$-wise independent, for some fixed ϵ (or $\epsilon = n^{-O(1)}$) (cf. [15]).

In the preprocessing stage, we may enumerate all monic polynomials of degree m and discard those which have non-trivial divisors. In case such a preprocessing is too costly we select a sample of monic polynomials so that we are guaranteed that, with overwhelmingly high probability, at least one of these polynomials is irreducible. A straight forward sample will require m^2 independently selected random polynomials, meaning that we use $m^3 + m$ unbiased bits to select an element of S_n^m (instead of $2m$ bits). An alternative procedure is suggested below.

Construction 2 (sample space for irreducible polynomials):

- *Use pairwise-independent sampling to specify m monic polynomials of degree m. With probability at least $\frac{1}{2}$, at least one of these polynomials is irreducible. The pairwise independent sampling requires $2m$ bits (cf. [6]). Call the resulting sample space P_m.*
- *Use an expander-path of lenght $2m$ to specify $2m$ points in the sample space P_m. With probability at least $1 - 2^{-m}$, at least one of these*

points specifies a sequence of m polynomials containing at least one irreducible polynomial (cf. [2, 7, 11, 9]). This sampling requires $O(m)$ bits. Call the resulting sample space E_m.

- *A sample point in E_m specifies m^2 polynomials and with overwhelming probability at least one of them is irreducible. Say we use the first irreducible polynomial among these m^2 polynomials (to specify the feedback rule). We now select a starting sequence which, together with the above feedback rule, specifies a sample point $r \in S_n^m$. Note that we used $O(m)$ bits to specify this sample point.*

Although this choice *does not* specify a uniformly selected irreducible polynomial, it is easy to see that the probability that the polynomial selected in this manner divides a fixed n degree polynomial is bounded above by $m^2 \cdot \frac{n}{2^m}$. Hence, the above construction gives

Proposition 2 : *For any nonzero α the random variable $(\alpha, r)_2$ is $nm^2 2^{-m}$-biased when r is selected in S_n^m with distribution induced by Construction 2.*

Theorem 3 *For any $k \leq n$, a string r selected in the set R_N^m (defined as in Theorem 2) according to the distribution induced by Construction 2 is $(k\lceil \log N \rceil 2^{-m}, k)$-independent. Again, r is specified using $O(m)$ bits.*

Finally, observe that one can get the ith bit by $O(\log i)$ matrix multiplications.

5 Concluding Remarks

This paper may be viewed as an explanation for the popularity of using linear feedback shift registers for sampling purposes. We showed that when both the feedback rule and the starting sequence are selected at random the resulting feedback sequence enjoys "almost independence" comparable to the length of the register.

Acknowledgment: We wish to thank Guy Even for collaboration in the early stages of this research.

References

[1] Adleman, L.M., and M-D.A. Huang, "Recognizing Primes in Random Polynomial Time", *Proc. 19th STOC*, 1987, pp. 462–470.

[2] M. Ajtai, J. Komlos, E. Szemeredi, "Deterministic Simulation in LOGSPACE", *Proc. 19th STOC*, 1987, pp. 132–140.

[3] N. Alon, L. Babai, and A. Itai, "A fast and Simple Randomized Algorithm for the Maximal Independent Set Problem", *J. of Algorithms*, Vol. 7, 1986, pp. 567–583.

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

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

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

[7] A. Cohen and A. Wigderson, "Dispensers, Deterministic Amplification, and Weak Random Sources", *30th FOCS*, 1989, pp. 14–19.

[8] O. Gabber, Z. Galil, "Explicit Constructions of Linear Size Superconcentrators", *JCSS*, **22** (1981), pp. 407-420.

[9] O. Goldreich, R. Impagliazzo, L.A. Levin, R. Venkatesan, D. Zuckerman, "Security Preserving Amplification of Hardness", submitted to *31st FOCS*, 1990.

[10] S. Goldwasser, and J. Kilian, "Almost All Primes Can Be Quickly Certified", *Proc. 18th STOC*, 1986, pp. 316–329.

[11] R. Impagliazzo, and D. Zuckerman, "How to Recycle Random Bits", *30th FOCS*, 1989, pp. 248-253.

[12] M. Luby, "A simple parallel algorithm for the maximal independent set problem", *Proc. 17th STOC*, 1985, pp. 1–10.

[13] A. Lubotzky, R. Phillips, P. Sarnak, "Explicit Expanders and the Ramanujan Conjectures", *STOC 86.*

[14] K. Mulmuley, U.V. Vazirani and V.V. Vazirani, "Matching is as easy as Matrix Inversion", *Proc. 19th STOC*, 1987, pp. 345–354.

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

[16] M.O. Rabin, "Probabilistic Algorithms for Testing Primality", *J. of Num. Th.*, 12, pp. 128–138, 1980.

[17] R. Solovay, and V. Strassen, "A Fast Monte-Carlo Test for Primality", *SIAM J. of Comp.*, 6, pp. 84–85, 1977.

[18] U.V. Vazirani, "Randomness, Adversaries and Computation", Ph.D. Thesis, EECS, UC Berkeley, 1986.