Sparse Pseudorandom Distributions

Oded Goldreich* **and Hugo Krawczyk†**
Computer Science Department, Technion, Haifa, Israel

ABSTRACT

The existence of *sparse pseudorandom* distributions is proved. These are probability distributions concentrated in a very small set of strings, yet it is infeasible for any polynomial-time algorithm to distinguish between truly random coins and coins selected according to these distributions. It is shown that such distributions can be generated by (nonpolynomial) probabilistic algorithms, while probabilistic polynomial-time algorithms cannot even approximate all the pseudorandom distributions. Moreover, we show the existence of *evasive* pseudorandom distributions which are not only sparse, but also have the property that no polynomial-time algorithm may find an element in their support, except for a negligible probability. All these results are proved independently of any intractability assumption.

1. INTRODUCTION

In recent years, randomness has become a central notion in diverse fields of computer science. Randomness is used in the design of algorithms in fields such as computational number theory, computational geometry and parallel and distributed computing, and is crucial to cryptography. Since in most cases the interest is in the behavior of efficient algorithms (modeled by polynomial-time computations), the fundamental notion of pseudorandomness arises. Pseudorandom distributions are those distributions which cannot be efficiently distinguished from the uniform distribution on strings of the same length.

The importance of pseudorandomness is in the fact that any efficient probabilistic algorithm performs essentially as well when substituting its source of unbiased coins by a pseudorandom sequence. Algorithms can therefore be

* Supported by Grant No. 86-00301 from the United States – Israel Binational Science Foundation (BSF), Jerusalem, Israel.
† Current address: IBM T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598.

Random Structures and Algorithms, Vol. 3, No. 2 (1992)

analyzed assuming they use unbiased coin tosses, and later implemented using pseudorandom sequences. Such an approach is practically beneficial if pseudorandom sequences can be generated more easily than "truly random" ones. This gave rise to the notion of a pseudorandom generator – an efficient deterministic algorithm which *expands* random seeds into longer pseudorandom sequences.

Most of the previous work on pseudorandomness has in fact focused on pseudorandom generators. Blum and Micali [1] and Yao [14] suggested the basic definitions and showed that pseudorandom generators can be constructed under certain (necessary) intractability assumptions.* Several works [4, 12, 10, 11, 6, 9, 7, 3] further developed this direction. An important aspect of pseudorandom generation, namely its utility for deterministic simulation of randomized complexity classes, is further studied in [13].

In our article we investigate the notion of pseudorandomness when decoupled from the notion of efficient generation. This investigation is carried out using no unproven assumptions. The first question we address is the existence of nontrivial pseudorandom distributions, that is, pseudorandom distributions which substantially differ from the uniform distribution. Yao [14] suggests a particular example of such a distribution. Further properties of such distributions are developed here.

We prove the existence of *sparse* pseudorandom distributions. A distribution is called sparse if it is concentrated on a negligible part of the set of all strings of a given length. For example, given a positive constant $\delta < 1$ we construct a probability distribution concentrated on $2^{\delta k}$ of the strings of length k which cannot be distinguished from the uniform distribution on the set of all k-bit strings (and hence is pseudorandom).

We show that sparse pseudorandom distributions can even be uniformly generated by probabilistic algorithms (that run in nonpolynomial time). These generating algorithms use less random coins than the number of pseudorandom bits they produce. Viewing these algorithms as generators which expand randomly selected short strings into much longer pseudorandom sequences, we can exhibit generators achieving subexponential expansion rate. This expansion is optimal as we show that no generator expanding strings into exponential longer ones can induce a pseudorandom distribution (which passes nonuniform tests). On the other hand, we use the subexponential expansion property in order to construct nonuniform generators of size slightly super-polynomial. An improvement to this result, namely, a proof of existence of nonuniform polynomial-size generators would separate nonuniform-P from nonuniform-NP, which would be a major breakthrough in Complexity Theory.

We also prove the existence of sparse pseudorandom distributions that cannot be generated or even approximated by efficient algorithms. Namely, there exist pseudorandom distributions that are statistically far from any distribution which is induced by any probabilistic polynomial-time algorithm. In other words, even if efficiently generable pseudorandom distributions exist, they do not exhaust (nor even in an approximative sense) all the pseudorandom distributions.

Finally, we introduce the notion of *evasive* probability distributions. These

* Intractability assumptions for constructing (polynomial-time) pseudorandom generators are unavoidable as long as we cannot prove the existence of one-way functions and, in particular, that $P \neq NP$. We stress that such a generator constitutes by itself a one-way function.

probability distributions have the property that any efficient algorithm will fail to find strings in their support* (except with a negligible probability). Certainly, evasive probability distributions are sparse, and cannot be efficiently approximated by probabilistic algorithms. We show the existence of *evasive pseudorandom* distributions.

Interestingly, we have applied the "abstract-flavored" results presented here in order to resolve two open questions concerning the sequential and parallel composition of zero-knowledge interactive proofs. This application is presented in a companion paper [5].

2. DEFINITIONS

The formal definition of pseudorandomness (given below) is stated in asymptotical terms, so we shall not discuss single distributions but rather collections of probability distributions called probability ensembles.

Definition. A *probability ensemble* Π is a collection of probability distributions $\{\pi_k\}_{k\in K}$, such that K is an infinite set of indices (nonnegative integers) and for every $k \in K$, π_k is a probability distribution on the set of (binary) strings of length k. In particular, an ensemble $\{\pi_k\}_{k\in K}$ in which π_k is a uniform distribution on $\{0,1\}^k$ is called a *uniform ensemble.*

Next, we give a formal definition of a pseudorandom ensemble. This is done in terms of polynomial indistinguishability between ensembles.

Definition. Let $\Pi = \{\pi_k\}$ and $\Pi' = \{\pi'_k\}$ be two probability ensembles. Let T be a probabilistic polynomial time algorithm outputting 0 or 1 (T is called a *statistical test*). Denote by $p_T(k)$ the probability that T outputs 1 when fed with an input selected according to the distribution π_k. Similarly, $p'_T(k)$ is defined with respect to π'_k. The test T *distinguishes* between Π and Π' if and only if there exists a constant $c>0$ and infinitely many k's such that $|p_T(k) - p'_T(k)| > k^{-c}$. The ensembles Π and Π' are called *polynomially indistinguishable* if there exists no polynomial-time statistical test that distinguishes between them.

Definition. A probabilistic ensemble is called *pseudorandom* if it is polynomially indistinguishable from a uniform ensemble.

Remark. Some authors define pseudorandomness by requiring that pseudorandom ensembles be indistinguishable from uniform distributions even by *nonuniform* (polynomial) tests. We stress that the results (and proofs) in this article also hold for these stronger definitions.

Notice that since probabilistic algorithms are fed with random bits chosen according to the uniform distribution, it is trivial for them to output uniform ensembles. Here we are interested in the question of whether nontrivial pseudorandom ensembles can be effectively sampled by means of probabilistic algorithms. The following definition captures the notion of "samplability."

* The support of a probability distribution is the set of elements that it assigns nonzero probability.

Definition. A *sampling algorithm* is a probabilistic algorithm A that on input a string of the form 1^n, outputs a string of length n. The *probabilistic ensemble* $\Pi^A = \{\pi_n^A\}_n$ *induced by a sampling algorithm* A is defined by $\pi_n^A(y) = Prob(A(1^n) = y)$, where the probability is taken over the coin tosses of algorithm A. A *samplable* ensemble is a probabilistic ensemble induced by a sampling algorithm. If the sampling algorithm uses, on input 1^n, less than n random bits, then we call the ensemble *strongly-samplable.*

Traditionally, pseudorandom generators are defined as efficient *deterministic* algorithms expanding short seeds into longer bit strings. Using the above terminology we can view them as strong-sampling algorithms (the seed is viewed as the random coins for the sampling algorithm).

We consider a pseudorandom ensemble to be trivial if it is "close" to a uniform ensemble. The meaning of "close" is formalized in the next definition.

Definition. Two probabilistic ensembles Π and Π' are *statistically close* if for any positive c and any sufficiently large n,

$$\sum_{x \in \{0,1\}^n} |\pi_n(x) - \pi'_n(x)| < n^{-c}.$$

A special case of nontrivial pseudorandom ensembles are those ensembles we call "sparse."

Definition. A probabilistic ensemble is called *sparse* if (for sufficiently large n's) the support of π_n is a set of *negligible* size relative to the set $\{0,1\}^n$ (i.e., for every $c > 0$ and sufficiently large n,

$$|\{x \in \{0,1\}^n : \pi_n(x) > 0\}| < n^{-c} 2^n).$$

Clearly, a sparse pseudorandom ensemble cannot be statistically close to a uniform ensemble.

Notation. I_k will denote the set $\{0,1\}^k$.

3. THE EXISTENCE OF SPARSE PSEUDORANDOM ENSEMBLES

The main result in this section is the following Theorem.

Theorem 1. *There exist strongly-samplable sparse pseudorandom ensembles.*

In order to prove this theorem we present an ensemble of sparse distributions which are pseudorandom even against nonuniform distinguishers. These distributions assign equal probability to the elements in their support. We use the following definitions.

Definition. Let C be a (probabilistic) circuit with k inputs and a single output. We say that a set $S \subseteq I_k$ is $\epsilon(k)$-*distinguished* by the circuit C if

$$|p_C(S) - p_C(I_k)| > \epsilon(k)$$

where $p_C(S)$ (resp. $p_C(I_k)$) denotes the probability that C outputs 1 when given elements of S (resp. I_k), chosen with uniform probability.

A set $S \subseteq I_k$ is called $(\tau(k), \epsilon(k))$-*pseudorandom* if it is *not* $\epsilon(k)$)-distinguished by any circuit of size at most $\tau(k)$.

Note that a collection of uniform distributions on a sequence of sets $S_1, S_2, \ldots$ where each S_k is a $(\tau(k), \epsilon(k))$-pseudorandom set, constitutes a pseudorandom ensemble, provided that both functions $\tau(k)$ and $\epsilon^{-1}(k)$ are super-polynomial (i.e., grow faster than any polynomial). Our goal is to prove the existence of such a collection for which the ratio $|S_k|/2^k$ is negligibly small.

Remark. In the following we consider only deterministic circuits (tests). The ability to toss coins does not add power to nonuniform tests. Using a standard averaging argument one can show that a deterministic nonuniform distinguisher C' with distinguishing probability $\delta'(k)$ can be obtained from a probabilistic nonuniform distinguisher C with distinguishing probability $\delta(k)$, where $\delta'(k) \geq \delta(k)$. The circuit C' is obtained from C by setting the bits on the random tape of C to values that achieve the largest distinguishing probability among all assignments of values to the random tape.

The next lemma measures the number of sets which are $\epsilon(k)$- distinguished by a given circuit. Notice that this result does not depend on the circuit size.

Lemma 2. *For any k-input Boolean circuit C, the probability that a random set $S \subseteq I_k$ of size N is $\epsilon(k)$-distinguished by C is at most $2e^{-2N\epsilon^2(k)}$.*

Proof. Let $L_C(k)$ be the set $\{x \in I_k : C(x) = 1\}$. Thus,

$$p_C(I_k) = \frac{|L_C(k)|}{2^k} \quad \text{and} \quad p_C(S) = \frac{|S \cap L_C(k)|}{|S|} .$$

Consider the set of strings of length k as an urn containing 2^k balls. Let those balls in $L_C(k)$ be painted white and the others black. The proportion of white balls in the urn is clearly $p_C(I_k)$, and the proportion of white balls in a sample S of N balls from the urn is $p_C(S)$. (We consider here a sample *without* replacement, i.e., sampled balls are not replaced in the urn.)

Lemma 2 follows by using the Chernoff-type inequality due to W. Hoeffding [8] (see Appendix)

$$Prob(|p_C(S) - p_C(I_k)| \geq \epsilon(k)) < 2e^{-2N\epsilon^2(k)} .$$

where the probability is taken over all the subsets $S \subseteq I_k$ of size N, with uniform probability. ∎

Corollary 3. *For any positive integers k and N, and functions $\tau(\cdot)$ and $\epsilon(\cdot)$, the proportion of subsets of I_k of size N which are $(\tau(k), \epsilon(k))$-pseudorandom is at least $1 - 2^{\tau^2(k) - 2N\epsilon^2(k)}$.*

Proof. The number of Boolean circuits of size $\tau(k)$ is at most $2^{\tau^2(k)}$. Therefore, using Lemma 2 we get that the proportion of sets $S \subseteq I_k$ of size N which are $\epsilon(k)$-distinguished by any k-input Boolean circuit of size $\tau(k)$ is at most

$$2^{\tau^2(k)} \cdot 2e^{-2N\epsilon^2(k)} < 2^{\tau^2(k)-2N\epsilon^2(k)}. \quad \blacksquare$$

The following Corollary shows there are pseudorandom ensembles composed of uniform distributions with very sparse support.

Corollary 4. *Let $k(n)$ be any subexponential function of n (i.e., $k(n) = 2^{o(n)}$).* There are super-polynomial functions $\tau(\cdot)$ and $\epsilon^{-1}(\cdot)$, and a sequence of sets $S_1, S_2, \ldots$ such that S_n is a $(\tau(k(n)), \epsilon(k(n)))$-pseudorandom subset of $I_{k(n)}$ and $|S_n| = 2^n$.*

Proof. Using Corollary 3 we get that a $(\tau(k(n)),\epsilon(k(n)))$-pseudorandom set $S \subseteq I_{k(n)}$ of size 2^n exists provided that

$$2^n\epsilon^2(k(n)) > \tau^2(k(n)) \tag{1}$$

It is easy to see that for any subexponential function $k(n)$ we can find super-polynomial functions $\epsilon^{-1}(\cdot)$ and $\tau(\cdot)$ such that inequality (1) holds for each value of n. $\blacksquare$

The following lemma states that the sparse pseudorandom ensembles presented above are strongly-samplable. This proves Theorem 1.

Lemma 5. *Let $k(n)$ be any subexponential function of n. There are (nonpolynomial) generators which expand random strings of length n into pseudorandom strings of length $k(n)$.*

Proof. Let $\tau(\cdot)$ and $\epsilon(\cdot)$ be as in Corollary 4. We construct a generator which on input of a seed of length n finds the $(\tau(k(n)), \epsilon(k(n)))$-pseudorandom set $S_n \subseteq I_{k(n)}$ whose existence is guaranteed by Corollary 4, and uses the n input bits in order to choose a random element from S_n. Clearly, the output of the generator is pseudorandom.

To see that the set S_n can be effectively found, note that it is effectively testable whether a given set S of size 2^n is $(\tau(k), \epsilon(k))$-pseudorandom. This can be done by enumerating all the circuits of size $\tau(k)$ and computing for each circuit C the quantities $p_C(S)$ and $p_C(I_k)$. Thus, our generator will test all the possible sets $S \subseteq I_k$ of size 2^n until S_n is found. $\blacksquare$

Remark 1. Inequality (1) implies a tradeoff between the expansion function $k(n)$ and the size of the tests (circuits) resisted by the generated ensemble. The pseudorandom ensembles we construct may be "very" sparse, in the sense that the expansion function $k(n)$ can be chosen to be very large (e.g., $2^{\sqrt{n}}$). On the other hand, if we consider "moderate" expansion functions such as $k(n) = 2n$, we can resist rather powerful tests, e.g., circuits of size $2^{n/4}$.

* $o(n)$ denotes any function $f(n)$ such that $\lim_{n\to\infty} f(n)/n = 0$.

Remark 2. The subexponential expansion, as allowed by our construction, is optimal since there is no generator which expands strings of length n into strings of length $k(n) = 2^{O(n)}$. To see this, consider a circuit C of size $k(n)^{O(1)}$ $(=(2^n)^{O(1)})$ which incorporates the (at most) 2^n strings of length $k(n)$ output by the generator. On input a string of length $k(n)$ the circuit C looks up whether this input appears in the incorporated list of strings output by the generator. Clearly, this circuit C constitutes a (nonuniform) test (of size polynomial in $k(n)$) which distinguishes the output of this generator from the uniform distribution on $I_{k(n)}$.

Remark 3. The subexponential expansion implies that the supports of the resultant pseudorandom distributions are very sparse. More precisely, our construction implies the existence of generators which induce on strings of length k a support of size *slightly* super-polynomial (i.e., of size $k^{\omega(k)}$ for an arbitrary nondecreasing unbounded function $\omega(k)$). Thus, by wiring this support into a Boolean circuit, we are able to construct *nonuniform* generators of size slightly super-polynomial. (On input of a seed s the circuit (generator) outputs the sth element in this "pseudorandom" support.) Let us point out that an improvement of this result, i.e., a proof of the existence of nonuniform pseudorandom generators of polynomial size, will imply that nonuniform-P $\neq$ nonuniform-NP. This follows by considering the language $\{x \in I_k : x \text{ is in the image of } G\}$, where G is a pseudorandom generator in nonuniform-P. Clearly, this language is in nonuniform-NP, but not in nonuniform-P, otherwise a decision procedure for it can be transformed into a test distinguishing the output of G from the uniform distribution on I_k.

Remark 4. The (uniform) complexity of the generators constructed in Lemma 5 is slightly super-exponential, i.e., $2^{k^{\omega(k)}}$, for unbounded $\omega(\cdot)$. (The complexity is, up to a polynomial factor, $2^{\tau^2(k)} \cdot (2^n + 2^k) \cdot \binom{2^k}{2^n}$, and 2^n is, as in Remark 3, slightly super-polynomial in k.) We stress that the existence of pseudorandom generators running in exponential time, and with arbitrary polynomial expansion function, would have interesting consequences in Complexity Theory as BPP $\subseteq \bigcap_{\epsilon>0}$ DTIME(2^{n^ϵ}) [14, 13].

4. THE COMPLEXITY OF APPROXIMATING PSEUDORANDOM ENSEMBLES

In the previous section we have shown sparse pseudorandom ensembles which can be sampled by probabilistic algorithms running in super-exponential time. The question of whether it is possible to sample at least *some* pseudorandom ensemble by polynomial-time (or even exponential-time) algorithms can only be answered today in the affirmative by making a complexity assumption. This raises the natural question of whether or not *all* pseudorandom ensembles can be sampled by polynomial-time (or exponential-time) algorithms. We give here a negative answer to this question, proving (without any assumptions) that for any complexity function $\phi(\cdot)$ there exists a samplable pseudorandom ensemble which cannot be sampled nor even "approximated" by algorithms in RTIME(ϕ). The notion of approximation is defined next.

Definition. A probabilistic ensemble Π is *approximated* by a sampling algorithm A if the ensemble Π^A induced by A is statistically close to Π. (See Section 2 for the definition of "statistically close).''

The main result of this section is stated in the following theorem.

Theorem 6. *For any complexity (constructive) function $\phi(\cdot)$, there is a strongly samplable pseudorandom ensemble that cannot be approximated by any algorithm whose running time is bounded by ϕ.*

Proof. We say that two probability distributions π and π' on a set X are $\frac{1}{2}$-*close* if

$$\sum_{x \in X} |\pi(x) - \pi'(x)| < \tfrac{1}{2}.$$

We say that a sampling algorithm M $\frac{1}{2}$-approximates a set $S \subseteq I_k$ if the probability distribution π_k^M induced by M on I_k and the uniform distribution U_S on S are $\frac{1}{2}$-close.

We show that for any sampling algorithm M most subsets of I_k of size 2^n are not $\frac{1}{2}$-approximated by M (for k sufficiently large with respect to n). This follows from the next lemma.

Lemma 7. *Let π be a probability distribution on I_k. The probability that π and U_S are $\frac{1}{2}$-close, for S randomly chosen over the subsets of I_k of size 2^n, is less than $(1/2)^{k-n-1}$.*

Proof. Notice that if two different sets S and T are $\frac{1}{2}$-close, for S randomly chosen over the subsets of I_k of size 2^n, is less than $(1/2)^{k-n-1}$.

$$\sum_{x \in I_k} |U_S(x) - \pi(x)| < \frac{1}{2} \quad \text{and} \quad \sum_{x \in I_k} |U_T(x) - \pi(x)| < \frac{1}{2}.$$

Using the triangle inequality we conclude that

$$\sum_{x \in I_k} |U_S(x) - U_T(x)| < 1.$$

Denoting the last sum by σ and the symmetric difference of S and T by D, we have that $|D| \cdot \frac{1}{2^n} < \sigma < 1$ (this follows from the fact that U_S and U_T assign uniform probability to the 2^n elements of S and T, respectively). But this implies that $|D| < 2^n$, and then (using $|S| + |T| = |D| + 2 \cdot |S \cap T|$) we get $|S \cap T| > 2^n/2$. Let T be a particular subset of I_k of size 2^n which is $\frac{1}{2}$-close to π. From the above argument it follows that the collection of subsets of size 2^n which are $\frac{1}{2}$-close to π is included in the collection $\{S \subseteq I_k : |S| = 2^n, |S \cap T| > 2^n/2\}$. Thus, we are able to bound the probability that π is $\frac{1}{2}$-close to a random set S of size 2^n, by the probability of the following experiment. Fix a set $T \subseteq I_k$ of size 2^n, and take at random a set S of 2^n elements among all the strings in I_k. We are interested in the probability that $|S \cap T| > 2^n/2$. Clearly, the expectation of $|S \cap T|$ is $\frac{|S| \cdot |T|}{2^k}$.

Using Markov inequality for nonnegative random variables we have

$$Prob\left(|S \cap T| > \frac{2^n}{2}\right) \cdot \frac{2^n}{2} < \frac{|S| \cdot |T|}{2^k}$$

and then

$$Prob(|S \cap T| > 2^n/2) < 2/2^{k-n} \tag{2}$$

The lemma follows. □

We now extend the pseudorandom generator constructed in Lemma 5, in order to obtain a generator for a pseudorandom ensemble which is not approximated by any ϕ-time sampling algorithm. On input a string of length n, the generator proceeds as in Lemma 5. Once a $(\tau(k(n)), \epsilon(k(n)))$-pseudorandom subset S_n is found, the generator checks whether S_n is $\frac{1}{2}$-approximated by all of the first n Turing machines, in some canonical enumeration, by running each of them as a sampling algorithm for $\phi(k(n))$ steps. Clearly, it is effectively testable whether a given machine M $\frac{1}{2}$-approximates a given set S. If the set S_n is $\frac{1}{2}$-approximated by any one of these machines, it is discarded and the next $S \subseteq I_k$, $|S| = 2^n$ is checked (first for pseudorandomness and then for approximation).

By Corollary 3 we have that for a suitable choice of the functions $(\tau(\cdot)$ and $\epsilon(\cdot)$ the probability that a set S is $(\tau(k(n)), \epsilon(k(n)))$-pseudorandom is almost 1. On the other hand, the probability that a set S is $\frac{1}{2}$-approximated by n sampling machines is, using Lemma 7, less than $n/2^{k(n)-n-1}$. For suitable $k(\cdot)$, e.g., $k(n) \geq 2n$, this probability is negligible. Thus, we are guaranteed to find a set S_n which is $(\tau(k(n)), \epsilon(k(n)))$-pseudorandom as well as not $\frac{1}{2}$-approximated by the first n sampling algorithms running ϕ-time. The resultant ensemble is as stated in the theorem. ■

Remark. The result in Theorem 6 relies on the fact that the sampling algorithms we have run are uniform ones. Nevertheless, if we use Hoeffding inequality (see Appendix) to bound the left side in (2), we derive a much better bound, which implies that for any constant $\alpha < 1$, there exist strongly-samplable pseudorandom ensembles that cannot be approximated by Boolean circuits of size $2^{\alpha n}$.

5. EVASIVE PSEUDORANDOM ENSEMBLES

In this section we prove the existence of pseudorandom ensembles which have the property that no polynomial-time sampling algorithm will output an element in their support, except for a negligible probability.

Definition. A probability ensemble $\Pi = \{\pi_k\}_{k \in K}$ is called *polynomial-time evasive* if for any polynomial-time sampling algorithm A, any constant c and sufficiently large k,

$$Prob(A(1^k) \in support(\pi_k)) < k^{-c}$$

($support(\pi_k)$ denotes the set $\{x \in I_k : \pi_k(x) > 0\}$).

Notice that evasiveness does not imply pseudorandomness. For example, any evasive ensemble remains evasive if we add to each string in the support a leading "0," while the resultant distributions are obviously not pseudorandom. On the other hand, an evasive pseudorandom ensemble is clearly sparse.

The following is the main result of this section.

Theorem 8. *There are (strongly-samplable) polynomial-time evasive pseudorandom ensembles.*

Proof. The proof outline is similar to the proof of Theorem 6. We again extend the generator of Lemma 5 by testing whether the $(\tau(k(n)), \epsilon(k(n))$-pseudorandom set S_n, found by that generator on input of length n, evades the first n Turing machines (run as polynomial-time sampling algorithms). We have to show that for each sampling algorithm M there is a small number of sets $S \subseteq I_k$ of size 2^n for which machine M outputs an element of S with significant probability. Throughout this proof we shall consider as "significant" a probability that is greater than $2^{2n}/2^k$. (Any negligible portion suffices here.) Thus, we are assuming $k \geq 3n$. We need the following technical lemma.

Lemma 9. *Let π be a fixed probability distribution on a set U of size K. For any $S \subseteq U$ denote $\pi(S) = \sum_{s \in S} \pi(s)$. Then*

$$Prob(\pi(S) > \epsilon) < \frac{N}{\epsilon K}$$

where the probability is taken over all the sets $S \subseteq U$ of size N with uniform probability.

Proof. Consider a random sample of N *distinct* elements from the set U. Let X_i, $1 \leq i \leq N$, be random variables so that X_i assumes the value $\pi(u)$ if the ith element chosen in the sample is u. Define the random variable X to be the sum of the X_i's $\left(\text{i.e., } X = \sum_{i=1}^{N} X_i\right)$. Clearly, each X_i has expectation $1/K$ and then the expectation of X is N/K. Using Markov inequality for nonnegative random variables we get

$$Prob(X > \epsilon) < \frac{E(X)}{\epsilon} = \frac{N}{\epsilon K}$$

proving the lemma. □

Let π_k^M be the probability distribution induced by the sampling algorithm M on I_k. Consider a randomly chosen $S \subseteq I_k$ of size 2^n. Lemma 9 states that

$$Prob\left(\pi_k^M(S) > \frac{2^{2n}}{2^k}\right) < \frac{1}{2^n}$$

Thus, we get that only $1/2^n$ of the subsets S fail the evasivity test for a single machine. Running n such tests the portion of failing sets is at most $n/2^n$.

Therefore, there exists a set passing all the distinguishing and evasivity tests. (Actually, using Corollary 3, we get that most of the sets of size 2^n pass these tests.) This completes the proof of the Theorem. ∎

Remark. Actually, we have proven that for any uniform time-complexity class **C**, there exist pseudorandom ensembles which evades any sampling algorithm of the class **C**. Notice that no restriction on the running time of the sampling machines is required. Thus, the results in these sections imply the results of Section 4 on unapproximability by uniform algorithms, but not the unapproximability by nonuniform circuits (see remark after the proof of Theorem 6). We stress that we cannot find ensembles evading the output of nonuniform circuits of polynomial-size, since for each set S there exists a circuit which outputs an element of S with probability 1. In [5] we construct *collections* of pseudorandom sets which are also "evasive" for nonuniform polynomial-time algorithms, in the sense that such an algorithm cannot find, for *most* sets in the collection, an element in that set, except for a negligible probability.

ACKNOWLEDGMENTS

The authors would like to thank Micha Hofri for referring them to Hoeffding inequality, and to Benny Chor and Eyal Kushilevitz for helpful comments.

APPENDIX: HOEFFDING INEQUALITY [8]

Suppose an urn contains u balls of which w are white and $u - w$ are black. Consider a random sample of s balls from the urn (without replacing any balls in the urn at any stage).

Hoeffding inequality states that the proportion of white balls in the sample is close, with high probability, to its expected value, i.e., to the proportion of white balls in the urn. More precisely, let x be a random variable assuming the number of white balls in a random sample of size s. Then, for any ϵ, $0 \le \epsilon \le 1$

$$Prob\left(\left|\frac{x}{s} - \frac{w}{u}\right| \ge \epsilon\right) < 2e^{-2s\epsilon^2}$$

This bound is often used for the case of binomial distributions (i.e., when drawn balls are replaced in the urn). The inequality for that case is due to H. Chernoff [2]. More general inequalities appear in Hoeffding's [8], as well as a proof that these bounds apply also for the case of samples without replacement.

REFERENCES

[1] M. Blum and S. Micali, How to generate cryptographically strong sequences of pseudo-random bits, *SIAM J. Comput.*, **13**, 850–864 (1984).

[2] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Math. Stat.*, **23**, 493–507 (1952).

[3] O. Goldreich, A note on computational indistinguishability, *IPL*, **34**, 277–281 (1990).

[4] O. Goldreich, S. Goldwasser, and S. Micali, How to construct random functions, **33** (4), 792–807 (1986).

[5] O. Goldreich and H. Krawczyk, On the composition of zero-knowledge proof systems, *17th ICALP Proceedings*, *Lecture Notes in Computer Science*, Vol. 443, Springer Verlag, New York, 1990, 268–282. Extended version: DIMACS Tech. Rep. 91–60, August 1991.

[6] O. Goldreich, H. Krawczyk, and M. Luby, On the existence of pseudorandom generators, *Proc. 29th IEEE Symp. on Foundation of Computer Science*, 1988, pp. 12–24.

[7] J. Hastad, Pseudo-random generators under uniform assumptions, *Proc. 22nd STOC*, 1990.

[8] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Am. Stat. Assoc.* **58**, 13–30 (1963).

[9] R. Impagliazzo, L. A. Levin, and M. G. Luby, Pseudo-random generation from one-way functions, *Proc. 21st STOC*, 1989, pp. 12–24.

[10] L. A. Levin, One-way function and pseudorandom generators, *Combinatorica*, **7** (4), 357–363 (1987).

[11] L. A. Levin, Homogeneous measures and polynomial time invariants, *Proc. 29th IEEE Symp. on Foundation of Computer Science*, 1988, pp. 36–41.

[12] M. Luby and C. Rackoff, How to construct pseudorandom permutations from pseudorandom functions, *SIAM J. Comput.*, **17**, 373–386 (1988).

[13] N. Nisan and A. Wigderson, Hardness vs. randomness, *Proc. 29th IEEE Symp. on Foundation of Computer Science*, 1988, pp. 2–11.

[14] A. C. Yao, Theory and applications of trapdoor functions, *Proc. 23rd IEEE Symp. on Foundation of Computer Science*, 1982, pp. 80–91.

Received April 18, 1990
Revised October 24, 1991