Unbiased Bits from Sources of Weak Randomness and Probabilistic Communication Complexity

(Abstract, Introduction and References only)

Benny Chor * Oded Goldreich **

MIT – Laboratory for Computer Science
Cambridge, Massachusetts 02139

ABSTRACT – A new model for weak random physical sources is presented. The new model strictly generalizes previous models (*e.g.* the Santha and Vazirani model [24]). The sources considered output strings according to probability distributions in which *no single string is too probable.*

The new model provides a fruitful viewpoint on problems studied previously as:

- *Extracting almost perfect bits from sources of weak randomness:* the question of possibility as well as the question of efficiency of such extraction schemes are addressed.
- *Probabilistic Communication Complexity:* it is shown that most functions have linear communication complexity in a very strong probabilistic sense.
- *Robustness of BPP* with respect to sources of weak randomness (generalizing a result of Vazirani and Vazirani [27]).

The paper has appeared in *SIAM Journal on Computing*, Vol. 17, No. 2, April 1988, pp. 230–261.

* Research supported in part by an IBM Graduate Fellowship and a Bantrell Postdoctoral Fellowship.
** Research supported in part by a Weizmann Postdoctoral Fellowship. On leave from the Computer Science Dept., Technion, Israel.

1. INTRODUCTION

The notion of randomness is central to the theory of computation. Thus the question of whether and how randomness can be implemented in a computer is of major importance. Our intention is not to address the metaphysical aspect of the above question. We rather assume that there are physical phenomenon which appear to be "somewhat random", and study the consequences of such assumption.

In reality, there is a variety of physical sources, the output of which appears to be unpredictable in some sense (e.g. noise diodes, Geiger counters, etc.). However, these sources do not seem to be perfect (i.e. they do not output a uniform distribution). This phenomena is amplified when trying to convert the analogue signal to a digital one, and in particular when sampling the physical source very frequently.

The main contribution of this paper is in presenting a general model for sources of weak randomness. This model not only generalizes previous models, but is also very convenient to manipulate and analyze. The new model provides a new viewpoint on several problems studied previously, and enables us to obtain interesting new results:

- *Extracting almost perfect bits from sources of weak randomness:* It is shown that almost all functions can be used for extracting many "almost unbiased" bits from two independent sources of "weak" randomness. An explicit function which performs almost as good is also presented. These results yield an extraction scheme which is efficient both in terms of output entropy and computational complexity.
- *Probabilistic Communication Complexity:* It is shown that most Boolean functions have linear communication complexity in a very strong probabilistic sense. This resolves an open problem of Yao [29].
- *Robustness of BPP* with respect to sources of weak randomness. It is shown that any probabilistic polynomial-time algorithm can be modified so that it works with bits supplied by a *single* source of weak randomness.

1.1 Previous Models

Previous works on extracting unbiased bits from non-perfect sources have implicitly or explicitly proposed models of "weak randomness". Von Neumann's classic algorithm [16] deals with sequences of bits generated by independent tosses of a single coin with fixed bias. This model is totally memoryless. Blum [4] models physical sources as finite state markov chains (with unknown transition probabilities). In this model, one can describe a dependency of the next bit (output by the source) on the previous c bits (for any fixed c).

Santha and Vazirani [24] have further relaxed the restrictions on the physical source. Their model, hereafter referred to as the *SV-model*, is the start point for our investigations. In the SV-model each bit in the output sequence is "slightly random" in the sense that it is 0 with probability at least δ and 1 with probability at least δ, where $\delta \leq 1/2$ is a constant. This allows to model a probabilistic dependency of the next bit (output by the source) on all previous bits. However, no bit of the output may be totally determined by the previous bits. It follows that in the SV-model, every bit sequence is output with some positive probability. This restriction could be violated by some "random" physical sources, which are constrained in a way that prevents certain bit sequences.

1.2 The New Model

We introduce and study a general model for physical sources, hereafter referred to as the *model of Probability-Bounded sources (PRB-sources)*. Loosely speaking, the probability that a PRB-source will output a particular *string* is bounded above by some parameter. This allows the source to be very imperfect, still it may not concentrate its probability mass on too few strings.

The PRB-model is formalized using two constants l (length parameter) and b (probability bound). A physical source S is a device which outputs an infinite sequence of bits. We say that S is a (l, b)-source if for every prefix α of the output sequence, and every l-bit string β, the conditional probability that the next l bits output by S equal β is at most 2^{-b} (i.e. $Pr(\beta|\alpha) \leq 2^{-b}$).

The PRB-model is a strict generalization of the SV-model. To see the inclusion, note that any SV-source with parameter δ is a $(1, \log_2(1-\delta)^{-1})$-source. To see that the inclusion is proper, consider the $(2,1)$-source which outputs 11 with probability 1/2 and 10 with probability 1/2. Clearly, this source is not a SV-source. Thus, *all positive results* (with respect to the PRB-model) presented *in this paper – apply also to the SV-model.*

1.3 Extracting Unbiased Bits From Sources of Weak Randomness

Algorithms for extracting unbiased bits from non-perfect sources depend on the underlying source model. Von Neumann's algorithm [16] for generating a sequence of unbiased bits by using a coin with fixed bias, is a well-known classic:

1) Toss the biased coin twice. Denote the outcome by $\sigma\tau \in \{HH, HT, TH, TT\}$.
2) If $\sigma = \tau$ then goto step (1). (nothing is output.)
3) If $\sigma\tau = HT$ output 0; If $\sigma\tau = TH$ output 1; Goto step (1).

Elias [8] improved upon von Neumann algorithm, showing how to nearly achieve the entropy of the one coin source. He also considered special type of visible finite Markov chains. His algorithm produces perfect bits from such sources.

Blum [4] has considered extracting (perfect) unbiased bits from general finite Markov chains with unknown structure and transition probabilities. He gave algorithms which work in linear expected time. Using Elias's techniques [8], the extracted bits reach the entropy of the source in the limit.

It seems that as far as extracting *perfect* unbiased bits, Blum schemes are optimal. However, as pointed out by Santha and Vazirani [24], for practical purposes one may lower the standards and settle for "almost" unbiased bits. Having this goal in mind, they further relaxed the restrictions on the physical source and introduced the SV-model (see sec 1.1). Santha and Vazirani showed that a single SV-source cannot be used to extract almost unbiased bits, while sufficiently many independent SV-sources can be used for this purpose. Vazirani [26] showed that by applying inner-product mod 2 to strings of length $C_\delta \cdot \log_2 \varepsilon^{-1}$ output by two independent SV-sources, a bit with bias $\leq \frac{1}{2} + \varepsilon$ is produced.

Summarizing the results in [24] and [26], we conclude that the SV-model presents a sufficient condition for the extraction of almost unbiased bits from two independent physical sources. We substantially relax this condition.

In this paper we show that almost all functions can be used to extract many independent unbiased bits from the output of any two independent (l, b)-sources. To be more specific, let $m = (b - 3 - \log l)/3 > 0$, and consider extraction functions from $l+l$ bits to m bits. The m extracted bits are *almost unbiased and independent* in the sense that each m-bit string appears with probability at least $(1-\frac{1}{2^m})\cdot 2^{-m}$ and at most $(1+\frac{1}{2^m})\cdot 2^{-m}$. This is achieved by a $1-2^{-2^b}$ fraction of all functions from $2l$-bit strings to m-bit strings. Notice that the number of bits we extract from the two sources is within a constant factor ($\approx \frac{1}{6}$) of the information theoretic bound, a feature not achieved in previous works [24, 26].

We also prove that, for all $b_1+b_2 \geq l+2+2\log_2 \varepsilon^{-1}$, all functions corresponding to 2^l-by-2^l Hadamard matrices can be used to extract a single bit with bias $\leq \varepsilon$ from any two independent PRB-sources which are (l, b_1) and (l, b_2) distributed respectively.

A new result contained in this paper, resolves a problem left open in the preliminary version of this work [7]: *an extraction scheme which is efficient both in terms of information rate and computation complexity.* The core of the new method is the discrete logarithm function, and its analysis is based on the method of trigonometric sums.

1.4 Probabilistic Communication Complexity

Vazirani pointed out that "good" bit-extraction functions have high communication complexity [26]. We establish further connections between the two notions. We show that functions which can be used for extracting an almost unbiased bit from two probability-bounded sources have *linear* communication complexity in a very strong sense. It follows that almost all functions, and in particular all functions corresponding to Hadamard matrices, have linear communication complexity. This resolves Yao's open problem [29] regarding the probabilistic communication complexity of random functions and of the set intersection function. (Related lower bounds on the communication complexity of random functions were presented independently by Alon, Frankl and Rödl [3] and by Orlitsky and El-Gamal [18]. Our linear ($\Omega(n)$) lower bound on the inner product modulo 2 function, improves over Vazirani's $\Omega(n/\log n)$ bound presented in [26].)

Another contribution in the field of communication complexity is the presentation of definitions and results for the case that the inputs are taken from probability-bounded distributions (i.e. distributions in which no string is too likely). This contribution is in the spirit of Vazirani's suggestion to analyze the communication complexity with respect to inputs chosen by a SV-model [26]. However, we feel that probability-bounded distributions are more natural in the context of communication complexity. We consider *randomized* protocols where the objective is to guess the value of the function with *average* success probability exceeding $\frac{1}{2}+\varepsilon$. Both the average length of a run and the average success probability are taken with respect to the "best" (for the protocol) probability-bounded distribution. We show that, even with respect to such protocols and distributions, the *average* communication complexity of almost all functions is linear in the probability bound b (where no input appear with probability greater than 2^{-b}).

1.5 On the Robustness of BPP

The class R [1] and its symmetric version BPP [10] consist of problems which can be solved with high probability in polynomial time. The probability is taken over the tosses of an unbiased coin. Umesh Vazirani raised the question whether BPP problems can be efficiently solved if a (single) SV-source is producing the coin tosses. Recently, Vazirani and Vazirani have answered this question affirmatively [27]. In this paper, we generalize their result by showing that BPP problems can be efficiently solved if a (single) PRB-source is producing the coin tosses. The underlying principles of our proof originate from Vazirani and Vazirani [27], but our proof is significantly simpler.

The main idea of the proof is that while a single PRB-source is useless for producing a *single* unbiased bit, it can nevertheless be used for producing polynomially many bits, most of which are unbiased. Our key observation is that *any* function which extracts almost unbiased bits from any two independent PRB-sources, can be used for this purpose.

1.6 Organization

In Section 2, we present our basic definitions and results concerning the extraction of unbiased bits from sources of weak randomness. These results are the basis for the rest of the paper. Subsection 2.1 consists of definitions. In subsection 2.2 we present impossibility results. In subsection 2.3 we introduce the notion of flat distributions and demonstrate its importance. In subsection 2.4 we show that almost all functions extract unbiased bits from any two independent PRB-sources, and in subsection 2.5 we show that functions corresponding to Hadamard matrices also perform well.

Each of the next three sections is based on Section 2 only, and can be read indepedently of the others. In Section 3, we further study the problem of extracting unbiased bits from probability-bounded sources. In subsection 3.1 we analyze extraction schemes with respect to two efficiency measures: rate and computation complexity. In subsection 3.2 we present and analyze the "discrete logarithm" extraction scheme. In subsection 3.3 we consider extraction from slightly dependence sources. In subsection 3.4 we consider various extensions of our model and results.

In Section 4, we present results concerning probabilistic communication complexity. In subsection 4.1 we present old ans new definitions of probabilistic communication complexity. In subsection 4.2 we prove linear lower bounds on the communication complexity of functions, and in subsection 4.3 we present almost matching upper bounds. In subsection 4.4 we suggest and investigate a robust notion of communication complexity.

In Section 5, we deal with the robustness of BPP with respect to probability-bounded sources. Conclusions and open problems appear in Section 6.

6. CONCLUSIONS

We have presented a new model of sources of weak randomness, and distilled its "hard core": the class of probability-bounded distributions. Probability-bounded distributions constitute a natural and wide class, which is convenient to analyze and yields strong results.

ACKNOWLEDGMENTS

First of all, we wish to thank Baruch Awerbuch for collaborating with us at the early stages of this research. We are very grateful to Noga Alon and László Babai for sharing with us some of their combinatorial knowledge. Noga Alon has referred us to the properties of Hadamard matrices, which turned out to be very suitable for our purposes. László Babai has referred us to the k-th residues construction used in section 3. This construction is the basis for the extraction scheme which is efficient in both rate and computation measures.

We are grateful to Jeff Lagarias for suggesting to consider "extraction at the limit", Nati Linial for pointing out Theorem 4, Mike Luby for pointing out that the entropy bound does suffice, Andrew Odlyzko for referring us to the Paley Graph Conjecture, and Avi Wigderson for his interpretation of the result in [27].

We also wish to thank Reuven Bar-Yehuda, Shimon Even, Michael Fischer, Michali Gereb, Shafi Goldwasser, Abraham Lempel, Leonid Levin, Silvio Micali, Ron Rivest, David Shmoys, and Umesh Vazirani for very helpful discussions.

REFERENCES

[1] Adleman, L., "Two Theorems on Random Polynomial Time", *Proc. 19th FOCS*, Oct. 1978, pp. 75-83.

[2] Ajtai, M., L. Babai, P. Hajnal, J. Komolós, P. Pudlák, V. Rödl, E. Szemerédi, and G. Turán, "Two Lower Bounds for Branching Programs", *Proc. 18th STOC*, May 1986, pp. 30-38.

[2] Alon, N., private communication (1985).

[3] Alon, N., P. Frankl, and V. Rödl, "Geometric Realization of Set Systems and Probabilistic Communication Complexity", *Proc. 26the FOCS*, Oct. 1985, pp. 277-280.

[4] Blum, M., "Independent Unbiased Coin Flips from a Correlated Biased Source: a Finite State Markov Chain", *Proc. 25th FOCS*, Oct. 1984, pp. 425-433.

[5] Bondy, J.A., and U.S.R. Murty, *Graph Theory with Applications*, American Elsevier Publishing Co., Inc, (1976).

[6] Canfield, E.R., P. Erdös, and C. Pomerance, "On a Problem of Oppenheim Concerning "Factorisatio Numerorum"," *Jour. of Number Theory*, Vol. 17, No. 1, 1983, pp. 1-28.

[7] Chor, B., and O. Goldreich, "Unbiased Bits from Sources of Weak Randomness and Probabilistic Communication Complexity", *Proc. 26th FOCS*, Oct. 1985, pp. 429-442.

[8] Elias, P., "The Efficient Construction of an Unbiased Random Sequence", *Ann. Math. Statist.*, Vol. 43, No. 3, 1972, pp. 865-870.

[9] Erdös, P., and J. Spencer, *Probabilistic Methods in Combinatorics*, Academic Press, (1974).

[10]Gill, J., "Complexity of Probabilistic Turing Machines", *SIAM Jour. on Computing*, Vol. 6, No. 4, 1977, pp. 675-695.

[11]Hall, M. Jr., *Combinatorial Theory*, Blaisdell Publishing Co., (1967).

[12]Hardy, G.H., and E.M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, (1960).

[13]Ja'Ja', J., V.K. Prasanna Kumar, and J. Simon, "Information Transfer under Different Sets of Protocols", *SIAM Jour. on Computing*, Vol. 13, No. 4, 1984, pp. 840-849.

[14]Johnson, N.J., and S. Kotz, *Distributions in Statistics*, Vol. 1, *Discrete Distributions*, John Wiley & Sons, 1969.

[15]McWilliams, F.J., and N.J.A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland Publishing Company, 1977.

[16]von Neumann, J., "Various Techniques Used in Connection with Random Digits", notes by Forsythe G.E., National Bureau of Standards, Applied Math. Series, 1951, Vol. 12, pp. 36-38. Reprinted in *von Neumann's Collected Works*, Vol. 5, Pergamon Press (1963), pp. 768-770.

[17]Odlyzko, A.M., "Discrete Logarithms in Finite Fields and their Cryptographic Significance", *Advances in Cryptology: Proceedings of EuroCrypt84*, T. Beth et al., eds., Springer–Verlag, 1985, pp. 224–314.

[18]Orlitsky, A., and A. El Gamal, "Randomized Communication Complexity", preprint, (1985).

[19]Papadimitriou, C.H., and M. Sipser, "Communication Complexity", *Jour. of Comp. and Sys. Sci.*, Vol. 28, No. 2, 1984, pp. 260-269.

[20]Papadimitriou, C.H., and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Inc. (1984).

[21]Paturi, R., and J. Simon, "Probabilistic Communication Complexity", *Proc. 25th FOCS*, Oct. 1984, pp. 118-126.

[22]Pohlig, R.C. and M. Hellman, "An improved algorithm for computing logarithms over $GF(p)$ and its cryptographic significance", *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 106–110, 1978.

[23]Rényi, A., *Probability Theory*, North-Holland Publishing Company (1970).

[24]Santha, M., and U.V. Vazirani, "Generating Quasi-Random Sequences from Slightly-Random Sources", *Proc. 25th FOCS*, Oct. 1984, pp. 434-440.

[25]Schmidt, W.M., *Equations over Finite Fields: An Elementary Approach*, Lecture Notes in Mathematics, Vol. 536, Springer-Verlag, Berlin, 1976.

[26]Vazirani, U.V., "Towards a Strong Communication Complexity Theory or Generating Quasi-Random Sequences from Two Communicating Slightly-Random Sources", *Proc. 17th STOC*, May 1985, pp. 366-378.

[27]Vazirani, U.V., and V.V. Vazirani, "Random Polynomial Time is Equal to Slightly-random Polynomial Time", *Proc. 26th FOCS*, Oct. 1985, pp. 417-428.

[28]Yao, A.C., "Probabilistic Computations: Towards a Unified Measure of Complexity", *Proc. 18th FOCS*, Oct. 1977, pp. 222-227.

[29]Yao, A.C., "Some Complexity Questions related to Distributive Computing", *Proc. 11th STOC*, April 1979, pp. 209-213.

[30]Yao, A.C., "Lower Bounds by Probabilistic Arguments", *Proc. 24th FOCS*, Nov. 1983, pp. 420-428.