Kolmogorov Complexity in Randomness Extraction

John M. Hitchcock[1]*, **A. Pavan**[2]†, **N. V. Vinodchandran**[3]‡

[1]Department of Computer Science, University of Wyoming
jhitchco@cs.uwyo.edu

[2]Department of Computer Science, Iowa State University
pavan@cs.iastate.edu

[3]Department of Computer Science and Engineering, University of Nebraska-Lincoln
vinod@cse.unl.edu

ABSTRACT.
We clarify the role of Kolmogorov complexity in the area of randomness extraction. We show that a computable function is an almost randomness extractor if and only if it is a Kolmogorov complexity extractor, thus establishing a fundamental equivalence between two forms of extraction studied in the literature: Kolmogorov extraction and randomness extraction. We present a distribution $\mathcal{M}_k$ based on Kolmogorov complexity that is complete for randomness extraction in the sense that a computable function is an almost randomness extractor if and only if it extracts randomness from $\mathcal{M}_k$.

1 Introduction

The problem of extracting pure randomness from weak random sources has received intense attention in the last two decades producing several exciting results. The main goal in this topic is to give explicit constructions of functions that are known as *randomness extractors;* functions that output almost pure random bits given samples from a weak source of randomness which may be correlated and biased. Randomness extractors have found applications in several areas of theoretical computer science including complexity theory and cryptography. The body of work on randomness extractors is vast and we do not attempt to list them here. Instead, we refer the readers to survey articles by Nisan and Ta-Shma [10] and Shaltiel [13], and Rao's thesis [11] for an extensive exposition on the topic (with the caveat that some of the recent advances are not reported in these articles).

We will focus on a type of randomness extractors known as *multi-source* extractors. These are multi-input functions with the property that if the inputs come from independent distributions with certain guaranteed randomness, typically measured by their *minentropy*, then the output distribution will be close to the uniform distribution. A distribution over n-bit strings is said to have minentropy k, if any element in the support of the distribution

*This research was supported in part by NSF grants 0515313 and 0652601. Part of this research work done while the author was on sabbatical at CWI.

†This research was supported in part by NSF grants 0830479 and 0916797.

‡This research was supported in part by NSF grants 0830730 and 0916525.

Foundations of Software Technology and Theoretical Computer Science (Kanpur) 2009.
Editors: Ravi Kannan and K. Narayan Kumar; pp 215–226
Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany.
Digital Object Identifier: 10.4230/LIPIcs.FSTTCS.2009.2320

has a probability $\leq 2^{-k}$. A function $f : \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\}^m$ is a 2-source extractor for minentropy k if for any two independent distributions X and Y on $\{0,1\}^n$ with minentropy k, the output $f(X,Y)$ is statistically close to the uniform distribution. It is known that such extractors exist for all minentropy levels with optimal parameters [3, 4], but explicitly constructing 2-source extractors for sources with low minentropy is a very active research question.

While minentropy characterizes the amount of randomness present in a probability distribution, Kolmogorov complexity characterizes the amount of randomness present in *individual strings*. The Kolmogorov complexity of a string x, denoted by $K(x)$, is the the length of the shortest program that outputs x. If $K(x) = m$, then x can be viewed as containing m bits of randomness. A string x is *Kolmogorov random* if its Kolmogorov complexity is close to the length of x. A natural notion that arises is that of *Kolmogorov extractors*: explicit functions that extract Kolmogorov complexity from strings that need not be Kolmogorov random. More formally, a 2-string Kolmogorov extractor for complexity k is a function $f : \Sigma^n \times \Sigma^n \rightarrow \Sigma^m$ such that $K(f(x,y))$ is close to m whenever $K(x), K(y) \geq k$ and x and y are Kolmogorov independent ($K(xy) \simeq K(x) + K(y)$). Kolmogorov extractors have recently been of interest to researchers [1, 5, 14, 15]. One of the main observations that emerged from this research is that a randomness extractor is also a Kolmogorov extractor. In particular, in [5], the authors show that the construction due to Barak, Impagliazzo and Wigderson [2] of a multisource extractor is also a Kolmogorov extractor. Zimand takes this approach further and gives constructions of Kolmogorov extractors in other settings [14, 15]. Thus, this line of research uses randomness extractors as a tool in Kolmogorov complexity research. However, the role of Komogorov complexity in the area of randomness extraction has not yet been explored by researchers. We take a step in this direction.

We ask the following question. Is it true that a Kolmogorov extractor is also a randomness extractor? While randomness extractors concern information-theoretic randomness, Kolmogorov extractors concern computational randomness. Thus intuitively it appears that Kolmogorov extractors are weaker objects than randomness extractors. Moreover, if we use the strict definition of extraction, it is easy to come up with a counterexample to this converse. Let f be a Kolmogorov extractor, then $f \circ 1$ (output of f concatenated with bit 1) is also a Kolmogorov extractor. But $f \circ 1$ is not a randomness extractor for any function f because it never outputs 50% of the strings - strings that end with 0. The reason for this counterexample is that any Kolmogorov complexity measure is precise only up to a small additive term. Consequently, a string x of length n is considered Kolmogorov random even if its Kolmogorov complexity is only $n - a(n)$ for a slow growing function $a(n)$ such as a constant multiple of $\log n$ [5]. Thus a more fruitful question is to ask whether a Kolmogorov extractor is also an *almost* randomness extractor. An almost randomness extractor is like a traditional randomness extractor except that we only require the output of an almost extractor to be close to a distribution with minentropy $m - O(\log n)$. For a traditional extractor, the output has to be close to the uniform distribution - the only distribution with minentropy m. Such almost extractors have been considered in the literature (see for example [12]).

Our first contribution is to show an equivalence between Kolmogorov extraction and the above-mentioned slightly relaxed notion of randomness extraction. The following statement is very informal and Section 3 is devoted to giving a precise statement with a proof.

RESULT 1. *A computable function f is a Kolmogorov extractor if and only if f is an almost randomness extractor.*

A randomness extractor is a universal object in the sense that it should extract randomness from *all* distributions with certain minentropy. Can this universality be shifted to a distribution? That is, is there a distribution D so that a computable function f is an extractor if and only if f extracts randomness from D? We call such a distribution a *complete* distribution for randomness extraction. Kolmogorov complexity has proved to be useful in the discovery of distributions with a similar universality property in other areas of computer science including average-case analysis [8] and learning theory [7].

Our second contribution is to present a complete distribution, based on Kolmogorov complexity, for randomness extraction. Fix an input length n. For a number k consider the distribution $\mathcal{M}_k$ that puts uniform weight on all strings of length n with Kolmogorov complexity $\leq k$. Motivated by the proof of our first result we show that the distribution $\mathcal{M}_k$ is a complete distribution for almost extractors. The following statement is informal and the full details are in Section 4.

RESULT 2. *For any k, there is a $k' = k + O(\log n)$ so that $\mathcal{M}_{k'}$ is complete for almost extractors with minentropy parameter k.*

2 Preliminaries, Definitions, and Basic Results

Kolmogorov Extractors

We only review the essentials of Kolmogorov complexity and refer to the textbook by Li and Vitányi [9] for a thorough treatment of the subject. For a string $x \in \{0,1\}^*$, $l(x)$ denotes the length of x. We use the following standard encoding function where a pair $\langle x, y\rangle$ is encoded as $1^{l(l(x))}0l(x)xy$. By viewing $\langle x, y, z\rangle$ as $\langle x, \langle y, z\rangle\rangle$, this encoding can be extended to 3-tuples (and similarly for any k-tuple).

Let U be a universal Turing machine. Then for any string $x \in \{0,1\}^*$, the Kolmogorov complexity of x is defined as

$$K(x) = \min\{l(p) \mid U(p) = x\},$$

that is, the length of a shortest program p that causes U to print x and halt. If we restrict the set of programs to be prefix-free, then the corresponding measure is known as prefix-free Kolmogorov complexity. These two complexity measures only differ by an additive logarithmic factor. We will work with the above-defined standard measure. Since we are flexible about additive logarithmic factors in this paper, our results will hold with the prefix-free version also.

Kolmogorov extractors are computable functions which convert strings that have a guaranteed amount of Kolmogorov complexity into a Kolmogorov random string. We give a general definition of Kolmogorov extractors involving a parameter for *dependency* between the input strings. Consequently, instead of aiming for maximum complexity in the output string, we will consider extractors which lose an additive factor equal to the dependency in the inputs. The following notion of dependency we use is equivalent to the well-studied notion of *mutual information* in the Kolmogorov complexity literature up to an additive log factor. However, we prefer to use the term dependency in this paper.

DEFINITION 1.[Dependency] *For two strings x and y of the same length, the* dependency between x and y *is*

$$\mathrm{dep}(xy) = K(x) + K(y) - K(xy).$$

DEFINITION 2.[Kolmogorov Extractor] *An* $(n, m(n), k(n), \alpha(n))$ Kolmogorov extractor *is a uniformly computable family $\{f_n\}_n$ of functions $f_n : \Sigma^n \times \Sigma^n \to \Sigma^{m(n)}$ where there is a constant c such that for all n, for all $x, y \in \Sigma^n$ with $K(x) \geq k(n)$, $K(y) \geq k(n)$, and* $\mathrm{dep}(xy) \leq \alpha(n)$, *we have*

$$K(f_n(x,y)) \geq m(n) - \mathrm{dep}(xy) - c \log n.$$

The computability restriction is required to make the definition nontrivial. Otherwise it is easy to come up with Kolmogorov extractors: for any pair of inputs at length n, just output a fixed string of length $m(n)$ that has maximal Kolmogorov complexity.

Randomness Extractors

Randomness extractors are functions which convert weak random sources to a distribution that is statistically close to the uniform distribution. A weak random source is characterized by its minentropy which is defined as follows.

DEFINITION 3. *For a probability distribution X over a universe S, the* minentropy *of X is*

$$-\log\left(\max_{s \in S} X(s)\right) = \min_{s \in S}\left(\log \frac{1}{X(s)}\right).$$

Here we are writing $X(s)$ for the probability that distribution X assigns to outcome s. For an event $T \subseteq S$, $X(T) = \sum_{s \in T} X(s)$ is the probability of T under X.

DEFINITION 4. *For any two distributions X and Y on a universe S, their* statistical distance $|X - Y|$ *is*

$$|X - Y| = \max_{T \subseteq S} |X(T) - Y(T)| = \frac{1}{2} \sum_{s \in S} |X(s) - Y(s)|$$

. If $|X - Y| \leq \epsilon$, we say X and Y are ϵ-close to each other.

DEFINITION 5.[Almost Randomness Extractor] *An* $(n, m(n), k(n), \epsilon(n))$ almost randomness extractor *is a family $\{f_n\}_n$ of functions $f_n : \Sigma^n \times \Sigma^n \to \Sigma^{m(n)}$ where there is a constant c such that for all n, for every pair of independent distributions X and Y over Σ^n with minentropy at least $k(n)$, the distribution $f_n(X, Y)$ is $\epsilon(n)$-close to a distribution with minentropy at least $m(n) - c \log n$. Moreover, f is uniformly computable.*

A distribution X over Σ^n is called a *flat distribution* if it is uniform over some subset of Σ^n. For a flat distribution X, we will use X also to denote the support of the distribution X. The following useful theorem due to Chor and Goldreich [3] states that every function that extracts randomness from flat distributions is a randomness extractor.

THEOREM 6.[3] *Let f be a function from $\Sigma^n \times \Sigma^n$ to Σ^m. Suppose for every pair of independent flat distributions X and Y with minentropy k, $f(X,Y)$ is ϵ-close to having minentropy $m - c\log n$. Then f is a (n,m,k,ϵ) almost randomness extractor.*

Let D be a distribution over Σ^m induced by a distribution over $\Sigma^n \times \Sigma^n$. We call D a *nice distribution* if for all $z \in \Sigma^m$, $D(z)$ is a rational number of the form p/q with $q \leq 2^{2n}$. This restriction allows us to effectively cycle through all nice distributions. For any distribution D with minentropy k, there is a nice distribution D' with the same minentropy so that the statistical distance between D and D' is at most $1/2^n$. Because of this we will assume that distribution are nice whenever necessary.

The following lemma due to Guruswami, Umans, and Vadhan [6] is useful to obtain a bound on the minentropy of a distribution. We will state it for nice distributions although the original statement and the proof do not have such a restriction. Their proof can be easily modified to prove this case also.

LEMMA 7.[6] *Let D be a nice distribution and s be an integer. Suppose that for every set S of size s, $D(S) \leq \epsilon$. Then D is ϵ-close to a nice distribution with minentropy at least* $\log(s/\epsilon)$.

Remarks and Clarifications

Although it is typical requirement for the extractors to be *efficiently* computable, the only requirement we need in our proofs is that the extractors are computable. Hence, we will not mention any resource restrictions here. Here we only focus on extractors with 2 inputs. The connection we prove here also holds for extractors with k inputs for any constant $k \geq 2$ with identical proofs. Although the parameters in the definition of the extractors depend on the input length n, we will omit it in the rest of the paper. For instance, a $(n, m(n), k(n), \alpha(n))$ Kolmogorov extractor will be denoted as an (n,m,k,α) extractor. In addition, we also assume that the parameters that depend on input length n are computable functions of n. Finally, henceforth by a randomness extractor we mean an almost randomness extractor unless otherwise mentioned.

Why is there a dependency parameter in the definition of Kolmogorov extractor? Our aim is to establish a tight connection between randomness extractors and Kolmogorov extractors. Randomness extractors typically have four parameters; input length n, output length m, minentropy bound k, and the error parameter ϵ. Except for the error parameter, there is an obvious mapping of parameters between Kolmogorov and randomness extractors. But there appears to be no natural notion of "error" in Kolmogorov extraction. What is a choice for the parameter in the definition of Kolmogorov extractor analogous to the error parameter? Our theorems indicate that the dependency is a good choice.

3 The Equivalence

3.1 Kolmogorov Extractor is a Randomness Extractor

In this subsection we show that for appropriate settings of parameters, a Kolmogorov extractor is also a randomness extractor. First we will give a simple argument for the special case when the dependency parameter is $O(\log n)$. In this case we get a inverse polynomial error for the randomness extractor. We will only give a sketch of the proof since the subsequent theorem for the general case subsumes this case.

A Special Case

The proof of this special case is a simple application of the following well known coding theorem.

THEOREM 8.*[Coding Theorem] Let D be a probability distribution over $\{0,1\}^*$ that is computable by a program P, there is a constant c such that*

$$\frac{1}{2^{K(x)}} \geq \frac{c}{2^{|P|}} D(x).$$

THEOREM 9. *Let f be a (n, m, k, α) Kolmogorov extractor with $\alpha = O(\log n)$. Then f is a (n, m, k', ϵ) almost randomness extractor where $k' = k + O(\log n)$ and $\epsilon = 1/\texttt{poly}(n)$.*

PROOF. We provide a proof sketch. Let c be the constant associated with the Kolmogorov extractor f. That is, $K(f(x,y)) \geq m - c\log n - \mathrm{dep}(xy)$ provided $K(x) \geq k$, $K(y) \geq k$, and $\mathrm{dep}(xy) \leq \alpha$.

We will show that for every pair of flat distributions X and Y with minentropy k', $f(X,Y)$ is ϵ-close to a nice distribution with minentropy at least $m - (c+6)\log n$. Then by Theorem 6, it will follow that f is an almost randomness extractor for minentropy k'. For the purpose of contradiction, suppose there are flat distributions X and Y with minentropy k' so that $f(X,Y)$ is ϵ far from all nice distributions with minentropy at least $m - (c+6)\log n$. Let X and Y be the first such distributions (in some fixed ordering of distributions).

The number of flat distributions with minentropy k' is finite, and the number of nice distributions over Σ^m with minentropy at least $m - (c+6)\log n$ is also finite. Thus there is a program p which given n as input, produces the distributions X and Y. Thus the size of p is at most $2\log n$ for large enough n. Let D denote the distribution $f(X,Y)$.

The idea of the rest proof is as follows. Consider the following set S.

$$S = \{\langle x,y\rangle \in X \times Y \mid K(x) \geq k, K(y) \geq k, \text{ and } \mathrm{dep}(xy) \leq 10\log n\}.$$

First using a simple counting argument it is easy to show that S is a large set and hence probability of the complement of S with respect to $X \times Y$ is small. Since f is a Kolmogorov extractor, for all elements $(x,y) \in S$, $K(z)$ is close to m where $z = f(x,y)$. Since D is computable, by the coding theorem, it follows that $D(z) \leq 1/2^{m-O(\log n)}$. Thus, except for a small fraction of strings in $f(\overline{S})$, the strings in the range of f satisfies the minentropy condition. Hence D must be close to a distribution with minentropy $m - c\log n$. ∎

The General Case

We now state and prove the theorem for a general setting of parameters. The proof follows the line of argument of the proof of the special case. But we will use Lemma 7 instead of the coding theorem.

THEOREM 10. *Let f be a (n,m,k,α) Kolmogorov extractor. Then f is a (n,m,k',ϵ) almost randomness extractor where*

(a) if $k'-k > \alpha - 4\log n + 1$, then $\epsilon \le \frac{1}{2^{\alpha-4\log n-1}}$.

(b) if $k'-k \le \alpha - 4\log n + 1$, then $\epsilon \le \frac{1}{2^{k'-k-2}}$.

PROOF. Let c be the constant associated with the Kolmogorov extractor f. That is, $K(f(x,y)) \ge m - c\log n - \text{dep}(xy)$ provided $K(x) \ge k$, $K(y) \ge k$, and $\text{dep}(xy) \le \alpha$.

We will show that for every pair of flat distributions X and Y with minentropy k', $f(X,Y)$ is ϵ-close to a nice distribution with minentropy at least $m-(c+9)\log n$ where ϵ is as given in the statement of the theorem. Then by Theorem 6, it will follow that f is an almost randomness extractor for minentropy k'. For the purpose of contradiction, suppose there are flat distributions X and Y with minentropy k' so that $f(X,Y)$ is ϵ far from all nice distribution with minentropy at least $m-(c+9)\log n$. Let X and Y be the first such distributions (in some fixed ordering of distributions). For simplicity, we will denote the supports of distributions X and Y also by X and Y, respectively. Let D denote the distribution $f(X,Y)$. D is a nice distribution.

The number of flat distributions with minentropy k' is finite, and the number of nice distributions over Σ^m with minentropy at least $m-(c+9)\log n$ is also finite. Thus there is a program p which given n,c and a code for f as input, produces the flat distributions X and Y by brute-force search method. The size of p is at most $2\log n$ for large enough n. We will split the rest of the proof into two cases.

Case (a). $k'-k > \alpha - 4\log n + 1$.

Define the "good set" S as

$$S = \{\langle x,y\rangle \in X\times Y \mid K(x)\ge k, K(y) \ge k, \text{ and } \text{dep}(xy) \le \alpha\}.$$

Let S' be the compliment of S within $X\times Y$. That is $S' = X\times Y\setminus S$. We will bound the size of S'. Observe that S' is a subset of the union of following sets:

$$S_1 = \{\langle x,y\rangle \in X\times Y \mid K(x) < k\},$$

$$S_2 = \{\langle x,y\rangle \in X\times Y \mid K(y) < k\},$$

$$S_3 = \{\langle x,y\rangle \in X\times Y \mid \text{dep}(xy) > \alpha\}.$$

Clearly, sizes of S_1 and S_2 are bounded by $2^{k+k'}$. We will bound $|S_3|$. Since the program p produces X and Y and $|X|=|Y|=2^{k'}$, every string in $X\cup Y$ has Kolmogorov complexity at most $k'+2\log n$. Thus for any $\langle x,y\rangle\in S_3$ we have that $K(xy) = K(x)+K(y)-\text{dep}(xy) \le 2k'+4\log n-\alpha$. So $|S_3| \le 2^{2k'+4\log n-\alpha}$. Hence $|S'| \le| S_1\cup S_2\cup S_3| \le| S_1|+|S_2|+|S_3| \le 2^{k+k'+1}+2^{2k'+4\log n-\alpha}$. Since $k'-k>\alpha-4\log n+1$, this sum is $\le 2^{2k'+4\log n-\alpha+1}$. Thus we have the following bound on the probability of S'.

CLAIM 11. *If $k'-k > \alpha-4\log n+1$ then $\Pr_{X\times Y}(S') \le \frac{1}{2^{\alpha-4\log n-1}}$*

We assumed that f is not an almost randomness extractor. That is the distribution is ϵ-far from any nice distribution with minentropy $m-(c+9)\log n$. By Lemma 7, there is a set $U\subseteq\Sigma^m$ of size $2^{m-\alpha-(c+4)\log n}$ such that $D(U) > 1/2^{\alpha-5\log n}$. Since a program of size

$2\log n$ produces distributions X and Y and f is computable, there is a program of size at most $3\log n$ that produces the set U. Thus for all $u \in U$, $K(u) < m - \alpha - c\log n$.

Since $\Pr_{X\times Y}(S') \leq \frac{1}{2^{\alpha-4\log n-1}} \leq \frac{1}{2^{\alpha-5\log n}}$ and $D(U) > \frac{1}{2^{\alpha-5\log n}}$, there must exist a tuple $\langle x,y\rangle \in S$ so that $f(x,y) \in U$ and for this tuple we have $K(f(x,y)) < m - \alpha - c\log n$. This is a contradiction since f is a Kolmogorov extractor and for all elements $\langle x,y\rangle \in S$, $K(f(x,y)) \geq m - \mathrm{dep}(xy) - c\log n \geq m - \alpha - c\log n$.

Case (b). $k' - k \leq \alpha - 4\log n + 1$.

The proof is very similar. Define the "good set" S as

$$S = \{\langle x,y\rangle \in X \times Y \mid K(x) \geq k, K(y) \geq k, \text{ and } \mathrm{dep}(xy) \leq k' - k + 4\log n\}.$$

In this case, we can bound the size of S' (the compliment of S within $X \times Y$) by considering the following sets.

$$S_1 = \{\langle x,y\rangle \in X \times Y \mid K(x) < k\},$$

$$S_2 = \{\langle x,y\rangle \in X \times Y \mid K(y) < k\},$$

$$S_3 = \{\langle x,y\rangle \in X \times Y \mid \mathrm{dep}(xy) > k' - k + 4\log n\}.$$

Sizes of S_1 and S_2 are bounded by $2^{k+k'}$. We will bound $|S_3|$. Since the program p produces X and Y and $|X| = |Y| = 2^{k'}$, every string in $X \cup Y$ has Kolmogorov complexity at most $k' + 2\log n$. Thus for any $\langle x,y\rangle \in S_3$ we have that $K(xy) = K(x) + K(y) - \mathrm{dep}(xy) \leq 2k' + 4\log n - (k' - k + 4\log n) = k' + k$. So $|S_3| \leq 2^{k'+k}$. Hence $|S'| \leq| S_1| + |S_2| + |S_3| \leq 2^{k+k'+1} + 2^{k'+k} \leq 2^{k+k'+2}$. Thus in this case we have the following bound on the probability of S'.

CLAIM 12. *If* $k' - k \leq \alpha - 4\log n + 1$ *then* $\Pr_{X\times Y}(S') \leq \frac{1}{2^{k'-k-2}}$

We assumed that distribution D is ϵ-far from any nice distribution with minentropy $m - (c+9)\log n$. By Lemma 7, there is a set $U \subseteq \Sigma^m$ of size $2^{m-(k'-k+4\log n)-(c+4)\log n}$ such that $D(U) > 1/2^{k'-k-\log n}$. Since a program of size $2\log n$ produces distributions X and Y and f is computable, there is a program of size at most $3\log n$ that produces the set U. Thus for all $u \in U$, $K(u) < m - (k' - k + 4\log n) - c\log n$. But since $\Pr_{X\times Y}(S') \leq \frac{1}{2^{k'-k-2}} \leq \frac{1}{2^{k'-k-\log n}}$ and $D(U) > \frac{1}{2^{k'-k-\log n}}$, there must exist a tuple $\langle x,y\rangle \in S$ so that $f(x,y) \in U$. This contradicts the fact that f is a Kolmogorov extractor with the prescribed parameters.

∎

3.2 Randomness Extractor is a Kolmogorov Extractor

In this subsection we show that an almost randomness extractor is also a Kolmogorov extractor. We follow the line of proof presented in [5] where it is shown that the construction of a multisource extractor in [2] is also a Kolmogorov extractor. Here we note that in fact the argument works even for almost randomness extractors.

THEOREM 13. *An (n,m,k,ϵ) almost extractor is also a (n,m,k',α) Kolmogorov extractor for $\alpha < \log\frac{1}{\epsilon} - 6\log n$ and $k' = k + 3\log n$.*

PROOF. Let $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ be an (n,m,k,ϵ) almost extractor. Let c be the the associated constant. That is, the minentropy guarantee of the output of f is $m - c\log n$.

Let x_1 and x_2 be two strings with $K(x_1) = k_1 \geq k'$, $K(x_2) = k_2 \geq k'$ and $\mathrm{dep}(x_1x_2) \leq \alpha$. Let X_1 and X_2 be subsets of $\{0,1\}^n$ with Kolmogorov complexity at most k_1 and k_2 respectively. That is, $X_1 = \{x \in \{0,1\}^n | K(x) \leq k_1\}$ and $X_2 = \{x \in \{0,1\}^n | K(x) \leq k_2\}$. We will also use X_1 and X_2 to denote the flat distributions that put uniform weight on sets X_1 and X_2 respectively (in the next section, we give specific notation for these distributions).

For $t = m - \mathrm{dep}(x_1x_2) - (c+6)\log n$, let $T \subseteq \{0,1\}^m$ be the set of strings with Kolmogorov complexity at most t. That is, $T = \{z \mid K(z) < t\}$. We will show that for all u, v so that $f(u,v) \in T$, $K(uv) < k_1 + k_2 - \mathrm{dep}(x_1x_2)$. This will show the theorem as this will imply $f(x_1,x_2) \notin T$ and hence $K(f(x_1,x_2)) > m - \mathrm{dep}(x_1x_2) - (c+6)\log n$.

CLAIM 14. *For all $u \in X_1$ and $v \in X_2$ so that $f(u,v) \in T$, $K(uv) < k_1 + k_2 - \mathrm{dep}(\mathrm{x}_1\mathrm{x}_2)$.*

PROOF. It is clear that $|X_i| \leq 2^{k_i}$. Since each string in the set $0^{(n-k)}\{0,1\}^k$ has Kolmogorov complexity $\leq k + 2\log n + O(\log\log n) \leq k_i$ (for large enough n), we also have that $|X_i| \geq 2^k$. Thus $\Pr_{X_i}(x) \leq \frac{1}{2^k}$ for any $x \in X_i$, X_i has minentropy at least k and f works for $X_1 \times X_2$.

Consider the output distribution $f(X_1, X_2)$ on $\{0,1\}^m$. Let us call this distribution D. Since f is an almost extractor the distribution D is ϵ-close to a distribution with minentropy $m - c\log n$.

Since $|T| \leq 2^t = 2^{m-\mathrm{dep}(x_1x_2)-(c+6)\log n}$ and D is ϵ-close to a distribution with minentropy $m - c\log n$, we have the following.

$$\begin{aligned}\Pr_D(T) &\leq \frac{|T|}{2^m} \times n^c + \epsilon \\ &\leq 2^{-\mathrm{dep}(x_1x_2)-6\log n} + 2^{-\alpha-6\log n} \\ &\leq 2^{-\mathrm{dep}(x_1x_2)-6\log n+1}\end{aligned}$$

The last two inequalities follow because $\alpha \leq \log(\frac{1}{\epsilon}) - 6\log n$ and $\mathrm{dep}(x_1x_2) \leq \alpha$.

Consider the set $S = f^{-1}(T) \cap X_1 \times X_2 \subseteq \{0,1\}^n \times \{0,1\}^n$. We will first bound $|S|$. Every tuple from S gets a weight of $\geq 1/2^{k_1+k_2}$ according to the joint distribution $X_1 \times X_2$. Thus we have

$$\begin{aligned}\frac{|S|}{2^{k_1+k_2}} &\leq \Pr_{(X_1,X_2)}(S) \\ &= \Pr_D(T) \\ &\leq (2^{-\mathrm{dep}(x_1x_2)-6\log n+1})\end{aligned}$$

Hence $|S| \leq 2^{k_1+k_2-\mathrm{dep}(x_1x_2)-6\log n+1}$.

The sets X_1, X_2, and T are recursively-enumerable and f is computable. Hence there is a program that given $n, k_1, k_2, \mathrm{dep}(x_1x_2)$, a code for f, and c, enumerates the elements of S. Hence for any $\langle u,v\rangle \in S$, $K(uv) \leq \log|S| + 4\log n + O(\log\log n) \leq \log|S| + 5\log n$ for large enough n. Since $|S| \leq 2^{k_1+k_2-\text{dep}(x_1x_2)-6\log n+1}$, $K(uv) < k_1 + k_2 - \text{dep}(x_1x_2)$ and the claim follows. ∎

3.3 The Error Parameter vs the Dependency Parameter

Theorem 13 suggests that there is a nice logarithmic relation between error of an almost extractor and the dependency parameter of the corresponding Kolmogorov extractor. In particular, in Theorem 13, we show that an (n, m, k, ϵ) almost randomness extractor is a (n, m, k', α) Kolmogorov extractor for $\alpha = \log(1/\epsilon) - O(\log n)$ for k' slightly larger than k ($k' = k + O(\log n)$). On the other hand, the parameters we get in the proof of the converse direction (Kolmogorov extractor ⇒ randomness extractor) are not fully satisfactory. Ideally we would like to prove that every (n, m, k, α) Kolmogorov extractor is a (n, m, k', ϵ) almost randomness extractor with $k' = k + O(\log n)$ and $\epsilon = 1/2^{\alpha - O(\log n)}$ which will be a true converse to Theorem 13. We note that this is not possible in general. In particular, we show that for a (n, m, k, α) Kolmogorov extractor to be a (n, m, k', ϵ) almost randomness extractor with $\epsilon = 2^{\alpha - O(\log n)}$, k' has to be greater than $k + \alpha$ (upto a log factor).

THEOREM 15. *Let f be a (n, m, k, α) Kolmogorov extractor. Then there exists a computable function g which is also a (n, m, k, α) Kolmogorov extractor but g is not a (n, m, k', ϵ) almost randomness extractor for $\epsilon < \frac{1}{2^{k'-k+4\log n}}$ for any k' where $k' < m + k - c\log n$ for all constants c.*

PROOF. Let f be a (n, m, k, α) Kolmogorov extractor. Consider the set $U \subseteq \{0,1\}^n$ defined as $U = \{0,1\}^{k-3\log n}0^{n-k+3\log n}$. For any string $x \in U$, $K(x) < k$. Define the function g as follows: $g(x, y) = 0^m$ if $x \in U$ and $g(x, y) = f(x, y)$ otherwise.

Since membership in the set U is easy to decide and f is computable, g is computable. Also, by definition of g, for all pair of strings x, y so that $K(x) \geq k$, $K(y) \geq k$ and $\text{dep}(x, y) \leq \alpha$, $g(x, y) = f(x, y)$. Hence g is a (n, m, k, α) Kolmogorov extractor.

Now consider two flat distributions X and Y of size $2^{k'}$ such that $U \subseteq X$. Let D denotes the distribution $g(X \times Y)$. Notice that $\Pr_D(0^m) \geq \Pr_X(x \in U) \geq \frac{1}{2^{k'-k+3\log n}}$. Now an easy calculation (omitted because of space constraints) proves the theorem. ∎

4 A Complete Distribution for Randomness Extraction

For integers k and n, let $\mathcal{M}^n_{k'}$ denote the distribution that places uniform weight on the set $\{x \in \{0,1\}^n \mid K(x) \leq k\}$. That is $\mathcal{M}^n_k$ is uniform over all the strings with Kolmogorov complexity $\leq k$. As n will be clear from the context, we will omit n from the notation and call it $\mathcal{M}_k$. We show that $\mathcal{M}_k$ is a complete distribution for randomness extraction in the sense that a computable function f is an almost randomness extractor if and only if it extracts randomness from two independent copies of $\mathcal{M}_k$.

This result is motivated by the proof of the equivalence theorem. Notice that in the proof that a randomness extractor f is also a Kolmogorov extractor, we essentially show that if f extracts randomness from the class of distributions $\{\mathcal{M}_l\}_{l \geq k}$, then it is a Kolmogorov extractor. The other implication shows that if f is a Kolmogorov extractor then it is also a randomness extractor. Thus intuitively we get that the class $\{\mathcal{M}_l\}_{l \geq k}$ is complete. Below we give a simple argument for completeness.

THEOREM 16. *A computable function f is a (n, m, k, ϵ) almost extractor if and only if there is a constant c so that $f(\mathcal{M}_{k'} \times \mathcal{M}_{k'})$ is ϵ' close to a distribution with minentropy $m - c\log n$ where $k' = k + 2\log n$ and $\epsilon' = \epsilon/n^4$.*

PROOF. The set $0^{(n-k)}\{0,1\}^k$ is a subset of $\mathcal{M}_k$ since every strings in this set has Kolmogorov complexity $\leq k + \log n + O(\log\log n) < k'$. Hence $\mathcal{M}_{k'}$ has minentropy $\geq k$ and since f is an almost extractor for minentropy k it should also extract randomness from $\mathcal{M}_{k'} \times \mathcal{M}_{k'}$.

For the other direction, let f be a function that extracts from $\mathcal{M}_{k'} \times \mathcal{M}_{k'}$. Hence there is a constant c so that $f(\mathcal{M}_{k'} \times \mathcal{M}_{k'})$ is ϵ' close to a distribution with minentropy $m - c\log n$.

For the sake of contradiction suppose f is not an almost extractor for minentropy k. Let X and Y be first two flat distributions over $\{0,1\}^n$ for which the distribution $D = f(X, Y)$ is ϵ-far from all nice distributions with minentropy $m - (c+4)\log n$. Observe that there is a program p which given n, c, and a code for f produces the distributions X and Y. Thus for any $x \in X$, we have $K(x) \leq k + \log n + O(\log\log n) \leq k'$. Similarly for $y \in Y$. Hence we have the following claim.

CLAIM 17. *For all $x \in X$, $K(x) \leq k'$. Similarly for all $y \in Y$, $K(y) \leq k'$. Hence $X \subseteq \mathcal{M}_{k'}$ and $Y \subseteq \mathcal{M}_{k'}$.*

We will show that for all $T \subseteq \{0,1\}^m$, $\Pr_D(T) \leq \frac{|T|}{2^m} \times n^{c+4} + \epsilon$. This suffices to show that D is ϵ-close to a distribution with minentropy $m - (c+4)\log n$.

$$\begin{aligned}
\Pr_D(T) &= \Pr_{X\times Y}(f^{-1}(T) \cap X \times Y) \\
&= \frac{|f^{-1}(T) \cap X \times Y|}{2^{2k}} \\
&\leq \Pr_{f(\mathcal{M}_k \times \mathcal{M}_k)}(T) \times n^4 \\
&\leq (\frac{|T|}{2^m} n^c + \epsilon') \times n^4 \\
&= \frac{|T|}{2^m} n^{c+4} + \epsilon
\end{aligned}$$

The inequality second from the last is because of the assumption that $f(\mathcal{M}_k \times \mathcal{M}_k)$ is ϵ' close to a distribution with minentropy $m - c\log n$. ∎

5 Acknowledgments

We thank Elvira Mayrdomo for useful discussions during the initial phase of this research. We thank the staff of Iowa Western Community College, Atlantic, Iowa for providing us meeting facilities throughout the course of this research.

References

[1] H. Buhrman, L. Fortnow, I. Newman, and N. Vereshchagin. Increasing Kolmogorov complexity. In *Proceedings of the 22nd Symposium on Theoretical Aspects of Computer Science*, volume 3404 of *LNCS*, pages 412–421, 2005.

[2] B. Barak, R. Impagliazzo, and A. Wigderson. Extracting randomness using few independent sources. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 384–393. IEEE Computer Society, 2004.

[3] B. Chor and O. Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988.

[4] Y. Dodis and R. Oliveira. On extracting randomness over a public channel. In *Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM)*, volume 2764 of *LNCS*, pages 252–263, 2003.

[5] L. Fortnow, J. M. Hitchcock, A. Pavan, N. V. Vinodchandran, and F. Wang. Extracting Kolmogorov complexity with applications to dimension zero-one laws. In *Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming*, number 4051 in LNCS, pages 335–345, 2006.

[6] V. Guruswami, C. Umans, and S. Vadhan. Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. In *IEEE Conference on Computational Complexity*, pages 96–108, 2007.

[7] M. Li and P. Vitányi. Learning simple concept under simple distributions. *SIAM Journal on Computing*, 20(5):911–935, 1991.

[8] M. Li and P. Vitányi. Average case complexity under the universal distribution equals worst-case complexity. *Information Processing Letters*, 42(3):145–149, 1992.

[9] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.

[10] N. Nisan and A. Ta-Shma. Extracting randomness: A survey and new constructions. *Journal of Computer and System Sciences*, 42(2):149–167, 1999.

[11] A. Rao. *Randomness Extractors for Independent Sources and Applications*. PhD thesis, University of Texas, Austin, 2006.

[12] A. Rao. A 2-source almost-extractor for linear entropy. In *APPROX-RANDOM*, pages 549–556, 2008.

[13] A. Shaltiel. *Current trends in theoretical computer science. Vol 1 Algorithms and Complexity*, chapter Recent Developments in extractors. World Scientific, 2004.

[14] M. Zimand. Two sources are better than one for increasing the Kolmogorov complexity of infinite sequences. In *Computer Science Symposium in Russia*, pages 326–338, 2008.

[15] M. Zimand. Extracting the Kolmogorov complexity of strings and sequences from sources with limited independence. In *Symposium on Theoretical Aspects of Computer Science*, pages 607–708, 2009.