Randomness extraction from defective sources

Domingos Dellamonica Jr. , Prof. Dr. Yoshiharu Kohayakawa (adviser)

[1]Instituto de Matemática e Estatística – Universidade de São Paulo
Rua do Matão, 1010 – Cidade Universitária, CEP 05508-090 – São Paulo – SP – Brasil

{ddj|yoshi}@ime.usp.br

Abstract. *Recently, Barak et al. (2004) constructed explicit deterministic* extractors *and* dispersers *(these are polynomial-time computable functions) with much better parameters than what was known before. We introduce the concepts involved in such a construction and mention some of its applications; in particular, we describe how it is possible to obtain much better bounds for the* bipartite Ramsey *problem (a very hard problem) using the machinery developed in that paper.*
We also present some original results that improve on these constructions. They are inspired by the work of Anup Rao (2005) and uses the recent breakthrough of Jean Bourgain (2005) in obtaining 2*-source extractors that break the "*1/2*-barrier".*

1. Introduction

This dissertation studies in depth many results in the area of randomness extraction. In order to understand this area, some concepts have to be introduced first.

A *randomized algorithm* can be viewed as an algorithm receiving two inputs. The first is an encoding of an instance and the second is a sequence of random bits. The algorithm then uses these random bits to make random decisions. Most analysis of randomized algorithms assume that the sequence of random bits it receives is perfectly uniform. In the application of such algorithms, however, these bits are generated by unknown probability distributions which may be far from uniform.

To make things even worse, it is impossible to check if a distribution on a large number of bits is close to uniform. In the randomness extraction setting, only a mild assumption over the distribution, which we call *source*, is made.

Definition 1. *Given a distribution $\mathcal{D}$ over $X = \{0,1\}^n$, the* min-entropy *of $\mathcal{D}$ is given by*

$$H^\infty(\mathcal{D}) = -\log\bigl(\max_{a \in X} \mathcal{D}(a)\bigr).$$

If X is a random variable with distribution $\mathcal{D}$ we may write $H^\infty(X) = H^\infty(\mathcal{D})$. A k-source *is a source with min-entropy at least k.*

Note that the min-entropy of a source basically measures the weight of the most likely element in the distribution. Anticipating the definitions that follow, any randomized algorithm A that uses $d \ll n$ random bits is incapable to convert a distribution $\mathcal{D}$ having a heavy weight element into a uniform distribution with many bits. Aiming for a perfectly uniform distribution in the output of A is not realistic. Therefore, one needs a notion of distance between distribution so that it is possible to compare the output of A against the uniform distribution.

Definition 2. *The* statistical difference *between two sources* $\mathcal{D}_1, \mathcal{D}_2 \subseteq X$*, is defined as*[1]

$$\frac{1}{2}\|\mathcal{D}_1 - \mathcal{D}_2\|_1 = \frac{1}{2}\sum_{a\in\Lambda}\bigl|\mathcal{D}_1(a) - \mathcal{D}_2(a)\bigr|.$$

We say that $\mathcal{D}_1$ *is* α-close *to* $\mathcal{D}_2$ *if the statistical distance between* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *is at most* α.

We can now state formally the definition of extractors. Let us denote by U_m the uniform distribution over $\{0,1\}^m$.

Definition 3 (Seeded extractor)**.** *A function* $E\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a* seeded extractor *for min-entropy* k *and error* $\varepsilon > 0$ *if, for any* k*-source* X*, we have that* $E(X, U_d)$*—the distribution obtained by computing* $E(x, s)$ *with* $x \in_R X$ *and an independent uniform seed* $s \in_R U_d$*—is* ε*-close to* U_m.

Definition 4 (Multiple sources extractor)**.** *An* l-source extractor *for min-entropy* k *and error* ε *is a function*

$$E\colon \{0,1\}^{n\times l} \to \{0,1\}^m$$

such that if $X_1, \ldots, X_l$ *are independent* k*-sources, then* $E(X_1, \ldots, X_l)$ *is* ε*-close to* U_m.

One can generalize the above definition to allow the lengths and min-entropies of the sources to be different.

2. Additive combinatorics and extractors

The field of randomness extraction experienced a revolution when breakthroughs in additive combinatorics provided the tools for powerful extractors which were not obtained constructively before (although probabilistically the existence of optimal extractors is known for a long time). The dissertation describes how these results in additive combinatorics can lead to such new construction in a fairly complete way, even going into proofs of several of them.

In turn, these new extractors, and other different constructions discussed in the thesis, can be used to explicitly construct Ramsey bipartite graphs which greatly improve the previous bound (which was almost twenty five years old).

The connection between additive combinatorics and randomness extractor is somewhat technical and a large chapter (Chapter 4) of the dissertation is devoted to it. Here, we shall present a high level idea of this connection, trying to keep things as simple and non-technical as possible.

The sum-product estimate over finite fields is the statement that for any given $A \subseteq \mathbb{F}$, with $|A| < |\mathbb{F}|^{.99}$, where $\mathbb{F}$ is a suitable field (such as $\mathbb{F} = \mathbb{Z}/(p)$ for a prime p), for some $\varepsilon > 0$,

$$\max\{|A+A|, |A\cdot A|\} \geq |A|^{1+\varepsilon}, \tag{1}$$

where $A * A = \{x * y \ : \ x, y \in A\}$.

Barak, Impagliazzo and Wigderson noticed that they could use the sum-product estimate to show that $f(x, y, x) = x \cdot y + z$ is a function such that if X, Y, Z are independent k-sources then $f(X, Y, Z)$ has min-entropy at least $\min\{0.99, (1+\varepsilon')k\}$. Iterating such a construction, they were able to obtain an extractor such that for any $\delta > 0$,

[1] The $1/2$ factor is just to keep the statistical distance in the range $[0, 1]$.

there exists a constant $l = l(\delta)$ and an explicit l-source extractor $E\colon \{0,1\}^{n\times l} \to \{0,1\}^n$ for min-entropy δn and error $2^{-\Omega(n)}$.

In order to prove that f has such a property, they had to deal with a several problems. The sum-product estimate deals with sets and not general distributions. In some sense, this is not a big problem, since one can decompose a distribution with min-entropy k into several distributions which are uniform over sets of cardinality 2^k. Moreover, the same set is used in the bound (1), while we do not make the assumption that X, Y and Z are identically distributed. Luckily, the celebrated Plunnecke–Rusza inequality shows that if a set A has $|A+A|$ large then $|A+B|$ is also large for all B with $|B| = |A|$.

Most importantly, even if we assume that X, Y and Z are independent and uniformly distributed over $A \subseteq \mathbb{F}$, there is no obvious connection between a set size estimate like (1) and the min-entropy of $f(X,Y,Z)$. That is where a recent result of Gowers debuts. This lemma of Gowers allows one to pass from a "density statement" to a "set size statement". Very roughly, the strategy is as follows: if $f(X,Y,Z)$ is not ε-close to uniform, then there is a dense hypergraph $\mathcal{H}$ consisting of many triples $(x,y,z) \in \mathbb{F}^3$ such that $|f(\mathcal{H})|$ is small. From $\mathcal{H}$ being dense one can obtain a subset $A' \subseteq A$ that violates the sum-product estimate, which is a contradiction.

In a recent paper accepted to LATIN–2008, I propose a more direct approach to the general strategy above which dispenses much of the machinery of additive combinatorics and instead focus on a proof entirely based on a *density to set size* argument over hypergraphs. Although several ideas behind this work were not mature enough at the time of the writing of the dissertation, they certainly were sparkled by the research carried during the preparation of my Master's Thesis.

3. Ramsey constructions

The bipartite Ramsey problem can be described as follows. Obtain an N by N bipartite graph with no K by K induced subgraph which is either empty or complete.

Suppose that $N = 2^n$ and $K = 2^k$, and let E be a two-source extractor for sources of n bits with min-entropy k, say $E\colon \{0,1\}^{n\times 2} \to \{0,1\}$. Then, for any pair of sets $X, Y \subseteq \{0,1\}^n$ with $|X|, |Y| \geq K$, we have $E(X,Y) = \{0,1\}$. If our bipartite graph is constructed by considering two copies of $\{0,1\}^n$ as the vertex classes and an edge uv exists iff $E(u,v) = 1$, there can be no K by K induced subgraph which is either complete or empty.

Actually, we have shown something stronger: the number of edges $(x,y) \in X \times Y$ contained in the graph should be approximately $\frac{1}{2}|X|\,|Y|$. If we restrict ourselves to the weaker definition that $E(X,Y) = \{0,1\}$ we have an object which is called a *disperser* in the randomness extraction literature.

Definition 5. *A function $D\colon \{0,1\}^{n\times l} \to \{0,1\}^m$ is an l-source* disperser *for min-entropy k with error ε if, for any independent k-sources $X_1, \ldots, X_l$, we have $\left|D(X_1,\ldots,X_l)\right| \geq (1-\varepsilon)2^m$. In particular, when $\varepsilon = 0$ we have $D(X_1,\ldots,X_l) = \{0,1\}^m$.*

Barak et al. constructed two-source dispersers where the min-entropy requirement is any linear function on the length of the inputs. This breakthrough used much of the machinery developed with the use of additive combinatorics. The philosophy behind their construction is, roughly speaking, to make two independent sources work in the same place as four independent sources would be needed. This "magic" is accomplished after providing a procedure to test the two inputs and decide a point where the inputs would be partitioned into two (hence one gets 4 inputs of possibly different sizes). The analysis is rather difficult since one has to show that, with some positive probability (over an arbitrary distribution for which the only thing one knows is the min-entropy), the test procedure will be partitioned in such a way that the four inputs are almost independent.

As a toy example of the above argument, let as assume one has three independent sources X_1, X_2, X_3. We say that a two-source extractor E for min-entropy k and error ε is *strong* if, for any k-sources X and Y, with probability $(1-\varepsilon)$ over the random choice $y \in_R Y$, we have that $E(X, y)$ is ε-close to uniform (and similarly for X). Therefore, a strong extractor is one for which we can fix one of the inputs to a "typical" value and the output will still be close to uniform. We can show that, although $E(X_1, X_2)$ and $E(X_3, X_2)$ cannot be said to be independent, if E is a strong extractor, they can be analysed as they were independent: fixing $x_2 \in X_2$, we have that $E(X_1, x_2)$ and $E(X_3, x_2)$ are independent (since x_2 is a constant).

It is possible to decompose the product distribution $Z = \big(E(X_1, X_2), E(X_3, X_2)\big)$ into distributions where $x_2 \in X_2$ is fixed. After discarding those x_2 for which $E(X_i, x_2)$ is not close to uniform for some $i \in \{1, 2\}$, we get a distribution which is close to Z and close to uniform at the same time. By the triangular inequality, Z is close to uniform and, hence, it is close to having the first element of the pair independent from the second element of the pair (since a uniform sequence of bits implies total independence between bits).

Clearly, the case where only two sources are available is considerably harder. There is no obvious choice for which value should be fixed as one of the inputs for the extractors. The strategy employed by Barak et al. is called the *challenge-response mechanism*. Loosely speaking, they define several possibilities for partitioning the two inputs (for instance, if $n = rs$, they can partition $x = x_1 \ldots x_n \in \{0, 1\}^n$ into $x_1 \ldots x_{js}$ and $x_{js+1} \ldots, x_n$ for some $j \in [r]$) and must decide which way they should be partitioned in order to simulate four independent sources. A string called the *challenge* is computed for each candidate partition. A set of *guesses* is also computed from the original inputs. If one of the guesses includes the challenge, the partition is *acceptable*.

After testing the possible partitions, a minimal (with respect to some partial order) acceptable partition is chosen and the four strings obtained from such a partition are used as if they were independent. This, however, does not work as well as in our toy case and the analysis can only prove that the challenge-response mechanism works with some probability under the initial space (which can be an arbitrary product of two independent k-sources). For the purposes of obtaining a disperser, this is more than enough: clearly, the support of the output includes the support of the smaller space in which the challenge-response can simulate independence. Since in the smaller space we actually obtain closeness to uniformity, the conclusion is immediate.

4. Our contributions

In Chapter 10 of the thesis, we present a strategy to obtain many more bits from the output of the constructions of extractors and dispersers in the work of Barak et al. This is accomplished by replacing a brute-force optimal extractor that is used in the composition of their constructions. We remark, however, that our strategy employs the use of extractors due to Bourgain which were not available by the time their results were announced (the dissertation was written based on a preliminary version of their paper). Independently, Rao[2] obtained these same improvements and several other interesting results.

We also worked on the construction of better constant-seed condensers, a tool used to obtain the extractors and dispersers of Barak et al. This work is not completely included in the dissertation and appears in a separate paper to be presented in LATIN 2008.

This Master's Thesis can be found on `http://www.teses.usp.br/teses/disponiveis/45/45134/tde-04052007-160412/`.

References

Barak, B., Impagliazzo, R., and Wigderson, A. (2004). Extracting randomness using few independent sources. In *FOCS*, pages 384–393. IEEE Computer Society.

Barak, B., Kindler, G., Shaltiel, R., Sudakov, B., and Wigderson, A. (2005). Simulating independence: new constructions of condensers, Ramsey graphs, dispersers, and extractors. In *STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 1–10, New York, NY, USA. ACM Press.

Barak, B., Rao, A., Shaltiel, R., and Wigderson, A. (2006). 2-source dispersers for $n^{o(1)}$ entropy and Ramsey graphs beating the Frankl-Wilson construction. In *STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, New York, NY, USA. ACM Press.

Bourgain, J. (2005). More on the sum-product phenomenon in prime fields and its applications. *International Journal of Number Theory*, 1:1–32.

Dellamonica Jr., D. (2008). Simpler constant-seed condensers. In E. S. Laber et al., editor, *Proceedings of Latin American Symposium on Theoretical Informatics (LATIN 2008)*, volume 4957, pages 664–675. Springer.

Rao, A. (2005). Extractors for a constant number of polynomial min-entropy independent sources. *Electronic Colloquium on Computational Complexity (ECCC)*, 5(106).

Raz, R. (2004). Extractors with weak random seeds. *Electronic Colloquium on Computational Complexity (ECCC)*, 4(099).

Tao, T. and Vu, V. H. (2006). *Additive Combinatorics*. Cambridge Studies in Advanced Mathematics.

[2] The first version of his paper, which appeared as an extended abstract in STOC 2006, did not include such results.