A 2-Source Almost-Extractor for Linear Entropy

Anup Rao[⋆,1]

School of Mathematics, Institute for Advanced Study
arao@ias.edu

Abstract. We give an explicit construction of a function that is almost a 2-source extractor for linear entropy, it is a condenser where the output has almost full entropy. Given 2 sources with entropy δn, the output of the condenser is a distribution on m-bit strings that is ϵ-close to having min-entropy $m - \text{poly}(\log(1/\epsilon), 1/\delta)$, where here m is linear in n.

1 Introduction

This paper is about constructing efficiently computable 2-source extractors. These are efficiently computable functions of the type $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ with the property that for any 2 independent distributions X, Y, each with entropy[1] k, the output $\mathsf{Ext}(X, Y)$ is close to uniform. Another way to view this object is as a coloring of the edges of the $N \times N$ complete bipartite graph with M colors that guarantees that in every $K \times K$ complete bipartite subgraph, every set of colors is hit with roughly the right frequency.

This problem was first suggested in the work of Chor and Goldreich [CG88] (see also [SV86]), who gave a simple argument that shows that the inner product function over $GF(2)$ is a good 2 source extractor as long as $k/n > 1/2 + \Omega(1)$. It is easy to generalize this to get many random bits (simply take the inner product over a large enough field). Since then, most work was diverted to the special case of seeded extractors (introduced in [NZ96]), where it is assumed that the second source is much shorter than the first source and is uniformly distributed (a 2-source extractor can be used in this situation just by padding the second source). Here almost optimal results are now known [LRVW03,GUV07].

There was no progress in reducing the entropy requirements for the general case of 2 source extractors until the work of Bourgain [Bou05], almost 20 years after [CG88]. Bourgain used recent results from arithmetic combinatorics [BKT04] to show that if the inputs are viewed as elements of a carefully chosen finite field, and ψ is any non-trivial additive character, the function $\psi(xy+x^2y^2)$ is an extractor even for entropy $0.499n^2$. Bourgain's result, while seemingly a minor improvement over the previous result, had at least one application that

⋆ Supported in part by NSF Grant CCR-0324906.

1 The definition of entropy we use is *min-entropy*, rather than Shannon entropy.

2 Note that the inner product function mentioned above can also be viewed as $\psi(xy)$, where x, y are interpreted as elements of $GF(2^n)$ and ψ is a suitably chosen additive character.

would not have been possible using just the ideas of Chor and Goldreich: it led to new constructions of Ramsey graphs with much better parameters than were previously known [BRSW06].

The problem of constructing 2 source extractors for arbitrary linear min-entropy remains open. In this paper we describe some partial progress towards this goal, obtaining an object that seems tantalizingly close to being a 2 source extractor.

1.1 Our Results and Techniques

We prove the following theorem:

Theorem 1. *For every $\delta > 0$ and every ϵ, there exists a polynomial time computable function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\}^m$ such that if X, Y are independent sources with min-entropy rate δ, $\mathsf{Ext}(X, Y)$ is ϵ close to having min-entropy $m - \mathrm{poly}(1/\delta, \log(1/\epsilon))$, with $m = \Omega(\delta n)$.*

The output of this algorithm is close to having such a high min-entropy that we hope that it may still be sufficient for applications where 2-source extractors are required. For instance, if we are willing to make cryptographic assumptions that rely only on secret keys with such high entropy, this extractor may be used in lieu of a 2-source extractor for generating secret keys.

Our result follows by composing several previous explicit constructions. Specifically, we rely on two types of explicit functions from previous work:

2 independent sources $\rightarrow$ SR-source A first observation (already made in [BKS+05]) is that it is possible to use arithmetic combinatoric results to get an explicit function $\mathsf{SExt} : \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\}^t$, that converts two independent sources into a *somewhere random source*. A distribution on strings is somewhere random if at least one of the strings is distributed uniformly. The above construction combined with some ideas from [Rao06] gives an algorithm that can carry out such a conversion, outputting a somewhere random source with only a constant number of strings, each of length linear in n.

2 independent sources + independent SR-source $\rightarrow$ uniform source It is also easy to use previous work [Raz05,DR05,BKS+05] to get an explicit function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^t \rightarrow \{0,1\}^m$ that can extract randomness from two independent sources with linear entropy and an additional *independent* small somewhere random source.

Our final construction is $\mathsf{Ext}'(X, Y, \mathsf{SExt}(X, Y))$, i.e. we use the somewhere random source generated by the original source to extract random bits from X, Y. At first it may seem like this has very little chance of working, since the somewhere random source is *not independent* of the original sources (in fact it is determined by them). Still, we show that if our goal is just to show that the output has high entropy, something can be salvaged from this approach, giving us our main result. Ideas that superficially seem similar to this one have been used in previous work [GRS04,Sha06]. It is hard to describe the many ideas in those papers succinctly, so in the discussion here we shall be slightly inaccurate in order to convey the gist of the differences between the techniques of those works and the present paper. In the earlier works, the authors first construct a function $\mathsf{DExt} : \{0,1\}^n \rightarrow \{0,1\}^t$ that extracts a few random bits from some class of sources. They then use the extracted random bits to extract many more random bits from the original source. Thus the final algorithm looks like $\mathsf{Ext}'(X, \mathsf{DExt}(X))$ for some carefully chosen function Ext'.

The major difference between the previous works and ours is in the analysis. The previous works carefully controlled the correlations between the extracted bits ($\mathsf{DExt}(X)$) and the original source X. In particular, they carefully chose a random variable in the probability space they were considering and fixed it. Conditioned on this fixing, they were able to argue that the extracted bits (or some subset of the extracted bits) became independent of the original source (or some part of the original source). In this way, after fixing this magic random variable, they were able to obtain two random variables that could be treated as being independent (without paying a too heavy price in terms of lost entropy). In order to make this approach work, they had to carefully exploit the properties of the class of distributions they were building extractors for and the properties of the functions they were constructing.

The ideas in this paper are less delicate and less intricate. In particular, they do not apply just to the case of independent sources. They can be generalized[3] to be used in any situation where we know how to construct an explicit function SExt that can convert a distribution from class $\mathcal{C}_1$ into one from class $\mathcal{C}_2$ with small support size, and an explicit function that can extract random bits (or even high entropy bits) from two independent distributions, one from class $\mathcal{C}_1$ and the other from $\mathcal{C}_2$. In our particular application, $\mathcal{C}_1$ is the class of two independent sources and $\mathcal{C}_2$ is the class of somewhere random sources. In this situation, we simply show how to use the union bound to get a result of the type of Theorem 1.

It is easy to see that if a distribution is far from having high min-entropy, then there must be a small set of the support that has an unusually high probability under it. Fix any subset of the support. In order to show that $\mathsf{Ext}'(X, Y, \mathsf{SExt}(X, Y))$ does not hit this set with such a high probability, we consider the set of *bad* outputs of SExt. Say z is bad if $\mathsf{Ext}'(X, Y, z)$ hits the set with high probability. Then the properties of Ext' guarantee that any somewhere random source has only a small probability of giving such a bad z. On the other hand, since the total number of z's is so small (the output is only a constant number of bits), we can argue that with high probability $\mathsf{Ext}'(X, Y, z)$ does not land in the set for *every good* z. Thus, by the union bound, we can argue that any small enough set is avoided with significant probability.

Since the above argument requires us to use the union bound on as many events as there are elements in the support of SExt, it is crucial that the error

[3] Shaltiel [Sha06] also generalized the ideas of [GRS04] to several classes of sources, but there each class he considered required a different construction and a different analysis, though there was a very significant overlap in the various cases.

of the extractor Ext' be significantly small in terms of the number of elements in the support of SExt. Luckily, explicit constructions that we rely on already provide such strong guarantees.

2 Preliminaries

We will be concerned with the treatment of various kinds of distributions that are *nice* in that they contain a lot of usable randomness. Here we discuss some ways to measure this niceness:

Definition 1. *The* min-entropy *of a distribution R is defined to be: $H_\infty(R) = -\log(\max_{x\in R}(R(x))$. The* min-entropy rate *of a distribution R on $\{0,1\}^n$ is $H_\infty(R)/n$.*

Definition 2. *An* (n,k)-source *denotes some random variable X over $\{0,1\}^n$ with $H_\infty(X) \geq k$.*

Definition 3. *Let D and F be two distributions on a set S. Their* statistical distance *is*

$$|D - F| \stackrel{def}{=} \max_{T\subseteq S}(|D(T) - F(T)|) = \frac{1}{2}\sum_{s\in S}|D(s) - F(s)|$$

If $|D - F| \leq \epsilon$ we shall say that D is ϵ-close *to F.*

This measure of distance is nice because it is robust in the sense that if two distributions are close in this distance, then applying any functions to them cannot make them go further apart.

Proposition 1. *Let D and F be any two distributions over a set S s.t. $|D-F| \leq \epsilon$. Let g be any function on S. Then $|g(D) - g(F)| \leq \epsilon$.*

A block source is a source broken up into a sequence of blocks, with the property that each block has min-entropy even conditioned on previous blocks.

Definition 4 (Block sources). *A distribution $X = X_1, X_2, \cdots, X_{\mathsf{C}}$ is called a $(k_1, k_2, \ldots, k_{\mathsf{C}})$-block source if for all $i = 1, \ldots, \mathsf{C}$, we have that for all $x_1 \in X_1, \ldots, x_{i-1} \in X_{i-1}$, $H_\infty(X_i|X_1 = x_1, \ldots, X^{i-1} = x_{i-1}) \geq k_i$, i.e., each block has high min-entropy even conditioned on the previous blocks. If $k_1 = k_2 = \cdots = k_{\mathsf{C}} = k$, we say that X is a k-block source.*

We have the following standard lemma:

Lemma 1. *Suppose X is a source with min-entropy k and $f : \{0,1\}^n \to \{0,1\}^t$ is a function such that $f(X)$ is ϵ close to having min-entropy k'. Then for every ℓ, $(f(X), X)$ is $\epsilon + 2^{-\ell}$ close to being a $k', k - t - \ell$ block source.*

We shall need the concept of a somewhere random distribution.

Definition 5. *A source X is $(t \times r)$* somewhere-random *if it is distribution on $t \times r$ boolean matrices s.t. X is distributed uniformly randomly over one of the rows. Every other row may depend on the random row in arbitrary ways. We say that X has somewhere min-entropy k if at least one of the rows has min-entropy k.*

3 Previous work needed

Our work relies on several previous constructions. The first object we shall need is the additive number theory based condensers independently constructed by Barak et al. [BKS⁺05] and Raz [Raz05]:

Lemma 2 ([Raz05,BKS⁺05]). *For every $\delta > 0$, there exists a polynomial time computable function* $\mathsf{Cond} : \{0,1\}^n \to (\{0,1\}^{n/\mathrm{poly}(1/\delta)})^{\mathrm{poly}(1/\delta)}$*, where the output is interpreted as a* $\mathrm{poly}(1/\delta) \times n/\mathrm{poly}(1/\delta)$ *boolean matrix, such that if X is a source with min-entropy rate δ,* $\mathsf{Cond}(X)$ *is* $2^{-\Omega(\delta^2 n)}$ *close to a convex combination of distributions, each of which has some row with min-entropy rate* 0.9*.*

When this lemma is combined with the merger from Raz's work [Raz05] and the improved analysis of Dvir and Raz (Lemma 3.2 in [DR05]), we get the following lemma:

Lemma 3 ([DR05,Raz05,BKS⁺05]). *For every $\delta > 0$ and $\epsilon > 2^{-n/10}$, there exists a polynomial-time computable function*
$\mathsf{Cond} : \{0,1\}^n \to (\{0,1\}^{n/\mathrm{poly}(1/\delta)})^{2^{\mathrm{poly}(1/\delta)}/\epsilon}$*, where the output is treated as a* $2^{\mathrm{poly}(1/\delta)/\epsilon} \times n/\mathrm{poly}(1/\delta)$ *boolean matrix, such that if X has min-entropy rate δ,* $\mathsf{Cond}(X)$ *is* $2^{-\Omega(\delta^2 n)}$ *close to a convex combination of distributions, each of which has at most an ϵ fraction of rows with min-entropy rate less than* 0.9*.*

We need the following two source extractor of Chor and Goldreich:

Theorem 2 ([CG88]). *For every constant $\delta > 1/2$ there exists a strong two source extractor* $\mathsf{Had} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^{\Omega(n)}$ *with error* $2^{-\Omega(n)}$ *for two independent sources with min-entropy δn.*

We can use X, Y to generate a somewhere random source Z. The following theorem was proved in [BKS⁺05]:

Theorem 3. *For every δ, there exists $c(\delta) = \mathrm{poly}(1/\delta)$ and a polynomial time computable function* $\mathsf{SExt} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^{cn/\mathrm{poly}(1/\delta)}$*, where the output is treated as a* $c \times n/\mathrm{poly}(1/\delta)$ *boolean matrix, such that if X, Y are independent sources with min-entropy rate δ,* SExt *is* $2^{-\Omega(n/\mathrm{poly}(1/\delta))}$ *close to a convex combination of somewhere random sources.*

Proof. Define the $(i,j)'th$ row $\mathsf{SExt}(X,Y)_{i,j} = \mathsf{Had}(\mathsf{Cond}(X)_i, \mathsf{Cond}(Y)_j)$, where Cond is as in Lemma 2 and Had is as in Theorem 2. The theorem follows directly.

Finally, we need the following two source extractor for block sources, that follows from the work of [BKS⁺05,Rao06]:

Theorem 4 ([BKS⁺05,Rao06]). *For every $\delta > 0$, there exists a constant $\gamma > 0$ and a polynomial time computable function* $\mathsf{Ext} : (\{0,1\}^n)^4 \to \{0,1\}^m$ *such that if X_1, X_2 is a $\delta n, \delta n$ block source and Y_1, Y_2 is an independent $\delta n, \delta n$ block source,*

$$\Pr_{x_1,x_2}[|\mathsf{Ext}(x_1, x_2, Y_1, Y_2) - U_m| > 2^{-\gamma n}] < 2^{-\gamma n}$$

and

$$\Pr_{y_1,y_2}[|\mathsf{Ext}(X_1,X_2,y_1,y_2)-U_m|>2^{-\gamma n}]<2^{-\gamma n}$$

where here $m=\Omega(n)$ and U_m denotes the uniform distribution on m bit strings.

4 The Condenser

First we show that if we were given a small independent somewhere random source, we can use it to extract random bits from two linear min-entropy independent sources. The idea is that the somewhere random source can be used to turn both of the other sources into block sources, using Lemma 3.

Theorem 5. *For every $1>\delta,\epsilon_2>0$ and $c>0$, there exists a $t(c,\delta,\epsilon_2)=\mathrm{poly}(c,1/\delta,\log(1/\epsilon_2))$, a constant $\gamma(\delta)$ and a polynomial time computable function $\mathsf{Ext}:\{0,1\}^n\times\{0,1\}^n\times\{0,1\}^{c\times t}\rightarrow\{0,1\}^{\delta n-o(1)}$ such that if X,Y are independent min-entropy rate δ sources and Z is an independent $c\times t$ somewhere random source,*

$$\Pr_z[\Pr_y[\mathsf{Ext}(X,y,z) \text{ is } 2^{-\gamma n} \text{ close to uniform}]>1-2^{-\gamma n}]>1-\epsilon_2$$

$$\Pr_z[\Pr_x[\mathsf{Ext}(x,Y,z) \text{ is } 2^{-\gamma n} \text{ close to uniform}]>1-2^{-\gamma n}]>1-\epsilon_2$$

Proof. Let $\delta'<\delta$ be a small enough constant so that length of the rows output by Cond in Lemma 3 for error ϵ_2 and min-entropy rate δ' is at most $\delta^2 n/c$. Let 2^t be the number of rows output by Cond for this setting of parameters (so that $t=\mathrm{poly}(1/\delta,c,\log(1/\epsilon_2))$.

Now we treat each row of Z as the name of a row of $\mathsf{Cond}(X)$. Let X_Z denote the string $\mathsf{Cond}(X)_{Z_1},\ldots,\mathsf{Cond}(X)_{Z_c}$. Similarly let Y_Z denote $\mathsf{Cond}(Y)_{Z_1},\ldots,\mathsf{Cond}(Y)_{Z_t}$. Then note that X_Z and Y_Z are of length $\delta^2 n$. Further, by the properties of Cond, with high probability over the choice of z, X_z and Y_z are $2^{-\Omega(n)}$ close to having min-entropy rate $0.9/c$. Since X_Z is so short, Lemma 1 implies that (X_z,X) and (Y_z,Y) are $2^{-\Omega(n)}$ close to independent block sources with entropy $0.9\delta^2 n/c,(\delta-\delta^2)n\geq\delta^2 n$. So we can apply the extractor from Theorem 4 to get the result of the lemma.

Now although we don't have access to a somewhere random source Z as above, Theorem 3 tells us that we can generate such a source in polynomial time using the function SExt. So let us define the function $\mathsf{Ext}'(X,Y)\stackrel{def}{=}\mathsf{Ext}(X,Y,\mathsf{SExt}(X,Y))$. It is not at all clear that this function is an extractor, since now X,Y are not independent of the somewhere random source being used (in fact they determine it!). Still, we show that the output of this function must be close to having very high min-entropy.

Before we show this, we need two simple lemmas:

Lemma 4. *Let A be a distribution that is ϵ-far from having min-entropy k. Then, there must be a set H of size at most 2^k such that* $\Pr[A\in H]\geq\epsilon$.

Proof. Set $H = \{h : \Pr[A = h] \geq 2^{-k}\}$. This set clearly has at most 2^k element. The lemma is immediate from the definition of statistical distance.

Lemma 5. *Let $A_1, \ldots, A_l$ be random variables taking values in $\{0,1\}^n$, Z be a random variable taking values in $[l]$ and $G \subset [l]$ be a set such that:*

– *For every $z \in G$, $|A_z - U_n| < \tau$.*
– $\Pr[Z \in G] > 1 - \epsilon$.

Then for every integer d, A_Z is $\epsilon + l(\tau + 2^{-d})$ close to having min-entropy $n - d$.

Proof. Suppose not. Then, by Lemma 4, there must be some set of heavy elements $H \subset \{0,1\}^n$ of size at most 2^{n-d} such that $\Pr[A_Z \in H] \geq \epsilon + l(\tau + 2^{-d})$. Now note that $A_Z \in H$ implies that either $Z \notin G$ or one of the good A_i's must have hit H. Thus, by the union bound,

$$\begin{aligned}\Pr[A_Z \in H] &< \Pr[Z \notin G] + \Pr[\exists z \in G \text{ with } A_z \in H] \\ &\leq \epsilon + |G|(\tau + 2^{-d}) \\ &< \epsilon + l(\tau + 2^{-d})\end{aligned}$$

We can now prove the main theorem of this paper.

Proof (Theorem 1). Let A_z denote the random variable $\mathsf{Ext}(X, Y, z)$. Let $\gamma = \Omega_\delta(1)$ be as in Theorem 5, with $\epsilon_2 = \epsilon$. Define

$$G = \{z : \Pr_y[\mathsf{Ext}(X, y, z) \text{ is } 2^{-\gamma n} \text{ close to uniform}] > 1 - 2^{-\gamma n}\}$$

Then we see that $\Pr[Z \in G] > 1 - \epsilon_2$, if Z is somewhere random and independent. Instead we set $Z = \mathsf{SExt}(X, Y)$ (truncating each row to be of length $t = \text{poly}(c, 1/\delta, \log(1/\epsilon))$ as required by Theorem 5).

Thus we have that $\Pr[Z \in G] > 1 - \epsilon_2 - 2^{-\Omega_\delta(n)}$, since $\mathsf{SExt}(X, Y)$ is $2^{-\Omega_\delta(n)}$ close to a convex combination of somewhere random sources. Further, for every $z \in G$, $|A_z - U_n| < 2^{-\Omega_\delta(n)}$. The total number of z's is at most $2^{ct} = 2^{\text{poly}(1/\delta, \log(1/\epsilon_2))}$. Thus, by Lemma 5, setting $d = 100ct/\log(1/\epsilon)$, we have that $\mathsf{Ext}'(X, Y)$ is $2\epsilon_2$ close to having min-entropy $m - \text{poly}(1/\delta, \log(1/\epsilon))$.

5 Acknowledgements

I would like to thank Boaz Barak, Ronen Shaltiel and Avi Wigderson for useful discussions.

References

[BKS+05] Boaz Barak, Guy Kindler, Ronen Shaltiel, Benny Sudakov, and Avi Wigderson. Simulating independence: New constructions of condensers, Ramsey graphs, dispersers, and extractors. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 1–10, 2005.

[BRSW06] Boaz Barak, Anup Rao, Ronen Shaltiel, and Avi Wigderson. 2 source dispersers for $n^{o(1)}$ entropy and Ramsey graphs beating the Frankl-Wilson construction. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, 2006.

[Bou05] Jean Bourgain. More on the sum-product phenomenon in prime fields and its applications. *International Journal of Number Theory*, 1:1–32, 2005.

[BKT04] Jean Bourgain, Nets Katz, and Terence Tao. A sum-product estimate in finite fields, and applications. *Geometric and Functional Analysis*, 14:27–57, 2004.

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988.

[DR05] Zeev Dvir and Ran Raz. Analyzing linear mergers. Technical Report TR05-25, ECCC: Electronic Colloquium on Computational Complexity, 2005.

[GRS04] Ariel Gabizon, Ran Raz, and Ronen Shaltiel. Deterministic extractors for bit-fixing sources by obtaining an independent seed. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, 2004.

[GUV07] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from parvaresh-vardy codes. In *Proceedings of the 22nd Annual IEEE Conference on Computational Complexity*, 2007.

[LRVW03] C. J. Lu, Omer Reingold, Salil Vadhan, and Avi Wigderson. Extractors: Optimal up to constant factors. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pages 602–611, 2003.

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *Journal of Computer and System Sciences*, 52(1):43–52, 1996.

[Rao06] Anup Rao. Extractors for a constant number of polynomially small min-entropy independent sources. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, 2006.

[Raz05] Ran Raz. Extractors with weak random seeds. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 11–20, 2005.

[SV86] Miklos Santha and Umesh V. Vazirani. Generating quasi-random sequences from semi-random sources. *Journal of Computer and System Sciences*, 33:75–87, 1986.

[Sha06] Ronen Shaltiel. How to get more mileage from randomness extractors. In *Proceedings of the 21th Annual IEEE Conference on Computational Complexity*, pages 49–60, 2006.