Electronic Colloquium on Computational Complexity, Report No. 67 (2005)

An Improved Analysis of Mergers

Zeev Dvir* Amir Shpilka†

Abstract

Mergers are functions that transform k (possibly dependent) random sources into a single random source, in a way that ensures that if one of the input sources has min-entropy rate δ then the output has min-entropy rate close to δ. Mergers have proven to be a very useful tool in explicit constructions of *extractors* and *condensers*, and are also interesting objects in their own right. In this work we present a new analysis of the merger construction of [LRVW03]. Our analysis shows that the min-entropy rate of this merger's output is actually $0.52 \cdot \delta$ instead of $0.5 \cdot \delta$, where δ is the min-entropy rate of one of the inputs. To obtain this result we deviate from the usual linear algebra methods that were used by [LRVW03] and introduce a new technique that involves results from additive number theory.

1 Introduction

Mergers are functions that take as input k samples, taken from k (possibly dependent) random sources, each of length n-bits. It is assumed that one of these random sources, who's index is unknown, is sufficiently random, in the sense that it has min-entropy $\geq \delta n$ (A source has min-entropy $\geq b$ if none of its values is obtained with probability larger than 2^{-b}). We want the merger to output an n'-bit string (n' could be smaller than n) that will be close to having min-entropy at least $\delta' n'$, where δ' is not considerably smaller than δ. To achieve this, the merger is allowed to use an additional small number of truly random bits called a *seed*. The goals in merger constructions are a) to minimize the seed length, b) to maximize the min-entropy of the output and c) to minimize the error (that is, the statistical distance between the merger's output and some high min-entropy source).

The notion of *merger* was first introduced by Ta-Shma [TS96], in the context of explicit constructions of *extractors*. An extractor is a function that transforms a source with min-entropy b into a source which is close to uniform, with the aid of an additional random seed. Extractors are a very important tool in derandomization. For a more detailed discussion of extractors see [Sha02]. Recently, Lu, Reingold, Vadhan and Wigderson [LRVW03] gave a very simple and beautiful construction of mergers based on Locally-Decodable-Codes. This construction was used in [LRVW03] as a building block in an explicit construction of extractors with nearly optimal parameters. More recently, [Raz05] generalized the construction of [LRVW03], and showed how this construction (when combined with other techniques) can be used to construct *condensers* (a condenser is a function that

*Department of Computer Science, Weizmann institute of science, Rehovot, Israel.
Email: zeev.dvir@weizmann.ac.il. Research supported by Israel Science Foundation (ISF) grant.
†Department of Computer Science, Weizmann institute of science, Rehovot, Israel.
Email: amir.shpilka@weizmann.ac.il. Research supported by the Koshland fellowship.

ISSN 1433-8092

transforms a source with min-entropy rate δ into a source which is close to having min-entropy rate $\delta' > \delta$, with the aid of an additional random seed) with constant seed length. The analysis of the merger constructed in [Raz05] was subsequently refined in [DR05].

The merger constructed by [LRVW03] takes as input k strings of length n, one of which has min-entropy b, and outputs a string of length n that is close to having min-entropy at least $0.5 \cdot b$. Loosely speaking, the output of the merger is computed as follows: treat each input block as a vector in the vector space F^m, where F is some small finite field, and output a uniformly chosen linear combination of these k vectors. The analysis of this construction is based on the following simple idea: In every set of linear combinations, who's density is at least γ (where γ is determined by the size of F), there exist two linear combinations that, when put together, determine the 'good' source (that is, the 'good' source can be computed from both of them deterministically). Therefore, one of these linear combinations must have at least half the entropy of the 'good' source (this reasoning extends also to min-entropy). As a results we get that for most seed values (linear combinations) the output has high min-entropy, and the result follows. This is of course an over-simplified explanation, but it gives the general idea behind the proof.

In this paper we present an alternative analysis to the one just described. Our analysis relies on two results from the field of additive-number theory. The first is Szemeredi's theorem ([Sze75],[Gow01]) on arithmetic progressions of length three (also known as Roth's theorem [Rot53]). This theorem states that in every subset of $\{1, \ldots, N\}$, who's density is at least $\delta(N)$, there exits an arithmetic progression of length three. For our purposes we use a quantitative version of this theorem as proven by Bourgain [Bou99b]. The second result that we rely on is a lemma of Bourgain [Bou99a] that deals with sum-sets and difference-sets of integers. Roughly speaking, the lemma says that if the sum-set of two sets of integers is very small, then their difference-set cannot be very large (for a precise formulation see Section 3). It is interesting to note that this is not the first time that results from additive number-theory are used in the context of randomness extraction. A recent result of Barak, Imagliazzo and Wigderson [BIW04] uses results from this field to construct multi-source extractors.

Using these two results we are able to show that the min-entropy outputted by the aforementioned merger is $0.52 \cdot b$ and not $0.5 \cdot b$ as was previously known. One drawback of our analysis is that the length of the seed is required to be $O(k \cdot \gamma^{-2})$ in order for the output error to be γ, where in the conventional analysis the seed length can be as short as $O(k \cdot \log(\gamma^{-1}))$. This however does not present a problem in many of the current applications of mergers, where the error parameter and the number of input sources are both constants and the seed length is also required to be a constant. One place where our analysis can be used in order to simplify an existing construction is in the extractor construction of [Raz05]. There, the output of the merger is used as an input to an extractor that requires the min-entropy rate of its input to be larger than one-half. In [Raz05] this problem is addressed by a more complicated merger construction who's output length is shorter than n. our analysis shows that the more simple construction of [LRVW03] could be used instead, since its output min-entropy rate is larger than one-half.

1.1 Somewhere-Random-Sources

An *n-bit random source* is a random variable X that takes values in $\{0,1\}^n$. We denote by $\mathrm{supp}(X) \subset \{0,1\}^n$ the support of X (i.e. the set of values on which X has non-zero probability). For two n-bit

sources X and Y, we define the statistical distance (or simply distance) between X and Y to be

$$\Delta(X,Y) \triangleq \frac{1}{2} \sum_{a \in \{0,1\}^n} \left| \mathbf{Pr}[X=a] - \mathbf{Pr}[Y=a] \right|.$$

We say that a random source X (of length n bits) has min-entropy $\geq b$ if for every $x \in \{0,1\}^n$ the probability for $X = x$ is at most 2^{-b}.

Definition 1.1 (Min-entropy). *Let X be a random variable distributed over $\{0,1\}^n$. The min-entropy of X is defined as* [1]

$$H^{\infty}(X) \triangleq \min_{x \in supp(X)} \log\left(\frac{1}{\mathbf{Pr}[X=x]}\right).$$

Definition 1.2 ((n,b)-Source). *We say that X is an (n,b)-source, if X is an n-bit random source, and $H^{\infty}(X) \geq b$.*

A *somewhere-(n,b)-source* is a source comprised of several blocks, such that at least one of the blocks is an (n,b)-source. Note that we allow the other source blocks to depend arbitrarily on the (n,b)-source, and on each other.

Definition 1.3 ($(n,b)^{1:k}$-Source)). *A k-places-somewhere-(n,b)-source, or shortly, an $(n,b)^{1:k}$-source, is a random variable $X = (X_1, \ldots, X_k)$, such that every X_i is of length n bits, and at least one X_i is of min-entropy $\geq b$. Note that $X_1, \ldots, X_k$ are not necessarily independent.*

Comment: It is possible to define a somewhere-(n,b)-source in a more general way to also include convex combinations of sources of the type described by Definition 1.3. However, it suffices to consider sources of this simpler type for the task of merger constructions.

1.2 Mergers

A merger is a function transforming an $(n,b)^{1:k}$-source into a source which is γ-close (i.e. it has statistical distance $\leq \gamma$) to an (m,b')-source. Naturally, we want b'/m to be as large as possible, and γ to be as small as possible. We allow the merger to use an additional small number of truly random bits, called a *seed*. A Merger is *strong* if for almost all possible assignments to the seed, the output is close to be an (m,b')-source. A merger is *explicit* if it can be computed in polynomial time.

Definition 1.4 (Merger). *A function $M : \{0,1\}^d \times \{0,1\}^{n \cdot k} \to \{0,1\}^m$ is a $[d, (n,b)^{1:k} \mapsto (m,b') \sim \gamma]$-merger if for every $(n,b)^{1:k}$-source X, and for an independent random variable Z uniformly distributed over $\{0,1\}^d$, the distribution $M(Z,X)$ is γ-close to a distribution of an (m,b')-source. We say that M is **strong** if the average over $z \in \{0,1\}^d$ of the minimal distance between the distribution of $M(z,X)$ and a distribution of an (m,b')-source is $\leq \gamma$.*

We now give a formal definition of the merger we wish to analyze. To simplify the analysis we will assume in several places that certain quantities are integers. This, however, will not affect our results in any significant way.

[1] All logarithms in this paper are taken base 2.

Construction 1.5 ([LRVW03]). *Let n, k be integers, p a prime number, and let $l = \frac{n}{\log(p)}$. We define a function*

$$M : \{0,1\}^d \times \{0,1\}^{n \cdot k} \to \{0,1\}^n,$$

with

$$d = \log(p) \cdot k,$$

in the following way: Let F denote the field GF(p)*. Given $z \in \{0,1\}^d$, we think of z as a vector $(z_1, \ldots, z_k) \in F^k$. Given $x = (x_1, \ldots, x_k) \in \{0,1\}^{n \cdot k}$, we think of each $x_i \in \{0,1\}^n$ as a vector in F^l. The function M is now defined as*

$$M(z, x) = \sum_{i=1}^{k} z_i \cdot x_i \in F^l,$$

where the operations are preformed in the vector space F^l. Intuitively, one can think of M as

$$M : F^k \times \left(F^l\right)^k \to F^l.$$

1.3 Our Results

We prove the following theorem:

Theorem 1. *Let $0 < \gamma < 1$ be any constant, $k > 0$ a constant integer, and let p be a prime larger than*[2] $\exp(\gamma^{-2})$*. Let*

$$M : \{0,1\}^d \times \{0,1\}^{n \cdot k} \to \{0,1\}^n,$$

be as in Construction 1.5, where $d = \log(p) \cdot k$ (and underlying field GF(p)*). Then for any constant $\alpha > 0$ there exists a constant b_0 such that for all $n \geq b \geq b_0$, M is a $[d, (n, b)^{1:k} \mapsto (n, b') \sim \gamma]$-strong merger with*

$$b' = (0.52 - \alpha) \cdot b.$$

From Theorem 1 we see that in order to get a merger with error γ we need to choose the underlying field to be of size at least $\exp(\gamma^{-2})$. It is well known that for every integer n, there is a prime between n and $2n$. Therefore we can take p to be $O\left(\exp(\gamma^{-2})\right)$ and have that the length of the random seed is

$$d = \log(p) \cdot k = O\left(k \cdot \gamma^{-2}\right)$$

bits long. Hence, for constant γ and k the length of the random seed used by the merger is constant. Notice that finding the prime p that answers our demands can be done by testing all integers in the range of interest for primality (if γ is constant then this will take a constant amount of time).

1.4 Organization

In Section 2 we give our analysis of Construction 1.5 and prove Theorem 1. The analysis presented in Section 2 relies on two central claims that we prove in Section 3.

[2] In writing $\exp(f)$ we mean $2^{O(f)}$.

2 Analysis of Construction 1.5

In this section we present our improved analysis of Construction 1.5, and prove Theorem 1. The analysis will go along the same lines as in [DR05] and will differ from it in two claims that we will prove in Section 3. We begin with some notations that will be used throughout the paper.

Let $X = (X_1, \ldots, X_k) \in \{0,1\}^{n \cdot k}$ be a somewhere (n, b)-source, and let us assume w.l.o.g. that $\mathrm{H}^{\infty}(X_1) \geq b$. Let $0 < \gamma < 1$ be any constant, and let $p \geq \exp(\gamma^{-2})$ be a prime number. Let $M : F^k \times \left(F^l\right)^k \to F^l$, be as in Construction 1.5, where $F = \mathrm{GF(p)}$, $d = \log(p) \cdot k$ and $n = \log(p) \cdot l$. Our goal is to analyze the min-entropy of $M(Z, X)$ where Z will denote a random variable uniformly distributed over F^k. In particular, we would like to show that the random variable $M(Z, X)$ is γ-close to having min-entropy $\geq (0.52 - \alpha) \cdot b$ for all constant α.

For every $z \in F^k$ we denote by $Y_z \triangleq M(z, X)$ the random variable given by the output of M on the fixed seed value z (recall that, in Construction 1.5, every seed value corresponds to a specific linear combination of the source blocks). Let $u \triangleq 2^d = p^k$ be the number of different seed values, so we can treat the set $\{0,1\}^d$ as the set $[u] \triangleq \{1, 2, \ldots, u\}$. We can now define $Y \triangleq (Y_1, \ldots, Y_u) \in (F^l)^u$. The random variable Y is a function of X, and is comprised of u blocks, each one of length $n = \log(p) \cdot l$ bits, representing the output of the merger on all possible seed values. We will first analyze the distribution of Y as a whole, and then use this analysis to describe the output of M on a uniformly chosen seed.

Definition 2.1. *Let $D(\Omega)$ denote the set of all probability distributions over a finite set Ω. Let $\mathcal{P} \subset D(\Omega)$ be some property. We say that $\mu \in D(\Omega)$ is γ-close to a convex combination of distributions with property $\mathcal{P}$, if there exists constants $\alpha_1, \ldots, \alpha_t, \gamma > 0$, and distributions $\mu_1, \ldots, \mu_t, \mu' \in D(\Omega)$ such that the following three conditions hold:*

1. $\mu = \sum_{i=1}^{t} \alpha_i \mu_i + \gamma \mu'$.
2. $\sum_{i=1}^{t} \alpha_i + \gamma = 1$.
3. $\forall i \in [t] \quad , \quad \mu_i \in \mathcal{P}$.

Note that in condition 1, we require that the convex combination of the μ_i's will be strictly smaller than μ. This is not the most general case, but it will be convenient for us to use this definition.

Let Y be the random variable defined above, and let $\mu : (F^l)^u \to [0,1]$ be the probability distribution of Y (i.e. $\mu(y) = \mathbf{Pr}[Y = y]$). We would like to show that μ is exponentially (in b) close to a convex combination of distributions, each having a certain property which will be defined shortly.

Given a probability distribution μ on $(F^l)^u$ we define for each $z \in [u]$ the distribution $\mu_z : F^l \to [0,1]$ to be the restriction of μ to the z's block. More formally, we define

$$\mu_z(y) \triangleq \sum_{y_1, \ldots, y_{z-1}, y_{z+1}, \ldots, y_u \in F^l} \mu(y_1, \ldots, y_{z-1}, y, y_{z+1}, \ldots, y_u).$$

Next, let $\alpha > 0$. We say that a distribution $\mu : (F^l)^u \to [0,1]$ is α-good if for at least $(1 - \gamma/2) \cdot u$ values of $z \in [u]$, μ_z has min-entropy at least $(0.52 - \alpha) \cdot b$. The statement that we would like to prove is that the distribution of Y is close to a convex combination of α-good distributions (see

Definition 2.1). As we will see later, this is good enough for us to be able to prove Theorem 1. The following lemma states this claim in a more precise form.

Lemma 2.2 (Main Lemma). *Let $Y = (Y_1, \ldots, Y_u)$ be the random variable defined above, and let μ be its probability distribution. Then, for any constant $\alpha > 0$, μ is $2^{-\Omega(b)}$-close to a convex combination of α-good distributions.*

We prove Lemma 2.2 in subsection 2.1. The proof of Theorem 1, which follows quite easily from Lemma 2.2, is essentially the same as in [DR05] (with modified parameters). For completeness we include the proof of Theorem 1 in Appendix A.

2.1 Proof of Lemma 2.2

In order to prove Lemma 2.2 we prove the following slightly stronger lemma.

Lemma 2.3. *Let $X = (X_1, \ldots, X_k)$ be an $(n, b)^{1:k}$-source, and let $Y = (Y_1, \ldots, Y_u)$ and μ be as in Lemma 2.2. Then for any constant $\alpha > 0$ there exists an integer $t \geq 1$, and a partition of $\{0,1\}^{n \cdot k}$ into $t + 1$ sets $W_1, \ldots, W_t, W'$, such that:*

1. $\mathbf{Pr}_X[X \in W'] \leq 2^{-\Omega(b)}$.
2. *For every $i \in [t]$ the probability distribution of $Y \,|\, X \in W_i$ (that is - of Y conditioned on the event $X \in W_i$) is α-good. In other words: for every $i \in [t]$ there exist at least $(1 - \gamma/2) \cdot u$ values of $z \in [u]$ for which*

$$H^{\infty}(Y_z | X \in W_i) \geq (0.52 - \alpha) \cdot b.$$

Before proving Lemma 2.3 we show how this lemma can be used to prove Lemma 2.2.

Proof of Lemma 2.2: The lemma follows immediately from Lemma 2.3 and from the following equality, which holds for every partition $W_1, \ldots, W_t, W'$, and for every y.

$$\mathbf{Pr}[Y = y] = \sum_{i=1}^{t} \mathbf{Pr}[X \in W_i] \cdot \mathbf{Pr}[Y = y \,|\, X \in W_i] + \mathbf{Pr}[X \in W'] \cdot \mathbf{Pr}[Y = y \,|\, X \in W'].$$

If the partition $W_1, \ldots, W_t, W'$ satisfies the two conditions of Lemma 2.3 then from Definition 2.1 it is clear that Y is exponentially (in b) close to a convex combination of α-good distributions. □

Proof of Lemma 2.3: Every random variable Y_z is a function of X, and so it partitions $\{0,1\}^{n \cdot k}$ in the following way:

$$\{0,1\}^{n \cdot k} = \bigcup_{y \in \{0,1\}^n} (Y_z)^{-1}(y),$$

where $(Y_z)^{-1}(y) \triangleq \left\{x \in \{0,1\}^{n \cdot k} \,|\, Y_z(x) = y\right\}$. For each $z \in [u]$ we define the set

$$\begin{aligned} B_z &\triangleq \bigcup_{\left\{y \,|\, \mathbf{Pr}[Y_z = y] > 2^{-(0.52 - \alpha/2) \cdot b}\right\}} (Y_z)^{-1}(y) \\ &= \left\{x' \in \{0,1\}^{n \cdot k} \;\middle|\; \mathbf{Pr}_X[Y_z(X) = Y_z(x')] > 2^{-(0.52 - \alpha/2) \cdot b}\right\}. \end{aligned}$$

Intuitively, B_z contains all values of x that are "bad" for Y_z, where in "bad" we mean that $Y_z(x)$ is obtained with relatively high probability in the distribution $Y_z(X)$.

Definition 2.4 (good triplets). *Let $(z_1, z_2, z_3) \in [u]^3$ be a triplet of seed values. Since each seed value is actually a vector in F^k we can write each z_i ($i = 1, 2, 3$) as a vector $(z_{i1}, \ldots, z_{ik})$, where each z_{ij} is an integer in the range $0...(p-1)$. We say that the triplet (z_1, z_2, z_3) is* **good** *if the following two conditions hold:*

1. *For all $2 \leq j \leq k$, $\quad z_{1j} = z_{2j} = z_{3j}$.*
2. *There exists a positive integer $0 < a < p$ such that $z_{21} = z_{11} + a$ and $z_{31} = z_{11} + 2a$, where the equalities are over $F = \mathrm{GF(p)}$.*

That is, the vectors z_1, z_2, z_3 are identical in all coordinates different from one, and their first coordinates form an arithmetic progression of length three in $F = \mathrm{GF(p)}$.

The next two claims are the place where our analysis differs from that of [LRVW03] and [DR05]. We devote Section 3 to the proofs of these two claims. The first claim shows that the intersection of the "bad" sets $B_{z_1}, B_{z_2}, B_{z_3}$ for a good triplet (z_1, z_2, z_3) is small:

Claim 2.5. *For every good triplet (z_1, z_2, z_3) it holds that*

$$\mathbf{Pr}_X[X \in B_{z_1} \cap B_{z_2} \cap B_{z_3}] \leq C_\alpha \cdot 2^{-(\alpha/4)\cdot b},$$

where C_α is a constant depending only on α.

The second claim shows that every set of seed values who's density is larger than $\gamma/2$ contains a good triplet.

Claim 2.6. *Let $T \subset [u]$ be such that $|T| > (\gamma/2) \cdot u$. Then T contains a good triplet.*

The proof of Lemma 2.3 continues along the same lines as in [DR05]. We define for each $x \in \{0,1\}^{n \cdot k}$ a vector $\pi(x) \in \{0,1\}^u$ in the following way :

$$\forall z \in [u] \quad , \quad \pi(x)_z = 1 \iff x \in B_z.$$

For a vector $\pi \in \{0,1\}^u$, let $w(\pi)$ denote the weight of π (i.e. the number of 1's in π). Since the weight of $\pi(x)$ denotes the number of seed values for which x is "bad", we would like to somehow show that for most x's $w(\pi(x))$ is small. This can be proven by combining Claim 2.5 with Claim 2.6, as shown by the following claim.

Claim 2.7.

$$\mathbf{Pr}_X[w(\pi(X)) > (\gamma/2) \cdot u] \leq u^3 \cdot C_\alpha \cdot 2^{-(\alpha/4)\cdot b}.$$

Proof. If x is such that $w(\pi(x)) > (\gamma/2) \cdot u$ then, by Claim 2.6, we know that there exists a good triplet (z_1, z_2, z_3) such that $x \in B_{z_1} \cap B_{z_2} \cap B_{z_3}$. Therefore we have

$$\mathbf{Pr}_X[w(\pi(X)) > (\gamma/2) \cdot u] \leq \mathbf{Pr}_X[\exists \text{ a good triplet } \quad (z_1, z_2, z_3) \quad s.t \quad x \in B_{z_1} \cap B_{z_2} \cap B_{z_3}].$$

Now, using the union bound and Claim 2.5 we can bound this probability by $u^3 \cdot C_\alpha \cdot 2^{-(\alpha/4)\cdot b}$. $\square$

From Claim 2.7 we see that every x (except for an exponentially small set) is contained in at most $(\gamma/2) \cdot u$ sets B_z. The idea is now to partition the space $\{0,1\}^{n \cdot k}$ into sets of x's that have the same $\pi(x)$. If we condition the random variable Y on the event $\pi(X) = \pi_0$, where π_0 is of small weight, we will get an α-good distribution. We now explain this idea in more details. We define the following sets

$$BAD_1 \triangleq \left\{ \pi' \in \{0,1\}^u \quad | \quad w(\pi') > (\gamma/2) \cdot u \right\},$$

$$BAD_2 \triangleq \left\{ \pi' \in \{0,1\}^u \quad | \quad \mathbf{Pr}_X[\pi(X) = \pi'] < 2^{-(\alpha/2) \cdot b} \right\},$$

$$BAD \triangleq BAD_1 \cup BAD_2.$$

The set $BAD \subset \{0,1\}^u$ contains values $\pi' \in \{0,1\}^u$ that cannot be used in the partitioning process described in the last paragraph. There are two reasons why a specific value $\pi' \in \{0,1\}^u$ is included in BAD. The first reason is that the weight of π' is too large (i.e. larger than $(\gamma/2) \cdot u$), these values of π' are included in the set BAD_1. The second less obvious reason for π' to be excluded from the partitioning is that the set of x's for which $\pi(x) = \pi'$ is of extremely small probability. These values of π' are bad because we can say nothing about the min-entropy of Y when conditioned on the event $\pi(X) = \pi'$. For example consider the extreme case where there is only one $x_0 \in \{0,1\}^{n \cdot k}$ with $\pi(x_0) = \pi'$. In this case the min-entropy of Y, when conditioned on the event $X \in \{x_0\}$, is zero, even if the weight of $\pi(x_0)$ is small.

Having defined the set BAD, we are now ready to define the partition required by Lemma 2.3. Let $\{\pi^1, \ldots, \pi^t\} = \{0,1\}^u \backslash BAD$. We define the sets $W_1, \ldots, W_t, W' \subset \{0,1\}^{n \cdot k}$ as follows:

- $W' = \{x \,|\, \pi(x) \in BAD\}$.
- $\forall i \in [t] \quad , \quad W_i = \{x \,|\, \pi(x) = \pi^i\}$.

Clearly, the sets $W_1, \ldots, W_t, W'$ form a partition of $\{0,1\}^{n \cdot k}$. We will now show that this partition satisfies the two conditions required by Lemma 2.3. To prove the first part of the lemma note that the probability of W' can be bounded by (using Claim 2.7 and the union-bound)

$$\begin{aligned} \mathbf{Pr}_X[X \in W'] &\leq \mathbf{Pr}_X[\pi(X) \in BAD_1] + \mathbf{Pr}_X[\pi(X) \in BAD_2] \\ &\leq u^3 \cdot C_\alpha \cdot 2^{-(\alpha/4) \cdot b} + 2^u \cdot 2^{-(\alpha/2) \cdot b} = 2^{-\Omega(b)}. \end{aligned}$$

We now prove that $W_1, \ldots, W_t$ satisfy the second part of the lemma. Let $i \in [t]$, and let $z \in [u]$ be such that $(\pi^i)_z = 0$ (there are at least $(1 - \gamma/2) \cdot u$ such values of z). Let $y \in \{0,1\}^n$ be any value. If $\mathbf{Pr}[Y_z = y] > 2^{-(0.52 - \alpha/2) \cdot b}$ then $\mathbf{Pr}[Y_z = y \,|\, X \in W_i] = 0$ (this follows from the way we defined the sets B_z and W_i). If on the other hand $\mathbf{Pr}[Y_z = y] \leq 2^{-(0.52 - \alpha/2) \cdot b}$ then

$$\begin{aligned} \mathbf{Pr}[Y_z = y \mid X \in W_i] &\leq \frac{\mathbf{Pr}[Y_z = y]}{\mathbf{Pr}[X \in W_i]} \\ &\leq 2^{-(0.52 - \alpha/2) \cdot b} / 2^{-(\alpha/2) \cdot b} \\ &= 2^{-(0.52 - \alpha) \cdot b}. \end{aligned}$$

Hence, for all values of y we have $\mathbf{Pr}[Y_z = y \mid X \in W_i] \leq 2^{-(0.52 - \alpha) \cdot b}$. We can therefore conclude that for all $i \in [t]$, $H^\infty(Y_z | X \in W_i) \geq (0.52 - \alpha) \cdot b$. This completes the proof of Lemma 2.3. $\square$

3 Proving Claim 2.5 and Claim 2.6 Using Results From Additive Number Theory

In this section we prove Claim 2.5 and Claim 2.6. These two claims are the only place in which our analysis differs from that of [LRVW03] and [DR05]. In the proofs we use two results from additive number theory. The first is a quantitative version of Roth's theorem [Rot53] given by Bourgain [Bou99a]. The second is a Lemma of Bourgain that deals with sum-sets and difference-sets.

3.1 Proof of Claim 2.5

The proof of the claim relies on the following result from additive number theory due to Bourgain [Bou99a]. Bourgain proved this result with respect to subsets of $\mathbb{Z}^d$. However, it is easily seen from the proof that it holds for subsets of any abelian group.

Lemma 3.1 ([Bou99a]). *For every $\epsilon > 0$ there exists a constant C_ϵ such that the following holds: Let A, B be subsets of an abelian group G. Let $\Gamma \subset A \times B$, and define*

$$S \triangleq \{a + b \quad | \quad (a, b) \in \Gamma\},$$

$$D \triangleq \{a - b \quad | \quad (a, b) \in \Gamma\}.$$

Suppose that there exists $K > 0$ such that $|A|, |B|, |S| \leq K$, then

$$|D| < C_\epsilon \cdot K^{(1/0.52)+\epsilon}.$$

Before we can apply Lemma 3.1 we need some notations. Let $U \triangleq B_{z_1} \cap B_{z_2} \cap B_{z_3}$. We define for every $i = 1, 2, 3$ the set

$$V_i \triangleq \{Y_{z_i}(x) \quad | \quad x \in U\}.$$

Next, we define a subset $\Gamma \subset V_1 \times V_3$ as follows

$$\Gamma \triangleq \{(v_1, v_3) \quad | \quad \exists x \in U \quad s.t \quad Y_{z_1}(x) = v_1 \quad and \quad Y_{z_3}(x) = v_3\}.$$

We now define the sets S and D as in Lemma 3.1, where the roles of A and B are taken by V_1 and V_3.

$$S \triangleq \{v_1 + v_3 \quad | \quad (v_1, v_3) \in \Gamma\},$$

$$D \triangleq \{v_1 - v_3 \quad | \quad (v_1, v_3) \in \Gamma\}.$$

We also define

$$K \triangleq 2^{(0.52-\alpha/2)\cdot b},$$

and

$$\hat{U} \triangleq \{x_1 \in \{0,1\}^n \quad | \quad \exists x_2, \ldots, x_k \in \{0,1\}^n \quad s.t \quad (x_1, \ldots, x_k) \in U\}.$$

the following claim states several facts that, when combined, will enable us to use Lemma 3.1 on the sets we have defined.

Claim 3.2. *the following is true:*

1. $|V_1|, |V_2|, |V_3| \leq K$.

2. $|S| \leq |V_2| \leq K$.

3. $|\hat{U}| \leq |D|$.

Proof. 1. Follows directly from the definition of the sets B_{z_i} and V_i. Each value $v \in V_i$ is a "heavy element" of the random variable Y_{z_i}. That is, the probability that $Y_{z_i} = v$ is at least $2^{-(0.52-\alpha/2)\cdot b} = K^{-1}$, and so there can be at most K such values.

2. What we will show is that the set S is contained in the set $2V_2 \triangleq \{2 \cdot v \,|\, v \in V_2\}$ (these two sets are actually equal, but we will not need this fact). To see this, recall that from the definition of a good triplet we have that for every $x \in \{0,1\}^{n \cdot k}$

$$Y_{z_1}(x) + Y_{z_3}(x) = 2 \cdot Y_{z_2}(x). \tag{1}$$

Let $v \in S$. From the definition of S (and of Γ) we know that there exists $x \in U$ and $v_1 \in V_1, v_3 \in V_3$ such that $Y_{z_1}(x) = v_1, Y_{z_3}(x) = v_3$ and $v = v_1 + v_3$. From Eq.1 we now see that $v = 2 \cdot Y_{z_2}(x)$, and therefore $v \in 2V_2$. The inequality now follows from the fact that $|V_2| = |2V_2|$.

3. This follows in a similar manner to 2. We will show that the set $\hat{U}$ is contained in the set $c \cdot D \triangleq \{c \cdot v \,|\, v \in D\}$, for some $0 < c < p$ (again, the two sets are actually equal, but we will not use this fact). From the definition of a good triplet we know that there exists $0 < c < p$ such that for every $x = (x_1, \ldots, x_k) \in \{0,1\}^{n \cdot k}$

$$c \cdot (Y_{z_1}(x) - Y_{z_3}(x)) = x_1. \tag{2}$$

Let $x_1 \in \hat{U}$. From the definition of $\hat{U}$ we know that there exist $x_2, \ldots, x_k \in \{0,1\}^n$ such that $x = (x_1, \ldots, x_k) \in U$. From Eq.2 it follows that $x_1 \in c \cdot D$, since $Y_{z_1}(x) - Y_{z_3}(x) \in D$ by definition.

□

Let

$$\epsilon \triangleq \frac{\alpha}{4} \cdot \frac{1}{(0.52 - \alpha/2)}.$$

From the first two parts of Claim 3.2 we see that we can apply Lemma 3.1 with $A = V_1$ and $B = V_3$ to get that

$$|D| < C_\epsilon \cdot K^{(1/0.52)+\epsilon},$$

(where C_ϵ depends only on ϵ, which in turn depends only on α). Substituting the values of ϵ and K we see that (we can write C_α instead of C_ϵ)

$$\begin{aligned} |D| &< C_\alpha \cdot 2^{b \cdot (0.52-\alpha/2) \cdot (1/0.52+\epsilon)} \\ &= C_\alpha \cdot 2^{b \cdot \left(1+\epsilon(0.52-\alpha/2) - \frac{\alpha}{2} \cdot \frac{1}{0.52}\right)} \\ &= C_\alpha \cdot 2^{b \cdot \left(1+\frac{\alpha}{4} - \frac{\alpha}{2} \cdot \frac{1}{0.52}\right)} \\ &\leq C_\alpha \cdot 2^{b \cdot \left(1-\frac{\alpha}{4}\right)}, \end{aligned} \tag{3}$$

where C_α depends only on α.

Using the third part of Claim 3.2 and Eq. 3 we conclude that

$$|\hat{U}| \leq |D| \leq C_\alpha \cdot 2^{b \cdot \left(1-\frac{\alpha}{4}\right)}. \tag{4}$$

We can therefore bound the probability of U by

$$\mathbf{Pr}_X[X \in U] \leq \mathbf{Pr}_{X_1}[X_1 \in \hat{U}] \leq 2^{-b} \cdot |\hat{U}| \leq 2^{-b} \cdot \left(C_\alpha \cdot 2^{b \cdot \left(1 - \frac{\alpha}{4}\right)}\right) = C_\alpha \cdot 2^{-(\alpha/4) \cdot b},$$

(the second inequality follows from the fact that the min-entropy of X_1 is at least b). This completes the proof of Claim 2.5. □

3.2 Proof of Claim 2.6

The claim follows from Roth's theorem [Rot53] on arithmetic progressions of length three. For our purposes we require the quantitative version of this theorem as proven by Bourgain [Bou99b].

Theorem 3.3 ([Bou99b]). *Let $\delta > 0$, let $N \geq \exp(\delta^{-2})$ and let $A \subset \{1, \ldots, N\}$ be a set of size at least δN. Then A contains an arithmetic progression of length three.*

Each element in T is a vector in F^k (recall that $F = \mathrm{GF(p)}$). A simple counting argument shows that T must contain a subset T' such that

1. $|T'| > (\gamma/2) \cdot p$.
2. All vectors in T' are identical in all coordinates different than one.

Using Theorem 3.3 and using the fact that p was chosen to be greater than $\exp(\gamma^{-2})$, we conclude that there exists a triplet in T' such that the first coordinates of this triplet form an arithmetic progression. This is a good triplet, since in T' the vectors are identical in all coordinates different than one. □

4 Acknowledgements

The authors would like to thank Ran Raz, Omer Reingold and Avi Wigderson for helpful conversations. A.S. would also like to thank Oded Goldreich for helpful discussions on related problems.

References

[BIW04] Boaz Barak, Russell Impagliazzo, and Avi Wigderson. Extracting randomness using few independent sources. In *45th Symposium on Foundations of Computer Science (FOCS 2004)*, pages 384–393, 2004.

[Bou99a] Jean Bourgain. On the dimension of kakeya sets and related maximal inequalities. *Geom. Funct. Anal.*, (9):256–282, 1999.

[Bou99b] Jean Bourgain. On triples in arithmetic progression. *Geom. Funct. Anal.*, (9):968–984, 1999.

[DR05] Zeev Dvir and Ran Raz. Analyzing linear mergers. *Electronic Colloquium on Computational Complexity (ECCC)*, (025), 2005.

[Gow01] Timothy Gowers. A new proof of szemeredi's theorem. *Geom. Funct. Anal.*, (11):465588, 2001.

[LRVW03] Chi-Jen Lu, Omer Reingold, Salil Vadhan, and Avi Wigderson. Extractors: optimal up to constant factors. In *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 602–611. ACM Press, 2003.

[Raz05] Ran Raz. Extractors with weak random seeds. *STOC 2005 (to appear)*, 2005.

[Rot53] Klaus F Roth. On certain sets of integers. *J. Lond. Math. Soc.*, (28):104–109, 1953.

[Sha02] Ronen Shaltiel. Recent developments in extractors. *Bulletin of the European Association for Theoretical Computer Science*, 77:67–95, 2002.

[Sze75] Endre Szemeredi. On sets of integers containing no k elements in arithmetic progression. *Acta. Arith.*, (27):299–345, 1975.

[TS96] Amnon Ta-Shma. On extracting randomness from weak random sources (extended abstract). In *STOC '96: Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 276–285. ACM Press, 1996.

A Proof of Theorem 1

let $Y = (Y_1, \ldots, Y_u)$ and μ be as in Lemma 2.2. Using Lemma 2.2 we can write μ as a convex combination of distributions

$$\mu = \sum_{i=1}^{t} \alpha_i \mu_i + \gamma' \mu', \tag{5}$$

with $\gamma' = 2^{-\Omega(b)}$, and such that for every $i \in [t]$ the distribution μ_i is α-good. Denote with $(\mu_i)_z$ the restriction of the distribution μ_i to the block indexed by z. It follows that for at least $(1 - \gamma/2) \cdot u$ values of $z \in [u]$, the distribution $(\mu_i)_z$ has min-entropy at least $b' = (0.52 - \alpha) \cdot b$. Next, define for every $z \in [u]$ the set $H_z \subset [t]$ as follows:

$$H_z \triangleq \{i \in [t] \,:\, \mathrm{H}^{\infty}\left((\mu_i)_z\right) < b'\}.$$

That is, $H_z \subset [t]$ is the set of indices of all distributions among $\{\mu_1, \ldots, \mu_t\}$, for which $(\mu_i)_z$ has min-entropy smaller than b'. Additionally, define for every $z \in [u]$,

$$e_z \triangleq \sum_{i \in H_z} \alpha_i.$$

Claim A.1. *Let $\Delta(Y_z, (n, b'))$ denote the minimal (statistical) distance between Y_z and an (n, b')-source. Then for every $z \in [u]$*

$$\Delta(Y_z, (n, b')) \leq e_z + \gamma'.$$

Proof. For every $z \in [u]$ let $\mu_z(y) = \mathbf{Pr}[Y_z = y]$ be the probability distribution of Y_z. From Eq.5 we can write μ_z as a convex combination

$$\begin{aligned}\mu_z &= \sum_{i=1}^{t} \alpha_i \cdot (\mu_i)_z + \gamma' \mu'_z \\ &= \left(\sum_{i \notin H_z} \alpha_i \cdot (\mu_i)_z\right) + \left(\sum_{i \in H_z} \alpha_i \cdot (\mu_i)_z + \gamma' \mu'_z\right) \\ &= (1 - e_z - \gamma') \cdot \mu'' + (e_z + \gamma') \cdot \mu''',\end{aligned}$$

where μ'' is the probability distribution of an (n, b') source (as a convex combination of (n, b')-sources is an (n, b')-source), and μ''' is some other distribution. Clearly, the statistical distance $\Delta(\mu_z, \mu'')$ is at most $e_z + \gamma'$, and since μ'' is an (n, b') source, we have that $\Delta(Y_z, (n, b')) \leq e_z + \gamma'$. □

Claim A.2. *Let Z be a random variable uniformly distributed over $[u]$. Then, the expectation of e_Z is at most $\gamma/2$:*

$$\mathbb{E}[e_Z] \leq \gamma/2.$$

Proof. For each $i \in [t]$ define the following indicator random variable

$$\chi_i = \begin{cases} 1, & i \in H_Z; \\ 0, & i \notin H_Z. \end{cases}$$

We can thus write

$$e_Z = \sum_{i=1}^{t} \chi_i \cdot \alpha_i.$$

By linearity of expectation we have

$$\mathbb{E}[e_Z] = \sum_{i=1}^{t} \mathbb{E}[\chi_i] \cdot \alpha_i,$$

and since for each $i \in [t]$ we have that

$$\mathbb{E}[\chi_i] = \mathbf{Pr}_Z[i \in H_Z] < \gamma/2$$

(this follows from the fact that each μ_i is α-good), we conclude that

$$\mathbb{E}[e_Z] \leq \epsilon \cdot \sum_{i=1}^{t} \alpha_i \leq \gamma/2.$$

□

Combining Claim A.1 and Claim A.2, and recalling that $\gamma' = 2^{-\Omega(b)}$, we see that

$$\mathbb{E}[\Delta(Y_Z, (n, b'))] \leq \mathbb{E}[e_Z] + \gamma' \leq \gamma/2 + 2^{-\Omega(b)},$$

where the expectations are taken over Z, which is chosen uniformly in $[u]$. Now, for values of b larger than some constant b_0, this expression is smaller than γ. This completes the proof of Theorem 1. □

ECCC ISSN 1433-8092
http://www.eccc.uni-trier.de/eccc
ftp://ftp.eccc.uni-trier.de/pub/eccc
ftpmail@ftp.eccc.uni-trier.de, subject 'help eccc'