AN IMPROVED ANALYSIS OF LINEAR MERGERS

Zeev Dvir and Amir Shpilka

Abstract. Mergers are procedures that, with the aid of a short random string, 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 were first introduced by Ta-Shma [28th STOC, pp. 276-285, 1996] and have proven to be a very useful tool in explicit constructions of *extractors* and *condensers*. In this work we present a new analysis of the merger construction of Lu et al [35th STOC, pp. 602-611, 2003]. We prove that the merger's output is close to a distribution with min-entropy rate of at least $\frac{6}{11}\delta$. We show that the distance from this distribution is polynomially related to the number of additional random bits that were used by the merger (i.e its seed). We are also able to prove a bound of $\frac{4}{7}\delta$ on the min-entropy rate at the cost of increasing the statistical error. Both results are improvements to the previous known lower bound of $\frac{1}{2}\delta$ (however, in the $\frac{1}{2}\delta$ result the error decreases exponentially in the length of the seed). To obtain our results we deviate from the usual linear algebra methods that were used by Lu et al and introduce techniques from additive number theory.

Keywords. Mergers, Extractors, Kakeya, Randomness.

Subject classification. 68W20 Randomized algorithms

1. Introduction

Mergers are procedures that take as input k samples, taken from k (possibly dependent) random sources, each ranging over n-bit long strings. It is assumed that one of these random sources, whose index is unknown, is sufficiently random, in the sense that it has min-entropy at least δn (A source has min-entropy at least 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 (1) to minimize the seed length, (2) to maximize the min-entropy of the output, and (3) 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 (1996), 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. For a more detailed discussion of extractors see Shaltiel (2002). Recently, Lu, Reingold, Vadhan & Wigderson (2003) gave a very simple and beautiful construction of mergers based on Locally-Decodable-Codes. This construction was used in Lu *et al.* (2003) as a building block in an explicit construction of extractors with nearly optimal parameters. More recently, Raz (2005) generalized the construction of Lu *et al.* (2003), and showed how this construction (when combined with other techniques) can be used to construct *condensers* with constant seed length. (A condenser is a function that 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.) The analysis of the merger constructed in Raz (2005) was subsequently refined in Dvir & Raz (2005).

The merger constructed by Lu *et al.* (2003) 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 $\frac{1}{2}b$. Loosely speaking, the output of the merger is computed as follows: treat each input block as a vector in the vector space $\mathbb{F}^m$, where $\mathbb{F}$ is some small finite field, and output a uniformly chosen linear combination of these k vectors. The k scalars defining this linear combination are the seed of the merger. The analysis of this construction is based on the following simple idea: In every set of linear combinations with density larger than $\frac{1}{|F|}$ there exist two linear combinations that, when put together, determine the 'good' source (that is, the 'good' source can be computed as a linear combination from both of them deterministically). More precisely, such sets must contain two linear combinations that differ only in the coefficient multiplying the 'good' source. 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 result we get that for most seed values (at least $1 - \frac{1}{|F|}$ fraction) 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 additive-number theory. The first is Roth's Theorem on arithmetic progressions of length three (Roth 1953). This theorem states that there exists a function $\delta(N)$ that tends to zero when N goes to infinity such that in every subset of $\{1, \ldots, N\}$, that has density at least $\delta(N)$, there exists an arithmetic progression of length three. For our purposes we use a quantitative version of this theorem proven by Bourgain (1999b), that gives the best bound on $\delta(N)$ known today. The second result that we rely on is a lemma of Bourgain (1999a) that deals with "sum-sets" and "difference-sets" of integers (we actually use a stronger version of the lemma that was proved by Katz & Tao (1999)). Roughly speaking, the lemma says that if A, B are two subsets of integers and their sum-set $A + B = \{a + b \mid a \in A,\ b \in B\}$ is very small, then their difference-set $A - B = \{a - b \mid a \in A,\ b \in B\}$ cannot be very large (for a precise formulation see Section 4). We 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, Impagliazzo & Wigderson (2004) uses results from this field to construct multi-source extractors.

The analysis in our case is somewhat more involved then the one in Lu *et al.* (2003). Let us identify a fixed linear combination of the source blocks with a vector of coefficients. Each such vector is a "seed" of the merger. Let us also assume that the first source is the one with entropy at least b (i.e the "good" source). The analysis of Lu *et al.* (2003) argues that every pair of seeds that differ only in the first coordinate cannot be both "bad" (a seed is considered "bad" if the output of the merger on this seed has entropy lower than $\frac{1}{2}b$). This is because together they determine the "good" source. In the new analysis a seed is considered "bad" if the entropy of the output of the merger on this seed is lower than $\frac{6}{11}b$. The general approach is the same as in Lu *et al.* (2003), we will show that every set of seeds of density larger than some γ must contain at least one "good" seed. The argument for showing this proceeds in two steps: In the first step we use Roth's Theorem to claim that every large enough set of seeds contains three seeds which are identical in all coordinates other than the first coordinate, and such that the values appearing in the first coordinate in each seed form an arithmetic progression of length three. The second step of the analysis uses the lemma of Katz & Tao (1999) to claim that at least one of the seeds in this triple must be "good".

To see why the lemma of Katz & Tao (1999) is relevant consider three seeds s_1, s_2, s_3 of the form just described (each s_i represents a vector of coefficients). Let Y_1, Y_2, Y_3 denote the random variables representing the output of the merger on these three seeds respectively. Since $s_1 + s_3 = 2s_2$ and $s_1 - s_3 \in \mathbb{F} \times \{0\}^{k-1}$ we have that the sum $Y_1 + Y_3$ equals $2Y_2$ and that the difference $Y_1 - Y_3$ is equal to some constant times the "good" source. If all three seeds were "bad" we could construct two sets (namely the supports of Y_1 and Y_3) such that (a) their sum-set is small, since Y_2 has low entropy (few values) and (b) their difference-set is large, since the good source has high entropy (many values) . Choosing the right parameters we get a contradiction to the lemma of Katz & Tao (1999) stated above. To summarize:

1. By Roth's Theorem, in every set of seeds (linear combinations) with density larger than some constant γ we can find three elements with some nice structure (arithmetic progression in the first coordinate and identical in all the rest).

2. Using Katz & Tao (1999), we show that every three seeds with this structure cannot be all "bad".

Combining these two facts we conclude that at most an γ fraction of the seeds can be "bad", and the result follows.

The end result of this new analysis is that, assuming the min-entropy of the "good" source is b, the output of the merger described above is close (in statistical distance) to a distribution with min-entropy at least $\frac{6}{11}b$ improving over the lower bound of $\frac{1}{2}b$ established by Lu *et al.* (2003). Using a more involved argument (using longer arithmetic progressions) we are able to show that the output distribution is close (but with a worse bound on the distance) to a distribution with min-entropy at least $\frac{4}{7}b$. One drawback of our analysis is that in our first result the length of the seed is required to be $O(k \cdot \gamma^{-2})$ in order for the output distribution to be γ-close to a distribution with high min-entropy, where in the conventional analysis (i.e. in Lu *et al.* 2003) the seed length can be as short as $O(k \cdot \log(\gamma^{-1}))$. In our second result we demand that the seed is even longer.[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 Raz (2005). 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 Raz (2005) this problem is addressed by a more complicated merger construction whose output length is shorter than n. our analysis shows that the more simple construction of Lu *et al.* (2003) could be used instead, since its output min-entropy rate is larger than one-half.

[1] To understand the tradeoff between the distance and the seed length in our second result the reader should read Theorem 2.7.

Organization. In Section 2 we give a precise formulation of the problem and state our results, as well as discussing the relation between linear mergers and the Kakeya problem. In Section 3 we present in detail our analysis of the linear merger construction and prove the 6/11 bound. The analysis presented in Section 3 relies on two central claims, which we prove in Section 4. The improved bound of 4/7 is proved in Section 5. Section 6 deals with encoding binary inputs as vectors over $\mathbb{F}_p$.

2. Formal Setting

2.1. Somewhere-Random-Sources. Let Γ denote a finite alphabet. A Γ^n *random source* is a random variable X that takes values in Γ^n. We denote by $\mathrm{supp}(X) \subset \Gamma^n$ the support of X (i.e. the set of values on which X has non-zero probability). For two Γ^n random sources X and Y, we define the statistical distance (or simply distance) between X and Y to be

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

We say that a Γ^n random source X has min-entropy $\geq b$ if for every $x \in \Gamma^n$ the probability for $X = x$ is at most 2^{-b}.

Definition 2.1. Min-entropy.

Let X be a Γ^n random source. The min-entropy of X is defined as

$$\mathrm{H}^{\infty}(X) \triangleq \min_{x\in\mathrm{supp}(X)} \log_2\left(\frac{1}{\mathbf{Pr}[X=x]}\right).$$

Definition 2.2. (Γ^n, b)-Source.

We say that X is a (Γ^n, b)-source, if X is a Γ^n random source, and $\mathrm{H}^{\infty}(X) \geq b$.

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

Definition 2.3. $(\Gamma^n, b)^{1:k}$-Source.

A k-places-somewhere-(Γ^n, b)-source, or shortly, an $(\Gamma^n, b)^{1:k}$-source, is a random variable $X = (X_1, \ldots, X_k)$, such that every X_i is a Γ^n random source, and at least one X_i is of min-entropy $\geq b$.

We note that any merger construction that applies to the sources of Definition 2.3 extends also to a convex combination of such sources.

2.2. Mergers. A merger is a function transforming a $(\Gamma^n, b)^{1:k}$-source into a source that is γ-close (i.e. it has statistical distance $\leq \gamma$) to an (Γ^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 a (Γ^m, b')-source. A merger is *explicit* if it can be computed in polynomial time.

Definition 2.4. Merger.

A function $M : \{0,1\}^d \times (\Gamma^n)^k \to \Gamma^m$ is a $[d, (\Gamma^n, b)^{1:k} \mapsto (\Gamma^m, b') \sim \gamma]$-merger if for every $(\Gamma^n, b)^{1:k}$-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 (Γ^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 (Γ^m, b')-source is $\leq \gamma$.

We now present the merger of Lu *et al.* (2003), which we wish to analyze. We will be interested only in the case were the underlying field is $\mathbb{F}_p$ for a prime p.

Construction 2.5. (Lu *et al.* 2003).

Let n, k be integers, p a prime number. We define a function

$$M : \{0,1\}^d \times \left(\mathbb{F}_p^n\right)^k \to \mathbb{F}_p^n,$$

with

$$d = \lfloor k \cdot \log_2 p \rfloor,$$

in the following way: Let $\phi : \{0,1\}^d \mapsto \mathbb{F}_p^k$ be some injective mapping (such a ϕ exists since $2^d \leq p^k$ and can be computed in polynomial time). We map each seed $z \in \{0,1\}^d$ into the vector $\phi(z) = (z_1, \ldots, z_k) \in \mathbb{F}_p^k$. Let $x = (x_1, \ldots, x_k) \in \left(\mathbb{F}_p^n\right)^k$. The value of $M(z, x)$ is computed as follows:

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

where the operations are preformed in the vector space $\mathbb{F}_p^n$. That is, the merger M outputs a different linear combination of the blocks of x for every seed z.

2.3. Our Results. Our first theorem improves the bound of $1/2$ on the min-entropy rate of the merger from Construction 2.5 to $6/11$. We write $\exp(f)$ to denote $2^{O(f)}$.

THEOREM 2.6. Let $0 < \gamma < 1$ be any constant, $k > 0$ a constant integer, and let p be a prime larger than $\exp(\gamma^{-2})$. Let

$$M : \{0,1\}^d \times \left(\mathbb{F}_p^n\right)^k \to \mathbb{F}_p^n,$$

be as in Construction 2.5, where $d = \lfloor k \cdot \log_2 p \rfloor$. 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 $[d, (\mathbb{F}_p^n, b)^{1:k} \mapsto (\mathbb{F}_p^n, b') \sim \gamma]$-strong merger with

$$b' = (6/11 - \alpha) \cdot b.$$

From Theorem 2.6 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 m, there is a prime between m and $2m$. Therefore we can take p to be $O\left(\exp(\gamma^{-2})\right)$ and have that the length of the random seed is

$$d = \lfloor k \cdot \log_2 p \rfloor = O\left(k \cdot \gamma^{-2}\right)$$

bits long. Hence, for constant γ and k, the length of the random seed used by the merger is constant.

We can further improve the bound on the min-entropy rate to $4/7$ at the cost of worse error dependency. We write $a \uparrow b$ for a^b. Also $a \uparrow b \uparrow c$ should be interpreted as $a \uparrow (b \uparrow c)$.

THEOREM 2.7. Let $0 < \gamma < 1$ be any constant, $k > 0$ a constant integer, and let p be a prime larger than $F(\gamma) \triangleq 2 \uparrow 2 \uparrow (\gamma/2)^{-1} \uparrow 2 \uparrow 2 \uparrow 16$. Let

$$M : \{0,1\}^d \times \left(\mathbb{F}_p^n\right)^k \to \mathbb{F}_p^n,$$

be as in Construction 2.5, where $d = \lfloor k \cdot \log_2 p \rfloor$. 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 $[d, (\mathbb{F}_p^n, b)^{1:k} \mapsto (\mathbb{F}_p^n, b') \sim \gamma]$-strong merger with

$$b' = (4/7 - \alpha) \cdot b.$$

2.4. Relation to the Kakeya problem. The Kakeya problem is a long standing open problem in mathematics: A set $S \subset \mathbb{R}^l$ is called Besicovitch if it contains a unit line segment in every direction. It is conjectured, e.g. Bourgain (1991, 1999a); Wolff (1995), that such a set must have Hausdorff dimension l. A weaker version of the conjecture asserts that these sets must have upper Minkowski dimension l (see Bourgain 1991 for definitions of Hausdorff and Minkowski dimension).

The finite field analog of the problem is the following. Let $\mathbb{F}$ be a finite field. A set $S \subset \mathbb{F}^l$ is called Besicovitch if for every $u \in \mathbb{F}^l$ there exist $x \in S$ such that the line $x + t \cdot u$, where t runs over all the elements of $\mathbb{F}$, is contained in S. The Kakeya set conjecture for finite fields asserts that every Besicovitch set has cardinality $|\mathbb{F}|^{l-o(1)}$ (see Mockenhaupt & Tao 2004). Informally, this means that it is impossible to compress lines in distinct directions into a small set. This conjecture is proven in two dimensions but is open in higher dimensions. The best bound is $|S| \geq |\mathbb{F}|^{l/\alpha}$, where $1 < \alpha < 2$ satisfies $\alpha^3 - 4\alpha + 2 = 0$, specifically $\alpha = 1.67513....$ (see Katz & Tao 2002).

Consider the merger of Construction 2.5. It takes a random linear combination of the k random variables $X_1, \ldots, X_k$. Assume w.l.o.g. that the k-th random variable is completely random in $\mathbb{F}^l$. Then when we run over all the linear combinations we get all vectors of the form $(\sum_{i=2}^{k-1} z_i \cdot X_i) + z_k \cdot X_k$ where the z_i-s are elements of $\mathbb{F}$. Fixing $z_1, \ldots, z_{k-1}$ we get the line in direction X_k. As X_k is completely random we get that the output of this merger is a Besikovitch set. Thus the Kakeya conjecture asserts that the output size is at least $|\mathbb{F}|^{l-o(1)}$. This shows the intimate connection of linear mergers to the Kakeya problem.

However for our purpose it is not enough to obtain a lower bound on the size of the output of the merger. We have to show that the output is close to a distribution with high *min-entropy* and not just to a distribution with a large support. Moreover, we are also interested in the case where X_k is not fully random but rather has high min-entropy. It turns out though that the techniques that are used in order to prove some of the lower bounds on the size of Besikovitch sets over finite fields can be applied to our scenario as well, after some modifications. We stress again that we do not know how to prove a general theorem that says that every lower bound for the Kakeya problem yields a lower bound on the min-entropy of this merger.

3. Analysis of Construction 2.5

In this section we present our improved analysis of Construction 2.5, and prove Theorem 2.6. The analysis will go along the same lines as in Dvir & Raz (2005) and will differ from it in two claims that we will prove in Section 4. We begin with some notations that will be used throughout the paper.

3.1. Notations. For an integer n, we write $[n] \triangleq \{1, 2, \ldots, n\}$. Let $0 < \gamma < 1$ be any constant, and let $p \geq \exp(\gamma^{-2})$ be a prime number. Let $X = (X_1, \ldots, X_k) \in \left(\mathbb{F}_p^n\right)^k$ be a somewhere $(\mathbb{F}_p^n, b)$-source, and let us assume w.l.o.g. that $\mathrm{H}^{\infty}(X_1) \geq b$. Let $M : \{0,1\}^d \times \left(\mathbb{F}_p^n\right)^k \to \mathbb{F}_p^n$, be as in Construction 2.5, where $d = \lfloor k \cdot \log_2 p \rfloor$. Our goal is to analyze the min-entropy of $M(Z, X)$ where Z will denote a random variable uniformly distributed over $\{0,1\}^d$. In particular, we would like to show that the random variable $M(Z, X)$ is γ-close to having min-entropy $\geq (6/11 - \alpha) \cdot b$ for all constant α (see Theorem 2.6 for the exact order of quantifiers).

We can extend the function M to be defined over $\mathbb{F}_p^k \times \left(\mathbb{F}_p^n\right)^k$ in a natural way by considering all possible linear combinations instead of just the 2^d indexed by $\phi\left(\{0,1\}^d\right)$ (see Construction 2.5 for the definition of ϕ). In the rest of this section we will analyze the output of M when the seed is uniform over $\mathbb{F}_p^k$. Later, in Section 3.3, in the proof of Theorem 2.6, we will use the results of this section to claim that the output behaves roughly the same when the seed is distributed over $\{0,1\}^d$.

For every $z \in \mathbb{F}_p^k$ we denote by $Y_z \triangleq M(z, X)$ the random variable given by the output of M on the fixed seed value z. Let $u = p^k$ be the number of different seed values. Let $Y \triangleq (Y_1, \ldots, Y_u) \in (\mathbb{F}_p^n)^u$. The random variable Y is a deterministic function of X, and is comprised of u blocks. The block Y_z is an $\mathbb{F}_p^n$ random source representing the output of the merger on the fixed seed value z. 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 3.1. 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. $\forall i \in [t] \quad , \quad \mu^i \in \mathcal{P}$.

Let Y be the random variable defined above, and let $\mu : (\mathbb{F}_p^n)^u \to [0,1]$ 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 $(\mathbb{F}_p^n)^u$ we define for each $z \in [u]$ the distribution $\mu_z : \mathbb{F}_p^n \to [0,1]$ to be the restriction of μ to the z's block. More formally, we define

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

Definition 3.2. α-good distribution.

We say that a distribution $\mu : (\mathbb{F}_p^n)^u \to [0,1]$ is α-good if for at least $(1-\gamma/2) \cdot u$ values of $z \in [u]$, μ_z has min-entropy at least $(6/11 - \alpha) \cdot b$.

The statement that we would like to prove is that the distribution of Y is close to a convex combination of α-good distributions. As we will see later, this will be enough to prove Theorem 2.6.

Lemma 3.3. 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 3.3 in Section 3.2. The proof of Theorem 2.6, which follows quite easily from Lemma 3.3, is very similar to the proof appearing in Dvir & Raz (2005) and is deferred to Section 3.3.

3.2. Proof of Lemma 3.3.

In order to prove Lemma 3.3 we prove the following slightly stronger lemma.

Lemma 3.4. Let $X = (X_1, \ldots, X_k)$ be an $(\mathbb{F}_p^n, b)^{1:k}$-source, and let Y and μ be as in Lemma 3.3. Then for any constant $\alpha > 0$ there exists an integer $t \geq 1$, and a partition of $(\mathbb{F}_p^n)^k$ 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

$$H^\infty(Y_z | X \in W_i) \geq (6/11 - \alpha) \cdot b.$$

Before proving Lemma 3.4 we show how this lemma can be used to prove Lemma 3.3.

Proof of Lemma 3.3: The lemma follows immediately from Lemma 3.4 and from the following equality, which holds for every partition $W_1, \ldots, W_t, W'$, and for every y.

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

If the partition $W_1, \ldots, W_t, W'$ satisfies the two conditions of Lemma 3.4 then from Definition 3.1 it is clear that Y is exponentially (in b) close to a convex combination of α-good distributions.

Proof of Lemma 3.4: Every random variable Y_z is a function of X, and so it partitions the set $(\mathbb{F}_p^n)^k$ in the following way:

$$(\mathbb{F}_p^n)^k = \bigcup_{y \in \mathbb{F}_p^n} (Y_z)^{-1}(y),$$

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

$$\begin{aligned} B_z &\triangleq \bigcup_{\left\{y \,\mid\, \mathbf{Pr}[Y_z = y] > 2^{-(6/11-\alpha/2)\cdot b}\right\}} (Y_z)^{-1}(y) \\ &= \left\{x' \in (\mathbb{F}_p^n)^k \;\middle|\; \mathbf{Pr}_X[Y_z(X) = Y_z(x')] > 2^{-(6/11-\alpha/2)\cdot b}\right\}. \end{aligned}$$

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 3.5. 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 $\mathbb{F}_p^k$ we can write each z_i $(i = 1, 2, 3)$ as a vector $(z_{i1}, \ldots, z_{ik})$, where each z_{ij} is in $\mathbb{F}_p$. 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 $\mathbb{F}_p$.

That is, the triplet (z_1, z_2, z_3) is good if 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 $\mathbb{F}_p$.

The next two claims are the place where our analysis differs from that of Lu *et al.* (2003) and Dvir & Raz (2005). We devote Section 4 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 3.6. For every good triplet (z_1, z_2, z_3) it holds that

$$\mathbf{Pr}_X[X \in B_{z_1} \cap B_{z_2} \cap B_{z_3}] \leq 2^{-\left(\frac{11}{12}\alpha\right)\cdot b}.$$

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

Claim 3.7. Let $T \subset [u]$ be such that $|T| > (\gamma/2) \cdot u$. Then T contains a good triplet.

We continue the proof along the same lines as in Dvir & Raz (2005). We define for each $x \in (\mathbb{F}_p^n)^k$ 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 show that for a random value of x, $w(\pi(x))$ is small with high probability. This can be proven by combining Claim 3.6 with Claim 3.7, as shown by the following claim.

Claim 3.8.

$$\mathbf{Pr}_X[w(\pi(X)) > (\gamma/2) \cdot u] \leq u \cdot (p-1) \cdot 2^{-\left(\frac{11}{12}\alpha\right)\cdot b}.$$

Proof. If x is such that $w(\pi(x)) > (\gamma/2)\cdot u$ then, by Claim 3.7, 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

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

Now, using the union bound and Claim 3.6 we can bound this probability by $u \cdot (p-1) \cdot 2^{-\left(\frac{11}{12}\alpha\right)\cdot b}$, (the number of good triplets is trivially bounded by $u \cdot (p-1)$). □

From Claim 3.8 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 $(\mathbb{F}_p^n)^k$ into sets according to the value of $\pi(x)$. 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

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

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

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

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 $\pi(X) = \pi'$.

Having defined the set BAD, we are now ready to define the partition required by Lemma 3.4. Let $\{\pi^1, \ldots, \pi^t\} = \{0,1\}^u \backslash BAD$. We define the sets $W_1, \ldots, W_t, W' \subset (\mathbb{F}_p^n)^k$ 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 $(\mathbb{F}_p^n)^k$. We will now show that this partition satisfies the two conditions required by Lemma 3.4. To prove the first part of the lemma note that the probability of W' can be bounded by (using Claim 3.8 and the union-bound)

$$\begin{aligned}\mathbf{Pr}_X[X \in W'] &\leq \mathbf{Pr}_X[\pi(X) \in BAD_1] + \mathbf{Pr}_X[\pi(X) \in BAD_2] \\ &\leq u \cdot (p-1) \cdot 2^{-\left(\frac{11}{12}\alpha\right)\cdot b} + 2^u \cdot 2^{-(\alpha/2)\cdot b} = 2^{-\Omega(b)}\end{aligned}$$

(recall that $u = p^k$ is a constant). We now prove that $W_1, \ldots, W_t$ satisfy the second part of the lemma. Let $i \in [t]$. We know that for at least $(1 - \gamma/2) \cdot u$ values of $z \in [u]$ it holds that $(\pi^i)_z = 0$. Let $z \in [u]$ be such that $(\pi^i)_z = 0$. Let $y \in \mathbb{F}_p^n$ be any value. If $\mathbf{Pr}[Y_z = y] > 2^{-(6/11-\alpha/2)\cdot b}$ 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 $\mathbf{Pr}[Y_z = y] \le 2^{-(6/11-\alpha/2)\cdot b}$ then

$$\begin{aligned}\mathbf{Pr}[Y_z = y \mid X \in W_i] &\le \frac{\mathbf{Pr}[Y_z = y]}{\mathbf{Pr}[X \in W_i]} \\ &\le 2^{-(6/11-\alpha/2)\cdot b}/2^{-(\alpha/2)\cdot b} \\ &= 2^{-(6/11-\alpha)\cdot b}.\end{aligned}$$

Hence, for all values of y we have $\mathbf{Pr}[Y_z = y \mid X \in W_i] \le 2^{-(6/11-\alpha)\cdot b}$. We can therefore conclude that for all $i \in [t]$, $H^{\infty}(Y_z | X \in W_i) \ge (6/11 - \alpha) \cdot b$ for at least $(1-\gamma/2)\cdot u$ values of $z \in [u]$. This completes the proof of Lemma 3.4. □

3.3. Proof of Theorem 2.6. Let $Y = (Y_1, \dots, Y_u)$ and μ be as in Lemma 3.3. Using Lemma 3.3 we can write μ as a convex combination of distributions

$$\mu = \sum_{i=1}^{t} \alpha_i \mu^i + \gamma' \mu', \tag{3.1}$$

with $\gamma' = 2^{-\Omega(b)}$, and such that for every $i \in [t]$ the distribution μ^i is α-good. That is, for at least $(1-\gamma/2)\cdot u$ values of $z \in [u]$, the distribution $(\mu^i)_z$ has min-entropy at least $b' = (6/11 - \alpha)\cdot b$ (when writing $(\mu^i)_z$, the superscript i denotes the index of the distribution, and the subscript z denotes its restriction to the block indexed by z). 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 $\{\mu^1, \dots, \mu^t\}$, 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 3.9. Let $\Delta(Y_z, (\mathbb{F}_p^n, b'))$ denote the minimal distance between Y_z and an $(\mathbb{F}_p^n, b')$-source. Then for every $z \in [u]$

$$\Delta(Y_z, (\mathbb{F}_p^n, b')) \le e_z + \gamma'.$$

Proof. For every $z \in [u]$ let $\mu_z(y) = \mathbf{Pr}[Y_z = y]$ be the probability distribution of Y_z. From (3.1) 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 $(\mathbb{F}_p^n, b')$ 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 $(\mathbb{F}_p^n, b')$ source, we have that $\Delta(Y_z, (\mathbb{F}_p^n, b')) \leq e_z + \gamma'$. □

The next claim analyzes the behavior of the merger when the seed is sampled as in Construction 2.5. That is, when it is distributed over a subset of $\mathbb{F}_p^k$ of size 2^d.

Claim 3.10. Let $\phi : \{0,1\}^d \mapsto \mathbb{F}_p^k$ be the mapping from construction 2.5 and let Z be a random variable uniformly distributed over $\phi(\{0,1\}^d) \subset [u]$. Then, the expectation of e_Z is at most γ.

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

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

Since

$$2^d \geq 2^{\log_2(p) \cdot k - 1} = \frac{1}{2} \cdot p^k,$$

we have that for every $i \in [t]$ the probability that i is in H_Z is at most twice the probability that i is in $H_{Z'}$ for Z' uniformly distributed over $\mathbb{F}_p^k$. This last probability is bounded by $\gamma/2$ and so we can conclude that for every $i \in [t]$, $\mathbb{E}[\chi_i] \leq \gamma$. We can thus write

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

By linearity of expectation we have

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

□

Combining Claim 3.9 and Claim 3.10, and recalling that $\gamma' = 2^{-\Omega(b)}$, we see that

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

where the expectations are taken over Z, which is chosen uniformly in $\phi(\{0,1\}^d) \subset \mathbb{F}_p^k$. Now, for values of b larger than some constant b_0, this expression is smaller than 2γ. This completes the proof of Theorem 2.6. □

4. Proving Claim 3.6 and Claim 3.7 Using Results From Additive Number Theory

In this section we prove Claim 3.6 and Claim 3.7. These two claims are the only place in which our analysis differs from that of Lu *et al.* (2003) and Dvir & Raz (2005). In the proofs we use two results from additive number theory. The first is a quantitative version of Roth's theorem (Roth 1953) given by Bourgain (1999b). The second is a Lemma of Katz & Tao (1999) that deals with sum-sets and difference-sets.

4.1. Proof of Claim 3.6. The proof of the claim relies on the following result from additive number theory due to Katz & Tao (1999).

Lemma 4.1. (Katz & Tao 1999).
Let A, B be subsets of any abelian group. 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

$$|D| \leq K^{11/6}.$$

Before we can apply Lemma 4.1 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 4.1, 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

$$K \triangleq 2^{(6/11-\alpha/2)\cdot b},$$

and

$$U_1 \triangleq \{x_1 \in \mathbb{F}_p^n \quad | \quad \exists x_2, \ldots, x_k \in \mathbb{F}_p^n \quad s.t \quad (x_1, \ldots, x_k) \in U\}.$$

The following claim states several facts that, when combined, will enable us to use Lemma 4.1 on the sets we have defined.

Claim 4.2. The following is true:

1. $|V_1|, |V_2|, |V_3| \leq K$.
2. $|S| \leq |V_2| \leq K$.
3. $|U_1| \leq |D|$.

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 $2^{-(6/11-\alpha/2)\cdot b} = K^{-1}$, 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 $x \in (\mathbb{F}_p^n)^k$

$$Y_{z_1}(x) + Y_{z_3}(x) = 2 \cdot Y_{z_2}(x). \tag{4.1}$$

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 (4.1) 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 U_1 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 $x = (x_1, \ldots, x_k) \in (\mathbb{F}_p^n)^k$

$$c \cdot (Y_{z_1}(x) - Y_{z_3}(x)) = x_1. \tag{4.2}$$

Let $x_1 \in U_1$. From the definition of U_1 it follows that there exist $x_2, \ldots, x_k \in \mathbb{F}_p^n$ such that $x = (x_1, \ldots, x_k) \in U$. Using (4.2) we see that $x_1 \in c \cdot D$, since $Y_{z_1}(x) - Y_{z_3}(x) \in D$ by definition. Again, the inequality now follows from $|D| = |c \cdot D|$.

□

From the first two parts of Claim 4.2 we see that we can apply Lemma 4.1 with $A = V_1$ and $B = V_3$ to get that $|D| \leq K^{11/6}$. Substituting K we see that:

$$|D| \leq 2^{b \cdot (6/11 - \alpha/2) \cdot 11/6} = 2^{b \cdot \left(1 - \frac{11}{12}\alpha\right)}, \tag{4.3}$$

Using the third part of Claim 4.2 and (4.3) we conclude that

$$|U_1| \leq |D| \leq 2^{b \cdot \left(1 - \frac{11}{12}\alpha\right)}. \tag{4.4}$$

We can therefore bound the probability of U by

$$\mathbf{Pr}_X[X \in U] \leq \mathbf{Pr}_{X_1}[X_1 \in U_1] \leq 2^{-b} \cdot |U_1| \leq 2^{-b} \cdot 2^{b \cdot \left(1 - \frac{11}{12}\alpha\right)} = 2^{-\left(\frac{11}{12}\alpha\right) \cdot b}$$

(the second inequality follows from the fact that the min-entropy of X_1 is at least b). This completes the proof of Claim 3.6. □

4.2. Proof of Claim 3.7. The claim follows from Roth's theorem (Roth 1953) on arithmetic progressions of length three. For our purposes we require the quantitative version of this theorem as proven by Bourgain (1999b).

Theorem 4.3. (Bourgain 1999b).
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 $\mathbb{F}_p^k$. 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 4.3 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.

5. Improving the bound to 4/7

In this section we prove Theorem 2.7, which gives a stronger bound of 4/7 on the min-entropy rate of the merger from Construction 2.5. In order to achieve this bound we need the size of the underlying field, p, to be much larger than before (as a function of the error parameter γ). However, for constant error (which is an interesting case by itself) this stronger bound also requires a field of constant size. The proof is very similar to the proof of Theorem 2.6 and so the proof given in this section will be less detailed than the proof given in the last two sections.

The key to the proof of the 4/7 bound is the following lemma of Katz and Tao which is similar in spirit to Lemma 4.1.

Lemma 5.1. (Katz & Tao 1999).

Let A, B be subsets of any abelian group. Let $\Gamma \subset A \times B$, and define

$$S_1 \triangleq \{a+b \ \ | \ \ (a,b) \in \Gamma\},$$

$$S_2 \triangleq \{a+2b \ \ | \ \ (a,b) \in \Gamma\},$$

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

Suppose that there exists $K > 0$ such that $|A|, |B|, |S_1|, |S_2| \leq K$, then

$$|D| \leq K^{7/4}.$$

It turns out that by making some minor changes to the proof of Theorem 2.6 we can use this lemma in our proof to get the bound of 4/7. The main change needed is to consider arithmetic projections of length seven instead of length three. Luckily we have Szemeredi's theorem for arithmetic projections of any length. For our purposes we require a quantitative version of this theorem due to Gowers (2001).

Theorem 5.2. (Gowers 2001).

Let $0 < \delta \leq 1/2$, let k be a positive integer, let $N \geq 2 \uparrow 2 \uparrow \delta^{-1} \uparrow 2 \uparrow 2 \uparrow (k+9)$ and let $A \subset \{1, \ldots, N\}$ be a set of size at least δN. Then A contains an arithmetic progression of length k.

We use the same notations as in Section 3. We "re-define" α-good distributions. This time with 6/11 replaced with 4/7.

Definition 5.3. α-good distribution.

We say that a distribution $\mu : (\mathbb{F}_p^n)^u \to [0,1]$ is α-good if for at least $(1-\gamma/2) \cdot u$ values of $z \in [u]$, μ_z has min-entropy at least $(4/7 - \alpha) \cdot b$.

As before, it is enough to prove the following lemma.

Lemma 5.4. Let $X = (X_1, \ldots, X_k)$ be an $(\mathbb{F}_p^n, b)^{1:k}$-source, and let Y and μ be as in Section 3. Then for any constant $\alpha > 0$ there exists an integer $t \geq 1$, and a partition of $(\mathbb{F}_p^n)^k$ 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
$$H^\infty(Y_z | X \in W_i) \geq (4/7 - \alpha) \cdot b.$$

5.1. Proof of Lemma 5.4.

Every Y_z partitions $(\mathbb{F}_p^n)^k$ in the following way:

$$(\mathbb{F}_p^n)^k = \bigcup_{y \in \{0,1\}^n} (Y_z)^{-1}(y).$$

For each $z \in [u]$ we define the set

$$B_z \;\triangleq\; \bigcup_{\left\{y \;\middle|\; \mathbf{Pr}[Y_z = y] > 2^{-(4/7-\alpha/2)\cdot b}\right\}} (Y_z)^{-1}(y)$$

As mentioned in the beginning of this section, we need to consider arithmetic progressions of length seven instead of three. This motivates the following definition.

Definition 5.5. good 7-tuple.

Let $(z_1, ..., z_7) \in [u]^7$ be a 7-tuple of seed values. Write each z_i $(i = 1, ..., 7)$ as a vector $(z_{i1}, \ldots, z_{ik})$. We say that the 7-tuple $(z_1, ..., z_7)$ is **good** if the vectors $z_1, ..., z_7$ are identical in all coordinates different from one, and their first coordinates form an arithmetic progression of length seven in $\mathbb{F}_p$.

The next claim replaces Claim 3.6.

Claim 5.6. For every good 7-tuple $(z_1, ..., z_7)$ it holds that

$$\mathbf{Pr}_X[X \in B_{z_1} \cap ... \cap B_{z_7}] \leq 2^{-\left(\frac{7}{8}\alpha\right)\cdot b}.$$

We defer the proof of this claim to the end of this section and continue with the proof of Lemma 5.4. The next claim replaces Claim 3.7.

Claim 5.7. Let $T \subset [u]$ be such that $|T| > (\gamma/2) \cdot u$. Then T contains a good 7-tuple.

Proof. Same as the proof of Claim 3.7, but using Theorem 5.2 (for $k = 7$) instead of Theorem 4.3. □

The rest of the proof of Lemma 5.4 is exactly the same as in the proof of Lemma 3.4, and follows from combining Claim 5.6 with Claim 5.7. □

5.2. Proof of Claim 5.6. As in the proof of Claim 3.6 we let $K = 2^{(4/7-\alpha/2)\cdot b}$ and define the sets

$$U \triangleq B_{z_1} \cap ... \cap B_{z_7},$$

$$U_1 \triangleq \{x_1 \in \mathbb{F}_p^n \quad | \quad \exists x_2, \dots, x_k \in \mathbb{F}_p^n \quad s.t \quad (x_1, \dots, x_k) \in U\},$$

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

$$\Gamma \triangleq \{(v_1, v_7) \quad | \quad \exists x \in U \quad s.t \quad Y_{z_1}(x) = v_1 \quad and \quad Y_{z_7}(x) = v_7\},$$

$$S_1 \triangleq \{v_1 + v_7 \quad | \quad (v_1, v_7) \in \Gamma\},$$

$$S_2 \triangleq \{v_1 + 2v_7 \quad | \quad (v_1, v_7) \in \Gamma\},$$

$$D \triangleq \{v_1 - v_7 \quad | \quad (v_1, v_7) \in \Gamma\}.$$

The following claim replaces Claim 4.2, and will enable us to use Lemma 5.1 on the sets we have defined.

Claim 5.8. the following is true:

1. For $i = 1, ..., 7$, $|V_i| \leq K$.
2. $|S_1|, |S_2| \leq K$.
3. $|U_1| \leq |D|$.

Proof. The proofs of (1) and (3) are exactly the same as in Claim 4.2. To prove (2) notice that S_1 is contained in $2V_4$ and that S_2 is contained in $3V_5$. □

We apply Lemma 5.1 with $A = V_1$ and $B = V_7$ to get that $|D| \leq K^{7/4}$. Substituting K and using part 3 of the Claim 5.8 we get

$$|U_1| \leq |D| \leq 2^{b\cdot\left(1-\frac{7}{8}\alpha\right)}. \tag{5.1}$$

Therefore,

$$\mathbf{Pr}_X[X \in U] \leq \mathbf{Pr}_{X_1}[X_1 \in U_1] \leq 2^{-b} \cdot |U_1| \leq 2^{-b} \cdot 2^{b\cdot\left(1-\frac{7}{8}\alpha\right)} = 2^{-\left(\frac{7}{8}\alpha\right)\cdot b}.$$

□

6. Encoding Binary Sources as Vectors in $\mathbb{F}_p^n$

The merger from Construction 2.5 works when its inputs are vectors in $\mathbb{F}_p^n$. It is usually desirable to construct mergers that take binary strings as inputs. In this section we prove analogs of Theorem 2.6 and Theorem 2.7 for mergers over binary inputs. We note that the issues dealt with in this section are common to many papers on extractors and are not new to this paper.

Let $n > 1$ be an integer, p a prime number, and set

$$\tilde{n} \triangleq \left\lceil \frac{n}{\log_2 p} \right\rceil$$

We first define a mapping from binary strings to vectors over $\mathbb{F}_p$

$$\varphi \ : \ \{0,1\}^n \mapsto \mathbb{F}_p^{\tilde{n}}$$

in the following way: for $x \in \{0,1\}^n$ we treat x as an integer in $[2^n - 1]$. Since $x < 2^n \leq p^{\tilde{n}}$ there exist $a_1, \ldots, a_{\tilde{n}} \in \mathbb{F}_p$ such that

$$x = a_1 + a_2 p + a_3 p^2 + \ldots + a_{\tilde{n}} p^{\tilde{n}-1}.$$

The mapping φ simply outputs the vector

$$\varphi(x) \triangleq (a_1, \ldots, a_{\tilde{n}}).$$

Since base p expansion is unique we get that φ is an injection. This proves the following claim:

Claim 6.1. Let $\varphi \ : \ \{0,1\}^n \mapsto \mathbb{F}_p^{\tilde{n}}$ be the mapping defined above and let X be a $(\{0,1\}^n, b)$ random source. Then $\varphi(X)$ is an $(\mathbb{F}_p^{\tilde{n}}, b)$ random source.

Next, we define a mapping which takes vectors over $\mathbb{F}_p$ and outputs binary vectors. Let n, p and $\tilde{n}$ be as before. We define a mapping

$$\psi \ : \ \mathbb{F}_p^{\tilde{n}} \mapsto \{0,1\}^n$$

as follows:

$$\psi(a_1, \ldots, a_{\tilde{n}}) \triangleq \left(\sum_{i=1}^{\tilde{n}} a_i \cdot p^{i-1} \right) \bmod 2^n,$$

(since the output is a number smaller than 2^n we can write it in binary using n bits). Of course, ψ is not one-to-one, but the loss of entropy when applying ψ on a random source can be shown to be bounded by $\log_2 p$. Since in our case p is a constant, this loss will not be noticeable.

Claim 6.2. Let $\psi : \mathbb{F}_p^{\tilde{n}} \mapsto \{0,1\}^n$ be the mapping defined above and let X be an $(\mathbb{F}_p^{\tilde{n}}, b)$ random source, with $b > \log_2 p$. Then $\psi(X)$ is a $(\{0,1\}^n, b - \log_2 p)$ random source.

Proof. We have

$$\tilde{n} \leq \frac{n}{\log_2 p} + 1$$

Or

$$2^n \geq p^{\tilde{n}} \cdot \frac{1}{p}.$$

Therefore, for every y in the range of ψ there are at most p elements that ψ maps to it. This implies that ψ can reduce the min entropy of its input by at most $\log_2 p$. □

The following Corollary is immediate:

Corollary 6.3. Let $\psi : \mathbb{F}_p^{\tilde{n}} \mapsto \{0,1\}^n$ be the mapping defined above. Let $1 > \gamma > 0$ and let X be γ-close to an $(\mathbb{F}_p^{\tilde{n}}, b)$ random source, with $b > \log_2 p$. Then $\psi(X)$ is γ-close to a $(\{0,1\}^n, b - \log_2 p)$ random source.

We conclude by composing φ and ψ with the merger from Construction 2.5 to get a merger over $\{0,1\}^n$.

Construction 6.4. Let n, p and $\tilde{n}$ be as before. Let k be a constant integer and let $d = \lfloor k \cdot \log_2 p \rfloor$. Let

$$M : \{0,1\}^d \times \left(\mathbb{F}_p^{\tilde{n}}\right)^k \to \mathbb{F}_p^{\tilde{n}}$$

Be as in Construction 2.5. We define

$$\tilde{M} : \{0,1\}^d \times (\{0,1\}^n)^k \to \{0,1\}^n,$$

as follows:

$$\tilde{M}(z, x_1, \ldots, x_k) \triangleq \psi(M(z, \varphi(x_1), \ldots, \varphi(x_k))).$$

From the two claims above we can easily prove the following analog of Theorem 2.6.

Theorem 6.5. Let $0 < \gamma < 1$ be any constant, $k > 0$ a constant integer, and let p be a prime larger than $\exp(\gamma^{-2})$. Let

$$\tilde{M} : \{0,1\}^d \times (\{0,1\}^n)^k \to \{0,1\}^n,$$

be as in Construction 6.4, where $d = \lfloor \log_2 p \cdot k \rfloor$. 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 $[d, (\{0,1\}^n, b)^{1:k} \mapsto (\{0,1\}^n, b') \sim \gamma]$-strong merger with

$$b' = (6/11 - \alpha) \cdot b.$$

Proof. Let $X = (X_1, \ldots, X_k)$ be a somewhere $(\{0,1\}^n, b)$ source. Then, from Claim 6.1, we have that $\varphi(X) \triangleq (\varphi(X_1), \ldots, \varphi(X_k))$ is a somewhere $(\mathbb{F}_p^{\tilde{n}}, b)$ source. We apply Theorem 2.6 with the same γ but with α replaced by $\frac{\alpha}{2}$ to get that the average (over z) distance between $M(z, \varphi(X))$ and an $(\mathbb{F}_p^n, (\frac{6}{11} - \frac{\alpha}{2})b)$-source is at most γ. Now, using Corollary 6.3 we have that the average distance between $\psi(M(z, \varphi(X)))$ and a $(\{0,1\}^n, b')$ source is at most γ where

$$b' = (\frac{6}{11} - \frac{\alpha}{2})b - \log_2 p \geq (\frac{6}{11} - \alpha)b,$$

if $b \geq \frac{2 \cdot \log_2 p}{\alpha}$. Taking b_0 to be larger than $\frac{2 \cdot \log_2 p}{\alpha}$ (this is still a constant) we are done. □

An analog of Theorem 2.7 can be proved in the same way:

Theorem 6.6. Let $0 < \gamma < 1$ be any constant, $k > 0$ a constant integer, and let p be a prime larger than $F(\gamma) \triangleq 2 \uparrow 2 \uparrow (\gamma/2)^{-1} \uparrow 2 \uparrow 2 \uparrow 16$. Let

$$\tilde{M} \, : \, \{0,1\}^d \times (\{0,1\}^n)^k \to \{0,1\}^n,$$

be as in Construction 6.4, where $d = \lfloor k \cdot \log_2 p \rfloor$. 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 $[d, (\{0,1\}^n, b)^{1:k} \mapsto (\{0,1\}^n, b') \sim \gamma]$-strong merger with

$$b' = (4/7 - \alpha) \cdot b.$$

Acknowledgements

The authors would like to thank Ran Raz and Omer Reingold for helpful conversations. We thank Avi Wigderson for bringing to our attention the connection between linear mergers and the Kakeya conjecture. We thank Oded Goldreich for valuable comments that improved the presentation of the results. A.S. would also like to thank Oded Goldreich for helpful discussions on related problems. Z.D was supported by Israel Science Foundation (ISF) grant.

References

Boaz Barak, Russell Impagliazzo & Avi Wigderson (2004). Extracting Randomness Using Few Independent Sources. In *45th Symposium on Foundations of Computer Science (FOCS 2004)*, 384–393.

Jean Bourgain (1991). Besicovitch-type maximal operators and applications to Fourier analysis. *Geom. Funct. Anal.* **22**, 147–187.

Jean Bourgain (1999a). On the dimension of Kakeya sets and related maximal inequalities. *Geom. Funct. Anal.* (9), 256–282.

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

Zeev Dvir & Ran Raz (2005). Analyzing Linear Mergers. *Electronic Colloquium on Computational Complexity (ECCC)* (025).

Timothy Gowers (2001). A new proof of Szemeredi's theorem. *Geom. Funct. Anal.* (11), 465–588.

Nets Katz & Terence Tao (1999). Bounds on arithmetic projections, and applications to the Kakeya conjecture. *Math. Res. Letters* **6**, 625–630.

Nets Katz & Terence Tao (2002). New bounds on Kakeya problems. *Journal d'Analyse de Jerusalem* **87**, 231–263.

Chi-Jen Lu, Omer Reingold, Salil Vadhan & Avi Wigderson (2003). Extractors: optimal up to constant factors. In *35th Symposium on Theory of Computing (STOC 2003)*, 602–611. ACM Press. ISBN 1-58113-674-9.

Gerd Mockenhaupt & Terence Tao (2004). Restriction and Kakeya phenomena for finite fields. *Duke Math. J.* **121**, 35–74.

Ran Raz (2005). Extractors with Weak Random Seeds. In *37th Symposium on Theory of Computing (STOC 2005)*, 11–20.

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

Ronen Shaltiel (2002). Recent Developments in Extractors. *Bulletin of the European Association for Theoretical Computer Science* **77**, 67–95.

Amnon Ta-Shma (1996). On extracting randomness from weak random sources (extended abstract). In *28th Symposium on Theory of Computing (STOC 1996)*, 276–285. ACM Press. ISBN 0-89791-785-5.

Thomas Wolff (1995). An improved bound for Kakeya type maximal functions. *Revista Matemática Iberoamericana* **11**, 651–674.

Manuscript received February 27, 2006

Zeev Dvir
Department of Computer Science,
Weizmann institute of science,
Rehovot, Israel.
zeev.dvir@weizmann.ac.il.

Amir Shpilka
Faculty of Computer Science,
Technion,
Haifa, Israel.
shpilka@cs.technion.ac.il.