Better condensers and new extractors from Parvaresh-Vardy codes

Amnon Ta-Shma*
Tel-Aviv University
amnon@tau.ac.il

Christopher Umans†
California Institute of Technology
umans@cs.caltech.edu.

December 7, 2011

Abstract

We give a new construction of condensers based on Parvaresh-Vardy codes [PV05]. Our condensers have entropy rate $(1-\alpha)$ for subconstant α (in contrast to [GUV09] which required constant α) and suffer only sublinear entropy loss.

Known extractors can be applied to the output to extract all but a subconstant fraction of the minentropy. The resulting (k,ε) extractor $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ has output length $m = (1-\alpha)k$ with $\alpha = 1/\text{poly}\log(n)$, and seed length $d = O(\log n)$, when $\varepsilon \geq 1/2^{\log^\beta n}$ for any constant $\beta < 1$. Thus we achieve the same "world-record" extractor parameters as [DKSS09], with a more direct construction.

*This research was supported by Grant No. 2010120 from the United States-Israel Binational Science Foundation (BSF), and by Grant No. 1090/10 from the Israel Science foundation (ISF).

†Research supported by NSF CCF-0846991, CCF-1116111 and BSF grant 2010120.

1 Introduction

A (k, ε) *extractor* is a function $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with the property that for every distribution X over $\{0,1\}^n$ with *minentropy* at least k, the distribution $E(X, U_d)$ is ε-close to uniform. In applications one typically needs *explicit* constructions, i.e., ones in which E can be computed in time $\mathrm{poly}(n/\varepsilon)$. Non-explicit (probabilistic) constructions suggest that this should be possible with output length $m = k + d - 2\log(1/\varepsilon) - O(1)$ and seed length $d = \log(n-k) + 2\log(1/\varepsilon) + O(1)$.

Following an extensive line of research, Lu, Reingold, Vadhan, and Wigderson [LRVW03] were the first to achieve extractors "optimal up to constant factors." These extractors have $m = (1-\alpha)k$ for any constant $\alpha > 0$, and $d = O(\log n)$, for constant $\varepsilon > 0$. As is common in the area, the [LRVW03] construction combines several components in a delicate composition.

In 2007, Guruswami, Umans, and Vadhan [GUV09] discovered a simpler construction (that also handles arbitrary error ε) based on Parvaresh-Vardy codes [PV05]. The main object they construct is a *condenser*, which is a weakening of an extractor in which the output distribution is only required to be ε-close to a distribution with large min-entropy k'. The [GUV09] condenser is *lossless* (in the sense that the output minentropy k' is exactly $k + d$), and has output length $m = (1+\alpha)k'$ for any constant $\alpha > 0$, and $d = O(\log n/\varepsilon)$. Thus these condensers have constant *minentropy rate* (defined to be k'/m). Known extractors that work for constant minentropy rate can be applied to the output to extract any constant fraction of the minentropy. The overall construction is more direct than that of [LRVW03], having essentially two stages: condense, and then extract.

In 2008, Dvir and Wigderson [DW11] obtained an alternate construction achieving essentially the same parameters as [GUV09] (but requiring $\varepsilon \geq 1/\mathrm{poly}(n)$). The main new object in [DW11] was a beautiful new construction of *mergers*, which can be inserted into the program outlined by Ta-Shma [TS96] to obtain extractors. The basic extractor construction has four stages: obtain a *somewhere block-source*, extract to get a *somewhere random source*, combine using the merger to get a source with constant minentropy rate, and then extract.

In 2009, Dvir, Kopparty, Saraf, and Sudan [DKSS09] showed how to extend the polynomial method at the heart of the proof in [DW11], by introducing multiplicities. Their so-called *extended method of multiplicities* has several applications, one of which is an improvement to the parameters of the merger from [DW11]. Following the same sequence of steps as in the basic extractor construction of [DW11], they obtain an extractor having output length $m = (1-\alpha)k$ with $\alpha = 1/\mathrm{poly}\log(n)$, and seed length $d = O(\log n)$ when $\varepsilon \geq 1/\mathrm{poly}\log(n)$. In fact, a more careful analysis[1] can obtain the same result when $\varepsilon \geq 1/2^{\log^\beta n}$ for any constant $\beta < 1$. This is the current best extractor construction, and indeed the first extractor construction with *sublinear entropy loss* (entropy loss for an extractor is defined to be $m + d - k$).

In this work we revisit the [GUV09] condenser construction, and show how to achieve condensers with entropy rate $(1-\alpha)$ for *subconstant* α ([GUV09] required α to be a constant), while suffering only *sublinear entropy loss*, thus improving the current best *condenser* construction.

Known extractors can be applied to the output to extract all but a subconstant fraction of the minentropy. As with [GUV09] our overall extractor construction has only two self-contained stages: condense, and then extract. By way of comparison with [DW11, DKSS09], our condenser takes us immediately to the second-to-last stage of their four stages, making for a more direct extractor construction. After using the repeated extraction trick of [WZ99], we obtain an extractor achieving the same "world-record" parameters as [DKSS09], namely, having output length $m = (1-\alpha)k$ with $\alpha = 1/\mathrm{poly}\log(n)$, and seed length $d = O(\log n)$ when $\varepsilon \geq 1/2^{\log^\beta n}$ for any constant $\beta < 1$.

[1] One only needs to set $\delta = \log^{1-\beta} n$ and $\varepsilon = 2^{-\log^\beta n}$ in Step 3 in the proof of their Theorem 20.

To obtain our result we introduce three new ideas into the basic framework of [GUV09], which is based on the codes of Parvaresh and Vardy [PV05]. PV codes are bundles of correlated Reed-Solomon codewords, with the correlation described by a simple algebraic operation. The first new idea is to use a different correlation based on what we call the *covering curve*, whose algebraic properties we need elsewhere in the argument. Second, we bound the *total degree* of the interpolating polynomial used in the proof rather than the individual degrees, and compensate by using *multiplicities* in the style of [GS99]. Finally, and most critically, we employ "two levels" of the main component of the proof in [PV05], in which they deduce that the interpolating polynomial vanishes at certain points in an extension field from knowledge of its zeros over the base field. The three new ideas work in concert, and it appears that no proper subset of them leads to any substantive improvement.

2 The GUV framework and our new ideas

Here we present the basic construction[2] of [GUV09] in a somewhat more general setting, in order to be able to describe our new ideas. The formal presentation follows in Section 3 (although the informal discussion is necessarily somewhat technical since our improvement concerns a low-level relationship between parameters of the construction).

2.1 The lossy GUV condenser

We begin with the formal definition of a (lossy) condenser:

Definition 2.1. *A function $C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a $k \to_\varepsilon k'$* condenser *if for all distributions X with minentropy at least k, the distribution $C(X, U_d)$ is ε-close to a distribution with minentropy at least k'. The condenser is* explicit *if C can be computed in time poly$(n, 1/\varepsilon)$. The* entropy loss *of the condenser is $k + d - k'$ and the* entropy rate *is $\frac{k'}{m}$.*

To prove that a function is a lossy condenser, we use the "list-decoding" approach described in [GUV09], captured by the following lemma:

Lemma 2.1 (Lemma 5.4 in [GUV09])**.** *Let $C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be a function. If for all $T \subseteq \{0,1\}^m$ of size at most L, we have that the set*

$$\mathit{LIST}(T, \varepsilon) = \{x : \Pr_y[C(x, y) \in T] \geq \varepsilon\}$$

has cardinality at most H, then C is a $\log(H/\varepsilon) \to_{2\varepsilon} \log(L/\varepsilon) - 1$ condenser.

The basic construction. Let n, m, ε be parameters. Choose a field $\mathbb{F}_h$ and a polynomial $E(X)$, irreducible over $\mathbb{F}_h$ of degree n. Let $\mathbb{F} = \mathbb{F}_h[X]/E(X)$. Let $C = (C_0, \ldots, C_{m-1})$ ("the curve") be a function from $\mathbb{F}$ to $\mathbb{F}^m$, to be fixed later, that can be computed by a polynomial-size arithmetic circuit. Consider the function:

$$G(f \in \mathbb{F}, x \in \mathbb{F}_h) = (C_0(f)(x), \ldots, C_{m-1}(f)(x)). \tag{1}$$

As per Lemma 2.1, to prove that this is a lossy condenser we need to show that for every set $T \subseteq \mathbb{F}_h^m$ up to size L, the set

$$\text{LIST}(T, \varepsilon) = \{f \in \mathbb{F} : \Pr_{x \in \mathbb{F}_h}[G(f, x) \in T] \geq \varepsilon\}$$

[2] We present the construction and analysis of *lossy* condensers, since that is what we will use later.

is smaller than H. To maximize the entropy rate we want L large (as close to h^m as possible), and to minimize the entropy loss, we want H small (as close to L/h as possible). We will see below that in [GUV09] L can be no larger than $(h^m)^{1-\alpha}$ for a constant $\alpha > 0$. Our new construction breaks this barrier (achieving subconstant α) while keeping H comparably small.

The main argument. The main argument in [GUV09] is quite compact. It allows us to conclude that every $f \in \text{LIST}(T, \varepsilon)$ is a root of a low degree univariate polynomial, as follows. Let $Q \in \mathbb{F}_h[W_0, \ldots, W_{m-1}]$ be a non-zero polynomial that vanishes on T, and for which $Q \circ C$ (a univariate polynomial over $\mathbb{F}$) is not the zero polynomial. By definition, for every $f \in \text{LIST}(T, \varepsilon)$, the polynomial

$$R_f(X) = Q(C_0(f)(X), \ldots, C_{m-1}(f)(X))$$

vanishes on an ε fraction of the $x \in \mathbb{F}_h$, which implies it is the zero polynomial, provided

$$\deg(R_f) = n \deg(Q) < \varepsilon h. \tag{2}$$

Then, $R_f = 0$ implies f is a root of $Q \circ C$, and therefore $|\text{LIST}(T, \varepsilon)| \leq \deg(Q \circ C)$.

Choice of the curve and the main constraint. In [GUV09], for a parameter $\ell < h$, they choose $C_i(Z) = Z^{\ell^i}$ (the "PV curve"), and they require that Q satisfies $\deg_i(Q) \leq \ell - 1$, where $\deg_i(Q)$ is the degree of W_i in Q. This combination ensures that $Q \circ C$ is not the zero polynomial, if $Q \neq 0$. This construction can then handle sets T of size $L = \ell^m - 1$ since Q with the specified degrees has more degrees of freedom than the number of homogenous constraints needed to force Q to vanish on T. Eq. (2) in the argument above then requires

$$\ell < \varepsilon h/(nm) \tag{3}$$

to ensure $n \deg(Q) \leq \varepsilon h$. The resulting upper bound on $|\text{LIST}(T, \varepsilon)|$ is $H = \deg(Q \circ C) \leq h^m - 1$.

Notice that Eq. (3) forces $L < \ell^m \leq (h^m)^{1-\alpha}$ for some constant $\alpha > 0$, since we need $h \leq \text{poly}(n)$ in order to maintain seed length $O(\log n)$. Thus the entropy rate can be no better than a constant.

2.2 The new ideas

The new ideas work together to replace the constraint $L \leq (\varepsilon h/(nm))^m$ with the constraint $L \leq O(\frac{\varepsilon h}{2})^m$, which results in entropy loss about $\frac{k}{\log^\beta n}$ and entropy rate $(1 - O(1/\log^\beta n))$ when $\varepsilon > 1/2^{\log^{1-\beta} n}$ (and when choosing $h \geq nm^2$).

The "two-layered" construction. In Eq. (2) and (3) the n comes from the fact that in forming $R_f(X)$, the substituted polynomials $C_i(f)$ have degree n. We show how to augment the basic construction and the main argument so that the bottleneck inequality arises when only substituting *linear* polynomials, effectively replacing this n with 1.

Construct the degree 2 extension field $\mathbb{F}_q = \mathbb{F}_h[Y]/P(Y)$ where P is irreducible over $\mathbb{F}_h$ and of degree 2, and the degree n extension field $\mathbb{F} = \mathbb{F}_q[X]/E(X)$ where E is irreducible over $\mathbb{F}_q$ and of degree n. As before we have a function $C = (C_0, \ldots, C_{m-1})$ from $\mathbb{F}$ to $\mathbb{F}^m$ (to be fixed later) that can be computed by a polynomial-size arithmetic circuit. The new function is:

$$G(f \in \mathbb{F}; x \in \mathbb{F}_q, y \in \mathbb{F}_h) = (C_0(f)(x)(y), \ldots, C_{m-1}(f)(x)(y)). \tag{4}$$

As before, to bound the size of LIST(T, ε), we argue that every $f \in$ LIST(T, ε) is a root of $Q \circ C$. Now there are two steps to the argument. First, for every $f \in$ LIST(T, ε), for at least an $\varepsilon/2$ fraction of $x \in \mathbb{F}_q$, the polynomial

$$S_{f,x}(Y) = Q(C_0(f)(x)(Y), \ldots, C_{m-1}(f)(x)(Y))$$

vanishes on an $\varepsilon/2$ fraction of the $y \in \mathbb{F}_h$ (by an averaging argument). This implies $S_{f,x}$ is the zero polynomial, provided $\deg(Q) < \varepsilon h/2$. For each such x, the polynomial

$$R_f(X) = Q(C_0(f)(X), \ldots, C_{m-1}(f)(X))$$

vanishes on x and so R_f is the zero polynomial, provided $n \deg(Q) < \varepsilon q/2$. As before we conclude that f is a root of $Q \circ C$ as desired. As long as $h > n$, the two new constraints ($\deg(Q) < \varepsilon h/2$ and $n \deg(Q) < \varepsilon q/2$) are both implied by $\deg(Q) < \varepsilon h/2$, and we can see that the n in Eq. (2) has disappeared.

Total degree and multiplicities. In Eq. (3) the m comes from the fact that the total degree of Q is ℓm. A more efficient version of the argument would replace the bound on the individual degrees of the variables of Q with a bound on the total degree. However, the constraint $\deg(Q) < \varepsilon h/2$ allows only

$$\binom{\varepsilon h/2 + m}{m} \approx (\varepsilon h/2)^m/m!$$

degrees of freedom, and the $m!$ in the denominator prevents a direct gain without further modification.

As pioneered by [GS99], we use *multiplicities* to address this problem. We require that Q vanishes on the set T *with multiplicity* t, and then our main constraint becomes $\deg(Q) < \varepsilon h t/2$, since when arguing that $S_{f,x}$ vanishes we count the zeros in their multiplicity. Because forcing Q to vanish at a point in $\mathbb{F}_h^m$ with multiplicity t entails $\binom{t+m}{m}$ constraints, we can now handle sets T up to size

$$\binom{\varepsilon h t/2 + m}{m} / \binom{t+m}{m} \approx (\varepsilon h/2)^m,$$

(provided $t \geq m^2$) as promised. However, this modification comes at a price: we can no longer easily argue that $Q \circ C$ is not the zero polynomial. In [GUV09] the PV curve was selected with ℓ greater than the individual degrees of Q, which ensured that distinct monomials in Q map to distinct monomials of $Q \circ C$ (and thus $Q \circ C \neq 0$ if $Q \neq 0$). Now, if we choose $\ell > \deg(Q) = \varepsilon h t/2$, we find that $\deg(Q \circ C) \geq (\varepsilon h t/2)^m$, while we are shooting for $\deg(Q \circ C) \leq (\varepsilon h/2)^m$. In fact with $t \geq m^2$, the weaker bound spoils the sublinear entropy loss. Nevertheless, with a different curve C, we are able to show (by a less obvious argument) that Q can be chosen so that it simultaneously vanishes on T with multiplicity t, and $Q \circ C \neq 0$. These ideas are described next.

The covering curve. In our construction we replace the PV curve used in [GUV09] with the "covering curve," which has the following form (it is a *linearized polynomial*):

$$C_i(Z) = \sum_{j=0}^{m-1} \alpha_j^{(i)} Z^{h^j}.$$

We show how to choose the $\alpha_j^{(i)}$ in $\mathbb{F}$ so that $\mathbb{F}_h^m \subseteq \mathrm{Im}(C)$, which is the key property.

Since the covering curve passes through all the points in $\mathbb{F}_h^m$, one way to find a Q that vanishes on a specified set T with multiplicity t, while $Q \circ C \neq 0$, is to ensure that Q is non-zero at *some* point in $\mathbb{F}_h^m$ outside of T. Indeed this is what we are able to do. Also, since the covering curve has degree h^{m-1} just like the PV curve, the overall construction and proof yields a bound on H comparable to the one obtained in [GUV09] (while significantly improving the parameter L).

3 The improved condenser

In this section we give the formal presentation of our new construction and proof.

The covering curve. We begin with a short proof implying that the desired "covering curve" exists and can be found efficiently.

Lemma 3.1. *Let $\beta_0, \ldots, \beta_{\ell-1}$ be a basis for $\mathbb{F}_{h^\ell}$ over $\mathbb{F}_h$. Then for every $1 \leq m \leq \ell$, there exist elements $\alpha_j^{(i)} \in \mathbb{F}_{h^\ell}$ (for $i, j = 0, \ldots, m-1$) such that $C = (C_0, \ldots, C_{m-1})$ given by*

$$C_i(Z) = \sum_{j=0}^{m-1} \alpha_j^{(i)} Z^{h^j}$$

satisfies $C_i(\sum_{j=0}^{m-1} c_j \beta_j) = c_i$ for all i and $c_j \in \mathbb{F}_h$. In particular, $\mathbb{F}_h^m$ is contained in the image Im(C).

Proof. First note the C_i is a linearized polynomial, i.e., for any $v_1, v_2 \in \mathbb{F}_{h^m}$ and any $\alpha, \beta \in \mathbb{F}_h$, $C_i(\alpha v_1 + \beta v_2) = \alpha C_i(v_1) + \beta C_i(v_2)$. Next notice that the requirements $C_i(\beta_i) = 1$ and $C_i(\beta_j) = 0$ for $0 \leq j < m$, $j \neq i$ place m linear constraints on the m coefficients of C_i. The $m \times m$ matrix B corresponding to this linear system has $B[i,j] = \beta_i^{h^j}$ and the constraints can be expressed as $B \begin{pmatrix} \alpha_0^{(i)} \\ \vdots \\ \alpha_{m-1}^{(i)} \end{pmatrix} = e_i$, where e_i is the column vector with a 1 in the i'th coordinate and zero elsewhere. It is well-known (see, e.g., Theorem 3.51 in [LN94]) that B is invertible if and only if the β_i are linearly independent, which they are in our case. Thus the desired coefficients for C_i may be found efficiently by solving this linear system.

Then, note that for each i (using the fact that C_i is a linearized polynomial):

$$C_i\left(\sum_{j=0}^{m-1} c_j \beta_j\right) = \sum_{j=0}^{m-1} C_i(c_j \beta_j) = \sum_{j=0}^{m-1} c_j C_i(\beta_j) = c_i$$

as claimed. In particular,

$$\left\{ C\left(\sum_{j=0}^{m-1} c_j \beta_j\right) : c_0, \ldots, c_{m-1} \in \mathbb{F}_h \right\} = \mathbb{F}_h^m$$

and so $\mathbb{F}_h^m$ is contained in the image of C. □

The new condenser. Now, we can describe the new condenser. We are given parameters n, m, ε. Let h be a prime power (a parameter to be chosen below). Select a degree two polynomial P, irreducible over $\mathbb{F}_h$, and construct the extension field $\mathbb{F}_q = \mathbb{F}_h[Y]/P(Y)$. Select a degree n polynomial E, irreducible over $\mathbb{F}_q$, and construct the extension field $\mathbb{F} = \mathbb{F}_q[X]/E(X)$. Let $C = (C_0, \ldots, C_{m-1})$ be the "covering curve" guaranteed by Lemma 3.1 that contains $\mathbb{F}_h^m$ in its image. The new condenser is given by:

$$G(f \in \mathbb{F}; x \in \mathbb{F}_q, y \in \mathbb{F}_h) = (C_0(f)(x)(y), \ldots, C_{m-1}(f)(x)(y)). \tag{5}$$

where we understand $f_i = C_i(f)$ to be the canonical representative in $\mathbb{F}$ (a polynomial of degree at most $n-1$), and we understand $f_i(x)$ to be the canonical representative in $\mathbb{F}_q$ (a polynomial of degree at most 1).

Our main theorem is

Theorem 3.2 (main)**.** *Let G be the function given by Eq. (5). Then G is a*

$$[m \log h + 2 \log m - 1] \to_{4\varepsilon} [m \log(\varepsilon h/2) + \log(1/\varepsilon) - \log(2e)]$$

condenser, provided $h > nm^2$.

Note that (provided $\varepsilon < 1/2e$), this $k \to_{4\varepsilon} k'$ condenser has entropy rate:

$$\frac{k'}{\log(h^m)} = \frac{m \log(\varepsilon h/2) + \log(1/\varepsilon) - \log(2e)}{\log(h^m)} \geq \frac{\log(\varepsilon h/2)}{\log h} = 1 - \frac{\log(2/\varepsilon)}{\log h} \geq 1 - \frac{\log(2/\varepsilon)}{\log n},$$

and sublinear entropy loss because

$$k - k' \quad \leq \quad m \log h + 2 \log m - m \log(\varepsilon h/2) = 2 \log m - m \log(\varepsilon/2) \leq \frac{k}{\log n} \log(2/\varepsilon) + 2 \log m,$$

because $k \geq m \log h \geq m \log n$.

3.1 Proof of main theorem (Theorem 3.2)

We need the following definition of *Hasse derivatives* and multiplicity:

Definition 3.1. *For $Q \in \mathbb{F}_h[W_0, \ldots, W_{m-1}]$, and a vector $\vec{a} = (a_0, \ldots, a_{m-1})$ of non-negative integers, the* $\vec{a}$-th Hasse derivative of Q*, denoted $Q^{(\vec{a})}$, is the coefficient of the monomial $Z_0^{a_0} Z_1^{a_1} \cdots Z_{m-1}^{a_{m-1}}$ in the polynomial*

$$Q(W_0 + Z_0, \ldots, W_{m-1} + Z_{m-1}).$$

We say that $\sum_i a_i$ is the weight *of vector $\vec{a}$, denoted wt($\vec{a}$). The* multiplicity *of $\alpha \in \mathbb{F}_h^m$ in Q, denoted mult(Q, α) is the maximum i such that $Q^{(\vec{a})}(\alpha) = 0$ for all $\vec{a}$ of weight less than i.*

It is clear that $\deg(Q^{(\vec{a})}) \leq \deg(Q)$. We will use two other basic property of Hasse derivatives: mult$(Q^{(\vec{a})}, \alpha) \geq$ mult$(Q, \alpha) -$ wt$(\vec{a})$ (Lemma 5 in [DKSS09]), and mult$(Q \circ C, \alpha) \geq$ mult$(Q, C(\alpha))$ (Proposition 6 in [DKSS09]).

We now proceed with the proof.

Choosing Q. Set $t = m^2$, and consider a set $T \subseteq \mathbb{F}_h^m$ of size at most $L = (\varepsilon h/2)^m/e$. Choose a non-zero $Q_0 \in \mathbb{F}_h[W_0, \ldots, W_{m-1}]$ of total degree $D = \varepsilon th/2 - 1$, that vanishes with multiplicity t on T. Since

$$\frac{\binom{D+m}{m}}{\binom{t+m}{m}} > \frac{D^m}{m!} \cdot \frac{m!}{t^m(1+m/t)^m} \geq (\varepsilon h/2)^m/e \geq |T|$$

such an interpolating polynomial exists, by solving a homogeneous linear system in the coefficients of Q_0.

By the multiplicity version of Schwartz-Zippel [DKSS09], it *cannot* be the case that Q_0 vanishes with multiplicity at least $t/2$ on all of $\mathbb{F}_h^m$ (the sum of the multiplicities over $\mathbb{F}_h^m$ can be at most $\frac{D}{h}h^m = h^{m-1}D < (t/2) \cdot h^m$). Therefore, by the definition of multiplicity, there exists $\vec{a}$ of weight at most $t/2$ for which $Q_0^{(\vec{a})}$ *does not* vanish on all of $\mathbb{F}_h^m$. We set

$$Q = Q_0^{(\vec{a})}.$$

Since the weight of $\vec{a}$ is at most $t/2$, we also know that Q still vanishes on T with multiplicity at least $t - t/2 = t/2$, and of course $\deg(Q) \leq \deg(Q_0) \leq D$.

We claim that $Q \circ C$, as a univariate polynomial in $\mathbb{F}[Z]$, is not the zero polynomial. To see that, notice that $\mathbb{F}$ is an extension field of $\mathbb{F}_h$ (namely, isomorphic to $\mathbb{F}_{h^{2n}}$) and by Lemma 3.1, $\mathbb{F}_h^m$ is contained in $Im(C)$. As Q does not vanish on all of $\mathbb{F}_h^m$, $Q \circ C$ is not identically zero.

The "two-layered" analysis. Recall that $\text{LIST}(T, 2\varepsilon) = \{f : \Pr_{x,y}[G(f; x, y) \in T] \geq 2\varepsilon\}$. We have the following claim:

Claim 3.2.1. *Every $f \in$ LIST$(T, 2\varepsilon)$ is a root of $Q \circ C$; i.e., $Q(C(f)) = 0$.*

Proof. Fix an $f \in \text{LIST}(T, 2\varepsilon)$. By averaging, we have that $f \in \text{LIST}(T, 2\varepsilon)$ implies

$$\Pr_x[\Pr_y[G(f; x, y) \in T] \geq \varepsilon] \geq \varepsilon.$$

Fix an $x \in \mathbb{F}_q$ for which $\Pr_y[G(f; x, y) \in T] \geq \varepsilon$. Consider the univariate polynomial

$$S_{f,x}(Y) = Q(C_0(f)(x)(Y), \ldots, C_{m-1}(f)(x)(Y))$$

which has degree at most D. For at least εh distinct $y \in \mathbb{F}_h$, we have that $\text{mult}(S_{f,x}, y) \geq t/2$. Since $\varepsilon ht/2 > D$, we conclude that $S_{f,x} = 0$ in $\mathbb{F}_h[Y]$, and therefore also $S_{f,x} \bmod P(Y) = 0$ in $\mathbb{F}_q$.

Now, we view Q as an element of $\mathbb{F}_q[W_0, \ldots, W_{m-1}]$, and we see that $S_{f,x} = 0$ implies that the univariate polynomial

$$R_f(X) = Q(C_0(f)(X), \ldots, C_{m-1}(f)(X))$$

has a root at x. This holds for εq distinct $x \in \mathbb{F}_q$, and then because

$$\varepsilon q = \varepsilon hnm^2 = \varepsilon hnt > nD \geq \deg(R_f),$$

we find that $R_f = 0$ in $\mathbb{F}_q[X]$ and therefore also $R_f \bmod E(X) = 0$ in $\mathbb{F}$.

Now we view Q as an element of $\mathbb{F}[W_0, \ldots, W_{m-1}]$ and we see that $R_f = 0$ implies that the univariate polynomial

$$(Q \circ C)(Z) = Q(C_0(Z), \ldots, C_{m-1}(Z))$$

has a root at f, as claimed. □

We conclude that $|\text{LIST}(T, 2\varepsilon)| \leq \deg(Q \circ C) \leq h^{m-1}D \leq \varepsilon th^m/2 = H$. By Lemma 2.1, this means that G is a $\log(\frac{H}{\varepsilon}) \to_{4\varepsilon} \log(\frac{L}{\varepsilon})$ condenser, i.e., a

$$\log(th^m/2) \to_{4\varepsilon} \log\left((\varepsilon h/2)^m/(e\varepsilon)\right) - 1$$

condenser. This concludes the proof of Theorem 3.2.

4 From condensers to extractors

At this point we have a very dense source, and the extraction task we have is *identical* to the one in [DKSS09] after they use their new mergers in the framework of [TS96] (in contrast, we arrived at this point directly, with one application of our new condenser). To describe our extractors, we follow the nice presentation of [DKSS09] (their Section 6.2). We emphasize that from now on we are using standard tools, and in particular we are not using the *new* results of [DKSS09] (but we are mirroring their exposition in this section for easy comparison of the two works).

Condense and extract. Our main theorem (Theorem 3.2) constructs a good condenser with high output entropy rate, namely,

Theorem 4.1 (Theorem 3.2 rephrased)**.** *For all positive integers $k < n$ and $\varepsilon > 0$, there exists an explicit function*

$$C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$$

that is a $k \to_\varepsilon (1-\delta)m$ condenser, for $\delta = \frac{\log(8/\varepsilon)}{\log n}$. The condenser has seed length $d = O(\log n)$ and entropy loss $\delta k + O(\log m) + d$.

Thus, applying our condenser we get a highly condensed source. The following lemma from [DKSS09] (which is a simple consequence of a block-source extractor constructed in [RSW06]) describes an excellent extractor for highly condensed sources:

Lemma 4.2 ([DKSS09, Lemma 27])**.** *For any k and $\delta > 0$, there exists an explicit $((1-\delta)k, k^{-\Omega(1)})$ extractor $E : \{0,1\}^k \times \{0,1\}^d \to \{0,1\}^{(1-3\delta)k}$ with seed length $d = O(\log k)$.*

Combining the two we obtain the following extractor:

Theorem 4.3. *For all positive integers $k < n$ and $\varepsilon > k^{-\Omega(1)}$ there exists an explicit (k, ε) extractor $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$, with $d = O(\log n)$ and $m - k \leq O(\delta k)$, for $\delta = \frac{\log(8/\varepsilon)}{\log n}$ as above.*

Setting $\varepsilon = 8 \cdot 2^{-\log^\beta n}$, we find that δ is $\log^{\beta-1} n$. Next, we reduce the entropy loss factor to any inverse polynomial in $\log n$, as in [DKSS09], using repeated extraction [WZ99].

Reducing the entropy loss by repeated extraction. We can make the extractor of Theorem 4.3 *strong* using techniques from [RSW06]. We can also reduce δ to δ^c by applying c independent extraction steps [WZ99]. Both transformations are described in detail in [DKSS09, Section 6.2], and have the following cost:

- The transformation from an extractor to a strong extractor has the following cost:
  - It enlarges the error from ε to $\sqrt{\varepsilon}$,
  - It enlarges the seed length from d to $O(d)$, and,
  - It increases the entropy loss by an additive factor of $2\log(1/\varepsilon) + O(1)$.

  See [DKSS09, Theorem 30].
- The transformation reducing the entropy loss using c independent extraction steps (for some integer $c \geq 1$) has the following cost:

– It enlarges the error from ε to $O(c\varepsilon)$,
– It enlarges the seed length from d to cd, and,
– It decreases the entropy loss from δk to $\delta^c k + O(c \log(1/\varepsilon))$.

See [DKSS09, Theorem 31], taking $r = \log(1/\varepsilon)$.

Altogether, we get:

Theorem 4.4. *For all constants $1 > \beta > 0$ and $b \geq 1$, there exists a constant $c \geq 1$ for which the following holds: For all positive integers $k < n$, there exists an explicit $(k, \varepsilon = 2^{-\log^\beta n})$ strong extractor*

$$E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m,$$

with

- *seed length $O(\log n)$, and*
- *entropy loss $k + d - m = O(\log n + \delta k)$, where $\delta = \frac{1}{\log^b n}$,*

provided $k \geq (1/\varepsilon)^c$.

For smaller min-entropies k an even better (smaller entropy loss) explicit extractor exists:

Theorem 4.5. *For all constants $1 > \beta > 0$ and $c \geq 1$, the following holds: For all positive integers $k < n$, there exists an explicit $(k, \varepsilon = 2^{-\log^\beta n})$ strong extractor*

$$E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m,$$

with

- *seed length $O(\log n)$, and*
- *entropy loss $k + d - m = O(\log n + \delta k)$, where $\delta = 2^{-\log^{1-\beta} n}$,*

provided $k \leq (1/\varepsilon)^c$.

Proof. Our starting point is [GUV09, Theorem 4.17] which extracts half the entropy with seed length $O(\log n/\varepsilon)$. We apply repeated extraction $O(\log(1/\delta))$ times to get a $(1-\delta)m$ output length, using $O(\log n) + O(\log(k/\varepsilon)\log(1/\delta))$ seed length (the $O(\log n)$ additive term is because we first apply the GUV lossless condenser to reduce the input length to $O(k)$). For the specified parameters (i.e., for such a small min-entropy k) the seed length is $O(\log n)$ as required. □

5 Conclusions and open problems

We wonder whether the "two-layered" construction and the accompanying proof ideas have applications elsewhere. As one example, we mention that this technique can improve the [DKSS09] merger, which has seed length $d = \frac{1}{\delta}\log(\frac{2\Lambda}{\varepsilon})$, where Λ is the number of sources, ε the statistical error and δ is the entropy loss (expressed as a fraction). Using the "two-level" construction we can reduce the seed length to $O(\max\left\{\frac{1}{\delta}\log(\frac{1}{\varepsilon}), \log(\Lambda)\right\})$. We do not see, however, how to translate this improvement into a better extractor construction.

Another important question is whether the condenser in this paper can be improved further. For example achieving an $O(\log n)$ entropy rate *deficiency* together with $O(\log n)$ entropy loss would lead to *optimal* output length extractors (with $O(\log n)$ seed length) via standard techniques. In this direction, we note that the existence of non-trivial finite-field *Kakeya set* constructions presents a potential pitfall that must be avoided.

A connection to Kakeya sets. To improve the entropy rate of our condensers, one wishes to handle sets T with size approaching $h^m/\text{poly}(n)$. Currently we handle sets T with size no larger than $h^m/2^m$, which limits our output entropy rate to $1 - 1/\log h \leq 1 - 1/O(\log n)$. Ignoring ε for this discussion, our current methodology is to show that there exists a degree $ht - 1$ polynomial Q that vanishes on T with multiplicity t, and yet Q does *not* vanish on all of $\mathbb{F}_h^m$. This property allows us to argue that such a Q does not vanish on the covering curve. One may wonder whether or not larger sets T can be handled in a similar way; here we give evidence that they cannot.

We use the existence of large finite-field Kakeya sets, which are sets containing a line in every direction. From, e.g., [SS08], we have

Theorem 5.1. *For every finite field $\mathbb{F}$ with h elements and every $m \geq 2$, there exists a Kakeya set $K \subseteq \mathbb{F}^m$ of size at least $h^m/2^{m-1}$.*

Kakeya sets represent a potential barrier in the following sense:

Theorem 5.2. *Let $K \subseteq \mathbb{F}_h^m$ be a Kakeya set. Suppose $Q \in \mathbb{F}_h[W_0, \ldots, W_{m-1}]$ is homogeneous of degree $D \leq ht - 1$ and vanishes on K with multiplicity t. Then Q vanishes on $\mathbb{F}_h^m$.*

Proof. By the definition of a Kakeya set, for every $b \in \mathbb{F}_h^m$, there exists $a \in \mathbb{F}_h^m$ for which $\{a + bx : x \in \mathbb{F}_h\} \subseteq K$. For such a pair (a, b), we find that $Q(a + bX)$ is a univariate polynomial of degree $D < ht$ that vanishes with multiplicity t on $\mathbb{F}_h$; thus it is the zero polynomial. The coefficient on X^D in $Q(a + bX)$ is $Q(b)$, and thus $Q(b) = 0$. As b was arbitrary in $\mathbb{F}_h^m$, we conclude that Q vanishes on $\mathbb{F}_h^m$ as claimed. □

Thus, in the context of GUV-style condenser constructions and proofs, sets T of size larger than about $h^m/2^m$ (needed for entropy rates larger than $1 - 1/O(\log n)$) seem to be qualitatively different, requiring new ideas.

6 Acknowledgements

We thank Zeev Dvir, Ariel Gabizon, Swastik Kopparty, and Ronen Shaltiel for useful discussions. Thanks for Zeev Dvir for confirming that the extractors of [DKSS09] work with smaller ε than claimed in the paper.

References

[DKSS09] Zeev Dvir, Swastik Kopparty, Shubhangi Saraf, and Madhu Sudan. Extensions to the method of multiplicities, with applications to Kakeya sets and mergers. In *FOCS*, pages 181–190. IEEE Computer Society, 2009.

[DW11] Zeev Dvir and Avi Wigderson. Kakeya sets, new mergers, and old extractors. *SIAM J. Comput.*, 40(3):778–792, 2011.

[GS99] Venkatesan Guruswami and Madhu Sudan. Improved decoding of Reed-Solomon and algebraic-geometry codes. *IEEE Transactions on Information Theory*, 45(6):1757–1767, 1999.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil P. Vadhan. Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes. *J. ACM*, 56(4), 2009.

[LN94] Rudolf Lidl and Harald Niederreiter. *Finite fields and their applications*. Cambridge University Press, 1994.

[LRVW03] Chi-Jen Lu, Omer Reingold, Salil P. Vadhan, and Avi Wigderson. Extractors: optimal up to constant factors. In Lawrence L. Larmore and Michel X. Goemans, editors, *STOC*, pages 602–611. ACM, 2003.

[PV05] Farzad Parvaresh and Alexander Vardy. Correcting errors beyond the Guruswami-Sudan radius in polynomial time. In *FOCS*, pages 285–294. IEEE Computer Society, 2005.

[RSW06] Omer Reingold, Ronen Shaltiel, and Avi Wigderson. Extracting randomness via repeated condensing. *SIAM J. Comput.*, 35(5):1185–1209, 2006.

[SS08] Shubangi Saraf and Madhu Sudan. Improved lower bound on the size of Kakeya sets over finite fields. 2008. arXiv:0808.2499v2.

[TS96] Amnon Ta-Shma. On extracting randomness from weak random sources (extended abstract). In *STOC*, pages 276–285, 1996.

[WZ99] Avi Wigderson and David Zuckerman. Expanders that beat the eigenvalue bound: Explicit construction and applications. *Combinatorica*, 19(1):125–138, 1999.