Electronic Colloquium on Computational Complexity, Report No. 134 (2006)

# Extractors and condensers from univariate polynomials

Venkatesan Guruswami*
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195
venkat@cs.washington.edu

Christopher Umans†
Computer Science Department
California Institute of Technology
Pasadena, CA 91125
umans@cs.caltech.edu

Salil Vadhan‡
Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
salil@eecs.harvard.edu

October 18, 2006

**Abstract**

We give new constructions of randomness extractors and lossless condensers that are optimal to within constant factors in both the seed length and the output length. For extractors, this matches the parameters of the current best known construction [LRVW03]; for lossless condensers, the previous best constructions achieved optimality to within a constant factor in one parameter only at the expense of a polynomial loss in the other.

Our constructions are based on the Parvaresh-Vardy codes [PV05], and our proof technique is inspired by the list-decoding algorithm for those codes. The main object we construct is a condenser that loses *only* the entropy of its seed plus one bit, while condensing to entropy rate $1 - \alpha$ for any desired constant $\alpha > 0$. This construction is simple to describe, and has a short and completely self-contained analysis. Our other results only require, in addition, standard uses of randomness-efficient hash functions (to obtain a lossless condenser) or expander walks (to obtain an extractor).

Our techniques also show for the first time that a natural construction based on univariate polynomials (i.e., Reed-Solomon codes) yields a condenser that retains a $1 - \alpha$ fraction of the source min-entropy, for any desired constant $\alpha > 0$, while condensing to constant entropy rate and using a seed length that is optimal to within constant factors.

*Supported by NSF CCF-0343672, a Sloan Research Fellowship, and a David and Lucile Packard Foundation Fellowship.
†Supported by NSF CCF-0346991, BSF 2004329, a Sloan Research Fellowship, and an Okawa Foundation research grant.
‡Supported by NSF grant CCF-0133096, ONR grant N00014-04-1-0478, and US-Israel BSF grant 2002246.



ISSN 1433-8092

# 1 Introduction

In this paper, we construct randomness extractors and condensers with the best parameters to date. Perhaps more importantly, we do this by introducing a new algebraic construction based on the ingenious variant of Reed-Solomon codes discovered by Parvaresh and Vardy [PV05]. Our proof technique is inspired by the list-decoding algorithm for the Parvaresh-Vardy codes, which builds on the list-decoding results of [Sud97, GS99]. The resulting extractors and condensers are simple to describe and have short, self-contained analyses. In the remainder of the introduction. we describe our results more precisely, and place them in context within the large body of literature on extractors and related objects.

A long line of research beginning in the late 1980s has been devoted to the goal of constructing explicit *randomness extractors*. (See the survey of Shaltiel [Sha02].) Extractors are efficient functions that take an $n$-bit string sampled from a "weak" random source together with a short truly random seed, and output a nearly uniform distribution.

The randomness in the source is measured by *minentropy*: a random variable $\mathbf{X}$ has minentropy at least $k$ iff $\Pr[\mathbf{X} = x] \leq 2^{-k}$ for all $x$. A random variable $\mathbf{Z}$ is *$\varepsilon$-close* to a distribution $D$ if for all events $A$, $\Pr[\mathbf{Z} \in A]$ differs from the probability of $A$ under the distribution $D$ by at most $\varepsilon$. An extractor is defined as follows:

**Definition 1.1 ([NZ96]).** *A $(k, \varepsilon)$* extractor *is a function $E : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ with the property that for every $\mathbf{X}$ with minentropy at least $k$, $E(\mathbf{X}, \mathbf{Y})$ is $\varepsilon$-close to uniform, when $\mathbf{Y}$ is uniformly distributed on $\{0,1\}^t$. An extractor is* explicit *if it is computable in polynomial time.*

The competing goals when constructing extractors are to obtain a short seed, and a long output length. Nonconstructively, it is possible to simultaneously have a seed length $t = \log n + 2\log(1/\varepsilon) + O(1)$ and an output length of $m = k + t - 2\log(1/\varepsilon) - O(1)$. It remains open to match these parameters with an explicit construction.

A major theme in extractor constructions since the breakthrough result of Trevisan [Tre01], has been the use of error-correcting codes. Trevisan's extractor construction, which is based on the Nisan-Wigderson pseudorandom generator [NW94], encodes the source with an error-correcting code with good distance, and uses the seed to select (via certain combinatorial designs) a subset of $m$ bits of the codeword to output.

A more algebraic approach, exploiting the specific structure of polynomial error-correcting codes was pioneered by Ta-Shma, Zuckerman and Safra [TZS06]. There the source is encoded with a multivariate polynomial code (Reed-Muller code), the seed is used to select a starting point, and the extractor outputs $m$ successive symbols along a line[1]. Better parameters were achieve with a variant introduced by Shaltiel and Umans [SU05], which exploits the fact that Reed-Muller codes are *cyclic*. There the $m$ output symbols are simply $m$ successive coordinates of the codeword, when written in the cyclic ordering. A common feature of these algebraic constructions is that their analysis relies crucially on the *local-decodability* properties of the underlying error-correcting code. This paper diverges from the previous works on exactly this point, as our constructions use only univariate polynomial codes, which are not locally decodable.

A second major theme dating to [RSW06] and [RR99][2] is the use of a relaxation of extractors, called *condensers*, as an intermediate goal:

**Definition 1.2.** *A $(n, k) \to_\varepsilon (m, k')$* condenser *is a function $C : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ with the property that for every $\mathbf{X}$ with minentropy at least $k$, $C(\mathbf{X}, \mathbf{Y})$ is $\varepsilon$-close to a distribution with minentropy*

[1] In this discussion we are ignoring the distinction between outputting $m$ symbols from a large alphabet and outputting $m$ bits.

[2] Actually, since the formal definition we give does not explicitly require that the min-entropy rate increase, such objects were already considered as far back as the original papers of [Zuc96, NZ96]. However, we will be interested in condensers that do actually increase the min-entropy rate.

*$k'$, when $\mathbf{Y}$ is uniformly distributed on $\{0,1\}^t$. A condenser is* explicit *if it is computable in polynomial time. A condenser is called* lossless *if $k' = k + t$.*

Observe that a $(n,k) \to_\varepsilon (m,m)$ condenser is an extractor, because the unique distribution on $\{0,1\}^m$ with minentropy $m$ is the uniform distribution. Condensers are a natural stepping-stone to constructing extractors, as they can be used to increase the *entropy rate* (the ratio of the minentropy in a random variable to the length of the strings over which it is distributed), and it is often easier to construct extractors when the entropy rate is high. Condensers have also been used extensively in less obvious ways to build extractors, often as part of complex recursive constructions (e.g., [ISW00, RSW06, LRVW03]). Nonconstructively, one can hope for *lossless* condensers with seed length $t = \log n + \log(1/\varepsilon) + O(1)$, and output length $m = k + t + \log(1/\varepsilon) + O(1)$.

Our central result is a completely elementary construction of a condenser that retains all but the seed min-entropy (plus one bit), and condenses to *any* constant entropy rate using a seed length that is optimal up to constant factors. This is the most basic object from which we derive most of the other results:

**Theorem 1.1 (main).** *For every $1 \geq \alpha > 0$: for all positive integers $n, k$ and all $\varepsilon > 0$, there is an explicit construction of a*

$$(n, k' = kt + \log(1/\varepsilon)) \to_{3\varepsilon} (n' = (1+\alpha)kt, k' - 1)$$

*condenser $C : \{0,1\}^n \times \{0,1\}^{(1+\alpha)t} \to \{0,1\}^{n'}$ with $t = \left\lceil \frac{1}{\alpha}(2\log n + \log(\frac{2}{\varepsilon})) \right\rceil$.*

In recent years, condensers have been studied in their own right. Lossless condensers are of particular interest, as they are equivalent to unbalanced bipartite *expander graphs* with extremely good expansion (of greater than half the left degree of the graph). This turns out to be useful in a number of applications; constructions of lossless condensers appear in [RR99, TUZ01, CRVW02, TU06].

For lossless condensers, the competing goals are short seed length, and *short* output length (thus achieving the greatest "condensing" of the source minentropy). Constructions are known that achieve essentially optimal parameters for very large $k$ [CRVW02], and very small $k$ [RR99], but for general $k$, the best known constructions can achieve optimality to within a constant factor in one parameter only at the expense of a polynomial loss in the other. Specifically, the best known constructions (stated here for constant $\varepsilon$) achieve seed length $t = O(\log^2 n)$ and output length $m = O(k)$ [TUZ01], or seed length $t = O(\log n)$ and output length $m = k^{1+\alpha}$ for any constant $\alpha > 0$ [TUZ01]. Recently Ta-Shma and Umans [TU06] showed that if optimal *derandomized curve samplers* can be constructed, then a construction of lossless condensers based on [SU05] would achieve seed length $t = O(\log n)$ and output length $m = k \cdot \operatorname{poly}\log(n)$; they obtain near-optimal derandomized curve samplers that produce lossless condensers with somewhat worse parameters.

Using Theorem 1.1, we obtain a new construction of lossless condensers that are optimal to within constant factors in both the seed length and the output length. This uses an idea from [RR99]: because the condenser of Theorem 1.1 is only missing a small amount of minentropy, it can be made lossless by appending a hash from an "almost-2-universal" hash family; we pay only with a constant factor increase in the seed length. We obtain:

**Theorem 1.2 (lossless condenser).** *For every constant $\alpha > 0$: for all positive integers $n, k$ and all $\varepsilon > 0$, there is an explicit construction of a*

$$(n, k + \log(1/\varepsilon)) \to_{6\varepsilon} (m = (1+\alpha)k, k + d + \log(1/\varepsilon))$$

*lossless condenser $C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = O(\log n + \log(1/\varepsilon))$, provided $k \geq cd/\alpha$ for a universal constant $c$.*

We now return to extractors. There is a great diversity of extractor constructions; see Shaltiel's survey [Sha02] for a nearly-up-to-date summary. The current champion is the construction of Lu, Reingold, Vadhan, and Wigderson [LRVW03] which achieves optimality to within a constant factor in the seed length and output length simultaneously, for any minentropy $k$. (As with lossless condensers, for small $k$, better constructions are known; e.g., [SZ99, TUZ01]). Again using the condenser of Theorem 1.1, we can match this best known construction with a simple, direct, and self-contained construction and analysis. We simply need to "finish" the condenser of Theorem 1.1 with an extractor that extracts any desired constant fraction of the minentropy, with a seed length that is optimal up to constant factors. Since this extractor can start from a constant entropy rate arbitrarily close to 1, we can even use a standard extractor based on expander walks [IZ89, CW89, Gil98, Zuc06]. When $\varepsilon$ is sub-constant, we use Zuckerman's extractor [Zuc97] to obtain the proper dependence on $\varepsilon$. Altogether we obtain:

**Theorem 1.3 (extractor).** *For all constants $\alpha, \gamma > 0$: for all positive integers $n, k$ and all $\varepsilon > \exp(-n^{1-\gamma})$, there is an explicit construction of a $(k, \varepsilon)$ extractor $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = O(\log n + \log \frac{1}{\varepsilon})$ and $m = (1-\alpha)k$, provided $k \geq cd/\alpha$ for a universal constant $c$.*

In fact this result slightly improves upon [LRVW03], for general error $\varepsilon = \varepsilon(n)$. They can handle error as small as $n^{-1/\log^{(c)} n}$ for any constant $c$, but for general $\varepsilon$, they must pay with either a larger seed length of $t = O((\log^* n)^2 \log n + \log(\frac{1}{\varepsilon}))$, or a smaller output length of $m = \Omega(k/\log^{(c)} n)$ for any constant $c$.

## 1.1 Our technique

In this section we give a high-level description of our construction and proof technique. Our condensers are based on Parvaresh-Vardy codes [PV05], which in turn are based on Reed-Solomon codes. A Reed-Solomon codeword is a univariate degree $n$ polynomial $f \in \mathbb{F}_q[Y]$, evaluated at all points in the field. A Parvaresh-Vardy codeword is a bundle of several related degree $n$ polynomials $f_0, f_1, f_2, \ldots, f_{m-1}$, each evaluated at all points in the field. The evaluations of the various $f_i$ at a given field element are packaged into a symbol from the larger alphabet $\mathbb{F}_{q^m}$. The purpose of this extra redundancy is to enable a better list-decoding algorithm than is possible for Reed-Solomon codes.

The main idea in [PV05] is to view degree $n$ polynomials as elements of the extension field $\mathbb{F} = \mathbb{F}_q[Y]/E(Y)$, where $E$ is some irreducible polynomial of degree $n+1$. The $f_i$ (now viewed as elements of $\mathbb{F}$) are chosen so that $f_i = f_0^{h_i}$ for $i \geq 1$, and positive integers $h_i$. In order to list-decode, one produces a nonzero univariate polynomial $Q'$ over $\mathbb{F}$ from the received word, with the property that $f_0$ is a root of $Q'$ whenever the codeword has sufficient agreement with the received word. We use the same technique in the analysis of our condenser, and below we describe how the interpolating polynomial is set up and how the relationship between the $f_i$'s helps in the context of our analysis.

Our condenser construction works as follows. We view the source string $x$ as describing a degree $n$ polynomial $f(Y) \in \mathbb{F}_q[Y]$. We then define $f_i \stackrel{\text{def}}{=} f^{h^i} \bmod E$ for some parameter $h$, and irreducible $E$. Given a seed $y \in \mathbb{F}_q$, our output is $f_0(y), f_1(y), \ldots, f_{m-1}(y)$.

Since [Tre01], a common technique in analyzing extractors has been to show that for every subset $D \subseteq \{0,1\}^m$, there are very few, say $\ll 2^k$, source strings $x$ that are "bad" with respect to $D$; i.e., much fewer than $2^k$ strings $x$ satisfy

$$\left| \Pr_y[E(x,y) \in D] - \Pr_z[z \in D] \right| > \varepsilon.$$

From this, it follows that a source with min-entropy $k$ is unlikely output a string that is bad with respect to any given $D$. Thus the output of $E$ on such a source must hit all $D$'s with close to the "proper" probability,

and so $E$ is an extractor for minentropy $k$. We use the same general outline to show that our construction is a condenser. We only wish to show that the output is close to having minentropy $k'$, rather than close to being uniform, and this is equivalent to showing that the output hits sets $S$ of size about $2^{k'}$ with less than $\varepsilon$ probability (see Section 2.1 for a precise statement of this fact). We do this by arguing that there are very few source strings $x$ that are "bad" with respect to $S$; i.e., very few $x$ satisfy $\Pr_y[C(x,y) \in S] > \varepsilon$.

Let's consider what $\Pr_y[C(x,y) \in S] > \varepsilon$ means for our construction. First of all, $x$ is interpreted as a degree $n$ polynomial $f_0$. Then, $f_0$ being "bad" means that for more than $\varepsilon q$ of the seeds $y$, we have

$$(f_0(y), f_1(y), \ldots, f_{m-1}(y)) \in S.$$

The first step in our analysis is to produce a non-zero polynomial $Q : \mathbb{F}_q^m \to \mathbb{F}_q$ that vanishes on $S$. We arrange to have $n \deg Q < \varepsilon q$, so that the univariate polynomial $Q(f_0(Y), f_1(Y), \ldots, f_{m-1}(Y))$ is *identically zero* for bad $f_0$. Viewing the $f_i$ as elements of the extension field $\mathbb{F} = \mathbb{F}_q[Y]/E(Y)$, and $Q$ as a polynomial over $\mathbb{F}$, we have that $(f_0, f_1, \ldots, f_{m-1})$ is a *root* of $Q$. Just as in the list-decoding algorithm of [PV05], we define the polynomial $Q'(Z) \stackrel{\text{def}}{=} Q(Z, Z^h, Z^{h^2}, \ldots, Z^{h^{m-1}})$, and observe that every bad $f_0$ is a root of this *univariate* polynomial. Thus the degree of $Q'$ is a bound on the number of such $f_0$, and it turns out that this bound is nearly optimal: the number of bad $f_0$ is shown to be at most the size of $S$.

To summarize, the analysis has two main steps: first, we encode $S$ into a low-degree multivariate polynomial $Q$, and argue that for every bad polynomial $f_0(Y)$, $Q(f_0(Y), \ldots, f_{m-1}(Y))$ is in fact identically zero. Then, we produce a univariate polynomial $Q'$ from $Q$ that has all of the bad $f_0$ as roots (when everything is viewed over the extension field $\mathbb{F}$). The degree of $Q'$ is an upper bound on the number of bad strings.

### 1.2 Additional results

In Section 6 we discuss some variations on the basic construction.

Using the "multiple roots" idea from Guruswami-Sudan [GS99], we optimize the seed length of our condenser, making it $(1+\gamma)$ times the optimal seed length, while still retaining almost all the entropy and outputting a source with a constant entropy rate of $\Omega(\gamma)$ (Theorem 6.2). For constant error $\varepsilon$, one can then extract almost all the entropy using the extractor from [Zuc06] which uses an additional seed of at most $\log k + O(1)$ bits. The total seed length is thus $(1+\gamma)\log n + \log k + O(1)$, which approaches the optimal $\log n + O(1)$ bound for $k = n^{o(1)}$. This result appears as Theorem 6.5. A different setting of the condenser parameters (Corollary 6.3) allows us to obtain an *exactly* optimal seed length, while retaining a constant fraction (arbitrarily close to 1) of the entropy, at the expense of an output entropy rate of $\Omega(1/\log(n/\varepsilon))$, which is still quite good.

With a small change to the original proof, we can say something about the variant of the main condenser in which the seed is included in the output. One can hope to capture the entire seed entropy (which we do in Theorem 1.2, but that involves the extra step of appending a hash); here we are able to capture all but $O(\log(1/\varepsilon))$ bits of the seed entropy directly.

Finally, using one of the main ideas from the Guruswami-Rudra codes [GR06], we argue that a variant of our main construction is the natural precursor of [SU05], in which that basic construction is applied Reed-Solomon codes. It has been an intriguing question for some time to determine what (if any) pseudorandom object(s) can be obtained from this very natural construction. This question is studied in [KU06], where they show that the Reed-Solomon construction "fools" certain kinds of low-degree tests. Our results in this paper, which show that this construction is a very good condenser, seem to provide the correct (or nearly-correct) answer, as we also describe an example that shows that the entropy rate and the constant factor entropy loss for this construction cannot be improved substantively.

## 2 Preliminaries

Throughout this paper, we use boldface capital letters for random variables (e.g., "$\mathbf{X}$"), capital letters for indeterminates, and lower case letters for elements of a set. Also throughout the paper, $\mathbf{U_t}$ is the random variable uniformly distributed on $\{0,1\}^t$. All logs are base 2.

We record some standard facts about minentropy:

**Proposition 2.1.** *A distribution $D$ has minentropy at least $k$ iff $D$ is a convex combination of flat distributions on sets of size exactly $2^k$.*

**Proposition 2.2.** *The distance from a distribution $D$ to a closest distribution with minentropy $k$ is exactly $\sum_{a:D(a)\geq 2^{-k}}(D(a)-2^{-k})$.*

**Proposition 2.3.** *A distribution $D$ with minentropy $\log(K-c)$ is $c/K$-close to a distribution with minentropy $\log K$.*

*Proof.* By Proposition 2.1, it suffices to prove the statement for flat distributions $D$. By Proposition 2.2, the distance from $D$ to the closest distribution with minentropy $\log K$ is exactly $\sum_{a:D(a)\geq 1/K}(D(a)-1/K) = (K-c)(1/(K-c)-1/K) = c/K$. □

### 2.1 Analysis of condensers

The next lemma gives a useful sufficient condition for a distribution to be close to having large minentropy:

**Lemma 2.4.** *Let $\mathbf{Z}$ be a random variable. If for all sets $S$ of size $K$, $\Pr[\mathbf{Z}\in S]\leq\varepsilon$ then $\mathbf{Z}$ is $\varepsilon$-close to having minentropy at least $\log(K/\varepsilon)$.*

*Proof.* Let $S$ be a set of the $K$ heaviest elements $x$ (under the distribution of $\mathbf{Z}$). Let $2^{-\ell}$ be the average weight of the elements in $S$. Then $\varepsilon\geq\Pr[\mathbf{Z}\in S]=2^{-\ell}K$, so $\ell\geq\log(K/\varepsilon)$. But every element outside $S$ has weight at most $2^{-\ell}$, and with all but probability $\varepsilon$, $\mathbf{Z}$ hits elements outside $S$. □

This lemma establishes the framework within which we will prove our constructions are condensers:

**Lemma 2.5.** *Let $C:\{0,1\}^n\times\{0,1\}^d\to\{0,1\}^m$ be a function. For each subset $S$, define*

$$\mathit{BAD}(S,\varepsilon)=\left\{x:\Pr_y[C(x,y)\in S]>\varepsilon\right\}.$$

*Let $B(K,\varepsilon)=\max_{S:|S|=K}|\mathit{BAD}(S,\varepsilon)|$. Then the function $C$ is a*

$$(n,\log(B(K,\varepsilon)/\varepsilon))\to_{2\varepsilon}(m,\log(K/\varepsilon)-1)$$

*condenser.*

*Proof.* We have a random variable $\mathbf{X}$ with minentropy $\log(B(K,\varepsilon)/\varepsilon)$. For a fixed $S$ of size $K$, the probability that $\mathbf{X}$ is in BAD$(S,\varepsilon)$ is at most $\varepsilon$; if that does not happen, then the probability $C(\mathbf{X},\mathbf{U_t})$ lands in $S$ is at most $\varepsilon$. Altogether the probability $C(\mathbf{X},\mathbf{U_t})$ falls in $S$ is at most $2\varepsilon$. Now apply Lemma 2.4. □

## 3 The main construction

Fix the field $\mathbb{F}_q$ and let $E(Y)$ be an irreducible polynomial of degree $n+1$ over $\mathbb{F}_q$. View elements of $\{0,1\}^n$ as describing univariate polynomials over $\mathbb{F}_q$ with degree at most $n$. Fix an integer parameter $h$.

We describe a function $C : \{0,1\}^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ that is the basis of all of our constructions:

$$C(f,y) \stackrel{\text{def}}{=} f(y) \circ (f^h \bmod E)(y) \circ (f^{h^2} \bmod E)(y) \circ \cdots \circ (f^{h^{m-1}} \bmod E)(y).$$

For ease of notation, we will refer to $(f^{h^i} \bmod E)$ as "$f_i$."

**Lemma 3.1.** *Defining BAD$(S,\varepsilon)$ and $B(K,\varepsilon)$ with respect to $C$ as in Lemma 2.5, we have*

$$B(K = h^m - 1, \varepsilon) \leq K,$$

*provided $q \geq nm(h-1)/\varepsilon$.*

*Proof.* Fix a set $S \subseteq \mathbb{F}_q^m$ of size at most $K$. Let $Q \in \mathbb{F}_q[Z_1, Z_2, \ldots, Z_m]$ be a nonzero $m$-variate polynomial that vanishes on $S$, and with individual degrees at most $h-1$. By definition, for every $f(Y) \in \text{BAD}(S,\varepsilon)$, it holds that

$$\Pr_y[Q(f_0(y), f_1(y), \ldots, f_{m-1}(y)) = 0] > \varepsilon.$$

Therefore, the univariate polynomial $R(Y) \stackrel{\text{def}}{=} Q(f_0(Y), \ldots, f_{m-1}(Y))$ has more than $\varepsilon q$ zeroes, and degree at most $nm(h-1)$. Since $nm(h-1) \leq \varepsilon q$, $R(Y)$ must be identically zero, and so

$$Q(f_0(Y), \ldots, f_{m-1}(Y)) = 0$$

for every $f(Y) \in \text{BAD}(S,\varepsilon)$.

Now, view $Q$ as a multivariate polynomial over the extension field $\mathbb{F} = \mathbb{F}_q[Y]/E(Y)$, and define

$$Q'(Z) \stackrel{\text{def}}{=} Q(Z, Z^h, Z^{h^2}, \ldots, Z^{h^{m-1}}).$$

Because the individual degrees of $Q$ were all less than $h$, $Q'$ is a non-zero polynomial (because distinct monomials in $Q$ map to distinct monomials in $Q'$).

For every $f(Y) \in \text{BAD}(S,\varepsilon)$, now viewed as an element of $\mathbb{F}$, we have

$$Q'(f) = Q(f_0, f_1, f_2, \ldots, f_{m-1}) = 0,$$

i.e., $f$ is a root of $Q'$. Thus $|\text{BAD}(S,\varepsilon)| \leq \deg(Q')$. The degree of $Q'$ is at most

$$(h-1)(1 + h + h^2 + \cdots + h^{m-1}) = h^m - 1 = K.$$

□

We can now prove our main theorem (restated here):

**Theorem 1.1 (restated).** *For every $1 \geq \alpha > 0$: for all positive integers $n, k$ and all $\varepsilon > 0$, there is an explicit construction of a*

$$(n, k' = kt + \log(1/\varepsilon)) \to_{3\varepsilon} (n' = (1+\alpha)kt, k' - 1)$$

*condenser $C : \{0,1\}^n \times \{0,1\}^{(1+\alpha)t} \to \{0,1\}^{n'}$ with $t = \left\lceil \frac{1}{\alpha}(2\log n + \log(\frac{2}{\varepsilon})) \right\rceil$.*

*Proof.* We describe how to set parameters, and then apply Lemmas 3.1 and 2.5. Set $h = 2^t$, note that $h \geq (2n^2/\varepsilon)^{1/\alpha}$. Let $q$ be the largest prime less than or equal to $h^{1+\alpha}$. By Bertrand's Postulate, first proved by Chebyshev, we have $h^{1+\alpha}/2 \leq q \leq h^{1+\alpha}$. Since we may assume $m \leq n$, we have $q \geq nmh/\varepsilon$ as required. Set $m = k$.

The function $C$ has output length

$$m \log q \leq m(1+\alpha) \log h = (1+\alpha)kt$$

as claimed (we can pad the condenser with dummy bits to make the output length exactly $(1+\alpha)kt$). By Lemma 3.1, and Lemma 2.5, $C$ is a

$$(n, \log((h^m - 1)/\varepsilon)) \to_{2\varepsilon} ((1+\alpha)kt, \log((h^m - 1)/\varepsilon) - 1)$$

condenser. All that remains is numerical manipulation to express this in the same way as it is stated in the theorem. First, note that

$$\log((h^m - 1)/\varepsilon) < \log(h^m/\varepsilon) = kt + \log(1/\varepsilon)\ .$$

Also, by Proposition 2.3, a distribution with $\log((h^m - 1)/\varepsilon) - 1$ minentropy is $1/h^m$ close to having minentropy

$$\log(h^m/\varepsilon) - 1 = m \log h + \log(1/\varepsilon) - 1 = kt + \log(1/\varepsilon) - 1.$$

Since $1/h^m$ is always at most $\varepsilon$, $C$ is a $(n, kt + \log(1/\varepsilon)) \to_{3\varepsilon} ((1+\alpha)kt, kt + \log(1/\varepsilon) - 1)$ condenser as claimed. The seed length is $\log q \leq (1+\alpha) \log h = (1+\alpha)t$. □

**Remark 1.** *In this proof we work in a prime field $\mathbb{F}_q$. The same proof works over any field $\mathbb{F}_q$, with a minor adjustment to the inequality describing how close $q$ is to $h^{1+\alpha}$.*

# 4 Lossless condensers that are optimal up to constant factors

We begin with the general method to recover "missing" minentropy, first used by [RR99]. Given a $(n, k) \to_\varepsilon (m, k')$ condenser, we say it has entropy loss $d = k + t - k'$. We can make the condenser lossless by appending a random hash into $O(d + \log(1/\varepsilon))$ bits. When $d$ is small, the extra randomness can also be small, provided we use a randomness-efficient family of hash functions. Next, we describe the "almost 2-universal" hash family that we will use:

**Theorem 4.1 ([AGHP92, SZ99]).** *For every $n', m'$, there exists an explicit family $H$ of hash functions from $n'$ to $m'$ bits, of cardinality $O((n'm'2^{m'})^2)$, that satisfies the following property:*

$$\forall w_1 \neq w_2 \quad \Pr_{h \in H}[h(w_1) = h(w_2)] \leq 2 \cdot 2^{-m'}. \tag{1}$$

*A random $h \in H$ can be sampled using $\log |H|$ bits, and given these bits, $h$ can be computed in poly$(n', m')$ time.*

Note that a truly 2-universal hash function would satisfy (1) with the right-hand-side replaced by $2^{-m'}$ – but the price would be that $|H| \geq 2^{n'}$, which is far too large to be useful for us. Now we show that appending a random hash makes a condenser lossless.

**Lemma 4.2.** *Let $C : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ be a $(n,k) \to_\varepsilon (m,k')$ condenser. Let $H$ be a family of hash functions from $n' = n + t$ bits to $m' = 2(k + t - k') + \log(1/\varepsilon) + 1$ bits satisfying (1). Then the function $C' : \{0,1\}^n \times \{0,1\}^{t'=t+\log|H|} \to \{0,1\}^{m+\log|H|+m'}$ defined by:*

$$C'(x; y, h \in H) \stackrel{\text{def}}{=} C(x,y) \circ h \circ h(x,y)$$

*is a $(n,k) \to_{2\varepsilon} (m + \log|H| + m', k + t')$ lossless condenser.*

*Proof.* Let $\mathbf{X}$ be a random variable distributed uniformly on an arbitrary set of size $2^k$. We prove that $C'$ is the stated condenser when its source is $\mathbf{X}$, which by Proposition 2.1 suffices. We denote by $\mathbf{H}$, the random variable that is uniformly distributed over the hash functions in $H$. We also take $\mathbf{Y}$ to be a random variable uniformly distributed on $\{0,1\}^t$.

Call $z \in \{0,1\}^m$ "good" if $\Pr[C(\mathbf{X},\mathbf{Y}) = z] \le 2^{-k'}$. Observe that by Proposition 2.2, $C(\mathbf{X},\mathbf{Y})$ is good with all but $\varepsilon$ probability.

Define $S_z = \{(x,y) : C(x,y) = z\}$, and call $h$ "good with respect to $z$" if $h$ is 1-1 on $S_z$. Notice that for an arbitrary set $S$,

$$\Pr[\mathbf{H} \text{ is not 1-1 on } S] \le \sum_{w_1,w_2 \in S, w_1 \neq w_2} \Pr[\mathbf{H}(w_1) = \mathbf{H}(w_2)] \le \frac{|S|^2}{2^{m'-1}}.$$

Since $|S_z| = 2^{k+t}\Pr[C(\mathbf{X},\mathbf{Y}) = z]$, we have that for good $z$, $|S_z| \le 2^{k+t-k'}$. Therefore, for good $z$, $\mathbf{H}$ is good with respect to $z$ with all but $\varepsilon$ probability.

We now argue that the output distribution of $C'$ is $2\varepsilon$-close to having minentropy $k + t'$. Fix an output string $(z,h,z')$. If $z$ is good, and $h$ is good with respect to $z$, then

$$\begin{aligned}
&\Pr[C(\mathbf{X},\mathbf{Y}) = z \wedge \mathbf{H} = h \wedge \mathbf{H}(\mathbf{X},\mathbf{Y}) = z'] \\
&= \Pr[C(\mathbf{X},\mathbf{Y}) = z] \cdot \frac{1}{|H|} \cdot \Pr[\mathbf{H}(\mathbf{X},\mathbf{Y}) = z' | \mathbf{H} = h, C(\mathbf{X},\mathbf{Y}) = z] \\
&= \Pr[C(\mathbf{X},\mathbf{Y}) = z] \cdot \frac{1}{|H|} \cdot \frac{1}{|S_z|} \\
&= \Pr[C(\mathbf{X},\mathbf{Y}) = z] \cdot \frac{1}{|H|} \cdot \frac{1}{2^{k+t}\Pr[C(\mathbf{X},\mathbf{Y}) = z]} = \frac{1}{2^{k+t}|H|} = 2^{-(k+t')}.
\end{aligned}$$

As we have argued, we hit a good $z$ with all but $\varepsilon$ probability, and then $\mathbf{H}$ is good with respect to $z$ with all but $\varepsilon$ probability. Overall, with all but $2\varepsilon$ probability, we hit an output string with weight $2^{-(k+t')}$, as required. □

Applying this transformation to the condenser from Theorem 1.1, we obtain our second main theorem, restated here:

**Theorem 1.2 (restated).** *For every constant $\alpha > 0$: for all positive integers $n, k$ and all $\varepsilon > 0$, there is an explicit construction of a*

$$(n, k + \log(1/\varepsilon)) \to_{6\varepsilon} (m = (1+\alpha)k, k + d + \log(1/\varepsilon))$$

*lossless condenser $C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = O(\log n + \log(1/\varepsilon))$, provided $k \ge cd/\alpha$ for a universal constant $c$.*

*Proof.* Consider the condenser of Theorem 1.1 with its parameter $\alpha$ set to half the present $\alpha$, which has seed length $(1+\alpha/2)t$ where

$$t = \left\lceil \frac{2}{\alpha}(2\log n + \log(2/\varepsilon)) \right\rceil .$$

We set that condenser's parameter $k$ to the present $k$ divided by $t$, rounded down. It then has entropy deficiency at most $(2+\alpha/2)t+1$ (up to $t$ is attributable to the rounding down, and the $(1+\alpha/2)t$ seed bits are lost, plus one).

Now apply Lemma 4.2. The output length of the hash is $m' = O(\log n + \log(1/\varepsilon))$, and the number of bits needed to sample from $H$ is $2m' + O(\log n) + O(\log\log(1/\varepsilon))$. The resulting condenser is lossless, and it has the stated seed length. Its output length is at most

$$(1+\alpha/2)k + \log(1/\varepsilon) + m' + \log|H| \le (1+\alpha/2)k + O(\log n + \log(1/\varepsilon)),$$

which by our lower bound on $k$ is at most $(1+\alpha)k$. □

# 5 Extractors that are optimal up to constant factors

Once we have condensed all (or almost all) of the entropy into a source with entropy rate close to 1, extracting (most of) that entropy is not that difficult. All we need to do is to compose the condenser with an extractor that works for entropy rates close to 1. The following standard fact makes this formal:

**Proposition 5.1.** *Suppose $C : \{0,1\}^n \times \{0,1\}^{t_1} \to \{0,1\}^{n'}$ is an $(n,k) \to_{\varepsilon_1} (n',k')$ condenser, and $E : \{0,1\}^{n'} \times \{0,1\}^{t_2} \to \{0,1\}^m$ is a $(k',\varepsilon_2)$-extractor, then $E \circ C : \{0,1\}^n \times \{0,1\}^{t_1+t_2} \to \{0,1\}^m$ defined by $(E\circ C)(x,y_1,y_2) \stackrel{\text{def}}{=} E(C(x,y_1),y_2)$ is a $(k,\varepsilon_1+\varepsilon_2)$-extractor.*

For the best dependence on the error parameter $\varepsilon$, the extractor we will use is due to Zuckerman:

**Theorem 5.2 ([Zuc97]).** *For all constants $\alpha,\delta,\gamma > 0$: for all positive integers $n,k$ and all $\varepsilon > \exp(-n^{1-\gamma})$, there is an explicit construction of a $(k=\delta n,\varepsilon)$ extractor $E : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ with $t = O(\log n + \log\frac{1}{\varepsilon})$ and $m = (1-\alpha)k$.*

We now prove our main extractor theorem, restated here:

**Theorem 1.3 (restated).** *For all constants $\alpha,\gamma > 0$: for all positive integers $n,k$ and all $\varepsilon > \exp(-n^{1-\gamma})$, there is an explicit construction of a $(k,\varepsilon)$ extractor $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = O(\log n + \log\frac{1}{\varepsilon})$ and $m = (1-\alpha)k$, provided $k \ge cd/\alpha$ for a universal constant $c$.*

*Proof.* Consider the condenser of Theorem 1.1, with its parameter $\varepsilon$ set to the one sixth of the present $\varepsilon$, and its parameter $\alpha$ set to (say) $1/2$. This condenser has seed length $3t/2$ where

$$t = \lceil 2\cdot(2\log n + \log(12/\varepsilon))\rceil ,$$

and we set its parameter $k$ to be the present $k$ divided by $t$, rounded down, minus $\log(6/\varepsilon)$. The result is a

$$(n,k) \to_{\varepsilon/2} ((3/2)(k-t-1), k-t-1)$$

condenser (the loss of up to $t$ bits comes from the rounding). By the lower bound on $k$, we know that $k-t-1 \ge (1-\alpha/2)k$. Applying Proposition 5.1 to this condenser and the extractor of Theorem 5.2 (with its error parameter $\varepsilon$ set to half the present $\varepsilon$) gives the claimed extractor. □

In the fairly common case that $\varepsilon$ is a constant, we can use the much simpler "expander-walk" extractor (in place of the extractor of Theorem 5.2) which extracts almost all of the entropy for entropy rates close to 1. Note that our condenser from Theorem 1.1 achieves a constant entropy rate arbitrarily close to 1, and so can be combined with any extractor for such high min-entropy rates. A standard construction achieving this is based on expander walks [IZ89, CW89, Gil98]; the following version can be found in [Zuc06]:

**Theorem 5.3.** *For every constant $\alpha > 0$, there is a constant $\delta < 1$ for which the following holds: for all positive integers $n$ and all constant $\varepsilon > 0$, there is an explicit construction of a $(k = \delta n, \varepsilon)$ extractor $E : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ with $t = \log(\alpha n)$ and $m \geq (1-\alpha)n$.*

For completeness, we present the short proof:

*Proof.* Let $m = (1-\alpha)n$, and for some absolute constant $c > 1$, let $G$ be an explicit $2^c$-regular expander on $2^m$ vertices (identified with $\{0,1\}^m$) with second eigenvalue $\lambda = \lambda(G) < 1$. The extractor $E$ is constructed as follows. Its first argument $x$ is used to describe a walk $v_1, v_2, \ldots, v_L$ of length $L$ in $G$ by picking $v_1$ based on the first $m$ bits of $x$, and each further step of the walk from the next $c$ bits of $x$ — so in all, $L$ must satisfy $n = m + (L-1)c$. The seed $y$, which contains more than $\lceil \log L \rceil$ bits, is used to pick one of the vertices of the walk at random. The output $E(x,y)$ of the extractor is the $m$-bit label of the chosen vertex.

Let $\mathbf{X}$ be a random variable with minentropy $k = \delta n$. We wish to prove that for any $S \subseteq \{0,1\}^m$, the probability that $E(\mathbf{X}, \mathbf{U_t})$ is a vertex in $S$ is in the range $\mu \pm \varepsilon$ where $\mu = |S|/2^m$. Fix any such subset $S$. Call an $x \in \{0,1\}^n$ "bad" if

$$\left| \Pr_y[E(x,y) \in S] - \mu \right| > \varepsilon/2.$$

The known Chernoff bounds for random walks on expanders [Gil98] imply that the number of bad $x$'s is at most

$$2^n \cdot e^{-\Omega(\varepsilon^2(1-\lambda)L)} = 2^n \cdot e^{-\Omega(\varepsilon^2(1-\lambda)\alpha n/c)} = 2^n 2^{-\Omega(\varepsilon^2 \alpha n)}$$

(since $c, \lambda$ are absolute constants). Therefore the probability that $\mathbf{X}$ is bad is at most $2^{(1-\delta)n} 2^{-\Omega(\varepsilon^2\alpha n)}$, which is exponentially small for large enough $\delta < 1$. Therefore

$$|\Pr[E(\mathbf{X}, \mathbf{U_t}) \in S] - \mu| \leq \varepsilon/2 + 2^{-\Omega(n)} \leq \varepsilon,$$

implying that $E$ is a $(k, \varepsilon)$-extractor. □

Combining Theorem 1.1 with Theorem 5.3 via Proposition 5.1, as in the proof of Theorem 1.3, we obtain the following extractor, which has the advantage that its proof is short and entirely self-contained:

**Theorem 5.4.** *For every constant $\alpha > 0$: for all positive integers $n, k$, and all constant $\varepsilon > 0$, there is an explicit construction of a $(k, \varepsilon)$ extractor $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d = O(\log n + \log \frac{1}{\varepsilon})$ and $m = (1-\alpha)k$, provided $k \geq cd/\alpha$ for a universal constant $c$.*

# 6 Variations on the main condenser

In this section we show how minor modifications to the proof allow us to optimize the seed length or the output entropy. We also show that a small modification to the construction yields condensers from Reed-Solomon codes.

## 6.1 Optimizing the seed length

The condenser of Theorem 1.1 retains all the source minentropy (except for 1 bit) and achieves an entropy rate of $1-\delta$ for any desired $\delta > 0$. Its main shortcoming is the large seed length, which is greater than $(\log n)/\delta$, whereas the optimal condenser achieves a seed length of $\log n + \log(1/\varepsilon) + O(1)$.

We now show that the seed length can be improved to $(1+\gamma)(\log n + \log(1/\varepsilon))$ — the new condenser still retains a $(1 - O(\frac{1}{\log n}))$ fraction of the input entropy and the output entropy rate is $\Omega(\gamma)$. While the entropy rate is not close to 1 as it was before, it is still a constant, and extractors with seed length of $1 \cdot \log n + O(1)$ were recently constructed for sources of any constant minentropy rate, and constant error $\varepsilon$ [Zuc06] (Theorem 6.4 below). Composing the condenser with such an extractor gives an extractor that extracts $(1-\alpha)k$ bits from a source with minentropy $k$, using seed length $(1+\gamma)\log n + \log k + O(1)$, for arbitrary constants $\alpha, \gamma > 0$. Note that when $k = n^{o(1)}$, the seed length is near-optimal.

The improved analysis that permits us to optimize the seed length is in the following lemma (compare to Lemma 3.1):

**Lemma 6.1.** *Defining BAD$(S,\varepsilon)$ and $B(K,\varepsilon)$ with respect to $C$ as in Lemma 2.5, for any integer parameter $s \geq 1$, we have*

$$B\left(K = \left\lfloor \frac{h^m - 1}{\binom{m+s-1}{s-1}} \right\rfloor, \varepsilon\right) \leq h^m - 1,$$

*provided $q \geq nm(h-1)/(s\varepsilon)$.*

*Proof.* Let $S \subseteq \mathbb{F}_q^m$ be an arbitrary set of size at most $K$. The proof follows along the lines of the proof of Theorem 1.1, with the main change being that we make sure that the interpolated polynomial $Q(Z_1, Z_2, \ldots, Z_m)$ has a root of multiplicity at least $s$ at each element $(\alpha_1, \alpha_2, \ldots, \alpha_m) \in S$. (Note that Theorem 1.1 is the special case of the current theorem with $s = 1$.) This is equivalent to the condition that $Q(Z_1 - \alpha_1, \ldots, Z_m - \alpha_m)$ has no monomials of degree $s-1$ or smaller with nonzero coefficients, which amounts to $\binom{m+s-1}{s-1}$ homogeneous linear constraints on the coefficients of $Q$. Since $h^m > |S|\binom{m+s-1}{s-1}$, such a nonzero polynomial $Q$ of degree at most $(h-1)$ in each variable exists. Fix $Q$ to be any such nonzero polynomial.

Suppose $f(Y) \in \text{BAD}(S,\varepsilon)$. Let $y \in \mathbb{F}_q$ be such that $C(f,y) \in S$. Then certainly

$$Q(f_0(y), f_1(y), \ldots, f_{m-1}(y)) = 0.$$

In fact, since $Q$ has $s$ roots at each element of $S$, the polynomial $R(Y) \stackrel{\text{def}}{=} Q(f_0(Y), f_1(Y), \ldots, f_{m-1}(Y))$ has a root of multiplicity $s$ at $y$. We conclude that if $f(Y) \in \text{BAD}(S,\varepsilon)$, i.e., if

$$\Pr_y[Q(f_0(y), f_1(y), \ldots, f_{m-1}(y)) = 0] > \varepsilon\ ,$$

then $R(Y)$ has more than $\varepsilon sq$ roots counting multiplicities. On the other hand the degree of $R(Y)$ is at most $nm(h-1)$. Therefore, since $\varepsilon sq \geq nm(h-1)$, we must have $R(Y) = 0$.

¿From this point on, the proof proceeds identically to that of Theorem 1.1, leading to the desired conclusion $|\text{BAD}(S,\varepsilon)| \leq h^m - 1$. □

Picking parameters suitably, and following the outline of the proof of Theorem 1.1, we obtain the following condenser:

**Theorem 6.2.** *For every $\gamma > 0$: for all positive integers $n, k$ and all $\varepsilon > 0$, there is an explicit construction of a*

$$(n, k' = kt + \log(1/\varepsilon)) \to_{2\varepsilon} (n' = (1 + 1/\gamma)kt, k' - 3k - 1)$$

*condenser $C : \{0,1\}^n \times \{0,1\}^{\frac{1+\gamma}{\gamma}t} \to \{0,1\}^{n'}$ with $t = \lceil \gamma \log(2n/\varepsilon) \rceil$, provided $t \geq 4$.*

*Proof.* We describe how to set parameters, and then apply Lemmas 6.1 and 2.5. For $t = \lceil \gamma \log(2n/\varepsilon) \rceil$, set $h = 2^t$ and note that note that $h^{1/\gamma} \geq 2n/\varepsilon$. Let $q$ be the largest prime less than or equal to $h^{1+1/\gamma}$. By Bertrand's Postulate, we have $h^{1+1/\gamma}/2 \leq q \leq h^{1+1/\gamma}$. Set $m = s = k$. We have $q \geq nmh/(\varepsilon s) = nh/\varepsilon$ as required.

With this parameter setting, the function $C$ has output length

$$m \log q \leq m(1 + 1/\gamma) \log h = (1 + 1/\gamma)kt$$

as claimed. By Lemma 6.1, and Lemma 2.5, $C$ is a

$$(n, \log((h^m - 1)/\varepsilon)) \to_{2\varepsilon} ((1 + 1/\gamma)kt, \log(K/\varepsilon) - 1)$$

condenser. Now, $K = \lfloor (h^m - 1)/\binom{2m-1}{m-1} \rfloor \geq (h^m - 1)/2^{2m-1} - 1 \geq (h/8)^m$, as long as $h \geq 10$. The theorem follows, using the fact that $\log(h^m) = kt$ and $\log(h/8)^m = k(t - 3)$. □

In the previous theorem, $\gamma$ may be subconstant, and in the following corollary we show that it can be set to produce an exactly optimal seed length (up to the additive constant), while still retaining a constant fraction of the minentropy, at the expense of an entropy rate of $\Omega(1/\log(n/\varepsilon))$, which is non-constant, but still quite good.

**Corollary 6.3.** *For every integer constant $c \geq 4$: for all positive integers $n, k$ and all $\varepsilon > 0$, there is an explicit construction of a*

$$(n, k' = kc + \log(1/\varepsilon)) \to_{2\varepsilon} \left(n' = \left(1 + \frac{\log(2n/\varepsilon)}{c}\right) kc, \left(1 - \frac{3}{c}\right) k' - 1\right)$$

*condenser $C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{n'}$ with $d = \log n + \log(1/\varepsilon) + O(1)$.*

*Proof.* Set $\gamma = c/\log(2n/\varepsilon)$ in Theorem 6.2. □

We now combine the condenser of Theorem 6.2 with Zuckerman's recent extractor. (This extractor in turn starts by applying a condenser due to Raz [Raz05] that has constant seed length and can increase the entropy rate from $\delta$ to $1 - \delta$ for any constant $\delta > 0$, while retaining a constant fraction of the minentropy.)

**Theorem 6.4 ([Zuc06]).** *For all constants $\alpha, \delta > 0$: for all positive integers $n$ and all constant $\varepsilon > 0$, there is an explicit construction of a $(k = \delta n, \varepsilon)$ extractor $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with seed length $d = \log n + O(1)$ and output length $m = (1 - \alpha)k$.*

Combining Theorem 6.2 with Theorem 6.4 via Proposition 5.1, as in the proof of Theorem 1.3, we obtain the following extractor, which has a near-optimal seed length:

**Theorem 6.5.** *For all constants $\alpha, \gamma > 0$: for all positive integers $n, k$ and all constant $\varepsilon > 0$, there is an explicit construction of a $(k, \varepsilon)$ extractor $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with seed length $d = (1 + \gamma) \log n + \log k + O(1)$ and output length $m = (1 - \alpha)k$, provided $k \geq cd/\alpha$ for a universal constant $c$.*

## 6.2 Increasing the output entropy

The condenser of Theorem 1.1 is missing only the entropy of the seed, which is small enough that it can be "recovered" using the hashing technique of Lemma 4.2. However, one can ask how far our new proof technique can go in isolation. More precisely, we modify the function $C$ as follows

$$C'(f, y) \stackrel{\text{def}}{=} (y, C(f, y)),$$

and ask how much entropy is retained for this "strong" variant of the basic construction. It is not hard to see that in the language of Lemma 2.5, we could hope for $B(K, \varepsilon) \leq K/q$, when the seed length is $\log q$. This would correspond to recovering all of the entropy of the source and seed together.

In this section we show that a minor modification to the proof allows us to argue that $B(K, \varepsilon) \leq K/r$ for $r$ approaching $\varepsilon q$. This corresponds to recovering all but $\log(1/\varepsilon) + O(1)$ of the total entropy, although we don't know of a direct use for this improvement. We show the improved result by recording a variant of Lemma 3.1 for $C'$ as defined above:

**Lemma 6.6.** *Defining BAD$(S, \varepsilon)$ and $B(K, \varepsilon)$ with respect to $C'$ as in Lemma 2.5, we have*

$$B(K = rh^m - 1, \varepsilon) < K/r,$$

*where $r = (1 - 1/c)\varepsilon q$, provided $q \geq cnm(h-1)/\varepsilon$, for any $c > 0$.*

*Proof.* Fix a set $S \subseteq \mathbb{F}_q \times \mathbb{F}_q^m$ of size at most $K$. Let $Q \in \mathbb{F}_q[Y, Z_1, Z_2, \ldots, Z_m]$ be a nonzero $m+1$-variate polynomial that vanishes on $S$, with degree $r-1$ in $Y$, and individual degrees at most $h-1$ for the remaining $m$ variables. By definition, for every $f(Y) \in \text{BAD}(S, \varepsilon)$, it holds that

$$\Pr_y[Q(y, f_0(y), f_1(y), \ldots, f_{m-1}(y)) = 0] > \varepsilon.$$

Therefore, the univariate polynomial $R(Y) \stackrel{\text{def}}{=} Q(Y, f_0(Y), \ldots, f_{m-1}(Y))$ has more than $\varepsilon q$ zeroes, and degree at most $r + nm(h-1)$. Since $r + nm(h-1) \leq \varepsilon q$, $R(Y)$ must be identically zero, and so $Q(Y, f_0(Y), \ldots, f_{m-1}(Y)) = 0$ for every bad $f(Y)$.

Now, view $Q$ as a polynomial in $\mathbb{F}_q[Y][Z_1, Z_2, \ldots, Z_m]$, and factor out the largest power of $E(Y)$. Since $E(Y)$ has no roots in $\mathbb{F}_q$, the resulting polynomial still vanishes on $S$. Also, the resulting polynomial is non-zero modulo $E(Y)$; let $Q'$ be the resulting polynomial after reducing modulo $E(Y)$.

Now, view $Q'$ as a multivariate polynomial (in variables $Z_1, Z_2, \ldots, Z_m$) over the extension field $\mathbb{F} = \mathbb{F}_q[Y]/E(Y)$, and define

$$Q''(Z) = Q'(Z, Z^h, Z^{h^2}, \ldots, Z^{h^{m-1}}).$$

Because the individual degrees of $Q'$ are all less than $h$, $Q''$ is a non-zero polynomial (because distinct monomials in $Q'$ map to distinct monomials in $Q''$).

For every $f(Y) \in \text{BAD}(S, \varepsilon)$, now viewed as an element of $\mathbb{F}$, we have $Q''(f) = 0$; i.e., $f$ is a root of $Q''$. Thus $|\text{BAD}(S, \varepsilon)| \leq \deg(Q'')$. The degree of $Q''$ is at most

$$(h-1)(1 + h + h^2 + \cdots + h^{m-1}) = h^m - 1 < K/r.$$

□

### 6.3 Reed-Solomon version

We use one of the main ideas from [GR06] to argue that a small modification to our construction gives a good condenser from Reed-Solomon codes, answering a question raised in [KU06].

Let $q$ be an arbitrary prime power, and let $\zeta \in \mathbb{F}_q$ be a generator of the multiplicative group $\mathbb{F}_q^*$. It is well known, and not hard to show, that $E(Y) = Y^{q-1} - \zeta$ is irreducible over $\mathbb{F}_q$ [LN86, Chap. 3, Sec. 5]. The following identity holds for all $f(Y) \in \mathbb{F}_q[Y]$:

$$(f(Y))^q \bmod E(Y) = f(Y^q) \bmod E(Y) = f(Y^{q-1}Y) \bmod E(Y) = f(\zeta Y) \bmod E(Y) \; .$$

In this case, if we modify our basic function $C : \{0,1\}^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ slightly so that we raise $f$ to successive powers of $q$ rather than $h$, we get:

$$\begin{aligned} C(f,y) &\stackrel{\text{def}}{=} f(y) \circ (f^q \bmod E)(y) \circ (f^{q^2} \bmod E)(y) \circ \cdots \circ (f^{q^{m-1}} \bmod E)(y) \\ &= f(y) \circ f(\zeta y) \circ \cdots \circ f(\zeta^{m-1} y). \end{aligned} \tag{2}$$

In other words, our function interprets its first argument as describing a univariate polynomial over $\mathbb{F}_q$ of degree at most $n$ (i.e., a Reed-Solomon codeword), it uses the seed to select a random location in the codeword, and it outputs $m$ successive symbols of the codeword. This is precisely the analog of the Shaltiel-Umans $q$-ary extractor construction [SU05] for univariate polynomials, rather than multivariate polynomials.

With a minor modification to the proof of Lemma 3.1, we show that this is good condenser:

**Lemma 6.7.** *Defining BAD$(S,\varepsilon)$ and $B(K,\varepsilon)$ with respect to the function $C$ of Equation (2) as in Lemma 2.5, we have*

$$B(K = h^m - 1, \varepsilon) \leq (q^m - 1)(h-1)/(q-1),$$

*provided $q \geq nm(h-1)/\varepsilon$.*

*Proof.* The proof is the same as the proof of Lemma 3.1 except that we define $Q'$ differently:

$$Q'(Z) \stackrel{\text{def}}{=} Q(Z, Z^q, Z^{q^2}, \ldots, Z^{q^{m-1}}).$$

As before, every $f(Y) \in \text{BAD}(S,\varepsilon)$, is a root of $Q'$. Thus $|\text{BAD}(S,\varepsilon)| \leq \deg(Q')$. The degree of $Q'$ is at most

$$(h-1)(1 + q + q^2 + \cdots + q^{m-1}) = (h-1)((q^m - 1)/(q-1)).$$

□

We obtain the following condenser:

**Theorem 6.8 (Reed-Solomon condenser).** *For every constant $1 \geq \alpha > 0$: for all positive integers $n, k$ and all $\varepsilon > 0$, there is an explicit construction of a*

$$(n, (1+\alpha)kt + \log(1/\varepsilon)) \to_{3\varepsilon} (n' = (1+\alpha)kt, kt + \log(1/\varepsilon) - 1)$$

*condenser $C : \{0,1\}^n \times \{0,1\}^{(1+\alpha)t} \to \{0,1\}^{n'}$ with $t = \left\lceil \frac{1}{\alpha}(2\log n + \log(\frac{2}{\varepsilon})) \right\rceil$.*

*Proof.* We describe how to set parameters, and then apply Lemmas 6.7 and 2.5. Set $h = 2^t$, note that $h \geq (2n^2/\varepsilon)^{1/\alpha}$. Let $q$ be the largest prime less than or equal to $h^{1+\alpha}$. By Bertrand's Postulate, we have $h^{1+\alpha}/2 \leq q \leq h^{1+\alpha}$. Since we may assume $m \leq n$, we have $q \geq nmh/\varepsilon$ as required. Set $m = k$.

The function $C$ has output length

$$m \log q \leq m(1+\alpha) \log h = (1+\alpha)kt$$

as claimed. By Lemma 6.7, and Lemma 2.5, $C$ is a

$$(n, \log(q^m/\varepsilon)) \rightarrow_{2\varepsilon} ((1+\alpha)kt, \log((h^m-1)/\varepsilon) - 1)$$

condenser (using the fact that $q^m > (q^m - 1)(h-1)/(q-1)$). Now,

$$\log(q^m/\varepsilon) \leq m \log q + \log(1/\varepsilon) \leq (1+\alpha)kt + \log(1/\varepsilon).$$

And, as in the proof of Theorem 1.1, a distribution with $\log((h^m-1)/\varepsilon) - 1$ minentropy is $1/h^m < \varepsilon$ close to having minentropy $kt + \log(1/\varepsilon) - 1$. Thus $C$ is the claimed condenser. The seed length is $\log q \leq (1+\alpha) \log h = (1+\alpha)t$. □

For the Reed-Solomon-based construction, a relatively simple argument shows that the entropy rate and the ratio of output minentropy to input minentropy must both be constants less than 1. The example below comes from [GHSZ02, TSZ04]:

**Theorem 6.9.** *For every positive integer $p$ such that $p|(q-1)$, there is a source $\mathbf{X}$ with minentropy at least $\lfloor n/p \rfloor \log q$ for which $C(\mathbf{X}, \mathbf{U_t})$, as defined in Equation (2), is* not *$\varepsilon$-close to having minentropy $\log(\frac{1}{1-\varepsilon} w^m)$, where $w = (q-1)/p + 1$.*

*Proof.* Take the source to be $p$-th powers of all degree $\lfloor n/p \rfloor$ polynomials. Every output symbol of $C$ is an evaluation of such a polynomial, and therefore must be a $p$-th power, or 0. There are thus only $w = (q-1)/p + 1$ possible output symbols, so the output is contained within a set of size $w^m$, which by Proposition 2.2 is not $\varepsilon$-close to any distribution with minentropy $\log(\frac{1}{1-\varepsilon} w^m)$. □

This example can be interpreted as follows. For any $m \leq \lfloor n/p \rfloor$, we have enough entropy to hope for $C$'s output (which has length $m \log q$) to be close to uniform. However, if we choose $p = n^\delta$ for some constant $\delta > 0$, then the output minentropy can be no larger than $\log(O(w^m)) = m \log(q^{1-\delta'})$, for some constant $\delta' > 0$, as long as $q = \text{poly}(n)$ (which is required for seed length $O(\log n)$). So this setting of parameters shows that an entropy rate that is a constant less than 1 is unavoidable, and also that the output minentropy must be a constant factor smaller than the input minentropy, in this case.

# 7 Conclusions

This paper introduces a new proof technique for analyzing algebraic extractor constructions, which does not rely on local decodability of the underlying error-correcting codes. It is thus natural to ask whether these new techniques can help in other settings. For example, can we use them to argue about *computational* analogs of the objects in this paper – pseudorandom generators and pseudoentropy generators? Or, can variants of our constructions yield so-called "2-source" objects, in which both the source and the seed are only weakly random?

Of course a significant remaining open problem is to construct truly optimal extractors, ones that are optimal up to *additive* constants in the seed length and/or output length. Towards this end, we wonder if there is some variant of our constructions with a better entropy rate – the next natural threshold is to have entropy *deficiency* only $k^{o(1)}$. Another interesting question is whether some variant of these constructions can give a block-wise source directly. Depending on the actual parameters, either of these two improvements have the potential to lead to extractors with optimal output length (i.e. ones that extract all the minentropy). Alternatively, if we can find an extractor with optimal output length for high min-entropy (say $.99n$), then, by composing it with our condenser, we would get one for arbitrary min-entropy.

**Acknowledgements.** This paper began with a conversation at the BIRS workshop "Recent Advances in Computation Complexity." The authors would like to thank the organizers for inviting them, and BIRS for hosting the workshop.

## References

[AGHP92] N. Alon, O. Goldreich, J. Hastad, and R. Peralta. Simple constructions of almost $k$-wise independent random variables. *Random Structures and Algorithms*, (3):289–304, 1992.

[CRVW02] M. Capalbo, O. Reingold, S. Vadhan, and A. Wigderson. Randomness conductors and constant-degree expansion beyond the degree/2 barrier. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pages 659–668, 2002.

[CW89] A. Cohen and A. Wigderson. Dispersers, deterministic amplification, and weak random sources (extended abstract). In *Proceedings of the 30th Annual IEEE Symposium on Foundations of Computer Science*, pages 14–19, 1989.

[GHSZ02] V. Guruswami, J. Hastad, M. Sudan, and D. Zuckerman. Combinatorial bounds for list decoding. *IEEE Transactions on Information Theory*, 48(5):1021–1035, 2002.

[Gil98] D. Gillman. A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220 (electronic), 1998.

[GR06] V. Guruswami and A. Rudra. Explicit capacity-achieving list-decodable codes. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 1–10, 2006.

[GS99] V. Guruswami and M. Sudan. Improved decoding of Reed-Solomon and Algebraic-Geometry codes. *IEEE Transactions on Information Theory*, 45(6):1757–1767, 1999.

[ISW00] R. Impagliazzo, R. Shaltiel, and A. Wigderson. Extractors and pseudo-random generators with optimal seed length. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, pages 1–10, 2000.

[IZ89] R. Impagliazzo and D. Zuckerman. How to recycle random bits. In *Proceedings of the 30th Annual IEEE Symposium on Foundations of Computer Science*, pages 248–253, 1989.

[KU06] S. Kalyanaraman and C. Umans. On obtaining pseudorandomness from error-correcting codes. *Electronic Colloquium on Computational Complexity (ECCC)*, (128), 2006.

[LN86] R. Lidl and H. Niederreiter. *Introduction to Finite Fields and their applications*. Cambridge University Press, 1986.

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

[NW94] N. Nisan and A. Wigderson. Hardness vs. randomness. *Journal of Computer and System Sciences*, 49:149–167, 1994.

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

[PV05] F. Parvaresh and A. Vardy. Correcting errors beyond the Guruswami-Sudan radius in polynomial time. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 285–294, 2005.

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

[RR99] R. Raz and O. Reingold. On recycling the randomness of states in space bounded computation. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 159–168, 1999.

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

[Sha02] R. Shaltiel. Recent developments in explicit constructions of extractors. *Bulletin of the European Association for Theoretical Computer Science*, 77:67–, June 2002. Columns: Computational Complexity.

[SU05] R. Shaltiel and C. Umans. Simple extractors for all min-entropies and a new pseudorandom generator. *Journal of the ACM*, 52(2):172–216, 2005. Conference version appeared in FOCS 2001.

[Sud97] M. Sudan. Decoding of Reed Solomon codes beyond the error-correction bound. *J. Complexity*, 13(1):180–193, 1997.

[SZ99] A. Srinivasan and D. Zuckerman. Computing with very weak random sources. *SIAM Journal on Computing*, 28:1433–1459, 1999.

[Tre01] L. Trevisan. Extractors and pseudorandom generators. *Journal of the ACM*, 48(4):860–879, 2001.

[TSZ04] A. Ta-Shma and D. Zuckerman. Extractor codes. *IEEE Transactions on Information Theory*, 50(12):3015–3025, 2004.

[TU06] A. Ta-Shma and C. Umans. Better lossless condensers through derandomized curve samplers. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, 2006. To appear.

[TUZ01] A. Ta-Shma, C. Umans, and D. Zuckerman. Loss-less condensers, unbalanced expanders, and extractors. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, pages 143–152, 2001.

[TZS06] A. Ta-Shma, D. Zuckerman, and S. Safra. Extractors from Reed-Muller codes. *J. Comput. Syst. Sci.*, 72(5):786–812, 2006.

[Zuc96] D. Zuckerman. Simulating BPP using a general weak random source. *Algorithmica*, 16(4-5):367–391, 1996.

[Zuc97] D. Zuckerman. Randomness-optimal oblivious sampling. *Random Struct. Algorithms*, 11(4):345–367, 1997.

[Zuc06] D. Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 681–690, 2006.



ECCC ISSN 1433-8092

http://eccc.hpi-web.de/