Increasing the Output Length of Zero-Error Dispersers

Ariel Gabizon * Ronen Shaltiel†

January 13, 2010

Abstract

Let $\mathcal{C}$ be a class of probability distributions over a finite set Ω. A function $D : \Omega \mapsto \{0,1\}^m$ is a *disperser* for $\mathcal{C}$ with *entropy threshold* k and *error* ϵ if for any distribution X in $\mathcal{C}$ such that X gives positive probability to at least 2^k elements we have that the distribution $D(X)$ gives positive probability to at least $(1-\epsilon)2^m$ elements. A long line of research is devoted to giving explicit (that is polynomial time computable) dispersers (and related objects called "extractors") for various classes of distributions while trying to maximize m as a function of k. For several interesting classes of distributions there are explicit constructions in the literature of *zero-error* dispersers with "small" output length m.

In this paper we develop a general technique to improve the output length of zero-error dispersers. This strategy works for several classes of sources and is inspired by a transformation that improves the output length of extractors (which was given in [32] building on earlier work by [17]). Our techniques are different than those of [32] and in particular give non-trivial results in the errorless case.

Using our approach we construct improved zero-error 2-source dispersers. More precisely, we show that for any constant $\delta > 0$ there is a constant $\eta > 0$ such that for sufficiently large n there is a poly-time computable function $D : \{0,1\}^n \times \{0,1\}^n \mapsto \{0,1\}^{\eta n}$ such that for every two independent distributions X_1, X_2 over $\{0,1\}^n$ such that support at least $2^{\delta n}$ elements, the output distribution $D(X_1, X_2)$ has full support. This improves the output length of previous constructions by [2] and has applications in Ramsey Theory and in constructing certain data structures [15].

We also use our techniques to give explicit constructions of zero-error dispersers for bit-fixing sources and affine sources over polynomially large fields. These constructions improve the best known explicit constructions due to [28, 16] and achieve $m = \Omega(k)$ for bit-fixing sources and $m = k - o(k)$ for affine sources over polynomial size fields.

*Department of Computer Science, Columbia University, New York, NY and Department of Computer Science, Austin, TX . ariel.gabizon@gmail.com. Research supported by DARPA award HR0011-08-1-0069.

†Department of Computer Science, University of Haifa, Haifa, Israel, ronen@cs.haifa.ac.il. Research supported by Binational US-Israel Science Foundation (BSF) grant 2004329 and Israel Science Foundation (ISF) grant 686/07.

1 Introduction

1.1 Background

Randomness extractors and dispersers are functions that refine the randomness in "weak sources of randomness" that "contain sufficient entropy". Various variants of extractors and dispersers are closely related to expander graphs, error correcting codes and objects from Ramsey theory. A long line of research is concerned with explicit constructions of these objects and these constructions have many applications in many areas of computer science and mathematics (e.g. network design, cryptography, pseudorandomness, coding theory, hardness of approximation, algorithm design and Ramsey theory).

1.1.1 Randomness extractors and dispersers

We start with formal definitions of extractors and dispersers. (We remark that in this paper we consider the "seedless version" of extractors and dispersers).

Definition 1.1 (min-entropy and statistical distance)**.** *Let Ω be a finite set. The* min-entropy *of a distribution X on Ω is defined by $H_\infty(X) = min_{x\in\Omega} \log_2 \frac{1}{\Pr[X=x]}$. For a class $\mathcal{C}$ of distributions on Ω we use $\mathcal{C}_k$ to denote the class of all distributions $X \in \mathcal{C}$ such that $H_\infty(X) \geq k$. We say that two distributions X, Y on Ω are* ϵ-close *if $\frac{1}{2}\sum_{w\in\Omega} |\Pr[X = w] - \Pr[Y = w]| \leq \epsilon$.*

When given a class $\mathcal{C}$ of distributions (which we call "sources") the goal is to design one function that refines the randomness of any distribution X in $\mathcal{C}$. An *extractor* produces a distribution that is (close to) uniform whereas a *disperser* produces a distribution with (almost) full support. A precise definition follows:

Definition 1.2 (Extractors and Dispersers)**.** *Let $\mathcal{C}$ be a class of distributions on a finite set Ω.*

- *A function $E : \Omega \mapsto \{0,1\}^m$ is an* extractor *for $\mathcal{C}$ with* entropy threshold *k and* error *$\epsilon > 0$ if for every $X \in \mathcal{C}_k$, $E(X)$ is ϵ-close to the uniform distribution on $\{0,1\}^m$.*
- *A function $D : \Omega \mapsto \{0,1\}^m$ is a* disperser *for $\mathcal{C}$ with* entropy threshold *k and* error *$\epsilon > 0$ if for every $X \in \mathcal{C}_k$, $|\mathrm{Supp}(D(X))| \geq (1-\epsilon)2^m$ (where $\mathrm{Supp}(Z)$ denotes the support of the random variable Z).*

We remark that every extractor is in particular a disperser and that the notion of dispersers only depends on the support of the distributions in $\mathcal{C}$. A long line of research is concerned with designing extractors and dispersers for various classes of sources. For a given class $\mathcal{C}$ we are interested in designing extractors and dispersers with as small as possible entropy threshold k, as large as possible output length m and as small as possible error ϵ. (We remark that it easily follows that $m \leq k$ whenever $\epsilon < 1/2$).

It is often the case that the probabilistic method gives that a randomly chosen function E is an excellent extractor. (This is in particular true whenever the class $\mathcal{C}$ contains "not too many" sources). However, most applications of extractors and dispersers require *explicit constructions*, namely functions that can be computed in time polynomial in their input length. Much of the work done in this area can be described as an attempt of matching the parameters obtained by existential results using explicit constructions.

1.1.2 Some related work

Classes of sources. Various classes $\mathcal{C}$ of distributions were studied in the literature: The first construction of deterministic extractors can be traced back to von Neumann [35] who showed how to use many independent tosses of a biassed coin (with unknown bias) to obtain an unbiased coin. Blum [5] considered sources that are generated by a finite Markov-chain. Santha and Vazirani [30], Vazirani [30, 34], Chor and Goldreich [9], Dodis et al. [12], Barak, Impagliazzo and Wigderson [1], Barak et al. [2], Raz [29], Rao [27], Bourgain [6], Barak et al. [3], and Shaltiel [32] studied sources that are composed of several independent samples from "high entropy" distributions. Chor et al. [10], Ben-Or and Linial [4], Cohen and Wigderson [11], Mossel and Umans [24], Kamp and Zuckerman [22], Gabizon, Raz and Shaltiel [17], and Rao [28] studied bit-fixing sources which are sources in which a subset of the bits are uniformly distributed. Trevisan and Vadhan [33] and Kamp et al. [21] studied sources which are "samplable" by "efficient" procedures. Barak et al. [2], Bourgain [7], Gabizon and Raz [16], and Rao [28] studied sources which are uniform over an affine subspace. Dvir, Gabizon and Wigderson [13] studied a generalization of affine sources to sources which are sampled by low degree multivariate polynomials.

Seeded extractors and dispersers. A different variant of extractors and dispersers are *seeded* extractors and dispersers (defined by Nisan and Zuckerman [25]). Here the class $\mathcal{C}$ is the class of all distributions on $\Omega = \{0,1\}^n$. It is easy to verify that there do not exist extractors or dispersers for $\mathcal{C}$ (even when $k = n-1$, $m = 1$ and $\epsilon < 1/2$). However, if one allows the extractor (or disperser) to receive an additional independent uniformly distributed input (which is called "a seed") then extraction is possible as long as the seed is of length $\Theta(\log(n/\epsilon))$. More precisely, a seeded extractor (or disperser) with entropy threshold k and error ϵ is a function $F : \{0,1\}^n \times \{0,1\}^t \mapsto \{0,1\}^m$ such that for any distribution X on $\{0,1\}^n$ with $\mathrm{H}_\infty(X) \geq k$ the distribution $F(X,Y)$ (where Y is an independent uniformly distributed variable) satisfies the guarantees of Definition 1.2. A long line of research is concerned with explicit constructions of seeded extractors and dispersers (the reader is referred to [31] for a survey article and to [23, 20, 14] for the current milestones in explicit constructions of extractors).

1.1.3 Zero-error dispersers

In this paper we are interested in *zero-error dispersers*. These are dispersers where the output distribution has full support. That is for every source X in the class $\mathcal{C}$:

$$\{D(x) : x \in \mathrm{Supp}(X)\} = \{0,1\}^m$$

We also consider a stronger variant which we call *strongly-hitting disperser* in which every output element $z \in \{0,1\}^m$ is obtained with "not too small" probability. A precise definition follows:

Definition 1.3 (Zero-error dispersers and strongly hitting dispersers)**.** *Let $\mathcal{C}$ be a class of distributions on a finite set Ω.*

- *A function D is a* zero-error disperser *for $\mathcal{C}$ with* entropy threshold *k if it is a disperser for $\mathcal{C}$ with entropy threshold k and error $\epsilon = 0$.*
- *A function $D : \Omega \mapsto \{0,1\}^m$ is a* μ-strongly hitting disperser *for $\mathcal{C}$ with* entropy threshold *k if for every $X \in \mathcal{C}_k$ and for every $z \in \{0,1\}^m$, $\Pr[D(X) = z] \geq \mu$.*

Note that a μ-strongly hitting disperser with $\mu > 0$ is in particular a zero-error disperser and that any μ-strongly hitting disperser has $\mu \leq 2^{-m}$. The following facts immediately follow:

Fact 1.4. *Let $f : \Omega \mapsto \{0,1\}^m$ be a function and let $\epsilon \leq 2^{-(m+1)}$.*

- *If f is a disperser with error ϵ then f is a zero-error disperser (for the same class $\mathcal{C}$ and entropy threshold k).*
- *If f is an extractor with error ϵ then f is a $2^{-(m+1)}$-strongly hitting disperser (for the same class $\mathcal{C}$ and entropy threshold k).*

It follows that extractors and dispersers with small ϵ immediately translate into zero-error dispersers (as one can truncate the output length to $m' = \log(1/\epsilon) - 1$ bits and such a truncation preserves the output guarantees of extractors and dispersers).

1.2 Increasing the output length of zero-error dispersers

For several interesting classes of sources there are explicit constructions of dispersers with "large" error (which by Fact 1.4 give zero-error dispersers with "short" output length). In this paper we develop techniques to construct zero-error dispersers with large output length.

1.2.1 The composition approach

The following methodology for increasing the output length of extractors was suggested in [17, 32]: When given an extractor E' with "small" output length t (for some class $\mathcal{C}$) consider the function $E(x) = F(x, E'(x))$ where F is a seeded extractor. Shaltiel [32] (building on earlier work by Gabizon et al. [17]) shows that if E' and F fulfill certain requirements then this construction yields an extractor for $\mathcal{C}$ with large output length. The high level idea is that if certain conditions are fulfilled then the distribution $F(X, E(X))$ (in which the two inputs of F are *dependent*) is close to the distribution $F(X, Y)$ (where Y is an independent uniformly distributed variable) and note that the latter distribution is close to uniform by the definition of seeded extractors. This technique proved useful for several interesting classes of sources.

We would like to apply an analogous idea to obtain zero-error dispersers. However, by the lower bounds of [25, 26] if F is a seeded extractor (or seeded disperser) then its seed length is at least $\log(1/\epsilon)$. This means that if we want $F(X, Y)$ to output m bits with error $\epsilon < 1/2^m$ we need seed length larger than m. This in turn means that we want E' to have output length $t > m$ which makes the transformation useless.

There are also additional problems. The argument in [32] requires the "original function" E' to be an extractor (and it does not go through if E' is a disperser) and furthermore the error of the "target function" E is at least as large as that of the "original function" E' (and once again we don't gain when shooting for zero-error dispersers).

Summing up we note that if we want to improve the output length of a zero-error disperser D' by a composition of the form $D(x) = F(x, D'(x))$ we need to use a function F with different properties (a seeded extractor or disperser will not do) and we need to use a different kind of analysis.

1.2.2 Composing zero-error dispersers

In this paper we imitate the method of [32] and give a general method to increase the output length of zero-error dispersers. That is when given:

- A zero-error disperser $D' : \Omega \mapsto \{0,1\}^t$ for a class $\mathcal{C}$ and "small" output length t.

- A function $F : \Omega \times \{0,1\}^t \mapsto \{0,1\}^m$ for "large" output length m.

We identify properties of F that are sufficient so that the construction

$$D(x) = F(x, D'(x))$$

gives a zero-error disperser. (The argument is more general and transforms $2^{-(t+O(1))}$-strongly hitting dispersers into $2^{-(m+O(1))}$-strongly hitting dispersers). We then use this technique to give new constructions of zero-error dispersers and strongly-hitting dispersers.

1.3 Subsource hitters

As explained earlier we cannot choose F to be a seeded extractor. Instead, we introduce a new object which we call a *subsource hitter*. The definition of subsource hitters is somewhat technical and is tailored so that the construction $D(x) = F(x, D'(x))$ indeed produces a disperser.

Definition 1.5 (subsource hitter)**.** *A distribution X' on Ω is a* subsource *of a distribution X on Ω if there exist $\alpha > 0$ and a distribution X'' on Ω such that X can be expressed as a convex combination $X = \alpha X' + (1-\alpha)X''$.*

Let $\mathcal{C}$ be a class of distributions on Ω. A function $F : \Omega \times \{0,1\}^t \mapsto \{0,1\}^m$ is a subsource-hitter *for $\mathcal{C}$ with* entropy threshold *k and* subsource entropy *$k - v$ if for any $X \in \mathcal{C}_k$ and $z \in \{0,1\}^m$ there exists a $y \in \{0,1\}^t$ and a distribution $X' \in \mathcal{C}_{k-v}$ that is a subsource of X such that for every $x \in \mathrm{Supp}(X')$ we have that $F(x,y) = z$.*

A subsource hitter has the property that for any $z \in \{0,1\}^m$ there exist $y \in \{0,1\}^t$ and $x \in \mathrm{Supp}(X)$ such that $F(x,y) = z$ and in particular

$$\{F(x,y) : x \in \mathrm{Supp}(X), y \in \{0,1\}^t\} = \{0,1\}^m$$

In addition a subsource hitter has the stronger property that there exists a subsource X' of X (which is itself a source in $\mathcal{C}$) such that for any $z \in \{0,1\}^m$ there exists $y \in \{0,1\}^t$ such that for *any* $x \in \mathrm{Supp}(X') \subseteq \mathrm{Supp}(X)$, $F(x,y) = z$.

This property allows us to show that $D(x) = F(x, D'(x))$ is a zero-error disperser with entropy threshold k whenever D' is a zero-error disperser with entropy threshold $k - v$. This is because when given a source $X \in \mathcal{C}_k$ and $z \in \{0,1\}^m$ we can consider the seed $y \in \{0,1\}^t$ and subsource X' guaranteed in the definition. We have that D' is a zero-error disperser and that X' meets the entropy threshold of D'. It follows that there exist $x \in \mathrm{Supp}(X') \subseteq \mathrm{Supp}(X)$ such that $D'(x) = y$ and therefore

$$D(x) = F(x, D'(x)) = F(x,y) = z.$$

The precise argument is given in Section 4. In that section we also define a generalized version of subsource hitters that applies to strongly hitting dispersers.

It is interesting to note that this argument is significantly simpler than that of [32]. Indeed, the definition of subsource hitters is specifically tailored to make the composition argument go through and the more complicated task is to design subsource hitters. This is in contrast to [32] in which the function F is in most cases an "off the shelf" seeded extractor and the difficulty is to show that the composition succeeds.

1.4 Outline of the paper

In Section 2 we survey our results for specific classes of sources, provide background and compare our results to previous work. In Section 3 we define the notations used in this paper. In Section 4 we present our main composition theorem. In Section 5 we prove our results for multiple independent sources. In Section 6 we prove our results on bit-fixing sources. In Section 7 we prove our results on affine sources. Finally, in Section 8 we give some open problems.

2 An overview of our results and technique

We use the new composition technique to construct zero-error dispersers with large output length for various classes of sources. In this section we survey our results for various classes of sources. For each class, we provide a high level overview of our construction.

2.1 Zero-error 2-source dispersers

The class of 2-*sources* is the class of distributions $X = (X_1, X_2)$ on $\Omega = \{0,1\}^n \times \{0,1\}^n$ such that X_1, X_2 are independent. It is common to consider the case where each of the two distributions X_1, X_2 has min-entropy at least some threshold k. A function $f : \{0,1\}^n \times \{0,1\}^n \mapsto \{0,1\}^m$ is a 2-source extractor (resp. disperser) with entropy threshold $2 \cdot k$ and error $\epsilon \geq 0$ if for every two independent distributions X_1, X_2 on $\{0,1\}^n$ both having min-entropy at least k, $f(X_1, X_2)$ is ϵ-close to the uniform distribution on $\{0,1\}^m$ (resp. $|\mathrm{Supp}(f(X_1, X_2))| \geq (1-\epsilon)2^m$). We say that f is a zero-error disperser if it is a disperser with error $\epsilon = 0$. We say that f is a μ-strongly hitting disperser if for every X_1, X_2 as above and every $z \in \{0,1\}^m$, $\Pr[f(X_1, X_2) = z] \geq \mu$.

Background. The probabilistic method gives 2-source extractors with $m = 2 \cdot k - O(\log(1/\epsilon))$ for any $k \geq \Omega(\log n)$. However, until 2005 the best explicit constructions [9, 34] only achieved $k > n/2$. The current best extractor construction [6] achieves entropy threshold $k = (1/2 - \alpha)n$ for some constant $\alpha > 0$. Improved constructions of dispersers for entropy threshold $k = \delta n$ (for an arbitrary constant $\delta > 0$) were given in [2]. These dispersers can output any constant number of bits with zero error (and are μ-strongly hitting for some constant $\mu > 0$).[1] Subsequent work by [3] achieved entropy threshold to $k = n^{o(1)}$ and gives zero-error dispersers that output one bit.

Our results. We use our composition techniques to improve the output length in the construction of [2]. We show that:

Theorem 2.1 (2-source zero-error disperser)**.** *For every $\delta > 0$ there exists a $\nu > 0$ and $\eta > 0$ such that for sufficiently large n there is a poly(n)-time computable $(\nu 2^{-m})$-strongly hitting 2-source disperser $D : (\{0,1\}^n)^2 \mapsto \{0,1\}^m$ with entropy threshold $2 \cdot \delta n$ and $m = \eta n$.*

Note that our construction achieves an output length that is optimal up to constant factors for this entropy threshold. For lower entropy threshold our techniques gives that any explicit construction of a zero-error 2-source disperser D' with entropy threshold k and output length $t = \Omega(\log n)$ can be transformed into an explicit construction of a zero-error 2-source disperser D with entropy threshold $2 \cdot k$ and output length $m = \Omega(k)$. (See Section 5 for a precise formulation

[1] In [27] it is pointed out that by enhancing the technique of [2] using ideas from [3] and replacing some of the components used in the construction with improved components that are constructed in [27] it is possible to increase the output length and achieve a zero-error disperser with output length $m = k^{\Omega(1)}$ for the same entropy threshold k.

that also considers strongly hitting dispersers). This cannot be applied on the construction of [3] that achieves entropy threshold $k = n^{o(1)}$ as this construction only outputs one bit. Nevertheless, this means that it suffices to extend the construction of [3] so that it outputs $\Theta(\log n)$ bits in order to obtain an output length of $m = \Omega(k)$ for low entropy threshold k.

We prove Theorem 2.1 by designing a subsource hitter for 2-sources and using our composition technique. The details are given in Section 5 and a high level outline appears next.

Outline of the argument. We want to design a function $F : \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^t \mapsto \{0,1\}^m$ such that for any 2-source $X = (X_1, X_2)$ with sufficient min-entropy and for any $z \in \{0,1\}^m$ there exists a "seed" $y \in \{0,1\}^t$ and a subsource X' of X such that $X' = (X'_1, X'_2)$ is a 2-source with roughly the same min-entropy as X, and furthermore $\Pr[F(X'_1, X'_2, y) = z] = 1$. We will be shooting for $m = \Omega(n)$ and $t = O(\log n)$.

We construct the seed obtainer F using ideas from [2, 3]. Let E be a seeded extractor with seed length $t = O(\log n)$, output length $v = \Omega(k)$ and error $\epsilon_E = 1/100$ (such extractors were constructed in [23, 20]). When given inputs x_1, x_2, y we consider $r_1 = E(x_1, y)$ and $r_2 = E(x_2, y)$. By using a stronger variant of seeded extractors called "strong extractors" it follows that there exists a "good seed" $y \in \{0,1\}^t$ such that $R_1 = E(X_1, y)$ and $R_2 = E(X_2, y)$ are ϵ_E-close to uniform. We then use a 2-source extractor $H : \{0,1\}^v \times \{0,1\}^v \mapsto \{0,1\}^m$ for *very high* entropy threshold (say entropy threshold $2 \cdot 0.9v$) and very low error (say error $2^{-(m+1)}$ for output length $m = \Omega(v) = \Omega(k)$). Such extractors were constructed in [34]. Our final output is given by:

$$F(x_1, x_2, y) = H(E(x_1, y), E(x_2, y))$$

This seems strange at first sight as it is not clear why running H on inputs R_1, R_2 that are already close to uniform helps. Furthermore, the straightforward analysis only gives that $H(R_1, R_2)$ is ϵ-close to uniform for *large* error $\epsilon \geq \epsilon_E = 1/100$ and this means that the output of F may miss a large fraction of strings in $\{0,1\}^m$.

The point to notice is that both R_1, R_2 are close to uniform and therefore have large support $(1-\epsilon_E)2^v \geq 2^{0.9v}$. Using Fact 1.4 we can think of H as a zero-error disperser. Recall that dispersers are oblivious to the precise probability distribution of R_1, R_2 and it is sufficient that R_1, R_2 have large support. It follows that indeed every string $z \in \{0,1\}^m$ is hit by $H(R_1, R_2)$.

This does not suffice for our purposes as we need that any string z is hit with probability one on a subsource $X' = (X'_1, X'_2)$ of X in which the two distributions X'_1 and X'_2 are independent. For any output string $z \in \{0,1\}^m$ we consider a pair of values (r_1, r_2) for R_1, R_2 on which $H(r_1, r_2) = z$ (we have just seen that such a pair exists) and set $X'_1 = (X_1 | E(X_1, y) = r_1)$ and $X'_2 = (X_2 | E(X_2, y) = r_2)$. Note that these two distributions are indeed independent (as each depends only on one of the original distributions X_1, X_2) and that on every $x'_1 \in \mathrm{Supp}(X'_1)$ and $x'_2 \in \mathrm{Supp}(X'_2)$ we have that:

$$F(x'_1, x'_2, y) = H(E(x'_1, y), E(x'_2, y)) = H(r_1, r_2) = z$$

Furthermore, for a typical choice of (r_1, r_2) we can show that both X'_1, X'_2 have min-entropy roughly $k - v$. Thus, setting v appropriately, X' is a subsource of X with the required properties. (A more careful version of this argument can be used to preserve the "strongly hitting" property).

2.1.1 Interpretation in Ramsey Theory

A famous theorem in Ramsey Theory (see [19]) states that for sufficiently large N and any 2-coloring of the edges of the complete graph on N vertices there is an induced subgraph on $K = \Theta(\log N)$ vertices which is "monochromatic" (that is all edges are of the same color).

Zero-error 2-source dispersers (with output length $m = 1$) can be seen as providing counterexamples to this statement for larger values of K in the following way: When given a zero-error 2-source disperser $D : \{0,1\}^n \times \{0,1\}^n \mapsto \{0,1\}^m$ with entropy threshold $2 \cdot k$ we can consider coloring the edges of the full graph on $N = 2^n$ vertices with 2^m colors by coloring an edge (v_1, v_2) by $D(v_1, v_2)$. (A technicality is that $D(v_1, v_2)$ may be different than $D(v_2, v_1)$ and to avoid this problem the coloring is defined by ordering the vertices according to some order and coloring the edge (v_1, v_2) where $v_1 \leq v_2$ by $D(v_1, v_2)$). The disperser guarantee can be used to show that any induced subgraph with $K = 2^{k+1}$ vertices contains edges of *all* 2^m colors.[2]

Note that dispersers with $m > 1$ translate into colorings with more colors and that in this context of Ramsey Theory the notion of a zero-error disperser seems more natural than one that allows error. Our constructions achieve $m = \Omega(k)$ and thus the number of colors in the coloring approaches the size of the induced subgraph.

Generalizing this relation between dispersers and Ramsey theory we can view any zero-error disperser for a class $\mathcal{C}$ as a coloring of all $x \in \Omega$ such that any set S that is obtained as the support of a distribution in $\mathcal{C}$ is colored by all possible 2^m colors.

2.1.2 Rainbows and implicit $O(1)$-probe search

As we now explain, explicit constructions of zero-error 2-source dispersers can be used to construct certain data structures (this connection is due to [15]).

Consider the following problem: We are given a set $S \subseteq \{0,1\}^n$ of size 2^k. We want to store the elements of S in a table T of the same size where every entry in the table contains a single element of S (and so the only freedom is in ordering the elements of S in the table T). We say that T supports q-queries if given $x \in \{0,1\}^n$ we can determine whether $x \in S$ using q queries to T (note for example that ordered tables and binary search support $q = k$ queries). Yao [36] and Fiat and Naor [15] showed that it is impossible to achieve $q = O(1)$ when n is large enough relative to k. (This result can be seen as a kind of Ramsey Theorem).

Fiat and Naor [15] gave explicit constructions of tables that support $q = O(1)$ queries when $k = \delta \cdot n$ for any constant $\delta > 0$. This was achieved by reducing the implicit probe search problem to the task of explicitly constructing a certain combinatorial object that they call a "rainbow".

Loosely speaking a rainbow is a zero-error disperser for the class of distributions X that are composed of q independent copies of a high min-entropy distribution. We stress that for this application one needs (strongly-hitting) dispersers with large output length. More precisely, in order to support $q = O(1)$ queries one requires such dispersers that have output length m that is a *constant fraction* of the entropy threshold.

Our techniques can be used to explicitly construct rainbows which in turn allow implicit probe schemes that support $q = O(1)$ queries for smaller values of k than previously known. More precisely for any constant $\delta > 0$ and $k = n^\delta$ there is a constant q and a scheme that supports q queries. The precise details are given in Section 5.5.

2.2 Zero-error dispersers for bit-fixing sources

The class of *bit-fixing sources* is the class of distributions X on $\Omega = \{0,1\}^n$ such that there exists a set $S \subseteq [n]$ such that X_S (that is X restricted to the indices in S) is uniformly distributed and $X_{[n]\setminus S}$ is constant. Note that for such a source X, $\mathrm{H}_\infty(X) = |S|$. (We remark that these sources are

[2] In fact, Dispersers translate into a significantly stronger guarantee that discusses colorings of the edges of the complete N by N bipartite graph such that any induced K by K subgraph has all colors.

sometimes called "oblivious bit-fixing sources" to differentiate them from "non-oblivious bit-fixing sources" in which $X_{[n]\setminus S}$ is allowed to be a function of X_S).

Background. The function $Parity(x)$ (that is the exclusive-or of the bits of x) is obviously an extractor for bit-fixing sources with entropy threshold $k = 1$, error $\epsilon = 0$ and output length $m = 1$. It turns out that there are no errorless extractors for $m = 2$. More precisely, [10] showed that for $k < n/3$ there are no extractors for bit-fixing sources with $\epsilon = 0$ and $m = 2$. For larger values of k, [10] give constructions with $m > 1$ and $\epsilon = 0$. For general entropy threshold k the current best explicit construction of extractors for bit-fixing sources is due to [28] (in fact, this extractor works for a more general class of "low weight affine sources"). These extractors work for any entropy threshold $k \geq (\log n)^c$ for some constant c, and achieve output length $m = (1 - o(1))k$ for error $\epsilon = 2^{-k^{\Omega(1)}}$. Using Fact 1.4 this gives a zero-error disperser with output length $m = k^{\Omega(1)}$.

Our results. We use our composition techniques to construct zero-error dispersers for bit-fixing sources with output length $m = \Omega(k)$. We show that:

Theorem 2.2 (Zero-error disperser for bit-fixing sources)**.** *There exist $c > 1$ and $\eta > 0$ such that for sufficiently large n and $k \geq (\log n)^c$ there is a poly(n)-time computable zero-error disperser $D : \{0,1\}^n \mapsto \{0,1\}^m$ for bit-fixing sources with entropy threshold k and output length $m = \eta k$.*

Note that our construction achieves an output length that is optimal up to constant factors. We prove Theorem 2.2 by designing a subsource hitter for bit-fixing sources and using our composition technique. The details are given in Section 6 and a high level outline appears next.

Outline of the argument. Our goal is to design a subsource hitter $G : \{0,1\}^n \times \{0,1\}^t \mapsto \{0,1\}^m$ for bit-fixing sources with entropy threshold k, output length $m = \Omega(k)$ and "seed length" $t = O(\log n)$. We make use of the subsource hitter for 2-sources $F : \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^{O(\log n)} \mapsto \{0,1\}^m$ that we designed earlier. We apply it for entropy threshold $k' = k/8$ and recall that it has output length $m = \Omega(k') = \Omega(k)$.

When given a seed $y \in \{0,1\}^t$ for G we think about it as a pair of strings (y', y'') where y' is a seed for F and y'' is a seed for an explicit construction of pairwise independent variables $Z_1, \ldots, Z_n$ where for each i, Z_i takes values in $\{1,2,3\}$ (indeed there are such constructions with seed length $O(\log n)$). When given such a seed y'' we can use the values $Z_1, \ldots, Z_n$ to partition the set $[n]$ into three disjoint sets T_1, T_2, T_3 by having each index $i \in [n]$ belong to T_{Z_i}. We construct G as follows:

$$G(x, (y', y'')) = F(x_{T_1}, x_{T_2}, y')$$

In words, we use y'' to partition the given n bit string into three strings and we run F on the first two strings (padding each of them to length n) using the seed y'.

We need to show that for any bit-fixing source X of min-entropy k and for any $z \in \{0,1\}^m$ there exist a seed $y = (y', y'')$ and a subsource X' of X such that X' is a bit-fixing source with roughly the same min-entropy as X and $\Pr[G(X', (y', y'')) = z] = 1$.

We have that X is a bit-fixing source and let $S \subseteq [n]$ be the set of its "good indices". Note that $|S| \geq k$. By the "sampling properties" of pairwise independent distributions (see e.g. [18] for a survey on "averaging samplers") it follows that there exists a y'' such that for every $i \in [3]$, $|S \cap T_i| \geq k/8$. It follows that $X_{T_1}, X_{T_2}, X_{T_3}$ are bit-fixing sources with min-entropy at least $k/8$ (and note that these three distributions are independent). Thus, by the properties of the subsource hitter F there exist x_1, x_2, y' such that $F(x_1, x_2, y') = z$ (note that here we're only using the

property that F "hits z" and do not use the stronger property that F "hits z on a subsource"). Consider the distribution

$$X' = (X | X_{T_1} = x_1 \wedge X_{T_2} = x_2)$$

This is a subsource of X which is a bit-fixing source with min-entropy at least $k/8$ (as we have not fixed the $k/8$ good bits in T_3). It follows that for every $x \in \mathrm{Supp}(X')$

$$G(x, (y', y'')) = F(x_1, x_2, y') = z$$

and G is indeed a subsource hitter for bit-fixing sources.

2.3 Zero-error dispersers for affine sources

The class of *affine sources* is the class of distributions X on $\Omega = \mathbb{F}_q^n$ (where $\mathbb{F}_q$ is the finite field of q elements) such that X is uniformly distributed over an affine subspace V in $\mathbb{F}_q^n$. Note that such a source X has min-entropy $\log q \cdot dim(V)$. Furthermore, any bit-fixing source is an affine source over $\mathbb{F}_2$.

Background. For $\mathbb{F}_2$ the best explicit construction of extractors for affine sources was given in [7]. This construction works for entropy threshold $k = \delta n$ (for any fixed $\delta > 0$) and achieves output length $m = \Omega(k)$ with error $\epsilon < 2^{-m}$.

Extractors for lower entropy thresholds were given by [16] in the case that $q = n^{\Theta(1)}$. For any entropy threshold $k > \log q$ these extractors can output $m = (1 - o(1))k$ bits with error $\epsilon = n^{-\Theta(1)}$. Using Fact 1.4 this gives zero-error dispersers with output length $m = \Theta(\log n)$.

Our results. Our composition techniques can be applied on affine sources. We focus on the case of polynomial size fields (as in that case we can improve the results of [16]). We prove the following theorem:

Theorem 2.3. *Fix any prime power q and integers n, k such that $q \geq n^{18}$ and $2 \leq k < n$. There is a poly$(n, \log q)$-time computable zero-error disperser $D : \mathbb{F}_q^n \mapsto \{0,1\}^m$ for affine sources with entropy threshold $k \cdot \log q$ and $m = (k-1) \cdot \log q$.*

Outline of the argument. We use our composition techniques to give a different analysis of the construction of [16] which shows that this construction also gives a zero-error disperser. The construction of [16] works by first constructing an affine source extractor D' with small output length $m = \Theta(\log n)$ and then composing it with some function F to obtain an extractor $D(x) = F(x, D'(x))$ that extracts many bits (with rather large error). We observe that the function F designed in [16] is in fact a subsource hitter for affine sources and therefore our composition technique gives that the final construction is a zero-error disperser.

3 Preliminaries

In this section we explain the notation used in this paper. Note that some definitions from the earlier sections are repeated in more precise form.

General Notation: We use $[n]$ to denote the set $\{1, \ldots, n\}$. We use $\mathrm{P}(S)$ to denote the set of subsets of a given set S. Given a string $x \in \{0,1\}^n$ and a set $S \subseteq [n]$ we use x_S to denote the string obtained by restricting x to the indices in S. We denote the length of a string x by $|x|$. Logarithms will always be taken with base 2.

Asymptotic Conventions: When stating formal statements in theorems and lemmas we use the Ω and O signs only to denote *absolute* constants, i.e., not depending on any parameters even if these parameters are considered constants.

Notation for probability distributions: Let Ω be some finite set and let P be a distribution on Ω. (All the probability distributions considered in this paper are on finite sets). For $B \subseteq \Omega$, we denote the probability of B according to P by $\Pr_P[B]$ or using "random variable notation" by $\Pr[P \subseteq B]$. Given a function $A : \Omega \to U$, we denote by $A(P)$ the distribution induced on U when sampling t according to P and calculating $A(t)$. We denote by U_Ω the uniform distribution on Ω. For an integer n, we denote by U_n the uniform distribution on $\{0,1\}^n$. For a distribution P on Ω^d and $j \in [d]$, we denote by P_j the restriction of P to the j'th coordinate. We denote by $\mathrm{Supp}(P)$ the support of P. A distribution P is *flat* if it P assigns the same probability to all the elements in $\mathrm{Supp}(P)$.

The *statistical distance* between two distributions P and Q on Ω, is defined as

$$\max_{S \subseteq \Omega} |\Pr_P[S] - \Pr_Q[S]| = \frac{1}{2} \sum_{w \in \Omega} |\Pr_P[w] - \Pr_Q[w]| .$$

We say that P is ϵ-*close* to Q, if the statistical distance between P and Q is at most ϵ.

Definition 3.1 (Conditional distributions)**.** *Let P be a distribution on Ω. Let $C \subseteq \Omega$ be an event such that $\Pr_P(C) > 0$. We define the distribution $(P|C)$ by*

$$\Pr_{(P|C)}[B] = \frac{\Pr_P[B \cap C]}{\Pr_P[C]}$$

for any $B \subseteq \Omega$. Given a function $A : \Omega \to U$, we denote by $(A(P)|C)$ the distribution $A((P|C))$.

We need the notion of a convex combination of distributions.

Definition 3.2 (Convex combination of distributions)**.** *Given distributions $P_1, \ldots, P_t$ on a set Ω and coefficients $\mu_1, \ldots, \mu_t \geq 0$ such that $\sum_{i=1}^t \mu_i = 1$, we define the distribution $P \triangleq \sum_{i=1}^t \mu_i \cdot P_i$ by*

$$\Pr_P[B] = \sum_{i=1}^{t} \mu_i \cdot \Pr_{P_i}[B]$$

for any $B \subseteq \Omega$.

min-entropy. The *min-entropy* of a distribution X on Ω is defined as

$$\mathrm{H}_\infty(X) \triangleq \min_{x \in \Omega} \log_2 \frac{1}{\Pr[X = x]}.$$

For a class of distributions $\mathcal{C}$ on Ω, we denote by $\mathcal{C}_k$ the set of distributions in $\mathcal{C}$ that have min-entropy at least k. We need the following standard fact:

Fact 3.3. *Let $k' \geq k$ and let X be a distribution with min-entropy at least k' then X is a convex combination of flat distributions with min-entropy exactly k.*

We also need the following standard lemma.

Lemma 3.4. *Let X be a distribution on Ω that is ϵ-close to a distribution with min-entropy k. Let $B = \{x \in \Omega : \Pr[X = x] \geq 2^{-(k-1)}\}$ then $\Pr[X \in B] \leq 2\epsilon$.*

Subsources We make use of the following notion of subsources.

Definition 3.5. *Let X be a distribution on a set Ω. A distribution X' on Ω is a subsource of X with* measure *δ if $X = \delta \cdot X' + (1-\delta) \cdot X''$ for some $\delta > 0$ and distribution X''. If X' is a subsource of X with measure $\delta \geq 2^{-v} > 0$ we say that X' is a subsource of X with* deficiency *v.*

We remark that this definition is more general than the one considered in [2, 3]. We use it as it is more convenient in this paper.[3]
We also need the following easy lemma:

Lemma 3.6. *Let X be a distribution on Ω such that $H_\infty(X) \geq k$ and let X' be a subsource of X with deficiency v then $H_\infty(X') \geq k - v$.*

Proof. We know that $X = \delta \cdot X' + (1-\delta) \cdot X''$ for some $\delta \geq 2^{-v} > 0$. Thus, for any $x \in \mathrm{Supp}(X')$

$$2^{-k} \geq \Pr[X = x] \geq 2^{-v} \cdot \Pr[X' = x] \Rightarrow \Pr[X' = x] \geq 2^{-(k-v)}.$$

Thus, $\mathrm{H}_\infty(X') \geq k - v$. □

Extractors, Dispersers and related objects:

Definition 3.7 (Extractors and dispersers)**.** *Let $\mathcal{C}$ be a class of distributions on Ω.*

- *A function $E : \Omega \mapsto \{0,1\}^m$ is an* extractor *for $\mathcal{C}$ with* entropy threshold *k and* error *$\epsilon > 0$ if for every $X \in \mathcal{C}_k$, $E(X)$ is ϵ-close to U_m.*
- *A function $D : \Omega \mapsto \{0,1\}^m$ is a* disperser *for $\mathcal{C}$ with* entropy threshold *k and* error *$\epsilon > 0$ if for every $X \in \mathcal{C}_k$, $|\mathrm{Supp}(D(X))| \geq (1-\epsilon)2^m$.*
- *A disperser D for $\mathcal{C}$ with entropy threshold k is a* zero-error disperser *with* entropy threshold *k if it has error $\epsilon = 0$.*
- *A function $D : \Omega \mapsto \{0,1\}^m$ is a* μ-strongly hitting disperser *for $\mathcal{C}$ with* entropy threshold *k if for every $X \in \mathcal{C}_k$ and for every $z \in \{0,1\}^m$, $\Pr[D(X) = z] \geq \mu$.*

We now observe that all the objects above allow the source X to be a convex combination of distributions in $\mathcal{C}$:

Fact 3.8. *Let $\mathcal{C}$ be a class of distributions on Ω. Let X be a distribution on Ω that is a convex combination of distributions from $\mathcal{C}_k$. Let f be an extractor/disperser/strongly hitting disperser with entropy threshold k. Applying f on X gives the same output guarantee as applying f on distributions in $\mathcal{C}_k$.*

[3] The definition in [2, 3] makes the additional requirement that there exists a function $f : \Omega \mapsto \{0,1\}$ such that $X' = (X | f(X) = 1)$.

Seeded extractors, dispersers and condensers. We use the following definition of seeded objects.

Definition 3.9 (seeded objects)**.**

- *A function* $E : \{0,1\}^n \times \{0,1\}^t \mapsto \{0,1\}^m$ *is a* strong seeded extractor *with* entropy threshold k *and* error ϵ *if for every distribution* X *on* $\{0,1\}^n$ *with* $H_\infty(X) \geq k$*, a* $(1-\epsilon)$ *fraction of* $y \in \{0,1\}^t$ *have that* $E(X,y)$ *is* ϵ*-close to uniform.*
- *A function* $D : \{0,1\}^n \times \{0,1\}^t \mapsto \{0,1\}^m$ *is a* seeded disperser *with* entropy threshold k *and* error ϵ *if for every distribution* X *on* $\{0,1\}^n$ *with* $H_\infty(X) \geq k$*,* $|\{D(x,y) : x \in \mathrm{Supp}(X), y \in \{0,1\}^t\}| \geq (1-\epsilon)2^m$*.*
- *A function* $C : \{0,1\}^n \times \{0,1\}^t \mapsto \{0,1\}^m$ *is a* strong seeded condenser *with* entropy threshold k*,* entropy guarantee k' *and* error ϵ *if for every distribution* X *on* $\{0,1\}^n$ *with* $H_\infty(X) \geq k$*, a* $(1-\epsilon)$ *fraction of* $y \in \{0,1\}^t$ *have that* $C(X,y)$ *is* ϵ*-close to some distribution with min-entropy* k'*.*

4 A Composition Theorem

In this section we present a general method for increasing the output length of zero-error dispersers. This is achieved by composing a zero-error disperser with a type of seeded function we call a *subsource hitter*. Our composition applies to both zero-error dispersers and strongly hitting dispersers. We start with the case of zero-error dispersers.

4.1 Zero-error dispersers

The key component in our composition theorem is the following new object which we call a "subsource hitter". In the next definition we rephrase Definition 1.5.

Definition 4.1 (Subsource hitters)**.** *Let* $\mathcal{C}$ *be a class of distributions on* Ω*. A function* $F : \Omega \times \{0,1\}^t \mapsto \{0,1\}^m$ *is a* subsource hitter *for* $\mathcal{C}$ *with* entropy threshold k *and* subsource entropy $k-v$ *if for every* $X \in \mathcal{C}_k$ *and every* $z \in \{0,1\}^m$ *there exists a* $y \in \{0,1\}^t$ *and a subsource* X' *of* X *such that* $X' \in \mathcal{C}_{k-v}$ *and* $\Pr[F(X',y) = z] = 1$*.*

The following theorem shows that subsource hitters are tailored to increase the output length of zero-error dispersers.

Theorem 4.2. *Let* $\mathcal{C}$ *be a class of distributions on* Ω*.*

- *Let* $D' : \Omega \mapsto \{0,1\}^t$ *be a zero-error disperser for* $\mathcal{C}$ *with entropy threshold* $k-v$*.*
- *Let* $F : \Omega \times \{0,1\}^t \mapsto \{0,1\}^m$ *be a subsource hitter with entropy threshold* k *and subsource entropy* $k-v$*.*

Define $D : \Omega \mapsto \{0,1\}^m$ *by* $D(x) \triangleq F(x, D'(x))$*. Then* D *is a zero-error disperser for* $\mathcal{C}$ *with entropy threshold* k*.*

Proof. Let X be a distribution in $\mathcal{C}_k$ and $z \in \{0,1\}^m$. By the guarantee on F we have that there exists a $y \in \{0,1\}^t$ and a subsource X' of X such that $\Pr[F(X',y) = z] = 1$ and $X' \in \mathcal{C}_{k-v}$. Note

that X' meets the entropy threshold of D' and therefore there exists $x \in \mathrm{Supp}(X') \subseteq \mathrm{Supp}(X)$ such that $D'(x) = y$. It follows that

$$D(x) = F(x, D'(x)) = F(x, y) = z$$

□

4.2 Strongly hitting dispersers

In this Section we generalize the composition argument so that it preserves the strongly hitting property. We start by generalizing the notion of subsource hitters:

Definition 4.3 (Generalized subsource hitters)**.** *Let $\mathcal{C}$ be a class of distributions on Ω. A function $F : \Omega \times \{0,1\}^t \mapsto \{0,1\}^m$ is a* generalized subsource hitter *for $\mathcal{C}$ with* entropy threshold k*,* subsource entropy $k - v$*,* measure α *and* error ϵ *if for every $X \in \mathcal{C}_k$ and $z \in \{0,1\}^m$ at least a $1 - \epsilon$ fraction of $y \in \{0,1\}^t$ have the property that there exists a subsource X' of X of measure α such that X' is a convex combination of distributions in $\mathcal{C}_{k-v}$ and* $\Pr[F(X', y) = z] = 1$.

The generalized version differs from the original version in two respects:

- We require that there are *many* seeds y that hit z rather than requiring that there exists *one* seed y that hits z.
- We allow X' to be a convex combination of sources in $\mathcal{C}_{k-v}$ rather than requiring that X' itself is in $\mathcal{C}_{k-v}$. This allows X' to have larger measure in the original source X.

Note that any generalized subsource hitter is a subsource hitter with the same entropy threshold and subsource entropy. (This is because we can replace the subsource X' with one of the components in the convex combination and this component is a subsource of X that meets the requirements of Definition 4.1). The following theorem is analogous to Theorem 4.2 for the case of strongly hitting dispersers.

Theorem 4.4. *Let $\mathcal{C}$ be a class of distributions on Ω.*

- *Let $D' : \Omega \mapsto \{0,1\}^t$ be a μ-strongly hitting disperser for $\mathcal{C}$ with entropy threshold $k - v$.*
- *Let $F : \Omega \times \{0,1\}^t \mapsto \{0,1\}^m$ be a subsource hitter with entropy threshold k, subsource entropy $k - v$, measure α and error ϵ.*

Define $D : \Omega \mapsto \{0,1\}^m$ by $D(x) \triangleq F(x, D'(x))$. Then D is a $((1-\epsilon)2^t\alpha\mu)$-strongly hitting disperser for $\mathcal{C}$ with entropy threshold k.

Before proving the theorem, let us discuss some of the parameters. Note that any μ-strongly hitting disperser with output length m has $\mu \leq 2^{-m}$. Let us suppose that D' that has output length t comes close to this bound (say that D' is μ-strongly hitting for $\mu = 2^{-t-O(1)}$). If F is also close to optimal in the sense that it has measure close to 2^{-m} (say $\alpha = 2^{-m-O(1)}$) then the "new disperser" D is ν-strongly hitting for $\nu = ((1-\epsilon)2^t\alpha\mu) = 2^{-m-O(1)}$. This means that when composing a "near optimal" strongly hitting disperser using a "near optimal" generalized subsource hitter we indeed obtain a "near optimal" strongly hitting disperser with large output length. We now give the proof of the theorem.

Proof. (of Theorem 4.4) Let X be a distribution in $\mathcal{C}_k$ and $z \in \{0,1\}^m$. By the guarantee on F we have that there exists a set $G \subseteq \{0,1\}^t$ of size $(1-\epsilon)2^t$ such that for every $y \in \{0,1\}^t$ there exists a subsource X'_y of X with measure α such that $\Pr[F(X'_y, y) = z] = 1$ and X'_y is a convex combination of distributions from $\mathcal{C}_{k-v}$. For every $y \in G$ we consider applying D' on X'_y (note that X'_y is a convex combination of distributions in $\mathcal{C}_{k-v}$ which meet the entropy threshold of D'). By Fact 3.8 we have that $\Pr[D'(X'_y) = y] \geq \mu$. Let

$$E_y = \{x : D'(x) = y \wedge F(x, y) = z\}.$$

We can rephrase the former statement and conclude that for every $y \in G$, $\Pr[X'_y \in E_y] \geq \mu$.

Note that for $x \in E_y$ we have that $D(x) = z$. Summing up we have that:

$$\begin{aligned}
\Pr[D(X) = z] &\geq \textstyle\sum_{y \in G} \Pr[D(X) = z | X \in E_y] \cdot \Pr[X \in E_y] \\
&= \textstyle\sum_{y \in G} \Pr[X \in E_y] \\
&\geq \textstyle\sum_{y \in G} \alpha \cdot \Pr[X'_y \in E_y] \\
&\geq \textstyle\sum_{y \in G} \alpha \cdot \mu \\
&\geq (1-\epsilon) \cdot 2^t \cdot \alpha \cdot \mu.
\end{aligned}$$

□

5 Zero-error dispersers for multiple independent sources

In this section we apply our composition techniques for the class of "multiple independent sources".

5.1 Formal definition of multiple independent sources

We now give a formal definition of the class of "multiple independent sources". We consider sources that are composed of ℓ independent high min-entropy distributions. We use the following notation.

Definition 5.1 (ℓ-sources)**.** *A distribution $X = (X_1, \ldots, X_\ell)$ on $\Omega = (\{0,1\}^n)^\ell$ is an ℓ-source if the ℓ distributions $X_1, \ldots, X_\ell$ are independent. An ℓ-source X is a* balanced-ℓ-source *if*

$$H_\infty(X_1) = H_\infty(X_2) = \ldots = H_\infty(X_\ell).$$

We say that an ℓ-source X has block-entropy *at least k if for every $1 \leq i \leq \ell$, $H_\infty(X_i) \geq k$. We say that an ℓ-source X has* block entropy *exactly k if for every $1 \leq i \leq \ell$, $H_\infty(X_i) = k$.*

Note that a balanced-ℓ-source X has min-entropy $k \cdot \ell$ if and only if X has block entropy exactly k. The following lemma is an immediate corollary of Fact 3.3

Lemma 5.2. *Every ℓ-source X with block entropy at least k is a convex combination of ℓ-sources with block entropy exactly k.*

By Fact 3.8 we can restrict our attention to designing dispersers for ℓ-sources with block entropy exactly k (or equivalently to balanced-ℓ-sources with min-entropy $\ell \cdot k$) and these dispersers can also be applied on ℓ-sources with block entropy at least k.

Remark 5.3. *Definition 5.1 uses a less standard terminology that is in particular slightly different than that used in Section 2. We use this terminology to capture ℓ-sources within our general framework of sources. The terminology above allows us to work with one "entropy threshold" parameter k, whereas the standard terminology requires ℓ thresholds (one for each block). For example, the general notion of disperser for balanced-ℓ-sources with entropy threshold $\ell \cdot k$ now coincides with the more standard notion of ℓ-source disperser that requires threshold k in each one of its blocks.*

5.2 A subsource hitter for 2-sources

In this section we construct a subsource hitter for balanced-2-sources. We make use of the "Hadamard extractor" constructed by [34, 9] (see also [12]).

Theorem 5.4. *There exists a constant $c_0 > 0$ such that for sufficiently large p there is an a poly(p)-time computable extractor $H : (\{0,1\}^p)^2 \mapsto \{0,1\}^m$ for balanced-2-sources with entropy threshold $2 \cdot 0.8p$ and error 2^{-2m} for $m = c_0 p$.*

Our construction of subsource hitters also uses a strong seeded condenser (see Definition 3.9). For different settings of parameters we use different choices of off the shelf condensers. We elaborate on these choices later on. We now present our construction.

Theorem 5.5. *Let c_0 be the constant from Theorem 5.4. Let n, k, p be integers such that $n \geq k \geq p \geq 100/c_0$, and let $m = c_0 p$.*

- *Let $C : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^p$ be a strong condenser with entropy threshold k, entropy guarantee $0.9p$ and error $1/100$.*
- *Let $H : (\{0,1\}^p)^2 \to \{0,1\}^m$ be the 2-source extractor from Theorem 5.4. (This extractor has entropy threshold $2 \cdot 0.8p$ and error 2^{-2m}).*

Define the function $F : (\{0,1\}^n)^2 \times \{0,1\}^t \to \{0,1\}^m$ by $F(x,y) = H(C(x_1,y), C(x_2,y))$ then

- *F is a subsource hitter for balanced-2-sources with entropy threshold $2 \cdot k$ and subsource entropy $2 \cdot (k - 3p)$.*
- *F is a generalized subsource hitter for balanced-2-sources with entropy threshold $2 \cdot k$, subsource entropy $2 \cdot (k - 3p)$, measure $2^{-(m+1)}$ and error $1/10$.*

Proof. Let X be a balanced-2-source on $(\{0,1\}^n)^2$ with min-entropy at least $2k$. Note that this means that X_1, X_2 are independent distributions with min-entropy k. We have that C is a strong condenser with this entropy threshold and therefore for any distribution V with min-entropy k a $(1 - 1/100)$ fraction of $y \in \{0,1\}^t$ are good in the sense that $C(V, y)$ is $(1/100)$-close to having min-entropy $0.9p$. By a union bound it follows that a $1 - 2/100$ fraction of $y \in \{0,1\}^t$ satisfy this property for both X_1, X_2 simultaneously, namely that: both $C(X_1, y)$ and $C(X_2, y)$ are $(1/100)$-close to having min-entropy $0.9p$. We call such $y \in \{0,1\}^t$ "good seeds". Fix some good seed y and let $R_1 = C(X_1, y)$ and $R_2 = C(X_2, y)$. We define:

$$B_1' = \{r \in \{0,1\}^p : \Pr[R_1 = r] < 2^{-(p+10)}\}$$

Note that:

$$\Pr[R_1 \in B_1'] \leq \sum_{r \in B_1'} \Pr[R_1 = r] \leq 2^p \cdot 2^{-(p+10)} \leq 2^{-10}.$$

We define:

$$B_1'' = \{r \in \{0,1\}^p : \Pr[R_1 = r] > 2^{-(0.9p-1)}\}$$

By Lemma 3.4 we have that $\Pr[R_1 \in B_1''] \leq 2/100$. Let $B_1 = B_1' \cup B_1''$ and note that $\Pr[R_1 \in B_1] \leq 2/100 + 2^{-10} \leq 1/10$.

We can repeat the same argument for R_2 and define subsets B_2, B_2', B_2'' in an analogous way and conclude that $\Pr[R_2 \in B_2] \leq 1/10$. Let us consider the events $E_1 = \{R_1 \notin B_1\}$, $E_2 = \{R_2 \notin B_2\}$ and $E = E_1 \cap E_2$ (note that we think about these events as events over the original distribution X). Let $V = (X|E)$. Note that V_1 is the distribution $(X_1|E_1)$ and V_2 is the distribution $(X_2|E_2)$. Moreover, the two distributions V_1, V_2 are independent. Let us estimate the min-entropy of the distribution $C(V_1, y)$: For any r in the support of $C(V_1, y)$ we have that:

$$\Pr[C(V_1,y) = r] = \Pr[C(X_1,y) = r|E_1] = \Pr[R_1 = r|E_1] \leq \frac{\Pr[R_1 = r]}{\Pr[E_1]} \leq \frac{2^{-(0.9p-1)}}{9/10} \leq 2^{-(0.9p-2)}.$$

Thus, we conclude that $C(V_1, y)$ has min-entropy at least $0.9p - 2 \geq 0.8p$. We can do the same argument for $C(V_2, y)$. We have that the two distributions $C(V_1, y), C(V_2, y)$ are independent and meet the entropy threshold of the extractor H. We conclude that $H(C(V_1, y), C(V_2, y))$ is 2^{-2m}-close to uniform. Fix some string $z \in \{0,1\}^m$. It follows that:

$$\Pr[H(C(V_1,y), C(V_2,y)) = z] \geq 2^{-m} - 2^{-2m}$$

It follows that:

$$\begin{aligned}
\Pr[E \wedge H(R_1,R_2) = z] &= \Pr[E] \cdot \Pr[H(R_1,R_2) = z|E] \\
&= \Pr[E_1] \cdot \Pr[E_2] \cdot \Pr[H(C(V_1,y), C(V_2,y)) = z] \\
&\geq (9/10)^2 \cdot (2^{-m} - 2^{-2m}) \\
&\geq 2^{-(m+1)}
\end{aligned}$$

We say that a pair $(r_1, r_2) \in (\{0,1\}^p)^2$ is *useful* (with respect to a good seed $y \in \{0,1\}^t$ and a $z \in \{0,1\}^m$) if $r_1 \notin B_1$, $r_2 \notin B_2$ and $H(r_1, r_2) = z$. Summing up what we did so far, we have that a $(1 - 2/100)$-fraction of $y \in \{0,1\}^t$ are good seeds and for any such good seed $y \in \{0,1\}^t$ and $z \in \{0,1\}^m$ we have that with probability $2^{-(m+1)}$ the pair $(C(X_1, y), C(X_2, y))$ is useful.
For any useful pair (r_1, r_2) we define a subsource $X^{(r_1,r_2)}$ of X by

$$X^{(r_1,r_2)} = (X|C(X_1,y) = r_1 \wedge C(X_2,y) = r_2)$$

We claim that:

Claim 5.6. *For every $(r_1, r_2) \in (\{0,1\}^p)^2$ that are useful with respect to a good seed y and $z \in \{0,1\}^m$ we have that:*

- $\Pr[F(X^{(r_1,r_2)}, y) = z] = 1$.
- *$X^{(r_1,r_2)}$ is a convex combination of balanced-2-sources with min-entropy exactly $2 \cdot (k - 3p)$.*

Proof. (of Claim 5.6) The first item follows because for every $x \in \mathrm{Supp}(X^{(r_1,r_2)})$ we have that:

$$F(x,y) = H(C(x_1,y), C(x_2,y)) = H(r_1,r_2) = z$$

For the second item, note that the two distributions $X_1^{(r_1,r_2)}$ and $X_2^{(r_1,r_2)}$ are independent. Furthermore:

$$\Pr[C(X_1,y) = r_1 \wedge C(X_2,y) = r_2] = \Pr[C(X_1,y) = r_1] \cdot \Pr[C(X_2,y) = r_2] \geq (2^{-(p+10)})^2 \geq 2^{-3p}$$

It follows that $X^{(r_1,r_2)}$ is a deficiency $3p$ subsource of X. By Lemma 3.6 we have that $\mathrm{H}_\infty\left(X^{(r_1,r_2)}\right) \geq 2k - 3p$. It follows that $X^{(r_1,r_2)}$ has block-entropy at least $k - 3p$ and by Lemma 5.2 it is a convex combination of balanced-2-sources with block entropy exactly $k - 3p$ (or equivalently balanced-2-sources with min-entropy $2 \cdot (k - 3p)$). □

We are now ready to prove Theorem 5.5.

Let us first prove the first item that says that F is a subsource hitter. Fix some good seed $y \in \{0,1\}^t$ and $z \in \{0,1\}^m$. Let (r_1, r_2) be a useful pair with respect to y and z. By the first item of Claim 5.6 we have that $X^{(r_1,r_2)}$ is a convex combination of balanced-2-source with min-entropy exactly $2 \cdot (k - 3p)$. Let X' be one of the components in this convex combination that appears with a positive coefficient. We have that X' is a subsource of $X^{(r_1,r_2)}$ which is in turn a subsource of X. Furthermore by the second item of Claim 5.6 and as $\mathrm{Supp}(X') \subseteq \mathrm{Supp}(X^{(r_1,r_2)})$ we have that $\Pr[F(X',y) = z] = 1$.

We now prove the second item. That is that F is a generalized subsource hitter. We have that a $(1 - 1/10)$ fraction of $y \in \{0,1\}^t$ are good seeds. Fix some good seed y and $z \in \{0,1\}^m$. We define:

$$X' = (X | (C(X_1,y), C(X_2,y)) \text{ are a useful pair})$$

We have already seen before that that X' has measure $2^{-(m+1)}$ as a subsource of X. Furthermore, X' is a convex combination of the sources $X^{(r_1,r_2)}$ for useful pairs (r_1, r_2). By Claim 5.6 each one of the latter sources is a convex combination of balanced-2-sources with min-entropy $2 \cdot (k - 3p)$. Thus, overall X' is a convex combination of balanced-2-sources with min-entropy $2 \cdot (k - 3p)$. For every $x \in \mathrm{Supp}(X')$ there exists a useful pair (r_1, r_2) such that $x \in \mathrm{Supp}(X^{r_1,r_2})$ and we already showed that for such x we have that $F(x,y) = z$. □

5.3 Zero-error dispersers for 2-sources

We now plug in specific choices of strong seeded condensers to obtain specific results.

5.3.1 High entropy threshold

Our first choice is a condenser by Raz [29]. This condenser has the advantage that it has a constant length seed. However it only works when the entropy threshold is a constant fraction of the length.

Theorem 5.7. *[29] For every $\delta > 0$ there is a $\beta > 0$ and integer t such that for sufficiently large n there is a poly(n)-time computable strong seeded condenser $C : \{0,1\}^n \times \{0,1\}^t \mapsto \{0,1\}^p$ with $p = \beta n$ entropy threshold δn, entropy guarantee $0.9p$ and error $1/100$.*

Plugging in Theorem 5.7 in Theorem 5.5 we obtain the following Corollary.

Corollary 5.8. *For every $\delta > 0$ there is an $\eta > 0$ and an integer t such that for sufficiently large n and $m = \eta n$:*

- *There is a poly(n)-time computable generalized subsource hitter* $F : (\{0,1\}^n)^2 \times \{0,1\}^t \mapsto \{0,1\}^m$ *for balanced-2-sources with entropy threshold* $2 \cdot \delta n$*, subsource entropy* δn*, measure* $2^{-(m+1)}$ *and error* $1/10$*.*

- *Any poly(n)-time computable* μ*-strongly hitting disperser* $D' : (\{0,1\}^n)^2 \mapsto \{0,1\}^t$ *for balanced-2-sources with entropy threshold* δn *can be transformed into a poly(n)-time computable* $(\mu 2^{t-m-2})$*-strongly hitting disperser* $D : (\{0,1\}^n)^2 \mapsto \{0,1\}^m$ *for balanced-2-sources with entropy threshold* $2\delta n$*.*

We can apply the second item in the Corollary above on the strongly hitting disperser of Barak et al. [2].

Theorem 5.9. *[2] For every* $\delta > 0$ *and integer* t *there exists a* $\mu > 0$ *such that for sufficiently large* n *there is a poly(n)-time computable* μ*-strongly hitting disperser* $D : (\{0,1\}^n)^2 \mapsto \{0,1\}^t$ *with entropy threshold* δn*.*

Applying the aforementioned transformation we get the following Theorem which implies Theorem 2.1 as a special case.

Theorem 5.10. *For every* $\delta > 0$ *there exists a* $\nu > 0$ *and* $\eta > 0$ *such that for sufficiently large* n *there is a poly(n)-time computable* $(\nu 2^{-m})$*-strongly hitting disperser* $D : (\{0,1\}^n)^2 \mapsto \{0,1\}^m$ *with entropy threshold* δn *and* $m = \eta n$*.*

5.3.2 Arbitrary entropy threshold

In order to handle lower entropy thresholds we use a strong seeded extractor (which is in particular a strong seeded condenser).

Theorem 5.11. *[23, 20] There exists a number* c *such that for every sufficiently large* k, n *there is a poly(n)-time computable strong seeded extractor* $E : \{0,1\}^n \times \{0,1\}^{c \log n} \mapsto \{0,1\}^m$ *for entropy threshold* k*, error* $1/100$ *and* $m = k/2$*.*

Plugging in Theorem 5.11 in Theorem 5.5 we obtain the following Corollary.

Corollary 5.12. *There exist* $\eta > 0$ *and* c *such that for every sufficiently large* k, n *and* $m = \eta k$*:*

- *There is a poly(n)-time computable generalized subsource hitter* $F : (\{0,1\}^n)^2 \times \{0,1\}^{t=c \log n} \mapsto \{0,1\}^m$ *for balanced-2-sources with entropy threshold* $2 \cdot k$*, subsource entropy* k*, measure* $2^{-(m+1)}$ *and error* $1/10$*.*

- *Any poly(n)-time computable* μ*-strongly hitting disperser* $D' : (\{0,1\}^n)^2 \mapsto \{0,1\}^{c \log n}$ *for balanced-2-sources with entropy threshold* k *can be transformed into a poly(n)-time computable* $(\mu 2^{t-m-2})$*-strongly hitting disperser* $D : (\{0,1\}^n)^2 \mapsto \{0,1\}^m$ *for balanced-2-sources with entropy threshold* $2 \cdot k$*.*

Barak et al. [3] construct zero-error dispersers for entropy threshold $k = n^{o(1)}$. One can hope to apply Corollary 5.12 to increase the output length of these dispersers. However, the construction of [3] only achieves output length $m = 1$. We note that by Corollary 5.12 improving the output length to $m = c \log n$ will immediately give further improvement to $m = \Omega(k)$.

5.4 Zero-error dispersers for $O(1)$-sources

In the previous section we constructed zero-error dispersers for balanced-2-sources with entropy threshold $k = \delta n$ for any constant $\delta > 0$. We now give constructions that has the disadvantage that they require $\ell > 2$ sources for $\ell = O(1)$. However, they achieve lower entropy thresholds.

We use an ℓ-source extractor constructed by Rao [27]. The version we use here has better analysis that provides low error and is due to Barak et al. [3].

Theorem 5.13. *[27, 3] There is a $\gamma > 0$ such that for every sufficiently large $k \leq n$ there are integers $\ell = O(\frac{\log n}{\log k})$, $m = k^\gamma$ and a poly(n)-time computable extractor $E : (\{0,1\}^n)^\ell \mapsto \{0,1\}^m$ for balanced-ℓ-sources with entropy threshold $\ell \cdot k$ and error $\epsilon < 2^{-(m+1)}$.*

Note that by Fact 1.4 such an extractor is in particular a μ-strongly hitting disperser for $\mu = 2^{-(m+1)}$. We now show how to improve the output length to $m = \Omega(k)$ while preserving this property.

Theorem 5.14. *There are numbers $c', \eta > 0$ such that for every sufficiently large k, n such that $k \geq (\log n)^{c'}$ there are integers $\ell = O(\frac{\log n}{\log k})$, $m = \eta k$ and a poly(n)-time computable $2^{-(m+3)}$-strongly hitting disperser $D : (\{0,1\}^n)^\ell \mapsto \{0,1\}^m$ for balanced-ℓ-sources with entropy threshold $\ell \cdot k$.*

Proof. By Corollary 5.12 there exist $\eta > 0$ and c such that for sufficiently large $k \leq n$ and $m = \eta k$ there is a poly(n)-time computable generalized subsource hitter $F : (\{0,1\}^n)^2 \times \{0,1\}^{c \log n} \mapsto \{0,1\}^m$ for balanced-2-sources with entropy threshold $2 \cdot k$, subsource entropy k, measure $2^{-(m+1)}$ and error $1/10$. Let $t = c \log n$.

Let E be the extractor from Theorem 5.13 (for the same k, n) and let γ, ℓ, m be the parameters associated with it. The extractor E has output length k^γ by choosing c' to be a sufficiently large constant as a function of the constants c, γ we have that $k \geq (\log n)^{c'}$ and so $k^\gamma \geq c \log n$. We can thus chop the output of E to length $t = c \log n$. Note that E is a $2^{-(t+1)}$-strongly hitting disperser. Let $\ell' = \ell + 2$. We construct a zero-error disperser D for balanced-ℓ'-sources with entropy threshold $\ell' \cdot k$ by

$$D(x_1, \ldots, x_{\ell'}) = F(x_{\ell+1}, x_{\ell+2}, E(x_1, \ldots, x_\ell))$$

Indeed, let $X = (X_1, \ldots, X_{\ell'})$ be a balanced-ℓ'-source with min-entropy at least $\ell' \cdot k$. We consider the balanced-2-source $(X_{\ell+1}, X_{\ell+2})$. By the properties of F we have that for every $z \in \{0,1\}^m$ a $9/10$ fraction of $y \in \{0,1\}^t$ (which we call good seeds) have that

$$\Pr[F(X_{\ell+1}, X_{\ell+2}, y) = z] \geq 2^{-(m+1)} \tag{1}$$

(Note that here we're not using the property that F hits z on a well structured subsource. We're only using the fact that F hits z with positive probability.) We also consider the balanced-ℓ-source $(X_1, \ldots, X_\ell)$. As E is a $2^{-(t+1)}$-strongly hitting disperser we have that for every $y \in \{0,1\}^t$

$$\Pr[E(X_1, \ldots, X_\ell) = y] > 2^{-(t+1)} \tag{2}$$

For every good seed $y \in \{0,1\}^t$ we have that the two events in (1) and (2) are independent and therefore the probability that they occur simultaneously is at least $2^{-(t+1)} \cdot 2^{-(m+1)}$. Whenever this happens we have that $D(X_1, \ldots, X_{\ell'}) = z$. Summing up over the $\frac{9}{10} \cdot 2^t$ good seeds y we have that:

$$\Pr[D(X_1, \ldots, X_{\ell'}) = z] \geq \frac{9}{10} \cdot 2^t \cdot 2^{-(t+1)} \cdot 2^{-(m+1)} \geq 2^{-(m+3)}$$

□

5.5 Rainbows and implicit $O(1)$ probe search

In this section we discuss an application of zero-error dispersers to the problem of *implicit probe search*. Loosely speaking, this is the problem of searching for an element in a table with few probes, when no additional information but the elements themselves is stored.

Definition 5.15 (Implicit probe search scheme)**.** *For integer parameters n, k, q, the implicit probe search problem is as follows: Store a subset $S \subseteq \{0,1\}^n$ of size 2^k in a table T of size 2^k, (where every table entry holds only a single element of S), such that given $x \in \{0,1\}^n$ we can determine whether $x \in S$ using q queries to T. A solution to this problem is called an* implicit q-probe scheme with table size 2^k and domain size 2^n.

Fiat and Naor [15] investigated implicit $O(1)$-probe schemes, i.e., schemes where the number of queries is a constant not depending on n and k. They showed that this problem is unsolvable when n is large enough relative to k (this improves a previous bound by Yao [36]). They also gave an efficient implicit $O(1)$-probe scheme whenever $k = \delta \cdot n$ for any constant $\delta > 0$. They did this by reducing the problem to the task of constructing a combinatorial object called a *rainbow*.

Definition 5.16. *[15]*

- *A* t-sequence *over a set U is a sequence of length t without repetitions, of elements in U.*
- *An (n, k, t)-rainbow is a coloring of all t-sequences over $\{0,1\}^n$ with 2^k colors such that for any $S \subseteq \{0,1\}^n$ of size 2^k, the t-sequences over S are colored in all colors.*

Fiat and Naor showed that rainbows imply implicit probe schemes.

Theorem 5.17. *[15] Fix any integers n, k with $\log n \leq k \leq n$. Given a polynomial time computable (n, k, t)-rainbow we can construct a polynomial time computable implicit $O(t)$-probe scheme with table size 2^k and domain size 2^n. In particular, when t is constant we get an implicit polynomial time computable $O(1)$-probe scheme.*

The following easy theorem shows that we can construct rainbows from strongly hitting dispersers for multiple independent sources.

Theorem 5.18. *Let $0 < \eta < 1$ be any constant, and let n, k and t be integers with $\log n \leq k \leq n$. Let $m = \eta k$ and let $G : \{0,1\}^{t \cdot n} \mapsto \{0,1\}^m$ be a polynomial time computable $t^2/2^k$-strongly hitting disperser for balanced-t-sources with entropy threshold $t \cdot k$, then there is a polynomial time computable $(n, k, O(t/\eta))$-rainbow.*

Proof. Fix a subset $S \subseteq \{0,1\}^n$ with $|S| = 2^k$. Let X be the uniform distribution on $S \subseteq \{0,1\}^n$. Thus X has min-entropy k. Let X^{*t} denote the distribution defined by t independent copies of X. X^{*t} is a t-source with block entropy exactly k. Therefore for any $z \in \{0,1\}^m$

$$\Pr[G(X^{*t}) = z] \geq t^2/2^k$$

We say that a sequence $(x_1, \ldots, x_t) \in (\{0,1\}^n)^t$ has repetitions if there exists $i \neq j$ such that $x_i = x_j$. The probability that X^{*t} produces a sequence with repetitions is smaller than $t^2/2^k$. Therefore there must be $(x_1, \ldots, x_t) \in \mathrm{Supp}(X^{*t})$ without repetitions such that $G(x_1, \ldots, x_t) = z$. Note that this exactly means there is a t-sequence of elements of S that G 'colors' z. Let $\ell = 1/\eta$ and consider a function $\bar{G} : (\{0,1\}^n)^{t\ell} \to \{0,1\}^k$ which partitions the input sequence into ℓ sequences of length t, applies G on each sequence, concatenates the outputs and truncates the output string to length k if necessary. By the previous analysis we have that $\bar{G}$ is a $(n, k, t\ell)$-rainbow. This is because every $t\ell$-sequence is colored by all possible 2^k colors. □

Plugging in our strongly hitting disperser for multiple independent sources we get the following implicit probe scheme.

Corollary 5.19. *Fix any constant $0 < \delta < 1$. For every sufficiently large n and $k = n^\delta$ there is a poly(n)-time computable implicit $O(1/\delta)$-probe scheme with table size 2^k and domain size 2^n.*

Proof. Fix k and n with $k = n^\delta$ for some constant $\delta > 0$. Using Theorem 5.14, we get a 2^{-m+3}-strongly hitting disperser $D : (\{0,1\}^n)^\ell \mapsto \{0,1\}^m$ for balanced-ℓ-sources with entropy threshold $k \cdot \ell$ where $\ell = O(\frac{\log n}{\log k}) = O(1/\delta)$ and $m = \eta \cdot k$ for some constant $0 < \eta < 1$ (not depending on δ). Applying Theorem 5.18 we get an $(n, k, O(1/\delta))$-rainbow and therefore by Theorem 5.17 an implicit $O(1/\delta)$-probe scheme with table size 2^k and domain size 2^n. □

We remark that using the technique of [15] and plugging in recent constructions of seeded dispersers seems to also give an implicit $O(1)$ probe scheme for the case that $k = n^{\Omega(1)}$.

6 Zero-error dispersers for bit-fixing sources

In this section we construct dispersers for the family of *bit-fixing sources* introduced by Chor et al. [10]. A distribution X over $\{0,1\}^n$ is a bit-fixing source if there is a subset $S \subseteq [n]$ of "good indices" such that the bits X_i for $i \in S$ are independent fair coins and the rest of the bits are fixed.

Definition 6.1 (bit-fixing sources)**.** *A distribution X over $\{0,1\}^n$ is an (n,k)-*bit-fixing source *if there exists a subset $S = \{i_1, \ldots, i_k\} \subseteq [n]$ such that $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ is uniformly distributed over $\{0,1\}^k$ and for every $i \notin S$, X_i is constant.*
The class of bit-fixing sources over $\{0,1\}^n$ *is the class of all (n,k)-bit-fixing sources for some $1 \leq k \leq n$.*

We will construct subsource-hitters for bit-fixing sources and these will allow us to obtain improved zero-error dispersers. An ingredient in the construction of subsource-hitters is the following easy Lemma on sampling properties of pairwise independence.

Lemma 6.2. *For any integers k and n with $64 < k \leq n$, there is a poly(n)-time computable function $P : \{0,1\}^{2\log n} \mapsto (\mathrm{P}([n]))^4$ returning a partition of $[n]$ into 4 disjoint sets $P(y)_1 \cup P(y)_2 \cup P(y)_3 \cup P(y)_4 = [n]$ such that for any (n,k)-bit-fixing source X, there exists a $y \in \{0,1\}^{2\log n}$ such that for every $i \in [4]$, $X_{P(y)_i}$ is an (n', k')-bit-fixing source for some $n' \leq n$ and $k' \geq k/8$.*

Proof. We use y as a random seed to generate pairwise independent variables $Z_1, \ldots, Z_n \in [4]$ (there are constructions which use $2\log n$ bits to generate such variables [8]). For $i = 1, \ldots, 4$ define the subset $P(y)_i \subseteq [n]$ by $P(y)_i \triangleq \{j : Z_j = i\}$. Assume without loss of generality that the 'good indices' of X are $\{1, \ldots, k\}$. Fix any $i \in [4]$. For $j \in [n]$ define the random variable X_j by $X_j = 1$ if $Z_j = i$ and 0 otherwise. Then for $j \in [n]$, $E(X_j) = 1/4$ and $Var(X_j) \leq 1/4$. Furthermore, for $j \neq l$ X_j and X_l are independent and $cov(X_j, X_l) = 0$. Define $Y = \sum_{j=1}^k X_j$. We have $E(Y) = k/4$ and $Var(Y) = \sum_{j=1}^k Var(X_j) \leq k/4$. Therefore by Chebychev's inequality

$$\Pr[|Y - k/4| \geq k/8] \leq k/4 \cdot (8/k)^2 \leq 16/k.$$

Note that Y is exactly the number of good indices in $P(Y)_i$. Thus, using the union bound, with probability $1 - 64/k$ over y, for every $i \in [4]$ $P(Y)_i$ contains at least $k/8$ good indices of X. In particular, when $k > 64$ there exists a y such that for every $i \in [4]$, $X_{P(y)_i}$ is an (n', k')-bit-fixing source for some $n' \leq n$ and $k' \geq k/8$. □

Using the above lemma, we show how to construct a subsource hitter for bit-fixing sources from a subsource hitter for 2-sources.

Lemma 6.3. *Fix any integers k and n with $64 < k \leq n$.*

- *Let $G : (\{0,1\}^n \times \{0,1\}^n) \times \{0,1\}^t \mapsto \{0,1\}^m$ be a subsource hitter for balanced-2-sources with entropy threshold $2 \cdot k/8$ and subsource entropy l.*
- *Let $P : \{0,1\}^{2\log n} \mapsto (\mathrm{P}([n]))^4$ be the partitioning function from Lemma 6.2.*

Define the function $F : \{0,1\}^n \times \{0,1\}^{2\log n + t} \mapsto \{0,1\}^m$ by

$$F(x,(y,y')) \triangleq G((x_{P(y)_1}, x_{P(y)_2}), y').$$

(we pad $x_{P(y)_1}$ and $x_{P(y)_2}$ with zeros to make them n-bit strings.) Then F is a subsource hitter for bit-fixing sources with entropy threshold k and subsource entropy $k/4$.

Proof. Let X be an (n,k)-bit-fixing source. Using Lemma 6.2, we can fix a $y \in \{0,1\}^{2\log n}$ such that for every $i \in [4]$, $X_i \triangleq X_{P(y)_i}$ is an (n',k')-bit-fixing source for some $n' \leq n$ and $k' \geq k/8$. Note that (X_1, X_2) is a 2-source with block entropy at least $k/8$. Fix any $z \in \{0,1\}^m$, and fix $y' \in \{0,1\}^t$ such that there is a subsource (X_1', X_2') of (X_1, X_2) with min-entropy at least l such that $\Pr[G((X_1', X_2'), y') = z] = 1$. (Such a subsource exists as by Lemma 5.2 (X_1, X_2) has a subsource (X_1^*, X_2^*) which is a balanced-2-source with entropy threshold $2 \cdot k/8$. (X_1^*, X_2^*) contains such a subsource (X_1', X_2') by the guarantee of G and (X_1', X_2') is also a subsource of (X_1, X_2)). Fix an arbitrary $x' \in \mathrm{Supp}(X_1', X_2')$ and let $X' \triangleq (X | (X_1, X_2) = x')$. Note that X' is an (n,k')-bit-fixing source for some $k' \geq k/4$ as $P(y)_3 \cup P(y)_4$ contain at least $k/4$ good indices. For any $x \in X'$, we have

$$F(x,(y,y')) = G(x',y') = z$$

and thus $\Pr[F(X',(y,y')) = z] = 1$. As X' is a subsource of X with min-entropy at least $k/4$ this proves that F is a subsource hitter for bit-fixing sources with entropy threshold k and subsource entropy $k/4$. □

Plugging in the subsource hitter for 2-sources from Corollary 5.12 we get the following.

Corollary 6.4. *There exist constants $c > 0$ and $0 < \eta < 1$ such that for every sufficiently large $k \leq n$ there is a poly(n)-time computable subsource hitter $F : \{0,1\}^n \times \{0,1\}^{c \log n} \mapsto \{0,1\}^m$ for bit-fixing sources with entropy threshold k and subsource entropy $k/4$, where $m = \eta \cdot k$.*

We use this subsource-hitter to improve the output length of the following zero-error disperser of [28].

Theorem 6.5. *[28] There exist constants $c > 0$ and $0 < d < 1$ s.t. for every $k \leq n$ with $k \geq log^c n$, there is a poly(n)-time computable zero-error disperser $D : \{0,1\}^n \mapsto \{0,1\}^t$ for bit-fixing sources with entropy threshold k, where $t = k^d$.*

We remark that the object constructed in [28] is stronger than stated in Theorem 6.5. It is an extractor that achieves error 2^{-k^d} for a more general class of sources called "low-weight affine sources" in [28]. We can now prove our main result for bit-fixing sources. The following Theorem is a formal restatement of Theorem 2.2.

Theorem 6.6. *There exist constants $c > 0$ and $0 < \eta < 1$ such that for every sufficiently large $k \leq n$ with $k \geq \log^c n$ there is a poly(n)-time computable zero-error disperser $D : \{0,1\}^n \mapsto \{0,1\}^m$ for bit-fixing sources with entropy threshold k, where $m = \eta \cdot k$.*

Proof. Using Theorem 6.5 and Corollary 6.4 we can choose a large enough constant c' such that for some constants $0 < d, \eta < 1$, for any $k \geq \log^{c'} n$, we have the following explicit components:

- A zero-error disperser $D' : \{0,1\}^n \mapsto \{0,1\}^{(k/4)^d}$ for bit-fixing sources with entropy threshold $k/4$.
- A subsource hitter $F : \{0,1\}^n \times \{0,1\}^{c' \log n} \mapsto \{0,1\}^{\eta \cdot k}$ for bit-fixing sources with entropy threshold k and subsource entropy $k/4$.

To use Theorem 4.2, we need to make sure D''s output is as long as F's seed. Assuming $k \geq \log^{2/d} n$ we have

$$(k/4)^d \geq (\log^{2/d} n/4)^d \geq (\log^2 n)/4^d \geq c' \cdot \log n,$$

for large enough n. Thus, using Theorem 4.2, we get a zero-error disperser $D : \{0,1\}^n \mapsto \{0,1\}^{\eta \cdot k}$ for bit-fixing sources with entropy threshold k. Taking $c = \max\{c', 2/d\}$ we are done. □

We remark that a more careful analysis can be used to construct a strongly hitting disperser along the same lines.

7 Zero-error dispersers for affine sources

In this section we construct dispersers for affine sources over polynomial size fields. Let q be a prime power. Let $\mathbb{F}_q$ denote the finite field with q elements. Let $\mathbb{F}_q^n$ denote the n-dimensional vector space over $\mathbb{F}_q$. The definition of affine sources is given below.

Definition 7.1 (affine source)**.** *A distribution X over $\mathbb{F}_q^n$ is an $(n,d)_q$-*affine source *if it is uniformly distributed over an affine subspace of dimension d. That is, X is sampled by choosing $t_1, \ldots, t_d$ uniformly and independently in $\mathbb{F}_q$ and calculating*

$$\sum_{j=1}^{d} t_j \cdot a^{(j)} + b$$

for some $a^{(1)}, \ldots, a^{(d)}, b \in \mathbb{F}_q^n$ such that $a^{(1)}, \ldots, a^{(d)}$ are linearly independent.
The class of affine sources over $\mathbb{F}_q^n$ *is the class of all $(n,d)_q$-affine sources for some $1 \leq d \leq n$.*

Note that an $(n,d)_q$-affine source has min-entropy $k = d \cdot \log q$. We will improve the output length of the following zero error disperser of [16]:

Theorem 7.2. *[16] Fix any sufficiently large prime power q and any integer n such that $q \geq n^9$. There is a poly$(n, \log q)$-time computable[4] zero-error disperser $D : \mathbb{F}_q^n \mapsto \{0,1\}^t$ for affine sources with entropy threshold $\log q$, where $t = \lceil (1/6) \log q \rceil$.*

[4]When we say that D is poly$(n, \log q)$-time computable we mean that computing D requires poly(n) field operations in $\mathbb{F}_q$. Thus, assuming we have a representation of $\mathbb{F}_q$ in which addition and multiplication can be done in poly$(\log q)$ time (which is true for all standard representations) we get that D is poly$(n, \log q)$-time computable.

We remark that the construction of [16] actually gives an extractor with error 2^{-t} and the theorem stated above follows using Fact 1.4. Another component in [16] is a way to improve the output length of the extractor. This gives an extractor which extracts many bits with the same error. We are interested in zero error dispersers and show that this component can be seen as a subsource hitter for affine sources. This will allow us to improve the output length while preserving the zero-error property.

We use the following construction from [16]. Given $u \in \mathbb{F}_q$ and an integer d, we define a $d \times n$ matrix $T_{u,d}$ by $(T_{u,d})_{j,i} = u^{ji}$ (where ji is an integer product). For $x \in \mathbb{F}_q^n$ we define $T_{u,d}(x)$ to be the multiplication of $T_{u,d}$ with x, that is:

$$T_{u,d}(x) = \left(\sum_{i=1}^{n} x_i \cdot u^i, \sum_{i=1}^{n} x_i \cdot u^{2i}, \ldots, \sum_{i=1}^{n} x_i \cdot u^{di} \right)$$

Lemma 7.3. *[Lemma 6.1 in [16]] Fix any field $\mathbb{F}_q$ and integers n, d such that $q \geq n \cdot d^2$. Fix any affine subspace $A \subseteq \mathbb{F}_q^n$ of dimension at least d. There are at most $n \cdot d^2$ elements $u \in \mathbb{F}_q$ such that $T_{u,d}(A) \subsetneq \mathbb{F}_q^d$.*

We now observe that this lemma gives a subsource hitter.

Corollary 7.4. *Fix any sufficiently large prime power q and any integers n, d such that $q \geq n^{18}$ and $2 \leq d < n$. Let $s = 2^t$ where $t = \lceil (1/6) \log q \rceil$. Let $U = \{u_1, \ldots, u_s\}$ be a set of distinct elements in $\mathbb{F}_q$. We identify U with $\{0,1\}^t$. The function $F : \mathbb{F}_q^n \times \{0,1\}^t \to \mathbb{F}_q^{d-1}$ defined by*

$$F(x, u) \triangleq T_{u,d-1}(x)$$

is a subsource hitter for affine sources with entropy threshold $d \cdot \log q$ and subsource entropy $\log q$.

Proof. X is uniformly distributed on an affine subspace A of dimension d, i.e., $\mathrm{Supp}(X) = A$. Since $|U| = s \geq q^{1/6} > n \cdot (d-1)^2$, by Lemma 7.3 there is $u \in U$ such that $T_{u,d-1}(A) = \mathbb{F}_q^{d-1}$. Fix such a u. Given any $z \in \mathbb{F}_q^{d-1}$, define $X' = (X | F(X, u) = z)$. $\mathrm{Supp}(X')$ is not empty by our choice of u. Moreover, since the conditioning $F(X, u) = z$ simply adds $d - 1$ affine constraints on $\mathrm{Supp}(X)$, $\mathrm{Supp}(X')$ is an affine subspace of dimension at least 1. Thus, X' is a subsource of X that is also an affine source with min-entropy at least $\log q$. Since $\Pr[F(X', u) = z) = 1]$, this proves the claim. □

We construct a zero-error disperser by using our composition technique.

Theorem 7.5. *Fix any sufficiently large prime power q and any integers n, d such that $q \geq n^{18}$ and $2 \leq d < n$. There is a poly$(n, \log q)$-time computable $D : \mathbb{F}_q^n \mapsto \{0,1\}^{(d-1) \cdot \log q}$ that is a zero-error disperser for affine sources over $\mathbb{F}_q^n$ with entropy threshold $d \log q$.*

Proof. Use Theorem 4.2 using the disperser from Theorem 7.2 and the subsource hitter from Corollary 7.4. □

8 Open problems

2-sources. One of the most important open problems in this area is to give constructions of extractors for entropy threshold $k = o(n)$. Such constructions are not known even for $m = 1$ and large error ϵ.

There are explicit constructions of zero-error dispersers with $k = n^{o(1)}$ [3]. However, these dispersers only output one bit. A consequence of Corollary 5.12 is that improving the output

length in these constructions to $\Theta(\log n)$ bits will allow our composition techniques to achieve output length $m = \Omega(k)$.

Another intriguing problem is that for the case of zero-error (or strongly hitting) dispersers we do not know whether the existential results proven via the probabilistic method achieve the best possible parameters. More precisely, a straightforward application of the probabilistic method gives zero-error 2-source dispersers which on entropy threshold $2 \cdot k$ output $m = k - \log(n-k) - O(1)$ bits. On the other hand the lower bounds of [25, 26] can be used to show that any zero-error 2-source disperser with entropy threshold $2 \cdot k$ has $m \leq k + O(1)$.[5]

O(1)-sources, rainbows and implicit probe search. When allowing ℓ-sources for $\ell = O(1)$ we give constructions of zero-error dispersers which on entropy threshold $k = n^{\Omega(1)}$ achieve output length $m = \Omega(k)$. An interesting open problem is to try and improve the entropy threshold. As explained in Subsection 5.5 this immediately implies improved implicit probe search schemes.

Bit-fixing sources. We give constructions of zero-error dispersers which on entropy threshold k achieve $m = \Omega(k)$. A straightforward application of the probabilistic method gives zero-error dispersers with $m = k - \log n - o(\log n)$. We do not know how to match these parameters with explicit constructions.

Affine sources. We constructed a subsource hitter for affine sources over relatively large fields (that is $q = n^{\Theta(1)}$). It is interesting to try and construct subsource hitters for smaller fields.

Dispersers for low entropy thresholds The technique developed in this paper increases the output length of zero error dispersers. A different goal is to try and reduce the entropy threshold of dispersers (for various classes) even for output length $m = 1$. We mention that in the past, extractors and dispersers with large output length turned out to be useful in constructions that output one bit for lower entropy threshold.

9 Acknowledgements

We are grateful to Ran Raz for his support.

References

[1] B. Barak, R. Impagliazzo, and A. Wigderson. Extracting randomness using few independent sources. *SIAM J. Comput*, 36(4):1095–1118, 2006.

[2] B. Barak, G. Kindler, R. Shaltiel, B. Sudakov, and A. Wigderson. Simulating independence: New consturctions of condenesers, ramsey graphs, dispersers, and extractors. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 1–10, 2005.

[5] Radhakrishnan and Ta-Shma [26] show that any seeded disperser $D : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ that is nontrivial in the sense that $m \geq t + 1$ has $t \geq \log(1/\epsilon) - O(1)$. A zero-error 2-source disperser D' with entropy threshold k can be easily transformed into a seeded disperser with seed length $t = k$ by setting $D(x, y) = D'(x, y')$ where y' is obtained by padding the k bit long "seed" y with $n - k$ zeroes. The bound follows as D' has error smaller than 2^{-m}.

[3] B. Barak, A. Rao, R. Shaltiel, and A. Wigderson. 2-source dispersers for sub-polynomial entropy and Ramsey graphs beating the Frankl–Wilson construction. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 671–680, 2006.

[4] M. Ben-Or and N. Linial. Collective coin flipping. *ADVCR: Advances in Computing Research*, 5:91–115, 1989.

[5] M. Blum. Independent unbiased coin flips from a correlated biased source-a finite stae markov chain. *Combinatorica*, 6(2):97–108, 1986.

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

[7] J. Bourgain. On the construction of affine extractors. *Geometric And Functional Analysis*, 17(1):33–57, 2007.

[8] I. L. Carter and M. N. Wegman. Universal classes of hash functions. In *Proceedings of the 9th Annual ACM Symposium on Theory of Computing*, pages 106–112, 1977.

[9] B. Chor and O. Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, April 1988. Special issue on cryptography.

[10] B. Chor, O. Goldreich, J. Hastad, J. Friedman, S. Rudich, and R. Smolensky. The bit extraction problem or t-resilient functions. In *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science*, pages 396–407, 1985.

[11] A. Cohen and A. Wigderson. Dispersers, deterministic amplification and weak random sources. In *Proceedings of the 30th Annual IEEE Symposium on Foundations of Computer Science*, pages 14–25, 1989.

[12] Y. Dodis, A. Elbaz, R. Oliveira, and R. Raz. Improved randomness extraction from two independent sources. In *RANDOM: International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 334–344, 2004.

[13] Z. Dvir, A. Gabizon, and A. Wigderson. Extractors and rank extractors for polynomial sources. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 52–62, 2007.

[14] Z. Dvir and A. Wigderson. Kakeya sets, new mergers and old extractors. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 625–633, 2008.

[15] A. Fiat and M.Naor. Implicit O(1) probe search. *SICOMP: SIAM Journal on Computing*, 22, 1993.

[16] A. Gabizon and R. Raz. Deterministic extractors for affine sources over large fields. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 407–418, 2005.

[17] A. Gabizon, R. Raz, and R. Shaltiel. Deterministic extractors for bit-fixing sources by obtaining an independent seed. *SICOMP: SIAM Journal on Computing*, 36(4):1072–1094, 2006.

[18] O. Goldreich. A sample of samplers – A computational perspective on sampling (survey). In *ECCCTR: Electronic Colloquium on Computational Complexity, technical reports*, 1997.

[19] R. L. Graham, B. L. Rothschild, and J. H. Spencer. *Ramsey Theory.* Wiley, 1980.

[20] V. Guruswami, C. Umans, and S. P. Vadhan. Unbalanced expanders and randomness extractors from parvaresh-vardy codes. In *Proceedings of the 22nd Annual IEEE Conference on Computational Complexity*, pages 96–108, 2007.

[21] J. Kamp, A. Rao, S. Vadhan, and D. Zuckerman. Deterministic extractors for small-space sources. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 691–700, 2006.

[22] J. Kamp and D. Zuckerman. Deterministic extractors for bit-fixing sources and exposure-resilient cryptography. *SIAM J. Comput*, 36(5):1231–1247, 2007.

[23] C. 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.

[24] E. Mossel and C. Umans. On the complexity of approximating the vc dimension. In *Sixteenth Annual IEEE Conference on Computational Complexity*, pages 220–225, 2001.

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

[26] J. Radhakrishnan and A. Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM Journal on Discrete Mathematics*, 13(1):2–24, 2000.

[27] A. Rao. Extractors for a constant number of polynomially small min-entropy independent sources. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 497–506, 2006.

[28] A. Rao. Extractors for low weight affine sources. *Unpublished Manuscript*, 2008.

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

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

[31] R. Shaltiel. Recent developments in explicit constructions of extractors. *Bulletin of the EATCS*, 77:67–95, 2002.

[32] R. Shaltiel. How to get more mileage from randomness extractors. In *CCC '06: Proceedings of the 21st Annual IEEE Conference on Computational Complexity*, pages 46–60, 2006.

[33] L. Trevisan and S. Vadhan. Extracting randomness from samplable distributions. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 32–42, 2000.

[34] U. Vazirani. Strong communication complexity or generating quasi-random sequences from two communicating semi-random sources. *Combinatorica*, 7:375–392, 1987.

[35] J. von Neumann. Various techniques used in connection with random digits. *Applied Math Series*, 12:36–38, 1951.

[36] A. C.-C. Yao. Should tables be sorted? *J. ACM*, 28(3):615–628, 1981.