Electronic Colloquium on Computational Complexity, Report No. 166 (2011)

Non-Malleable Extractors for Entropy Rate $< 1/2$

Xin Li*
Department of Computer Science
University of Washington
Seattle, WA 98905, U.S.A.
lixints@cs.washington.edu

Abstract

Dodis and Wichs [DW09] introduced the notion of a non-malleable extractor to study the problem of privacy amplification with an active adversary. A non-malleable extractor is a much stronger version of a strong extractor. Given a weakly-random string x and a uniformly random seed y as the inputs, the non-malleable extractor nmExt has the property that $\mathsf{nmExt}(x, y)$ appears uniform even given y as well as $\mathsf{nmExt}(x, \mathcal{A}(y))$, for an arbitrary function $\mathcal{A}$ with $\mathcal{A}(y) \neq y$. Dodis and Wichs showed that such an object can be used to give optimal privacy amplification protocols with an active adversary.

Previously, there are only two known explicit constructions of non-malleable extractors [DLWZ11, CRS11]. Both constructions only work for (n, k)-sources with $k > n/2$ and thus they also only achieve optimal privacy amplification protocols for $k > n/2$. In this paper, we give the first construction of non-malleable extractors for $k < n/2$. Specifically, we give two unconditional constructions for min-entropy $k = (1/2 - \delta)n$ for some constant $\delta > 0$, and a conditional construction that can potentially achieve $k = \alpha n$ for any constant $\alpha > 0$. Using our non-malleable extractor, we also obtain the first optimal privacy amplification protocol for min-entropy $k = (1/2 - \delta)n$, with an active adversary.

Our constructions mainly use the bilinear property of the inner product function, and involve appropriate encodings of the seed y as well as ideas from additive combinatorics.

*Partially supported by NSF Grants CCF-0634811, CCF-0916160, THECB ARP Grant 003658-0113-2007, and a Simons postdoctoral fellowship.

ISSN 1433-8092

1 Introduction

Seeded randomness extractors as defined by Nisan and Zuckerman [NZ96] is an object that has been studied extensively in computer science. Besides its original motivation in computing with imperfect random sources, seeded extractors have found applications in coding theory, cryptography, complexity and many other areas. We refer the reader to [FS02, Vad02] for a survey on this subject. Especially, seeded extractors have been used in cryptography to give protocols that are leakage resilient. Recently, a new kind of seeded extractors, called *non-malleable extractors* were introduced in [DW09] to give protocols for the problem of privacy amplification with an active adversary. We now give the definition of a non-malleable extractor below. As a comparison, we also give the definition of a strong extractor.

Notation. We let $[s]$ denote the set $\{1, 2, \ldots, s\}$. For ℓ a positive integer, U_ℓ denotes the uniform distribution on $\{0,1\}^\ell$, and for S a set, U_S denotes the uniform distribution on S. When used as a component in a vector, each U_ℓ or U_S is assumed independent of the other components. We say $W \approx_\varepsilon Z$ if the random variables W and Z have distributions which are ε-close in variation distance.

Definition 1.1. The *min-entropy* of a random variable X is

$$H_\infty(X) = \min_{x \in \mathsf{supp}(X)} \log_2(1/\Pr[X = x]).$$

For $X \in \{0,1\}^n$, we call X an $(n, H_\infty(X))$-source, and we say X has *entropy rate* $H_\infty(X)/n$. We say X is a flat source if it is the uniform distribution over some subset $S \subset \{0,1\}^n$.

Definition 1.2. A function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a *strong* (k, ε)*-extractor* if for every source X with min-entropy k and independent Y which is uniform on $\{0,1\}^d$,

$$(\mathsf{Ext}(X,Y), Y) \approx_\varepsilon (U_m, Y).$$

Definition 1.3. [1] A function $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k, ε)-non-malleable extractor if, for any source X with $H_\infty(X) \geq k$ and any function $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ such that $\mathcal{A}(y) \neq y$ for all y, the following holds. When Y is chosen uniformly from $\{0,1\}^d$ and independent of X,

$$(\mathsf{nmExt}(X,Y), \mathsf{nmExt}(X, \mathcal{A}(Y)), Y) \approx_\varepsilon (U_m, \mathsf{nmExt}(X, \mathcal{A}(Y)), Y).$$

As we can see from the definitions, a non-malleable extractor is a stronger version of the strong extractor, in the sense that it requires the output to be close to uniform even conditioned on both the seed Y and the output $\mathsf{nmExt}(X, \mathcal{A}(Y))$ on a different but arbitrarily correlated seed $\mathcal{A}(Y)$.

The motivation to study a non-malleable extractor, the privacy amplification problem, is a fundamental problem in symmetric cryptography that has been studied by many researchers. Bennett, Brassard, and Robert introduced this problem in [BBR88]. The basic setting is that, two parties (Alice and Bob) share an n-bit secret key X, which is weakly random. This could happen because the secret comes from a password or biometric data, which are themselves weakly random, or because an adversary Eve managed to learn some partial information about an originally uniform secret, for example via side channel attacks. We measure the entropy of X by the min-entropy

[1] Following [DLWZ11], we define worst case non-malleable extractors, which is slightly different from the original definition of average case non-malleable extractors in [DW09]. However, the two definitions are essentially equivalent up to a small change of parameters.

defined above. The goal is to have Alice and Bob communicate over a public channel so that they can convert X into a nearly uniform secret key. Generally, we also assume that Alice and Bob have local private uniform random bits. The problem is the presence of the adversary Eve, who can see every message transmitted in the channel and may or may not change the messages. We assume that Eve has unlimited computational power.

The case where Eve is *passive*, i.e., cannot change the messages, can be solved simply by using the above mentioned strong seeded extractors. The case where Eve is *active* (i.e., can change the messages in arbitrary ways), on the other hand, is much more difficult. In this case, the simple solution by using a strong seeded extractor no longer works, and stronger tools are required. Historically, Maurer and Wolf [MW97] gave the first non-trivial protocol in this case. Their protocol takes one round and works when the entropy rate of the weakly-random secret X is bigger than $2/3$. Dodis, Katz, Reyzin, and Smith [DKRS06] later improved this result to give protocols that work for entropy rate bigger than $1/2$. One drawback in both cases is that the final secret key R is much shorter than the min-entropy of X. Later, Dodis and Wichs [DW09] showed that no one-round protocol exists for entropy rate less than $1/2$. The first protocol that breaks the $1/2$ entropy rate barrier is due to Renner and Wolf [RW03], where they gave a protocol that works for essentially any entropy rate. However their protocol takes $O(s)$ rounds and only achieves entropy loss $O(s^2)$, where s in the security parameter of the protocol. Kanukurthi and Reyzin [KR09] simplified their protocol, but the parameters remain essentially the same.

In [DW09], Dodis and Wichs showed that explicit non-malleable extractors can be used to give privacy amplification protocols that take an optimal 2 rounds and achieve optimal entropy loss $O(s)$. They showed that non-malleable extractors exist when $k > 2m + 3\log(1/\varepsilon) + \log d + 9$ and $d > \log(n-k+1) + 2\log(1/\varepsilon) + 7$. However, they only constructed weaker forms of non-malleable extractors and they gave a protocol that takes 2 rounds but that still has entropy loss $O(s^2)$. Chandran, Kanukurthi, Ostrovsky and Reyzin [CKOR10] improved the entropy loss to $O(s)$ but the number of rounds becomes $O(s)$.

Dodis, Li, Wooley and Zuckerman [DLWZ11] constructed the first explicit non-malleable extractor. Their construction works for entropy $k > n/2$, but they use a large seed length $d = n$ and the efficiency when outputting more than $\log n$ bits relies on an unproven assumption. Cohen, Raz, and Segev [CRS11] later gave an alternative construction that also works for $k > n/2$, but uses a short seed length and does not rely on any unproven assumption. By using the non-malleable extractors, these two papers thus gave 2-round privacy amplification protocols that achieve optimal entropy loss $O(s)$. However, since both constructions of non-malleable extractors are only shown to work for entropy $k > n/2$,[2] the protocols also only work for $k > n/2$. Thus the natural open question is whether we can construct non-malleable extractors for smaller min-entropy, and whether there are 2-round privacy amplification protocols with optimal entropy loss for smaller min-entropy.

1.1 Our results

In this paper, we improve previous results by giving the first explicit non-malleable extractors that work for min-entropy $k < n/2$. We give two unconditional constructions that work for $k = (1/2-\delta)n$ for some universal constant $\delta > 0$. We also give a conditional construction that can potentially work for $k = \delta n$ for any constant $\delta > 0$. Specifically, we have the following theorems.

[2] We remark that the construction in [DLWZ11] is a special case of the construction in [CRS11]. Also, it is possible that the construction in [DLWZ11] can work for entropy $k \leq n/2$ (but until now nobody can prove it), but the construction in [CRS11] in general cannot work for entropy $k \leq n/2$.

Theorem 1.4. *There exist constants $0 < \delta, \gamma < 1$ such that for any $n \in \mathbb{N}$, $k = (1/2-\delta)n$ and any $\epsilon > 2^{-\gamma n}$, there exists an explicit (k, ϵ)-non-malleable extractor* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ *with $d = O(\log n + \log(1/\epsilon))$.*

Theorem 1.5. *There exists a constant $0 < \delta < 1$ such that for any $n \in \mathbb{N}$, $k = (1/2-\delta)n$, there exists an explicit (k, ϵ)-non-malleable extractor* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *with $m = \Omega(n)$ and $\epsilon = 2^{-\Omega(n)}$.*

Our third result needs to use an affine extractor and an assumption from additive combinatorics, as used in [BSZ11]. Thus we first define affine extractors and state the assumption.

Definition 1.6. An $[n, m, \rho, \epsilon]$ affine extractor is a deterministic function $f : \{0,1\}^n \to \{0,1\}^m$ such that whenever X is the uniform distribution over some affine subspace over $\mathbb{F}_2^n$ with dimension ρn, we have that for every $z \in \{0,1\}^m$,

$$|\Pr[f(X) = z] - 2^{-m}| < \epsilon.$$

Note that we bound the error by the ℓ^∞ norm instead of the traditional ℓ^1 norm, as in [BSZ11]. We will let λ denote the entropy loss rate, i.e., $\lambda = 1 - \frac{m}{\rho n}$. In this paper we will focus on $[n, (1-\lambda)\frac{2}{3}n, \frac{2}{3}, 2^{-m}]$ affine extractors and ideally we would like λ to be as small as possible (e.g., close to 0). We note that it is straightforward to show by the probabilistic method that such extractors exist for any constant $\lambda > 0$. However the state of art constructions only achieve $\lambda \approx \frac{3}{4}$.

Now we define the duality measure of two sets as in [BSZ11].

Definition 1.7. [BSZ11] Given two sets $A, B \subseteq \mathbb{F}_2^n$, their duality measure is defined as

$$\mu^\perp(A, B) = \left| E_{a \in A, b \in B}[(-1)^{\langle a,b \rangle}] \right|.$$

The following conjecture is introduced in [BSZ11] and is shown in that paper to be implied by the well-known Polynomial Freiman-Ruzsa Conjecture in additive combinatorics.

Conjecture 1.8. *(Approximate Duality (ADC)) [BSZ11] For every pair of constants $\alpha, \delta > 0$ there exist a constant $\zeta > 0$ and an integer r, both depending on α and δ such that the following holds for sufficiently large n. If $A, B \subseteq \mathbb{F}_2^n$ satisfy $|A|, |B| \geq 2^{\alpha n}$ and $\mu^\perp(A, B) \geq 2^{-\zeta n}$, then there exists a pair of subsets*

$$A' \subseteq A, A' \geq \frac{|A|}{2^{\delta n + 1}} \text{ and } B' \subseteq B, |B'| \geq \left(\frac{\mu^\perp(A, B)}{2}\right)^r \cdot \frac{|B|}{2^{\delta n}}$$

such that $\mu^\perp(A', B') = 1$.

We now have the following theorems.

Theorem 1.9. *Assume the ADC conjecture and that we have an explicit $[n, m, \frac{2}{3}, 2^{-m}]$ affine extractor with $m = (1-\lambda)\frac{2}{3}n$, then there exists a constant $0 < \gamma < 1$ such that for any $n \in \mathbb{N}$, $k = \frac{3\lambda}{1+2\lambda}n$ and any $\epsilon > 2^{-\gamma n}$, there exists a semi-explicit (k, ϵ)-non-malleable extractor* nmExt : $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ *with $d = O(\log n + \log(1/\epsilon))$.*

Theorem 1.10. *Assume the ADC conjecture and we have an explicit* $[n, m, \frac{2}{3}, 2^{-m}]$ *affine extractor with* $m = (1-\lambda)\frac{2}{3}n$*, then for any* $n \in \mathbb{N}$ *and* $k = \frac{3\lambda}{1+2\lambda}n$*, there exists a semi-explicit* (k, ϵ)*-non-malleable extractor* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $d = \frac{3}{2+4\lambda}n$*,* $m = \Omega(n)$ *and* $\epsilon = 2^{-\Omega(n)}$*.*

Remark 1.11. In these two theorems, we use the term *semi-explicit* to mean that the construction may run in time 2^n. It is semi-explicit in the sense that the running time is polynomial in the length of the truth table of the extractor (note that an exhaustive search takes time 2^{2^n}). If we have affine extractors with large output size so that $\lambda \to 0$, then we can essentially achieve $k = \alpha n$ for any constant $\alpha > 0$.

Remark 1.12. It is also shown in [BSZ11] that a weaker form of the ADC conjecture is true.

By plugging Theorem 1.5 into the protocol of Dodis and Wichs, we obtain a 2-round privacy amplification protocol with optimal entropy loss for min-entropy $k = (1/2 - \delta)n$.

Theorem 1.13. *There exist constants* $0 < \delta, \gamma < 1$ *such that for any* $n \in \mathbb{N}$ *and* $\epsilon > 2^{-\gamma n}$*, there exists an explicit 2-round privacy amplification protocol for min-entropy* $k \geq (1/2-\delta)n$ *with security parameter* $\log(1/\epsilon)$ *and entropy loss* $O(\log n + \log(1/\epsilon))$*, in the presence of an active adversary.*

Note that this protocol is truly optimal in both round complexity and entropy loss, since [DW09] shows that there can be no one-round protocol for $k < n/2$. As a comparison, for $k > n/2$ we do have one-round protocols, although the entropy loss is quite large.

2 Overview of The Constructions and Techniques

In this section we give an overview of our constructions and the techniques used. In order to give a clean description, we shall be informal and imprecise sometimes.

All of our constructions are based on the inner product function. Especially, we are going to make extensive use of the fact that the inner product function is a *bilinear function*. Note that the inner product function is a good strong extractor. In fact, it is also a good two-source extractor. For two independent sources on n bits, it works as long as the sum of the entropies of the two sources is greater than n. However, at first this function does not seem to be a good candidate for a non-malleable extractor. To see this, consider the inner product function over $\mathbb{F}_2$. Let X be a source that is obtained by concatenating the bit 0 with U_{n-1}, and let Y be an independent uniform seed over $\{0,1\}^n$. Now for any $y \in \{0,1\}^n$, let $\mathcal{A}(y)$ be y with the first bit flipped. Thus we see that for all x in the support of X, one has $\langle x, y \rangle = \langle x, \mathcal{A}(y) \rangle$. Therefore, the inner product function is not a non-malleable extractor even for weak sources with min-entropy $k = n - 1$.

Let IP denote the inner product function. In the above example, we have that for all x in the support of X, $\mathsf{IP}(x, y) = \mathsf{IP}(x, \mathcal{A}(y))$. Or equivalently, $\mathsf{IP}(x, y) \oplus \mathsf{IP}(x, \mathcal{A}(y)) = 0$. Since IP is bilinear, this means that $\mathsf{IP}(x, y + \mathcal{A}(y)) = 0$. How does this happen? Looking closely at this example, our key observation is that this is because the range of Y is *too large*. Indeed, in this example the range of Y is the entire $\{0,1\}^n$, thus for any y the adversary can choose a different $\mathcal{A}(y)$ such that $y + \mathcal{A}(y) = 10\cdots0$ so that $\forall x \in \mathsf{Supp}(X), \mathsf{IP}(x, y + \mathcal{A}(y)) = 0$.

This observation suggests that we should choose the range of Y to be a small subset $S \subset \{0,1\}^n$, so that for some y's, the adversary will be unable to choose the appropriate $\mathcal{A}(y)$ from S. In other words, we take a shorter seed length l, choose an independent and uniform $y \in \{0,1\}^l$ and map y

to an element in $\{0,1\}^n$. This mapping is essentially an encoding. Now let us see what properties we need the encoding to have.

We will start with a construction that works for min-entropy $k > n/2$. Assume that we have a weak source X with min-entropy $k = (1/2+\delta)n$ for some constant $\delta > 0$. We take an independent and uniform $y \in \{0,1\}^l$ and encode y to $\bar{y} \in \{0,1\}^n$. For any adversarial function $\mathcal{A}$, let $\bar{y}'$ be the encoding of $\mathcal{A}(y)$. We will use an injective encoding, so that $\forall y, \bar{y}' \neq \bar{y}$. The output of the non-malleable extractor is then $\mathsf{IP}(X, \bar{Y})$.

To show that $\mathsf{IP}(X, \bar{Y})$ is a non-malleable extractor, it suffices to show that $\mathsf{IP}(X, \bar{Y})$ is close to uniform, and that $\mathsf{IP}(X, \bar{Y}) \oplus \mathsf{IP}(X, \bar{Y}')$ is close to uniform. The first part is easy. If X has min-entropy $k > n/2$, then we can take Y to be the uniform distribution over some $l \geq n/2$ bits. Since the encoding is injective, $\bar{Y}$ will have min-entropy $l \geq n/2$. Thus $\mathsf{IP}(X, \bar{Y})$ is close to uniform. For the second part, note that $\mathsf{IP}(X, \bar{Y}) \oplus \mathsf{IP}(X, \bar{Y}') = \mathsf{IP}(X, \bar{Y} + \bar{Y}')$. Thus now we need $\bar{Y} + \bar{Y}'$ to have a large min-entropy. Indeed, in the above counterexample where $l = n$, the adversary can choose $\mathcal{A}$ such that $\bar{Y} + \bar{Y}'$ is always equal to $10\cdots 0$ and thus has entropy 0. Now when we take $l < n$ and map $\{0,1\}^l$ to $S \subset \{0,1\}^n$, we want $\bar{Y} + \bar{Y}'$ to have a large support size.

The ideal case would be that $\bar{Y} + \bar{Y}'$ also has support size $|S| = 2^l$. This can be achieved if the encoding has the following property: for every two different y_1, y_2, we have that $\bar{y}_1 + \bar{y}_1' \neq \bar{y}_2 + \bar{y}_2'$, or equivalently, $\bar{y}_1 + \bar{y}_1' + \bar{y}_2 + \bar{y}_2' \neq 0$. Indeed, if this is true then $\bar{Y} + \bar{Y}'$ also has min-entropy $l \geq n/2$, and thus $\mathsf{IP}(X, \bar{Y}) \oplus \mathsf{IP}(X, \bar{Y}')$ is close to uniform. Looking carefully at this property, we see that it can be ensured (at least almost ensured, as we will explain shortly) if we have another property: the elements in S (when viewed as vectors in $\mathbb{F}_2^n$) are 4-wise linearly independent. Indeed, assume that the elements in S are 4-wise linearly independent. Then if $\bar{y}_1 + \bar{y}_1' + \bar{y}_2 + \bar{y}_2' = 0$, the only possible situation is that $\bar{y}_1' = \bar{y}_2$ and $\bar{y}_2' = \bar{y}_1$. Thus there cannot be three different y_1, y_2, y_3 such that $\bar{y}_1 + \bar{y}_1' = \bar{y}_2 + \bar{y}_2' = \bar{y}_3 + \bar{y}_3'$. Thus the min-entropy of $\bar{Y} + \bar{Y}'$ is at least $l - 1$.

So now the question is to explicitly find a large subset $S \subset \{0,1\}^n$ such that the elements in S are 4-wise linearly independent. Note that in particular this implies that the sum of any two different pairs of elements in S cannot be the same. Thus we have $\binom{|S|}{2} \leq 2^n$. Therefore $|S|$ can be at most roughly $2^{n/2}$. On the other hand, in order to work for any min-entropy $k > n/2$, we will need $l \geq n/2$ and thus $|S| = 2^l \geq 2^{n/2}$. These are very tight upper and lower bounds. Luckily, we have explicit constructions that meet these bounds. We will think of the elements in S as columns in a parity check matrix of some binary linear code. Thus we basically need a code with block length $2^{n/2}$ and message length $2^{n/2} - n$. The 4-wise linearly independent property basically is equivalent to saying that the code has distance at least 5. This is precisely the $[2^{n/2}, 2^{n/2} - n, 5]$-BCH code. Note that although the parity check matrix has $2^{n/2}$ columns, each column is (a, a^3) for a different element $a \in \mathbb{F}^*_{2^{n/2}}$. Thus the encoding from y to $\bar{y}$ can be computed efficiently.

Once we have the encoding, we can choose $l = n/2$ and we know that $\bar{Y}$ has min-entropy l and $\bar{Y} + \bar{Y}'$ has min-entropy $l - 1$. Now it is straightforward to show that both $\mathsf{IP}(X, \bar{Y})$ and $\mathsf{IP}(X, \bar{Y} + \bar{Y}')$ are close to uniform. Thus we obtain a non-malleable extractor that works for min-entropy $k > n/2$.

2.1 Achieving min-entropy $k < n/2$

Next we show how we can improve the above construction to achieve min-entropy $k < n/2$. To this end, we borrow ideas from [Bou05], where the first and the only known unconditional explicit constructions of two source extractors for min-entropy $k < n/2$ were given.

Specifically, let X be a distribution over some vector space $\mathbb{F}_q^n$ and let cX be the distribution obtained by sampling $x_1, x_2, \cdots, x_c$ from c independent copies of X and computing $\sum x_i$. By Fourier analysis and the Cauchy-Schwarz inequality one can show that in order to prove $\mathsf{IP}(X,Y)$ is close to uniform, it suffices to prove that $\mathsf{IP}(cX,Y)$ is close to uniform with a smaller error, for some integer $c > 1$. In [Bou05], Bourgain showed that for a weak source X with min-entropy rate $1/2 - \delta$ for some constant $\delta > 0$, one can encode X to $\mathsf{Enc}(X)$ such that $3\mathsf{Enc}(X)$ is close to having min-entropy rate $1/2 + \delta$. He then used this encoding together with the inner product function to construct a two source extractor that works for min-entropy rate $1/2 - \delta$.

Here we want to do the same thing. When given a source X with min-entropy rate $1/2 - \delta$, we encode X using Bourgain's encoding, and we encode the seed Y using the parity check matrix we discussed before. The non-malleable extractor is given as $\mathsf{nmExt}(X,Y) = \mathsf{IP}(\mathsf{Enc}(X), \bar{Y})$. Thus we see that $3\mathsf{Enc}(X)$ is close to having min-entropy rate $1/2 + \delta$, and both $\bar{Y}$ and $\bar{Y} + \bar{Y}'$ have min-entropy rate roughly $1/2$. Therefore both $\mathsf{IP}(3\mathsf{Enc}(X), \bar{Y})$ and $\mathsf{IP}(3\mathsf{Enc}(X), \bar{Y} + \bar{Y}')$ are close to uniform. We thus conclude that both $\mathsf{IP}(\mathsf{Enc}(X), \bar{Y})$ and $\mathsf{IP}(\mathsf{Enc}(X), \bar{Y} + \bar{Y}')$ are close to uniform, and we obtain a non-malleable extractor with 1 bit output for min-entropy $k = (1/2 - \delta)n$.

To give our first construction that outputs $\Omega(n)$ bits, we use a different encoding for Y. In this case our construction is essentially Bourgain's two-source extractor. Given an (n,k)-source X with $k = (1/2 - \delta)n$, we treat X as an element in the field $\mathbb{F}_p$, for an $n+1$-bit prime p. We next take a uniform independent seed $Y \in \{0,1\}^n$ and also treat Y as an element in $\mathbb{F}_p$. We encode X to (X, X^2) and Y to (Y, Y^2), viewed as vectors in $\mathbb{F}_p^2$. The non-malleable extractor is given as $\mathsf{nmExt}(X,Y) = \mathsf{IP}((X,X^2),(Y,Y^2)) \bmod M$ for an appropriately chosen integer M, and the inner product is taken over $\mathbb{F}_p^2$.

Here the argument for the non-malleability is different. Since the output has multiple bits, we need to use a non-uniform XOR lemma. Specifically, we choose the characters $e_r(s) = e^{2\pi i r s/p}$. Let $Y' = \mathcal{A}(Y)$, $Z = XY + X^2Y^2$ and $Z' = XY' + X^2(Y')^2$, we need to show that for any non-trivial character e_t and any character $e_{t'}$, $|E_{X,Y}[e_t(Z)e_{t'}(Z')]|$ is bounded. Simple calculations show that when $t \neq 0$, $e_t(Z)e_{t'}(Z') = e_t(Z'')$, where $Z'' = \mathsf{IP}((X,X^2),(Y + rY', Y^2 + r(Y')^2))$ and $r = t'/t$. Thus it suffices to show that $\mathsf{IP}((X,X^2),(Y + rY', Y^2 + r(Y')^2))$ is close to uniform.

As in [Bou05], one can show that $3(X, X^2)$ is close to having min-entropy rate $1/2 + \delta$. If $t' = 0$ then $r = 0$ and $(Y + rY', Y^2 + r(Y')^2)) = (Y, Y^2)$, which has min-entropy rate roughly $1/2$. Thus $\mathsf{IP}((X,X^2),(Y + rY', Y^2 + r(Y')^2))$ is close to uniform. If $t' \neq 0$ then $r \neq 0$ and we show that $(Y + rY', Y^2 + r(Y')^2))$ has roughly the same min-entropy as Y (at least the min-entropy of Y minus 1). Thus $\mathsf{IP}((X,X^2),(Y + rY', Y^2 + r(Y')^2))$ is still close to uniform. We further show that the error is $2^{-\Omega(n)}$. Therefore by the non-uniform XOR lemma Lemma 3.4 we obtained a non-malleable extractor with $\Omega(n)$ bits of output.

2.2 Achieving any constant min-entropy rate

In [BSZ11], Ben-Sasson and Zewi showed that affine extractors with large output size can be used to construct two source extractors with min-entropy rate $< 1/2$. Their "preimage construction" can potentially achieve any constant min-entropy rate. We show that their techniques combined with ours can also potentially lead to non-malleable extractors for any constant min-entropy rate. Specifically, they showed that if we have an affine extractor with large output size, then there is an injective mapping $F : \{0,1\}^n \to \{0,1\}^{n'}$ that maps $\{0,1\}^n$ into the preimage of a certain output of the affine extractor, such that for any weak source X with a certain amount of min-entropy, $F(\mathsf{Supp}(X))$ is not contained in any affine subspace of dimension say $0.51n'$. Thus, we can take an independent uniform seed $Y \in \{0,1\}^{n'/2}$ and construct a non-malleable extractor $\mathsf{nmExt}(X,Y) = \mathsf{IP}(F(X),\bar{Y})$, where $\bar{Y}$ is encoded using the parity check matrix of a suitable BCH-code as before. Since both $\bar{Y}$ and $\bar{Y}+\bar{Y}'$ have min-entropy roughly $n'/2$, we have that both $\mathsf{IP}(F(X),\bar{Y})$ and $\mathsf{IP}(F(X),(\bar{Y}+\bar{Y}'))$ are non-constant. Next, similar as in [BSZ11], we use the ADC conjecture to argue that in fact both $\mathsf{IP}(F(X),\bar{Y})$ and $\mathsf{IP}(F(X),(\bar{Y}+\bar{Y}'))$ are close to uniform. Therefore we obtain a non-malleable extractor.

2.3 Reducing seed length

In all the constructions where we encode the seed Y by a parity check matrix, the seed length is linear in the source length. However the error is also $2^{-\Omega(n)}$. If we only need to achieve a bigger error, we can reduce the seed length by using the parity check matrix of a BCH code with larger distance. Specifically, when the distance is $2t+1$ the seed length is roughly n/t. However we need to guarantee something else. For example, in the construction for min-entropy $k > n/2$, we need to show that both $\mathsf{IP}(X,\bar{Y})$ and $\mathsf{IP}(X,(\bar{Y}+\bar{Y}'))$ are still close to uniform. This can be shown as follows. Since now the columns of the parity check matrix are $2t$-wise linearly independent, both $\frac{t}{2}\bar{Y}$ and $\frac{t}{2}(\bar{Y}+\bar{Y}')$ will now have min-entropy roughly $\frac{t}{2}H_\infty(Y) = n/2$. Thus we can conclude that both $\mathsf{IP}(X,\frac{t}{2}\bar{Y})$ and $\mathsf{IP}(X,\frac{t}{2}(\bar{Y}+\bar{Y}'))$ are close to uniform, and therefore both $\mathsf{IP}(X,\bar{Y})$ and $\mathsf{IP}(X,(\bar{Y}+\bar{Y}'))$ are also close to uniform, by the Cauchy-Schwarz inequality. However the error increases according to the seed length. Calculations show that we can get seed length $d = O(\log n + \log(1/\epsilon))$.

2.4 Increasing output size

We can also increase the output size to $\Omega(n)$ for all our constructions with 1 bit output. To do this, note that we encode the seed Y by using the columns of a parity check matrix of a BCH code. Equivalently, the encoding is that $\bar{Y} = (Y, Y^3)$ when we use a field $\mathbb{F}_{2^l}$ with $l = \Theta(n)$ and Y is viewed as an element in $\mathbb{F}^*_{2^l}$. Now treat $\mathbb{F}_{2^l}$ as the vector space $\mathbb{F}_2^l$ and we take l elements $b_1, \cdots, b_l \in \mathbb{F}_{2^l}$ that corresponds to a basis of $\mathbb{F}_2^l$. Now for each b_i we define one bit $Z_i = \mathsf{IP}(\mathsf{Enc}(X), b_i\bar{Y})$.

We then show that $\{Z_i\}$ satisfy the conditions of a non-uniform XOR lemma, Lemma 3.3. Specifically, let $Z'_i = \mathsf{IP}(\mathsf{Enc}(X), b_i\bar{Y}')$ where $Y' = \mathcal{A}(Y)$. For any non-empty subset $S_1 \subset [l]$ and any subset $S_2 \subset [l]$, by the linearity of the inner product function, the xor of Z_i's where $i \in S_1$ and Z'_j's where $j \in S_2$ is of the form $\mathsf{IP}(\mathsf{Enc}(X), t_1\bar{Y} + t_2\bar{Y}')$, with $t_1, t_2 \in \mathbb{F}_{2^l}$. Since S_1 is non-empty we have $t_1 \neq 0$. We then show that $t_1\bar{Y} + t_2\bar{Y}'$ roughly has the same min-entropy as Y (at least the min-entropy of Y minus $\log 3$). Now since for example $3\mathsf{Enc}(X)$ is close to having min-entropy rate $1/2+\delta$, we conclude that $\mathsf{IP}(\mathsf{Enc}(X), t_1\bar{Y} + t_2\bar{Y}')$ is close to uniform. We further show that the error is $2^{-\Omega(n)}$. Thus by Lemma 3.3 we can output $m = \Omega(n)$ bits with error $2^{-\Omega(n)}$.

Organization. The rest of the paper is organized as follows. We give some preliminaries in Section 3. Next, to illustrate our ideas, in Section 4 we give a construction of a non-malleable extractor for $k > n/2$, using the inner product function. In Section 5 we give non-malleable extractors for $k = (1/2 - \delta)n$. In Section 6 we give non-malleable extractors that can potentially achieve any constant min-entropy rate. In Section 7 we briefly describe how we can reduce the seed length by using a BCH-code with larger distance, and how we can increase the output size for all constructions with one bit output. Finally in Section 8 we conclude with some open problems.

3 Preliminaries

We often use capital letters for random variables and corresponding small letters for their instantiations. Let $|S|$ denote the cardinality of the set S. Let $\mathbb{Z}_r$ denote the cyclic group $\mathbb{Z}/(r\mathbb{Z})$, and let $\mathbb{F}_q$ denote the finite field of size q. All logarithms are to the base 2.

3.1 Probability distributions

Definition 3.1 (statistical distance)**.** Let W and Z be two distributions on a set S. Their *statistical distance* (variation distance) is

$$\Delta(W, Z) \stackrel{def}{=} \max_{T \subseteq S}(|W(T) - Z(T)|) = \frac{1}{2}\sum_{s \in S}|W(s) - Z(s)|.$$

We say W is ε-close to Z, denoted $W \approx_\varepsilon Z$, if $\Delta(W, Z) \leq \varepsilon$. For a distribution D on a set S and a function $h : S \to T$, let $h(D)$ denote the distribution on T induced by choosing x according to D and outputting $h(x)$. We often view a distribution as a function whose value at a sample point is the probability of that sample point. Thus $\|W - Z\|_{\ell^1}$ denotes the ℓ_1 norm of the difference of the distributions specified by the random variables W and Z, which equals $2\Delta(W, Z)$.

Definition 3.2. A function $\mathsf{TExt} : \{0,1\}^{n_1} \times \{0,1\}^{n_2} \to \{0,1\}^m$ is a *strong two source extractor* for min-entropy k_1, k_2 and error ϵ if for every independent (n_1, k_1) source X and (n_2, k_2) source Y,

$$|(\mathsf{TExt}(X, Y), X) - (U_m, X)| < \epsilon$$

and

$$|(\mathsf{TExt}(X, Y), Y) - (U_m, Y)| < \epsilon,$$

where U_m is the uniform distribution on m bits independent of (X, Y).

3.2 Fourier analysis

We give some basic and standard facts about Fourier analysis here. We normalize as in [DLWZ11]. For functions f, g from a set S to $\mathbb{C}$, we define the inner product $\langle f, g\rangle = \sum_{x \in S} f(x)\overline{g(x)}$. Let D be a distribution on S, sometimes we will also view it as a function from S to $\mathbb{R}$. Note that $\mathrm{E}_D[f(D)] = \langle f, D\rangle$. Now suppose we have functions $h : S \to T$ and $g : T \to \mathbb{C}$. Then

$$\langle g \circ h, D\rangle = \mathrm{E}_D[g(h(D))] = \langle g, h(D)\rangle.$$

Let G be a finite abelian group, we say ϕ is a character of G if it is a homomorphism from G to $\mathbb{C}^\times$. We call the character that maps all elements to 1 the trivial character. Define the Fourier coefficient $\widehat{f}(\phi) = \langle f, \phi\rangle$, and let $\widehat{f}$ denote the vector with entries $\widehat{f}(\phi)$ for all ϕ. Note that for a distribution D, one has $\widehat{D}(\phi) = \mathrm{E}_D[\phi(D)]$.

Since the characters divided by $\sqrt{|G|}$ form an orthonormal basis, the inner product is preserved up to scale: $\langle \widehat{f}, \widehat{g}\rangle = |G|\langle f, g\rangle$. As a corollary, we obtain Parseval's equality:

$$\|\widehat{f}\|_{\ell^2}^2 = \langle \widehat{f}, \widehat{f}\rangle = |G|\langle f, f\rangle = |G|\|f\|_{\ell^2}^2.$$

Hence by Cauchy-Schwarz,

$$\|f\|_{\ell^1} \le \sqrt{|G|}\|f\|_{\ell^2} = \|\widehat{f}\|_{\ell^2} \le \sqrt{|G|}\|\widehat{f}\|_{\ell^\infty}. \tag{1}$$

For functions $f, g : S \to \mathbb{C}$, we define the function $(f,g) : S \times S \to \mathbb{C}$ by $(f,g)(x,y) = f(x)g(y)$. Thus, the characters of the group $G \times G$ are the functions (ϕ, ϕ'), where ϕ and ϕ' range over all characters of G. We abbreviate the Fourier coefficient $\widehat{(f,g)}((\phi,\phi'))$ by $\widehat{(f,g)}(\phi,\phi')$. Note that

$$\widehat{(f,g)}(\phi,\phi') = \sum_{(x,y)\in G\times G} f(x)g(y)\phi(x)\phi'(y) = \left(\sum_{x\in G} f(x)\phi(x)\right)\left(\sum_{y\in G} g(x)\phi'(x)\right) = \widehat{f}(\phi)\widehat{g}(\phi').$$

In this paper, in the additive group of $\mathbb{F}_p$ we use the characters $e_r(s) = e^{2\pi i rs/p}$ for $r \in \mathbb{F}_p$. It is easy to verify that $\{e_r, r \in F_p\}$ indeed are characters and these characters divided by $\sqrt{p}$ form an orthonormal basis. Note that the trivial character corresponds to the case $r = 0$.

We next generalize the characters to the additive group of the field $\mathbb{F}_{p^l}$. In this case, for any $r \in \mathbb{F}_{p^l}$, we use the character $e_r(s) = e^{2\pi i (r\cdot s)/p}$, where r and s are viewed as vectors in $\mathbb{F}_p^l$ and $\cdot$ indicates the inner product function in $\mathbb{F}_p^l$. Again it is easy to verify that these indeed are characters and they form an orthonormal basis (up to a normalization factor of $p^{l/2}$).

3.3 Non-uniform XOR lemma

The following non-uniform XOR lemmas are proved in [DLWZ11].

Lemma 3.3. *Let (W, W') be a random variable on $G \times G$ for a finite abelian group G, and suppose that for all characters ψ, ψ' on G with ψ nontrivial, one has*

$$|\,\mathrm{E}_{(W,W')}[\psi(W)\psi'(W')]| \le \epsilon.$$

Then the distribution of (W, W') is $\epsilon|G|$ close to (U, W'), where U is the uniform distribution on G independent of W'. Moreover, for $f : G \times G \to \mathbb{R}$ defined as the difference of distributions $(W, W') - (U, W')$, we have $\|\widehat{f}\|_{\ell^\infty} \le \epsilon$.

Lemma 3.4. *For every cyclic group $G = \mathbb{Z}_N$ and every integer $M \le N$, there is an efficiently computable function $\sigma : \mathbb{Z}_N \to \mathbb{Z}_M = H$ such that the following holds. Let (W, W') be a random variable on $G \times G$, and suppose that for all characters ψ, ψ' on G with ψ nontrivial, one has*

$$|\,\mathrm{E}_{(W,W')}[\psi(W)\psi'(W')]| \le \epsilon.$$

Then the distribution $(\sigma(W), \sigma(W'))$ is $O(\epsilon M \log N + M/N)$-close to the distribution (U, W') where U stands for the uniform distribution over H independent of W'.

3.4 Strong non-malleable extractor

The following theorem is proved in [Rao07].

Theorem 3.5. *[Rao07] Let $\mathsf{TExt} : \{0,1\}^{n_1} \times \{0,1\}^{n_2} \to \{0,1\}^m$ be any two source extractor for min-entropy k_1, k_2 with error ϵ. Then if X is an (n_1, k_1) source and Y is an independent (n_1, k_2') source, we have*

$$|(\mathsf{TExt}(X,Y), Y) - (U_m, Y)| \le 2^m(2^{k_2 - k_2' + 1} + \epsilon).$$

Here we prove a similar theorem that will enable our non-malleable extractor to be "strong".

Theorem 3.6. *Let* $\mathsf{TExt} : \{0,1\}^{n_1} \times \{0,1\}^{n_2} \to \{0,1\}^m$ *be a two source extractor for min-entropy* k_1, k_2 *and* $\mathcal{A} : \{0,1\}^{n_2} \to \{0,1\}^{n_2}$ *be a deterministic function such that for any* (n_1, k_1) *source* X *and any independent* (n_2, k_2) *source* Y,

$$|(\mathsf{TExt}(X,Y), \mathsf{TExt}(X, \mathcal{A}(Y))) - (U_m, \mathsf{TExt}(X, \mathcal{A}(Y)))| \leq \epsilon.$$

Then for any (n_2, k_2') *source* Y' *independent of* X,

$$|(\mathsf{TExt}(X,Y'), \mathsf{TExt}(X, \mathcal{A}(Y')), Y') - (U_m, \mathsf{TExt}(X, \mathcal{A}(Y')), Y')| \leq 2^{2m}(2^{k_2 - k_2' + 1} + \epsilon).$$

Proof. Let $W = \mathsf{TExt}(X,Y)$ and $W' = \mathsf{TExt}(X, \mathcal{A}(Y))$. For any $(z, z') \in \{0,1\}^m \times \{0,1\}^m$, define the set of bad y's for (z, z') to be

$$B_{z,z'} = \{y : |\Pr[W = z, W' = z'] - 2^{-m}\Pr[W' = z']| > \epsilon\}.$$

Then we must have

Claim 3.7. *For every* (z, z'), $|B_{z,z'}| < 2 \cdot 2^{k_2}$.

To see this, assume for the sake of contradiction that $|B_{z,z'}| \geq 2 \cdot 2^{k_2}$ for some (z, z'). Let

$$B^+_{z,z'} = \{y : \Pr[W = z, W' = z'] - 2^{-m}\Pr[W' = z'] > \epsilon\}$$

and

$$B^-_{z,z'} = \{y : \Pr[W = z, W' = z'] - 2^{-m}\Pr[W' = z'] < -\epsilon\}.$$

Then $|B_{z,z'}| = |B^+_{z,z'}| + |B^-_{z,z'}|$ and thus one of them must have size $\geq 2^{k_2}$. Without loss of generality assume that $|B^+_{z,z'}| \geq 2^{k_2}$. Then we can let Y to be the uniform distribution over $|B^+_{z,z'}|$ and Y is independent of X, but $|(W, W') - (U, W')| > \epsilon$, which is a contradiction.

Let $B = \cup_{z,z'} B_{z,z'}$. We have $|B| < 2^{2m} \cdot 2 \cdot 2^{k_2} = 2^{2m+1}2^{k_2}$. Now we can bound $|(W, W', Y') - (U, W', Y')|$ when Y' is an independent (n_2, k_2') source, as follows.

$$\begin{aligned}
&|(W, W', Y') - (U, W', Y')| \\
\leq &\sum_{y \in \mathsf{Supp}(Y')} 2^{-k_2'} |(W, W')|_{Y'=y} - (U, W')|_{Y'=y}| \\
= &\sum_{y \in \mathsf{Supp}(Y') \cap B} 2^{-k_2'} |(W, W')|_{Y'=y} - (U, W')|_{Y'=y}| + \sum_{y \in \mathsf{Supp}(Y') \setminus B} 2^{-k_2'} |(W, W')|_{Y'=y} - (U, W')|_{Y'=y}| \\
< &2^{-k_2'} 2^{2m+1} 2^{k_2} + 2^{2m}\epsilon \\
= &2^{2m}(2^{k_2 - k_2' + 1} + \epsilon).
\end{aligned}$$

■

3.5 Basic properties of the inner product function

Here we prove some basic properties of the inner product function.

Lemma 3.8. *Let $\mathbb{F}_p$ be a field and X, Y be two independent random variables over $\mathbb{F}_p^l$. Assume that X has min-entropy k_1 and Y has min-entropy k_2. Let $Z = \mathsf{IP}(X, Y) = X \cdot Y$ be the inner product function where the operation is in $\mathbb{F}_p$. For any non-trivial character e_r where $r \in \mathbb{F}_p$,*

$$|E_{X,Y}[e_r(Z)]|^2 \leq p^l 2^{-(k_1+k_2)}.$$

Proof. Note that if a weak random source W has min-entropy k, then $\|W\|_{\ell^\infty} \leq 2^{-k}$, and $\|W\|_{\ell^2}^2 = \sum_w (\Pr[W = w])^2 \leq 2^{-k} \sum_w \Pr[W = w] = 2^{-k}$.

For a fixed $Y = y$,

$$E_X[e_r(x \cdot y)] = E_X[e_{ry}(X)] = \langle e_{ry}, X \rangle = \overline{\widehat{X}(e_{ry})}.$$

Thus

$$E_{X,Y}[e_r(Z)] = E_Y[E_X[e_r(x \cdot y)]] = E_Y[\overline{\widehat{X}(e_{ry})}] = \langle Y, \widehat{X} \rangle.$$

Therefore by Cauchy-Schwartz,

$$\begin{aligned}(E_{X,Y}[e_r(Z)])^2 &\leq \langle Y, Y \rangle \cdot \langle \widehat{X}, \widehat{X} \rangle \\ &= \|Y\|_{\ell^2}^2 \|\widehat{X}\|_{\ell^2}^2 = p^l \|Y\|_{\ell^2}^2 \|X\|_{\ell^2}^2 \\ &\leq p^l 2^{-k_1} 2^{-k_2} = p^l 2^{-(k_1+k_2)}.\end{aligned}$$

□

Now for any weak random source W, we let $2W = W + W$ stand for the distribution that is obtained by first sampling w_1, w_2 from two independent and identical distributions according to W, and then computing $w_1 + w_2$. Similarly $W - W$ is obtained by first sampling w_1, w_2 and then computing $w_1 - w_2$. Similarly we define cW to be the distribution by sampling w_i from c independent and identical distributions according to W, and then computing the sum. We now have the following lemma.

Lemma 3.9. *Let X, Y be two independent random variables over $\mathbb{F}_p^l$. For any two integers c_1, c_2, let $X_{c_1} = 2^{c_1} X - 2^{c_1} X$ and $Y_{c_2} = 2^{c_2} Y - 2^{c_2} Y$. Then for any non-trivial character ψ,*

$$|E_{X,Y}[\psi(X \cdot Y)]| \leq |E_{X_{c_1}, Y_{c_2}}[\psi(X_{c_1} \cdot Y_{c_2})]|^{1/2^{c_1+c_2+2}}.$$

Proof. First note

$$|E_{X,Y}[\psi(X \cdot Y)]| = |E_Y[E_X[\psi(X \cdot Y)]]| \leq E_Y |E_X[\psi(X \cdot Y)]|.$$

Note that $\psi(s) = e^{2\pi i r s/p}$ for some $r \in \mathbb{F}_p$. Thus by Jensen's inequality,

$$
\begin{aligned}
(E_{X,Y}[\psi(X\cdot Y)])^2 &\le E_Y|E_X[\psi(X\cdot Y)]|^2 = E_Y[E_X[\psi(X\cdot Y)]\overline{E_X[\psi(X\cdot Y)]}]\\
&= |E_Y \sum_{x_1,x_2} X(x_1)X(x_2)\psi((x_1-x_2)\cdot Y)|\\
&= |E_Y E_{X-X}[\psi((X-X)\cdot Y)]|\\
&= |E_{X_1,Y}[\psi(X_1\cdot Y)]|
\end{aligned}
$$

where $X_1 = X - X$.

Apply the above procedure again, we get that

$$
(E_{X,Y}[\psi(X\cdot Y)])^4 \le |E_{X_1,Y}[\psi(X_1\cdot Y)]|^2 \le |E_{X_2,Y}[\psi(X_2\cdot Y)]|,
$$

where $X_2 = X_1 - X_1 = 2X - 2X$.

Repeat the procedure for c_1 times, we get that

$$
(E_{X,Y}[\psi(X\cdot Y)])^{2^{c_1+1}} \le |E_{X_{c_1},Y}[\psi(X_{c_1}\cdot Y)]|,
$$

where $X_{c_1} = 2^{c_1}X - 2^{c_1}X$.

similarly, we can apply the argument to Y for another c_2 times, and we get

$$
(E_{X,Y}[\psi(X\cdot Y)])^{2^{c_1+c_2+2}} \le |E_{X_{c_1},Y_{c_2}}[\psi(X_{c_1}\cdot Y_{c_2})]|,
$$

where $X_{c_1} = 2^{c_1}X - 2^{c_1}X$ and $Y_{c_2} = 2^{c_2}Y - 2^{c_2}Y$. Thus the lemma is proved. □

3.6 Incidence theorems

We need the following theorems about point line incidences. For a field $\mathbb{F}$, we call a subset $\ell \subset F\times F$ a line if there exist $a, b \in \mathbb{F}$ such that $\ell = \{(x, ax+b)\}$ for all $x \in \mathbb{F}$. Let $P \subset F\times F$ be a set of points and L be a set of lines, we say that a point (x, y) has an incidence with a line ℓ if $(x, y) \in \ell$. The following theorem provides a bound on the number of incidences that can be generated from K points and K lines.

Theorem 3.10. *[BKT04, Kon03] There exist universal constants $\alpha > 0, 0.1 > \beta > 0$ such that for any field $\mathbb{F}_q$ where q is either prime or 2^p for p prime, if L, P are sets of K lines and K points respectively, with $K \le q^{2-\beta}$, the number of incidences $I(P, L) \le O(K^{3/2-\alpha})$.*

3.7 BCH codes

In this paper we will only focus on BCH codes over $\mathbb{F}_2$. Given two parameters $m, t \in \mathbb{N}$, a BCH code is a linear code with block length $n = 2^m - 1$, message length roughly $n - mt$ and distance $d \ge 2t+1$. Specifically, we have the following theorem.

Theorem 3.11. *For all integers m and t there exists an explicit $[n, n-mt, 2t+1]$-BCH code[3], with $n = 2^m - 1$.*

[3] In fact, the message length may not be exactly $n - mt$, but for simplicity we will assume that it is exactly $n - mt$. The small error does not affect our analysis. Also, for small t the message length is exactly $n - mt$.

Since a BCH code is a linear code, we can take its parity check matrix. Note that this is a $mt \times n$ matrix. Let α be a primitive element in $\mathbb{F}_{2^m}^*$, the i'th column of the parity check matrix is of the form $(\alpha^i, (\alpha^i)^3, (\alpha^i)^5, \cdots, (\alpha^i)^{2t-1})$, for $i = 0, 1, \cdots, n-1$. Since α is a generator in $\mathbb{F}_{2^m}^*$, equivalently, for $y \in \mathbb{F}_{2^m}^*$ we can think of the y'th column to be $(y, y^3, \cdots, y^{2t-1})$.

4 A Non-Malleable Extractor for Entropy Rate $> 1/2$

As a warm up, in this section we construct a non-malleable extractor for weak sources with min-entropy rate $> 1/2$, based on the inner product function over $\mathbb{F}_2$. To this end, we need to use the BCH code over $\mathbb{F}_2$. The construction is described as below.

Given an (n,k)-source X with $k = (1/2 + \delta)n$, we choose a BCH code with $t = 2$ and $m = n/2$, thus the block length is $n' = 2^{n/2} - 1$ and the parity check matrix is a $n \times (2^{n/2} - 1)$ matrix.

Take an independent uniform seed $Y \in \{0,1\}^{n/2-1}$ and let S_Y stand for the integer whose binary expression is Y. We encode Y to $\bar{Y}$ such that $\bar{Y} = \mathsf{Enc}(Y)$ is the S_Y'th column in the parity check matrix (i.e., $\bar{Y} = (Y, Y^3)$ when Y is viewed as an element in $\mathbb{F}_{2^{n/2}}^*$). Our non-malleable extractor is now defined as

$$\mathsf{nmExt}(X, Y) = \mathsf{IP}(X, \mathsf{Enc}(Y)) = \mathsf{IP}(X, \bar{Y}),$$

where IP is the inner product function over $\mathbb{F}_2$ and the output is just 1 bit.

To analyze our construction, we first have the following theorem.

Theorem 4.1. *[CG88, Vaz85] For every constant $\delta > 0$, if X is an (n, k_1) source, Y is an independent (n, k_2) source and $k_1 + k_2 \geq (1 + \delta)n$, then*

$$|(Y, \mathsf{IP}(X, Y)) - (Y, U)| < \epsilon$$

with $\epsilon = 2^{-\Omega(n)}$.

We now show that the construction is a non-malleable extractor.

Theorem 4.2. *For any constant $\delta > 0$, the function nmExt defined as above is a $((1/2+\delta)n, 2^{-\Omega(n)})$ non-malleable extractor.*

Proof. Let $Z = \mathsf{nmExt}(X, Y)$ and $Z' = \mathsf{nmExt}(X, Y')$ where $Y' = \mathcal{A}(Y)$ for any deterministic function $\mathcal{A}$ such that $\forall y, \mathcal{A}(y) \neq y$. To show the construction is a non-malleable extractor, by the xor lemma it suffices to show that

$$(Z, Y) \approx_\epsilon (U, Y)$$

and

$$(Z \oplus Z', Y) \approx_\epsilon (U, Y)$$

for some $\epsilon = 2^{-\Omega(n)}$.

Note that the BCH code has distance $2t + 1 = 5 > 4$, thus any 4 columns in the parity check matrix must be linearly independent. This in particular implies that every two different columns must be different. Thus $\bar{Y}$ has min-entropy $n/2 - 1$. Since $k + n/2 - 1 = n + \delta n - 1$, by Theorem 4.1 we have (note that there is a one to one correspondence between $\bar{Y}$ and Y)

$$(Z, Y) \approx_{\epsilon_1} (U, Y)$$

with $\epsilon_1 = 2^{-\Omega(n)}$. Thus we have that with probability $1 - \sqrt{\epsilon_1}$ over the fixing of $Y = y$,

$$|(Z|Y = y) - U| \leq \sqrt{\epsilon_1}.$$

Next, note that

$$Z \oplus Z' = \mathsf{IP}(X, \bar{Y}) \oplus \mathsf{IP}(X, \bar{Y'}) = \mathsf{IP}(X, \bar{Y} \oplus \bar{Y'}) = \mathsf{IP}(X, \bar{Y} + \bar{Y'}).$$

For two different y_1, y_2, if $\bar{y_1} + \bar{y_1}' = \bar{y_2} + \bar{y_2}'$, then $\bar{y_1}, \bar{y_2}, \bar{y_1}', \bar{y_2}'$ are linearly dependent. Note that $\bar{y_1}' = \mathsf{Enc}(y_1')$ and $\bar{y_2}' = \mathsf{Enc}(y_2')$ are also some columns of the parity check matrix. Since $y_1' \neq y_1$ and $y_2' \neq y_2$, we have that $\bar{y_1}' \neq \bar{y_1}$ and $\bar{y_2}' \neq \bar{y_2}$. Thus we must have $\bar{y_1}' = \bar{y_2}$ and $\bar{y_2}' = \bar{y_1}$.

Therefore, the min-entropy of $\bar{Y} + \bar{Y'}$ is at least $n/2 - 2$ since the probability of getting any particular element in the support is at most $2 \cdot 2^{-(n/2-1)} = 2^{-(n/2-2)}$. Since $k + n/2 - 2 = n + \delta n - 2$, by Theorem 4.1 we have

$$(Z \oplus Z', \bar{Y} + \bar{Y'}) \approx_{\epsilon_2} (U, \bar{Y} + \bar{Y'})$$

with $\epsilon_2 = 2^{-\Omega(n)}$. This means that with probability $1 - \sqrt{\epsilon_2}$ over the fixing of $\bar{Y} + \bar{Y'}$, $Z \oplus Z'$ is $\sqrt{\epsilon_2}$-close to uniform. Since $\bar{Y} + \bar{Y'}$ is a deterministic function of Y, this implies that with probability $1 - \sqrt{\epsilon_2}$ over the fixing of $Y = y$,

$$|(Z \oplus Z'|Y = y) - U| \leq \sqrt{\epsilon_2}.$$

Thus by the non-uniform xor lemma, Lemma 3.3, we have that with probability $1 - \sqrt{\epsilon_1} - \sqrt{\epsilon_2}$ over the fixing of $Y = y$,

$$|(Z, Z')|Y = y - (U, Z'|Y = y)| \leq 2max\{\sqrt{\epsilon_1}, \sqrt{\epsilon_2}\}.$$

Therefore we have that

$$|(Z, Z', Y) - (U, Z', Y)| \leq \epsilon,$$

where $\epsilon = 2max\{\sqrt{\epsilon_1}, \sqrt{\epsilon_2}\} + \sqrt{\epsilon_1} + \sqrt{\epsilon_2} = 2^{-\Omega(n)}$. ■

5 Non-Malleable Extractors for Entropy Rate $< 1/2$

In this section we give our main constructions, namely non-malleable extractors for weak sources with min-entropy rate $1/2 - \delta$ for some universal constant $\delta > 0$. We give two constructions, one that outputs one bit and one that outputs many bits.

5.1 A non-malleable extractor with 1 bit output

Given an (n, k)-source X with $k = (1/2 - \delta)n$, we first pick a prime p that is close to n. By Bertrand's postulate and Pierre Dusart's improvement, for every $n \geq 3275$, there exists a prime between n and $n(1 + \frac{1}{2\ln^2 n})$. We will pick a prime p in this range. Note that the prime can be found in polynomial time in n. Take the field $\mathbb{F}_q$ where $q = 2^p$ and let g be a generator in $\mathbb{F}_q^*$. The construction is as follows.

- Treat X as an element in $\mathbb{F}_q^*$ and encode X such that $\mathsf{Enc}(X) = (X, g^X)$.
- Take the parity check matrix of a $[2^p - 1, 2^p - 1 - 2p, 5]$-BCH code (note that the parity check matrix is a $2p \times (2^p - 1)$ matrix). Take an independent and uniform seed $Y \in \{0,1\}^{p-1}$ and let S_Y stand for the integer whose binary expression is Y. We encode Y to $\bar{Y}$ such that $\bar{Y}$ is the S_Y'th column in the parity check matrix (i.e., $\bar{Y} = (Y, Y^3)$ when Y is viewed as an element in $\mathbb{F}_{2^p}^*$).
- Output $\mathsf{nmExt}(X, Y) = \mathsf{IP}(\mathsf{Enc}(X), \bar{Y})$ where IP is the inner product function over $\mathbb{F}_2$.

To prove our construction is a non-malleable extractor, we are going to use the non-uniform XOR lemma. Specifically, we will first prove that $\mathsf{nmExt}(X, Y) \approx U$ and $\mathsf{nmExt}(X, Y) \oplus \mathsf{nmExt}(X, \mathcal{A}(Y)) \approx U$, for any function $\mathcal{A}$ such that $\forall y \in \{0,1\}^{p-1}, \mathcal{A}(y) \neq y$.

To this end, we first prove the following lemma.

Lemma 5.1. *There exists a constant $\delta > 0$ such that for any (n, k)-source X with $k = (1/2 - \delta)n$, and any independent $(2p, k_2)$ source Y with $k_2 \geq (1 - \delta)p$,*

$$|\mathsf{IP}(\mathsf{Enc}(X), Y) - U| \leq \epsilon,$$

where $\epsilon = 2^{-\Omega(n)}$.

Proof. We think of X as a distribution in $\mathbb{F}_q^*$ that has min-entropy k. This increases the error by at most 2^{-k} (for the element 0). By the XOR lemma, we only need to show that for the only non-trivial character ψ (since we only output 1 bit),

$$|E_{X,Y}[\psi(\mathsf{IP}(\mathsf{Enc}(X), Y))]| \leq 2^{-\Omega(n)}.$$

Let $X' = 4\mathsf{Enc}(X) - 4\mathsf{Enc}(X)$, by Lemma 3.9 we have

$$|E_{X,Y}[\psi(\mathsf{IP}(\mathsf{Enc}(X), Y))]| \leq |E_{X',Y}[\psi(X' \cdot Y)]|^{\frac{1}{8}}.$$

We next bound $|E_{X',Y}[\psi(X' \cdot Y)]|$. First we show that X' is close to a source with min-entropy rate $> 1/2$. We have the following claim.

Claim 5.2. *There is a universal constant $\delta > 0$ such that if X is any weak source with min-entropy $(1/2 - \delta)n$, $3\mathsf{Enc}(X)$ is $2^{-\Omega(n)}$-close to a source with min-entropy $(1/2 + \delta)(2p)$.*

Proof of the claim. Note that $k = (1/2 - \delta)n$ and p is between n and $n(1 + \frac{1}{2 \ln^2 n})$. Thus for sufficiently large n we have that $k \geq (1/2 - 1.01\delta)p$. Note that we choose the field $\mathbb{F}_q$ where $q = 2^p$. Thus the sum of $\mathsf{Enc}(X) + \mathsf{Enc}(X)$ when viewing $\mathsf{Enc}(X)$ as a vector in $\mathbb{F}_2^{2p}$ is the same as when viewing $\mathsf{Enc}(X)$ as a vector in $\mathbb{F}_q^2$. In the following we will view $\mathsf{Enc}(X)$ as a vector in $\mathbb{F}_q^2$. We show that $3\mathsf{Enc}(X)$ has a larger min-entropy rate.

First consider the distribution $2\mathsf{Enc}(X)$. Note that the distribution is of the form $(X + X, g^X + g^X)$. Let $\bar{X} = g^X$ and note that g^x is a bijection in F_q^*. Thus $\bar{X}$ has the same min-entropy as X. Now the support of $2\mathsf{Enc}(X)$ is of the form $(\log_g(\bar{x}_1 \bar{x}_2), \bar{x}_1 + \bar{x}_2)$. For any (b, a) in this support, we have that $\bar{x}_1 \bar{x}_2 = g^b$ and $\bar{x}_1 + \bar{x}_2 = a$. Thus there are at most 2 different pairs of $(\bar{x}_1, \bar{x}_2)$ that satisfy both equations. Therefore the min-entropy of $2\mathsf{Enc}(X)$ is at least $2H_\infty(X) - 1$. We can also assume that $a \neq 0$ since this only increases the error by at most $2^{-H_\infty(X)}$. Now let $k = H_\infty(X) - 1$, we have that $\mathsf{Enc}(X)$ has min-entropy at least k and $2\mathsf{Enc}(X)$ has min-entropy at least $2k$.

Now consider $3\mathsf{Enc}(X)$. Every element in the support of $3\mathsf{Enc}(X)$ has the form $(\log_g(\bar{x}_1\bar{x}_2\bar{x}_3), \bar{x}_1 + \bar{x}_2 + \bar{x}_3)$, which determines the point $(\bar{x}_1\bar{x}_2\bar{x}_3, \bar{x}_1 + \bar{x}_2 + \bar{x}_3)$. Let $a = \bar{x}_1 + \bar{x}_2$ and $b = \bar{x}_1\bar{x}_2$, this point is

$$(b\bar{x}_3, a + \bar{x}_3).$$

Let $\tilde{x}_3 = a + \bar{x}_3$, then

$$(a + \bar{x}_3, b\bar{x}_3) = (\tilde{x}_3, b\tilde{x}_3 - ab).$$

For a fixed $(a = \bar{x}_1 + \bar{x}_2, b = \bar{x}_1\bar{x}_2)$ define the line

$$\ell_{a,b} = \{(x, bx - ab) | x \in \mathbb{F}_q\}.$$

Thus we have a set of lines $L = \{\ell_{a,b}\}$. Note that $a \neq 0$ and $b \neq 0$. Thus for different (a, b), the line $\ell_{a,b}$ is also different. Note that x_3 is sampled from X_3, which has min-entropy k and (a, b) is sampled from $\mathsf{Enc}(X_1) + \mathsf{Enc}(X_2)$, which has min-entropy $2k$. Further note that these two distributions are independent. Since every weak source with min-entropy k is a convex combination of flat k sources, without loss of generality we can assume that X_3 and $\mathsf{Enc}(X_1) + \mathsf{Enc}(X_2)$ are both flat sources. Thus L has size 2^{2k}.

Now let α, β be the two constants in Theorem 3.10. Assume that $3\mathsf{Enc}(X)$ is ϵ-far from any source with min-entropy $(1 + \alpha/2)2k$. Since $3\mathsf{Enc}(X)$ determines the distribution $(A + \bar{X}_3, B\bar{X}_3)$, this distribution is also ϵ-far from any source with min-entropy $(1 + \alpha/2)2k$. Thus there must exist some set M of size at most $2^{(1+\alpha/2)2k}$ such that

$$\Pr_{(a,b)\leftarrow 2\mathsf{Enc}(X), x_3 \leftarrow X}[(a + \bar{x}_3, b\bar{x}_3) \in M] \geq \epsilon.$$

Note that whenever $(a + \bar{x}_3, b\bar{x}_3) \in M$, this point has an incidence with the line $\ell_{a,b}$. Further note that whenever (a, b) is different or x_3 is different, the incidence is also different. Thus by the above inequality the number of incidences between the set of points M and the set of lines L is at least

$$\Pr_{(a,b)\leftarrow 2\mathsf{Enc}(X), x_3 \leftarrow X}[(a + \bar{x}_3, b\bar{x}_3) \in M] 2^k 2^{2k} \geq \epsilon 2^{3k}.$$

On the other hand, since L has size 2^{2k} and M has size $2^{(1+\alpha/2)2k} \leq 2^{(1+\alpha/2)2(1/2-\delta)p} < 2^{(1+\alpha/2)p} \leq q^{2-\beta}$, by Theorem 3.10, the number of incidences between M and L is at most $O(2^{(3/2-\alpha)(2+\alpha)k}) < 2^{3k(1-\alpha/6)} = 2^{-\alpha k/2} 2^{3k}$.

Thus we must have $\epsilon < 2^{-\alpha k/2}$.

Thus we have shown that $3\mathsf{Enc}(X)$ is $2^{-\alpha k/2}$-close to having min-entropy $(1 + \alpha/2)2k$. By choosing δ appropriately, we get that $3\mathsf{Enc}(X)$ is $2^{-\Omega(n)}$-close to having min-entropy $(1/2+\delta)2p$. □

Now note that Y is a weak source over $\{0, 1\}^{2p}$ with min-entropy $k_2 \geq (1 - \delta)p$. Also note that the min-entropy of X' is at least the min-entropy of $3\mathsf{Enc}(X)$. Thus by Lemma 3.8 we have that

$$|E_{X',Y}[\psi(X' \cdot Y)]| \leq 2^{2p} 2^{-(1/2+\delta)2p} 2^{-(1-\delta)p} + 2^{-\Omega(n)} = 2^{-\Omega(n)}.$$

Therefore

$$|E_{X,Y}[\psi(\mathsf{IP}(\mathsf{Enc}(X), Y))]| \leq 2^{-\Omega(n)}.$$

□

Now we can prove our construction is a non-malleable extractor.

Theorem 5.3. *For any (n,k)-source X with $k=(1/2-\delta)n$, an independent seed Y and any deterministic function $\mathcal{A}$ such that $\forall y \in \{0,1\}^{p-1}, \mathcal{A}(y) \neq y$,*

$$|(\mathsf{nmExt}(X,Y), \mathsf{nmExt}(X,\mathcal{A}(Y)), Y) - (U, \mathsf{nmExt}(X,\mathcal{A}(Y)), Y)| \leq 2^{-\Omega(n)}$$

Proof. First let Y' be an independent source over $\{0,1\}^{p-1}$ with min-entropy $k_2 \geq (1-\delta)p+1$. Since the columns of the parity check matrix of the BCH code are 4-wise linearly independent, different y will be mapped to different $\bar{y}$. Thus $\bar{Y'}$ also has min-entropy $k_2 \geq (1-\delta)p+1$. By Lemma 5.1 we have

$$|\,\mathsf{nmExt}(X,Y') - U| \leq 2^{-\Omega(n)}.$$

Next, note that $\mathsf{nmExt}(X,Y') \oplus \mathsf{nmExt}(X,\mathcal{A}(Y')) = \mathsf{IP}(\mathsf{Enc}(X), \bar{Y'} + \bar{\mathcal{A}(Y')})$. By the same argument as in the proof of Theorem 4.2, $\bar{Y'} + \bar{\mathcal{A}(Y')}$ has min-entropy at least $k_2 - 1 \geq (1-\delta)p$. Thus again by Lemma 5.1 we have

$$|\,\mathsf{nmExt}(X,Y') \oplus \mathsf{nmExt}(X,\mathcal{A}(Y')) - U| \leq 2^{-\Omega(n)}.$$

Thus by the non-uniform XOR lemma, Lemma 3.3, we have

$$|(\mathsf{nmExt}(X,Y'), \mathsf{nmExt}(X,\mathcal{A}(Y'))) - (U, \mathsf{nmExt}(X,\mathcal{A}(Y')))| \leq 2^{-\Omega(n)}.$$

Now note that Y has min-entropy $p-1$, thus by Theorem 3.6,

$$|(\mathsf{nmExt}(X,Y), \mathsf{nmExt}(X,\mathcal{A}(Y)), Y) - (U, \mathsf{nmExt}(X,\mathcal{A}(Y)), Y)| \leq 2^2(2^{-\Omega(n)} + 2^{-\Omega(n)}) = 2^{-\Omega(n)}.$$

■

5.2 A non-malleable extractor with multiple bits output

Given an (n,k)-source X with $k=(1/2-\delta)n$, we first pick a prime p such that $2^n < p < 2^{n+1}$. By Bertrand's postulate, there is always such a prime. Now treat X as an element in the field $\mathbb{F}_p$. Next we take an independent and uniform seed $Y \in \{0,1\}^n$ and again treat Y as an element in $\mathbb{F}_p$. Encode X, Y such that $\mathsf{Enc}(X) = (X, X^2)$ and $\mathsf{Enc}(Y) = (Y, Y^2)$. The operations are in $\mathbb{F}_p$. Our non-malleable extractor is defined as

$$\mathsf{nmExt}(X,Y) = \mathsf{IP}(\mathsf{Enc}(X), \mathsf{Enc}(Y)) \bmod M$$

for some integer $M = 2^m$ that we will choose later. Note that $\mathsf{Enc}(X)$ and $\mathsf{Enc}(Y)$ are vectors in $\mathbb{F}_p^2$ and IP is the inner product function taken over $\mathbb{F}_p$.

Again, we show that for any weak source X with min-entropy $(1/2-\delta)n$, $3\mathsf{Enc}(X)$ is close to a weak source that has min-entropy $(1/2+\delta)\log(p^2)$.

Lemma 5.4. *Let $\mathbb{F} = \mathbb{F}_p$ for p prime and X be a random variable over $\mathbb{F}$. There is a universal constant $\delta > 0$ such that if X is any weak source with min-entropy $(1/2-\delta)n$, $3\mathsf{Enc}(X)$ is $p^{-\Omega(1)}$-close to a source with min-entropy $(1/2+\delta)\log(p^2)$.*

Proof. Note that X has min-entropy $(1/2-\delta)n > (1/2-\delta)\log p - 1$. First consider the distribution $2\mathsf{Enc}(X)$. Note that the distribution is of the form $(X+X, X^2+X^2)$. For any (a,b) in the support of $2\mathsf{Enc}(X)$, we have that $a = x_1 + x_2$ and $b = x_1^2 + x_2^2$. Thus there are at most 2 different pairs of (x_1, x_2) that satisfy both equations. Therefore the min-entropy of $2\mathsf{Enc}(X)$ is at least $2H_\infty(X) - 1$. Now let $k = H_\infty(X) - 1$, we have that $\mathsf{Enc}(X)$ has min-entropy at least k and $2\mathsf{Enc}(X)$ has min-entropy at least $2k$. We now have the following claim.

Claim 5.5. *Let α, β be the two constants in Theorem 3.10. Then $3\mathsf{Enc}(X)$ is $2^{-\Omega(k)}$-close to a source with min-entropy $(1+\alpha/2)2k$.*

Proof of the claim. Note that an element in the support of $3\mathsf{Enc}(X)$ has the form $(x_1+x_2+x_3, x_1^2+x_2^2+x_3^2)$. This determines the point

$$\begin{aligned}&(x_1+x_2+x_3, (x_1+x_2+x_3)^2 - (x_1^2+x_2^2+x_3^2))\\ =&((x_1+x_2)+x_3, 2(x_1+x_2)x_3 + (x_1+x_2)^2 - (x_1^2+x_2^2))\end{aligned}$$

Let $a = x_1 + x_2$ and $b = x_1^2 + x_2^2$, this point is

$$(a+x_3, 2ax_3 + a^2 - b).$$

Let $\bar{x}_3 = a + x_3$, then

$$(a+x_3, 2ax_3 + a^2 - b) = (\bar{x}_3, 2a\bar{x}_3 - a^2 - b).$$

For a fixed $(a = x_1 + x_2, b = x_1^2 + x_2^2)$ define the line

$$\ell_{a,b} = \{(x, 2ax - a^2 - b) | x \in \mathbb{F}\}.$$

Note that for different (a,b), the line $\ell_{a,b}$ is also different. Thus we have a set of lines $L = \{\ell_{a,b}\}$. Note that x_3 is sampled from X_3, which has min-entropy k and (a,b) is sampled from $\mathsf{Enc}(X_1) + \mathsf{Enc}(X_2)$, which has min-entropy $2k$. Further note that these two distributions are independent. Since every weak source with min-entropy k is a convex combination of flat k sources, without loss of generality we can assume that X_3 and $\mathsf{Enc}(X_1) + \mathsf{Enc}(X_2)$ are both flat sources. Thus L has size 2^{2k}.

Now assume that $3\mathsf{Enc}(X)$ is ϵ-far from any source with min-entropy $(1+\alpha/2)2k$. Since $3\mathsf{Enc}(X)$ determines the distribution $(A+X_3, 2AX_3 + A^2 - B)$, this distribution is also ϵ-far from any source with min-entropy $(1+\alpha/2)2k$. Thus there must exist some set M of size at most $2^{(1+\alpha/2)2k}$ such that

$$\Pr_{(a,b)\leftarrow 2\mathsf{Enc}(X), X_3 \leftarrow X}[(a+x_3, 2ax_3 + a^2 - b) \in M] \geq \epsilon.$$

Note that whenever $(a+x_3, 2ax_3 + a^2 - b) \in M$, this point has an incidence with the line $\ell_{a,b}$. Further note that whenever (a,b) is different or x_3 is different, the incidence is also different. Thus by the above inequality the number of incidences between the set of points M and the set of lines L is at least

$$\Pr_{(a,b)\leftarrow 2\mathsf{Enc}(X), X_3 \leftarrow X}[(a+x_3, 2ax_3 + a^2 - b) \in M]2^k 2^{2k} \geq \epsilon 2^{3k}.$$

On the other hand, since L has size 2^{2k} and M has size $2^{(1+\alpha/2)2k} \le 2^{(1+\alpha/2)2(1/2-\delta)\log p} < 2^{(1+\alpha/2)\log p} \le p^{2-\beta}$, by Theorem 3.10, the number of incidences between M and L is at most $O(2^{(3/2-\alpha)(2+\alpha)k}) < 2^{3k(1-\alpha/6)} = 2^{-\alpha k/2}2^{3k}$.

Thus we must have $\epsilon < 2^{-\alpha k/2}$. □

By choosing δ appropriately and noting that $k \ge (1/2-\delta)\log p - 2$, the lemma is proved. □

Now we can use the non-uniform XOR lemma to argue that our extractor is non-malleable. Specifically, we have the following lemma.

Lemma 5.6. *Let δ be the constant in Lemma 5.4. Given any (n,k)-source X with $k = (1/2-\delta)n$, and Y an independent source over $\{0,1\}^n$ with min-entropy $(1-\delta)n$, let $W = \mathsf{IP}(\mathsf{Enc}(X), \mathsf{Enc}(Y))$ and $W' = \mathsf{IP}(\mathsf{Enc}(X), \mathsf{Enc}(Y'))$ where $Y' = \mathcal{A}(Y)$ and $\forall y \in \{0,1\}^n, \mathcal{A}(y) \ne y$. For any two characters $\psi(s) = e^{2\pi i t s/p}$ and $\psi'(s) = e^{2\pi i t' s/p}$ where $t, t' \in \mathbb{F}_p$ and $t \ne 0$,*

$$|E_{W,W'}[\psi(W)\psi'(W')]| \le 2^{-\Omega(n)}.$$

Proof. Note that W, W' are deterministic functions of X, Y. Thus

$$E_{W,W'}[\psi(W)\psi'(W')] = E_{X,Y}[\psi(W)\psi'(W')].$$

Depending on whether ψ' is trivial, we have two cases.

Case 1: $t' = 0$. This corresponds to the case where ψ' is the trivial character. In this case $\psi'(W')$ is always 1. Thus

$$E_{W,W'}[\psi(W)\psi'(W')] = E_{X,Y}[\psi(W)] = E_{X,Y}[\psi(\mathsf{Enc}(X)\cdot\mathsf{Enc}(Y))].$$

Note that $\mathsf{Enc}(Y)$ has the same min-entropy as Y, which is $(1-\delta)n$. Now consider $\mathsf{Enc}(X)$. Since X has min-entropy $(1/2-\delta)n$, by Lemma 5.4 $3\mathsf{Enc}(X)$ is $p^{-\Omega(1)}$-close to having min-entropy $(1/2+\delta)\log(p^2)$. Now note that the min-entropy of $4\mathsf{Enc}(X) - 4\mathsf{Enc}(X)$ is at least the min-entropy of $4\mathsf{Enc}(X)$, and which in turn is at least the min-entropy of $3\mathsf{Enc}(X)$. Thus $4\mathsf{Enc}(X) - 4\mathsf{Enc}(X)$ is $p^{-\Omega(1)}$-close to having min-entropy $(1/2+\delta)\log(p^2)$. Since $(1/2+\delta)\log(p^2) + (1-\delta)n > (1+2\delta)\log p + (1-\delta)(\log p - 1) > (2+\delta)\log p - 1$, by Lemma 3.9 we have

$$|E_{W,W'}[\psi(W)\psi'(W')]| = |E_{X,Y}[\psi(\mathsf{Enc}(X)\cdot\mathsf{Enc}(Y))]| \le (p^2 2^{1-(2+\delta)\log p})^{1/16} + p^{-\Omega(1)} = 2^{-\Omega(n)}.$$

Case 2: $t' \ne 0$. This corresponds to the case where ψ' is non-trivial. In this case, note that

$$\psi(W)\psi'(W') = e^{2\pi i t(\mathsf{Enc}(X)\cdot\mathsf{Enc}(Y))} e^{2\pi i t'(\mathsf{Enc}(X)\cdot\mathsf{Enc}(Y'))} = e^{2\pi i t(\mathsf{Enc}(X)\cdot(\mathsf{Enc}(Y)+r\mathsf{Enc}(Y')))},$$

where $r = t'/t \in \mathbb{F}_p$ and $r \ne 0$ since $t \ne 0$ and $t' \ne 0$.

Let $\widetilde{\mathsf{Enc}(Y)} = \mathsf{Enc}(Y) + r\mathsf{Enc}(Y')$, then

$$E_{W,W'}[\psi(W)\psi'(W')] = E_{X,Y}[\psi(W)\psi'(W')] = E_{X,Y}[\psi(\mathsf{Enc}(X)\cdot\widetilde{\mathsf{Enc}(Y)})].$$

Now again by the same argument as above we have that $4\mathsf{Enc}(X) - 4\mathsf{Enc}(X)$ is $p^{-\Omega(1)}$-close to having min-entropy $(1/2+\delta)\log(p^2)$. Now we only need to bound the min-entropy of $\widetilde{\mathsf{Enc}(Y)}$.

If for every two different y_1, y_2, we have that $\mathsf{Enc}(y_1) + r\mathsf{Enc}(y'_1) \neq \mathsf{Enc}(y_2) + r\mathsf{Enc}(y'_2)$, then obviously $\widetilde{\mathsf{Enc}(Y)}$ will have the same min-entropy as Y. Now assume that for some two different y_1, y_2, we have $\mathsf{Enc}(y_1) + r\mathsf{Enc}(y'_1) = \mathsf{Enc}(y_2) + r\mathsf{Enc}(y'_2)$.

This gives us

$$y_1 + ry'_1 = y_2 + ry'_2$$

and

$$(y_1)^2 + r(y'_1)^2 = (y_2)^2 + r(y'_2)^2.$$

Hence we get

$$y_1 - y_2 = r(y'_2 - y'_1)$$

and

$$(y_1 + y_2)(y_1 - y_2) = r(y'_2 + y'_1)(y'_2 - y'_1).$$

Since $y_1 \neq y_2$ and $r \neq 0$, we must have that $y'_1 \neq y'_2$. Thus we get

$$y_1 + y_2 = y'_2 + y'_1.$$

Therefore we can completely solve the equations and get

$$y'_1 = ((r+1)y_2 + (r-1)y_1)/2r, \;\; y'_2 = ((r+1)y_1 + (r-1)y_2)/2r.$$

Thus any element in $\mathsf{Supp}(\widetilde{\mathsf{Enc}(Y)})$ can come from at most 2 elements in $\mathsf{Supp}(Y)$. To see this, assume for the sake of contradiction that we have $\mathsf{Enc}(y_1) + r\mathsf{Enc}(y'_1) = \mathsf{Enc}(y_2) + r\mathsf{Enc}(y'_2) = \mathsf{Enc}(y_3) + r\mathsf{Enc}(y'_3)$ for three different y_1, y_2, y_3. Thus by above we have

$$y'_1 = ((r+1)y_2 + (r-1)y_1)/2r$$

and

$$y'_1 = ((r+1)y_3 + (r-1)y_1)/2r.$$

Note that $r \neq -1$ since otherwise this would imply that $y'_1 = y_1$ which contradicts the assumption that $\forall y, \mathcal{A}(y) \neq y$. Thus we get $y_2 = y_3$, another contradiction.

Therefore the min-entropy of $\widetilde{\mathsf{Enc}(Y)}$ is at least $H_\infty(Y) - 1 = (1-\delta)n - 1$. Now since $(1/2 + \delta)\log(p^2) + (1-\delta)n - 1 > (1+2\delta)\log p + (1-\delta)(\log p - 1) - 1 > (2+\delta)\log p - 2$, by Lemma 3.9 we have

$$|E_{W,W'}[\psi(W)\psi'(W')]| = |E_{X,Y}[\psi(\mathsf{Enc}(X) \cdot \widetilde{\mathsf{Enc}(Y)})]| \leq (p^2 2^{2-(2+\delta)\log p})^{1/16} + p^{-\Omega(1)} = 2^{-\Omega(n)}.$$

□

Now we can prove the following theorem.

Theorem 5.7. *Let δ be the constant from Lemma 5.4. Given any (n, k) source X with $k = (1/2 - \delta)n$ and an independent uniform seed $Y \in \{0,1\}^n$, as well as any deterministic function $\mathcal{A} : \{0,1\}^n \to \{0,1\}^n$ such that $\forall y, \mathcal{A}(y) \neq y$,*

$$|(\mathsf{nmExt}(X, Y), \mathsf{nmExt}(X, \mathcal{A}(Y)), Y) - (U_m, \mathsf{nmExt}(X, \mathcal{A}(Y)), Y)| \leq \epsilon,$$

where $\epsilon = 2^{-\Omega(n)}$ and output size $m = \Omega(n)$.

Proof. Let $Z = \mathsf{nmExt}(X, Y)$ and $Z' = \mathsf{nmExt}(X, \mathcal{A}(Y))$. By Lemma 5.6 and Lemma 3.4, we can choose an $m = \Omega(n)$ and $M = 2^m$ such that when $\mathsf{nmExt}(X, Y) = \mathsf{IP}(\mathsf{Enc}(X), \mathsf{Enc}(Y)) \bmod M$ and Y is an $(n, (1-\delta)n)$ source independent of X, we have

$$|(Z, Z') - (U_m, Z')| \leq \epsilon',$$

where $\epsilon' = O(n 2^m 2^{-\Omega(n)} + 2^{m-n}) = 2^{-\Omega(n)}$.

Therefore when Y is an independent uniform distribution over $\{0,1\}^n$, by Theorem 3.6 we have

$$|(Z, Z', Y) - (U_m, Z', Y)| \leq \epsilon,$$

where $\epsilon = 2^{2m}(2^{1-\delta n} + \epsilon')$.

Note that $\epsilon' = O(n 2^m 2^{-\Omega(n)} + 2^{m-n})$. Thus we can take $m = \Omega(n)$ and $\epsilon = 2^{2m}(2^{1-\delta n} + \epsilon') = 2^{-\Omega(n)}$. Thus the theorem is proved. ■

6 Achieving Even Smaller Min-Entropy

In this section we show that we can construct non-malleable extractors for even smaller min-entropy rate (potentially any constant arbitrarily close to 0), if we assume that we have affine extractors with large enough output size, and the Approximate Duality Conjecture (or the Polynomial Freiman-Ruzsa Conjecture) as in [BSZ11].

Recall the definition of an affine extractor.

Definition 6.1. An $[n, m, \rho, \epsilon]$ affine extractor is a deterministic function $f : \{0,1\}^n \to \{0,1\}^m$ such that whenever X is the uniform distribution over some affine subspace over $\mathbb{F}_2^n$ with dimension ρn, we have that for every $z \in \{0,1\}^m$,

$$|\Pr[f(X) = z] - 2^{-m}| < \epsilon.$$

We now have the following construction.

Construction 6.2. Given any (n, k) source X and a constant $0 < \lambda < 1$, let $f : \{0,1\}^{n'} \to \{0,1\}^{m'}$ be an $[n', m' = (1-\lambda)\frac{2}{3}n', \frac{2}{3}, 2^{-m'}]$ affine extractor such that $n = n' - m'$. For any $z \in \{0,1\}^{m'}$, let $f^{-1}(z) = \{x : f(x) = z\}$. Then there exists $z \in \{0,1\}^{m'}$ such that $|f^{-1}(z)| \geq 2^n$. Let $F : \{0,1\}^n \to f^{-1}(z)$ be (any) injective map. Now take a BCH code as in Theorem 3.11 with $t = 2$ and $m = n'/2$, and the $n' \times (2^{n'/2} - 1)$ parity check matrix.

Take an independent uniform seed $Y \in \{0,1\}^{n'/2-1}$ and let S_Y stand for the integer whose binary expression is Y. We encode Y to $\bar{Y}$ such that $\bar{Y} = \mathsf{Enc}(Y)$ is the S_Y'th column in the parity check matrix (i.e., $\bar{Y} = (Y, Y^3)$ when Y is viewed as an element in $\mathbb{F}^*_{2^{n'/2}}$). Our non-malleable extractor is now defined as

$$\mathsf{nmExt}(X,Y) = \mathsf{IP}(F(X), \mathsf{Enc}(Y)) = \mathsf{IP}(F(X), \bar{Y}),$$

where IP is the inner product function taken over $\mathbb{F}_2$.

Remark 6.3. Note that here the function F may not be efficiently computable (in time poly(n)). However, the time to compute F is polynomial in the length of the truth table of our final extractor.

Let $W = \mathsf{nmExt}(X,Y)$ and $W' = \mathsf{nmExt}(X,Y')$ where $Y' = \mathcal{A}(Y)$ such that $\forall y, \mathcal{A}(y) \neq y$. Again we will show the non-malleability of our extractor by showing that $(W,Y) \approx_\epsilon (U,Y)$ and $(W \oplus W', Y) \approx_\epsilon (U,Y)$ for some ϵ. To this end, we first show the following.

Lemma 6.4. *For any (n,k) source X with $k = \frac{2.5\lambda}{1+2\lambda}n$ and an independent $(\frac{n'}{2}-1, \frac{n'}{3}+1)$ source Y, $\mathsf{nmExt}(X,Y)$ is non-constant.*

Proof. As usual we can assume without loss of generality that X and Y are flat sources. If $\mathsf{nmExt}(X,Y) = \mathsf{IP}(F(X), \mathsf{Enc}(Y))$ is a constant, then $\mathsf{Supp}(F(X))$ and $\mathsf{Supp}(\mathsf{Enc}(Y))$ must be contained in two affine subspaces with dimension d_1, d_2 such that $d_1 + d_2 \leq n'$. Note that $\mathsf{Enc}(Y)$ is an injective function. Thus $\mathsf{Supp}(\mathsf{Enc}(Y))$ has size $2^{\frac{n'}{3}+1}$. Therefore $d_2 > \frac{n'}{3}$. We next show that $d_1 > \frac{2}{3}n'$ and thus reach a contradiction.

To see this, let $S = \mathsf{Supp}(F(X))$. It suffices to show that S is not contained in any affine subspace of dimension $\frac{2}{3}n'$. Let A be such an affine subspace. We have

$$|A \cap S| \leq |A \cap f^{-1}(z)| < 2 \cdot 2^{-m'} 2^{\frac{2}{3}n'} = 2^{\frac{2\lambda}{3}n'+1},$$

where the last inequality follows from the fact that f is an affine extractor. Now note that $|S| = 2^{\frac{2.5\lambda}{1+2\lambda}n} = 2^{\frac{2.5\lambda}{3}n'}$. Thus we have that $|A \cap S| < |S|$ and therefore S cannot be contained in A. □

Recall the Approximate Duality conjecture introduced in [BSZ11].

Conjecture 6.5. *(Approximate Duality (ADC)) [BSZ11] For every pair of constants $\alpha, \delta > 0$ there exist a constant $\zeta > 0$ and an integer r, both depending on α and δ such that the following holds for sufficiently large n. If $A, B \subseteq \mathbb{F}_2^n$ satisfy $|A|, |B| \geq 2^{\alpha n}$ and $\mu^\perp(A,B) \geq 2^{-\zeta n}$, then there exists a pair of subsets*

$$A' \subseteq A, A' \geq \frac{|A|}{2^{\delta n+1}} \text{ and } B' \subseteq B, |B'| \geq \left(\frac{\mu^\perp(A,B)}{2}\right)^r \cdot \frac{|B|}{2^{\delta n}}$$

such that $\mu^\perp(A', B') = 1$.

Now we have the following lemma.

Lemma 6.6. *There exists a constant $\zeta = \zeta(\lambda)$ such that for any (n,k) source X with $k = \frac{3\lambda}{1+2\lambda}n$ and an independent $(\frac{n'}{2}-1, \frac{7n'}{15})$ source Y, $\mathsf{nmExt}(X,Y)$ is $2^{-\zeta n}$-close to uniform.*

Proof. Let $v = \frac{n}{n'} = \frac{1+2\lambda}{3}$, $\alpha = min\{\lambda, \frac{1}{3}\}$ and $\delta = \frac{\lambda}{8}$. Let ζ' and r be the constant and the integer guaranteed by conjecture 6.5 for α and δ. Let $\zeta = min\{\frac{\zeta'}{v}, \frac{1-\lambda}{8r}\}$. We will prove the lemma by way of contradiction.

Let X and Y be two independent sources as in the statement of the lemma. Again we assume without loss of generality that both X and Y are flat sources. Let $A = \mathsf{Supp}(X)$, $B = \mathsf{Supp}(Y)$

and $\bar{A} = \{F(a)|a \in A\} \subseteq \mathbb{F}_2^{n'}$, $\bar{B} = \{\mathsf{Enc}(b)|b \in B\} \subseteq \mathbb{F}_2^{n'}$. Note that both F and Enc are injective functions. Thus $|\bar{A}| = 2^{\frac{3\lambda}{1+2\lambda}n} = 2^{\lambda n'} \geq 2^{\alpha n'}$ and $|\bar{B}| = 2^{\frac{7n'}{15}} > 2^{\frac{n'}{3}} \geq 2^{\alpha n'}$.

Assume for the sake of contradiction that the error of $\mathsf{nmExt}(X, Y)$, which is equal to $\frac{1}{2}\mu^{\perp}(\bar{A}, \bar{B})$, is greater than $2^{-\zeta n} \geq 2^{-\zeta' n'}$. Then by the ADC conjecture (conjecture 6.5) there exist $A' \subseteq \bar{A}$ and $B' \subseteq \bar{B}$ such that

$$|A'| \geq \frac{|\bar{A}|}{2^{\delta n'+1}} \geq 2^{\frac{5\lambda}{6}n'} = 2^{\frac{2.5\lambda}{1+2\lambda}n} \text{ and } |B'| \geq \frac{|\bar{B}|}{2^{\delta n' + r\zeta n}} \geq \frac{2^{\frac{7n'}{15}}}{2^{\frac{n'}{8}}} > 2^{\frac{n'}{3}+1},$$

and $\mu^{\perp}(A', B') = 1$.

Let A'' and B'' be the preimages of A' and B' under F and Enc respectively. Since F and Enc are injective, we must have $|A''| \geq 2^{\frac{2.5\lambda}{1+2\lambda}n}$ and $|B''| \geq 2^{\frac{n'}{3}+1}$. Thus if we let X' and Y' be the uniform distribution over A'' and B'' respectively, we get two independent sources that satisfy the conditions in Lemma 6.4. However $\mathsf{IP}(F(X'), \mathsf{Enc}(Y'))$ is a constant, which contradicts Lemma 6.4. Thus we must have that $\mathsf{nmExt}(X, Y)$ is $2^{-\zeta n}$-close to uniform. □

Now we can prove the following theorem.

Theorem 6.7. *There exists a constant $\zeta = \zeta(\lambda)$ such that for any (n, k) source X with $k = \frac{3\lambda}{1+2\lambda}n$ and an independent uniform seed Y over $\frac{n'}{2} - 1 = \frac{3}{2+4\lambda}n - 1$ bits,*

$$|(W, W', Y) - (U, W', Y)| \leq 2^{-\Omega(\zeta n)}.$$

Proof. Let $\zeta = \zeta(\lambda)$ be as in Lemma 6.6. First let Y be an independent source with min-entropy $\frac{7n'}{15} + 1$. By Lemma 6.6 we have

$$|W - U| \leq 2^{-\zeta n}.$$

Now consider $W \oplus W' = \mathsf{nmExt}(X, Y) \oplus \mathsf{nmExt}(X, Y') = \mathsf{IP}(F(X), \mathsf{Enc}(Y) + \mathsf{Enc}(Y'))$. By the same argument as in the proof of Theorem 4.2, $\mathsf{Enc}(Y) + \mathsf{Enc}(Y')$ has min-entropy at least $H_\infty(Y) - 1 = \frac{7n'}{15}$. Thus again by Lemma 6.6 we have

$$|W \oplus W' - U| \leq 2^{-\zeta n}.$$

Therefore by the non-uniform XOR lemma, Lemma 3.3, we have

$$|(W, W') - (U, W')| \leq O(2^{-\zeta n}).$$

Now if we let Y be the uniform distribution over $\frac{n'}{2} - 1$ bits, by Theorem 3.6 we have

$$|(W, W', Y) - (U, W', Y)| \leq 2^2(2^{-\Omega(n)} + O(2^{-\zeta n})) = 2^{-\Omega(\zeta n)}.$$

■

7 Reducing the Seed Length and Increasing the Output Size

In this section we show that we can reduce the seed length and increase the output size for the constructions in Section 4, Subsection 5.1 and Section 6. All these constructions share the same pattern: the seed Y is encoded using the parity check matrix of a BCH code, and then the output is the inner product function of the encoded source and the encoded seed over $\mathbb{F}_2$.

We only discuss the construction in Subsection 5.1, and the method used can be applied to all the other constructions in the same way. We start by showing how to reduce the seed length.

7.1 Reducing the seed length

In the constructions mentioned above, we use a BCH code with distance 5. Thus the columns of the parity check matrix are 4-wise linearly independent. To reduce the seed length, we are going to use a BCH code with larger distance. Specifically, we will choose a $[2^\ell - 1, 2^\ell - 1 - 2t\ell, 4t+1]$-BCH code with $\ell = p/t$ for some parameter t to be chosen later. Note that the parity check matrix is a $2p \times (2^\ell - 1)$ matrix[4]. Thus the columns of the matrix are $D = 4t$-wise linearly independent. The detailed construction is as follows.

- Given an (n, k)-source X with $k = (1/2 - \delta)n$, pick a prime p such that $n \le p \le n(1 + \frac{1}{2\ln^2 n})$.
- Let $q = 2^p$ and g be a generator in $\mathbb{F}_q^*$. Treat X as an element in $\mathbb{F}_q^*$ and encode X such that $\mathsf{Enc}(X) = (X, g^X)$.
- Let $\ell = p/t$. Take the parity check matrix of a $[2^\ell - 1, 2^\ell - 1 - 2t\ell, 4t+1]$-BCH code. Note that it is a $2p \times (2^\ell - 1)$ matrix. Take an independent and uniform seed $Y \in \{0,1\}^{\ell-1}$ and let S_Y stand for the integer whose binary expression is Y. We encode Y to $\bar{Y}$ such that $\bar{Y}$ is the S_Y'th column in the parity check matrix.
- Output $\mathsf{nmExt}(X, Y) = \mathsf{IP}(\mathsf{Enc}(X), \bar{Y})$ where IP is the inner product function taken over $\mathbb{F}_2$.

As in Subsection 5.1, we have Claim 5.2. We now want to argue about the min-entropy of $t\bar{Y}$ and $t(\bar{Y} + \bar{\mathcal{A}(Y)})$.

Lemma 7.1. *Assume Y has min-entropy k_2, then $t\bar{Y}$ is $t^2 2^{-(k_2+1)}$-close to having min-entropy $t(k_2 - \log t)$, and $t(\bar{Y} + \bar{Y}')$ is $t^2 2^{-(k_2+2)} + t(t2^{-\frac{2}{3}k_2})^{\log t}$-close to having min-entropy $t((1 - \frac{\log t}{3t})k_2 - 3\log t)$.*

Proof. Without loss of generality assume that Y is a flat source. Let $K = 2^{k_2}$. First consider $t\bar{Y}$. Note that $\bar{Y}$ has the same min-entropy as Y and is also a flat source, since every two columns of the parity check matrix are different. The support of $t\bar{Y}$ has the form $\bar{y}_1 + \cdots + \bar{y}_t$. Consider the case where all $\bar{y}_i$'s are different. This takes up a probability mass of

$$\frac{K!}{(K-t)!} \cdot K^{-t} = 1 \cdot (1 - \frac{1}{K}) \cdots (1 - \frac{t-1}{K}) > 1 - \sum_{i=1}^{t-1} \frac{i}{K} > 1 - \frac{t^2}{2K}.$$

Since the columns of the parity check matrix are $4t$-wise linearly independent. For every two different sets $\{\bar{y}_i\}$'s, their sum cannot be the same. Therefore, the probability mass of getting a

[4]Actually p is not divisible by t, thus $\ell t < p$. However for simplicity we will assume that the matrix has $2p$ rows. For example we can add 0's in the end, the small error does not affect our analysis.

particular value is at most $t!K^{-t} \leq 2^{-t(k_2-\log t)}$. Thus $t\bar{Y}$ is $t^2 2^{-(k_2+1)}$-close to having min-entropy $t(k_2 - \log t)$.

Next consider $t(\bar{Y}+\bar{\mathcal{A}(Y)})$. Let $\mathcal{A}(Y) = Y'$ and $Y'' = \bar{Y}+\bar{Y'}$. Note that for every $s \in \mathsf{Supp}(Y'')$, $s \neq 0$ since $\forall y, \mathcal{A}(y) \neq y$. Also note that Y'' has min-entropy at least k_2-1 since if $\bar{y}_1+\bar{y'}_1 = \bar{y}_2+\bar{y'}_2$ for $y_1 \neq y_2$, then we must have $\bar{y'}_1 = \bar{y}_2$ and $\bar{y'}_2 = \bar{y}_1$. Without loss of generality assume that Y'' is a flat source with min-entropy $k_2 - 1$. Let $K_2 = 2^{k_2-1}$. Note that now in the support of Y'' there are no two different y_1, y_2 such that $\bar{y}_1 + \bar{y'}_1 = \bar{y}_2 + \bar{y'}_2$ (since this will be absorbed into the same element).

We now consider tY''. An element in its support has the form $\sum_{i=1}^t (\bar{y}_i + \bar{y'}_i)$. We first get rid of those elements in $\mathsf{Supp}(tY'')$ such that some of the $\{\bar{y}_i + \bar{y'}_i\}$'s are the same. By the same argument as above this takes up a probability mass of at most $\frac{t^2}{2K_2}$. Now, for a particular set $\{\bar{y}_i + \bar{y'}_i\}_{i\in[t]}$, we consider how many different sets can have the same sum.

Since the columns of the parity check matrix are $4t$-wise linearly independent, if the sum of two different sets $\{\bar{y}_i + \bar{y'}_i\}_{i\in[t]}$ are the same, then except those $\bar{y}_i + \bar{y'}_i$'s that are common in both sets, the rest of $\bar{y}_i + \bar{y'}_i$'s must form cycles. By cycle we mean a set of l elements such that $\bar{y'}_1 = \bar{y}_2, \bar{y'}_2 = \bar{y}_3, \cdots, \bar{y'}_l = \bar{y}_1$ so that the sum is 0. Note that $l \geq 3$ since the support of Y'' has no 2-cycles. Let S_1, S_2 be the two sets $\{\bar{y}_i + \bar{y'}_i\}_{i\in[t]}$. Now, the elements in a cycle can come from both sets or just from one set. If the elements from a cycle comes only from S_2, then this cycle can be replaced by any other cycle with the same length, and the sum of S_1 and S_2 are still the same. On the other hand, if the elements of a cycle comes from both S_1 and S_2, then the elements in this cycle are completely determined by S_1 since cycles are disjoint. Therefore, let r be the number of common elements in S_1, S_2, and let l be the total length of cycles whose elements only come from the rest elements of S_2, and note that cycles have length at least 3, we have that if $l \geq \log t$, then the total probability mass of these elements in $\mathsf{Supp}(tY'')$ is at most

$$\sum_{\log t \leq l \leq t} \binom{t}{l}\binom{\frac{K_2}{3}}{\frac{l}{3}} l!(K_2)^{-l} \leq \sum_{\log t \leq l \leq t} t^l (\frac{K_2}{3})^{\frac{l}{3}}(K_2)^{-l} < \sum_{\log t \leq l \leq t} (t(K_2)^{-\frac{2}{3}})^l < t(t2^{-\frac{2}{3}k_2})^{\log t}.$$

On the other hand, if $l < \log t$, then the probability that tY'' gets a particular value is at most

$$\sum_{0\leq l \leq \log t}\sum_{0 \leq r \leq t-l} \binom{t}{l}\binom{t-l}{r}\binom{\frac{K_2}{3}}{\frac{l}{3}} t!(K_2)^{-t} < t\log t \cdot t^{2t}(\frac{K_2}{3})^{\frac{\log t}{3}}(K_2)^{-t} < t \log t (t^2 (K_2)^{-(1-\frac{\log t}{3t})})^t.$$

Thus the min-entropy is at least $t((1-\frac{\log t}{3t})k_2 - 3\log t)$. □

Now for an (n, k)-source X with $k = (1/2-\delta)n$, we know that $\mathsf{3Enc}(X)$ is $2^{-\Omega(n)}$-close to having min-entropy $(1/2+\delta)(2p)$. Assume that we want our non-malleable extractor to have error $\epsilon \leq 1/n$. We'll choose a parameter $t < n/C \log n$ for a sufficiently large constant $C > 1$. When Y is uniform over $\ell = p/t$ bits, $t\bar{Y}$ is close to having min-entropy $t(k_2-\log t) > (1-1/C)p > (1/2-\delta/2)(2p)$, and $t(\bar{Y}+\bar{Y'})$ is close to having min-entropy $t((1-\frac{\log t}{3t})k_2-3\log t) > (1-1/C)p > (1/2-\delta/2)(2p)$. When $t\bar{Y}$ and $t(\bar{Y}+\bar{Y'})$ indeed have this min-entropy, by Lemma 3.8 we have that both $\mathsf{IP}(\mathsf{3Enc}(X), t\bar{Y})$ and $\mathsf{IP}(\mathsf{3Enc}(X), t(\bar{Y}+\bar{Y'}))$ are $2^{-\Omega(n)}$-close to uniform. Thus we can take $t = \Omega(n/(\log(1/\epsilon)))$ and by Lemma 3.9 and Theorem 3.6 we have that the error of the non-malleable extractor is at most ϵ, and the seed length is roughly $p/t = O(n/t) = O(\log(1/\epsilon))$. Thus we have the following theorem.

Theorem 7.2. *There exists a universal constant $\delta > 0$ such that for every $n \in \mathbb{N}$ and ϵ such that $2^{-\Omega(n)} \leq \epsilon \leq 1/\text{poly}(n)$, there exists an explicit (k, ϵ) non-malleable extractor* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ *for $k = (1/2 - \delta)n$ and seed length $d = O(\log n + \log(1/\epsilon))$.*

7.2 Increasing the output size

Here we show that we can modify all the constructions with 1 bit output to output $m = \Omega(n)$ bits. Again we only discuss the construction in Subsection 5.1, and the method can be applied to all the other constructions with 1 bit output.

Recall that in the construction we used a field $\mathbb{F}_{2^p}$ for a prime p. Given the finite filed $\mathbb{F}_{2^p}$, the elements of this field form a vector space of dimension p over $\mathbb{F}_2$. Let $b_1, \cdots, b_p \in \mathbb{F}_{2^p}$ be a basis for this vector space. Now recall that in the construction we encode the seed Y to $\bar{Y} = (Y, Y^3)$, when viewing Y as an element in $\mathbb{F}_{2^p}^*$. Now for each b_i, let $\bar{Y^i} = (b_i Y, b_i Y^3)$ and define one bit $Z_i = \mathsf{IP}(\mathsf{Enc}(X), \bar{Y^i})$. We now show that $\{Z_i\}$ satisfy the conditions of a non-uniform XOR lemma.

Lemma 7.3. *Given any (n, k)-source X with $k = (1/2 - \delta)n$ and an independent seed $Y \in \{0,1\}^{p-1}$ with min-entropy $(1-\delta)p$, let $\mathcal{A} : \{0,1\}^{p-1} \to \{0,1\}^{p-1}$ be any deterministic function such that $\forall y, \mathcal{A}(y) \neq y$. For any i, let $Z'_i = \mathsf{IP}(\mathsf{Enc}(X), \bar{Y^i}')$, where $\bar{Y^i}' = (b_i Y', b_i Y'^3)$ and $Y' = \mathcal{A}(Y)$. Then for any non-empty subset $S_1 \subseteq [p]$ and any subset $S_2 \subseteq [p]$, we have that*

$$|\bigoplus_{i \in S_1} Z_i \oplus \bigoplus_{j \in S_2} Z'_j - U| \leq 2^{-\Omega(n)}.$$

Proof. Note that

$$\bigoplus_{i \in S_1} Z_i = \mathsf{IP}(\mathsf{Enc}(X), \sum_{i \in S_1} \bar{Y^i}) = \mathsf{IP}(\mathsf{Enc}(X), t_1(Y, Y^3)),$$

where $t_1 = \sum_{i \in S_1} b_i \in \mathbb{F}_{2^p}$, and

$$\bigoplus_{j \in S_2} Z'_j = \mathsf{IP}(\mathsf{Enc}(X), \sum_{j \in S_2} \bar{Y^j}') = \mathsf{IP}(\mathsf{Enc}(X), t_2(Y', Y'^3)),$$

where $t_2 = \sum_{j \in S_2} b_j \mathbb{F}_{2^p}$.

Thus

$$\bigoplus_{i \in S_1} Z_i \oplus \bigoplus_{j \in S_2} Z'_j = \mathsf{IP}(\mathsf{Enc}(X), \tilde{Y}),$$

where $\tilde{Y} = t_1(Y, Y^3) + t_2(Y', Y'^3)$.

As usual, it suffices to prove that for the only non-trivial character ψ (since we only have 1 bit),

$$|E_{X,Y}[\psi(\mathsf{IP}(\mathsf{Enc}(X), \tilde{Y}))]| \leq 2^{-\Omega(n)}.$$

Again, let $X' = 4\mathsf{Enc}(X) - 4\mathsf{Enc}(X)$, by Lemma 3.9 we have

$$|E_{X,Y}[\psi(\mathsf{IP}(\mathsf{Enc}(X), \tilde{Y}))]| \leq |E_{X',Y}[\psi(X' \cdot \tilde{Y})]|^{\frac{1}{8}}.$$

Now by the same argument as in the proof of Lemma 5.1, X' is $2^{-\Omega(n)}$-close to a source with min-entropy $(1/2 + \delta)(2p)$. Thus we only need to bound the min-entropy of $\tilde{Y}$. We have two cases.

Case 1: $S_2 = \phi$. In this case $\tilde{Y} = t_1(Y, Y^3)$. Since $S_1 \neq \phi$, we have $t_1 \neq 0$. Thus $\tilde{Y}$ has the same min-entropy as Y, which is $(1-\delta)p$. Since $(1/2+\delta)(2p) + (1-\delta)p = (2+\delta)p$, by Lemma 3.8 we have that

$$|E_{X',Y}[\psi(X' \cdot \tilde{Y})]| \leq 2^{2p} 2^{-(2+\delta)p} + 2^{-\Omega(n)} = 2^{-\Omega(n)}.$$

Therefore $|E_{X,Y}[\psi(\mathsf{IP}(\mathsf{Enc}(X), \tilde{Y}))]| \leq 2^{-\Omega(n)}$.

Case 2: $S_2 \neq \phi$. In this case we have $t_1 \neq 0$ and $t_2 \neq 0$. We need to bound the min-entropy of $\tilde{Y} = t_1(Y, Y^3) + t_2(Y', Y'^3)$. Again, if for every two different y_1, y_2, we have $t_1(y_1, y_1^3) + t_2(y_1', y_1'^3) \neq t_1(y_2, y_2^3) + t_2(y_2', y_2'^3)$, then $\tilde{Y}$ will have the same min-entropy of Y. We now have the following claim.

Claim 7.4. *Any element in* $\mathsf{Supp}(\tilde{Y})$ *can come from at most 3 different elements in* $\mathsf{Supp}(Y)$*.*

Proof of the claim. Assume for the sake of contradiction that there are 4 different y_1, y_2, y_3, y_4 such that $t_1(y_i, y_i^3) + t_2(y_i', y_i'^3)$ are the same for $i = 1, 2, 3, 4$. First consider y_1, y_2, we have $t_1(y_1, y_1^3) + t_2(y_1', y_1'^3) = t_1(y_2, y_2^3) + t_2(y_2', y_2'^3)$. Since $t_1 \neq 0$, let $r = t_2/t_1 \mathbb{F}_{2^p}$. Thus $r \neq 0$ and we have $(y_1, y_1^3) + r(y_1', y_1'^3) = (y_2, y_2^3) + r(y_2', y_2'^3)$. We first consider the case where $r = 1$. In this case, the vectors (y_1, y_1^3), $(y_1', y_1'^3)$, (y_2, y_2^3) and $(y_2', y_2'^3)$ are linearly dependent over $\mathbb{F}_2$. However we know that the columns of the parity check matrix of the BCH code are 4-wise linearly independent. Thus we must have $y_1' = y_2$ and $y_2' = y_1$. Thus in this case the element in $\mathsf{Supp}(\tilde{Y})$ comes from at most 2 different elements in $\mathsf{Supp}(Y)$. Now if $r \neq 1$, we have

$$y_1 + ry_1' = y_2 + ry_2'$$

and

$$(y_1)^3 + r(y_1')^3 = (y_2)^3 + r(y_2')^3.$$

Hence we get

$$y_1 - y_2 = r(y_2' - y_1')$$

and

$$(y_1^2 + y_1y_2 + y_2^2)(y_1 - y_2) = r(y_2'^2 + y_1'y_2' + y_1'^2)(y_2' - y_1').$$

Since $y_1 \neq y_2$ and $r \neq 0$, we must have that $y_1' \neq y_2'$. Thus we get

$$y_1^2 + y_1y_2 + y_2^2 = y_2'^2 + y_1'y_2' + y_1'^2.$$

Similarly we can get

$$y_1^2 + y_1y_3 + y_3^2 = y_3'^2 + y_1'y_3' + y_1'^2.$$

Thus

$$(y_1 + y_2 + y_3)(y_2 - y_3) = (y_1' + y_2' + y_3')(y_2' - y_3').$$

Also, from $y_2 + ry_2' = y_3 + ry_3'$ we get

$$y_2 - y_3 = r(y_3' - y_2').$$

Since $y_2 \neq y_3$, we have

$$y_1' + y_2' + y_3' = -r(y_1 + y_2 + y_3).$$

Similarly we can get

$$y_1' + y_2' + y_4' = -r(y_1 + y_2 + y_4).$$

Therefore

$$y_4' - y_3' = r(y_3 - y_4).$$

On the other hand, from $y_3 + ry_3' = y_4 + ry_4'$ we get

$$y_4' - y_3' = 1/r(y_3 - y_4).$$

Thus

$$(r^2 - 1)(y_3 - y_4) = 0.$$

Since $r \neq 1$, $r^2 - 1 \neq 0$. Thus we have $y_3 = y_4$, a contradiction. □

Therefore, the min-entropy of $\tilde{Y}$ is at least $H_\infty(Y) - \log 3 = (1-\delta)p - \log 3$. Now by Lemma 3.8 we have that

$$|E_{X',Y}[\psi(X' \cdot \tilde{Y})]| \leq 2^{2p} 2^{-(1+2\delta)p} 2^{-(1-\delta)p + \log 3} + 2^{-\Omega(n)} = 2^{-\Omega(n)}.$$

Therefore $|E_{X,Y}[\psi(\mathsf{IP}(\mathsf{Enc}(X), \tilde{Y}))]| \leq 2^{-\Omega(n)}$. □

Now if we have a uniform random seed $Y \in \{0,1\}^{p-1}$, by Lemma 3.3 and Theorem 3.6 we can choose $m = \Omega(n)$ bits from $\{Z_i\}$ such that when we output $Z_1 \circ \cdots \circ Z_m$ we get a non-malleable extractor with error $2^{-\Omega(n)}$ and output size $m = \Omega(n)$. Specifically, we have the following theorem.

Theorem 7.5. *There exists a constant $0 < \delta < 1$ such that for any $n \in \mathbb{N}$, $k = (1/2 - \delta)n$, there exists an explicit (k, ϵ)-non-malleable extractor $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ with $m = \Omega(n)$ and $\epsilon = 2^{-\Omega(n)}$.*

8 Conclusions and Open Problems

In this paper we give the first explicit constructions of non-malleable extractors for min-entropy $k < n/2$. We give two unconditional constructions for $k = (1/2 - \delta)n$ for some constant $\delta > 0$ and one conditional construction that can potentially achieve any constant min-entropy rate. Using our non-malleable extractor, we also obtain the first optimal privacy amplification protocol for $k = (1/2 - \delta)n$, with an active adversary.

There are several natural open problems left. First, two of our constructions achieve an optimal seed length, but only output 1 bit. It will be interesting to see if we can output more than 1 bit in these cases. Second, the constructions that can output multiple bits have a large seed length

$d = n$. It is natural to ask if we can reduce the seed length in these cases. Also, one obvious open problem is to construct non-malleable extractors for smaller min-entropy, or to obtain an unconditional construction for any constant min-entropy rate.

Finally, we want to point out that all known constructions of non-malleable extractors, including [DLWZ11, CRS11] and our constructions, seem to be some variant of known constructions of two-source extractors. This seems to suggest that there is some connection between non-malleable extractors and two-source extractors. We feel that this is interesting. It would be very nice if such a connection can be established, and thus help us gain knowledge and insights about both non-malleable extractors and two-source extractors.

Acknowledgments

We are grateful to David Zuckerman for many valuable discussions, especially for his suggestion to use the BCH code.

References

[BBR88] C.H. Bennett, G. Brassard, and J.-M. Robert. Privacy amplification by public discussion. *SIAM Journal on Computing*, 17:210–229, 1988.

[BKT04] Jean Bourgain, Nets Katz, and Terence Tao. A sum-product estimate in finite fields, and applications. *Geometric and Functional Analysis*, 14:27–57, 2004.

[Bou05] Jean Bourgain. More on the sum-product phenomenon in prime fields and its applications. *International Journal of Number Theory*, 1:1–32, 2005.

[BSZ11] Eli Ben-Sasson and Noga Zewi. From affine to two-source extractors via approximate duality. In *Proceedings of the 43nd Annual ACM Symposium on Theory of Computing*, 2011.

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988.

[CKOR10] N. Chandran, B. Kanukurthi, R. Ostrovsky, and L. Reyzin. Privacy amplification with asymptotically optimal entropy loss. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pages 785–794, 2010.

[CRS11] Gil Cohen, Ran Raz, and Gil Segev. Non-malleable extractors with short seeds and applications to privacy amplification. Technical Report TR11-096, ECCC, 2011.

[DKRS06] Y. Dodis, J. Katz, L. Reyzin, and A. Smith. Robust fuzzy extractors and authenticated key agreement from close secrets. In *CRYPTO*, pages 232–250, 2006.

[DLWZ11] Yevgeniy Dodis, Xin Li, Trevor D. Wooley, and David Zuckerman. Privacy amplification and non-malleable extractors via character sums. In *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, 2011.

[DW09] Yevgeniy Dodis and Daniel Wichs. Non-malleable extractors and symmetric key cryptography from weak secrets. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, page 601610, 2009.

[FS02] Lance Fortnow and Ronen Shaltiel. Recent developments in explicit constructions of extractors, 2002.

[Kon03] S. Konyagin. A sum-product estimate in fields of prime order. Technical report, Arxiv, 2003. `http://arxiv.org/abs/math.NT/0304217`.

[KR09] B. Kanukurthi and L. Reyzin. Key agreement from close secrets over unsecured channels. In *EUROCRYPT*, pages 206–223, 2009.

[MW97] Ueli M. Maurer and Stefan Wolf. Privacy amplification secure against active adversaries. In *CRYPTO '97*, 1997.

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

[Rao07] Anup Rao. An exposition of Bourgain's 2-source extractor. Technical Report TR07-34, ECCC: Electronic Colloquium on Computational Complexity, 2007. `http://eccc.hpi-web.de/eccc-reports/2007/TR07-034/index.html`.

[RW03] R. Renner and S. Wolf. Unconditional authenticity and privacy from an arbitrarily weak secret. In *CRYPTO*, pages 78–95, 2003.

[Vad02] Salil Vadhan. Randomness extractors and their many guises: Invited tutorial. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, 2002.

[Vaz85] Umesh Vazirani. Towards a strong communication complexity theory or generating quasi-random sequences from two communicating slightly-random sources (extended abstract). In *Proceedings of the 17th Annual ACM Symposium on Theory of Computing*, 1985.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de