Improved Extractors for Recognizable and Algebraic Sources

Fu Li
Department of Computer Science, University of Texas at Austin, USA
fuli2015@cs.utexas.edu

David Zuckerman
Department of Computer Science, University of Texas at Austin, USA
diz@cs.utexas.edu

Abstract

We study the task of seedless randomness extraction from recognizable sources, which are uniform distributions over sets of the form $\{x : f(x) = 1\}$ for functions f in some specified class $\mathcal{C}$. We give two simple methods for constructing seedless extractors for $\mathcal{C}$-recognizable sources.

Our first method shows that if $\mathcal{C}$ admits XOR amplification, then we can construct a seedless extractor for $\mathcal{C}$-recognizable sources by using a mildly hard function for $\mathcal{C}$ as a black box. By exploiting this reduction, we give polynomial-time, seedless randomness extractors for three natural recognizable sources: (1) constant-degree algebraic sources over any prime field, where constant-degree algebraic sources are uniform distributions over the set of zeros of a system of constant degree polynomials; (2) sources recognizable by randomized multiparty communication protocols of cn bits, where $c > 0$ is a small enough constant; (3) halfspace sources, or sources recognizable by linear threshold functions. In particular, the new extractor for each of these three sources has linear output length and exponentially small error for min-entropy $k \geq (1 - \alpha)n$, where $\alpha > 0$ is a small enough constant.

Our second method shows that a seed-extending pseudorandom generator with exponentially small error for $\mathcal{C}$ yields an extractor with exponentially small error for $\mathcal{C}$-recognizable sources, improving a reduction by Kinne, Melkebeek, and Shaltiel [16]. Using the hardness of the parity function against AC^0 [13], we significantly improve Shaltiel's extractor [25] for AC^0-recognizable sources. Finally, assuming sufficiently strong one-way permutations, we construct seedless extractors for sources recognizable by BPP algorithms, and these extractors run in quasi-polynomial time.

2012 ACM Subject Classification Theory of computation → Pseudorandomness and derandomization; Theory of computation

Keywords and phrases Extractor, Pseudorandomness

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2019.72

Category RANDOM

Related Version https://eccc.weizmann.ac.il/report/2018/110/

Funding Supported by NSF Grant CCF-1526952, NSF Grant CCF-1705028, and a Simons Investigator Award (#409864, David Zuckerman).

Acknowledgements We wish to thank Salil Vadhan, Ronen Shaltiel, Avishay Tal, and William Hoza for helpful discussions and comments.

1 Introduction

Randomness is needed for many applications, such as statistics, algorithms and cryptography. However, most physical sources are not truly random, in the sense that they can have substantial biases and correlations. Weak random sources can also arise in cryptography when an adversary can learn partial information about a uniformly random string.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019).
Editors: Dimitris Achlioptas and László A. Végh; Article No. 72; pp. 72:1–72:22

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A natural approach to dealing with weak random sources is to apply a randomness extractor – a function that transforms a weak random source into an almost-perfect random source. However, it is impossible to give a single function that extracts even one bit of randomness from sufficiently general classes of sources [24]. There are two ways to combat this. One is to extract with the help of another short random string. An object constructed in this manner is called a seeded extractor [21]. The focus of this paper is the second way: to extract from more structured sources (without using additional random bits). Such a function is called a seedless, or deterministic, extractor.

More formally, a random source X is modeled as a probability distribution over n bit strings with some entropy k. In the context of randomness extraction, the standard measure of entropy is the so called min-entropy – the min-entropy k of a source X is defined as $H_\infty(X) = \min_s(\log(1/\Pr[X = s]))$. Then, the definition of a seedless extractor can be presented as follows.

▶ **Definition 1** (Seedless extractors for structured sources)**.** *Let $\mathcal{D}$ be a class of distributions over $\{0,1\}^n$. We say a function $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ is a (k,ϵ)-extractor for $\mathcal{D}$ if for any distribution $D \in \mathcal{D}$ with min-entropy at least k, we have*

$$\mathsf{Ext}(D) \approx_\epsilon U_m,$$

where U_m denotes the uniform distribution over $\{0,1\}^m$ and $\approx_\epsilon$ stands for ϵ-close in statistical distance (Definition 16).

By the probabilistic method, it is known that for any constant $\alpha > 0$ and any distribution family $\mathcal{D}$ of at most $2^{2^{(1-\alpha)k}}$ sources of min-entropy k, there is a seedless extractor outputting $m = (1-\alpha)k$ bits with error $2^{-\alpha k/3}$.

A large body of research has been devoted to constructing explicit seedless extractors for various structured sources. There are mainly two natural perspectives to limit the structure of a distribution: an algebraic perspective and a computational perspective.

The algebraic perspective is to impose some algebraic structure on the distribution, such as an affine source [5]. Later, affine sources were generalized to distributions defined using low-degree polynomials. On one hand, Dvir, Gabizon and Wigderson [10] studied polynomial sources, which are the images of low-degree polynomial maps. On the other hand, viewing an affine source as the kernel, or set of zeros, of an affine mapping, Dvir [9] introduced the class of sources sampled uniformly from kernels or sets of common zeros of one or more polynomials, which he called algebraic sources[1].

The computational perspective is to assume a distribution has "low complexity". This started with Trevisan and Vadhan [27], who considered distributions that can be sampled by efficient algorithms. They showed that constructing a seedless extractor for this class is closely related to proving lower bound for circuits and gave a conditional construction of such an extractor based on lower bound assumptions. Later, in [15], an unconditional extractor was constructed for sources generated by space-bounded algorithms. More recently, Viola [29] constructed a seedless extractor for AC^0-samplable sources.

[1] For clarification, in [9], Dvir mentioned sources which are distributed uniformly on varieties. A variety is also a set of common zeros of one or more polynomials, but it is often defined to require the ground field to be algebraically closed.

1.1 Recognizable sources

We focus on recognizable sources, first suggested by Shaltiel [25]. Recognizable sources are uniform distributions over sets of the form $\{x : f(x) = v\}$ for functions f coming from some specified class. Formally, for any boolean function $f : \{0,1\}^n \to \{0,1\}$, define the source recognizable by f, denoted by U_f, as the uniform distribution over $f^{-1}(1)$. For short, we call this distribution the f-recognizable source. For any boolean function family $\mathcal{C}$, the set of $\mathcal{C}$-recognizable sources is the set of f-recognizable sources, for each $f \in \mathcal{C}$.

This notion naturally interacts with the algebraic and computational perspectives to limit the structure of a distribution, and also captures several distributions that were widely studied. For example, distributions with algebraic structures are those distributions recognizable by algebraic classes – affine sources are distributions recognizable by affine functions and algebraic sources are distributions recognizable by products of low-degree polynomials. Moreover, distributions that have "low complexity" could also be the distributions recognizable by low-complexity classes, such as small circuits.

Shaltiel [25] initially proposed an extractor for recognizable sources. He showed that such extractors produced randomness that was in some sense not correlated with the input and hence could be used for derandomization. In particular, to derandomize any class of randomized algorithms, he needed to explicitly construct an extractor for distributions recognizable by the class. He showed that without further changes, some appropriate known extractors could work for distributions recognizable by decision trees, streaming algorithms, and AC^0. What's more, assuming average-case hardness against polynomial-size circuits, he showed that applying the hard function on disjoint blocks of the input was an extractor for distributions recognizable by general polynomial-time algorithms.

Later, Kinne, Melkebeek and Shaltiel [16] improved the derandomization results in [25] by using "seed-extending pseudorandom generators", which are pseudorandom generators that reveal their seed. They gave reductions between seed-extending PRGs and extractors for recognizable sources. However, both Shaltiel [25] and this later paper [16] focused on derandomization rather than constructing new extractors.

1.2 XOR Amplification

Given a boolean function $f : \{0,1\}^n \to \{0,1\}$, let $f^{\oplus m}(x_1, \ldots, x_m) := \bigoplus_{i \in [m]} f(x_i)$ denote the XOR of m independent copies of f. The XOR Amplification Lemma[2] states that if a function f is hard on average for some computational class $\mathcal{C}$, (i.e., f cannot be computed correctly by any function in $\mathcal{C}$ on at most a $(1/2 + p)$-fraction of of the inputs), then $f^{\oplus m}$ cannot be computed correctly on at most a $(1/2 + p^{\Omega(m)})$-fraction of of the inputs. Loosely speaking, the hardness of f is amplified when the outputs of independent copies of f are XOR together. Indeed, this idea is analogous to the information theoretic setting. If f is a biased coin with $\Pr[f = 1] = 1/2 + p$, then the XOR of m independent biased coins, $f^{\oplus m}$, induces a coin with $\Pr[f^{\oplus m} = 1] = 1/2 - (-2p)^m/2$. However, showing that such an idea holds in the computational setting is significantly more involved.

There are several works dedicated to proving XOR amplification for computational models. Yao [31] first suggested XOR amplification, and proved that XOR (hardness) amplification held for polynomial-size circuits. Unfortunately, the amplification stops when XORing more than logarithmically many copies, which makes it not so useful for us. Later, Viola and

[2] This is usually called simply the XOR lemma, or Yao's XOR lemma, but we want to distinguish it from a different XOR lemma.

Wigderson [30] showed XOR amplification for multi-party communication complexity and polynomials over GF(2). Subsequently, their proof was extended by Bogdanov, Kawachi and Tanaka [4], to prove XOR amplification for polynomials over any prime field.

In this paper, we give a new application of XOR amplification – constructing seedless extractors for recognizable sources.

2 Overview of our results

2.1 From XOR amplification to Extractors for recognizable sources

It is folklore that one can use correlation bounds to extract a single bit. In this paper, we use XOR amplification to extend the output length from one bit to linear in the input length.

Intuitively, XOR amplification states that if a function f is hard on average for some complexity class $\mathcal{C}$ of Boolean functions, then $f^{\oplus m}(x_1, \ldots, x_m) = f(x_1) \oplus \cdots \oplus f(x_m)$ is exponentially harder on average. We actually only need a weaker condition: that there exists some h for which $h^{\oplus k}$ gets exponentially harder.

More precisely, let $\mathcal{C} \subseteq \{\{0,1\}^* \to \{0,1\}\}$ be a class of Boolean functions. For a positive constant α, we say $\mathcal{C}$ has α-XOR amplification if there exists a function $h : \{0,1\}^t \to \{0,1\}$ such that for any positive integer k, the correlation between $h^{\oplus k}$ and g is no more than $2^{-\alpha k}$, for any $g \in \mathcal{C}$.

We show that if $\mathcal{C}$ is closed under restrictions and $\mathcal{C}$ has α-XOR amplification, then there is an efficient extractor for $\mathcal{C}_n$-recognizable sources, where $\mathcal{C}_n$ denotes the set of all n-variate functions in $\mathcal{C}$.

▶ **Theorem 2.** *Let $\mathcal{C} \subseteq \{\{0,1\}^* \to \{0,1\}\}$ be any boolean function class closed under restrictions and α be any positive constant. If $\mathcal{C}$ has α-XOR amplification, then for any positive integer n, there is an explicit seedless $((1-\beta)n, 2^{-\Omega(\alpha n)})$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ for $\mathcal{C}_n$-recognizable sources, where $\beta = \Theta(\alpha) > 0$, $m = \Omega(\alpha n)$, and $\mathcal{C}_n$ denotes the set of all n-variate functions in $\mathcal{C}$.*

Our construction uses $h : \{0,1\}^t \to \{0,1\}$ from the definition of XOR amplification. Since the function h is fixed, its input length t is a constant, and it is computable efficiently (by hardwiring it). We also use the generator matrix M of an asymptotically good $[l, m, r]$-code, where $l = n/t$, so the distance $r = \Omega(l) = \Omega(n/t)$. Then $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ is simply

$$\mathsf{Ext}(x) = h^{(l)}(x)M, \text{ where } h^{(l)}(x = (x_1, \ldots, x_l)) = (h(x_1), \ldots, h(x_l)).$$

Occasionally we will apply a variation of this theorem when t grows with n, in which case we need h to be computable in time polynomial in n. For example, if the input length of h is $t = O(\log n)$, then h should be computable in exponential time.

Li [18] uses a similar construction to extend the output length of two-source extractors from one bit to more. Raz [22] had a related but different way to extend the output length of his specific extractor using small biased spaces. Raz uses the XOR lemma, but his method would not work with any asymptotically good code (as ours and Li's does); in fact, Raz does not even mention codes or distance.

2.1.1 Algebraic sources

An algebraic set is a set of common zeros of one or more multivariate polynomials defined over a finite field $\mathbb{F}$. An *algebraic source* is a random variable distributed uniformly over an algebraic set, which was originally introduced by Dvir [9]. Algebraic sources are a natural

generalization of affine sources that have been widely studied. Furthermore, we say that an algebraic source has degree d if the algebraic source can be defined by polynomials of degree at most d.

▶ **Definition 3** (Algebraic extractor). *We say that* $\mathsf{Ext} : \mathbb{F}^n \to \mathbb{F}^m$ *is a* (k, d, ϵ)*-algebraic extractor over* $\mathbb{F}$ *if for any degree-*d *algebraic source* U_V *with* $|V| \geq |\mathbb{F}|^k$, $\mathsf{Ext}(U_V) \approx_\epsilon U_m$.

Dvir obtained explicit extractors for degree-d algebraic sources with entropy rate greater than $1/2$ over moderately sized fields, where $|\mathbb{F}| = \text{poly}(d)$, and with small entropy rate over large fields, where $|\mathbb{F}| = d^{\Omega(n^2)}$.

Golovnev and Kulikov [12] related the study of Boolean dispersers for quadratic algebraic sets to improving circuit lower bounds. A disperser is a relaxation of an extractor, which is only required to output a non-constant bit from a weak random source. They posed the open question of constructing a disperser for any algebraic set of size $2^{0.03n}$ and defined by using at most $1.78n$ quadratic polynomials. Such a disperser yields a new circuit lower bound.

Nevertheless, to our knowledge, there were only two papers on explicitly constructing dispersers or extractors for algebraic sources over GF(2). Cohen and Tal [8] constructed an extractor for algebraic sources defined by at most $(\log \log n)^{1/(2e)}$ quadratic polynomials. They also constructed dispersers for algebraic sources defined by at most n^α polynomials of degree at most $\log^{0.1}(n)$ for some constant $\alpha < 1$. Our extractor construction subsumes both their extractor and disperser, outputting n^γ random bits for algebraic sources with higher degree $c \log n$ and the same bound n^α for the number of defining polynomials, where γ, c are constants. Remscrim [23] constructed the best extractors before our work, outputting one bit with error $O(1/\sqrt{n})$ for min-entropy $n - n^c$ for any $c < 1/2$. It can handle fairly large degree, up to $n^{1/2-\alpha}$, where $\alpha > 0$ is a constant. Our construction significantly improves the extractor for constant-degree algebraic sources, outputting more bits and handling lower min-entropy.

Using Theorem 2, we construct a seedless extractor for algebraic sources of constant degree for some linear min-entropy. In particular, the new extractor has linear output length and exponentially small error for min-entropy $k \geq (1 - \alpha)n$, where $\alpha > 0$ is a small enough constant.

▶ **Theorem 4.** *For any positive integer* d*, there is an efficient* $\left((1 - 1/c_d)n, d, 2^{-\Omega(n/c_d)}\right)$*-algebraic extractor* $\mathsf{Ext} : \mathbb{F}_2^n \to \mathbb{F}_2^m$*, where* $c_d = \Theta(d^2 4^d)$*,* $m = \Omega(n/c_d)$.

Even for degree $c \log n$ for a small enough constant $c > 0$, our extractor outputs n^γ bits with error $2^{-\Omega(n^\alpha)}$ for $n - n^\alpha$ min-entropy, where $\gamma, \alpha > 0$ are some constants.

We can extend our algebraic extractor to any prime field $\mathbb{F}_q$.

▶ **Theorem 5.** *For any positive integer* d *and any prime field* $\mathbb{F}_q$*, there is an efficient* $\left((1 - 1/c_{d,q})n, d, q^{-\Omega(n/c_{d,q})}\right)$*-algebraic extractor* $\mathsf{Ext} : \mathbb{F}_q^n \to \mathbb{F}_q^m$*, where* $c_{d,q} = \Theta(d^2 2^{2d} q^3 \log q)$*,* $m = \Omega(n/c_{d,q})$.

2.1.2 Sources recognizable by communication protocols

We consider a boolean function class that has low communication complexity. Communication complexity was defined by Yao [32], who introduced a standard 2-party communication model. Later, Chandra, Furst, and Lipton [6] generalized this to the multiparty model. In a t-party communication NOF (number-on-forehead) model, each party holds a separate input and each party knows all but its own input. These parties attempt to compute (or approximate) a given function of these t inputs by exchanging few bits of communication. The

complexity of a communication protocol is the number of bits exchanged on the worst input. Both deterministic and randomized communication protocols are considered. A randomized protocol can be viewed as a distribution on deterministic protocols.

For deterministic 2-party protocols, Shaltiel [25] already constructed an efficient extractor that has linear output for linear min-entropy and exponentially small error. To do this, he proved that 2-source extractors are also extractors for sources recognizable by deterministic 2-party protocols, and hence some known constructions of 2-source extractors could be used. However, this approach is tailored to the 2-party case and does not generalize to the t-party case for some $t > 2$.

We construct an extractor for sources recognizable by randomized t-party protocols. Formally, we prove the following theorem.

▶ **Theorem 6.** *There exists an explicit seedless $((1 - 1/c_t)n,\ 2^{-c_1 n/c_t})$ extractor $\mathsf{Ext} : \left(\{0,1\}^{n/t}\right)^t \to \{0,1\}^{c_2 n/c_t}$ for sources recognizable by randomized t-party communication protocols of at most $c_3 n/4^t$ bits, where $c_t = \Theta(t4^t)$ and c_1, c_2, c_3 are some positive constants.*

This extractor has linear output for linear min-entropy and exponentially small error, and is simply $\mathsf{Ext}(x) = \left(\wedge_t^{(l)}(x)\right) M$, where $l = n/t$, $\wedge_t$ is the AND function over t variables and M is the $l \times (c_2 n/c_t)$ generator matrix of a good linear code.

2.1.3 Halfspace sources

Halfspace sources are sources recognizable by linear threshold functions. A linear threshold function (abbreviated LTF) is a boolean function $f : \{0,1\}^n \to \{0,1\}$ that can be represented as $f(x) = 1_{\sum_{i \in n} a_i x_i > a_0}$ for some constants $a_0, a_1, \ldots, a_n \in \mathbb{R}$. From a geometric perspective, a boolean LTF is a halfspace-indicator to the discrete cube $\{0,1\}^n$.

We construct an efficient extractor that has linear output for linear min-entropy and exponentially small error for halfspace sources.

▶ **Theorem 7.** *There exists an explicit seedless $((1 - c_1)n,\ 2^{-c_2 n})$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{c_3 n}$ for halfspace sources, where c_1, c_2, c_3 are some positive small enough constants.*

The construction of this extractor is simply $\mathsf{Ext}(x) = \left(\wedge_2^{(l)}(x)\right) M$, where $l = n/2$, M is the $l \times c_3 n$ generator matrix of a good linear code.

2.2 From Seed-extending PRGs to Extractors for recognizable sources

The Kinne et al. reductions between seed-extending pseudorandom generators and extractors for recognizable distributions were asymmetric. They showed that an extractor with exponentially small error yielded a seed-extending pseudorandom generator with exponentially small error. However, they proved a weak converse.

In this paper, we prove that a seed-extending pseudorandom generator with exponentially small error yields an extractor with exponentially small error. This applies to flip-invariant families of boolean functions, which are invariant under flipping input bits (see Definition 26).

▶ **Theorem 8.** *Let $\mathcal{C}$ be a flip-invariant family of boolean functions over n bits. If G is a seed-extending (d, ϵ)-pseudorandom generator $G : \{0,1\}^d \to \{0,1\}^n$ for $\mathcal{C}$, then for any $\Delta = \Delta(n) > 0$ we can construct an $(n - \Delta, 2^{\Delta}\epsilon)$-extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-d}$ for $\mathcal{C}$-recognizable sources. Specifically, if $G(x) = (x, E(x))$ fools any function in $\mathcal{C}$, then $\mathsf{Ext}(x \circ y) = y \oplus E(x)$ is an $(n - \Delta, 2^{\Delta}\epsilon)$-extractor for $\mathcal{C}$-recognizable sources, where $x \in \{0,1\}^d, y \in \{0,1\}^m$, where $m = n - d$.*

In particular, the reduction in [16] requires a tiny $\epsilon \leq 2^{-(m+2\Delta)}$ for the seed-extending PRG to get an $(n-\Delta, 2^{-\Delta})$-extractor. Moreover, the reduction in [16] breaks down for a seed-extending PRG, $G(x) = (x, E(x))$, where $E(x)$ is longer than x. We improve the reduction from seed-extending PRGs to extractors to require only $\epsilon \leq 2^{-2\Delta}$, without depending on the output length m. Furthermore, the new reduction can still work even for a seed-extending PRG, $G(x) = (x, E(x))$, where $E(x)$ is longer than x.

Based on this new reduction, we significantly improve extractors for two important types of recognizable sources as follows.

2.2.1 Circuit-recognizable sources

Kinne et al. proved that the well-known Nisan-Wigderson pseudorandom generator construction [20] can be made seed-extending. Therefore, assuming hardness against small circuits, we can construct an extractor for sources recognizable by small circuits.

▶ **Proposition 9.** *For any $\Delta = \Delta(n) > 0$ and positive integers $l < n$, if there is a function H that is ϵ-hard at input length $\sqrt{l}/2$ for circuits of size $s + (n-l)2^{O(\log(n-l)/\log l)}$ and depth $d+1$, then we can get an $(n-\Delta, (n-l)2^{\Delta}\epsilon)$-extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-l}$ for any sources recognizable by circuits of size s and depth d.*

Using the hardness of the parity function against AC^0 [13], we significantly improve Shaltiel's extractor [25] for AC^0-recognizable sources.

▶ **Theorem 10.** *For any $\Delta = \Delta(n) > 0$ and positive integers $l < n$, there exists a polynomial time computable $(n-\Delta, (n-l)2^{\Delta - \Omega(l^{1/(2d+2)})})$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-l}$ for any sources recognizable by circuits of size $2^{n^{1/d}}$ and depth d.*

In particular, for min-entropy $n - n^{1/(\alpha d)}$, our extractor outputs $n - n^{2/\alpha + O(1/d)}$ bits, whereas Shaltiel's extractor outputs only $n^{1/(\alpha d)}$ bits. When $\alpha > 2d/(d-1)$ is a large enough constant, our extractor outputs $n - o(n)$ bits whereas Shaltiel's extractor outputs only $n^{1/(\alpha d)}$ bits. For min-entropy $n - \text{polylog}(n)$ bits, our extractor outputs $n - \text{polylog}(n)$, whereas Shaltiel's extractor outputs only $\text{polylog}(n)$ bits.

Our methods also apply to formulas. Komargodski, Raz and Tal [17] constructed an explicit function $h : \{0,1\}^n \to \{0,1\}$ that is $2^{-\Omega(r)}$-hard for any deMorgan formula of size $n^{3-o(1)}/r^2$. Based on this hardness result, we can construct an efficient extractor for sources recognizable by deMorgan formulas of size close to $n^{3/2}$.

▶ **Theorem 11.** *For any $\Delta, r, \alpha > 0$ and $m \leq (1-\alpha)n$, there exists a polynomial time computable $(n-\Delta, m2^{\Delta - \Omega(r)})$-extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ for any sources recognizable by deMorgan formulas of size $n^{3/2-o(1)}/r^2$.*

2.2.2 Sources recognizable by efficient randomized algorithms

Note that there are no efficient seed-extending cryptographic PRGs. Otherwise, with revealed seeds, it is easy to efficiently distinguish the output of an efficient seed-extending PRG, $G(x) = (x, E(x))$, from a random string (x, y), by checking whether y equals $E(x)$.

We show that there is an inefficient seed-extending cryptographic PRG implied by the existence of one-way permutations. By our reduction, we show that a one-way permutation with exponentially small error yields an $(n - n^{\Omega(1)}, 2^{-n^{\Omega(1)}})$ extractor extracting $n - n^{O(1)}$ bits from sources recognizable by BPP algorithms. Formally, this follows by taking $\epsilon = 2^{-cn^{\alpha}}$ and $q(n) = n^{w(1)}$ in the following theorem.

▶ **Theorem 12.** *For any polynomial-time computable functions $t(\cdot)$ and $\epsilon(\cdot)$, assume that $f : \{0,1\}^* \to \{0,1\}^*$ is a one-way permutation with error $\epsilon(\cdot)$ against $t(\cdot)$-bounded inverters. Then for any $\Delta = \Delta(n) > 0$ and a positive constant $\delta < 1$, we can construct an $\left(n - \Delta, O\left(2^{\Delta}\epsilon(n^{\delta})^{c_\delta}\right)\right)$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-n^{\delta}}$ for sources recognizable by randomized algorithms running in time $\left(t(n^{\delta})\right)^{c_\delta}$, where c_δ is a constant depending on δ. The running time of the extractor is a polynomial times the time to compute the inverse function f^{-1} of the one-way permutation f with input length n^{δ}. Due to the space limitation, we prove the following theorem in the full version of this paper.*

Furthermore, the running time of such an extractors will be quasi-polynomial if there exists a sufficiently strong one-way permutations. In particular, by scaling down, we have the following corollary.

▶ **Corollary 13.** *For any constants $a, b, c, \delta > 0$, assume that there exists a one-way permutation invertible in time $O\left(2^{n^a}\right)$ with error 2^{-n^c} against $2^{\delta n^b}$-bounded inverters, Then, for any positive constants α and $\beta < 1$, we can get an $\left(n - c_\beta \log^{c\alpha}(n), O\left(2^{-c_\beta \log^{c\alpha}(n)}\right)\right)$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-n^{\beta}}$ for sources recognizable by randomized algorithms running in time $2^{c_\beta \delta \log^{b\alpha}(n)}$, where c_β is a constant depending on β.The running time of the extractor is $O\left(2^{\log^{a\alpha}(n^{\beta})}\right)$.*

3 Overview of our main constructions and proofs

3.1 From XOR amplification to Extractors

In this subsection, we describe how to construct a seedless extractor for $\mathcal{C}$-recognizable sources if there exists a function $h : \{0,1\}^t \to \{0,1\}$ such that for any $g \in \mathcal{C}$ and $k \leq n/t$, $Cor(h^{\oplus k}, g) \leq 2^{-\Omega(k)}$. Think of $t = O(1)$.

We start with the statistical XOR lemma[3], usually attributed to Vazirani. We say a random variable Z over $\{0,1\}$ is ϵ-biased if $bias(Z) = Cor(Z, 0) = |\Pr[Z = 0] - \Pr[Z = 1]| \leq \epsilon$.

▶ **Lemma 14** (Statistical XOR Lemma)**.** *Let $X_1, \ldots, X_m$ be 0-1 random variables such that for any nonempty $S \subseteq \{1, \ldots, m\}$, the random variable $\bigoplus_{i \in S} X_i$ is ϵ-biased. Then, the distribution of $(X_1, \ldots, X_m)$ is $\epsilon 2^{m/2}$-close to uniform.*

Let $g_i(x)$ be the i-th bit of $\mathsf{Ext}(x)$ for each $i \in [m]$. Thus, to show that the output of Ext is close to uniform, it suffices to show that for any non-empty set $S \subseteq [m]$, $g_S = \sum_{i \in S} g_i$ is low-biased conditioned on $f(x) = 1$ for each $f \in \mathcal{C}$. By XOR amplification, it is enough to guarantee that each g_S is the sum of $\Omega(n)$ independent copies of h, and hence g_S has $2^{-\Omega(n)}$ correlation with any function in $\mathcal{C}$.

A linear code is a natural candidate to guarantee that each g_S is the sum of $\Omega(n)$ independent copies. Let $h^{(l)} : \{0,1\}^{tl} \to \{0,1\}$ denote the concatenation of l copies of h and M be the generating matrix of an asymptotically good $[l, m, r]_2$ code. Our construction is simply

$$\mathsf{Ext}(x) = (g_1(x), \ldots, g_m(x)) = h^{(l)}(x)M.$$

Finally, we observe that the bias of g_S conditioned on $f(x) = 1$ can be bounded by the correlation between g_S and f plus the bias of g_S.

[3] The statistical XOR lemma is unrelated to the XOR amplification used in our proof.

▶ **Lemma 15.** $|\Pr[g_S(X)=1|f(X)=1]-\Pr[g_S(X)=0|f(X)=1]| \leq \frac{Cor(g_S,f)+bias(g_S)}{2\Pr[f(X)=1]}$.

That is, if we choose a good linear code, then $\mathsf{Ext}(x) = h^{(l)}(x)M$ is an extractor for $\mathcal{C}$-recognizable sources with exponentially small error.

For details, see Section 5.

3.2 Algebraic extractors over GF(2)

In this subsection, we describe our algebraic extractor construction.

Notice that to construct a degree-d algebraic extractor that outputs only one bit, it is enough to let the extractor have small correlation bounds with degree-d polynomials. This fact is implicitly proved by Dvir [9] and observed by others, e.g., Eshan Chattopadhyay and Avishay Tal (personal communication). Based on this fact, we combine XOR amplification and linear codes to extend the output length form one bit to more.

First we observe that an algebraic source over n bits defined by n-variate polynomials $p_1,\ldots,p_k$ is also a source recognizable by the product $\prod_{i\in[k]}(p_i+1)$. Let $\mathcal{V}_d$ denote the set of all products of polynomials of degree at most d. Thus, for any positive integer n, to get an extractor for n-bit algebraic sources of degree d, it suffices to construct an extractor for $\mathcal{V}_d$-recognizable sources over n bits. In particular, by the previous discussion, it suffices to show that XOR amplification holds for $\mathcal{V}_d$.

Second we observe that to show that a function f has low correlations with $\mathcal{V}_d$, it suffices to show that f has low correlation with any d-degree polynomials. This is because the L1 norm of the Fourier transform of the AND function is at most 2.

Viola and Wigderson [30] proved XOR amplification for low-degree polynomials over GF(2). Specifically, if a Boolean function h over $\{0,1\}^{O(d)}$ has correlation at most $1-1/2^d$ with degree-d polynomials, then the correlation between $h^{\oplus l}$ (see Section 1.2) and degree-d polynomials drops exponentially with l. Such h are known.

For details, see Section A.1.

3.3 From seed-extending PRGs to Extractors

We start with a new reduction from pseudorandom generators to seedless extractors. Observe that a seedless extractor $\mathsf{Ext}:\{0,1\}^n\to\{0,1\}^m$ partitions $\{0,1\}^n$ as $\bigcup_{z\in\{0,1\}^m}\mathsf{Ext}^{-1}(z)$. If Ext is a (k,ϵ)-extractor for $\mathcal{C}$-recognizable sources, then for every $f\in\mathcal{C}$ with $|f^{-1}(1)|\geq 2^k$, most intersections $\mathsf{Ext}^{-1}(z)\cap f^{-1}(1)$ should have almost the same size. That is, for most m-bit strings z, the preimage $\mathsf{Ext}^{-1}(z)$ is an ϵ-pseudorandom set against any $f\in\mathcal{C}$ with $|f^{-1}(1)|\geq 2^k$.

Now, given PRGs, how do we construct extractors? From the above observation, converting an ϵ-pseudorandom set into a partition of ϵ-pseudorandom sets is a possible way. If each preimage $\mathsf{Ext}^{-1}(z)$ of Ext is an ϵ-pseudorandom set for $\mathcal{C}$, Ext should be an extractor for $\mathcal{C}$-recognizable sources with a bit worse parameters.

To make $\mathsf{Ext}^{-1}(z)$ an ϵ-pseudorandom set for each z, we need a seed-extending PRG $G(x)$, i.e., $G(x)=x\circ E(x)$ for some function $E:\{0,1\}^d\to\{0,1\}^{n-d}$. By linearly shifting the set $\{(x,E(x))\}$, we can partition $\{0,1\}^n$ as $\bigcup_{z\in\{0,1\}^{n-d}}\{(x,(E(x)\oplus z)):x\in\{0,1\}^d\}$. We therefore define $\mathsf{Ext}(x,z)=E(x)\oplus z$. Since $\mathcal{C}$ is a flip-invariant function family, we have that the set $\mathsf{Ext}^{-1}(z)=\{(x,(E(x)\oplus z)):x\in\{0,1\}^d\}$ fools any function f in $\mathcal{C}$.

For details, see Section 6.

3.4 Algebraic extractors over prime fields

We remark that the main results used in building our algebraic extractor over GF(2) – the XOR amplification, the statistical XOR lemma and the asymptotically linear code – all have been extended to prime fields. Thus, to generalize our algebraic extractor, the remaining technical parts are not hard.

Bogdanov, Kawachi and Tanaka [4] proved XOR amplification for low-degree polynomials over prime fields, i.e., the sum of k independent copies of h was $q^{-\Omega(k)}$-hard for P_d if h was mildly hard. However, besides the sum of copies, we require the same hardness result for linear combinations of k copies of h. We prove this hardness result by using the original proof of Bogdanov, Kawachi and Tanaka with some slight modifications. The main revision of our proof uses the fact that the Gowers norm is multiplicative for functions over disjoint sets of input variables.

Furthermore, over a prime field $\mathbb{F}_q$, an algebraic source over n bits defined by n-variate polynomials $p_1, \ldots, p_k$ is a source recognizable by the product $\prod_{i\in[k]}(1-p_i^{q-1})$. We need to analyze the product of the special form $\prod_{i\in[k]}(1-x_i^{q-1})$, as an analog of the AND function over GF(2).

The reason we assume prime fields in our results is that XOR amplification for polynomials is known only over prime fields.

For details, please check the full version of this paper.

4 Preliminaries

In the following, for any two binary strings x, y, let $x \circ y$ denote their concatenation, and let $x \oplus y$ denote their bitwise XOR when x and y have the same length.

▶ **Definition 16** (Statistical distance). *Let D_1 and D_2 be two distributions over a set S. Define the* statistical distance *between D_1 and D_2 as $|D_1 - D_2| = \frac{1}{2}\sum_{s\in S}|\Pr[D_1 = s] - \Pr[D_2 = s]|$. We say D_1 is ϵ-close to D_2, denoted by $D_1 \approx_\epsilon D_2$, if $|D_1 - D_2| \le \epsilon$.*

▶ **Definition 17** (Recognizable source). *For any boolean function $f : \{0,1\}^n \to \{0,1\}$, define the source recognizable by f, denoted by U_f, as the uniform distribution over $f^{-1}(1)$. For short, we call this distribution the f-recognizable source.*

For any boolean function family $\mathcal{C}$, the set of $\mathcal{C}$-recognizable sources is the set of f-recognizable sources for $f \in \mathcal{C}$.

For $l \in \mathbb{N}$, let U_l denote the uniform distribution on l bits.

▶ **Definition 18** (Extractor for recognizable sources [25]). *Let $\mathcal{C}$ be a class of functions $C : \{0,1\}^n \to \{0,1\}$. We say that $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ is a (k,ϵ)-extractor for $\mathcal{C}$-recognizable sources if for every $f \in \mathcal{C}$ such that $\left|f^{-1}(1)\right| \ge 2^k$, $\mathsf{Ext}(U_f) \approx_\epsilon U_m$.*

Note that when the output length $m = 1$, the extractor is simply a boolean function which has low correlation with any function in $\mathcal{C}$.

4.1 Algebraic sources

An algebraic set is a set of common zeros of one or more multivariate polynomials defined over a finite field $\mathbb{F}$.

▶ **Definition 19** (Algebraic set). *For any s polynomials $f_1, \ldots, f_s \in \mathbb{F}[x_1, \ldots, x_n]$, the set $V(f_1, \ldots, f_s) = \{x \in \mathbb{F}^n | f_i(x) = 0, \forall i \in [s]\}$ is an algebraic set. We say V is an algebraic set of degree d, if each polynomial f_i has degree at most d.*

An *algebraic source* is a random variable distributed uniformly over an algebraic set as initially defined by Dvir [9].

▶ **Definition 20** (Algebraic source). *An algebraic source is the uniform distribution U_V over an algebraic set V. If V is a degree-d algebraic set, then we say U_V is an algebraic source of degree d.*

▶ **Definition 21** (Algebraic extractor). *We say that $\mathsf{Ext} : \mathbb{F}^n \to \mathbb{F}^m$ is a (k, d, ϵ)-algebraic extractor if for any degree-d algebraic source U_V with $|V| \geq |\mathbb{F}|^k$, $\mathsf{Ext}(U_V) \approx_\epsilon U_m$.*

▶ **Definition 22** (Linear codes over prime fields). *For a prime q, a linear code of length n and dimension k is a k-dimensional linear subspace C of the vector space $\mathbb{F}_q^n$. If the distance of the code C is d, i.e., the minimum number of two codewords in which they differ, we say that C is an $[n, k, d]_q$ code. A family of codes $\{C_n\}$ is asymptotically good if there exist constants $0 < \delta_1, \delta_2 < 1$ s.t. $k \geq \delta_1 n$ and $d \geq \delta_2 n$.*

Note that every linear code has an associated generating matrix $M \in \mathbb{F}_q^{k \times n}$, and every codeword can be expressed as vM, for some vector $v \in \mathbb{F}_q^k$. There are explicit constructions of asymptotically good linear codes, such as the Justensen codes over GF(2) constructed in [14] and the expander codes over GF(q) in [1] for any prime q.

▶ **Definition 23** (Correlation over prime fields). *Let $f, g : \mathbb{F}_q^n \to \mathbb{F}_q$ be two functions over n inputs. The correlation between f and g with respect to the uniform distribution is defined as*

$$Cor(f, g) := |Ee_q[f(x) + g(x)]| \in [0, 1],$$

where $e_q[x] := w^x$ for $x \in \{0, 1, \ldots, q-1\}$, where w denotes the q-th root of unity.

For a class $\mathcal{C}$ of functions, we denote by $Cor(f, \mathcal{C})$ the maximum of $Cor(f, C)$ over all $C \in \mathcal{C}$ whose domain is the same as f.

Furthermore, when $q = 2$, we have $e_2[x] = (-1)^x$, and $Cor(f, g) = |\Pr[f(x) = g(x)] - \Pr[f(x) \neq g(x)]|$. We often write $e_2[x]$ as $e[x]$ for convenience.

▶ **Definition 24** ($f^{(m)}, f^v$). *For any function $f : \mathbb{F}_q^n \to \mathbb{F}_q$, let $f^{(m)}$ denote the concatenation of m copies of f, i.e., $f^{(m)}(x_1, x_2, \ldots, x_m) := (f(x_1), \ldots, f(x_m))$, where $x_1, \ldots, x_m \in \mathbb{F}_q^n$. For each $v = (v_1, \ldots, v_m) \in \mathbb{F}_q^m$, let f^v denote the linear combination of m copies of f according to v, i.e., $f^v(x_1, x_2, \ldots, x_m) := \sum_{i \in [m]} v_i f(x_i)$.*

Let $\mathbb{F}_q^* = \mathbb{F}_q \setminus \{0\}$ denotes the set of non-zero elements in $\mathbb{F}_q$. We remark that the statistical XOR lemma has been generalized to prime fields by e.g., Goldreich [11].

▶ **Lemma 25** (Statistical XOR Lemma over $\mathbb{F}_q$). *Let $X = (X_1, \ldots, X_m)$ be random vector over $\mathbb{F}_q^m$ such that for any nonzero vector $v = (v_1, \ldots, v_m) \in \mathbb{F}_q^m \setminus \{0^m\}$, the random variable $v \cdot X = \sum_{i \in [m]} v_i X_i$ is ϵ-biased. Then, the distribution of $(X_1, \ldots, X_m)$ is $\epsilon q^{m/2}$-close to the uniform distribution over $\mathbb{F}_q^m$.*

For example, when $m = 1$, for a random variable X over $\mathbb{F}_q$, to show that $X \approx_\epsilon U_{\mathbb{F}_q}$, we need to show that $bias(\alpha X) \leq \epsilon/\sqrt{q}$ for each $\alpha \in \mathbb{F}_q^*$.

4.2 Seed-extending PRGs

▶ **Definition 26** (Flip-invariant family). *We say a boolean function family $\mathcal{C}$ over n bits is* flip-invariant *if for any string $s \in \{0, 1\}^n$, $f \in \mathcal{C}$ implies $f(x \oplus s) \in \mathcal{C}$.*

▶ **Definition 27** (Seed-extending pseudorandom generator). *A seed-extending pseudorandom generator is a generator G that outputs the seed as part of the pseudorandom string.*

Formally, a seed-extending (d, ϵ)-pseudorandom generator $G : \{0,1\}^d \to \{0,1\}^n$ for a class of functions over n bits, is a seed-extending function, i.e., $G(s) = (s, E(s))$ for some function E, such that

$$|\Pr[f(G(U_d)) = 1] - \Pr[f(U_n) = 1]| \leq \epsilon.$$

5 From XOR Amplification to Extractors for Recognizable Sources

First we define XOR amplification for a boolean function class that contains functions with various input lengths. Recall that $f^{\oplus m}(x_1, \ldots, x_m) = \bigoplus_{i \in [m]} f(x_i)$.

▶ **Definition 28** (α-XOR amplification for a boolean function class). *Let $\mathcal{C} \subseteq \{\{0,1\}^* \to \{0,1\}\}$ be a class of boolean functions. For a positive constant α, we say $\mathcal{C}$ has α-XOR amplification if there exists a function $h : \{0,1\}^t \to \{0,1\}$ such that for any positive integer k, $Cor(h^{\oplus k}, g) \leq 2^{-\alpha k}$, for any $g \in \mathcal{C}$.*

However, for constructing extractors for n-bit recognizable sources, we need to focus on the specific subset $\mathcal{C}_n \subseteq \mathcal{C}$ that contains all n-variate functions in $\mathcal{C}$. We define XOR amplification for $\mathcal{C}_n$ to also allow fixing some input bits.

▶ **Definition 29** ((α, w)-XOR amplification for functions with a fixed input length). *For a set $\mathcal{C}_n$ of n-variate functions $C : \{0,1\}^n \to \{0,1\}$ and a positive constant α, we say $\mathcal{C}_n$ has (α, w)-XOR amplification for a function $h : \{0,1\}^t \to \{0,1\}$ if for any vector $v \in \{0,1\}^{\lfloor n/t \rfloor}$ with at least w ones, $Cor(h^v, \mathcal{C}_n) \leq 2^{-\alpha w}$, where we add dummy variables to the input of h^v if h^v has less than n input variables.*

Moreover, we say C_n has α-XOR amplification for h, if C_n has (α, w)-XOR amplification for h for each positive integer $w \leq \lfloor n/t \rfloor$.

Note that if $\mathcal{C}$ is closed under restrictions, the fact that $\mathcal{C}$ has α-XOR amplification implies that $\mathcal{C}_n$ has also α-XOR amplification for every positive integer n. Formally,

▶ **Lemma 30.** *Let $\mathcal{C} \subseteq \{\{0,1\}^* \to \{0,1\}\}$ be a class of boolean functions closed under restrictions. Let $\mathcal{C}_n \subseteq \mathcal{C}$ denote the set of all n-variate functions in $\mathcal{C}$. If $\mathcal{C}$ has α-XOR amplification for a function $h : \{0,1\}^t \to \{0,1\}$, then $\mathcal{C}_n$ has also α-XOR amplification for h for every positive integer n.*

Proof. Assume that $\mathcal{C}$ has α-XOR amplification for a function $h : \{0,1\}^t \to \{0,1\}$, i.e., $Cor(h^{\oplus k}, \mathcal{C}) \leq 2^{-\alpha k}$ for each positive integer k. Then, we need to prove that for every positive integer n, $\mathcal{C}_n$ has also α-XOR amplification for h. In particular, fix n and let $l = \lfloor n/t \rfloor$. It suffices to prove that for any vector $v \in \{0,1\}^l$ with k ones, $Cor(h^v, \mathcal{C}_n) \leq Cor(h^{\oplus k}, \mathcal{C})$, as $Cor(h^{\oplus k}, \mathcal{C}) \leq 2^{-\alpha k}$.

To prove this, without loss of generality, assume that the first k coordinates of v are all 1's, and the remaining coordinates are all 0's. Thus, h^v depends only on the first kt variables. For any n-variate function $C(x_1, \ldots, x_n) \in \mathcal{C}_n$,

$$\begin{aligned} Cor(h^v, C) &= E_{X\sim U_{kt}, Y\sim U_{n-kt}} e[h^v(X,Y) + C(X,Y)] \\ &= E_{Y\sim U_{n-kt}} \left[E_{X\sim U_{kt}} e[h^v(X,Y) + C(X,Y)]\right] \\ &\le \frac{1}{2^{n-kt}} \sum_{Y_0\in\{0,1\}^{n-kt}} Cor\left(h^{\oplus k}(X), C(X,Y_0)\right) \\ &\le \frac{1}{2^{n-kt}} \sum_{Y_0\in\{0,1\}^{n-kt}} Cor(h^{\oplus k}, \mathcal{C}) \\ &= Cor(h^{\oplus k}, \mathcal{C}). \end{aligned}$$

The last inequality follows since $\mathcal{C}$ is closed under restrictions, i.e., $C(X, Y_0) \in \mathcal{C}$ for any $Y_0 \in \{0,1\}^{n-kt}$. ◀

▶ **Theorem 31.** *Let $\mathcal{C}_n$ be a family of boolean functions over n bits containing the constant function $f(x) = 0$. For any positive integers n, m, t, let M be the $l \times m$ generating matrix of an asymptotically good $[l, m, r_0]_2$ code, where $l = n/t$. Assume that $\mathcal{C}_n$ has (α, r)-XOR amplification for $h : \{0,1\}^t \to \{0,1\}$, where $r \le r_0$. Then, the function $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$,*

$$\mathsf{Ext}(x) = h^{(l)}(x)M,$$

is an $(n - \Delta, 2^{m/2+\Delta-\alpha r})$ extractor for $\mathcal{C}_n$-recognizable sources.

Proof. For convenience, let $(g_1(x), \ldots, g_m(x)) = h^{(l)}(x)M$. To show that the output of Ext is $2^{m/2+\Delta-\alpha r}$-closed to the uniform, by the statistical XOR Lemma, it suffices to show that for any non-empty set $S \subseteq [m]$, $g_S = \sum_{i\in S} g_i$ is $2^{\Delta-\alpha r}$-biased conditioned on $f(x) = 1$ for any $f \in \mathcal{C}_n$ with $|f^{-1}(1)| \ge 2^{n-\Delta}$.

First we observe that the bias of g_S conditioned on $f(x) = 1$ can be bounded by the correlation between g_S and f plus the bias of g_S.

▶ **Lemma 32** (Lemma 15, restated).

$$|\Pr[g_S(X) = 1 | f(X) = 1] - \Pr[g_S(X) = 0 | f(X) = 1]| \le \frac{Cor(g_S, f) + bias(g_S)}{2\Pr[f(X) = 1]}.$$

Proof. By multiplying $2\Pr[f(X) = 1]$ on both sides, it is equivalent to prove that

$$2\,|\Pr[g_S(X) = 1 \wedge f(X) = 1] - \Pr[g_S(X) = 0 \wedge f(X) = 1]| \le Cor(g_S, f) + bias(g_S).$$

Notice that

$$\begin{aligned} Cor(g_S, f) &= |\Pr[g_S(X) = f(X)] - \Pr[g_S(X) \ne f(X)]| \\ &= |\Pr[g_S(X) = 1 \wedge f(X) = 1] + \Pr[g_S(X) = 0 \wedge f(X) = 0] \\ &\quad - \Pr[g_S(X) = 0 \wedge f(X) = 1] - \Pr[g_S(X) = 1 \wedge f(X) = 0]|, \end{aligned}$$

and

$$\begin{aligned} bias(g_S) &= |\Pr[g_S(X) = 1] - \Pr[g_S(X) = 0]| \\ &= |\Pr[g_S(X) = 1 \wedge f(X) = 1] + \Pr[g_S(X) = 1 \wedge f(X) = 0] \\ &\quad - \Pr[g_S(X) = 0 \wedge f(X) = 1] - \Pr[g_S(X) = 0 \wedge f(X) = 0]|. \end{aligned}$$

Thus, by the triangle inequality,

$$\begin{aligned} bias(g_S) + Cor(g_S, f) &\geq |2\Pr[g_S(X) = 1 \wedge f(X) = 1] - 2\Pr[g_S(X) = 0 \wedge f(X) = 1]| \\ &= 2\left|\Pr[g_S(X) = 1 \wedge f(X) = 1] - \Pr[g_S(X) = 0 \wedge f(X) = 1]\right|. \end{aligned}$$ ◀

Then, observe that not only is each g_i a sum of at least r independent copies, but also so is any non-empty sum of the g_i, and hence has exponentially small correlation with degree-d polynomials.

▶ **Lemma 33.** *For any nonempty set $S \subseteq [m]$, $Cor(g_S, \mathcal{C}_n) \leq 2^{-\alpha r}$.*

Proof. Note that

$$g_S(x) = \sum_{i \in S} h^{(l)}(x) M_i = h^{(l)}(x) \left(\sum_{i \in S} M_i \right),$$

where M_i denotes the i-th row of the matrix M. As M is the generating matrix of an $[l, m, r]_2$ code and S is non-empty, $\sum_{i \in S} M_i$ is a codeword and hence has at least r 1's. Thus, g_S is the XOR of at least r_0 independent copies of h. By the assumed (α, r)-XOR amplification, we know $Cor(g_S, \mathcal{C}_n) \leq 2^{-\alpha r}$. ◀

Since the constant function $0 \in \mathcal{C}_n$, we also have that $bias(g_S) = Cor(g_S, 0) \leq 2^{-\alpha r}$. Thus, by Lemma 32, the bias of g_S conditioned on $f(x) = 1$ is at most $2^{-\alpha r}/p$, where $p = \Pr[f(X) = 1]$.

At last, we have $p = \frac{|f^{-1}(1)|}{2^n} \geq 2^{-\Delta}$ by the min-entropy requirement that $|f^{-1}(1)| \geq 2^{n-\Delta}$. Therefore, $g_S(x)$ is $2^{\Delta - \alpha r}$-biased conditioned on $f(x) = 1$. ◀

Combining with an explicit asymptotically good $[l, m, r]_2$ code, we prove the following theorem.

▶ **Theorem 34.** *Let $\mathcal{C} \subseteq \{\{0,1\}^* \to \{0,1\}\}$ be any boolean function class closed under restrictions and α be any positive constant. Let $\mathcal{C}_n$ denote the set of all n-variate functions in $\mathcal{C}$. If $\mathcal{C}_n$ has $(\alpha, \delta n)$-XOR amplification for $h : \{0,1\}^t \to \{0,1\}$, where $\delta < 1/t$ is a positive constant, then there is an explicit $(n - c_1\alpha l,\ 2^{-c_2\alpha l})$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{c_3\alpha l}$ for $\mathcal{C}_n$-recognizable sources, where $l = n/t$ and c_1, c_2, c_3 are some positive constants.*

Moreover, if $\mathcal{C}$ has α-XOR amplification for a function $h : \{0,1\}^t \to \{0,1\}$, then for any positive integer n, there is an explicit seedless $(n - c_1\alpha l,\ 2^{-c_2\alpha l})$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{c_3\alpha l}$ for $\mathcal{C}_n$-recognizable sources, where $l = n/t$ and c_1, c_2, c_3 are some positive constants.

Proof. Note that if $\mathcal{C}$ has α-XOR amplification for a function h, then by Lemma 30, $\mathcal{C}_n$ also has α-XOR amplification for h for every positive integer n, i.e., $\mathcal{C}_n$ also has $(\alpha, \delta l)$-XOR amplification for h by definition. Now, we start with the assumption that $\mathcal{C}_n$ has $(\alpha, \delta l)$-XOR amplification for h. We use an explicit $[l, \delta_1 l, \delta_2 l]_2$ linear code for some constants $\delta_1 > 0$ and $\delta_2 > \delta$ by Justesen [14]. Therefore, Theorem 31 yields an $(n - \Delta,\ 2^{m/2 + \Delta - \alpha\delta_2 l})$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ for $\mathcal{C}_n$-recognizable sources. That is, by setting $\Delta = c_1\alpha l$ and $m = c_3\alpha l$ for some small positive constants c_1, c_3, we get the desired $(n - c_1\alpha l,\ 2^{-c_2\alpha l})$ extractor, where $c_2 = -(c_3/2 + c_1 - \delta_2) > 0$ is also a positive constant. ◀

6 From Seed-Extending PRGs to Extractors for Recognizable Sources

Note that Kinne et al. [16] already showed reductions between extractors for recognizable sources and seed-extending PRGs.

▶ **Lemma 35** ([16, Theorem 7]). *Let $C : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}$ be a function. Let $\Delta = m + \log(1/\epsilon)$ and let $E : \{0,1\}^n \to \{0,1\}^m$ be an $(n-\Delta, 2^{-\Delta})$-extractor for $\mathcal{C}$-recognizable distributions, where each function in $\mathcal{C}$ is of the form $f_r(x) = C(x,r)$ where $r \in \{0,1\}^m$ is an arbitrary string. Then, $G(x) = (x, E(x))$ is ϵ-pseudorandom for $\mathcal{C}$.*

▶ **Lemma 36** ([16, Theorem 8]). *Let $f : \{0,1\}^n \to \{0,1\}$ be a function and let $E : \{0,1\}^n \to \{0,1\}^m$ be a function such that $G(x) = (x, E(x))$ is ϵ-pseudorandom for tests $T(x,r)$ of the form $T_z(x,r) = f(x) \wedge (r = z)$ where $z \in \{0,1\}^m$ is an arbitrary string. For any $\Delta > 0$, if $\epsilon \le 2^{-(m+2\Delta)}$ then E is an $(n-\Delta, 2^{-\Delta})$-extractor for the distribution recognized by f.*

The Lemma 3.2 requires a tiny $\epsilon \le 2^{-(m+2\Delta)}$ for the seed-extending PRG to get an $(n-\Delta, 2^{-\Delta})$-extractor. In the following, we improve the reduction from seed-extending PRGs to extractors to require only $\epsilon \le 2^{-2\Delta}$. Moreover, our extractor is even stronger – the output of our extractor is close to uniform with relative error, which will be defined as follows.

▶ **Definition 37** (Statistical distance with relative error). *We say that a distribution Z on $\{0,1\}^m$ is ϵ-close to uniform with relative error if for every event $A \subseteq \{0,1\}^m$,*

$$|\Pr[Z \in A] - \mu(A)| \le \epsilon \cdot \mu(A), \text{ where } \mu(A) = |A|/2^m.$$

Note that if Z is ϵ-close to uniform with relative error, then it is also ϵ-close to uniform. Next we define extractors with relative error analogously.

▶ **Definition 38** (Seedless extractor with relative error, [2, Definition 1.19]). *Let $\mathcal{C}$ be a class of distributions over $\{0,1\}^n$. A function $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ is a (k,ϵ)-relative-error extractor for $\mathcal{C}$ if for every distribution X in the class $\mathcal{C}$ such that $H_\infty(X) \ge k$, $\mathsf{Ext}(X)$ is ϵ-close to uniform with relative error.*

We remark that the notions of statistical distance and extractors with relative error were introduced by Applebaum, Artemenko, Shaltiel, and Yang [2]. They translate relative-error extractors for distributions recognizable by small circuits into incompressible functions. However, parameters of our relative-error extractors are not strong enough to get incompressible functions.

Now we prove the reduction lemma from seed-extending PRGs to seedless extractors with relative error, which directly implies the reduction from seed-extending PRGs to seedless extractors.

▶ **Lemma 39.** *Let $\mathcal{C}$ be a flip-invariant family of boolean functions over n bits. If G is a seed-extending (d,ϵ)-pseudorandom generator $G : \{0,1\}^d \to \{0,1\}^n$, then we can construct an $(n-\Delta, 2^\Delta\epsilon)$-relative-error extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-d}$ as follows. If $G(x) = (x, E(x))$ fools any function in $\mathcal{C}$, then $\mathsf{Ext}(x \circ y) = y \oplus E(x)$ is an extractor for $\mathcal{C}$-recognizable sources, where $x \in \{0,1\}^d, y \in \{0,1\}^{n-d}$.*

For intuition, observe that a seedless extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ partitions $\{0,1\}^n$ as $\bigcup_{z\in\{0,1\}^m} \mathsf{Ext}^{-1}(z)$. If Ext is a (k,ϵ)-relative-error extractor for $\mathcal{C}$-recognizable sources, then for every $f \in \mathcal{C}$ with $|f^{-1}(1)| \ge 2^k$, all intersections $\mathsf{Ext}^{-1}(z) \cap f^{-1}(1)$ should have almost the same size. That is, for most m-bit strings z, the preimage $\mathsf{Ext}^{-1}(z)$ is an ϵ-pseudorandom set against any $f \in \mathcal{C}$ with $|f^{-1}(1)| \ge 2^k$.

Now, given PRGs, how to construct extractors? From the above observation, converting an ϵ-pseudorandom set into a partition of ϵ-pseudorandom sets is a possible way. If each preimage $\mathsf{Ext}^{-1}(z)$ of Ext is an ϵ-pseudorandom set for $\mathcal{C}$, Ext should be a relative-error extractor for $\mathcal{C}$-recognizable sources with a bit worse parameters, which will be precisely calculated in the following formal proof.

To make $\mathsf{Ext}^{-1}(z)$ an ϵ-pseudorandom set for each z, we need a PRG of the specific form: $G(x) = B(x) \circ E(x)$, for some bijection $B : \{0,1\}^d \to \{0,1\}^d$ and some function $E : \{0,1\}^d \to \{0,1\}^{n-d}$. By linearly shifting the set $\{(B(x), E(x))\}$, we can partition $\{0,1\}^n$ as $\bigcup_{z\in\{0,1\}^{n-d}} \{(B(x), (E(x) \oplus z)) : x \in \{0,1\}^d\}$. Since $\mathcal{C}$ is a flip-invariant function family, we have that the set $\mathsf{Ext}^{-1}(z) = \{(B(x), (E(x) \oplus z)) : x \in \{0,1\}^d\}$ fools any function f in $\mathcal{C}$.

Note that to convert the PRG of the form $(B(x), E(x))$ into an extractor, the above intuition gives $\mathsf{Ext}(x) = E(B^{-1}(x))$. Thus, to get an efficient extractor, we have to assume that $E(B^{-1}(x))$ can be efficiently computed. That is, the PRG of the form $(B(x), E(x))$ also gives an efficient seed-extending PRG $(x, E(B^{-1}(x)))$. Therefore, for constructing extractors from the above intuition, we only need to focus on the seed-extending PRGs.

Proof. For convenience, let $m = n - d$ denote the output length of Ext.

First, we observe that, for any fixed z, $G_z(x) = (x, (E(x) \oplus z))$ fools any function $f(x, y)$ in $\mathcal{C}$. Notice that to prove $G_z(x)$ fools $f(x, y)$, it is equivalent to prove $(x, E(x))$ fools $f(x, y \oplus z)$. Because of the flip-invariant property of $\mathcal{C}$, we know if $f(x, y) \in C$, then $f(x, y \oplus z) \in \mathcal{C}$. So $G(x) = x \circ E(x)$ fools $f(x, y \oplus z)$. That is, $G_z(x)$ fools the function $f(x, y)$.

Note that $\mathsf{Ext}^{-1}(z)$ is the range of G_z. Then, we can get

$$
\begin{aligned}
&\Pr[\mathsf{Ext}(X \circ Y) = z | f(X \circ Y) = 1] \\
=&\frac{\Pr[\mathsf{Ext}(X \circ Y) = z \wedge f(X \circ Y) = 1]}{\Pr[f(X \circ Y) = 1]} \\
=&\frac{\Pr[\mathsf{Ext}(X \circ Y) = z]}{\Pr[f(X \circ Y) = 1]} \Pr[f(X \circ Y) = 1 | \mathsf{Ext}(X \circ Y) = z] \\
=&\frac{\Pr[\mathsf{Ext}(X \circ Y) = z]}{\Pr[f(X \circ Y) = 1]} \Pr[f(G_z(X)) = 1] \\
=&\frac{\Pr[\mathsf{Ext}(X \circ Y) = z]}{\Pr[f(X \circ Y) = 1]} (\Pr[f(X \circ Y) = 1] \pm \epsilon) \\
=&\frac{p \pm \epsilon}{p} \Pr[\mathsf{Ext}(X \circ Y) = z], \text{ where } p = \Pr[f(X \circ Y) = 1], \\
=&\frac{p \pm \epsilon}{p} \frac{1}{2^m}.
\end{aligned}
$$

For any nonempty subset $S \subseteq \{0,1\}^m$, summing over all $z \in S$, we deduce that the output of Ext is $\frac{\epsilon}{p}\mu(S)$-close to the uniform distribution over S. Furthermore, we have $\frac{\epsilon}{p} \le 2^{\Delta}\epsilon$, since $p = \frac{|f^{-1}(1)|}{2^n} \ge 2^{-\Delta}$ by the min-entropy requirement that $|f^{-1}(1)| \ge 2^{n-\Delta}$. Therefore, $\mathsf{Ext}(x \circ y) = y \oplus E(x)$ is an $(n - \Delta, 2^{\Delta}\epsilon)$-relative-error extractor for $\mathcal{C}$-recognizable sources. ◀

References

1 Noga Alon, Jehoshua Bruck, Joseph Naor, Moni Naor, and Ron M Roth. Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs. *IEEE Transactions on information theory*, 38(2):509–516, 1992.

2 Benny Applebaum, Sergei Artemenko, Ronen Shaltiel, and Guang Yang. Incompressible functions, relative-error extractors, and the power of nondeterministic reductions. *Computational complexity*, 25(2):349–418, 2016.

3 László Babai, Noam Nisant, and Márió Szegedy. Multiparty protocols, pseudorandom generators for logspace, and time-space trade-offs. *Journal of Computer and System Sciences*, 45(2):204–232, 1992.

4 Andrej Bogdanov, Akinori Kawachi, and Hidetoki Tanaka. Hard functions for low-degree polynomials over prime fields. *ACM Transactions on Computation Theory (TOCT)*, 5(2):5, 2013.

5 Jean Bourgain. On the construction of affine extractors. *GAFA Geometric And Functional Analysis*, 17(1):33–57, 2007.

6 Ashok K Chandra, Merrick L Furst, and Richard J Lipton. Multi-party protocols. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 94–99. ACM, 1983.

7 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.

8 Gil Cohen and Avishay Tal. Two Structural Results for Low Degree Polynomials and Applications. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, page 680, 2015.

9 Zeev Dvir. Extractors for varieties. *Computational complexity*, 21(4):515–572, 2012.

10 Zeev Dvir, Ariel Gabizon, and Avi Wigderson. Extractors and rank extractors for polynomial sources. *Computational Complexity*, 18(1):1–58, 2009.

11 O Goldreich. Three XOR-Lemmas – An exposition, 1995.

12 Alexander Golovnev and Alexander S Kulikov. Weighted gate elimination: Boolean dispersers for quadratic varieties imply improved circuit lower bounds. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 405–411. ACM, 2016.

13 Johan Håstad. *Computational limitations of small-depth circuits*. MIT Press, 1987.

14 Jørn Justesen. Class of constructive asymptotically good algebraic codes. *IEEE Transactions on Information Theory*, 18(5):652–656, 1972.

15 Jesse Kamp, Anup Rao, Salil Vadhan, and David Zuckerman. Deterministic extractors for small-space sources. *Journal of Computer and System Sciences*, 77(1):191–220, 2011.

16 Jeff Kinne, Dieter van Melkebeek, and Ronen Shaltiel. Pseudorandom generators, typically-correct derandomization, and circuit lower bounds. *Computational complexity*, 21(1):3–61, 2012.

17 Ilan Komargodski, Ran Raz, and Avishay Tal. Improved Average-Case Lower Bounds for De Morgan Formula Size: Matching Worst-Case Lower Bound. *SIAM Journal on Computing*, 46(1):37–57, 2017.

18 Xin Li. Improved two-source extractors, and affine extractors for polylogarithmic entropy. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 168–177. IEEE, 2016.

19 Noam Nisan. The communication complexity of threshold gates. *Combinatorics, Paul Erdos is Eighty*, 1:301–315, 1993.

20 Noam Nisan and Avi Wigderson. Hardness vs. randomness. In *Foundations of Computer Science, 1988., 29th Annual Symposium on*, pages 2–11. IEEE, 1988.

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

22 Ran Raz. Extractors with weak random seeds. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 11–20. ACM, 2005.

23 Zachary Remscrim. The Hilbert Function, Algebraic Extractors, and Recursive Fourier Sampling. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 197–208. IEEE, 2016.

24 Miklos Santha and Umesh V Vazirani. Generating quasi-random sequences from semi-random sources. *Journal of Computer and System Sciences*, 33(1):75–87, 1986.

25 Ronen Shaltiel. Weak derandomization of weak algorithms: explicit versions of Yao's lemma. *Computational complexity*, 20(1):87, 2011.

26 Roman Smolensky. Algebraic methods in the theory of lower bounds for Boolean circuit complexity. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 77–82. ACM, 1987.

27 Luca Trevisan and Salil Vadhan. Extracting randomness from samplable distributions. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 32–42. IEEE, 2000.

28 Emanuele Viola. Guest Column: correlation bounds for polynomials over {0 1}. *ACM SIGACT News*, 40(1):27–44, 2009.

29 Emanuele Viola. Extractors for circuit sources. *SIAM Journal on Computing*, 43(2):655–672, 2014.

30 Emanuele Viola and Avi Wigderson. Norms, XOR Lemmas, and Lower Bounds for Polynomials and Protocols. *Theory of Computing*, 4(1):137–168, 2008.

31 Andrew C Yao. Theory and application of trapdoor functions. In *Foundations of Computer Science, 1982. SFCS'08. 23rd Annual Symposium on*, pages 80–91. IEEE, 1982.

32 Andrew Chi-Chih Yao. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the eleventh annual ACM symposium on Theory of computing*, pages 209–213. ACM, 1979.

A Application of Theorem 2

A.1 Algebraic extractors over GF(2)

In this subsection, we will show that for any algebraic sources of constant degree over GF(2), there exists an efficient extractor that has linear output for linear min-entropy and exponentially small error. Formally, we will prove the following theorem:

▶ **Theorem 40.** *For any positive integer d, there is an efficient $\left((1-1/c_d)n, d, 2^{-\Omega(n/c_d)}\right)$-algebraic extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$, where $c_d = \Theta(d^2 4^d)$, $m = \Omega(n/c_d)$.*

Let P_d denote the set of all polynomials of degree at most d over GF(2). Let $\mathcal{V}_d$ denote the set of all products of polynomials in P_d and $\mathcal{V}_{d,n}$ denote the set of all products of n-variate polynomials in P_d.

Notice that an algebraic source of degree d over n bits is also a $\mathcal{V}_{d,n}$-recognizable source.

▶ **Lemma 41.** *An n-bit algebraic source of degree d iff it is a $\mathcal{V}_{d,n}$-recognizable source.*

Proof. Let U_V denote an arbitrary algebraic source, where $V = \{x \in \{0,1\}^n | p_i(x) = 0, p_i \in P_d, \forall i \in [k]\}$ is an algebraic set of degree d over n bits. Notice that V can be viewed as the set of 1-inputs of function $\prod_{i\in[k]}(p_i(x)+1)$. That is, the uniform distribution over V is also the source recognizable by $\prod_{i\in[k]}(p_i(x)+1) \in \mathcal{V}_{d,n}$. In other words, an algebraic source of degree d is a $\mathcal{V}_{d,n}$-recognizable source.

For the other direction, let U_f denote an arbitrary $\mathcal{V}_{d,n}$-recognizable source, where $f = \prod_{i\in[k]} p_i \in \mathcal{V}_{d,n}$ with $\deg(p_i) \le d$ for each $i \in [k]$. Note that $f^{-1}(1) = \{x \in \{0,1\}^n | p_i(x) = 1, \forall i \in [k]\} = \{x \in \{0,1\}^n | p_i(x) + 1 = 0, \forall i \in [k]\}$. Hence, $f^{-1}(1)$ is the algebraic set of $p_1(x)+1, \ldots, p_k(x)+1$. Since $\deg(p_i(x)+1) = \deg(p_i) \le d$ for each $i \in [k]$, $f^{-1}(1)$ is an algebraic set of degree d over n bits. Therefore, U_f is an n-bit algebraic source of degree d. ◀

Then, observe that $\mathcal{V}_d$ is closed under restrictions. Thus, by Theorem 31, to get an extractor for $\mathcal{V}_{d,n}$-recognizable sources, it is enough to show that $\mathcal{V}_d$ has α-XOR amplification for some positive constant α.

Note that to show that a function f has low correlations with $\mathcal{V}_d$, it suffices to show that f has low correlation with any polynomial of degree at most d. Recall that the correlation between a function f and a class $\mathcal{C}$ of functions is defined as the maximum of $Cor(f, C)$ over all $C \in \mathcal{C}$ whose input length is the same as f. In particular, to show that a function $f : \{0,1\}^t \to \{0,1\}$ has low correlations with $\mathcal{V}_d$, it suffices to show that f has low correlation with any t-variate polynomial of degree at most d.

▶ **Lemma 42.** *If a function $f : \{0,1\}^t \to \{0,1\}$ is ϵ-correlated with any polynomial of degree at most d in t variables, then f is at most 2ϵ-correlated with any product of polynomials of degree at most d in t variables.*

The lemma follows because the L1 norm of the Fourier transform of the AND function is at most 2.

Proof. We need to show that if for any t-variate $p \in P_d$ $Cor(f,p) = |Ee[f+p]| \leq \epsilon$, then for any product $\prod_{i\in[k]}(p_i+1) \in \mathcal{V}_{d,t}$ where $p_1+1, \ldots, p_k+1 \in P_{d,t}$, we have

$$Cor\left(f, \prod_{i\in[k]}(p_i(X)+1)\right) = \left|Ee\left[f + \prod_{i\in[k]}(p_i(X)+1)\right]\right| \leq 2\epsilon.$$

Consider the Fourier expansion of the function

$$e\left[\prod_{i\in[k]}(y_i+1)\right] = -\sum_{S\neq\emptyset}\frac{e\left[\sum_{i\in S} y_i\right]}{2^{k-1}} + (1 - 1/2^{k-1}).$$

Now, substituting each y_i by p_i, we have $e\left[\prod_{i\in[k]}(p_i+1)\right] = -\sum_{S\neq\emptyset}\frac{e\left[\sum_{i\in S} p_i\right]}{2^{k-1}} + (1 - 1/2^{k-1})$.

That is,

$$\left|Ee\left[f + \prod_{i\in[k]}(p_i(X)+1)\right]\right| \leq \sum_{S\neq\emptyset}\frac{\left|Ee\left[f + \sum_{i\in S} p_i(X)\right]\right|}{2^{k-1}}.$$

Notice that for each $S \neq \emptyset$, the sum $\sum_{j\in S} p_i$ is also a polynomial of degree at most d. For the polynomial of degree at most d, $\sum_{j\in S} p_i$, we have that $\left|Ee\left[f + \sum_{j\in S} p_i(X)\right]\right| \leq \epsilon$. In other words, $\left|Ee\left[f + \prod_{i\in[k]}(p_i(X)+1)\right]\right| \leq 2^k \frac{2\epsilon}{2^{k-1}} = 4\epsilon$. ◀

Moreover, Viola and Wigderson [30] proved XOR amplification for GF(2) polynomials, which implies XOR amplification for $\mathcal{V}_d$ by Lemma 42.

▶ **Theorem 43** ([30, Theorem 1.1])**.** *Let $h : \{0,1\}^n \to \{0,1\}$ be a function such that $Cor(h, P_{d,n}) \leq 1 - 1/2^d$. Then $Cor(h^{\oplus m}, P_d) \leq 2^{-\Omega(m/(4^d\cdot d))}$.*

Finally, by brute force search, it is easy to find a function h over $O(d)$ bits such that $Cor(h, P_d) \leq 1 - 1/2^d$ as d is a constant. That is, P_d has $\Omega(\frac{1}{4^d\cdot d})$-XOR amplification for the function $h : \{0,1\}^{O(d)} \to \{0,1\}$. This implies that V_d has $\Omega(\frac{1}{4^d\cdot d})$ -XOR amplification for the function $h : \{0,1\}^{O(d)} \to \{0,1\}$ by Lemma 42. Therefore, Theorem 34 yields our main theorem of this subsection, i.e., constructing an efficient $\left((1-1/c_d)n, d, 2^{-\Omega(n/c_d)}\right)$-algebraic extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$, where $c_d = \Theta(d^2 4^d)$, $m = \Omega(n/c_d)$.

We remark that an explicit example of h is the mod_3 function, which outputs 1 if and only if the number of input bits that are '1' is congruent to 1 modulo 3. Smolensky [26] proved that the mod_3 function over $O(d^2)$ bits is 2/3-hard for P_d (see Viola [28] for a proof), that is, P_d has $\Omega(\frac{1}{4^d \cdot d})$-XOR amplification for the function $\mathrm{mod}_3 : \{0,1\}^{O(d^2)} \to \{0,1\}$. Using the mod_3 function, Theorem 34 yields an efficient $\left((1-1/c'_d)n, d, 2^{-\Omega(n/c'_d)}\right)$-algebraic extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$, where $c'_d = \Theta(d^3 4^d)$, $m = \Omega(n/c'_d)$.

A.2 Sources recognizable by communication protocols

In this subsection, we construct an extractor for sources recognizable by randomized t-party protocols. Formally, we prove the following theorem.

▶ **Theorem 44.** *There exists an explicit seedless $((1-1/c_t)n,\ 2^{-c_1 n/c_t})$ extractor $\mathsf{Ext} : \left(\{0,1\}^{n/t}\right)^t \to \{0,1\}^{c_2 n/c_t}$ for sources recognizable by randomized t-party communication protocols of at most $c_3 n/c_t$ bits, where $c_t = \Theta(t4^t)$ and c_1, c_2, c_3 are some positive constants.*

Let $\mathcal{RCC}_{n,t,w}$ denote the class of n-variate randomized t-party protocols using at most w communication bits. Now, to construct extractors for $\mathcal{RCC}_{n,t,w}$-recognizable sources with exponentially small error, by Theorem 31, it suffices to show $\mathcal{RCC}_{n,t,w}$ has $(\alpha, r)-$XOR amplification for some function h, where $r = \Omega(n)$ is the distance of some good linear code.

Notice that, Babai, Nisan, and Szegedy [3] proved a lower bound for randomized t-party protocols for the Generalized Inner Product (GIP) function, which is the XOR of AND functions. Formally, let $\wedge_t : \{0,1\}^t \to \{0,1\}$ denote the AND function on t variables. Then, the GIP function $\mathrm{GIP}_{kt} : (\{0,1\}^t)^k \to \{0,1\}$ is defined as the function$\wedge_t^{\oplus k}$, i.e., $GIP_{kt}(x_1, \ldots, x_k) := \bigoplus_{i=1}^k \wedge_t(x_i)$. Moreover, let $R_{t,\epsilon}(f)$ denote the complexity of the best randomized t-party protocol correlating f with at least ϵ.

▶ **Theorem 45** ([3, Theorem 2])**.**

$$R_{t,\epsilon}(GIP_n) = \Omega\left(\frac{n}{4^t} - \log(1/\epsilon)\right).$$

Now, for any constant $0 < \delta < 1/t$ and some constant $c_t = \Theta(t4^t)$, we prove that $\mathcal{RCC}_{n,t,O(n/4^t)}$ has $(\Omega(1/c_t), \delta n)-$XOR amplification for $\wedge_t$, which directly yields Theorem 44 by Theorem 34.

▶ **Proposition 46.** *For any constant $0 < \delta < 1/t$, $\mathcal{RCC}_{n,t,c'n/4^t}$ has $(c/c_t, \delta n)-$XOR amplification for $\wedge_t$, where $c_t = \Theta(t4^t), c, c' > 0$ are constants.*

Proof. Assume by contradiction that $\mathcal{RCC}_{n,t,c'n/4^t}$ does not have $(c/c_t, \delta n)-$XOR amplification for $\wedge_t$, where c, c' are some constants to be decided later. That is, there exists some vector $v \in \{0,1\}^{n/t}$ with at least δn ones, $Cor(h^v, \mathcal{RCC}_{n,t,c'n/4^t}) \le 2^{-\frac{c}{c_t}\delta n}$, That is, there exists a $(c'n/4^t)$-bit randomized protocol that approximates h^v within $2^{-\frac{c}{c_t}\delta n}$ error. Furthermore, observe that h^v is the XOR of at least δn copies of $\wedge_t$, i.e, h^v depends on $\ge \delta nt$ variables. Therefore, by Theorem 45, we have

$$R_{t,2^{-\frac{c}{c_t}\delta n}}(h^v) \ge R_{t,2^{-\frac{c}{c_t}\delta n}}(GIP_{\delta n}) = \Omega\left(\delta\frac{n}{4^t} - \frac{c}{c_t}\delta nt\right).$$

That is, letting the constant c be small enough, we know there exists a positive constant c'' such that

$$R_{t,2^{-\alpha\delta n}}(h^v) \ge c''n/4^t.$$

Now letting $c' < c''$ yields a contraction. Therefore, $\mathcal{RCC}_{n,t,c'n/4^t}$ has $(c/c_t, \delta n)-$XOR amplification for $\wedge_t$. ◀

A.3 Halfspace sources

In this subsection, for halfspace sources, we construct an efficient extractor that has linear output for linear min-entropy and exponentially small error. Formally, we will prove the following theorem.

▶ **Theorem 47.** *There exists an explicit seedless* $((1-c_1)n,\ 2^{-c_2 n})$ *extractor* $\mathsf{Ext}: \{0,1\}^n \to \{0,1\}^{c_3 n}$ *for halfspace sources, where* c_1, c_2, c_3 *are some positive small enough constants.*

Note that Nisan already proved an exponentially small correlation bound for Inner Product function against LTFs. Formally, let $IP_n : (\{0,1\}^2)^{n/2} \to \{0,1\}$ denote the inner product function over n variables, i.e., $IP_n(x_1, \ldots, x_{n/2}) = \bigoplus_{i\in[n/2]} \wedge_2(x_i)$. Then, we have the following lemma.

▶ **Lemma 48.** *For any LTF* f *on* n *variables, we have*

$$Cor(IP_n, f) \le 2^{-\Omega(n)}.$$

Proof of sketch. Nisan proved that a LTF on n variables can be approximated within ϵ error by a randomized 2-party protocol of complexity $O(\log(n/\epsilon))$ by [19, Theorem 1]. Moreover, by Chor and Goldreich [7], we know at least $n/2 - \log(1/\epsilon)$ complexity needed for randomized 2-party protocol computing the function IP_n.

Therefore, for any LTF f over n variables, there is a protocol $\mathcal{P}$ of complexity cn bits approximating f within $2^{-\Omega(n)}$ error and $Cor(IP_n, \mathcal{P}) \le 2^{-\Omega(n)}$. That is, replacing f by IP_n in $Cor(IP_n, f)$, we can bound $Cor(IP_n, f) \le 2^{-\Omega(n)} + Cor(IP_n, \mathcal{P}) = 2^{-\Omega(n)}$. ◀

Let $\mathcal{LTF}_n$ denote the class of LTFs over n variables. Then, the above lemma directly yields that $\mathcal{LTF}_n$ has $(\alpha, \delta n)$-XOR amplification for $\wedge_2$ for any positive constant $\delta < 1/2$, where α is some positive constant. Hence Theorem 47 directly follows by Theorem 34.

B Application of Theorem 8

In this section, we construct extractors for sources recognized by several widely used function families. These constructions are all based on Lemma 39 proved in the previous section, which means we can convert seed-extending PRGs into extractors. In the following subsections, the main points are to construct seed-extending PRGs for some specific common function families.

B.1 Circuit-recognizable sources

Recall that we say a function $h : \{0,1\}^t \to \{0,1\}$ is ϵ-hard for $\mathcal{C}$ if $Cor(h, \mathcal{C}) \le \epsilon$.

For any circuit family, Nisan and Wigderson [20] already constructed a hardness-based PRG. Reviewing the NW generator, Kinne et al. [16] proved that it could be made seed-extending, and hence they gave a seed-extending PRG for circuits. In particular, they proved the following lemma.

▶ **Lemma 49** ([16, Lemma 2.9])**.** *Let* l *and* m *be positive integers and* $H : \{0,1\}^{\sqrt{l}/2} \to \{0,1\}$ *a function. If* H *is* $\frac{\epsilon}{m}$*-hard at input length* $\sqrt{l}/2$ *for circuits of size* $s + m \cdot 2^{O(\log m/\log l)}$ *and depth* $d+1$*, then there is a seed-extending* (l, ϵ)*-PRG* $NW_{H;l,m} : \{0,1\}^l \to \{0,1\}^{l+m}$ *for tests* $T : \{0,1\}^{l+m} \to \{0,1\}$ *computable by circuits of size* s *and depth* d*.*

Notice that the set of bounded-size circuits is flip-invariant since flipping the inputs of a circuit does not change its size. Thus, applying Lemma 39, we get an extractor.

▶ **Proposition 50.** *For any positive integer $l < n$, if there is a function H that is ϵ-hard at input length $\sqrt{l}/2$ for circuits of size $s + (n-l) \cdot 2^{O(\log(n-l)/\log l)}$ and depth $d+1$, then for any $\Delta = \Delta(n) > 0$ we can get an $(n - \Delta, (n-l)2^{\Delta}\epsilon)$-extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-l}$ for any sources recognizable by circuits of size s and depth d.*

We remark that, in the best case, the above lemma yields an $(n - \tilde{O}(\sqrt{l}), 2^{-\tilde{\Omega}(\sqrt{l})})$-extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-l}$, if we can get a function at input length $\sqrt{l}/2$ which is $2^{-\tilde{\Omega}(\sqrt{l})}$-hard for circuits of polynomial size.

B.2 AC^0-ecognizable sources

Hastad [13] proved that the parity function is $2^{-n^{1/(d+1)}}$-hard against any AC^0 circuit of size $2^{n^{1/(d+1)}}$ and depth d. Based on this hardness, Shaltiel [25] constructed extractors for AC^0-recognizable sources.

▶ **Theorem 51** (Corollary 4.25, [25])**.** *For any $\Delta = \Delta(n) > 0$, there is a constant $\alpha > 0$ such that for every sufficiently large n, $m \leq n^{1/(\alpha d)}$, and sources recognizable by circuits of size $2^{n^{1/(\alpha d)}}$ and depth d, we can construct an $(n - n^{1/(\alpha d)}, 2^{-100m})$-extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$.*

▶ **Theorem 52** (Theorem 4.21, [25])**.** *For any constants $c, d, e > 1$ there is a constant $d' > 1$ and a uniform family $E = \{E_n\}$ of circuits of polynomial-size and depth d' such that $E_n : \{0,1\}^n \to \{0,1\}^m$ for $m(n) = (\log n)^e$ and E_n is a $(n - 100m(n), 2^{-100m(n)})$-extractor for sources recognizable by circuits of size n^c and depth d.*

However, directly using the Lemma 50 with the hardness of parity function, we can get the following lemma.

▶ **Theorem 53.** *For any $\Delta = \Delta(n) > 0$, there exists a polynomial time computable $(n - \Delta, (n-l)2^{\Delta - \Omega(l^{1/(2d+2)})})$ extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^{n-l}$ for any sources recognizable by circuits of size $2^{n^{1/d}}$ and depth d.*

▶ **Proposition 54.** *For any constants $c, d, e > 1$ there is a constant $e' < e$ and a polynomial-time computable uniform family $E = \{E_n\}$ such that $E_n : \{0,1\}^n \to \{0,1\}^m$ for $m(n) = n - (\log n)^e$ and E_n is a $(n - 100(\log n)^{e'}, 2^{-100(\log n)^{e'}})$-extractor for sources recognizable by circuits of size n^c and depth d.*

In particular, for min-entropy $n - n^{1/(\alpha d)}$, our extractor outputs $n - n^{2/\alpha + O(1/d)}$ bits, whereas Shaltiel's extractor outputs only $n^{1/(\alpha d)}$ bits. When $\alpha > 2d/(d-1)$ is a large enough constant, our extractor outputs $n - o(n)$ bits whereas Shaltiel's extractor outputs only $n^{1/(\alpha d)}$ bits. For min-entropy $n - \mathrm{polylog}(n)$ bits, our extractor outputs $n - \mathrm{polylog}(n)$, whereas Shaltiel's extractor outputs only $\mathrm{polylog}(n)$ bits

For circuit sources, Viola [29] also constructed extractors for AC^0-samplable sources, extracting $k(k/n^{1+\gamma})^{O(1)}$ bits with super-polynomially small error from n-bit sources of min-entropy k, for any $\gamma > 0$. Nevertheless, AC^0-samplable sources are different from AC^0-recognizable sources.