Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 114 (2016)

Towards Optimal Two-Source Extractors and Ramsey Graphs

Gil Cohen[*]

November 1, 2016

Abstract

The main contribution of this work is a construction of a two-source extractor for quasi-logarithmic min-entropy. That is, an extractor for two independent n-bit sources with min-entropy $\widetilde{O}(\log n)$, which is optimal up to the $\mathrm{poly}(\log\log n)$ factor. A strong motivation for constructing two-source extractors for low entropy is for Ramsey graphs constructions. Our two-source extractor readily yields a $(\log n)^{(\log\log\log n)^{O(1)}}$-Ramsey graph on n vertices.

Although there has been exciting progress towards constructing $O(\log n)$-Ramsey graphs in recent years, a line of work that this paper contributes to, it is not clear if current techniques can be pushed so as to match this bound. Interestingly, however, as an artifact of current techniques, one obtains strongly explicit Ramsey graphs, namely, graphs on n vertices where the existence of an edge connecting any pair of vertices can be determined in time $\mathrm{polylog}\, n$. On top of our strongly explicit construction, in this work, we consider algorithms that output the entire graph in $\mathrm{poly}(n)$-time, and make progress towards matching the desired $O(\log n)$ bound in this setting. In our opinion, this is a natural setting in which Ramsey graphs constructions should be studied.

The main technical novelty of this work lies in an improved construction of an independence-preserving merger (IPM), a variant of the well-studied notion of a merger, that was recently introduced by Cohen and Schulman (FOCS'16). Our construction is based on a new connection to correlation breakers with advice. In fact, our IPM satisfies a stronger and more natural property than that required by the original definition, and we believe it may find further applications.

[*] Department of Computer Science, Princeton University, Princeton, USA. Email: gilc@cs.princeton.edu. Most of this work was completed while the author was a postdoctoral fellow at the Department of Computing and Mathematical Sciences, Caltech, Pasadena, USA.

ISSN 1433-8092

Contents

1 Introduction

Ramsey theory is a branch of combinatorics that investigates the unavoidable presence of local structure in globally unstructured objects. In the paper that marks the birth of Ramsey theory [Ram28], Ramsey considered this general phenomena in the graph-theoretic setting.

Definition 1.1 (Ramsey graphs [Ram28])**.** *An undirected graph is k-*Ramsey *if it contains neither a clique nor an independent set of size k.*

Ramsey [Ram28], and Erdaős, Szekeres [ES35] proved that there does not exist a graph on n vertices that is $0.5\log_2 n$-Ramsey. In his influential paper that inaugurated the probabilistic method, Erdős [Erd47] complemented Ramsey's result by proving the existence of a k-Ramsey graph on n vertices, where $k = 2\log_2 n + O(1)$. Unfortunately, Erdős' argument is non-constructive and one does not obtain from Erdős' proof an explicit example of a graph that is k-Ramsey–a challenge that was posed by Erdős and had attracted a significant attention in the literature since then (see Table 1.)

There are several ways to formalize what it means for a graph to be "explicit". First, one must consider not a single graph but rather a family of graphs $\mathcal{G} = \{G_n\}_{n\in N}$, where $N \subseteq \mathbb{N}$, and $G_n = (V_n, E_n)$ is an undirected graph with $|V_n| = n$. The family $\mathcal{G}$ is *strongly explicit* if there is an algorithm $\mathcal{A}_{\mathsf{strong}}$ that, on inputs $u, v \in V_n$, runs in time $\mathrm{polylog}(n)$ and decide whether $\{u, v\} \in E_n$. A natural relaxation would ask for an algorithm $\mathcal{A}_{\mathsf{semi}}$ that on input $n \in N$ runs in time $\mathrm{poly}(n)$ and outputs the adjacency matrix representation of G_n. In such case, the family $\mathcal{G}$ is called *semi-explicit*. We typically abuse notation and say that a graph $G = (V, E)$ is strongly explicit or semi-explicit while formally refer to a family of graphs, of which G is a representative.

Interestingly, as a byproduct of existing techniques, almost all of the known Ramsey graph constructions are strongly explicit. However, semi-explicit constructions are as natural as strongly explicit constructions in the setting of Ramsey graphs. In this paper, among other results, we make progress on both strongly explicit and semi-explicit constructions of Ramsey graphs. Our first result is an improved strongly explicit construction of Ramsey graphs.

Theorem 1.2. *There exists a universal constant $c \geq 1$ for which the following holds. For any integer n, there exists a strongly explicit $O((\log n)^{(\log\log\log n)^c})$-Ramsey graph on n vertices.*

Theorem 1.2 improves upon an exciting line of work [BKS+10, BRSW12, Coh16d, CZ16] that has been accumulated to a construction of $(\log n)^{2^{O(\sqrt{\log\log\log n})}}$-Ramsey graphs by Ben-Aroya, Doron, and Ta-Shma [BADTS16]. While Theorem 1.2 is fairly close to optimal, it is not clear if current techniques for constructing strongly explicit Ramsey graphs can be pushed further so as to match the desired $2\log_2 n + O(1)$ bound, or even to obtain $O(\log n)$-Ramsey graphs. The best construction of a $2\log_2 n + O(1)$-Ramsey graph on n vertices runs in quasi-polynomial time $n^{O(\log n)}$ [BKS+10]. This naturally leads us to consider semi-explicit constructions. To present our results in this direction, we proceed with a discussion on multi-source extractors. This will also allow us to present our two-source extractor construction (Theorem 1.4.)

Construction	$k(n)$	Bipartite
[Erd47] (non-constructive)	$2\log_2 n + O(1)$	✓
[Abb72]	$n^{\log_5 2}$	
[Nag75]	$n^{1/3}$	
[Fra77]	$n^{o(1)}$	
[Chu81]	$2^{O((\log n)^{3/4}\cdot(\log\log n)^{1/4})}$	
[FW81, Nao92, Alo98, Gro01, Bar06]	$2^{O(\sqrt{\log n \cdot \log\log n})}$	
The Hadamard matrix (folklore)	$\sqrt{n}$	✓
[PR04]	$n^{1/2-o(1)}$	✓
[BKS+10]	$n^{O(1/\log\log n)}$	✓
[BRSW12]	$2^{2^{(\log\log n)^{1-\alpha}}}$	✓
[Coh16d, CZ16]	$2^{(\log\log n)^{O(1)}}$	✓
[BADTS16]	$(\log n)^{2^{O(\sqrt{\log\log\log n})}}$	✓
Theorem 1.2	$(\log n)^{(\log\log\log n)^{O(1)}}$	✓

Table 1: Summary of Ramsey graphs constructions from the literature. Bipartite Ramsey graphs are the analog of Ramsey graphs for the bipartite setting (for the formal definition see, e.g., [BRSW12]). One can show that any bipartite Ramsey graph induces a Ramsey graph with comparable parameters.

1.1 Two-Source Extractors for Quasi-Logarithmic Entropy

Recall that a random variable X is said to have *min-entropy* k if $\forall x\ \mathbf{Pr}[X = x] \le 2^{-k}$. In such case, X is called a *k-source* or, more informatively, an *(n,k)-source* if X is supported on n-bit strings.

Definition 1.3 (Two-source extractors [CG88])**.** *A function* $\mathsf{Ext}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ *is called a* (k,ε) two-source extractor *if for any pair of independent* (n,k)*-sources* X, Y*,* $\mathsf{Ext}(X,Y) \approx_\varepsilon U$*.*

In this paper, we only consider extractors with a single output bit and with constant error guarantee ε. This regime of parameters is most interesting for Ramsey graphs constructions. It is easy to prove the existence of a (k,ε) two-source extractor for n-bit sources with $k = \log_2 n + O(1)$, and that this value is optimal up to the additive constant term. One can show that for any $\varepsilon < 1$, a (k,ε) two-source extractor $\mathsf{Ext}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ yields a 2^{k+1}-Ramsey graph on 2^n vertices. Moreover, if Ext can be evaluated in $\mathrm{poly}(n)$-time then the induced Ramsey graph is strongly explicit.

The main result of this work is an explicit two-source extractor for quasi-logarithmic min-entropy. By the connection to Ramsey graphs mentioned above, Theorem 1.4 readily implies Theorem 1.2.

Theorem 1.4. *For any integer n and constant $\varepsilon > 0$ there exists a $\mathrm{poly}(n)$-time computable (k, ε) two-source extractor with $k = \widetilde{O}(\log n)$.* [1]

Construction	Supported entropy	Comments
[CG88] (non-constructive)	$\log_2 n + O(1)$	
[CG88] (conditional)	$o(n)$	
[CG88]	$(1/2 + \delta)n$	for any constant $\delta > 0$
[Raz05]	$(1/2 + \delta)n, \ \ O(\log n)$	for any constant $\delta > 0$
[Bou05]	$(1/2 - \beta)n$	for some universal constant $\beta > 0$
[BSZ11] (conditional)	$(2/5 + \delta)n$	for any constant $\delta > 0$
[CZ16]	$(\log n)^c$	for some universal constant c
[BADTS16]	$\log n \cdot 2^{O(\sqrt{\log \log n})}$	
Theorem 1.4	$\widetilde{O}(\log n)$	

Table 2: Explicit constructions of two-source extractors from the literature.

1.2 Semi-Explicit Four-Source Extractors

In their influential paper, Barak, Impagliazzo, and Wigderson [BIW06] considered a generalization of two-source extractors.

Definition 1.5 (Multi-source extractors [BIW06])**.** *Let $s \geq 2$ be an integer. A function $\mathsf{Ext}\colon (\{0,1\}^n)^s \to \{0,1\}$ is a (k, ε)* s-source extractor *if for any independent (n, k)-sources $X_1, \ldots, X_s$, $\mathsf{Ext}(X_1, \ldots, X_s) \approx_\varepsilon U$.*

One can extend the argument for the existence of two-source extractors that is mentioned above so as to prove the following.

[1] We use the standard notation $\widetilde{O}(m)$ for $m \cdot (\log m)^{O(1)}$.

Fact 1.6. *For any constant integer $s \geq 2$ and constant $\varepsilon > 0$ there exists a (k, ε) s-source extractor* $\mathsf{Ext}\colon (\{0,1\}^n)^s \to \{0,1\}$, *where*

$$k = \frac{1}{s-1} \cdot \log_2 n + O(1).$$

This is tight up to the additive constant term.

As mentioned, although Theorem 1.4 comes fairly close to optimal and, as a result, yields close to optimal strongly explicit Ramsey graphs, it seems unlikely to us that current techniques will yield strongly explicit Ramsey graphs matching the existential $2\log_2 n + O(1)$ bound. For that one would need a two-source extractor for min-entropy $1 \cdot \log_2 n + O(1)$. In particular, one must insists on the tight constant factor that multiplies $\log_2 n$. This seems out of reach. Obtaining semi-explicit Ramsey graphs raises the problem of constructing extractors that are allowed to run in exponential-time in their input length. By allowing this exponential slowdown, which is completely natural in the setting of Ramsey graphs, we ask for optimal, or near optimal, constructions. In particular, we insists on obtaining the tight constant that multiplies $\log_2 n$, which is 1 in the case of two-source extractors and, more generally, is $\frac{1}{s-1}$ for s-source extractors.

In this work we make a step towards constructing optimal semi-explicit Ramsey graph by devising exponential-time multi-source extractors with the tight constant factor that multiplies $\log_2 n$. More precisely, we prove the following.

Theorem 1.7. *For any constant integer $s \geq 3$, constant $\varepsilon > 0$, and integer n, there exists an* $\exp\left(n + n^{\frac{2}{s-1}} \cdot \log^{10} n\right)$*-time computable (k, ε) s-source extractor, where*

$$k = \frac{1}{s-1} \cdot \log_2 n + 5\log_2 \log n + O(1).$$

Note that for any $s \geq 4$, Theorem 1.7 gives an exponential-time s-source extractor with the tight constant factor that multiplies $\log_2 n$. The extractor is optimal up to the *additive* $5\log_2 \log n + O(1)$ term. Moreover, Theorem 1.7 yields a 3-source extractor with similar parameters, however, its running-time is $\exp(n \cdot \log^{10} n)$ as opposed to the desired $\exp(n)$ running-time. Unfortunately, our techniques breaks at $s = 2$, and we do not obtain an improvement over the Barak *et al.* result [BKS+10], which is optimal up to an additive constant factor, and runs in time $\exp(n^2)$. The proof of Theorem 1.7 can be found in Section 7.

1.3 Non-Malleable Extractors

For the proof of Theorem 1.4, we construct an improved non-malleable extractor [DW09]. The original motivation for studying non-malleable extractors was for the problem of privacy amplification. A framework for constructing privacy amplification protocols was devised [DW09] that is instantiated with a non-malleable extractor, and where the parameters of the protocol inherits those of the extractor. In particular, via the Dodis-Wichs framework, an optimal non-malleable extractor readily induces an optimal privacy amplification protocol. For a discussion on the Dodis-Wichs framework, the reader may consult the original paper [DW09] or Section 2.3 of [Coh16a] for a brief and informal treatment. In [DW09] it was shown that non-malleable extractors exist, though the task of constructing such extractors was left for future research, and has gained a significant attention as summarized in Table 3.

Definition 1.8 (Non-malleable extractors [DW09])**.** *A function* $\mathsf{nmExt}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is called a* (k,ε)-non-malleable extractor *if for any* (n,k)*-source* X *and function* $\mathcal{A}\colon \{0,1\}^d \to \{0,1\}^d$ *with no fixed points,*

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

where Y *is uniformly distributed over* $\{0,1\}^d$ *independently of* X*. If* nmExt *is a* (k,ε)*-non-malleable extractor, we say that* nmExt *has error guarantee* ε *and that* nmExt *supports min-entropy* k*.*

It can be shown that, regardless of the computational aspect, any (k,ε)-non-malleable extractor for n-bit sources requires seed length $d = \Omega(\log(n/\varepsilon))$, can only support min-entropy $k = \Omega(\log(1/\varepsilon))$, and can output at most $k/2 - \Omega(\log(1/\varepsilon))$ bits. Prior to this work, the state of the art explicit non-malleable extractor [Coh16a] has seed length $d = O(\log n) + \log(1/\varepsilon) \cdot 2^{O(\sqrt{\log\log(1/\varepsilon)})}$, supports min-entropy $k = \Omega(d)$, and can output $m = (1/2-\alpha)k$ bits for any desired constant $\alpha > 0$. In this work we improve upon this result and obtain the following.

Theorem 1.9. *For any constant* $\alpha > 0$ *there exists a constant* $c \geq 1$ *such that for any integer* n *and any* $\varepsilon > 0$*, there exists an explicit* (k,ε)*-non-malleable extractor* $\mathsf{nmExt}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with seed length* $d = O(\log n) + \widetilde{O}(\log(1/\varepsilon))$ *for any* $k \geq cd$*, with* $m = (1/2-\alpha)k$ *output bits.*

By plugging our non-malleable extractor from Theorem 1.9 to the Dodis-Wichs framework, we obtain improved, near-optimal, privacy amplification protocols.

Corollary 1.10. *For all* n, λ*, there exists an explicit two-round privacy amplification protocol against an active adversary, that supports min-entropy* $k = \Omega(d)$*, with entropy-loss* $O(\lambda + \log n)$*, and communication complexity* $O(d + (\lambda + \log k) \cdot \log k)$*, where* $d = O(\log n) + \widetilde{O}(\lambda)$*.*

In the next two subsections we briefly review what is known about the connection between non-malleable extractors and two-source extractors–a connection we employ so as to deduce Theorem 1.4 from Theorem 1.9.

1.3.1 The [CZ16] Reduction From Two-Source Extractors to Non-Malleable Extractors

In a recent breakthrough [CZ16], Chattopadhyay and Zuckerman showed how to reduce the problem of constructing two-source extractors to that of constructing non-malleable

Construction	Seed length d (up to constants)	Supported entropy
[DW09] (non-constructive)	$\log(n) + O(1)$	$\Omega(\log\log n)$
[LWZ11]	n	$(0.5+\delta)\cdot n$
[CRS14, DLWZ14, Li12a]	$\log(n/\varepsilon)$	$(0.5+\delta)\cdot n$
[Li12b]	$\log(n/\varepsilon)$	$(0.5-\beta)\cdot n$
[CGL16]	$\log^2(n/\varepsilon)$	$\Omega(d)$
[Coh16b]	$\log(n/\varepsilon)\cdot\log(\log(n)/\varepsilon)$	$\Omega(d)$
[Coh16c]	$\log n + \log^3(1/\varepsilon)$	$\Omega(d)$
[CL16]	$\log(n/\varepsilon)\cdot 2^{\sqrt{\log\log(n/\varepsilon)}}$	$\Omega(d)$
[Coh16a]	$\log n + \log(1/\varepsilon)\cdot 2^{\sqrt{\log\log(1/\varepsilon)}}$	$\Omega(d)$
Theorem 1.9	$\log n + \widetilde{O}(\log(1/\varepsilon))$	$\Omega(d)$

Table 3: Explicit constructions of non-malleable extractors from the literature. We note that [Coh16b, CL16] offer several more constructions.

extractors. More precisely, it was shown how to construct a two-source extractor given a non-malleable extractor as well as an extractor for non-oblivious bit-fixing sources. The min-entropy supported by the two-source extractor is related to the seed length d and the supported min-entropy k of the non-malleable extractor when set with error guarantee $\varepsilon = 2^{-\log^c n}$ for some large enough constant $c > 1$. By plugging the state of the art non-malleable extractor that was available at the time [CGL16] to their reduction, the first n-bit two-source extractor for polylog(n) min-entropy sources was obtained.

Although exciting, the [CZ16] reduction from non-malleable extractors to two-source extractors cannot be used to obtain two-source extractors for min-entropy $O(\log n)$. In fact, as was observed by [CS16], ideas that were used at the time were stuck at min-entropy $\Omega(\log^2 n)$ for several different reasons, even if one has access to any $o(\log n)$ number of sources (as opposed to just 2 sources) and even if one would settle for a disperser. In a sequence of works [CS16, CL16] that has accumulated to [Coh16a], an extractor for 5 n-bit sources with min-entropy $\log n \cdot 2^{O(\sqrt{\log\log n})}$ was constructed. It was not clear, however, how to reduce the number of sources from 5 to 2.

1.3.2 The Improved [BADTS16] Reduction

Very recently, Ben-Aroya, Doron, and Ta-Shma [BADTS16] devised an improved reduction from two-source extractors to non-malleable extractors. The new reduction has two advantages over the original reduction of [CZ16]. First, as in [CS16, CL16, Coh16a], the fairly complicated extractor for non-oblivious bit-fixing sources was replaced with the simple majority function, significantly simplifying the overall construction. Second, the reduction applies the non-malleable extractor with error guarantee $\varepsilon = \mathrm{poly}(1/n)$. Thus, the [BADTS16] reduction paves the way for constructing two-source extractors for min-entropy $O(\log n)$.

For their reduction, Ben-Aroya *et al.* apply some of the new techniques that were developed in [CS16, Coh16a, CL16], as well as a variation on a classical error reduction technique for seeded extractors [RRV99] and a result by Dodis *et al.* [DPW14]. By plugging the explicit non-malleable extractor of [Coh16a] to their reduction, Ben-Aroya *et al.* obtained a two-source extractor for n-bit sources with min-entropy $\log n \cdot 2^{O(\sqrt{\log\log n})}$. By plugging our non-malleable extractor from Theorem 1.9 instead, we readily obtain Theorem 1.4.

2 Independence-Preserving Mergers and Correlation Breakers

Now that the notion of a non-malleable extractor and its applications were briefly discussed, we turn to consider the inner workings of our non-malleable extractor. To this end we discuss two recently introduced pseudorandom objects: independence-preserving mergers (IPM) [CS16], and correlation breakers (CBA) [Coh15, CGL16, Coh16b].

2.1 Independence-Preserving Mergers (IPM)

Informally speaking, a *merger* is a function that is given as input a sequence of random variables $M_1, \ldots, M_r$, one of which is uniform, while the others are arbitrary and may correlate with the former in arbitrary ways. As implied by its name, the task of a merger Merg is to "merge" the sequence to a new random variable $Z = \mathsf{Merg}(M_1, \ldots, M_r)$ that is close to uniform. We find it convenient to stack all M_i's as the rows of a matrix M. One can show that as we do not know which row M_g of M is uniform, and since all rows of M can correlate with M_g in arbitrary ways, for the merger to fulfil its task, it must have access to some "fresh" randomness, namely, to a random variable Y that is independent of M.

The problem of constructing *seeded-mergers*, namely, mergers with a uniformly distributed Y, attracted a significant attention in the literature [TS96a, TS96b, LRVW03, Raz05, DS07, DW09, DKSS09], mainly due to its role in some constructions of seeded extractors. Other works studied the problem of constructing *mergers with weak-seeds* [BRSW12, Coh15] in which Y is only assumed to be a weak-source.

Motivated by the problem of constructing multi-source extractors, the notion of an *independence-preserving merger (IPM)* was introduced in [CS16] and was further studied and used in other contexts [CL16, Coh16a]. This is a function IPM that, similarly to a "standard" merger, is given a matrix M and an auxiliary fresh randomness Y. Further, all rows of M are uniform (in which case, standard merging is trivial, deterministically). However, an adversary holds a matrix M' that is allowed to arbitrarily correlate with M but for the assumption that some row of M is uniform (even) conditioned on the corresponding

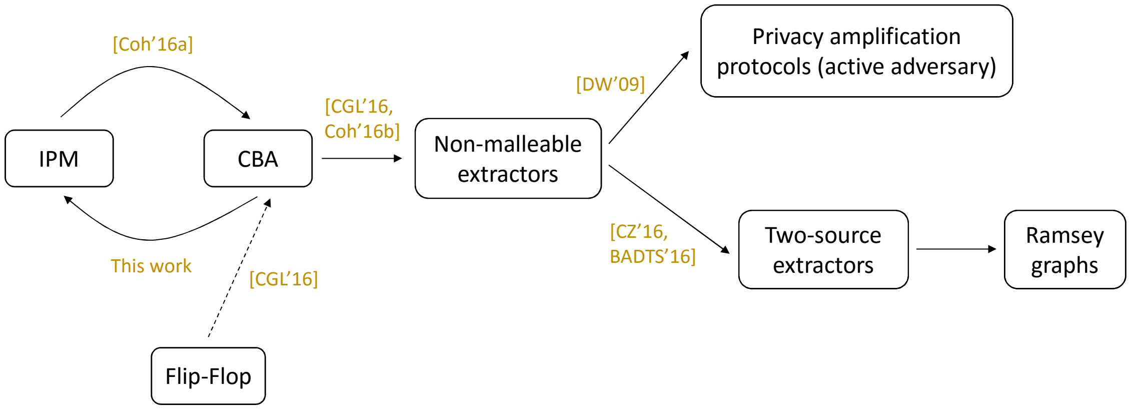

Figure 1: A schematic description of the connection between non-malleable extractors, their applications, and inner workings. The dashed arrow represents the first technique for constructing CBA [CGL16] via the flip-flop primitive [Coh15], which was subsumed by an IPM-based construction [Coh16a]. Among other ideas, our improved non-malleable extractors rely on a new, inverse, reduction from IPM to CBA.

row of M'. The guarantee of the independence-preserving merger is that $\mathsf{IPM}(M,Y)$ is close to uniform even when conditioned on $\mathsf{IPM}(M',Y')$ where Y' may correlate arbitrarily with Y. In that sense, IPM preserves the existing independence that one of the rows of M has with the corresponding row in M'.

Although seeded-IPM are natural objects, for current applications one is required to consider the stronger notion of an IPM with weak-seeds, namely, the IPM must work with Y that is not necessarily uniform and is only guaranteed to have some min-entropy k. The quantitative goal is to optimize k with respect to r and ε – the statistical distance of the output of IPM from uniform. In fact, for technical reasons, the formal definition (see Definition 4.2) is more involved, and we prefer to postpone it and carry out only a high-level discussion in this section.

Being somewhat imprecise for the sake of simplicity, in [CS16] an IPM was constructed for $k = r \cdot \log(1/\varepsilon)$. Subsequently, a strengthening of IPM was constructed in [CL16] for $k = 2^{\sqrt{\log r}} \cdot \log(1/\varepsilon)$. The main technical contribution of this work is the construction of an IPM with a lower min-entropy requirement. The formal statement is the content of Theorem 4.3. Here we settle for an informal statement.

Theorem 2.1 (Main technical contribution – informal statement)**.** *There exists an explicit IPM for r-row matrix with $k = \mathrm{polylog}(r) \cdot \log(1/\varepsilon)$.*

In fact, our construction yields a stronger and more natural variant of IPM as it does not require all rows of M to be uniform. The only requirement is that some row of M must be uniform when conditioned on the corresponding row of M'. We remark that, for different reasons, previous constructions [CL16, Coh16a] require all rows of M to be uniform. Throughout the paper we sometimes refer to our stronger notion of IPM as *IPM with no* *uniformity assumption.*

2.2 Correlation Breakers with Advice (CBA)

When constructing pseudorandom objects, one often faces undesired correlations between random variables. For examples, mergers are able to merge random variables despite their correlations, and IPM preserves, in some sense, an already acquired independence despite the presence of other correlations. Extractors can be thought of as breaking correlations between the different bits of the weak-source, etc.

As their name suggests, correlation breakers tackle the problem of breaking correlations between random variables heads on. Although a central issue, the problem of efficiently breaking arbitrary correlations some adversarial random variable has with a uniformly distributed random variable that we posses, using (unavoidably) an auxiliary source of randomness, was first explicitly studied by [Coh15] in the form of an object called a *local correlation breaker*, and was constructed based on techniques developed in [Li13] who obtained some restricted results on that direction. By adapting the construction of local correlation breakers, Chattopadhyay *et al.* [CGL16] gave a construction for a different type of correlation breakers, which was later explicitly defined and coined *correlation breakers with advice* [Coh16b]. This primitive is the main component, both conceptually and in terms of technical effort, in existing constructions of non-malleable extractors [CGL16, Coh16b, Coh16c]. Correlation breakers with advice found applications in other contexts as well [CS16].

The formal definition of CBA is fairly technical, and we choose to conduct an informal and high-level discussion here. For a formal treatment see Definition 3.16. The first construction of CBA [CGL16] was based on a sequential application of the so-called flip-flop primitive [Coh15]. The parameters of that construction are exponential in the advice length, which is the main parameter of complexity in these constructs. In [Coh16a], a reduction from CBA to IPM was established, which allowed for a construction of CBA with near-optimal parameters, and in particular with the optimal dependence on the advice length.

In this work we establish a reduction in the other direction, namely we show how to use a CBA for the construction of IPM. Combined with the original, inverse, reduction [Coh16a] we obtain CBA with improved parameters (see Theorem 5.1). We further remark that our reduction from non-malleable extractors to CBA has a slight twist on the original one [CGL16] and on its followup improvement [Coh16b] which allows us to save even further on randomness (see Section 6).

2.3 Independent Work

While writing this paper, we have learned that in a concurrent and independent work, Li [Li16] obtained results that are comparable to ours using different ideas. In particular, Li constructed a non-malleable extractor with seed length $d = O(\log n) + O(\log(1/\varepsilon) \cdot \log\log(1/\varepsilon))$. Li also obtained a 10-source extractor for n-bit sources with min-entropy $O(\log n)$ which is incomparable with our semi-explicit 4-source extractor for min-entropy $(1/3 + o(1)) \log_2 n$.

3 Preliminaries

In this section we set some notations that will be used throughout the paper and recall some of the more standard results from the literature that we make use of.

Setting some standard notations. Unless stated otherwise, the logarithm in this paper is always taken base 2. For every natural number $n \geq 1$, define $[n] = \{1, 2, \ldots, n\}$. We avoid the use of flooring and ceiling in order not to make the equations cumbersome. We say that a function is *explicit* or *efficiently-computable* when the corresponding family of functions can be computed by a (uniform) algorithm that runs in polynomial-time in the input length. In particular, when a real parameter ε is introduced, the running time is polynomial in $\log(1/\varepsilon)$ (as apposed to $1/\varepsilon$).

Random variables and distributions. We sometimes abuse notation and syntactically treat random variables and their distribution as equal, specifically, we denote by U_m a random variable that is uniformly distributed over $\{0,1\}^m$. Furthermore, if U_m appears in a joint distribution (U_m, X) then U_m should be understood as being independent of X. When m is clear from context, we omit it from the subscript and write U. The support of a random variable X is denoted by $\mathsf{supp}(X)$. Let X, Y be two random variables. We say that Y is a *deterministic function of* X if the value of X determines the value of Y. Namely, there exists a function f such that $Y = f(X)$.

Statistical distance. The *statistical distance* between two distributions X, Y on the same domain D is defined by

$$\mathsf{SD}\left(X, Y\right) = \max_{A \subseteq D} \left\{ \left| \mathbf{Pr}[X \in A] - \mathbf{Pr}[Y \in A] \right| \right\}.$$

If $\mathsf{SD}(X, Y) \leq \varepsilon$ we write $X \approx_\varepsilon Y$ and say that X and Y are ε-close.

3.1 Average Conditional Smooth Min-Entropy

Throughout the paper we make use of the notion of average conditional smooth min-entropy and some basic properties of it. We start by recalling the more basic notion of The *min-entropy*. The min-entropy of a random variable X, denoted by $\mathbf{H}_\infty(X)$, is defined by

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

If X is supported on $\{0,1\}^n$, we define the *min-entropy rate* of X by $\mathbf{H}_\infty(X)/n$. In such case, if X has min-entropy k or more, we say that X is an (n, k)-source. When wish to refer to an (n, k)-source without specifying the quantitative parameters, we sometimes use the standard terms *source* or *weak-source*.

Definition 3.1. *Let A, B be random variables. The* average conditional min-entropy *of A given B is defined as*

$$\mathbf{H}_\infty(A \mid B) = -\log_2\left(\mathop{\mathbf{E}}_{b\sim B}\left[\max_a \mathbf{Pr}\left[A = a \mid B = b\right]\right]\right).$$

Further, for an $\varepsilon > 0$ define

$$\mathbf{H}^\varepsilon_\infty(A \mid B) = \max \mathbf{H}_\infty(A' \mid B'),$$

where the maximum is taken over all (A', B') that are within statistical distance ε from (A, B). This quantity is referred to as the average conditional smooth min-entropy *of A given B, where ε is the* smoothness parameter*.*

Lemma 3.2 (Chain rule, [VDTR13])**.** *For any random variables A, B, C and for any $\varepsilon, \delta > 0$ it holds that*

$$\mathbf{H}^{\varepsilon+\delta}_\infty(A|BC) \geq \mathbf{H}^\varepsilon_\infty(AB|C) - |\mathsf{supp}(B)| - O(\log(1/\delta)),$$

where $\mathsf{supp}(B)$ is the support of B.

Theorem 3.3. *Let $f \colon D \to \{0,1\}$ be some function with domain D, and let $\mu = \mathbf{E}_x\left[f(x)\right]$ be the expectation of f, where x is sampled uniformly at random from D. Let $x_1, \ldots, x_t$ be elements that are sampled uniformly and independently at random from D. Then,*

$$\mathbf{Pr}\left[\left|\mu - \frac{1}{t} \cdot \sum_{i=1}^{t} f(x_i)\right| > \varepsilon\right] \leq 2 \cdot e^{-2\varepsilon^2 t}.$$

3.2 Building Blocks We Use

Throughout the paper we make use of several building blocks from the literature. We turn to state these results we use.

Extractors and condensers.

Definition 3.4 (Seeded extractors)**.** *A function $\mathsf{Ext} \colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is called a* (k, ε)-seeded extractor *if for any (n, k)-source X it holds that $\mathsf{Ext}(X, S) \approx_\varepsilon U_m$, where S is uniformly distributed over $\{0,1\}^d$ and is independent of X. We say that Ext is a* strong *if $(\mathsf{Ext}(X, S), S) \approx_\varepsilon U_{m+d}$.*

We sometimes say that an extractor Ext *supports* min-entropy k. By that we mean that Ext is an extractor for min-entropy k. Throughout the paper we make use of the following family of explicit strong seeded extractors.

Theorem 3.5 ([GUV09])**.** *There exists a universal constant $c_{\mathsf{GUV}} > 0$ such that the following holds. For all positive integers n, k and $\varepsilon > 0$, there exists an efficiently-computable (k, ε)-strong seeded-extractor $\mathsf{Ext} \colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ having seed length $d = c_{\mathsf{GUV}} \cdot \log(n/\varepsilon)$ and $m = k/2$ output bits. Further, one can have $m = (1-\alpha)k$ for any constant $\alpha > 0$ at the price of having a larger constant $c_{\mathsf{GUV}} = c_{\mathsf{GUV}}(\alpha)$.*

Theorem 3.6 ([GUV09], Theorem 4.4)**.** *For all integers n, k and $\varepsilon > 0$, there is a* $\mathrm{poly}(n, \log(1/\varepsilon))$*-time computable $k \to_\varepsilon k + d$ condenser* $\mathsf{Cond}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with*

$$\begin{aligned} d &= \log_2 n + \log_2 k + \log_2(1/\varepsilon) + 1, \\ m &= d(k+2). \end{aligned}$$

Lemma 3.7 ([Li11, CS16])**.** *Let* $\mathsf{Cond}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a $k \to_\varepsilon k'$ condenser. Let X be an (n,k)-source and let S be an independent random variable that is uniformly distributed over d-bit strings. Then, for any $\delta > 0$, with probability $1-\delta$ over $s \sim S$ it holds that* $\mathsf{Cond}(X, s)$ *is $(2\varepsilon/\delta)$-close to having min-entropy $k' - d - \log_2(2/\delta)$.*

Theorem 3.8. *There exist universal constants $c_{\mathsf{Raz}}, c'_{\mathsf{Raz}}$ such that the following holds. Let n, k be integers and let $\varepsilon > 0$. Set $d = c_{\mathsf{Raz}} \cdot \log(n/\varepsilon)$. For all $k \geq c'_{\mathsf{Raz}} d$, there exists an efficiently-computable function*

$$\mathsf{Raz}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{k/2}$$

with the following property. Let X be an (n,k)-source, and let Y be an independent $(d, 0.6d)$-source. Then, $(\mathsf{Raz}(X,Y), Y) \approx_\varepsilon (U, Y)$.

Theorem 3.9 ([BKS$^+$10, Raz05, Zuc07])**.** *For any constants $\delta_1, \delta_2 > 0$ there exists a constant integer $\Delta = \Delta(\delta_1, \delta_2) \geq 1$ such that the following holds. For any integer n there exists a sequence of efficiently computable functions* $\{\mathsf{Cond}_i\colon \{0,1\}^n \to \{0,1\}^{n/\Delta}\}_{i=1}^{\Delta}$ *such that the following holds. For any $(n, \delta_1 n)$-source X, the joint distribution of* $\{\mathsf{Cond}_i(X)\}_{i=1}^{\Delta}$ *is $2^{-n/\Delta}$-close to a convex combination such that for any participant $(Y_1, \ldots, Y_\Delta)$ in the combination, there exists $g \in [\Delta]$ such that Y_g has min-entropy rate $1 - \delta_2$.*

Definition 3.10 (Seeded condensers)**.** *A function* $\mathsf{Cond}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is said to be a* $k \to_\varepsilon k'$ condenser *if for any (n,k)-source X and for any independent random variable S that is uniformly distributed over d-bit strings, it holds that* $\mathsf{Cond}(X,S)$ *is ε-close to a random variable with min-entropy k'. The function* Cond *is called a* lossless condenser *if $k' = k + d$.*

We also make use of the following lemma.

Lemma 3.11 ([CG88])**.** *Any (n,k)-source X can be written as a convex combination of (n,k) flat sources.*

Error correcting codes. We also make us of the following standard definition of an error correcting code.

Definition 3.12. *Let Σ be some set. A mapping* $\mathsf{ECC}\colon \Sigma^k \to \Sigma^n$ *is called an error correcting code with relative-distance δ if for any $x, y \in \Sigma^k$, it holds that the Hamming distance between* $\mathsf{ECC}(x)$ *and* $\mathsf{ECC}(y)$ *is at least δn. The rate of the code, denoted by ρ, is defined by $\rho = k/n$. We say that the alphabet size of the code is $|\Sigma|$.*

Theorem 3.13 ([GS95] (see also [Sti09])). *Let p be any prime number and let m be an even integer. Set $q = p^m$. For every $\rho \in [0,1]$ and for any large enough integer n, there exists an efficiently-computable rate ρ linear error correcting code $\mathsf{ECC}\colon \mathbb{F}_q^{\rho n} \to \mathbb{F}_q^n$ with relative distance δ such that*

$$\rho + \delta \geq 1 - \frac{1}{\sqrt{q} - 1}.$$

Bounded-independence distributions.

Definition 3.14 ([NN93]). *Let n, k be integers such that $k \leq n$, and let $\delta > 0$. A random variable X over n-bit strings is called* (n, k, δ)-independent*, if for any $S \subseteq [n]$, with $|S| \leq k$, the marginal distribution X_S is δ-close to uniform.*

Theorem 3.15 ([NN93, AGHP92]). *For all $\delta > 0$ and integers n, k, there exists an explicit construction of an (n, k, δ)-independent sample space, with size $(2^k n/\delta)^{O(1)}$.*

Correlation breakers.

Definition 3.16 (Correlation breakers with advice). *A function*

$$\mathsf{CBA}\colon \{0,1\}^n \times \{0,1\}^\ell \times \{0,1\}^a \to \{0,1\}^m$$

*is called a (t, k, ε)-*correlation breaker with advice *(or (t, k, ε)-CBA for short) if the following holds. Let $\alpha, \alpha^1, \ldots, \alpha^t \in \{0,1\}^a$. Let $\mathcal{X} = (X, X^1, \ldots, X^t)$ be a sequence of n-bit random variables, $\mathcal{Y} = (Y, Y^1, \ldots, Y^t)$ a sequence of ℓ-bit random variables, and let $\mathcal{H}$ be a random variable for which the following holds:*

- *Conditioned on $\mathcal{H}$ the random variables $\mathcal{X}, \mathcal{Y}$ are independent;*
- *The strings $\alpha, \alpha^1, \ldots, \alpha^t \in \{0,1\}^a$ are fixed when conditioned on $\mathcal{H}$, and $\alpha \notin \{\alpha^i \mid i \in [t]\}$;*
- $\mathbf{H}_\infty^\varepsilon(X \mid \mathcal{H}) \geq k + \Omega(\log(1/\varepsilon))$*; and*
- $(Y, \mathcal{H}) \approx_\varepsilon (U, \mathcal{H})$*.*

Then,

$$\left(\mathsf{CBA}(X, Y, \alpha), \left\{\mathsf{CBA}\left(X^i, Y^i, \alpha^i\right)\right\}_{i=1}^t, \mathcal{Y}, \mathcal{H}\right) \approx_{O(\varepsilon)} \left(U, \left\{\mathsf{CBA}\left(X^i, Y^i, \alpha^i\right)\right\}_{i=1}^t, \mathcal{Y}, \mathcal{H}\right).$$

When considering $(t = 1, k, \varepsilon)$-CBA, we sometimes abbreviate and write (k, ε)-CBA. Further, we sometimes consider (t, k, ε)-CBA with $k = \delta n$ for some constant δ. We refer to such objects also as (t, δ, ε)-CBA, and note that this should never cause any confusion (as $\delta < 1 < k$). For our constructions we make use of the following construction of CBA.

Theorem 3.17 ([Coh16a]). *For any constant integers a, t there exists a constant $c = c(a, t) \geq 1$ such that the following holds. Let n, m be integers and let $\varepsilon > 0$. Then, there exists an explicit (t, k, ε)-CBA*

$$\mathsf{CBA}\colon \{0,1\}^n \times \{0,1\}^\ell \times \{0,1\}^a \to \{0,1\}^m$$

with $\ell = c \cdot \log(n/\varepsilon)$ and $k = c(m + \ell)$.

3.3 Hierarchy of Independence

Let n, b be integers and let $\varepsilon > 0$. Let c_{GUV} be the constant that is given by Theorem 3.5 and set $s = c_{\mathsf{GUV}} \cdot \log(n/\varepsilon)$. Note that s is sufficiently long so to be used as a seed for the strong seeded extractor that is given by Theorem 3.5 when fed with a sample from an n-bit source and when set with error guarantee ε. We make use of the following pair of extractors:

- Let $\mathsf{Ext}_{\mathsf{in}} \colon \{0,1\}^n \times \{0,1\}^s \to \{0,1\}^s$ be the $(2s, \varepsilon)$-strong seeded extractor that is given by Theorem 3.5.
- Let $\mathsf{Ext}_{\mathsf{out}} \colon \{0,1\}^n \times \{0,1\}^s \to \{0,1\}^b$ be the $(2b, \varepsilon)$-strong seeded extractor that is given by Theorem 3.5.

Define the pair of functions

$$\begin{aligned} \mathsf{a} &\colon \{0,1\}^s \times \{0,1\}^n \to \{0,1\}^b, \\ \mathsf{b} &\colon \{0,1\}^s \times \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^b, \end{aligned}$$

as follows. For $y \in \{0,1\}^s$ and $z, w \in \{0,1\}^n$,

$$\begin{aligned} \mathsf{a}(y, w) &= \mathsf{Ext}_{\mathsf{out}}(w, y), \\ \mathsf{b}(y, z, w) &= \mathsf{Ext}_{\mathsf{out}}(w, \mathsf{Ext}_{\mathsf{in}}(z, \mathsf{Ext}_{\mathsf{in}}(w, y))). \end{aligned}$$

The following lemma, in different forms and with different twists, appears in several previous works [DP07, DW09, Li13, Li15, Coh15, CS16, Coh16a].

Lemma 3.18. *Let $\mathcal{Y} = (Y, Y')$ be a pair of s-bit random variables, $\mathcal{Z} = (Z, Z')$ a pair of n-bit random variables, and let $\mathcal{W} = (W, W')$ be a pair of n-bit random variables. Let $\mathcal{H}$ be a random variable for which the following holds:*

- *Conditioned on $\mathcal{H}$, the random variable $\mathcal{W}$ is independent of $(\mathcal{Y}, \mathcal{Z})$;*
- *$(Y, \mathcal{H}) \approx_\delta (U, \mathcal{H})$;*
- *$\mathbf{H}_\infty^\varepsilon(Z \mid \mathcal{H}) \geq 4s + \Omega(\log(1/\varepsilon))$; and*
- *$\mathbf{H}_\infty^\varepsilon(W \mid \mathcal{H}) \geq 2b + 2s + \Omega(\log(1/\varepsilon))$.*

Write

$$\begin{aligned} \widehat{\mathcal{A}} &= \mathsf{a}(Y, W), \mathsf{a}(Y', W'), \\ \widehat{\mathcal{Z}} &= \mathsf{Ext}_{\mathsf{in}}(Z, \mathsf{Ext}_{\mathsf{in}}(W, Y)), \mathsf{Ext}_{\mathsf{in}}(Z', \mathsf{Ext}_{\mathsf{in}}(W', Y')). \end{aligned}$$

Then, the following holds:

1. $(\mathsf{a}(Y, W), \mathcal{Z}, \mathcal{Y}, \mathcal{H}) \approx_{\delta + 2\varepsilon} (U, \mathcal{Z}, \mathcal{Y}, \mathcal{H})$,

2. $\left(\mathsf{b}(Y,Z,W),\mathcal{Z},\widehat{\mathcal{Z}},\widehat{\mathcal{A}},\mathcal{Y},\mathcal{H}\right) \approx_{\delta+6\varepsilon} \left(U,\mathcal{Z},\widehat{\mathcal{Z}},\widehat{\mathcal{A}},\mathcal{Y},\mathcal{H}\right)$.

Furthermore,

3. $\mathbf{H}_{\infty}^{2\varepsilon}\left(Z \mid \widehat{\mathcal{Z}},\widehat{\mathcal{A}},\mathcal{Y},\mathcal{H}\right) \geq \mathbf{H}_{\infty}^{\varepsilon}\left(Z \mid \mathcal{H}\right) - 4s - O(\log(1/\varepsilon))$,

4. $\mathbf{H}_{\infty}^{2\varepsilon}\left(W \mid \widehat{\mathcal{Z}},\widehat{\mathcal{A}},\mathcal{Y},\mathcal{H}\right) \geq \mathbf{H}_{\infty}^{\varepsilon}\left(W \mid \mathcal{H}\right) - 2b - 2s - O(\log(1/\varepsilon))$.

4 IPM with No Uniformity Assumption

Definition 4.1 (Somewhere-independent matrices with no uniformity assumption)**.** *Let M, M' be a pair of random variables in the form of $r \times \ell$ matrices. Let $\mathcal{H}$ be a random variable and let $\delta > 0$. We say that M is $(\delta,\mathcal{H})$-*somewhere independent of *M' if there exists $g \in [r]$ such that*

$$\left(M_g, M'_g, \mathcal{H}\right) \approx_{\delta} \left(U, M'_g, \mathcal{H}\right).$$

Definition 4.2 (IPM with no uniformity assumption)**.** *A function*

$$\mathsf{IPM}\colon \{0,1\}^{r\times\ell} \times \{0,1\}^{d} \times \{0,1\}^{d} \to \{0,1\}^{\ell} \tag{4.1}$$

*is called a (k,ε)-*independence preserving merger (or (k,ε)-IPM for short) with no uniformity assumption *if the following holds. Let $\mathcal{X} = (X, X')$ be a pair of d-bit random variables, $\mathcal{Y} = (Y, Y')$ a pair of d-bit random variables, and let $\mathcal{M} = (M, M')$ be a pair of random variables in the form of $r \times \ell$ matrices. Let $\mathcal{H}$ be a random variable for which the following holds:*

- *Conditioned on $\mathcal{H}$ the random variable $\mathcal{X}$ is independent of $(\mathcal{M},\mathcal{Y})$;*
- $\mathbf{H}_{\infty}^{\varepsilon}\left(X \mid \mathcal{H}\right) \geq k + \Omega(\log(1/\varepsilon))$*;*
- $\mathbf{H}_{\infty}^{\varepsilon}\left(Y \mid \mathcal{M},\mathcal{H}\right) \geq k + \Omega(\log(1/\varepsilon))$*; and*
- *M is $(\varepsilon,\mathcal{H})$-somewhere independent of M'.*

Then,

$$\left(\mathsf{IPM}(M,X,Y), \mathsf{IPM}(M',X',Y'), \mathcal{M}, \mathcal{Y}, \mathcal{H}\right) \approx_{O(\varepsilon)} \left(U, \mathsf{IPM}(M',X',Y'), \mathcal{M}, \mathcal{Y}, \mathcal{H}\right).$$

Some remarks. Unlike previous works [CS16, CL16, Coh16a], our construction of independence-preserving mergers satisfies the stronger notion of being an independence-preserving merger with no uniformity assumption. That is, we do not require that $\forall i \in [r] \ \ (M_i, H) \approx_{\delta} (U, H)$. Thus, for the rest of this paper we simply use the term independence-preserving mergers (or IPM for short) when referring to the stronger notion that is introduced in Definition 4.2. Further, we sometimes consider (k,ε)-IPM as in (4.1) with $k = \delta d$ for some constant δ. We refer to such objects as (δ, ε)-IPM, and note that this should never cause any confusion (as $\delta < 1 < k$).

The main result proved in this section, which is the main technical contribution of this work, is the following theorem, which is a formal restatement of Theorem 2.1.

Theorem 4.3. *For any constant $\tau > 0$ there exists a constant $c = c(\tau) \geq 1$ such that the following holds. For all integers r, ℓ and for any $\varepsilon > 0$ such that $\ell = \Omega(\log(\log(r)/\varepsilon))$, there exists an explicit $(6/7 + \tau, \varepsilon)$-IPM*

$$\mathsf{IPM}\colon \{0,1\}^{r \times \ell} \times \{0,1\}^d \times \{0,1\}^d \to \{0,1\}^\ell$$

with $d = O(\ell \cdot \log^c r)$.

The construction of the IPM stated in Theorem 4.3 is recursive. For the base of the recursion we need an IPM with no uniformity assumption for a constant number of rows. We construct this base IPM in the following section. This is the content of Lemma 4.4. We then proceed to prove Theorem 4.3 in Section 4.2.

4.1 IPM for a Constant Number of Rows via CBA

Lemma 4.4. *For any constant integer r, any integers d, ℓ, and any $\varepsilon > 0$ such that $\ell = \Omega(\log(d/\varepsilon))$ there exists an explicit (k, ε)-IPM*

$$\mathsf{BaseIPM}\colon \{0,1\}^{r \times \ell} \times \{0,1\}^d \times \{0,1\}^d \to \{0,1\}^\ell$$

with $k = \Omega(\ell)$.

For the proof of Lemma 4.4 we first observe a property of CBA. Correlation breakers with advice are designated to break correlations between random variables when fed with *distinct* advices. In the following lemma we show that any CBA is also independence-preserving in the sense that if some random variable is already uniform conditioned on another, that independence is preserved even if one applies a CBA to both variables using the *same* advice string. We make this formal in the following lemma.

Lemma 4.5. *Let $\mathsf{CBA}\colon \{0,1\}^n \times \{0,1\}^\ell \times \{0,1\}^a \to \{0,1\}^m$ be a $(t, k, 2\varepsilon)$-CBA. Let $\alpha, \alpha^1, \ldots, \alpha^t \in \{0,1\}^a$, and set $I = \{i \mid \alpha = \alpha^i\}$. Let $\mathcal{X} = (X, X^1, \ldots, X^t)$ be a sequence of n-bit random variables, $\mathcal{Y} = (Y, Y^1, \ldots, Y^t)$ a sequence of ℓ-bit random variables, and let $\mathcal{H}$ be a random variable for which the following holds:*

- *Conditioned on $\mathcal{H}$, the random variables $\mathcal{X}, \mathcal{Y}$ are independent;*
- *The strings $\alpha, \alpha^1, \ldots, \alpha^t$ are fixed when conditioned on $\mathcal{H}$;*
- $\mathbf{H}_\infty^\varepsilon(X \mid \mathcal{H}) \geq k + m|I| + \Omega(\log(1/\varepsilon))$*; and*
- $(Y, \{Y^i\}_{i \in I}, \mathcal{H}) \approx_\varepsilon (U, \{Y^i\}_{i \in I}, \mathcal{H})$.

Then,

$$\left(\mathsf{CBA}\left(X,Y,\alpha\right),\left\{\mathsf{CBA}\left(X^i,Y^i,\alpha^i\right)\right\}_{i=1}^{t},\mathcal{Y},\mathcal{H}\right)\approx_{O(\varepsilon)}\left(U,\left\{\mathsf{CBA}\left(X^i,Y^i,\alpha^i\right)\right\}_{i=1}^{t},\mathcal{Y},\mathcal{H}\right). \tag{4.2}$$

Proof. By the hypothesis of the lemma,

$$\left(Y,\{Y^i\}_{i\in I},\mathcal{H}\right)\approx_{\varepsilon}\left(U,\{Y^i\}_{i\in I},\mathcal{H}\right).$$

Conditioned on $\{Y^i \mid i\in I\},\mathcal{H}$, the random variable Y is independent of the joint distribution of $\{X^i \mid i\in I\}$, and so we can adjoin the latter to the above equation and obtain

$$\left(Y,\{X^i,Y^i\}_{i\in I},\mathcal{H}\right)\approx_{\varepsilon}\left(U,\{X^i,Y^i\}_{i\in I},\mathcal{H}\right).$$

As $\mathsf{CBA}(X^i,Y^i,\alpha^i)$ is a deterministic function of X^i,Y^i, we conclude that

$$\left(Y,\mathcal{H}_1\right)\approx_{\varepsilon}\left(U,\mathcal{H}_1\right), \tag{4.3}$$

where $\mathcal{H}_1=\{\mathsf{CBA}(X^i,Y^i,\alpha^i),Y^i \mid i\in I\},\mathcal{H}$. Note that we removed the random variables $\{X^i \mid i\in I\}$ when deducing (4.3) and preserved only the corresponding set of outputs $\{\mathsf{CBA}(X^i,Y^i,\alpha^i) \mid i\in I\}$. This step is crucial for the following derivation. By Lemma 3.2,

$$\mathbf{H}_{\infty}^{2\varepsilon}(X\mid\mathcal{H}_1)\ge\mathbf{H}_{\infty}^{\varepsilon}(X\mid\mathcal{H})-m|I|-O(\log(1/\varepsilon))\ge k+\Omega(\log(1/\varepsilon)). \tag{4.4}$$

Let $\mathcal{X}'=\{X^i \mid i\notin I\}$, $\mathcal{Y}'=\{Y^i \mid i\notin I\}$. Note that conditioned on $\mathcal{H}_1$, the random variables $\mathcal{X}',\mathcal{Y}'$ are independent. By (4.3), (4.4) we may apply CBA to $\mathcal{X}',\mathcal{Y}'$ with $\mathcal{H}_1$ and the corresponding advices $\{\alpha^i \mid i\notin I\}$ to conclude that

$$\left(\mathsf{CBA}(X,Y,\alpha),\{\mathsf{CBA}(X^i,Y^i,\alpha^i)\}_{i\notin I},\mathcal{Y}',\mathcal{H}_1\right)\approx_{O(\varepsilon)}\left(U,\{\mathsf{CBA}(X^i,Y^i,\alpha^i)\}_{i\notin I},\mathcal{Y}',\mathcal{H}_1\right),$$

which readily concludes the proof. □

With Lemma 4.5 in hand, we are now ready to prove Lemma 4.4.

Proof of Lemma 4.4. Let c_{GUV} be the constant that is given by Theorem 3.5. Set $a=\log r$, $t=2r$, and $m=c_{\mathsf{GUV}}\cdot\log(d/\varepsilon)$. Note that a,t are constants.

Building blocks. For the construction of $\mathsf{BaseIPM}$ we make use of the following building blocks:

- Let $\mathsf{CBA}\colon\{0,1\}^d\times\{0,1\}^\ell\times\{0,1\}^a\to\{0,1\}^m$ be the $(t,k-m-O(\log(1/\varepsilon)),\varepsilon)$-CBA that is given by Theorem 3.17. Note that the hypothesis of Theorem 3.17 regarding ℓ,k is met.

- Let $\mathsf{Ext}_1\colon\{0,1\}^d\times\{0,1\}^m\to\{0,1\}^m$ be the $(2m,\varepsilon)$-strong seeded extractor that is given by Theorem 3.5. Note that m was chosen as required by Theorem 3.5.

- Let $\mathsf{Ext}_2\colon\{0,1\}^d\times\{0,1\}^m\to\{0,1\}^\ell$ be the $(2\ell,\varepsilon)$-strong seeded extractor that is given by Theorem 3.5.

The construction. Let $m \in \{0,1\}^{r \times \ell}$ and let $x, y \in \{0,1\}^d$. For $i \in [r]$ we define

$$z_i = \mathsf{CBA}(x, m_i, i), \tag{4.5}$$

where by feeding i as the third argument to CBA we formally mean the binary string obtained by writing the integer i to the base 2. Note that indeed the advice length is $\log r = a$. Define

$$s = \bigoplus_{i=1}^{r} z_i,$$
$$t = \mathsf{Ext}_1(y, s),$$

and set

$$\mathsf{BaseIPM}(m, x, y) = \mathsf{Ext}_2(x, t).$$

Analysis. Let $\mathcal{X} = (X, X')$ be a pair of d-bit random variables, $\mathcal{Y} = (Y, Y')$ a pair of d-bit random variables, and let $\mathcal{M} = (M, M')$ be a pair of random variables in the form of $r \times \ell$ matrices. Let $\mathcal{H}$ be a random variable for which the following holds:

- Conditioned on $\mathcal{H}$, the random variable $\mathcal{X}$ is independent of $(\mathcal{M}, \mathcal{Y})$;
- $\mathbf{H}_{\infty}^{\varepsilon}(X \mid \mathcal{H}) \geq k + \Omega(\log(1/\varepsilon))$;
- $\mathbf{H}_{\infty}^{\varepsilon}(Y \mid \mathcal{M}, \mathcal{H}) \geq k + \Omega(\log(1/\varepsilon))$; and
- $(M_g, M'_g, \mathcal{H}) \approx_{\varepsilon} (U, M'_g, \mathcal{H})$ for some $g \in [r]$.

Note that we set the advice strings in such a way that with every row in the pair of matrices M, M' we associate an advice that is distinct of all other advices but for one – the one associated with the corresponding row in the other matrix. Further, clearly the advices are fixed (also when conditioned on $\mathcal{H}$). Thus, we may apply Lemma 4.5 with Theorem 3.17 to conclude that

$$\left(Z_g, \{Z_i\}_{i \in [r] \setminus \{g\}}, \{Z'_i\}_{i=1}^{r}, \mathcal{M}, \mathcal{H}\right) \approx_{O(\varepsilon)} \left(U, \{Z_i\}_{i \in [r] \setminus \{g\}}, \{Z'_i\}_{i=1}^{r}, \mathcal{M}, \mathcal{H}\right).$$

It then follows that

$$(S, S', \mathcal{M}, \mathcal{H}) \approx_{O(\varepsilon)} (U, S', \mathcal{M}, \mathcal{H}),$$

where we used the simple fact that $(A, B, C) \approx_{\varepsilon} (U, B, C)$ implies $(A \oplus B, B, C) \approx_{\varepsilon} (U, B, C)$. Conditioned on $S', \mathcal{M}, \mathcal{H}$, the random variable S is independent of Y' and so we may adjoin Y' to obtain

$$(S, Y', S', \mathcal{M}, \mathcal{H}) \approx_{O(\varepsilon)} (U, Y', S', \mathcal{M}, \mathcal{H}).$$

As $T' = \mathsf{Ext}_1(Y', S')$ is a deterministic function of Y', S' we conclude that

$$(S, T', S', \mathcal{M}, \mathcal{H}) \approx_{O(\varepsilon)} (U, T', S', \mathcal{M}, \mathcal{H}).$$

Note that we removed Y when deducing the above equation. This is crucial for the following. Conditioned on $T', S', \mathcal{M}, \mathcal{H}$, the random variables S, Y are independent. Further, by Lemma 3.2,

$$\begin{aligned}\mathbf{H}^{2\varepsilon}_{\infty}(Y \mid T', S', \mathcal{M}, \mathcal{H}) &\geq \mathbf{H}^{\varepsilon}_{\infty}(Y \mid \mathcal{M}, \mathcal{H}) - m - O(\log(1/\varepsilon))\\ &\geq k - m - O(\log(1/\varepsilon))\\ &\geq 2m + \Omega(\log(1/\varepsilon)),\end{aligned}$$

and so

$$(T, S, T', S', \mathcal{M}, \mathcal{H}) \approx_{O(\varepsilon)} (U, S, T', S', \mathcal{M}, \mathcal{H})\,.$$

Conditioned on $S, T', S', \mathcal{M}, \mathcal{H}$, the random variables T, X' are independent, and so

$$(T, X', S, T', S', \mathcal{M}, \mathcal{H}) \approx_{O(\varepsilon)} (U, X', S, T', S', \mathcal{M}, \mathcal{H})\,.$$

The above equation together with the fact that $\mathsf{BaseIPM}(M', X', Y') = \mathsf{Ext}_2(X', T')$ is a deterministic function of X', T' implies that

$$(T, \mathsf{BaseIPM}(M', X', Y'), S, T', S', \mathcal{M}, \mathcal{H}) \approx_{O(\varepsilon)} (U, \mathsf{BaseIPM}(M', X', Y'), S, T', S', \mathcal{M}, \mathcal{H})\,. \tag{4.6}$$

Further,

$$\begin{aligned}\mathbf{H}^{2\varepsilon}_{\infty}(X \mid \mathsf{BaseIPM}(M', X', Y'), S, T', S', \mathcal{M}, \mathcal{H}) &\geq \mathbf{H}^{\varepsilon}_{\infty}(X \mid \mathcal{H}) - \ell - 2m\\ &\geq 2\ell + \Omega(\log(1/\varepsilon)).\end{aligned} \tag{4.7}$$

By Equations (4.6), (4.7), and by the fact that X, T are independent when conditioned on the output $\mathsf{BaseIPM}(M', X', Y')$, and $S, T', S', \mathcal{M}, \mathcal{H}$, we have that

$$\begin{aligned}&(\mathsf{BaseIPM}(M, X, Y), T, \mathsf{BaseIPM}(M', X', Y'), S, T', S', \mathcal{M}, \mathcal{H}) \approx_{O(\varepsilon)}\\ &(U, T, \mathsf{BaseIPM}(M', X', Y'), S, T', S', \mathcal{M}, \mathcal{H})\,.\end{aligned}$$

Conditioned on $T, \mathsf{BaseIPM}(M', X', Y'), S, T', S', \mathcal{M}, \mathcal{H}$, the random variable $\mathsf{BaseIPM}(M, X, Y)$ is independent of $\mathcal{Y}$, and so we can adjoin $\mathcal{Y}$ and remove the excess random variables to obtain

$$(\mathsf{BaseIPM}(M, X, Y), \mathsf{BaseIPM}(M', X', Y'), \mathcal{M}, \mathcal{Y}, \mathcal{H}) \approx_{O(\varepsilon)} (U, \mathsf{BaseIPM}(M', X', Y'), \mathcal{M}, \mathcal{Y}, \mathcal{H})\,,$$

which completes the proof. □

4.2 Proof of Theorem 4.3

Before proving Theorem 4.3 we prove the following lemma.

Lemma 4.6. *Let X, X' be n-bit random variables, and let $\mathcal{H}$ be a random variable such that $\mathbf{H}^{\varepsilon}_{\infty}(X \mid \mathcal{H}) \geq (1-\delta)n$. Let $\tau > 0$, and define X_1, X'_1 to be the length $n_1 = (\delta+\tau)n$ bit prefixes of X, X', respectively. Define X_2, X'_2 to be the length $n_2 = 3(\delta+\tau)n$ bit prefixes of X, X', respectively. Then, the following holds:*

- $\mathbf{H}_\infty^{2\varepsilon}(X_1 \mid \mathcal{H}) \geq \tau n_1 - O(\log(1/\varepsilon))$;
- $\mathbf{H}_\infty^{2\varepsilon}(X_2 \mid X_1, X_1', \mathcal{H}) \geq \tau n_2 - O(\log(1/\varepsilon))$; *and*
- $\mathbf{H}_\infty^{2\varepsilon}(X \mid X_2, X_2', \mathcal{H}) \geq (1 - 7\delta - 6\tau)n - O(\log(1/\varepsilon))$.

Proof. Write $X = X_1 \circ X_{>1} = X_2 \circ X_{>2}$, where $X_{>1}, X_{>2}$ are of length $n - n_1$, $n - n_2$, respectively. Using Lemma 3.2 we obtain

$$\begin{aligned}
\mathbf{H}_\infty^{2\varepsilon}(X_1 \mid \mathcal{H}) &\geq \mathbf{H}_\infty^{2\varepsilon}(X \mid X_{>1}, \mathcal{H}) \\
&\geq \mathbf{H}_\infty^{\varepsilon}(X \mid \mathcal{H}) - |X_{>1}| - O(\log(1/\varepsilon)) \\
&\geq (1-\delta)n - (1 - \delta - \tau)n - O(\log(1/\varepsilon)) \\
&= \tau n - O(\log(1/\varepsilon)) \\
&\geq \tau n_1 - O(\log(1/\varepsilon)),
\end{aligned}$$

which proves the first item. As for the second item,

$$\begin{aligned}
\mathbf{H}_\infty^{2\varepsilon}(X_2 \mid X_1, X_1', \mathcal{H}) &\geq \mathbf{H}_\infty^{2\varepsilon}(X \mid X_{>2}, X_1, X_1', \mathcal{H}) \\
&\geq \mathbf{H}_\infty^{\varepsilon}(X \mid \mathcal{H}) - |X_{>2}| - |X_1| - |X_1'| - O(\log(1/\varepsilon)) \\
&\geq (1-\delta)n - (n - 3(\delta+\tau)n) - 2(\delta+\tau)n - O(\log(1/\varepsilon)) \\
&= \tau n - O(\log(1/\varepsilon)) \\
&\geq \tau n_2 - O(\log(1/\varepsilon)),
\end{aligned}$$

which concludes the proof of the second item. The third item follows by a similar argument:

$$\begin{aligned}
\mathbf{H}_\infty^{2\varepsilon}(X \mid X_2, X_2', \mathcal{H}) &\geq \mathbf{H}_\infty^{\varepsilon}(X \mid \mathcal{H}) - 2 \cdot 3(\delta+\tau)n - O(\log(1/\varepsilon)) \\
&\geq (1-\delta)n - 6(\delta+\tau)n - O(\log(1/\varepsilon)) \\
&= (1 - 7\delta - 6\tau)n - O(\log(1/\varepsilon)).
\end{aligned}$$

□

Lemma 4.6 readily implies the following corollary by setting $\tau = 1/7 - \delta$.

Corollary 4.7. *Let $\tau > 0$. Let X, X' be n-bit random variables such that $\mathbf{H}_\infty^{\varepsilon}(X \mid \mathcal{H}) \geq (6/7+\tau)n + \Omega(\log(1/\varepsilon))$ for some random variable $\mathcal{H}$. Define X_1, X_1' to be the length $n_1 = n/7$ bit prefixes of X, X', respectively. Define X_2, X_2' to be the length $n_2 = 3n/7$ bit prefixes of X, X', respectively. Then, the following holds:*

- $\mathbf{H}_\infty^{2\varepsilon}(X_1 \mid \mathcal{H}) \geq \tau n_1 - O(\log(1/\varepsilon))$;
- $\mathbf{H}_\infty^{2\varepsilon}(X_2 \mid X_1, X_1', \mathcal{H}) \geq \tau n_2 - O(\log(1/\varepsilon))$; *and*
- $\mathbf{H}_\infty^{2\varepsilon}(X \mid X_2, X_2', \mathcal{H}) \geq \tau n - O(\log(1/\varepsilon))$.

We are now ready to prove Theorem 4.3.

Proof of Theorem 4.3. We construct the function IPM recursively. More precisely, for any integer r and any ℓ that meet the hypothesis of the theorem, we construct a $(6/7+\tau, \varepsilon(r))$-IPM

$$\mathsf{IPM}_r \colon \{0,1\}^{r \times \ell} \times \{0,1\}^{d(r)} \times \{0,1\}^{d(r)} \to \{0,1\}^{\ell}$$

given an explicit $(6/7+\tau, \varepsilon(\sqrt{r}))$-IPM

$$\mathsf{IPM}_{\sqrt{r}} \colon \{0,1\}^{\sqrt{r} \times \ell} \times \{0,1\}^{d(\sqrt{r})} \times \{0,1\}^{d(\sqrt{r})} \to \{0,1\}^{\ell},$$

where $\varepsilon(\sqrt{r})$ is set with hindsight to be $\Theta(\varepsilon(r))$. For ease of readability, we let $d = d(r)$. For a d-bit string s we let s_1, s_2 denote the length $d/7$, $3d/7$ prefixes of s, respectively.

Building blocks. On top of $\mathsf{IPM}_{\sqrt{r}}$, for the construction of IPM_r we make use of the following building blocks:

- Let $\{\mathsf{Cond}^1_i \colon \{0,1\}^{d/7} \to \{0,1\}^{d/(7\Delta)}\}_{i=1}^{\Delta}$ be the sequence of functions that is given by Theorem 3.9 set with $\delta_1 = \tau/2$ and $\delta_2 = 1/7 - 2\tau$. By Theorem 3.9, $\Delta = \Delta(\tau)$ is some constant.
- Let $\{\mathsf{Cond}^2_i \colon \{0,1\}^{3d/7} \to \{0,1\}^{3d/(7\Delta)}\}_{i=1}^{\Delta}$ be the sequence of functions that is given by Theorem 3.9 also set with $\delta_1 = \tau/2$ and $\delta_2 = 1/7 - 2\tau$.
- Let $\mathsf{BaseIPM} \colon \{0,1\}^{\Delta^4 \times \ell} \times \{0,1\}^{d} \times \{0,1\}^{d} \to \{0,1\}^{\ell}$ be the $(k, \varepsilon(r))$-IPM that is given by Lemma 4.4.

Note that the output length of the functions $\{\mathsf{Cond}^2_i\}_i$ is 3 times longer than that of the functions $\{\mathsf{Cond}^1_i\}_i$. For technical reasons, it will be simpler for these two sequences of functions to have a common output length. This can be easily achieved without any asymptotic affect on the parameters. Thus, from this point on we assume that the output length of the functions $\mathsf{Cond}^1_i, \mathsf{Cond}^2_i$ is $d' = \alpha d$ for some constant α (recall that Δ is constant). We further define $d'_1 = d'/7$ and $d'_2 = 3d'/7$.

The construction. Let $m \in \{0,1\}^{r \times \ell}$ and let $x, y \in \{0,1\}^{d}$. Let $m^1, \ldots, m^{\sqrt{r}}$ be $\sqrt{r} \times \ell$ matrices obtained by partitioning the r rows of m in an arbitrary manner. For concreteness, assume that m^i contains rows $(i-1)\sqrt{r}+1, \ldots, i\sqrt{r}$ of m. For $(i_1, j_1) \in [\Delta]^2$ define the $\sqrt{r} \times \ell$ matrix $z^{(i_1,j_1)}$ as follows. For $v \in [\sqrt{r}]$, row v of $z^{(i_1,j_1)}$ is defined as

$$z^{(i_1,j_1)}_v = \mathsf{IPM}_{\sqrt{r}}\left(m^v, \mathsf{Cond}^1_{i_1}(x_1), \mathsf{Cond}^1_{j_1}(y_1)\right). \tag{4.8}$$

Define the $\Delta^4 \times \ell$ matrix t, with rows indexed by $(i_1, j_1, i_2, j_2) \in [\Delta]^4$ by

$$t_{(i_1,j_1,i_2,j_2)} = \mathsf{IPM}_{\sqrt{r}}\left(z^{(i_1,j_1)}, \mathsf{Cond}^2_{j_2}(y_2), \mathsf{Cond}^2_{i_2}(x_2)\right). \tag{4.9}$$

Finally, define

$$\mathsf{IPM}_r(m, x, y) = \mathsf{BaseIPM}(t, x, y).$$

Note that in the above construction, and in particular in Equations (4.8),(4.9), we implicitly set $d(\sqrt{r}) = d' = \alpha d = \alpha d(r)$.

Analysis. Let $\mathcal{X} = (X, X')$ be a pair of d-bit random variables, $\mathcal{Y} = (Y, Y')$ a pair of d-bit random variables, and $\mathcal{M} = (M, M')$ a pair of random variables in the form of $r \times \ell$ matrices. Let $\mathcal{H}$ be a random variable such that the following holds:

- Conditioned on $\mathcal{H}$, the random variable $\mathcal{X}$ is independent of $(\mathcal{M}, \mathcal{Y})$;
- $\mathbf{H}_\infty^{\varepsilon(r)}(X \mid \mathcal{H}) \geq (6/7 + \tau)d + \Omega(\log(1/\varepsilon(r)))$;
- $\mathbf{H}_\infty^{\varepsilon(r)}(Y \mid \mathcal{M}, \mathcal{H}) \geq (6/7 + \tau)d + \Omega(\log(1/\varepsilon(r)))$; and
- M is $(\varepsilon(r), \mathcal{H})$-somewhere independent of M'.

By Corollary 4.7,

$$\mathbf{H}_\infty^{2\varepsilon(r)}(X_1 \mid \mathcal{H}) \geq \tau d_1 - O(\log(1/\varepsilon(r))) \geq (\tau/2)d_1,$$

where we used $\varepsilon(r) > 2^{-\Omega(d)}$, the fact that τ is constant, and that $d_1 = \Theta(d)$. Therefore, by Theorem 3.9 there exists $i_1^* \in [\Delta]$ such that

$$\mathbf{H}_\infty^{3\varepsilon(r)}\left(\mathsf{Cond}^1_{i_1^*}(X_1) \mid \mathcal{H}\right) \geq (6/7 + 2\tau)d_1' \geq (6/7 + \tau)d_1' + \Omega(\log(1/\varepsilon(r))), \tag{4.10}$$

where, again, we used $\varepsilon(r) > 2^{-\Omega(d)}$, and $d_1 = \Theta(d)$. Similarly, there exists $j_1^* \in [\Delta]$ such that

$$\mathbf{H}_\infty^{3\varepsilon(r)}\left(\mathsf{Cond}^1_{j_1^*}(Y_1) \mid \mathcal{M}, \mathcal{H}\right) \geq (6/7 + \tau)d_1' + \Omega(\log(1/\varepsilon(r))). \tag{4.11}$$

As M is $(\varepsilon(r), \mathcal{H})$-somewhere independent of M', by the way we defined $M^1, \ldots, M^{\sqrt{r}}$, there exist $g_1, g_2 \in [\sqrt{r}]$ such that

$$\left(M^{g_1}_{g_2}, (M')^{g_1}_{g_2}, \mathcal{H}\right) \approx_{\varepsilon(r)} \left(U, (M')^{g_1}_{g_2}, \mathcal{H}\right). \tag{4.12}$$

Recall that $\mathsf{IPM}_{\sqrt{r}}$ is a $(6/7 + \tau, \varepsilon(\sqrt{r}))$-IPM and that $\varepsilon(\sqrt{r}) = \Theta(\varepsilon(r))$. Equations (4.10), (4.11), (4.12) imply that

$$\left(Z^{(i_1^*, j_1^*)}_{g_1}, (Z')^{(i_1^*, j_1^*)}_{g_1}, \{M^{g_1}_i, (M')^{g_1}_i\}_{i=1}^{\sqrt{r}}, \mathsf{Cond}^1_{j_1^*}(Y_1), \mathsf{Cond}^1_{j_1^*}(Y_1'), \mathcal{H}\right) \approx_{O(\varepsilon(r))}$$
$$\left(U, (Z')^{(i_1^*, j_1^*)}_{g_1}, \{M^{g_1}_i, (M')^{g_1}_i\}_{i=1}^{\sqrt{r}}, \mathsf{Cond}^1_{j_1^*}(Y_1), \mathsf{Cond}^1_{j_1^*}(Y_1'), \mathcal{H}\right),$$

which readily implies that

$$\left(Z^{(i_1^*, j_1^*)}_{g_1}, (Z')^{(i_1^*, j_1^*)}_{g_1}, \mathcal{H}_1\right) \approx_{O(\varepsilon(r))} \left(U, (Z')^{(i_1^*, j_1^*)}_{g_1}, \mathcal{H}_1\right), \tag{4.13}$$

with $\mathcal{H}_1 = \mathcal{M}, Y_1, Y_1', \mathcal{H}$. Hence, the $\sqrt{r} \times \ell$ matrix $Z^{(i_1^*, j_1^*)}$ is $(O(\varepsilon(r)), \mathcal{H}_1)$-somewhere independent of the matrix $(Z')^{(i_1^*, j_1^*)}$.

Note that the random variable $Z^{(i_1^*,j_1^*)}$ (resp. $(Z')^{(i_1^*,j_1^*)}$) is a deterministic function of X_1 (resp. X_1') when conditioned on $\mathcal{H}_1$. This, together with Corollary 4.7, implies that

$$\begin{aligned}\mathbf{H}_\infty^{2\varepsilon(r)}\left(X_2 \mid Z^{(i_1^*,j_1^*)}, (Z')^{(i_1^*,j_1^*)}, \mathcal{H}_1\right) &\geq \mathbf{H}_\infty^{2\varepsilon(r)}\left(X_2 \mid X_1, X_1', \mathcal{H}_1\right)\\ &= \mathbf{H}_\infty^{2\varepsilon(r)}\left(X_2 \mid X_1, X_1', \mathcal{H}\right)\\ &\geq \tau d_2 - O(\log(1/\varepsilon(r)))\\ &\geq (\tau/2)d_2.\end{aligned}$$

Therefore, by Theorem 3.9 there exists $i_2^* \in [\Delta]$ such that

$$\mathbf{H}_\infty^{3\varepsilon(r)}\left(\mathsf{Cond}_{i_2^*}^2(X_2) \mid Z^{(i_1^*,j_1^*)}, (Z')^{(i_1^*,j_1^*)}, \mathcal{H}_1\right) \geq (6/7+\tau)d_2' + \Omega(\log(1/\varepsilon(r))). \tag{4.14}$$

By a similar argument, there exists $j_2^* \in [\Delta]$ such that

$$\mathbf{H}_\infty^{3\varepsilon(r)}\left(\mathsf{Cond}_{j_2^*}^2(Y_2) \mid \mathcal{H}_1\right) \geq (6/7+\tau)d_2' + \Omega(\log(1/\varepsilon(r))). \tag{4.15}$$

Recall that

$$T_{(i_1^*,j_1^*,i_2^*,j_2^*)} = \mathsf{IPM}_{\sqrt{r}}\left(Z^{(i_1^*,j_1^*)}, \mathsf{Cond}_{j_2^*}^2(Y_2), \mathsf{Cond}_{i_2^*}^2(X_2)\right).$$

By (4.13),(4.14),(4.15) and since $Z^{(i_1^*,j_1^*)}$, $(Z')^{(i_1^*,j_1^*)}$ are jointly independent of ($\mathsf{Cond}_{j_2^*}^2(Y_2)$, $\mathsf{Cond}_{j_2^*}^2(Y_2')$) when conditioned on $\mathcal{H}_1$, we have that

$$\begin{aligned}&\left(T_{(i_1^*,j_1^*,i_2^*,j_2^*)}, (T')_{(i_1^*,j_1^*,i_2^*,j_2^*)}, Z^{(i_1^*,j_1^*)}, (Z')^{(i_1^*,j_1^*)}, \mathsf{Cond}_{i_2^*}^2(X_2), \mathsf{Cond}_{i_2^*}^2(X_2'), \mathcal{H}_1\right) \approx_{O(\varepsilon(r))}\\ &\left(U, (T')_{(i_1^*,j_1^*,i_2^*,j_2^*)}, Z^{(i_1^*,j_1^*)}, (Z')^{(i_1^*,j_1^*)}, \mathsf{Cond}_{i_2^*}^2(X_2), \mathsf{Cond}_{i_2^*}^2(X_2'), \mathcal{H}_1\right).\end{aligned}$$

As $T_{(i_1^*,j_1^*,i_2^*,j_2^*)}$ is a deterministic function of $\mathsf{Cond}_{j_2^*}^2(Y_2)$ when conditioned on $Z^{(i_1^*,j_1^*)}$, $\mathsf{Cond}_{i_2^*}^2(X_2)$, we may adjoin X_2, X_2' and disregard the excess random variables to obtain

$$\left(T_{(i_1^*,j_1^*,i_2^*,j_2^*)}, (T')_{(i_1^*,j_1^*,i_2^*,j_2^*)}, \mathcal{H}_2\right) \approx_{O(\varepsilon(r))} \left(U, (T')_{(i_1^*,j_1^*,i_2^*,j_2^*)}, \mathcal{H}_2\right), \tag{4.16}$$

where $\mathcal{H}_2 = X_2, X_2', \mathcal{H}_1$. That is, T is $(O(\varepsilon(r)), \mathcal{H}_2)$-somewhere independent of T'. Further, for any $i_1, j_1, i_2, j_2 \in [\Delta]^4$, the random variables $T_{(i_1,j_1,i_2,j_2)}$, $T'_{(i_1,j_1,i_2,j_2)}$ are deterministic functions of Y_2, Y_2', respectively, when conditioned on $\mathcal{H}_2$, and are therefore independent of $\mathcal{X}$. This, together with Corollary 4.7, implies that

$$\begin{aligned}\mathbf{H}_\infty^{2\varepsilon(r)}(Y \mid T, T', \mathcal{H}_2) &\geq \mathbf{H}_\infty^{2\varepsilon(r)}(Y \mid Y_2, Y_2', \mathcal{H}_2)\\ &= \mathbf{H}_\infty^{2\varepsilon(r)}(Y \mid Y_2, Y_2', \mathcal{H}_1)\\ &= \mathbf{H}_\infty^{2\varepsilon(r)}(Y \mid Y_2, Y_2', \mathcal{M}, \mathcal{H})\\ &\geq \tau d - O(\log(1/\varepsilon(r)))\\ &\geq \tau d/2.\end{aligned} \tag{4.17}$$

Similarly,

$$\begin{aligned}\mathbf{H}_{\infty}^{2\varepsilon(r)}(X \mid \mathcal{H}_2) &= \mathbf{H}_{\infty}^{2\varepsilon(r)}(X \mid X_2, X_2', \mathcal{H}_1) \\ &= \mathbf{H}_{\infty}^{2\varepsilon(r)}(X \mid X_2, X_2', \mathcal{H}) \\ &\geq \tau d/2. \qquad (4.18)\end{aligned}$$

Recall that $\mathsf{IPM}_r(M, X, Y) = \mathsf{BaseIPM}(T, X, Y)$. By Equations (4.16),(4.17),(4.18) we conclude

$$(\mathsf{IPM}_r(M, X, Y), \mathsf{IPM}_r(M', X', Y'), T, T', \mathcal{Y}, \mathcal{H}_2) \approx_{O(\varepsilon(r))} (U, \mathsf{IPM}_r(M', X', Y'), T, T', \mathcal{Y}, \mathcal{H}_2)$$

which readily implies that

$$(\mathsf{IPM}_r(M, X, Y), \mathsf{IPM}_r(M', X', Y'), \mathcal{M}, \mathcal{Y}, \mathcal{H}) \approx_{O(\varepsilon(r))} (U, \mathsf{IPM}_r(M', X', Y'), \mathcal{M}, \mathcal{Y}, \mathcal{H})\,.$$

As the for parameters. The construction forces the recursive relation $d(r) = c \cdot d(\sqrt{r})$ for some constant $c = c(\tau)$. This solves for $d(r) = d(\Delta^4) \cdot \mathrm{polylog}(r)$, where we set the base of the recursion at $r = \Delta^4$ rows. To make sure that the applications of $\mathsf{BaseIPM}$ are all valid, we must meet the hypothesis of Lemma 4.4 which forces $\ell = \Omega(\log(d(r)/\varepsilon))$ (as we fix ℓ throughout the $O(\log\log r)$ steps of the recursion, and $d(r)$ increases with r) and $k = \Omega(\ell)$, where k is the min-entropy of the two sources, which in our applications are proportional to the lengths of these sources as τ is constant. As these lengths increase with r, it is enough to require $d(\Delta^4) = \Omega(\ell)$. Even after taking into account the deterioration of the error parameter throughout the recursion, all of the required conditions are met by the hypothesis of the theorem, namely, $d(r) = \ell \cdot \mathrm{polylog}(r)$ and $\ell = \Omega(\log(\log(r)/\varepsilon))$. □

5 Improved CBA via IPM

In this section we construct an improved CBA based on the IPM that was developed in the previous section. Our construction follows a similar construction from [Coh16a]. There are some technical differences between the two works and so we cannot rely on [Coh16a] and are required to give a complete proof. This is the content of the following theorem.

Theorem 5.1. *There exist universal constants $c_{\mathsf{ACB}} > 1 > \gamma_0 > 0$ such that for any $0 < \gamma \leq \gamma_0$ the following holds. For any integers n, a and for any $\varepsilon > 0$ that satisfy*

$$n = \Omega((\log a)^{c_{\mathsf{ACB}}} \cdot \log(1/\varepsilon)),$$

there exists an explicit $(1-\gamma, \varepsilon)$-CBA

$$\mathsf{CBA} \colon \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^a \to \{0,1\}^m$$

with $m = (1/2 - O(\gamma))n$.

Proof. Let $x, y \in \{0,1\}^n$ and $\alpha \in \{0,1\}^a$. Defining $\mathsf{CBA}(x,y,\alpha)$ will require some preparations, namely, introducing some notations and building blocks that we use. Let $c_{\mathsf{GUV}}, c_{\mathsf{IPM}}$ be the constants from Theorem 3.5 and Theorem 4.3, respectively. Let c_1 be a constant to be set later on. Set

$$\begin{aligned} n_1 &= c_1 \cdot (\log(n/\varepsilon) + \log\log a), \\ n_2 &= 2\gamma n, \\ n_3 &= 40\gamma n. \end{aligned}$$

By the hypothesis of the theorem, and by taking $\gamma_0 < 40$, we have that $n_1 < n_2 < n_3 < n$. For $i = 1,2,3$, let x_i (resp. y_i) be the length n_i prefix of x (resp. y).

Building blocks. For the construction of CBA we make use of the following building blocks:

- Let $\mathsf{a}\colon \{0,1\}^{n_1} \times \{0,1\}^{n_2} \to \{0,1\}^{n_1}$ and $\mathsf{b}\colon \{0,1\}^{n_1} \times \{0,1\}^{n_2} \times \{0,1\}^{n_2} \to \{0,1\}^{n_1}$ be the pair of functions that are given in Section 3.3, set with error guarantee ε. Note that by taking $c_1 \geq c_{\mathsf{GUV}}$, the parameter n_1 was chosen large enough as $n \geq n_2$.
- Let $\mathsf{IPM}\colon \{0,1\}^{(2a)\times n_1} \times \{0,1\}^{n_3} \times \{0,1\}^{n_3} \to \{0,1\}^{n_1}$ be the $(0.86, \varepsilon)$-IPM that is given by Theorem 4.3. This instantiation of Theorem 4.3 is valid when taking c, c_1 large enough, as:
  - $n_1 \geq c_1 \cdot \log(\log(a)/\varepsilon)$,
  - $n_3 = \Omega(n) = \Omega((\log a)^{c_{\mathsf{IPM}}} \cdot \log(1/\varepsilon))$,
  - $0.86 > 6/7$.
- Set $m = (1-82\gamma)n/2$. Let $\mathsf{Ext}\colon \{0,1\}^n \times \{0,1\}^{n_1} \to \{0,1\}^m$ be the $((1+\gamma)m, \varepsilon)$-strong seeded extractor that is given by Theorem 3.5. Note that by taking c_1 to be a large enough constant (as a function of the constant γ), the parameter n_1 is sufficiently large as required by Theorem 3.5.

The construction. We start by defining a $(2a) \times n_1$ matrix $m = m(x_2, y_2, \alpha)$ as follows. For $i \in [2a]$, row i of m is defined by

$$m_i = \begin{cases} \mathsf{a}(y_1, x_2), & i \neq \alpha_{\lceil i/2 \rceil} \pmod 2; \\ \mathsf{b}(y_1, y_2, x_2), & i = \alpha_{\lceil i/2 \rceil} \pmod 2. \end{cases}$$

We then define

$$s = \mathsf{IPM}(m, y_3, x_3),$$

and finally define

$$\mathsf{CBA}(x,y,\alpha) = \mathsf{Ext}(x,s).$$

Analysis. We now turn to the analysis. Let $\mathcal{X} = (X, X')$ be a pair of n-bit random variables, $\mathcal{Y} = (Y, Y')$ a pair of n-bit random variables, and let $\alpha, \alpha' \in \{0,1\}^a$. Let $\mathcal{H}$ be a random variable for which the following holds:

- Conditioned on $\mathcal{H}$, the random variables $\mathcal{X}, \mathcal{Y}$ are independent;
- α, α' are fixed distinct strings when conditioned on $\mathcal{H}$;
- $\mathbf{H}_\infty^\varepsilon(X \mid \mathcal{H}) \geq (1-\gamma)n + \Omega(\log(1/\varepsilon))$;
- $(Y, \mathcal{H}) \approx_\varepsilon (U, \mathcal{H})$.

To conclude the proof, we are required to show that

$$(\mathsf{CBA}(X, Y, \alpha), \mathsf{CBA}(X', Y', \alpha'), \mathcal{Y}, \mathcal{H}) \approx_{O(\varepsilon)} (U, \mathsf{CBA}(X', Y', \alpha'), \mathcal{Y}, \mathcal{H}).$$

Define $M = m(X_2, Y_2, \alpha)$ and $M' = m(X_2', Y_2', \alpha')$. We begin by showing that M is somewhere-independent of M'. More precisely, we establish the following claim.

Claim 5.2. *M is $(O(\varepsilon), \mathcal{H}_1)$-somewhere independent of M', where $\mathcal{H}_1 = Y_2', Y_2, \mathcal{H}$.*

Proof of Claim 5.2. Recall that $\alpha \neq \alpha'$ when conditioned on $\mathcal{H}$. Let $i = i(\mathcal{H}) \in [a]$ be such that $\alpha_i \neq \alpha_i'$, and set $g = 2i - \alpha_i$. Note that, by construction, $M_g = \mathsf{b}(Y_1, Y_2, X_2)$ whereas $M_g' = \mathsf{a}(Y_1', X_2')$. We can therefore apply Lemma 3.18 with $\mathcal{W} = (X_2, X_2')$, $\mathcal{Y} = (Y_1, Y_1')$, $\mathcal{Z} = (Y_2, Y_2')$, and $\mathcal{H}$, to conclude that

$$\left(M_g, M_g', Y_2, Y_2', \mathcal{H}\right) \approx_{O(\varepsilon)} \left(U, M_g', Y_2, Y_2', \mathcal{H}\right).$$

To justify this application of Lemma 3.18 we note that

- Conditioned on $\mathcal{H}$, the random variables X_2, X_2' are jointly independent of (Y_2, Y_2'), which also include Y_1, Y_1' as their respective prefixes;
- $(Y_1, \mathcal{H}) \approx_\varepsilon (U, \mathcal{H})$;
- $|Y_2| = n_2 \geq 4n_1 + \Omega(\log(1/\varepsilon))$ and $(Y_2, \mathcal{H}) \approx_\varepsilon (U, \mathcal{H})$, and so $\mathbf{H}_\infty^\varepsilon(Y_2 \mid \mathcal{H}) \geq 4n_1 + \Omega(\log(1/\varepsilon))$; and
- $\mathbf{H}_\infty^{2\varepsilon}(X_2 \mid \mathcal{H}) \geq 4n_1 + \Omega(\log(1/\varepsilon))$. To see this, set $X_{>2}$ to be the length $n - n_2$ suffix of X, and observe that

$$\begin{aligned}\mathbf{H}_\infty^{2\varepsilon}(X_2 \mid \mathcal{H}) &\geq \mathbf{H}_\infty^{2\varepsilon}(X \mid X_{>2}, \mathcal{H})\\ &\geq \mathbf{H}_\infty^{\varepsilon}(X \mid \mathcal{H}) - |X_{>2}| - O(\log(1/\varepsilon))\\ &\geq (1-\gamma)n - (n - n_2) - O(\log(1/\varepsilon))\\ &= \gamma n - O(\log(1/\varepsilon))\\ &\geq 4n_1 + \Omega(\log(1/\varepsilon)).\end{aligned}$$

This concludes the proof of the claim. □

Returning to the proof of Theorem 5.1, our next step is to show that

$$(\mathsf{IPM}(M,Y_3,X_3),\mathsf{IPM}(M',Y_3',X_3'),\mathcal{H}_2)\approx_{O(\varepsilon)}(U,\mathsf{IPM}(M',Y_3',X_3'),\mathcal{H}_2)\,, \tag{5.1}$$

where $\mathcal{H}_2 = M, M', X_3, X_3', \mathcal{H}_1$. To this end we prove the following claim which states that all the assumptions required by the application of IPM in the above equation are met.

Claim 5.3. *The following holds:*

- *Conditioned on $\mathcal{H}_1$, the random variables Y_3, Y_3' are jointly independent of X_3, X_3', M, M';*
- *M is $(O(\varepsilon),\mathcal{H}_1)$-somewhere independent of M';*
- $\mathbf{H}_\infty^{O(\varepsilon)}(Y_3 \mid \mathcal{H}_1) \geq 0.86n_3 + \Omega(\log(1/\varepsilon))$*;*
- $\mathbf{H}_\infty^{O(\varepsilon)}(X_3 \mid M, M', \mathcal{H}_1) \geq 0.86n_3 + \Omega(\log(1/\varepsilon))$*.*

Proof of Claim 5.3. Recall that $M = m(X_2, Y_2, \alpha)$, $M' = m(X_2', Y_2', \alpha')$ are deterministic functions of X_2, X_2', Y_2, Y_2'. Since $\mathcal{H}_1 = Y_2', Y_2, \mathcal{H}$, conditioned on $\mathcal{H}_1$, the random variables M, M' are deterministic functions of X_2, X_2', and therefore also of X_3, X_3', that are jointly independent of (Y_3, Y_3') when conditioned on $\mathcal{H}_1$. This proves the first item. The second item is the content of Claim 5.2.

As for the third item, let $Y_{>3}$ be the length $n - n_3$ suffix of Y. By Lemma 3.2, and by our choice of parameters,

$$\begin{aligned}
\mathbf{H}_\infty^{O(\varepsilon)}(Y_3 \mid \mathcal{H}_1) &\geq \mathbf{H}_\infty^{O(\varepsilon)}(Y \mid Y_{>3}, \mathcal{H}_1)\\
&= \mathbf{H}_\infty^{O(\varepsilon)}(Y \mid Y_{>3}, Y_2, Y_2', \mathcal{H})\\
&\geq \mathbf{H}_\infty^{\varepsilon}(Y \mid \mathcal{H}) - |Y_{>3}| - |Y_2| - |Y_2'| - O(\log(1/\varepsilon))\\
&\geq n - (n - n_3) - 2n_2 - O(\log(1/\varepsilon))\\
&= n_3 - 4\gamma n - O(\log(1/\varepsilon))\\
&\geq 0.9n_3 - O(\log(1/\varepsilon))\\
&\geq 0.86n_3 + \Omega(\log(1/\varepsilon)).
\end{aligned}$$

For the forth item, recall that M, M' are deterministic functions of X_2, X_2' when conditioned on $\mathcal{H}_1$, and so

$$\begin{aligned}
\mathbf{H}_\infty^{O(\varepsilon)}(X_3 \mid M, M', \mathcal{H}_1) &\geq \mathbf{H}_\infty^{O(\varepsilon)}(X_3 \mid X_2, X_2', \mathcal{H}_1)\\
&\geq \mathbf{H}_\infty^{O(\varepsilon)}(X \mid X_{>3}, X_2, X_2', \mathcal{H}_1)\\
&\geq \mathbf{H}_\infty^{\varepsilon}(X \mid \mathcal{H}_1) - |X_{>3}| - |X_2| - |X_2'| - O(\log(1/\varepsilon))\\
&= \mathbf{H}_\infty^{\varepsilon}(X \mid \mathcal{H}) - |X_{>3}| - |X_2| - |X_2'| - O(\log(1/\varepsilon))\\
&\geq (1-\gamma)n - (n - n_3) - 2n_2 - O(\log(1/\varepsilon))\\
&\geq n_3 - 5\gamma n - O(\log(1/\varepsilon))\\
&\geq (7/8)n_3 - O(\log(1/\varepsilon))\\
&\geq 0.86n_3 + \Omega(\log(1/\varepsilon)).
\end{aligned}$$

This concludes the proof of the claim. □

By Claim 5.3 we can apply Theorem 4.3 and conclude (5.1), that is,

$$(S, S', \mathcal{H}_2) \approx_{O(\varepsilon)} (U, S', \mathcal{H}_2)\,.$$

Conditioned on $S', \mathcal{H}_2$, the random variable $S = \mathsf{IPM}(M, Y_3, X_3)$ is a deterministic function of Y_3 whereas $\mathsf{Ext}(X', S')$ is a deterministic function of X', which is independent of Y_3. Thus, we may adjoin $\mathsf{Ext}(X', S')$ to the above equation and conclude that

$$(S, \mathsf{Ext}(X', S'), S', \mathcal{H}_2) \approx_{O(\varepsilon)} (U, \mathsf{Ext}(X', S'), S', \mathcal{H}_2)\,. \tag{5.2}$$

As X is independent of S' when conditioned on $\mathcal{H}_2$, and since M, M' are deterministic functions of X_2, X'_2, we have that

$$\begin{aligned}
\mathbf{H}_{\infty}^{O(\varepsilon)}(X \mid \mathsf{Ext}(X', S'), S', \mathcal{H}_2) &= \mathbf{H}_{\infty}^{O(\varepsilon)}(X \mid \mathsf{Ext}(X', S'), X_3, X'_3, \mathcal{H}) \\
&\geq \mathbf{H}_{\infty}^{\varepsilon}(X \mid \mathcal{H}) - |\mathsf{Ext}(X', S')| - |X_3| - |X'_3| - O(\log(1/\varepsilon)) \\
&\geq (1-\gamma)n - m - 2n_3 - O(\log(1/\varepsilon)) \\
&\geq (1-81\gamma)n - m - O(\log(1/\varepsilon)) \\
&\geq (1+\gamma)m + \Omega(\log(1/\varepsilon)),
\end{aligned} \tag{5.3}$$

where the last inequality follows as

$$(2+\gamma)m = (2+\gamma)\left(\frac{1-82\gamma}{2}\right)n < (1-81\gamma)n.$$

By equations (5.2),(5.3), and by the fact that X is independent of S when conditioned on $\mathsf{Ext}(X', S')$, $S', \mathcal{H}_2$, we have that

$$(\mathsf{Ext}(X, S), \mathsf{Ext}(X', S'), S, S', \mathcal{H}_2) \approx_{O(\varepsilon)} (U, \mathsf{Ext}(X', S'), S, S', \mathcal{H}_2)\,.$$

Recall that $\mathsf{CBA}(X, Y, \alpha) = \mathsf{Ext}(X, S)$ and $\mathsf{CBA}(X', Y', \alpha') = \mathsf{Ext}(X', S')$. Conditioned on $\mathsf{Ext}(X', S'), S, S', \mathcal{H}_2$, the random variable $\mathsf{Ext}(X, S)$ is independent of $\mathcal{Y}$ and so we may adjoin $\mathcal{Y}$ to the above equation and remove the excess random variables to obtain

$$(\mathsf{CBA}\left(X, Y, \alpha\right), \mathsf{CBA}\left(X', Y', \alpha'\right), \mathcal{Y}, \mathcal{H}) \approx_{O(\varepsilon)} (U, \mathsf{CBA}\left(X', Y', \alpha'\right), \mathcal{Y}, \mathcal{H})\,,$$

which concludes the proof. □

6 Non-Malleable Extractors via CBA

In this section we prove Theorem 1.9. As in [CGL16, Coh16b], our construction of non-malleable extractors relies on CBA. Besides using the improved CBA that we constructed in Theorem 5.1, we also make some improvements to the reduction itself. In particular, we show how to generate a shorter advice string.

Proof of Theorem 1.9. Let $c_{\mathsf{GUV}}, c_{\mathsf{Raz}}$ be the constants that are given by Theorem 3.5 and Theorem 3.8, respectively, and let γ_0 be the constant that is given by Theorem 5.1. Set

$$\begin{aligned} d_1 &= c_{\mathsf{GUV}} \cdot \log(n/\varepsilon), \\ d_2 &= \max\left(10 d_1, c_{\mathsf{Raz}} \cdot \log(n/\varepsilon)\right). \end{aligned}$$

For a d-bit string y, let y_1 denote the length d_1 prefix of y. Similarly, let y_2 denote the length d_2 prefix of y. We further assume that $d \geq (3/\gamma_0) \cdot d_2$. Note that this assumption is met by taking the hidden constant under the $O(\cdot)$ notation in the seed length d large enough with respect to the constants $c_{\mathsf{GUV}}, c_{\mathsf{Raz}}$.

Building blocks. For the construction of nmExt we make use of the following building blocks.

- Let q be the least even prime power of 2 that is larger or equal than $5/\varepsilon^2$. Note that $q \leq 20/\varepsilon^2$. Let r be the least integer such that $q^r \geq d$. We identify $[d]$ with an arbitrary subset of $\mathbb{F}_q^r$. Set $v = 2r/\varepsilon$ and let $\mathsf{ECC}\colon \mathbb{F}_q^r \to \mathbb{F}_q^v$ be the error correcting code that is given by Theorem 3.13, set with relative distance $1 - \varepsilon$. Theorem 3.13 gives an explicit code with these parameters.

- Let $\mathsf{Ext}_{\mathsf{AG}}\colon \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{\log v}$ be the $(2\log v, \varepsilon)$-strong seeded extractor that is given by Theorem 3.5. Note that d_1 was defined to be large enough so as to be used as a seed for $\mathsf{Ext}_{\mathsf{AG}}$. We identify the output of $\mathsf{Ext}_{\mathsf{AG}}$ as an element of $[v]$.

- Let $\mathsf{Raz}\colon \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^d$ be the $(2d, \varepsilon)$-extractor with weak-seeds that is given by Theorem 3.8. Note that d_2 was chosen large enough as required by Theorem 3.8.

- Set $a = \log(qv)$. Let $\mathsf{CBA}\colon \{0,1\}^d \times \{0,1\}^d \times \{0,1\}^a \to \{0,1\}^{d_1}$ be the $(1-\gamma_0, \varepsilon)$-CBA that is given by Theorem 5.1. By Theorem 5.1, the output length of CBA is $(1/2 - O(\gamma_0))d$, which is larger than d_1. Thus, we may truncate the output length to d_1 bits. Moreover, by the hypothesis of the theorem, the requirement $d = \Omega((\log a)^{c_{\mathsf{ACB}}} \cdot \log(1/\varepsilon))$ of Theorem 5.1 is met. Indeed,

$$\begin{aligned} (\log a)^{c_{\mathsf{ACB}}} \cdot \log(1/\varepsilon) &= (\log\log(qv))^{c_{\mathsf{ACB}}} \cdot \log(1/\varepsilon) \\ &\leq \left(\log\log\left(\frac{\log d}{\varepsilon}\right)\right)^{c_{\mathsf{ACB}}} \cdot \log(1/\varepsilon) \\ &= O(d). \end{aligned}$$

- Let $\mathsf{Ext}_{\mathsf{out}}\colon \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{(1/2-\alpha)k}$ be the $(k/2, \varepsilon)$-strong seeded extractor that is given by Theorem 3.5. Note that d_1 is large enough as required by Theorem 3.5 when taking the constant c_{GUV} large enough (as a function of the constant α).

The construction. On input $x \in \{0,1\}^n$, $y \in \{0,1\}^d$, we define $\mathsf{nmExt}(x,y)$ as follows. First we compute

$$i = i(x, y_1) = \mathsf{Ext}_{\mathsf{AG}}(x, y_1),$$

and define

$$\mathsf{AdvGen}(x,y) = \mathsf{ECC}(y)_i \circ i.$$

In the expression above, by $\mathsf{ECC}(y)_i$ we mean the following – we interpret $i \in \{0,1\}^{\log v}$ as an index in $[v]$ of the codeword $\mathsf{ECC}(y)$. Then, $\mathsf{ECC}(y)_i$ refers to the content in that i'th entry, when interpreted as a $(\log q)$-bit string. Define

$$z = \mathsf{CBA}\left(y, \mathsf{Raz}(x, y_2), \mathsf{AdvGen}(x,y)\right).$$

Finally, we define

$$\mathsf{nmExt}(x,y) = \mathsf{Ext}_{\mathsf{out}}(x,z).$$

Analysis. Let X be an (n,k)-source, Y a random variable that is uniformly distributed over d-bit strings, independently of X, and let $\mathcal{A}\colon \{0,1\}^d \to \{0,1\}^d$ be a function with no fixed points. Denote $Y' = \mathcal{A}(Y)$. We start by proving the following claim.

Claim 6.1. *Let C, C' be a pair of arbitrarily correlated random variables over n-bit strings such that the relative Hamming distance between C, C' is at least $1-\varepsilon_1$ (with probability 1). Let I, I' be a pair of arbitrarily correlated random variables over $[n]$ that are jointly independent of (C, C'). Assume that $I \sim_{\varepsilon_2} U$. Then,*

$$\mathbf{Pr}\left[C_I \circ I = C'_{I'} \circ I'\right] \le \varepsilon_1 + \varepsilon_2,$$

where C_I (resp. $C'_{I'}$) denotes the I'th entry of C (resp. (I')'th entry of C').

Proof. For $i \in \mathsf{supp}(I)$, let I'_i denote the random variable $I' \mid (I = i)$. Using the assumption that C, C' are jointly independent of I,

$$\mathbf{Pr}\left[C_I \circ I = C'_{I'} \circ I'\right] = \sum_{i \in \mathsf{supp}(I)} \mathbf{Pr}[I = i] \cdot \mathbf{Pr}\left[C_i \circ i = C'_{I'_i} \circ I'_i\right]. \tag{6.1}$$

Observe that for any $i \in \mathsf{supp}(I)$,

$$\begin{aligned}\mathbf{Pr}\left[C_i \circ i = C'_{I'_i} \circ I'_i\right] &\le \mathbf{Pr}\left[C_i \circ i = C'_{I'_i} \circ I'_i \mid I'_i = i\right] \\ &= \mathbf{Pr}\left[C_i = C'_i\right],\end{aligned}$$

where we have used the independence between (C, C') and (I, I'). Let J be a random variable that is uniformly distributed over $[n]$. By plugging the above equation back to

Equation (6.1), and using again the independence of I from (C, C'), we conclude that

$$\begin{aligned}\mathbf{Pr}\left[C_I \circ I = C'_{I'} \circ I'\right] &\leq \sum_{i \in \mathsf{supp}(I)} \mathbf{Pr}[I = i] \cdot \mathbf{Pr}\left[C_i = C'_i\right] \\ &= \sum_{i \in \mathsf{supp}(I)} \mathbf{Pr}[I = i] \cdot \mathbf{Pr}\left[C_i = C'_i \mid I = i\right] \\ &= \mathbf{Pr}\left[C_I = C'_I\right] \\ &\leq \mathbf{Pr}\left[C_J = C'_J\right] + \mathsf{SD}(I, J) \\ &\leq \varepsilon_1 + \varepsilon_2.\end{aligned}$$

□

Returning back to the proof of Theorem 1.9, we prove the following claim.

Claim 6.2.

$$\mathbf{Pr}_{(x,y) \sim (X,Y)}\left[\mathsf{AdvGen}(x, y) = \mathsf{AdvGen}(x, y')\right] = O(\sqrt{\varepsilon}).$$

Proof. Recall that $I = i(X, Y_1) = \mathsf{Ext}_{\mathsf{AG}}(X, Y_1)$ and $\mathsf{AdvGen}(X, Y) = \mathsf{ECC}(Y)_I \circ I$. As $\mathsf{Ext}_{\mathsf{AG}}$ is a (k, ε)-strong seeded extractor, $(I, Y_1) \approx_\varepsilon (U, Y_1)$. Conditioned on any fixing of Y_1, the random variables I, Y'_1 are independent and so we may adjoin Y'_1 to the latter equation and conclude that

$$(I, Y'_1, Y_1) \approx_\varepsilon (U, Y'_1, Y_1).$$

Therefore, by Markov's inequality, except with probability $\sqrt{\varepsilon}$ over $(y_1, y'_1) \sim (Y_1, Y'_1)$, it holds that $I \approx_{\sqrt{\varepsilon}} U$. By aggregating an error of $\sqrt{\varepsilon}$ to the total error, we condition on the event $(Y_1, Y'_1) = (y_1, y'_1)$ for which $I \approx_{\sqrt{\varepsilon}} U$ holds. Observe that for any fixing of (Y_1, Y'_1) to (y_1, y'_1), the random variables I, I' are jointly independent of $(\mathsf{ECC}(Y), \mathsf{ECC}(Y'))$. This, together with the fact that ECC has relative Hamming distance $1 - \varepsilon$, allows us to apply Claim 6.1, which readily concludes the proof of the claim. □

By Lemma 3.2, by our choice of parameters, and as ECC has alphabet size q,

$$\mathbf{H}^\varepsilon_\infty\left(Y_2 \mid \mathsf{AdvGen}(X, Y), \mathsf{AdvGen}(X, Y')\right) \geq d_2 - 2(d_1 + \log q) \geq 0.6 d_2.$$

Further,

$$\begin{aligned}\mathbf{H}^\varepsilon_\infty\left(X \mid \mathsf{AdvGen}(X, Y), \mathsf{AdvGen}(X, Y')\right) &\geq k - 2 \log v \\ &\geq \max\left(2d, c'_{\mathsf{Raz}} d_2\right) + \Omega(\log(1/\varepsilon)).\end{aligned}$$

Note that one can condition on $\mathsf{AdvGen}(X, Y)$, $\mathsf{AdvGen}(X, Y')$ while maintaining the independence between X and Y. Indeed, after conditioning on Y_1, Y'_1, the random variables $\mathsf{Ext}_{\mathsf{AG}}(X, Y_1)$, $\mathsf{Ext}_{\mathsf{AG}}(X, Y'_1)$ are deterministic functions of X, and so one can further condition on these random variables without introducing dependencies between X and Y. Conditioned on Y_1, Y'_1, $\mathsf{Ext}_{\mathsf{AG}}(X, Y_1)$, $\mathsf{Ext}_{\mathsf{AG}}(X, Y'_1)$, the random variables $\mathsf{AdvGen}(X, Y)$, $\mathsf{AdvGen}(X, Y')$ are deterministic functions of Y, and so conditioning on these variables does not introduce any dependencies between X, Y. By the above, we can apply Theorem 3.8 and conclude that

$$(\mathsf{Raz}(X, Y_2), Y_2, \mathsf{AdvGen}(X, Y), \mathsf{AdvGen}(X, Y')) \approx_{O(\varepsilon)} (U, Y_2, \mathsf{AdvGen}(X, Y), \mathsf{AdvGen}(X, Y'))\,.$$

As $\mathsf{Raz}(X, Y_2)$ is independent of Y_2' when conditioned on Y_2, $\mathsf{AdvGen}(X, Y)$, $\mathsf{AdvGen}(X, Y')$, we have that

$$(\mathsf{Raz}(X, Y_2), \mathcal{H}) \approx_{O(\varepsilon)} (U, \mathcal{H})\,, \tag{6.2}$$

where $\mathcal{H} = Y_2', Y_2, \mathsf{AdvGen}(X, Y), \mathsf{AdvGen}(X, Y')$.

Recall that

$$Z = \mathsf{CBA}\left(Y, \mathsf{Raz}(X, Y_2), \mathsf{AdvGen}(X, Y)\right).$$

By (6.2), the second argument to CBA is close to uniform, as required, when conditioned on $\mathcal{H}$. We now consider the first argument. By Lemma 3.2,

$$\begin{aligned}\mathbf{H}_\infty^\varepsilon\left(Y \mid \mathcal{H}\right) &\geq d - 2(d_2 + \log q) - O(\log(1/\varepsilon)) \\ &\geq (1-\gamma_0)d + \Omega(\log(1/\varepsilon)),\end{aligned} \tag{6.3}$$

where we have used the fact that $d \geq (3/\gamma_0) \cdot d_2$ and that $d_2 = \Omega(\log(1/\varepsilon))$.

By Equations (6.2),(6.3), we can apply Theorem 5.1 to conclude that

$$(Z, Z', \mathcal{H}') \approx_{O(\sqrt{\varepsilon})} (U, Z', \mathcal{H}')\,,$$

where $\mathcal{H}' = \mathsf{Raz}(X, Y_2), \mathsf{Raz}(X, Y_2'), \mathcal{H}$. Note that conditioned on $Z', \mathcal{H}'$, the random variables Z and $\mathsf{Ext}_{\mathsf{out}}(X, Z')$ are independent. Thus, we may adjoin $\mathsf{Ext}_{\mathsf{out}}(X, Z')$ to the above equation and conclude that

$$(Z, \mathsf{Ext}_{\mathsf{out}}(X, Z'), Z', \mathcal{H}') \approx_{O(\sqrt{\varepsilon})} (U, \mathsf{Ext}_{\mathsf{out}}(X, Z'), Z', \mathcal{H}')\,.$$

By Lemma 3.2 and since $\mathsf{Ext}_{\mathsf{out}}$ is set to have $(1/2-\alpha)k$ output bits,

$$\begin{aligned}\mathbf{H}_\infty^{O(\varepsilon)}(X \mid \mathsf{Ext}_{\mathsf{out}}(X, Z'), Z', \mathcal{H}') &\geq k - (1/2-\alpha)k - 2(d + \log v) - O(\log(1/\varepsilon)) \\ &= (1/2+\alpha)k - 2(d + \log v) - O(\log(1/\varepsilon)) \\ &\geq k/2 + \Omega(\log(1/\varepsilon)).\end{aligned}$$

Therefore,

$$(\mathsf{Ext}_{\mathsf{out}}(X, Z), Z, \mathsf{Ext}_{\mathsf{out}}(X, Z'), Z', \mathcal{H}') \approx_{O(\sqrt{\varepsilon})} (U, Z, \mathsf{Ext}_{\mathsf{out}}(X, Z'), Z', \mathcal{H}')\,.$$

By the definition of nmExt and since conditioned on Z, $\mathsf{Ext}_{\mathsf{out}}(X, Z')$, $Z', \mathcal{H}'$, the random variable $\mathsf{Ext}_{\mathsf{out}}(X, Z)$ is independent of Y, we may adjoin Y to the above equation and remove the excess random variables to conclude

$$(\mathsf{nmExt}(X, Y), \mathsf{nmExt}(X, Y'), Y) \approx_{O(\sqrt{\varepsilon})} (U, \mathsf{nmExt}(X, Y'), Y)\,,$$

as desired. □

7 Semi-Explicit Four-Source Extractors

In this section we prove Theorem 1.7. Our construction of the semi-explicit multi-source extractors is divided into several step. In Section 7.1, we construct semi-explicit seeded extractors with seed length and output length that are better than what is known for strongly explicit constructions (see Proposition 7.1). These savings are crucial for us. Building on our seeded extractor, in Section 7.2, we construct a semi-explicit two-source condenser. For any desired constant $\alpha > 0$, this algorithm converts two independent $(n, \alpha \log_2 n)$ sources to an $(n^{1-\alpha}, \alpha \log_2 n - O(1))$ source (see Proposition 7.3). Building on these, in Section 7.3, we can finally prove Theorem 1.7.

7.1 Seeded Extractors with Short Seeds and Optimal Entropy Loss

Proposition 7.1. *For any integer n and constants $0 < \alpha, \varepsilon < 1$, there exists an $\exp(n^\alpha \cdot \log^3 n)$-time (k, ε)-strong seeded extractor $\mathsf{Ext}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$, where*

$$\begin{aligned} k &= \alpha \log_2 n, \\ m &= k - O(1), \\ d &= \log_2 n + 3 \log_2 \log_2 n + O(1). \end{aligned}$$

For the proof of Proposition 7.1 we prove the following lemma.

Lemma 7.2. *For all integers n, k and constant $\varepsilon > 0$ such that $k \geq 2\log_2(1/\varepsilon) + \log_2 \log_2(1/\varepsilon) + 1$, there exists an $\exp(n \cdot 2^k k)$-time computable (k, ε)-strong seeded extractor $\mathsf{Ext}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with seed length $d = \log_2 n + O(1)$ and $m = k - O(1)$ output bits.*

Proof of Lemma 7.2. Let c be a constant to be chosen later on (as a function of the constant error guarantee ε). Set

$$\begin{aligned} m &= k - c, \\ d &= \log_2 n + c, \\ N &= 2^{n+d} \cdot m, \\ K &= 2^{k+d} \cdot m, \\ \delta &= 2^{-\varepsilon^3 \cdot 2^{k+d}}. \end{aligned}$$

Let Z be the (N, K, δ)-independent random variable that is given by Theorem 3.15. We index the N bits of Z by tuples (x, y, j) where $x \in \{0,1\}^n$, $y \in \{0,1\}^d$, and $j \in [m]$. We define the (random) function $F\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ by $F(x,y)_j = Z_{(x,y,j)}$. Note that, as a random variable, F is completely determined by Z. For any $z \in \mathsf{supp}(Z)$, we denote the (deterministic) function $F \mid (Z = z)$ by f_z.

We show that there exists $z \in \mathsf{supp}(Z)$ such that f_z is a (k, ε)-strong seeded extractor. By Lemma 3.11, it suffices to consider only (n, k)-flat sources. Fix such a flat source X. We slightly abuse notation and denote the support of X also by X. We further fix $\tau\colon \{0,1\}^m \to \{0,1\}$ and $y \in \{0,1\}^d$.

As $|X| = 2^k$, by our choice of K, the joint distribution of the random variables $\{F(x,y) \mid x \in X, y \in \{0,1\}^d\}$ is δ-close to uniform. Let $\{R_{(x,y)} \mid x \in X, y \in \{0,1\}^d\}$ be a sequence of uniform and independent m-bit random variables. We say that y is "bad" with respect to X, τ if

$$\left| \mathop{\mathbf{E}}_{x\sim X}\left[\tau(R_{(x,y)})\right] - \mathop{\mathbf{E}}_{v\sim U_m}\left[\tau(v)\right] \right| \geq \varepsilon,$$

and denote the set of bad y's with respect to X, τ by $B_{X,\tau}$. By Theorem 3.3, for any $y \in \{0,1\}^d$,

$$\mathbf{Pr}\left[y \in B_{X,\tau}\right] \leq 2 \cdot e^{-2\varepsilon^2 \cdot 2^k}.$$

Thus, by the union bound over all functions $\tau\colon \{0,1\}^m \to \{0,1\}$,

$$\mathbf{Pr}\left[y \in B_X\right] \leq 2^{2^m} \cdot 2 \cdot e^{-2\varepsilon^2 \cdot 2^k} \leq 2^{-\varepsilon^2 \cdot 2^k},$$

where $B_X = \cup_\tau B_{X,\tau}$. The last inequality follows by taking the constant $c = 2\log(1/\varepsilon) + c'$ for some universal constant c'. By the union bound,

$$\mathbf{Pr}\left[|B_X| \geq \varepsilon \cdot 2^d\right] \leq \binom{2^d}{\varepsilon \cdot 2^d} \cdot \left(2^{-\varepsilon^2 \cdot 2^k}\right)^{\varepsilon \cdot 2^d} \leq 2^{-\varepsilon^3 \cdot 2^{k+d}/2}$$

where the last equality follows as $k \geq 2\log(1/\varepsilon) + \log\log(1/\varepsilon) + 1$. Thus, by our choice of δ, the probability that the event $|B_X| \geq \varepsilon \cdot 2^d$ holds with respect to Z is bounded above by $2^{-\varepsilon^3 \cdot 2^{k+d}/2} + \delta \leq 2^{-\varepsilon^3 \cdot 2^{k+d}/3}$. By the union bound over all k-flat sources X, we have that except with probability $\binom{2^n}{2^k} \cdot 2^{-\varepsilon^3 \cdot 2^{k+d}/2}$ over $z \sim Z$, the function f_z has the following property. For any flat source X with support of size 2^k, at most ε fraction of the seeds are bad with respect to X. By our choice of parameters, the above expression is strictly smaller than 1 and so there exists some z such that f_z is a $(k, O(\sqrt{\varepsilon}))$-strong seeded extractor. The error guarantee can be easily reduced to ε.

Now that it has been established that there exists $z \in \mathsf{supp}(Z)$ for which f_z is a (k,ε)-strong seeded extractor, we are ready to define the function Ext. On input x, y, we find the first (under some order) $z^* \in \mathsf{supp}(Z)$ for which f_{z^*} is a (k,ε)-strong seeded extractor. We do so by iterating over all $z \in \mathsf{supp}(Z)$ according to the order, and for each z check whether f_z is a (k,ε)-strong seeded extractor. This is done in a brute force manner, by iterating over all k-flat sources X and functions $\tau\colon \{0,1\}^m \to \{0,1\}$, and for each, check whether f_z has the desired guarantee with respect to the current pair of X, τ. This can be done in time polynomial in

$$|\mathsf{supp}(Z)| \cdot 2^{2^m} \cdot \binom{2^n}{2^k} = \exp(n \cdot 2^k k).$$

Once z^* is found, the output $\mathsf{Ext}(x,y)$ is defined by $f_{z^*}(x,y)$. This evaluation can be computed in time that is dominated by the above running-time. □

Proof of Proposition 7.1. For the construction of Ext, we make use of the following building blocks:

- Set $d_1 = \log_2 n + \log_2 k + O(1)$, and let $\mathsf{Cond}\colon \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^\ell$ be the $k \to_\varepsilon k + d_1$ condenser that is given by Theorem 3.6. By Theorem 3.6, $\ell = d_1(k+2)$.
- Set $d_2 = \log_2 \ell + O(1)$, and let $\mathsf{Ext}'\colon \{0,1\}^\ell \times \{0,1\}^{d_2} \to \{0,1\}^m$ be the strong $(k - O(1), \varepsilon)$-seeded extractor that is given by Lemma 7.2. By Lemma 7.2, $m = k - O(1)$.

We partition the seed y to two consecutive strings $y = y_1 \circ y_2$ where $|y_1| = d_1$ and $|y_2| = d_2$. We then define

$$\mathsf{Ext}(x,y) = \mathsf{Ext}'(\mathsf{Cond}(x,y_1),y_2).$$

Let X be an (n,k)-source and Y a d-bit uniformly distributed random variable that is independent of X. By Theorem 3.6 and Lemma 3.7, with probability $1-\sqrt{\varepsilon}$ over $y_1 \sim Y_1$, the random variable $\mathsf{Cond}(X,y_1)$ is $2\sqrt{\varepsilon}$-close to an $(\ell, k - O(1))$-source. For any such y_1, with probability $1-\varepsilon$ over $y_2 \sim Y_2$, it holds that $\mathsf{Ext}'(\mathsf{Cond}(X,y_1),y_2) \approx_{2\sqrt{\varepsilon}+\varepsilon} U_m$. Thus, $\mathsf{Ext}(X,Y)$ is a $(k, O(\sqrt{\varepsilon}))$ strong seeded extractor. The error guarantee can be reduced to ε without changing the proposition statement.

Note that d, the seed length for Ext, is $d_1 + d_2 = \log_2 n + 3\log_2\log_2 n + O(1)$ as stated. Further, recall that $m = k - O(1)$. The running-time for computing $\mathsf{Ext}(x,y)$ is $\mathrm{poly}(n) + \exp(\ell \cdot 2^k k) = \exp(n^\alpha \cdot \log^3 n)$. □

7.2 A Two-Source Condenser

In this section we prove the following proposition. Roughly speaking, the proposition states that one can transform two independent weak-sources to a single, shorter, weak-source with comparable min-entropy. Moreover, this transformation is “strong” in one of the sources.

Proposition 7.3. *For any integer n and constants $0 < \alpha, \varepsilon < 1$, there exists an $\exp(n^{2\alpha})$-time computable function $\mathsf{TwoSourceCond}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$, where $m = O(n^{1-\alpha} \cdot \log^5 n)$, having the following property. For any pair of independent $(n, \alpha\log n)$-sources X, Y, except with probability ε over $y \sim Y$, the random variable $\mathsf{TwoSourceCond}(X,y)$ is ε-close to having min-entropy $\alpha \log n - O(1)$.*

For the proof of Proposition 7.3 we also require the following variant of a lemma by [BKS+10].

Lemma 7.4. *For any integer n and constant $\varepsilon > 0$, there exists an $\exp(n^2)$-time computable strong (k,ε) two-source extractor $\mathsf{TwoSourceExt}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$, where $k = \log_2 n + O(1)$ and $m = k - O(1)$.*

Proof. Let c be a constant to be chosen later on. Similarly to the proof of Lemma 7.2, set

$$\begin{aligned} m &= k - c, \\ N &= 2^{2n} \cdot m, \\ K &= 2^k \cdot m, \\ \delta &= 2^{-\varepsilon^3 \cdot 2^{2k}}. \end{aligned}$$

Let Z be the (N, K, δ)-independent random variable that is given by Theorem 3.15. We index the N bits of Z by tuples (x, y, j) where $x, y \in \{0,1\}^n$, and $j \in [m]$. We define the (random) function $F\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ by $F(x,y)_j = Z_{(x,y,j)}$. Note that, as a random variable, F is completely determined by Z. For any $z \in \mathsf{supp}(Z)$, we denote the (deterministic) function $F \mid (Z = z)$ by f_z.

We turn to show that with strictly positive probability over $z \sim Z$, the deterministic function f_z is a strong (k, ε) two-source extractor. We show the strongest in the second source. The proof for the strongest in the first source is similar (and we do not need it anyhow.) Let X be an (n, k) flat source. We slightly abuse notation and also denote that support of X by X. Let $\tau\colon \{0,1\}^m \to \{0,1\}$ be an arbitrary function.

As $|X| = 2^k$ and since F has m output bits, our choice of K implies that for any $y \in \{0,1\}^n$, the joint distribution of $\{F(X,y) \mid x \in X\}$ is δ-close to uniform. Thus, the expectation of $\tau(F(X,y))$ is within distance δ to the expectation of τ applied to 2^k uniform and independent random variables. By Theorem 3.3, the probability that the latter expression has bias exceeding ε is bounded above by $2 \cdot e^{-2\varepsilon^2 \cdot 2^k}$. By taking the union bound over all $\tau\colon \{0,1\}^m \to \{0,1\}$, the probability that the number of "bad" y's exceeds $\varepsilon \cdot 2^k$ is bounded above by

$$\binom{2^n}{\varepsilon \cdot 2^k} \cdot \left(2 \cdot e^{-2\varepsilon^2 \cdot 2^k} \cdot 2^{2^m}\right)^{\varepsilon \cdot 2^k} < 2^{n \cdot 2^k - \varepsilon^3 \cdot 2^{2k}/2}.$$

By taking the union bound over all $\binom{2^n}{2^k}$ (n,k)-flat sources X, we conclude that except with probability $2^{2n \cdot 2^k - \varepsilon^3 \cdot 2^{2k}/2}$, over $z \sim Z$, the function f_z is a $(k, O(\varepsilon))$ two-source extractor. By our choice of parameters, this probability is strictly smaller than 1, and so there exists $z \in \mathsf{supp}(Z)$ for which f_z is a $(k, O(\varepsilon))$ two-source extractor. As in Lemma 7.2, for the construction of the two-source extractor we first find such z and then apply f_z to the samples. It can be easily verified that the running-time is as stated. □

We are now ready to prove Proposition 7.3.

Proof of Proposition 7.3. We first describe the construction of TwoSourceCond and then turn to the analysis. To this end, we make use of the following components:

- Let $\mathsf{Ext}\colon \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^\ell$ be the (k, ε)-strong seeded extractor that is given by Proposition 7.1 with $k = \alpha \log_2 n$. By Proposition 7.1, $\ell = k - O(1)$, where the $O(1)$ term depends only on the constant error guarantee ε. Further, the seed length $d = \log_2 n + 3 \log_2 \log_2 n + O(1)$. We further set $r = 2^d = O(n \cdot \log^3 n)$ and identify $[r]$ with $\{0,1\}^d$.

- Set $r' = \gamma \cdot 2^k / k$ for some constant $0 < \gamma < 1$ that we fix later on (as a function of the constant ε). Let $\mathsf{TwoSourceExt}\colon \{0,1\}^{r'\ell} \times \{0,1\}^{r'\ell} \to \{0,1\}^{\ell'}$ be the (k', ε) two-source extractor that is given by Lemma 7.4. Note that $k' = \log_2(r'\ell) + O(1) = k - \log(1/\gamma) + O(1)$. By taking $\gamma > 0$ small enough, one can ensure that $\ell \geq k'$. Moreover, by Lemma 7.4, $\ell' = k' - O(1) = k - O(1)$.

On inputs $x, y \in \{0,1\}^n$ the function $\mathsf{TwoSourceCond}$ is defined as follows. First, we apply Ext to x with each possible seed $s \in \{0,1\}^d$. We stack all outputs as the rows of an $r \times \ell$ matrix denoted by m^x. That is, for $i \in [r]$, the i'th row of m^x is given by $m^x_i = \mathsf{Ext}(x,i)$. We define m^y in an analogous way, namely, $m^y_i = \mathsf{Ext}(y,i)$. We partition the rows of m^x (resp. m^y) to r/r' consecutive blocks, each consists of r' rows. We denote these blocks by $b^x_1, \ldots, b^x_{r/r'}$ (resp. $b^y_1, \ldots, b^y_{r/r'}$). For each $i \in [r/r']$, we compute $z_i = \mathsf{TwoSourceExt}(b^x_i, b^y_i)$, where we treat b^x_i, b^y_i as a pair of $(r'\ell)$-bit strings, ignoring their matrix form. The output $\mathsf{TwoSourceCond}(x,y)$ is defined as the concatenation of all r/r' outputs $z_1, \ldots, z_{r/r'}$.

We turn to the analysis. Let X, Y be a pair of independent (n,k)-sources. By Theorem 7.1, if we denote the random variable m^x, with $x \sim X$, by m^X, then all but one third of the rows of m^X are 3ε-close to uniform. A similar property holds for the matrix m^Y defined in an analogous way.

As at most one third of the rows of each of the matrices m^X, m^Y are not 3ε-close to uniform, there exists a common row index $g \in [r]$ such that each of m^X_g, m^Y_g os 3ε-close to uniform. Thus, there is some common $g' \in [r/r']$ such that each of the blocks $b^X_{g'}, b^Y_{g'}$ is 3ε-close to a block that contains a uniform row. We denote these blocks by B_1, B_2, respectively. As each of B_1, B_2 is an $r' \times \ell$ random variable in the form of a matrix that contains a uniform row, each of B_1, B_2 is an $(r'\ell, \ell)$-source.

As $\ell \geq k'$, except with probability ε over $b_2 \sim B_2$, the random variable $\mathsf{TwoSourceExt}(B_1, b_2)$ is ε-close to uniform. Thus, except with probability 4ε over $y \sim Y$, the random variable $\mathsf{TwoSourceExt}(b^X_{g'}, b^y_{g'})$ is 4ε-close to a uniform string on ℓ' bits. Thus, except with probability 4ε over $y \sim Y$, the output $\mathsf{TwoSourceCond}(X, y)$ is 4ε-close to having min-entropy $\ell' = k - O(1)$. Up to the factor of 4 in the error guarantee, this is the property we were set to prove. Clearly, this factor of 4 can be eliminated by using components Ext and $\mathsf{TwoSourceExt}$ with error guarantee $\varepsilon/4$. This does not affect the statement of the proposition.

As for the running-time, by Proposition 7.1, computing the matrices m^x, m^y given x, y can be done in time $\exp(n^\alpha \cdot \log^3 n)$. By Lemma 7.4, each application of $\mathsf{TwoSourceExt}$ can be carried out in time $\exp((r'\ell)^2) = \exp(n^{2\alpha})$, which dominates the total running-time. To conclude the proof, note that the output length is $(r/r') \cdot \ell' = O(n^{1-\alpha} \cdot \log^5 n)$. □

7.3 Proof of Theorem 1.7

For the proof of Theorem 1.7 we make use of the following lemma that generalizes, in a straightforward manner, a lemma by Barak *et al.* [BKS+10] who proved a similar lemma for the special case $s = 2$ (see Lemma 7.4.)

Lemma 7.5. *For any constant integer $s \geq 2$, constant $\varepsilon > 0$, and integer n, there exists an $\exp\left(n^{1+\frac{1}{s-1}}\right)$-time computable (k, ε) s-source extractor* $\mathsf{Ext}\colon (\{0,1\}^n)^s \to \{0,1\}$, *where $k = \frac{1}{s-1} \cdot \log_2 n + O(1)$.*

Proof of Lemma 7.5. Set $N = 2^{sn}$, $K = 2^{sk}$, and $\delta = 2^{-\varepsilon^2 2^{sk}}$. Let Z be the (N, K, δ)-independent random variable given by Theorem 3.15. We index the N bits of Z by tuples $(x_1, \ldots, x_s)$ where for each $i \in [s]$, $x_i \in \{0,1\}^n$. We define the (random) function

$F\colon (\{0,1\}^n)^s \to \{0,1\}$ by $F(x_1,\ldots,x_s) = Z_{(x_1,\ldots,x_s)}$. Note that F is completely determined by Z. For any $z \in \mathsf{supp}(Z)$, we denote the deterministic function $F \mid (Z = z)$ by f_z.

We turn to show that with strictly positive probability over $z \sim Z$, the function f_z is a (k,ε) s-source extractor. By Lemma 3.11, it is enough to consider only (n,k)-flat-sources. Let $X_1,\ldots,X_s$ be independent (n,k) flat sources. We slightly abuse notation and also denote that support of each flat source X_i by X_i.

As $|X_i| = 2^k$ for all $i \in [s]$, our choice of K implies that the joint distribution of $\{F(x_1,\ldots,x_s) \mid x_i \in X_i\}$ is δ-close to uniform. Thus, the probability that the bias of $F(X_1,\ldots,X_s)$ exceeds ε is bounded above by $\delta + 2 \cdot e^{-2\varepsilon^2 2^{sk}}$. By the union bound over all $\binom{2^n}{2^k}^s$ s-tuples of (n,k)-flat-sources, we conclude that except with probability

$$\binom{2^n}{2^k}^s \cdot \left(\delta + 2 \cdot e^{-2\varepsilon^2 2^{sk}}\right)$$

over $z \sim Z$, the function f_z is a (k,ε) s-source extractor. Note that by our choice of parameters, the above expression is strictly smaller than 1.

Now that it has been established that there exists a $z \in \mathsf{supp}(Z)$ for which the (deterministic) function f_z is a (k,ε) s-source extractor, we define the s-source extractor to be the function f_{z^*}, where $z^* \in \mathsf{supp}(Z)$ is the first element for which f_{z^*} has the above property.

In order to evaluate the extractor, one first needs to find z^*. This can be done by iterating over all $z \in \mathsf{supp}(Z)$, according to the chosen order, and check in a brute-force manner whether f_z has the desired property. The running-time for finding z^* is polynomial in $|\mathsf{supp}(Z)| \cdot 2^{sn \cdot 2^k}$, which is bounded above by $\exp\left(n^{1+\frac{1}{s-1}}\right)$. Once z^* is found, evaluating f_{z^*} on a given input can be done in time comparable to the time required by the computation that was carried out so far. □

We are now ready to prove Theorem 1.7.

Proof of Theorem 1.7. For the proof we make use of the following building blocks:

- Set $\alpha = \frac{1}{s-1} + \frac{5\log_2\log_2 n}{\log_2 n}$ [2], and let $\mathsf{TwoSourceCond}\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ be the two-source condenser that is given by Proposition 7.3 with this choice of α and with error guarantee $\varepsilon = 1/(10s)$. By Proposition 7.3 and by our choice of α, $m = O(n^{1-\alpha} \cdot \log^5 n) = O\left(n^{1-\frac{1}{s-1}}\right)$.
- Let $\mathsf{Ext}'\colon (\{0,1\}^m)^{s-1} \to \{0,1\}$ be the $(s-1)$-source extractor for min-entropy $k' = \frac{1}{s-2} \cdot \log_2 m + O(1)$, set with error guarantee $1/20$, that is given by Lemma 7.5. Note that $k' = \frac{1}{s-1} \cdot \log_2 n + O(1)$.

With these building blocks, we define

$$\mathsf{Ext}(x_1,\ldots,x_s) = \mathsf{Ext}'\left(\mathsf{TwoSourceCond}(x_1,x_s),\ldots,\mathsf{TwoSourceCond}(x_{s-1},x_s)\right).$$

[2] Throughout this section we assumed that α is constant. However, it is enough that α is bounded by a constant.

We turn to the analysis. Let $X_1, \dots, X_s$ be independent (n, k)-sources. For $i \in [s-1]$, let $Y_i = \mathsf{TwoSourceCond}(X_i, X_s)$. By Proposition 7.3, for each $i \in [s-1]$, except with probability $1/(10s)$ over $x_s \sim X_s$, it holds that Y_i is $1/(10s)$-close to having min-entropy $k - O(1)$. Thus, by the union bound and by the triangle inequality for statistical distance, except with probability $1/10$ over $x_s \sim X_s$, $(Y_1, \dots, Y_{s-1}) \approx_{1/10} (Z_1, \dots, Z_{s-1})$, where $Z_1, \dots, Z_{s-1}$ are independent $(m, k - O(1))$-sources (that depends on the choice of x_s). Consider any such x_s. By Lemma 7.5, $\mathsf{Ext}'(Z_1, \dots, Z_{s-1}) \approx_{1/20} U$, and so $\mathsf{Ext}'(Y_1, \dots, Y_{s-1}) \approx_{1/5} U$. By taking the "bad" x_s into account, we have that $\mathsf{Ext}(X_1, \dots, X_s) \approx_{1/3} U$.

The running-time is

$$\begin{aligned} \exp\left(n^{2\alpha}\right) + \exp\left(m^{1+\frac{1}{s-2}}\right) &= \exp\left(n^{\frac{2}{s-1}} \cdot \log^{10} n\right) + \exp\left(\left(n^{1-\frac{1}{s-1}}\right)^{1+\frac{1}{s-2}}\right) \\ &= \exp\left(n + n^{\frac{2}{s-1}} \cdot \log^{10} n\right). \end{aligned}$$

This concludes the proof. □

References

[Abb72] H. L. Abbott. Lower bounds for some Ramsey numbers. *Discrete Mathematics*, 2(4):289–293, 1972.

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

[Alo98] N. Alon. The Shannon capacity of a union. *Combinatorica*, 18(3):301–310, 1998.

[BADTS16] A. Ben-Aroya, D. Doron, and A. Ta-Shma. Explicit two-source extractors for near-logarithmic min-entropy. In *Electronic Colloquium on Computational Complexity (ECCC)*, number 088, 2016.

[Bar06] B. Barak. A simple explicit construction of an $n^{\tilde{o}(\log n)}$-Ramsey graph. *arXiv preprint math/0601651*, 2006.

[BIW06] B. Barak, R. Impagliazzo, and A. Wigderson. Extracting randomness using few independent sources. *SIAM Journal on Computing*, 36(4):1095–1118, 2006.

[BKS$^+$10] B. Barak, G. Kindler, R. Shaltiel, B. Sudakov, and A. Wigderson. Simulating independence: New constructions of condensers, Ramsey graphs, dispersers, and extractors. *Journal of the ACM (JACM)*, 57(4):20, 2010.

[Bou05] J. Bourgain. More on the sum-product phenomenon in prime fields and its applications. *International Journal of Number Theory*, 1(01):1–32, 2005.

[BRSW12] B. Barak, A. Rao, R. Shaltiel, and A. Wigderson. 2-source dispersers for $n^{o(1)}$ entropy, and Ramsey graphs beating the Frankl-Wilson construction. *Annals of Mathematics*, 176(3):1483–1544, 2012.

[BSZ11] E. Ben-Sasson and N. Zewi. From affine to two-source extractors via approximate duality. In *Proceedings of the 43rd annual ACM Symposium on Theory of computing*, pages 177–186. ACM, 2011.

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

[CGL16] E. Chattopadhyay, V. Goyal, and X. Li. Non-malleable extractors and codes, with their many tampered extensions. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016.

[Chu81] F.R.K. Chung. A note on constructive methods for Ramsey numbers. *Journal of Graph Theory*, 5(1):109–113, 1981.

[CL16] E. Chattopadhyay and X. Li. Explicit non-malleable extractors, multi-source extractors and almost optimal privacy amplification protocols. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016.

[Coh15] G. Cohen. Local correlation breakers and applications to three-source extractors and mergers. In *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 845–862. IEEE, 2015.

[Coh16a] G. Cohen. Making the most of advice: New correlation breakers and their applications. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2016.

[Coh16b] G. Cohen. Non-malleable extractors-new tools and improved constructions. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 50. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[Coh16c] G. Cohen. Non-malleable extractors with logarithmic seeds. In *Electronic Colloquium on Computational Complexity (ECCC)*, page 30, 2016.

[Coh16d] G. Cohen. Two-source dispersers for polylogarithmic entropy and improved ramsey graphs. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 278–284. ACM, 2016.

[CRS14] G. Cohen, R. Raz, and G. Segev. Nonmalleable extractors with short seeds and applications to privacy amplification. *SIAM Journal on Computing*, 43(2):450–476, 2014.

[CS16] G. Cohen and L. Schulman. Extractors for near logarithmic min-entropy. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, page 14, 2016.

[CZ16] E. Chattopadhyay and D. Zuckerman. Explicit two-source extractors and resilient functions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, pages 670–683. ACM, 2016.

[DKSS09] Z. Dvir, S. Kopparty, S. Saraf, and M. Sudan. Extensions to the method of multiplicities, with applications to Kakeya sets and mergers. In *50th Annual IEEE Symposium on Foundations of Computer Science*, pages 181–190. IEEE, 2009.

[DLWZ14] Y. Dodis, X. Li, T. D. Wooley, and D. Zuckerman. Privacy amplification and nonmalleable extractors via character sums. *SIAM Journal on Computing*, 43(2):800–830, 2014.

[DP07] S. Dziembowski and K. Pietrzak. Intrusion-resilient secret sharing. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 227–237, 2007.

[DPW14] Y. Dodis, K. Pietrzak, and D. Wichs. Key derivation without entropy waste. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 93–110. Springer, 2014.

[DS07] Z. Dvir and A. Shpilka. An improved analysis of linear mergers. *Computational Complexity*, 16(1):34–59, 2007.

[DW09] Y. Dodis and D. Wichs. Non-malleable extractors and symmetric key cryptography from weak secrets. In *Proceedings of the forty-first annual ACM Symposium on Theory of Computing*, pages 601–610. ACM, 2009.

[Erd47] P. Erdős. Some remarks on the theory of graphs. *Bulletin of the American Mathematical Society*, 53(4):292–294, 1947.

[ES35] P. Erdös and G. Szekeres. A combinatorial problem in geometry. *Compositio Mathematica*, 2:463–470, 1935.

[Fra77] P. Frankl. A constructive lower bound for some Ramsey numbers. *Ars Combinatoria*, 3:297–302, 1977.

[FW81] P. Frankl and R. M. Wilson. Intersection theorems with geometric consequences. *Combinatorica*, 1(4):357–368, 1981.

[Gro01] V. Grolmusz. Low rank co-diagonal matrices and Ramsey graphs. *Journal of combinatorics*, 7(1):R15–R15, 2001.

[GS95] A. Garcia and H. Stichtenoth. A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vladut bound. *Inventiones Mathematicae*, 121(1):211–222, 1995.

[GUV09] V. Guruswami, C. Umans, and S. Vadhan. Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes. *Journal of the ACM*, 56(4):20, 2009.

[Li11] X. Li. Improved constructions of three source extractors. In *IEEE 26th Annual Conference on Computational Complexity*, pages 126–136, 2011.

[Li12a] X. Li. Design extractors, non-malleable condensers and privacy amplification. In *Proceedings of the forty-fourth annual ACM Symposium on Theory of Computing*, pages 837–854, 2012.

[Li12b] X. Li. Non-malleable extractors, two-source extractors and privacy amplification. In *IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 688–697, 2012.

[Li13] X. Li. Extractors for a constant number of independent sources with polylogarithmic min-entropy. In *IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 100–109, 2013.

[Li15] X. Li. Three-source extractors for polylogarithmic min-entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 2015.

[Li16] X. Li. Improved non-malleable extractors, non-malleable codes and independent source extractors, 2016. Personal communication.

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

[LWZ11] X. Li, T. D. Wooley, and D. Zuckerman. Privacy amplification and nonmalleable extractors via character sums. *arXiv preprint arXiv:1102.5415*, 2011.

[Nag75] Zs. Nagy. A constructive estimation of the Ramsey numbers. *Mat. Lapok*, 23:301–302, 1975.

[Nao92] M. Naor. Constructing Ramsey graphs from small probability spaces. *IBM Research Report RJ 8810*, 1992.

[NN93] J. Naor and M. Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM J. on Computing*, 22(4):838–856, 1993.

[PR04] P. Pudlák and V. Rödl. Pseudorandom sets and explicit constructions of Ramsey graphs. *Quad. Mat*, 13:327–346, 2004.

[Ram28] F. P. Ramsey. On a problem of formal logic. *Proceedings of the London Mathematical Society*, 30(4):338–384, 1928.

[Raz05] R. Raz. Extractors with weak random seeds. In *Proceedings of the thirty-seventh annual ACM Symposium on Theory of Computing*, pages 11–20, 2005.

[RRV99] R. Raz, O. Reingold, and S. Vadhan. Error reduction for extractors. In *40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 191–201. IEEE, 1999.

[Sti09] H. Stichtenoth. *Algebraic function fields and codes*, volume 254. Springer Science & Business Media, 2009.

[TS96a] A. Ta-Shma. On extracting randomness from weak random sources. In *Proceedings of the twenty-eighth annual ACM Symposium on Theory of Computing*, pages 276–285, 1996.

[TS96b] A. Ta-Shma. *Refining randomness*. PhD thesis, Hebrew University of Jerusalem, 1996.

[VDTR13] Alexander Vitanov, Frederic Dupuis, Marco Tomamichel, and Renato Renner. Chain rules for smooth min-and max-entropies. *Information Theory, IEEE Transactions on*, 59(5):2603–2612, 2013.

[Zuc07] D. Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3:103–128, 2007.