Inverted Leftover Hash Lemma

Maciej Obremski and Maciej Skórski

[1] Aarhus University
[2] IST Austria

Abstract. Universal hashing found a lot of applications in computer science. In cryptography the most important fact about universal families is the so called Leftover Hash Lemma, proved by Impagliazzo, Levin and Luby. In the language of modern cryptography it states that almost universal families are good extractors. In this work we provide a somewhat surprising characterization in the *opposite direction.* Namely, every extractor with sufficiently good parameters yields a universal family on a noticeable fraction of its inputs.
Our proof technique is based on tools from extremal graph theory applied to the "collision graph" induced by the extractor, and may be of independent interest. We discuss possible applications to the theory of randomness extractors and non-malleable codes.

keywords Min-Entropy Extractors, Universal Hash Functions, Extremal Graph Theory

1 Introduction

1.1 Universal Hashing and Leftover Hash Lemma

Universal hashing, introduced by Carter and Wegman [CW79], has found many applications in computer science such as parallel computing [LSS12; KSV10], data structures [Sie04; ÖP03], randomized algorithms [IZ89], complexity theory [Sip83] and many others. For cryptography particularly important is one statement about universal families called Leftover Hash Lemma, proved by Impagliazzo, Levin and Luby in [ILL89][3]. It has been recognized as a very useful tool for (a) randomness extraction [IZ89] (b) pseudorandomness [ILL89; Nis92] and (c) privacy amplification [BBCM95], followed by many other applications [BHKKR99; DRS04; HK97; HP08; Hay11; RW05; TV15; WC81].

The Leftover Hash Lemma (LHL), formulated in the language of randomness extractors, states that universal hash families are (seeded) extractors with best entropy/security tradeoff[4]. While universal hash functions have been (relatively easily) shown to be equivalent to several other structures such as error-correcting

[3] The term "Leftover Hash Lemma" was used for the first time in [IZ89]

[4] Extracting m bits ϵ-close to uniform from a source of $m + 2\log(1/\epsilon)$ bits of entropy, which is optimal as shown in [RT00]

codes [BJKS93] or combinatorial designs (balanced incomplete block designs, difference matrices, orthogonal arrays) [Sti95], the link between universal hash families and extractors established by the LHL is one-way and somewhat incomplete: it is not clear if an arbitrary extractor is related to universal hash families in any way.

In this paper we complete the picture by providing the "missing" link to the relation between universal hash families and extractors. Namely (somewhat surprisingly), we prove that any extractor can be viewed as a universal family when restricted to a noticeable subset of inputs. This "reversed" LHL seems to be interesting in its own right, as it shows that universal hashing is, in some sense, necessary for designing extractors.

Our result follows from non-constructive techniques. If we could find an efficient and generic (we know how to solve this problem in some special instances) ways to truncate extractors domain our result would have other interesting consequences for example for flexible extractors and non-malleable codes.

1.2 Our Results and Techniques

Results Our main result shows that for every seeded extractor has a "core" being a universal hashing faimily: there exists a significant fraction of inputs where the extractor yields an almost universal hash family. More precisely, suppose that $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ extracts m bits that are ϵ-close to uniform from any n-bit source of min-entropy k, with the help of a d-bit uniform seed. Then the family of functions $h_s(x) = \mathsf{Ext}(x,s)$ (indexed by seeds s) is almost universal on a set of size 2^{n-k}. The exact statement is given Theorem 1 in Section 3, and the parameters and a comparison with the LHL is illustrated in Figure 1 below. The size of the "core" set is optimal as discussed in Section 4. In Section 5 we show how to amplify this result and apply it to flexible two-source extractors.

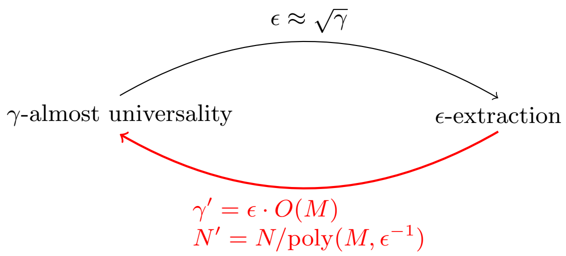
$\epsilon \approx \sqrt{\gamma}$
γ-almost universality
ϵ-extraction
$\gamma' = \epsilon \cdot O(M)$
$N' = N/\mathrm{poly}(M, \epsilon^{-1})$

Fig. 1. From γ-universality to ϵ-extraction and back. Both universal families and extractors are from $[N]$ to $[M]$. The extractors is assumed to work for uniform sources over $K = \mathrm{poly}(M, \epsilon^{-1})$ elements. When going from extractors to universal families a loss occurs in the universality parameter (γ') and in the domain size (N').

Proof overview and techniques In order to prove our result, we consider the "correlation" graph over the extractor inputs. Two inputs x, x' are considered correlated (and linked by an edge) when there is a lot of collisions $\mathsf{Ext}(x, s) = \mathsf{Ext}(x', s)$ over different seeds s. Since the output of the extractor should be distributed almost uniformly when x is sampled from a set of size 2^k (conditioned on the uniformly sampled seed) we conclude that no x can have more than 2^k neighbours x' (otherwise 2^k elements are mapped into the same output for many choices of the seed). By results from extremal graphs we obtain that there is an independent set of vertices (no link between any pair of vertices) of size roughly $2^n/2^k$. In this independent set any pair of vertices has a low number of collisions (because the lack of an edge), hence the restricted extractor is an almost universal family.

Applications We discuss the following consequences of our result

(a) two-source extractors: every two-source extractor is flexible on a restricted domain. Flexible two-source extractors are slight generalization of two-source extractors, they proved to be very useful tools in leakage and tamper resilient cryptography (see: [DW09; DDV10; DLWZ11; DF12; CRS12], and specifically for flexibility see: [DKO13; ADL14]). We discuss the implications in Section 5 and Section 6.1.

(b) non-malleable extractors and codes: our result allows trading a non-malleable extraction rate for better leakage rate which could improve the parameters of continuous non-malleable codes [DNO17]. For details, see Section 6.2.

(c) bounds for extractor seeds: our result implies a lower bound on the seed length for any extractor. While this bound is slightly worse than the RT bound [RT00], our proof offers a simple and intuitive explanation: the seed needs to be sufficiently long, because the extractor is partly a universal family and universal families require large key spaces (to be indexed upon). For details we refer to Section 6.3.

1.3 Organization

We prove the main result in Section 3. The lower bounds are shown in Section 4. In Section 5 we show how to amplify this result and apply it to flexible two-source extractors. Applications to leakage-resilient storage, non-malleable extractors and codes, and randomness extractors are discussed in Section 6.

2 Preliminaries

Basic notations The statistical distance of two random variables X, Y over a finite set $\mathcal{X}$ is denoted by $\mathrm{SD}\left(X;Y\right) = \frac{1}{2}\sum_x |\Pr[X = x] - \Pr[Y = x]|$. For any set S by U_S we understand a random variable uniformly distributed over S.

Min-entropy

Definition 1 (Min-entropy). *Let X be random variable, its min-entropy $\mathbf{H}_\infty(X)$ is defined as below:*

$$\mathbf{H}_\infty(X) = -\log \max_x \Pr(X = x).$$

Definition 2 (Average min-entropy). *Let X, Y be random variable, average min-entropy $\widetilde{\mathbf{H}}_\infty(X|Y)$ is defined as below:*

$$\widetilde{\mathbf{H}}_\infty(X|Y) = -\log \mathbb{E}_y \max_x \Pr(X = x|Y = y).$$

Lemma 1 (Decomposition). *Let k be such that $2^k \in \mathbb{N}$. Any random variable X with $\mathbf{H}_\infty(X) > k$ can be represented as convex combination of flat distributions each with support of size at least 2^k.*

Universal families

Definition 3 (Almost universality). *We say that a family of functions $\mathcal{H}$ from $\{0,1\}^n$ to $\{0,1\}^m$ is γ-almost universal if for a random member H of $\mathcal{H}$ and any different inputs $x, x' \in \{0,1\}^n$ we have*

$$\Pr[H(x) = H(x')] \leqslant \frac{1+\gamma}{M}.$$

We also say that $\mathcal{H}$ is γ-almost universal hash family, abbreviated to γ-UHF.

Extractors

Definition 4 (Extractors). *A function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k,ϵ)-seeded extractor if for any source $X \in \{0,1\}^n$ such that $\mathbf{H}_\infty(X) \geqslant k$ we have*

$$\mathrm{SD}\left((\mathsf{Ext}(X, U_{\{0,1\}^d}), U_{\{0,1\}^d}); (U_{\{0,1\}^m}, U_{\{0,1\}^d})\right) \leqslant \epsilon.$$

Lemma 2 (Extractors from universal hashing). *Every γ-almost universal family $\{h_y\}_y$ from $\{0,1\}^n$ to $\{0,1\}^m$ yields, by $\mathsf{Ext}(x,y) = h_y(x)$, a (k,ϵ)-extractor where $\epsilon = \frac{1}{2}\sqrt{2^{m-k} + \gamma}$.*

Two-source extractors

Definition 5 (Two-source extractors). *A function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ is a (k,ϵ)-seeded extractor if for any random variables X, Y such that $\mathbf{H}_\infty(X) > k$ and $\mathbf{H}_\infty(Y) > k$ we have*

$$\begin{aligned}&\mathrm{SD}\left((\mathsf{Ext}(X,Y), Y); (U_{\{0,1\}^m}, Y)\right) \leqslant \epsilon,\\ &\quad \textit{and}\\ &\mathrm{SD}\left((\mathsf{Ext}(X,Y), X); (U_{\{0,1\}^m}, X)\right) \leqslant \epsilon.\end{aligned}$$

Definition 6 (Flexible two-source extractors). *A function* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *is a* (k,ϵ)*-seeded extractor if for any any random variables* X, Y *such that* $\mathbf{H}_\infty(X) + \mathbf{H}_\infty(Y) > k > n$ *we have*

$$\begin{aligned} &\mathrm{SD}\left((\mathsf{Ext}(X,Y),Y);(U_{\{0,1\}^m},Y)\right) \leqslant \epsilon,\\ &\quad \textit{and}\\ &\mathrm{SD}\left((\mathsf{Ext}(X,Y),X);(U_{\{0,1\}^m},X)\right) \leqslant \epsilon. \end{aligned}$$

Definition 7 (Non-malleable extractors). *A function* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *is a* (k,ϵ)*-seeded extractor if for any random variables* X, Y *such that* $\mathbf{H}_\infty(X) > k$ *and* $\mathbf{H}_\infty(Y) > k$*, and any functions* $f, g : \{0,1\}^n \to \{0,1\}^n$ *we have*

$$\mathrm{SD}\left((\mathsf{Ext}(X,Y),\mathsf{Ext}(f(X),g(Y)));(U_{\{0,1\}^m},\mathsf{Ext}(f(X),g(Y)))\right) \leqslant \epsilon.$$

For a construction see [Li16].

3 Main Result

3.1 Auxiliary results from graph theory

We will use the following well-known results from graph theory.

Lemma 3 (Handshaking Lemma). *For any graph* G *with vertices* V *and edges* E *and we have* $\sum_{v\in V}\deg(v) = 2|E|$.

Lemma 4 (Turan's Theorem). *Every graph with* N *vertices and more than* $\left(1-r^{-1}\right)\binom{N}{2}$ *edges contains a clique of size* $r+1$.

3.2 Proof of the main result

Theorem 1 (Every extractor is a universal family on a restricted set of inputs). *Let* $f : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* (k,ϵ)*-extractor. Then there is a subset* $S \subset \{0,1\}^n$ *of size* $\frac{2^n}{2^k}$ *such that the family* $\{f(x,y)\}_{y\in\{0,1\}^d}$ *is* $2^m\epsilon$*-almost universal on* S.

Remark 1 (The universal subdomain is of noticable size). Note that best extractors achieve $k = O(\log(1/\epsilon))$. Most restrictive settings for applications of extractors require $\epsilon = (2^n)^{-O(1)}$ which gives that the subset S from the theorem above is $\Omega\left(\frac{1}{n}\right)$-dense. In particular for $\epsilon = 2^{-O(n)}$ we obtain $|S|/2^n = \Omega\left(n^{-1}\right)$.

Correlation coefficients Define the following *correlation coefficients*

$$\rho(x,x') \overset{def}{=} \Pr_{y\sim[D]}\left[f(x,y) = f(x',y)\right]$$

Correlation graph Consider the graph G with vertex set $[N]$ and the edge set E consisting of inputs (x, x') for which the extractor outputs coincide only for a small fraction of seeds

$$(x, x') \in E \Leftrightarrow \rho(x, x') \leqslant \frac{1+\gamma}{2^m}.$$

Intuitively, two inputs are considered uncorrelated if they don't collide to often in the extractor mapping once the seed changes.

Claim (Large universal subset for f $\Leftrightarrow$ G has a large clique). There exists a subset of size R on which the family of functions $\{f(x,y)\}_y$ is γ-almost universal, if and only if G has a clique of size R.

Proof. Follows from the definitions of ρ and G.

Claim (Extraction $\Rightarrow$ all degrees in G are large). Suppose that the minimum degree of G is at most $2^n - 2^k$. Then f cannot be a $\left(k, \frac{\gamma}{2^m}\right)$-extractor.

Proof. By definition of G, for some x and the set S of at least 2^k vertices x' not connected to x (counting x itself) we obtain $\rho(x, x') > \gamma$ for every $x' \in S$. In other words, for one fixed $x' \in S$ and all $x \in S$ we have

$$\Pr_{y \sim \{0,1\}^d} [f(x,y) = f(x',y)] > \frac{1+\gamma}{2^m}.$$

Since the above is true for every x, by the total probability law we obtain also

$$\Pr_{y \sim \{0,1\}^d, x \sim S} [f(x,y) = f(x',y)] > \frac{1+\gamma}{2^m}.$$

Fix x' and consider now the following distinguisher

$$D(u,y) = \begin{cases} 1, \, u = f(x',y) \\ 0, \, u \neq f(x',y) \end{cases}$$

By the discussion above we have

$$\Pr_{y \sim \{0,1\}^d, x \sim S} [D(f(x,y),y) = 1] > \frac{1+\gamma}{2^m}. \tag{1}$$

By the properties of the uniform distribution we have

$$\Pr_{y \sim \{0,1\}^d, u \sim U(\{0,1\}^m)} [D(u,y) = 1] = \Pr_{y \sim \{0,1\}^d} \Pr_{u \sim U(\{0,1\}^m)} [u = f(x',y)] = \frac{1}{2^m}. \tag{2}$$

Therefore we conclude that

$$\Pr_{y \sim \{0,1\}^d, x \sim S} [D(f(x,y),y) = 1] - \Pr_{y \sim \{0,1\}^d, u \sim U(\{0,1\}^m)} [D(u,y) = 1] > \frac{\gamma}{2^m} \tag{3}$$

which means that f is not a $(\log |S|, \gamma/2^m)$-extractor.

Claim (Large degrees in $G \Rightarrow G$ has a large clique). If the minimum degree of G is bigger than $2^n - 2^k$ then there is a clique of size $\frac{2^n}{2^k}$.

Proof. By the Handshaking Lemma (Lemma 3), we have at least $\frac{1}{2}2^n(2^n - 2^k + 1)$ edges in G. Note that

$$\frac{1}{2}2^n(2^n - 2^k + 1) = \binom{2^n}{2}\left(1 - \frac{2^k - 2}{2^n - 1}\right),$$

and hence by Turan's theorem (Lemma 4) there is a clique of size at least $\left\lfloor \frac{2^n-1}{2^k-2} - \alpha \right\rfloor + 1$ where α is any positive constant (to make sure that the inequality in the thereom is strict). It remains to observe that for sufficiently small α this is always at least $\frac{2^n-1}{2^k-2} \geq \frac{2^n}{2^k}$.

Theorem 1 follows now by combining the last two claims.

4 Discussing optimality

In Theorem 1 we showed that there exists a clique of size $\frac{2^n}{2^k}$, in this section we will discuss optimality of that theorem, namely we will show that there exists extractors with cliques of size at most $O\left(\frac{2^n}{2^k}\right)$. For that claim we could use the Turan's graphs however it is not obvious that it is possible to build extractor from any correlation graph. Instead we will define extractor that achieves abovementioned bound. By inspection of the correlation structure of that graph reader can notice it is very slight modification of Turan's graph which actually leads to better overall extraction parameters while maintaining small cliques sizes.

Definition 8. *Let $\mathsf{Ext}^{\mathsf{l,n}}_{bad} : \{0,1\}^n \times \{0,1\}^{n-l} \to \{0,1\}$ be seeded extractor with seed $\{0,1\}^{n-l}$ such that, $\mathsf{Ext}^{\mathsf{l,n}}_{bad}([L,X];Y) = \langle X, Y \rangle$, where $L \in \{0,1\}^l$ and $\langle .,. \rangle$ stands for standard inner product over $\{0,1\}$.*

Lemma 5 (Parameters of extraction). $\mathsf{Ext}^{\mathsf{l,n}}_{bad}$ *is $(k, 2^{-\frac{(k-l-1)}{2}})$-extractor.*

Proof. First notice that

$$\mathrm{SD}[(\mathsf{Ext}^{\mathsf{l,n}}_{\mathsf{bad}}([L,X],Y),Y);(U_{\{0,1\}},Y)] = \mathrm{SD}[(\langle X,Y \rangle, Y);(U_{\{0,1\}},Y)]$$

We know that inner product is Universal Hashing Family, thus by Lemma 2.4 from [DORS08] we obtain that

$$\mathrm{SD}[(\langle X,Y \rangle, Y);(U_{\{0,1\}},Y)] \leq \sqrt{\frac{2}{2^{\widetilde{\mathbf{H}}_\infty([L,X]|L)}}}$$

To finalize the proof we notice that $\widetilde{\mathbf{H}}_\infty([L,X]|L) \geq \mathbf{H}_\infty([L,X]) - l = k - l$.

Lemma 6 (Clique size for $\mathsf{Ext}^{\mathsf{l,n}}_{\mathbf{bad}}$). *Notice that the largest clique C in correlation graph of $\mathsf{Ext}^{\mathsf{l,n}}_{bad}$ is of size $|C| = 2^{n-l}$.*

Proof. Assume $|C| > 2^{n-l}$ then there exists 2 elements in C of the same suffixes: $[l_1, x]$ and $[l_2, x]$ which by definition of extractor are not connected by low-correlation-edge (they have full correlation coefficient).

Corollary 1. *For every k, n there exists extractor (f.e. $\mathsf{Ext}_{bad}^{\mathsf{k-1,n}}$) that has low-correlation cliques of size at most $2 \cdot \frac{2^n}{2^k}$.*

5 From extraction to flexible extraction

Introduction In this section we will show that for every two-source extractor we can "cut out" part of his domain and then it becomes a flexible two source extractor i.e. the extraction depends on sum of entropies of sources ($\mathbf{H}_\infty(X) + \mathbf{H}_\infty(Y)$) rather then on each source exceeding certain threshold ($\mathbf{H}_\infty(X) > k$ and $\mathbf{H}_\infty(Y) > k$). In the process the entropy-rate is improved.

Given a (k, ϵ)-two-source extractor with a sources in $\{0,1\}^n$ it is trivial that such extractor is a $(k + n, \epsilon)$-flexible extractor, simply because if $\mathbf{H}_\infty(X) + \mathbf{H}_\infty(Y) > k + n$ and X, Y are independent then $\mathbf{H}_\infty(X) > k$, and $\mathbf{H}_\infty(Y) > k$. We show that it is de facto possible to obtain something better then trivial solution.

Theorem 2. *Let $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ be a (k, ϵ)-two source extractor. Then for any $0 < d < k$, there exists subsets of the domain $A_x, A_y \in \{0,1\}^n$ such that*

- $|A_x| = |A_y| = 2^n \cdot 2^{-d}$
- $\mathsf{Ext}_{|A_x \times A_y}$ *is $(n + k - d, \epsilon')$-flexible extractor.*

Above theorem follows from Theorem 3 which we prove below. For the exact equation for ϵ' see Theorem 3.

Optimality of Theorem 2. Let us consider following two-source extractor: $\mathsf{Ext}_{\text{bad}}^{\mathsf{l,n}} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ such that for $X, Y \in \{0,1\}^{n-l}$,

$$\mathsf{Ext}_{\text{bad}}^{\mathsf{l,n}}([L, X]; [R, Y]) = \langle X, Y \rangle$$

Such extractor is a $(k + l + n, O(2^{-\frac{k}{2}}))$-flexible two-source extractor (proof is almost identical as in Lemma 5 thus we ommit it). To improve the entropy bound we have to fix the prefixes that are being "ignored" by the extractor. The size of a truncated domain and the entropy threshold are the same as in Theorem 2.

Technical part. We start with a small generalization of the Leftover Hash Lemma from [LLTT05], by Lemma 1 it is sufficient to consider only flat distributions.

Lemma 7. *Let $H : \{0,1\}^n \to \{0,1\}^m$ be a $\gamma-$almost Universal Hash Family. Let G be a flat distributed over some subset of functions in H and let X be a flat distribution over $\{0,1\}^n$, then:*

$$\mathrm{SD}[(G(X), G); (U_m, G)] \leq \frac{1}{2}\sqrt{\frac{2^m \cdot |H|}{|G| \cdot |X|} + \frac{|H| \cdot \gamma}{|G|}}$$

Proof. We will associate G with a support of G, same for X.

$$4 \cdot \mathrm{SD}[(G(X), G); (U_m, G)]^2 = \left(\sum_{h \in G} \sum_{z \in \{0,1\}^m} \frac{1}{|G|} \left| \Pr_{x \in X}[h(x) = z] - \frac{1}{2^m} \right| \right)^2 \leq$$ [5]

$$\leq \frac{2^m}{|G|} \cdot \sum_{h \in H} \sum_{z \in \{0,1\}^m} \left(\Pr_{x \in X}[h(x) = z] - \frac{1}{2^m} \right)^2 =$$ [6]

$$= \frac{2^m}{|G|} \left[\left(\sum_{h \in H} \sum_{z \in \{0,1\}^m} \Pr_{x, x' \in X}[h(x) = h(x') = z] \right) \right.$$

$$\left. -2 \cdot \sum_{h \in H} \sum_{z \in \{0,1\}^m} \left(\Pr[h(x) = z] \cdot \frac{1}{2^m} \right) + \frac{|H|}{2^m} \right] =$$ [7]

$$= \frac{2^m}{|G|} \left(\Pr_{h \in H; x, x' \in X}[h(x) = h(x')] \cdot |H| - \frac{|H|}{2^m} \right) =$$

$$= \frac{2^m \cdot |H|}{|G|} \left(\Pr[h(x) = h(x') | x \neq x'] + \Pr[x = x'] - \frac{1}{2^m} \right) \leq$$

$$\leq \frac{2^m \cdot |H|}{|G|} \left(\frac{1 + \gamma}{2^m} + \frac{1}{|X|} - \frac{1}{2^m} \right) = \frac{2^m \cdot |H|}{|G| \cdot |X|} + \frac{|H| \cdot \gamma}{|G|}$$

Theorem 3. *For every $(k, \epsilon)-$two source extractor* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *and for any $T > 1$ there exists sets $D_x, D_y \in \{0,1\}^n$ such that for every $\delta > 0$*

1. $\mathsf{Ext}_{|(\{0,1\}^n \setminus D_x) \times (\{0,1\}^n \setminus D_y)}$ *is strong flexible $(r, \epsilon')-$extractor where* $\epsilon' = \frac{2^m}{2} \sqrt{2^{m+n+\log T - r + \delta} + 2^{n+m-\frac{r-\delta}{2}} \cdot \epsilon} + 2^{-\delta}$.
2. $|D_x| = |D_y| = 2^n \cdot \frac{2^k}{2^k + T}$

In other words: if we limit source X to be distributed over $\{0,1\}^n \setminus D_x$, and source Y to be distributed over $\{0,1\}^n \setminus D_y$ then for any $\delta > 0$:

$$\mathrm{SD}((\mathsf{Ext}(X, Y), Y); (U_m, Y)) \leq$$
$$\leq \frac{2^m}{2} \sqrt{2^{m+n+\log T - (\mathbf{H}_\infty(X) + \mathbf{H}_\infty(Y) - \delta)} + 2^{n+m-\frac{(\mathbf{H}_\infty(X) + \mathbf{H}_\infty(Y) - \delta)}{2}} \cdot \epsilon} + 2^{-\delta}$$

Proof. The idea behind this proof is to treat $\mathsf{Ext}(., Y)$ as a seeded extractor for seed Y, then use Theorem 1 multiple times. Notice that after finding and cutting out the UHF clique the remaining part of domain has enough points to still extract, thus we can cut out another clique and so on. Assume we want to cut the cliques/UHFs of the size C at every step and we want T of such cliques. By Theorem 1 our only restrictions are:

[5] Follows from Jensen's inequality.

[6] Follows from decomposition of $(...)^2$.

[7] Follows from $\sum_z \Pr[h(x) = z] = 1$.

- Domain size drops after each cut, we need that after $T-1$ cuts there is still enough space left for large clique: $\frac{2^n-(T-1)\cdot C}{2^k} \geq C$. For simplicity assume $\frac{2^n-T\cdot C}{2^k} = C$.
- The other restriction is that after $T-1$ cuts there must be enough points for last extraction (and last cut): $2^n - (T-1)\cdot C \geq 2^k$. However this condition is trivially fulfilled by $\frac{2^n-T\cdot C}{2^k} = C$.

After finishing the above procedure we are left with $2^n - T\cdot C$ points which we assign to the set D_x. Notice that by $\frac{2^n-T\cdot C}{2^k} = C$ we get $|D_x| = C\cdot 2^k$ and thus

$$2^n - T\cdot\frac{|D_x|}{2^k} = |D_x|$$

$$|D_x| = 2^n\cdot\frac{2^k}{2^k+T}$$

Let us consider $\{0,1\}^n \setminus D_x$ as a domain of seeded extractor $\mathsf{Ext}(.,Y)$. The domain consists of T disjoint cliques each is a $M\epsilon-$Universal Hashing Family. Order the cliques $C_1,...,C_T$ and let $T(x) : \{0,1\}^n \setminus D_x \to [T]$ be such that $T(x) = i$ if and only if $x \in C_i$. Let X be a flat distribution on $\{0,1\}^n \setminus D_x$, let us calculate

$$\mathrm{SD}[(\mathsf{Ext}(X,Y),Y);(U_m,Y)] \leq \mathrm{SD}[(\mathsf{Ext}(X,Y),Y,T(X));(U_m,Y,T(X))]$$

Notice that $\widetilde{\mathbf{H}}_\infty(X|T(X)) = \mathbf{H}_\infty(X) - \log T$, also by Lemma 2.4 from [DORS08] every UHL-based extractor is also extractor for avg-min-entropy, thus by Lemma 7:

$$\begin{aligned}
&\mathrm{SD}[\mathsf{Ext}(X,Y);U_m] \leq \\
&\leq \mathrm{SD}[(\mathsf{Ext}(X,Y),Y,T(X));(U_m,Y,T(X))] \leq \\
&\leq \frac{1}{2}\sqrt{2^{m+n-(\widetilde{\mathbf{H}}_\infty(X|T(X))+\mathbf{H}_\infty(Y))} + 2^{n+m-\mathbf{H}_\infty(Y)}\cdot\epsilon} = \\
&= \frac{1}{2}\sqrt{2^{m+n+\log T-(\mathbf{H}_\infty(X)+\mathbf{H}_\infty(Y))} + 2^{n+m-\mathbf{H}_\infty(Y)}\cdot\epsilon} \qquad (4)
\end{aligned}$$

If we symmetrically repeat the procedure for $\mathsf{Ext}(X,.)$ seeded extractor with seed X we will obtain set D_y and for Y distributed over $\{0,1\}^n \setminus D_y$

$$\begin{aligned}
&\mathrm{SD}[\mathsf{Ext}(X,Y);U_m] \leq \\
&= \frac{1}{2}\sqrt{2^{m+n+\log T-(\mathbf{H}_\infty(X)+\mathbf{H}_\infty(Y))} + 2^{n+m-\mathbf{H}_\infty(X)}\cdot\epsilon} \qquad (5)
\end{aligned}$$

If we restrict domain of the extractor to $(\{0,1\}^n \setminus D_x) \times (\{0,1\}^n \setminus D_y)$ we can combine Equation (4) with Equation (5) and obtain

$$\mathrm{SD}[\mathsf{Ext}(X,Y);U_m] \leq \epsilon_{\mathsf{Ext}}$$

where

$$\begin{aligned}
\epsilon_{\mathsf{Ext}} &= \frac{1}{2}\sqrt{2^{m+n+\log T-(\mathbf{H}_\infty(X)+\mathbf{H}_\infty(Y))} + 2^{n+m-\max\{\mathbf{H}_\infty(X);\mathbf{H}_\infty(Y)\}}\cdot\epsilon} \leq \\
&\leq \frac{1}{2}\sqrt{2^{m+n+\log T-(\mathbf{H}_\infty(X)+\mathbf{H}_\infty(Y))} + 2^{n+m-\frac{\mathbf{H}_\infty(X)+\mathbf{H}_\infty(Y)}{2}}\cdot\epsilon}
\end{aligned}$$

To obtain a strong flexible extraction, i.e. to calculate $\mathrm{SD}[(ext(X,Y),Y);(U_m,Y)]$ we apply Claim 3 from [DKO13] (which states that every weak flexible extractor is also a strong flexible extractor with a slightly worse parameters) and obtain: that for any $\delta > 0$, extractor $ext_{|(\{0,1\}^n \setminus D_x)\times(\{0,1\}^n \setminus D_y)}$ is (r', ϵ') strong flexible extractor for

$$\epsilon' = \frac{2^m}{2}\sqrt{2^{m+n+\log T-(r'-\delta)} + 2^{n+m-\frac{(r'-\delta)}{2}}\cdot \epsilon} + 2^{-\delta}$$

6 Applications

In this section we assume that Theorem 2 can be realized efficiently, that is: Let $A_x, A_y \in \{0,1\}^n$ of size $|A_x| = |A_y| = 2^{n-d}$ be sets from Theorem 2. We will assume we have efficient maps $\mu_x, \mu_y : \{0,1\}^{n-d} \to \{0,1\}^n$ such that $\mu_x(\{0,1\}^{n-d}) = A_x$ and $\mu_y(\{0,1\}^{n-d}) = A_y$.

6.1 Leakage-resilient storage

In [DDV10] Davi, Dziembowski and Venturi introduced a *leakage-resilient storage*. In their construction message msg is encoded as follows:

- let $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ be (k,ϵ)-two-source extractor
- pick $X, Y \in \{0,1\}^n$ independent, uniformly random vectors.
- $\mathsf{Enc}(\mathrm{msg}) = (X; Y; \mathrm{msg} + \mathsf{Ext}(X,Y))$

Then X, Y, m' are stored on 3 separate servers. All servers leak information to the adversary, whos goal is to guess msg. Authors argue that as long as the adversary has leaked at most $n-k$ bits from first and second server (even if adversary learns the whole content of third server), the secret message is secure. It is, de facto, possible to leak a whole third state and a whole first (or simmetrically second) as long as adversary leaked at most $n-k$ bits from a second (or symmetrically first) state, the difference follows from the fact that [DDV10] uses a weak notion of extractors instead of a strong one.

Alternative option, used widely in non-malleable codes (see [DKO13; ADL14; ADKO15b; CZ14; ADKO15a; Li16]), is to choose X, Y independent, uniformly random such that $\mathsf{Ext}(X,Y) = \mathrm{msg}$. This way we only need two servers.

If for some reasons we need to use a specific extractor (more on that in next subsection) but we need a better leakage-resilience then we can apply Theorem 2 and improve leakage rate of the code.

More precisely, let $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ be (k,ϵ)-two-source extractor, then

$$\overline{\mathsf{Ext}} : \{0,1\}^{n-d} \times \{0,1\}^{n-d} \to \{0,1\}^m$$

defined as

$$\overline{\mathsf{Ext}}(X,Y) = \mathsf{Ext}(\mu_x(X); \mu_y(Y))$$

is a $(k-d+n, \epsilon')$-flexible extractor. When we use it as a leakage resilient storage, adversary can leak whole Y (or symmetrically X) and $n-k$ bits of information from X (or symmetrically Y) and secret remains secure. Leakage rate of original solution was $\frac{n-k}{n}$, while with $\overline{\mathsf{Ext}}$ it's $\frac{n-k}{n-d}$.

6.2 Non-malleable codes and extractors

In recent work [DNO17] Dottling, Nielsen and Obremski construct a continuous non-malleable code. Its a procedure of storing secret message on multiple servers when each of the servers is tampered independently and in continuous manner. Paper offers two instantiation options with a super-strong NMC (see [AKO16]) or with a non-malleable extractor (see [CZ14; Li16]). The authors of [DNO17] require from the underlying construction to have two main properties

1. to offer some version of non-malleability (only) in very high entropy regimes,
2. to offer leakage resilient storage in low entropy regimes.

The construction from [AKO16] has fairly bad code rate, however it is an excellent leakage-resilient storage.

Constructions from [CZ14; Li16] have good code rates and they fulfill the first condition very well, they have a very bad leakage-resilience parameters.

Because of that while instantiation with nm-extractor gives better code rate it also leads to complications and requires workaround. The bad leakage-parameters influence the security parameter of the final code.

Theorem 2 gives a way to trade non-malleable extraction rate for better leakage-resilience rate.

Corollary 2. *Given* $\mathsf{nmExt} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *which is* (k,ϵ)*-non-malleable extractor, let* $\overline{\mathsf{nmExt}} : \{0,1\}^{n-d} \times \{0,1\}^{n-d} \to \{0,1\}^m$ *be defined as* $\overline{\mathsf{nmExt}}(X,Y) = \mathsf{nmExt}(\mu_x(X);\mu_y(Y))$ *then:*

- $\overline{\mathsf{nmExt}}$ *is* (k,ϵ)*-non-malleable extractor*
- $\overline{\mathsf{nmExt}}$ *is* $(k-d+n,\epsilon')$*-flexible two-source extractor.*

6.3 Lower Bounds on Extractors Seeds

Suppose that we have a collection of functions $h_s : [N] \to [M]$ keyed by $s \in [D]$ such that every two different inputs are hashed into different outputs with probability at most $\delta < 1$. By iterating the pigeon-hole principle, for $t = 1, 2 \ldots$ one can obtain a subset of inputs of size $M^{-t} \cdot N$ such that hashes of every pair collide for at least t keys. Therefore, the key space must be at least

$$D \geqslant \delta^{-1} \cdot \log N \cdot \log^{-1} M$$

Combining this with our theorem we reprove the known bound $d = \log(n-k) + \Omega(\log(1/\epsilon))$ on the seed length of extractors [RT00]

Corollary 3 (Good extractors must have logarithmic seeds). *For every extractor such that* $m \leqslant \log(1/\epsilon)$ *the seed length satisfies* $d > \log(n-k) + \min(\log(\epsilon^{-1}, m))$.

While the original bound gives slightly stronger $d \geqslant \log(n-k) + 2\log(1/\epsilon) - O(1)$, our result is qualitatively the same: the seed space, even for one bit extraction ($m = 1$) needs to be as large as $2^d = \mathrm{poly}(n-k, \epsilon^{-1})$.

7 Conclusion

We have shown that the Leftover Hash Lemma can be (partially) reversed: every extractor needs to be a universal hash family on a large subdomain. We discussed the consequences of this result for non-malleable extractors, non-malleable codes, flexible extractors, and limitations of randomness extraction. An interesting open problem is to give a *constructive* version of our result.

References

[ADKO15a] D. Aggarwal, Y. Dodis, T. Kazana, and M. Obremski. “Leakage-Resilient Non-malleable Codes”. In: *The 47th ACM Symposium on Theory of Computing (STOC)*. 2015.

[ADKO15b] D. Aggarwal, S. Dziembowski, T. Kazana, and M. Obremski. “Leakage-Resilient Non-malleable Codes”. In: *Theory of Cryptography - 12th Theory of Cryptography Conference, TCC 2015, Warsaw, Poland, March 23-25, 2015, Proceedings, Part I*. Ed. by Y. Dodis and J. B. Nielsen. Vol. 9014. Lecture Notes in Computer Science. Springer, 2015, pp. 398–426.

[ADL14] D. Aggarwal, Y. Dodis, and S. Lovett. “Non-malleable Codes from Additive Combinatorics”. In: *STOC*. ACM, 2014.

[AKO16] D. Aggarwal, T. Kazana, and M. Obremski. *Inception Makes Non-malleable Codes Stronger*. Cryptology ePrint Archive, Report Report 2015/1013. http://eprint.iacr.org/. 2016.

[BBCM95] C. H. Bennett, G. Brassard, C. Crépeau, and U. M. Maurer. “Generalized privacy amplification”. In: *IEEE Trans. Information Theory* 41.6 (1995), pp. 1915–1923.

[BHKKR99] J. Black, S. Halevi, H. Krawczyk, T. Krovetz, and P. Rogaway. “UMAC: Fast and Secure Message Authentication”. In: *Advances in Cryptology - CRYPTO ’99, 19th Annual International Cryptology Conference, Santa Barbara, California, USA, August 15-19, 1999, Proceedings*. 1999, pp. 216–233.

[BJKS93] J. Bierbrauer, T. Johansson, G. Kabatianskii, and B. J. M. Smeets. “On Families of Hash Functions via Geometric Codes and Concatenation”. In: *Advances in Cryptology - CRYPTO ’93, 13th Annual International Cryptology Conference, Santa Barbara, California, USA, August 22-26, 1993, Proceedings*. 1993, pp. 331–342.

[CRS12] G. Cohen, R. Raz, and G. Segev. “Non-malleable extractors with short seeds and applications to privacy amplification”. In: *Computational Complexity (CCC), 2012 IEEE 27th Annual Conference on*. IEEE. 2012, pp. 298–308.

[CW79] L. Carter and M. N. Wegman. “Universal Classes of Hash Functions”. In: *J. Comput. Syst. Sci.* 18.2 (1979), pp. 143–154.

[CZ14] E. Chattopadhyay and D. Zuckerman. “Non-malleable Codes in the Constant Split-State Model”. In: *FOCS* (2014).

[DDV10] F. Davì, S. Dziembowski, and D. Venturi. "Leakage-Resilient Storage". In: *SCN*. Ed. by J. A. Garay and R. D. Prisco. Vol. 6280. Lecture Notes in Computer Science. Springer, 2010, pp. 121–137.

[DF12] S. Dziembowski and S. Faust. "Leakage-resilient circuits without computational assumptions". In: *Theory of Cryptography*. Springer, 2012, pp. 230–247.

[DKO13] S. Dziembowski, T. Kazana, and M. Obremski. "Non-malleable codes from two-source extractors". In: *Advances in Cryptology-CRYPTO 2013*. Springer. 2013.

[DLWZ11] Y. Dodis, X. Li, T. D. Wooley, and D. Zuckerman. "Privacy Amplification and Non-malleable Extractors via Character Sums". In: *FOCS*. Ed. by R. Ostrovsky. IEEE, 2011, pp. 668–677.

[DNO17] N. Dottling, J. B. Nielsen, and M. Obremski. *Information Theoretic Continuously Non-Malleable Codes in the Constant Split-State Model*. eprint. https://eprint.iacr.org/2017/357.pdf. 2017.

[DORS08] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith. "Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data". In: *SIAM Journal on Computing* 38.1 (2008), pp. 97–139.

[DRS04] Y. Dodis, L. Reyzin, and A. D. Smith. "Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data". In: *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*. 2004, pp. 523–540.

[DW09] Y. Dodis and D. Wichs. "Non-malleable extractors and symmetric key cryptography from weak secrets". In: *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*. Ed. by M. Mitzenmacher. Bethesda, MD, USA: ACM, 2009, pp. 601–610.

[Hay11] M. Hayashi. "Exponential Decreasing Rate of Leaked Information in Universal Random Privacy Amplification". In: *IEEE Trans. Information Theory* 57.6 (2011), pp. 3989–4001.

[HK97] S. Halevi and H. Krawczyk. "MMH: Software Message Authentication in the Gbit/Second Rates". In: *Fast Software Encryption, 4th International Workshop, FSE '97, Haifa, Israel, January 20-22, 1997, Proceedings*. 1997, pp. 172–189.

[HP08] H. Handschuh and B. Preneel. "Key-Recovery Attacks on Universal Hash Function Based MAC Algorithms". In: *Advances in Cryptology - CRYPTO 2008, 28th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2008. Proceedings*. 2008, pp. 144–161.

[ILL89] R. Impagliazzo, L. A. Levin, and M. Luby. "Pseudo-random Generation from one-way functions (Extended Abstracts)". In: *Proceedings of the 21st Annual ACM Symposium on Theory of Com-*

puting, May 14-17, 1989, Seattle, Washigton, USA. 1989, pp. 12–24.

[IZ89] R. Impagliazzo and D. Zuckerman. “How to Recycle Random Bits”. In: *30th Annual Symposium on Foundations of Computer Science, Research Triangle Park, North Carolina, USA, 30 October - 1 November 1989*. 1989, pp. 248–253.

[KSV10] H. Karloff, S. Suri, and S. Vassilvitskii. “A Model of Computation for MapReduce”. In: *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA ’10. Austin, Texas: Society for Industrial and Applied Mathematics, 2010, pp. 938–948.

[Li16] X. Li. *Improved Non-Malleable Extractors, Non-Malleable Codes and Independent Source Extractors*. arXiv Archive, arXiv:1608.00127. https://arxiv.org. 2016.

[LLTT05] C.-J. Lee, C.-J. Lu, S.-C. Tsai, and W.-G. Tzeng. “Extracting randomness from multiple independent sources”. In: *Information Theory, IEEE Transactions on* 51.6 (2005), pp. 2224–2227.

[LSS12] C. E. Leiserson, T. B. Schardl, and J. Sukha. “Deterministic Parallel Random-number Generation for Dynamic-multithreading Platforms”. In: *SIGPLAN Not.* 47.8 (Feb. 2012), pp. 193–204.

[Nis92] N. Nisan. “Pseudorandom generators for space-bounded computation”. In: *Combinatorica* 12.4 (1992), pp. 449–461.

[RT00] J. Radhakrishnan and A. Ta-Shma. “Bounds for Dispersers, Extractors, and Depth-Two Superconcentrators”. In: *SIAM J. Discrete Math.* 13.1 (2000), pp. 2–24.

[RW05] R. Renner and S. Wolf. “Simple and Tight Bounds for Information Reconciliation and Privacy Amplification”. In: *Advances in Cryptology - ASIACRYPT 2005, 11th International Conference on the Theory and Application of Cryptology and Information Security, Chennai, India, December 4-8, 2005, Proceedings*. 2005, pp. 199–216.

[Sie04] A. Siegel. “On Universal Classes of Extremely Random Constant-Time Hash Functions”. In: *SIAM J. Comput.* 33.3 (2004), pp. 505–543.

[Sip83] M. Sipser. “A Complexity Theoretic Approach to Randomness”. In: *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*. STOC ’83. New York, NY, USA: ACM, 1983, pp. 330–335.

[Sti95] D. R. Stinson. “On the Connections Between Universal Hashing, Combinatorial Designs and Error-Correcting Codes”. In: *Electronic Colloquium on Computational Complexity (ECCC)* 2.52 (1995).

[TV15] H. Tyagi and A. Vardy. “Universal Hashing for Information-Theoretic Security”. In: *Proceedings of the IEEE* 103.10 (2015), pp. 1781–1795.

[WC81] M. N. Wegman and L. Carter. “New Hash Functions and Their Use in Authentication and Set Equality”. In: *J. Comput. Syst. Sci.* 22.3 (1981), pp. 265–279.

[ÖP03] A. Östlin and R. Pagh. “Uniform hashing in constant time and linear space”. In: *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9-11, 2003, San Diego, CA, USA*. 2003, pp. 622–628.