Santha-Vazirani sources, deterministic condensers and very strong extractors

Dmitry Gavinsky*† Pavel Pudlák*

February 21, 2020

Abstract

The notion of *semi-random sources*, also known as *Santha-Vazirani (SV)* sources, stands for a sequence of n bits, where the dependence of the i'th bit on the *previous* $i-1$ *bits* is limited for every $i \in [n]$. If the dependence of the i'th bit on the *remaining* $n-1$ *bits* is limited, then this is a *strong SV-source*. Even the strong SV-sources are known not to admit (universal) *deterministic extractors*, but they have *seeded extractors*, as their min-entropy is $\Omega(n)$.

It is intuitively obvious that strong SV-sources are *more than just high-min-entropy sources*, and this work explores the intuition. *Deterministic condensers* are known not to exist for general high-min-entropy sources, and we construct for any constants $\varepsilon, \delta \in (0,1)$ a deterministic condenser that maps n bits coming from a strong SV-source with *bias* at most δ to $\Omega(n)$ bits of *min-entropy rate* at least $1-\varepsilon$.

In conclusion we observe that deterministic condensers are closely related to *very strong extractors* – a proposed strengthening of the notion of *strong (seeded) extractors*: in particular, our constructions can be viewed as *very strong extractors for the family of strong Santha-Vazirani distributions*. The notion of very strong extractors requires that the output remains unpredictable even to someone who knows not only the seed value (as in the case of strong extractors), but also the extractor's outputs corresponding to the same input value with each of the preceding seed values (say, under the lexicographic ordering). Very strong extractors closely resemble the original notion of SV-sources, except that the bits must satisfy the unpredictability requirement only on average.

*Institute of Mathematics of the Czech Academy of Sciences, Žitna 25, Praha 1, Czech Republic.

Partially funded by the grant 19-27871X of GA ČR.

†Part of this work was done while visiting the Centre for Quantum Technologies at the National University of Singapore, and was partially supported by the Singapore National Research Foundation, the Prime Minister's Office and the Ministry of Education under the Research Centres of Excellence programme under grant R 710-000-012-135.

1 Introduction

According to the principles of quantum mechanics, perfectly unbiased and independent random bits can be generated in a physical experiment; however, the imperfectness of practical implementations makes it impossible to deduce from the postulates of quantum mechanics perfect independence of the generated bit sequences. Another possibility is using various chaotic system as a source of of random bits, but in this case it also remains unclear whether perfect (or arbitrarily close to such) independence can be claimed.

The natural question then is: can we reduce the bias and the dependence to a negligible minimum by post-processing the generated bits, coming from a non-perfect source? The post-processing must, of course, be done by a deterministic algorithm. The answer, surprisingly, is that this is not possible (at least in some models of the real situation).

In 1986 Santha and Vazirani [SV86] have defined and studied the notion of *semi-random sources*, now known as SV-sources. These are sequences of n bits $X_1, \ldots, X_n$, whose values cannot be accurately predicted in the following sense: for some $\delta \in [0, 1)$ and any $z \in \{0,1\}^{i-1}$ for $i \in [n]$, it holds that

$$\frac{1-\delta}{2} \leq \boldsymbol{Pr}\big[X_i = 1 \big| X_1 = z_1, \ldots, X_{i-1} = z_{i-1}\big] \leq \frac{1+\delta}{2}.$$

If $\delta \in (0, 1)$ is fixed, we will denote the class of such sources by SV_δ.

In 2004 Reingold, Vadhan and Wigderson [RVW04] defined an even stronger class of entropic bits, which they called *strong Santha-Vazirani sources*, where for some $\delta \in [0, 1)$ and any $z \in \{0,1\}^n$:

$$\begin{aligned}\frac{1-\delta}{2} &\leq \boldsymbol{Pr}\big[X_i = 1 \big| X_1 = z_1, \ldots, X_{i-1} = z_{i-1}, X_{i+1} = z_{i+1}, X_n = z_n\big] \\ &\leq \frac{1+\delta}{2}\end{aligned}$$

for any $i \in [n]$. Similarly, for a fixed $\delta \in (0, 1)$, we will call such sources strong SV_δ.

A *deterministic extractor* is a function that maps n bits to (at least) 1 bit that is nearly-unbiased, as long as the input bits are coming from the corresponding type of entropy source. Santha and Vazirani demonstrated [SV86] that SV-sources did not admit a (universal) deterministic extractor. Later Reingold, Vadhan and Wigderson [RVW04] generalized this impossibility result to the case of strong SV-sources.

It is well known, on the other hand, that *seeded extractors* (see Sect. 2) exist for the class of bit sources whose min-entropy is $\Omega(n)$; since the SV-sources, obviously, belong to that class,[1] seeded extractors exist, in particular, for them.

[1] Throughout the work we will assume, unless stated otherwise, $\delta \in \Omega(1)$ in the context of SV-sources.

The definition of SV-sources is very natural, so it is both important and interesting to investigate this type of randomness. Intuitively, it is obvious that SV-sources – especially the strong form – are more structured than general sources with the same min-entropy and one should be able to use this fact. This work will explore this intuition.

Deterministic condensers are functions that map n bits to m ($< n$) bits, such that the *min-entropy rate* of the output is higher than the min-entropy rate of the input. (Min-entropy rate is the ratio of the min-entropy to the number of bits.) This is certainly not always possible, for instance, if the input has the maximum min-entropy. Typically we have a class of sources and a lower bound on their min-entropy rate and we want to achieve higher lower bound on the min-entropy of the output distributions. Similarly to deterministic extractors, deterministic condensers do not exist for the class of *all sources* whose min-entropy rate is at least a given number.

In this paper we prove two results about SV and strong SV sources. First we show that for SD_δ sources, there are no non-trivial condensers, which means that in general we cannot improve the min-entropy rate using condensers. On the other hand, we construct for any constants $\varepsilon, \delta \in (0, 1)$, a deterministic condenser that maps n bits coming from a *strong* SV_δ source to $\Omega(n)$ bits of min-entropy rate at least $1 - \varepsilon$.

As deterministic condensers are somewhat exotic objects (primarily due to their non-existence for the general class of high-min-entropy sources), this work continues by investigating that notion.

The familiar notion of *strong (seeded) extractors* can be strengthened further – we call the new type of distribution-transforming objects *very strong extractors* – here the output must remain unpredictable even to someone who knows not only the seed value (as in the case of strong extractors), but also the extractor's outputs corresponding to the same input value with each of the preceding seed values (see Sect. 2). We show that deterministic condensers can be easily transformed into very strong extractors;[2] therefore, deterministic condensers are "stronger" objects than strong extractors, but "weaker" than deterministic extractors. Via the same transformation, the main construction of this work gives a very strong extractor for the class of strong SV-sources.

2 Preliminaries

For an excellent survey of error correcting codes, see [MS77].

Let $[n] = \{1, \ldots, n\}$ and $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z} \simeq \{0, \ldots, n-1\}$. Let log be base-2 by default.

[2] The reverse transformation is almost possible: the resulting mapping can only guarantee high *entropy rate* of the output (see Sect. 5).

For $x \in \{0,1\}^n$ and $i \in [n]$, we will use both x_i and $x(i)$ to address the i'th bit of x. Let $|x|$ denote the Hamming weight of x. For $y \in \{0,1\}^n$, let $x \oplus y$ denote the bit-wise XOR of the two vectors. For $y \in \{0,1\}^m$, let $x \circ y \in \{0,1\}^{n+m}$ denote the corresponding concatenation.

We will often implicitly assume the arithmetic of $\mathcal{GF}_2$ and its generalization to $\mathcal{GF}_2^n$ (e.g., $x \oplus y$ is the two vectors' sum in the n-dimensional linear space). In the context of $\mathcal{GF}_2^n$ for $i \in [n]$ we will write e_i to denote the i'th unit vector in $\mathcal{GF}_2^n$ and let $e_0 \stackrel{\text{def}}{=} \bar{0} \in \mathcal{GF}_2^n$.

For sets A and B we will write $A \cup B$ to denote the union while implying the sets' disjointness (the notation, especially the indexed version "$\bigcup_{i=1}^{n} \dots$", is a convenient way of addressing partitions).

For a non-empty finite set A we will denote by $\mathcal{U}_A$ the uniform distribution on A. Let μ and ν be distributions on A, we will say that they are *ε-close* if the l_1-distance between them is at most 2ε.

Let X be a random variable, then

$$H_{\min}(\mu) = \underset{X\sim\mu}{H_{\min}}(X) \stackrel{\text{def}}{=} \min\left\{ \log\left(\frac{1}{\mu(a)}\right) \middle| a \in A \right\}$$

is the *min-entropy* of μ and

$$H(\mu) = \underset{X\sim\mu}{H}(X) \stackrel{\text{def}}{=} \sum_{a\in A} \mu(a) \cdot \log\left(\frac{1}{\mu(a)}\right)$$

is the *entropy* of μ. If $A = \{0,1\}^n$, then $\frac{H_{\min}(\mu)}{n}$ is the *min-entropy rate* and $\frac{H(\mu)}{n}$ is the *entropy rate* of μ.

It is intuitively obvious that a distribution is close to the uniform on its support if and only if the entropy of the distribution is close to the maximum (the logarithm of its support size). The following statement formalizes this intuition.

Fact 1. *Let μ be a distribution supported on A. Then*

$$\begin{aligned}\frac{\log e}{2} \cdot \|\mu - \mathcal{U}_A\|_1^2 \;&\le\; \log(|A|) - H(\mu)\\ &\le\; \frac{1}{2} \cdot \|\mu - \mathcal{U}_A\|_1 \cdot \log(|A|) + \sqrt{2\log e \cdot \|\mu - \mathcal{U}_A\|_1 \cdot \log(|A|)}.\end{aligned}$$

Proof. Let $d \stackrel{\text{def}}{=} \|\mu - \mathcal{U}_A\|_1$. By Pinsker's inequality, see [Pin60],

$$\begin{aligned}\frac{\log e}{2} \cdot d^2 \le d_{KL}\big(\mu\big\|\mathcal{U}_A\big) &= \sum_{a\in A} \mu(a) \cdot \log\left(\frac{\mu(a)}{\mathcal{U}_A(a)}\right)\\ &= \sum_{a\in A} \mu(a) \cdot \left(\log(|A|) - \log\left(\frac{1}{\mu(a)}\right)\right) = \log(|A|) - H(\mu),\end{aligned}$$

which establishes the desired lower bound on $\log(|A|) - H(\mu)$.

For every $t \geq 0$ let

$$B_t \stackrel{\text{def}}{=} \left\{ x \in A \middle| \mu(x) > \frac{1+t}{|A|} \right\}, \tag{1}$$

then

$$\begin{aligned} \mathcal{U}_A(B_t) \leq \frac{\mu(B_t)}{1+t} &\implies \frac{d}{2} \geq \mu(B_t) - \mathcal{U}_A(B_t) \geq \frac{t}{1+t} \cdot \mu(B_t) \qquad (2) \\ &\implies \mu(B_t) \leq \frac{d}{2} + \frac{d}{2t}. \end{aligned}$$

So,

$$\begin{aligned} H(\mu) &\geq \sum_{x \notin B_t} \mu(x) \cdot \log\left(\frac{1}{\mu(x)}\right) \geq \mu(A \setminus B_t) \cdot \log\left(\frac{|A|}{1+t}\right) \\ &\geq \log(|A|) \cdot \left(1 - \frac{d}{2} - \frac{d}{2t}\right) - \log(1+t) \\ &\geq \log(|A|) - \frac{d \cdot \log(|A|)}{2} - \frac{d \cdot \log(|A|)}{2t} - t \cdot \log e, \end{aligned}$$

where the second inequality is (1), the third is (2), and the last one holds, as $\log(1+t) \leq t \cdot \log e$. Choosing $t = \sqrt{\frac{d \cdot \log(|A|)}{2 \log e}}$ establishes the desired upper bound on $\log(|A|) - H(\mu)$. ■

The following are several families of *discrete distributions* that we will be interested in.

Definition 1 (*Santha-Vazirani sources*)**.** *Let $\delta \in [0,1)$ and $X = X_1, \ldots, X_n$ be a random variable distributed over $\{0,1\}^n$ according to a distribution μ. If*

$$\forall i \in [n],\, z \in \{0,1\}^{i-1} : \underset{\mu}{\boldsymbol{Pr}}\left[X_i = 1 \middle| \bigwedge_{j=1}^{i-1} X_j = z_j \right] \in \left[\frac{1-\delta}{2}, \frac{1+\delta}{2} \right],$$

then we call X a Santha-Vazirani source with bias δ *(SV_δ) and μ a* Santha-Vazirani distribution with bias δ*. If*

$$\forall i \in [n],\, z \in \{0,1\}^n : \underset{\mu}{\boldsymbol{Pr}}\left[X_i = 1 \middle| \bigwedge_{j \in [n] \setminus \{i\}} X_j = z_j \right] \in \left[\frac{1-\delta}{2}, \frac{1+\delta}{2} \right],$$

then we call X a strong Santha-Vazirani source with bias δ *and μ a* strong Santha-Vazirani distribution with bias δ*.*

The following are several types of *distribution transformations* that we will be interested in.

Definition 2 (*Deterministic condensers*)**.** *Let $\mathcal{F}$ be a family of distributions over $\{0,1\}^n$. A function $h : \{0,1\}^n \to \{0,1\}^m$ is a* deterministic k-condenser for $\mathcal{F}$ *if*

$$\underset{X\sim\mu}{H_{min}}(h(X)) \geq k$$

for every $\mu \in \mathcal{F}$.

The min-entropy rate *of the condenser is k/m. The condenser is* non-trivial *if*

$$\frac{k}{m} > \inf\left\{ \frac{H_{min}(\mu)}{n} \middle| \mu \in \mathcal{F} \right\}.$$

In the concluding Section 5 we will consider the *entropy rate* of a condenser, defined as

$$\frac{\inf\{H_{X\sim\mu}(h(X)) | \mu \in \mathcal{F}\}}{m}, \tag{3}$$

which is, obviously, at least as high as the min-entropy rate of the same condenser. An object, analogous to a deterministic condenser, but only guaranteeing certain entropy rate (as opposed to min-entropy) will be called *entropy condenser*.

The following statement must be folklore.

Fact 2. *Let $n \geq m \in \mathbb{N}$ and $l \in [0,n]$. No non-trivial deterministic condenser from $\{0,1\}^n$ to $\{0,1\}^m$ exists for the family of all distributions whose min-entropy is at least l.*

In Theorem 1 below we will prove the non-existence of non-trivial condenser even for a restricted class of distributions whose min-entropy rate is at least some bound l, namely for the distributions of SV_δ sources (where $l = \frac{2}{1+\delta}$).

Definition 3 (*Deterministic extractors*)**.** *Let $\varepsilon \in [0,1]$ and $\mathcal{F}$ be a family of distributions over $\{0,1\}^n$. A function $h : \{0,1\}^n \to \{0,1\}^m$ is a* deterministic ε-extractor for $\mathcal{F}$ *if the distribution of $h(X)$ is ε-close to $\mathcal{U}_{\{0,1\}^m}$ when $X \sim \mu$ for any $\mu \in \mathcal{F}$.*

It is well-known (and follows from Fact 2) that no non-trivial deterministic extractor exists for the family of distributions whose min-entropy is at least l. It was shown in [RVW04] that no non-trivial deterministic extractor exists for the family of strong SV-sources even for $m = 1$, which implies that no such extractors exist for any m. On the other hand, there are known constructions of *seeded extractors* for the family of high-min-entropy distributions.

Definition 4 (*Seeded extractors*)**.** *Let $\varepsilon \in [0,1]$ and $\mathcal{F}$ be a family of distributions over $\{0,1\}^n$.*

A function $h : \{0,1\}^n \times [D] \to \{0,1\}^m$ is a seeded ε-extractor for $\mathcal{F}$ *if the distribution of $h(X,S)$ is ε-close to $\mathcal{U}_{\{0,1\}^m}$ when $S \sim \mathcal{U}_{[D]}$ and $X \sim \mu$ for any $\mu \in \mathcal{F}$.*

Of the following two versions that strengthen the notion of seeded extractors, the first is standard and the second is, to the best of our knowledge, new.

Definition 5 (*Strong and very strong seeded extractors*)**.** *Let $\varepsilon \in [0,1]$, $\mathcal{F}$ be a family of distributions over $\{0,1\}^n$ and $h : \{0,1\}^n \times [D] \to \{0,1\}^m$.*

For $s \in [D]$ let ν^{μ}_{s} be the distribution of $h(X,s)$ when $X \sim \mu$ – that is, the distribution of $h(X,S)$, conditioned on $[S = s]$. For $y_1, \ldots, y_{s-1} \in \{0,1\}^m$ let $\nu^{\mu}_{s,y_1,\ldots,y_{s-1}}$ be the distribution of $h(X,S)$ when $X \sim \mu$, conditioned on $[S = s, h(X,1) = y_1, \ldots, h(X,s-1) = y_{s-1}]$ (let it be undefined if the conditioning is inconsistent).

We call h a strong ε-extractor for $\mathcal{F}$ *if*

$$\mathop{\boldsymbol{E}}_{T \sim \mathcal{U}_{[D]}} \left[\left\| \nu^{\mu}_{T} - \mathcal{U}_{\{0,1\}^m} \right\|_1 \right] \leq 2\varepsilon.$$

We call h a very strong ε-extractor for $\mathcal{F}$ *if*

$$\mathop{\boldsymbol{E}}_{T \sim \mathcal{U}_{[D]},\, Z \sim \mu} \left[\left\| \nu^{\mu}_{T,h(Z,1),\ldots,h(Z,T-1)} - \mathcal{U}_{\{0,1\}^m} \right\|_1 \right] \leq 2\varepsilon.$$

In other words, a seeded extractor is *strong* if its output $h(X,S)$ remains unpredictable even when the seed value S is known, and it is *very strong* if the output remains unpredictable even when the seed value S, as well as the outputs corresponding to the preceding seed values $(h(X,1), \ldots, h(X,S-1))$ are known.

The concept of very strong extractors somewhat resembles the original notion of SV-sources: if we write down the sequence $(h(X,1), \ldots, h(X,D))$, then *most* (not necessarily all, as would be the case for SV-sources) of the blocks will be sufficiently unpredictable, even conditioned on the previous blocks. Very strong extractors will be compared to deterministic condensers in the concluding part of this work (Sect. 5).

3 No condensers for SV sources

It [SV86] Santha and Vazirani proved that in general it is not possible to extract a bit from SV_δ sources that is biased less than δ. Later it was shown [RVW04] that this is also true for strong SV_δ-sources. This implies that there is no extractor that would produce a distribution that is guaranteed

to have the statistical distance from uniform less than δ. In other words, there are no ϵ-extractors for strong SV_δ sources with $\epsilon < \delta$.

In this section we prove another impossibility result for SV-sources (we will see in Section 4 that the same is not true for strong SV_δ-sources).

Theorem 1. *There is no non-trivial min-entropy condenser for SV_δ-sources for any $0 < \delta < 1$. Namely,*

1. *the min-entropy-rate of any SV_δ-source is at least $\log \frac{2}{1+\delta}$, and*
2. *for every $F : \{0,1\}^n \to \{0,1\}^m$, there exists an SV_δ-source X, such that the min-entropy-rate of $F(X)$ is at most $\log \frac{2}{1+\delta}$.*

Proof of Theorem 1. 1. Let X be an SV_δ-source. Then by definition,

$$\boldsymbol{Pr}\big[X_i = a_i \big| X_1 = a_1, \ldots, X_{i-1} = a_{i-1}\big] \leq \frac{1+\delta}{2}$$

for every i and $a_1, \ldots, a_{i-1}, a_i \in \{0,1\}$. Hence

$$\boldsymbol{Pr}[X = a] = \prod_{i=1}^{n} \boldsymbol{Pr}\big[X_i = a_i \big| X_1 = a_1, \ldots, X_{i-1} = a_{i-1}\big] \leq \left(\frac{1+\delta}{2}\right)^n$$

for every $a \in \{0,1\}^n$, which proves that the min-entropy rate of X is $\geq \frac{2}{1+\delta}$.

2. To prove the second claim, we need the following lemma.

Lemma 1. *For every $\delta \in (0,1)$ and any non-empty $A \subseteq \{0,1\}^n$, there exists an SV_δ-distribution μ, such that*

$$\mu(A) \geq \left(\frac{1+\delta}{2}\right)^{n - \log|A|}. \tag{4}$$

First we prove 2., assuming that the lemma is valid.

Let s be such that $|F^{-1}(s)| \geq 2^{n-m}$. Then Lemma 1 implies that for some SV_δ-source X it holds that

$$\boldsymbol{Pr}[F(X) = s] \geq \left(\frac{1+\delta}{2}\right)^{n-(n-m)} = \left(\frac{1+\delta}{2}\right)^m.$$

Hence

$$H_\infty[F(X)] \leq m \cdot \log \frac{2}{1+\delta},$$

as required. ■ *Theorem 1*

Now let us prove the lemma.

Proof of Lemma 1. Denote by $p = \frac{1-\delta}{2}$, $q = \frac{1+\delta}{2}$. The idea is simple: put as much weight as possible to A. So we define an SD_δ source X by

$$\begin{aligned} \boldsymbol{Pr}\left[X_i = 0 \middle| (X_1, \ldots, X_{i-1}) = u\right] &= q \quad \text{if } |\{v \mid u0v \in A\}| \geq |\{v \mid u1v \in A\}| \\ &= p \quad \text{otherwise.} \end{aligned}$$

We will prove (4) by induction on n. The base case is clear.

If all $a \in A$ start with 0 or all start with 1, the induction step is trivial. So suppose that this is not the case. Let A_0, respectively A_1, be those that start with 0, respectively with 1. Assume without loss of generality that $|A_0| \geq |A_1|$. We define X so that $\boldsymbol{Pr}[X_1 = 0] = q$ and for the conditional probabilities we will use the two sources that maximize the probabilities that $0v \in A_0$ and $1v \in A_1$. If we denote by $a = |A_0|$ and $b = |A_1|$, then for the induction step, it suffices to prove the following inequality

$$q \cdot q^{n-\log a} + p \cdot q^{n-\log b} \geq q^{n+1-\log(a+b)}. \tag{5}$$

This is equivalent to

$$q \cdot a^{-\log q} + p \cdot b^{-\log q} \geq q \cdot (a+b)^{-\log q}. \tag{6}$$

Let a be fixed and consider the function

$$f(x) := q \cdot a^{-\log q} + p \cdot x^{-\log q} - q \cdot (a+x)^{-\log q}$$

in the domain $0 \leq x \leq a$. We need to prove that $f(x)$ is non-negative in this domain. To this end, it suffices to prove:

1. $f(0) = 0$, which is immediate,
2. $f(a) = (q+p)a^{-\log q} - q(2a)^{-\log q} = a^{-\log q} - q2^{-\log q}a^{-\log q} = 0$,
3. $f'(0) > 0$ or $f'(a) < 0$,
4. $f'(x)$ has a unique root.

Concerning 3., both are true, but it suffices to prove one of these inequalities. We will check the second one.

$$\begin{aligned} f'(a) = &\quad p(-\log q)a^{-\log q - 1} - q(-\log q)(2a)^{-\log q - 1} \\ = &\quad (-\log q)a^{-\log q - 1}(p - q \cdot 2^{-\log q - 1}) \\ = &\quad (-\log q)a^{-\log q - 1}(p - \tfrac{1}{2}) < 0, \end{aligned}$$

because $-\log q > 0$, as well as $a^{-\log q - 1} > 0$, and $p < \frac{1}{2}$.

Concerning 4.,

$$\begin{aligned} & f'(x) = 0 && \Leftrightarrow \\ & p(-\log q)x^{-\log q-1} = q(-\log q)(a+x)^{-\log q-1} && \Leftrightarrow \\ & p^{\frac{1}{-\log q-1}}x = q^{\frac{1}{-\log q-1}}(a+x) && \Leftrightarrow \\ & (p^{\frac{1}{-\log q-1}} - q^{\frac{1}{-\log q-1}})x = q^{\frac{1}{-\log q-1}}a. \end{aligned}$$

Since $p \neq q$, the coefficient at x is non-zero and consequently the equation has a unique solution. This finishes the proof of the inequality (5) which was needed for the induction step. ■*Lemma 1*

4 A condenser for strong SV sources

In this section we will show that in contrast to the standard SV sources, non-trivial deterministic condensers do exist for *strong* SV sources.

4.1 The construction

In this part we construct a family of functions, which will be shown to act as deterministic condensers for strong Santha-Vazirani sources in the next part.

The following definition is, essentially, due to [Ham50].

Definition 6 (*Hamming code*)**.** *Let $d \in \mathbb{N}$ and $M_d \in \{0,1\}^{d\times(2^d-1)}$ be the matrix whose columns are the numbers $1,\ldots,2^d-1$ in their binary representation. Then the set*

$$Ham_d \stackrel{\text{def}}{=} \left\{ x \in \{0,1\}^{2^d-1} \middle| M_d \cdot x = \bar{0} \right\}$$

is the Hamming code *of length 2^d-1.*

The Hamming codes are known to have minimum distance 3, which is easy to see:

$$x_1 \neq x_2 \in Ham_d \implies M_d \cdot (x_1 \oplus x_2) = \bar{0} \implies |x_1 \oplus x_2| \geq 3, \qquad (7)$$

as the columns of M_d are linearly independent in $\mathcal{G}F_2^d$.

For $i \in \mathbb{Z}_{2^d}$, let

$$Ham_d^i \stackrel{\text{def}}{=} \left\{ x \oplus e_i \middle| x \in Ham_d \right\}. \qquad (8)$$

From (7) it follows that these 2^d sets are pairwise disjoint. As $|Ham_d| = 2^{2^d-1-d} = 2^{2^d-1}/2^d$,

$$\{0,1\}^{2^d-1} = \bigcupdot_{i=0}^{2^d-1} Ham_d^i. \tag{9}$$

This is the well-known fact that Hamming codes are *perfect*. Let

$$g_d : \{0,1\}^{2^d-1} \to \{0,1\}^d \tag{10}$$

point to the equivalence class of its argument: namely, $g_d(x)$ is the *binary representation* of such i that $x \in Ham_d^i$; from (9) it follows that g_d is well-defined.

Let $n = k\cdot(2^d-1)$ be a multiple of 2^d-1, define $f_d : \{0,1\}^n \to \{0,1\}^{\frac{d}{2^d-1}\cdot n}$ as follows:

$$f_d(y_1, \ldots, y_k) \stackrel{\text{def}}{=} g_d(y_1) \circ \cdots \circ g_d(y_k), \tag{11}$$

where every y_i is a block of length $2^d - 1$.

4.2 Analysis

Following [RVW04], we will call a distribution ν *δ-imbalanced* if for every x, y,

$$\boldsymbol{Pr}[X = x] \le \frac{1+\delta}{1-\delta} \cdot \boldsymbol{Pr}[Y = y].$$

Lemma 2. *If X is a strong SV_δ source, then the distribution of $g_d(X)$ is δ imbalanced.*

Proof. As $X \sim \mu$ is a strong SV_δ-source (see Def. 1), for all $x \in \{0,1\}^{2^d-1}$ and $i \in [2^d - 1]$,

$$\frac{\mu(x \oplus e_i)}{\mu(x)} \le \frac{1+\delta}{1-\delta}. \tag{12}$$

Since the Hamming code is a perfect code of distance 3, for every $i \neq j$, there exists k such that $e_i + e_j + e_k \in Ham_d$. Hence

$$Ham_d{}^i = Ham_d{}^j + e_k. \tag{13}$$

Now we can write for $u_1 \in Ham_d{}^i, u_2 \in Ham_d{}^j$,

$$\begin{aligned} \mu(g_d^{-1}(u_1)) = \sum_{x \in Ham_d{}^i} \mu(x) &\le \sum_{x \in Ham_d{}^i} \frac{1+\delta}{1-\delta} \cdot \mu(x \oplus e_k) \\ &\le \frac{1+\delta}{1-\delta} \cdot \sum_{x \in Ham_d{}^j} \mu(x) = \frac{1+\delta}{1-\delta} \cdot \mu(g_d^{-1}(u_2)), \end{aligned} \tag{14}$$

where the inequality is (12). Thus $g_d(X)$ is δ-imbalanced. ∎

Lemma 3. *Let ν be δ-imbalanced distribution on $\{0,1\}^d$. Then*

$$H_{min}(\nu) \geq d - \log\left(\frac{1+\delta}{1-\delta}\right). \tag{15}$$

Proof. As $\sum_{u\in\{0,1\}^d} \nu(u) = 1$, there exists some $u_0 \in \{0,1\}^d$, such that $\nu(u_0) \leq 2^{-d}$. Since ν is δ-imbalanced, for all $u \in \{0,1\}^d$,

$$\nu(u) \leq \frac{1+\delta}{1-\delta} \cdot \nu(u_0) \leq \frac{1+\delta}{1-\delta} \cdot 2^{-d},$$

which proves (15). ■

From the two, lemmas we get

$$\underset{X\sim\mu}{H_{\min}}(g_d(X)) \geq d - \log\left(\frac{1+\delta}{1-\delta}\right). \tag{16}$$

Now let $n = k \cdot (2^d - 1)$ for $k \in \mathbb{N}$ and $Y = (Y_1, \ldots, Y_k) \in \{0,1\}^n$ be sampled according to a strong SV_δ-distribution ν (every Y_i consists of $2^d - 1$ bits). We want to analyze the resulting distribution of $f_d(Y) \in \{0,1\}^{k\cdot d}$.

Fix any $i \in [k]$ and $y_1, \ldots, y_i \in \{0,1\}^{2^d-1}$. By the definition of (strong) SV-distributions (Def. 1), the distribution of Y_i remains strong SV_δ when conditioned upon $[Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}]$; accordingly, from (16) it follows that

$$\underset{Y\sim\nu}{\boldsymbol{Pr}}\left[Y_i = y_i \middle| Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}\right] \leq \frac{1+\delta}{1-\delta} \cdot 2^{-d}.$$

Via the trivial induction this implies

$$\underset{Y\sim\nu}{H_{\min}}(f_d(Y)) \geq kd - k \cdot \log\left(\frac{1+\delta}{1-\delta}\right). \tag{17}$$

Therefore, f_d is a deterministic condenser for the family of strong SV_δ-distributions, whose min-entropy rate is lower-bounded by

$$1 - \frac{1}{d} \cdot \log\left(\frac{1+\delta}{1-\delta}\right).$$

Since the min-entropy of a strong SV_δ-distribution over $\{0,1\}^n$ can be as low as $n \cdot \log\left(\frac{2}{1+\delta}\right)$ (as witnessed by the mutually independent distribution of n bits, each taking value "1" with probability $\frac{1+\delta}{2}$), the min-entropy rate of a strong SV_δ-distribution over $\{0,1\}^n$ can be as low as $\log\left(\frac{2}{1+\delta}\right)$ and therefore, f_d is a *non-trivial deterministic condenser for strong SV_δ-distributions* as long as $\delta \in (0,1)$.

We have established the following (via an explicit construction).

Theorem 2. *Let $d \in \mathbb{N}$ and n be a multiple of $2^d - 1$. A deterministic condenser for strong SV-distributions exists that maps n bits to $\frac{n \cdot d}{2^d - 1}$ bits and when the input distribution is a strong SV_δ for $\delta \in [0, 1)$ (whose min-entropy rate can be as low as $\log\left(\frac{2}{1+\delta}\right)$), the generated min-entropy rate is at least*

$$1 - \frac{1}{d} \cdot \log\left(\frac{1+\delta}{1-\delta}\right).$$

In particular, for any constants $\varepsilon, \delta \in (0, 1)$, a deterministic condenser for strong SV_δ-distributions exists that maps n bits to $\Omega(n)$ bits of min-entropy rate at least $1 - \varepsilon$.

5 Deterministic condensers vs. very strong extractors

As *deterministic condensers* are known not to exist for the family of high-min-entropy distributions (Fact 2), they are not considered in the literature very often. We conclude this work by discussing these elegant and natural objects.

Very strong extractors from deterministic condensers. Recall the notion of *entropy rate* (as opposed to *min*-entropy rate) of a condenser, given by (3).

Claim 1. *Let $\varepsilon \in [0, 1]$, $m < n$, D be such that $D|m$ and $\frac{\ln 2}{2}\varepsilon \leq \frac{D}{m}$. Let $\mathcal{F}$ be a family of distributions over $\{0,1\}^n$ and let $h : \{0,1\}^n \to \{0,1\}^m$ be a deterministic entropy condenser for $\mathcal{F}$ of entropy rate at least $1 - \varepsilon$. Then $g : \{0,1\}^n \times [D] \to \{0,1\}^{m/D}$, defined as*

$$g(x, s) = h(x) \restriction_{\frac{(s-1)m}{D}+1, \ldots, \frac{sm}{D}},$$

i.e., $h(x)$ restricted to bits $\frac{(s-1)m}{D} + 1, \ldots, \frac{sm}{D}$, is a very strong δ-extractor for $\mathcal{F}$, where $\delta = \sqrt{\frac{\ln 2}{2}\varepsilon \cdot \frac{m}{D}}$.

In particular, for $D = m$ this gives $g : \{0,1\}^n \times [m] \to \{0,1\}$, which is a very strong $\sqrt{\frac{\ln 2}{2}\varepsilon}$-extractor for $\mathcal{F}$.

As the entropy rate is always at least as high as the min-entropy rate, it follows from the above statement that the construction of Section 4.1, as summarized by Theorem 2, also gives *very strong extractors for the family of strong Santha-Vazirani distributions*, namely:

Corollary 1. *For any constants $\varepsilon, \delta \in (0, 1)$, a very strong single-bit ε-extractor of seed length $\log n - O(1)$ exists for strong SV_δ-distributions.*

Proof of Claim 1. Let $\mu \in \mathcal{F}$. Similarly to Definition 5, for every $s \in [D]$ and $y_1, \ldots, y_{s-1} \in \{0,1\}^{m/D}$ let $\nu^{\mu}_{s,y_1,\ldots,y_{s-1}}$ be the distribution of $g(X,S)$ when $X \sim \mu$, conditioned on $[S = s, g(X,1) = y_1, \ldots, g(X,s-1) = y_{s-1}]$ (let it be undefined if the conditioning is inconsistent).

Then

$$\begin{aligned}
&\mathop{\boldsymbol{E}}_{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}} \left[H\left(\nu^{\mu}_{T,g(Z,1),\ldots,g(Z,T-1)}\right)\right] \qquad (18)\\
&\qquad = \mathop{\boldsymbol{E}}_{T\sim\mathcal{U}_{[D]}} \left[\mathop{H}_{X\sim\mu} \left(g(X,T)\middle| g(X,1),\ldots,g(X,T-1)\right)\right]\\
&\qquad = \frac{1}{D}\cdot\sum_{t=1}^{D} \mathop{H}_{X\sim\mu} \left(g(X,t)\middle| g(X,1),\ldots,g(X,t-1)\right)\\
&\qquad = \frac{1}{D}\cdot \mathop{H}_{X\sim\mu}(h(X)) \geq \frac{1}{D}\cdot(1-\varepsilon)\cdot m = (1-\varepsilon)\cdot\frac{m}{D}.
\end{aligned}$$

Accordingly,

$$\begin{aligned}
&\mathop{\boldsymbol{E}}_{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}} \left[\left\| \nu^{\mu}_{T,h(Z,1),\ldots,h(Z,T-1)} - \mathcal{U}_{\{0,1\}^{m/D}} \right\|_1^2 \right]\\
&\qquad \leq 2\ln 2 \cdot \mathop{\boldsymbol{E}}_{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}} \left[\frac{m}{D} - H\left(\nu^{\mu}_{T,g(Z,1),\ldots,g(Z,T-1)}\right)\right]\\
&\qquad \leq 2\ln 2\cdot\varepsilon\cdot\frac{m}{D},
\end{aligned}$$

where the first inequality is Fact 1 and the second one is (18). The result follows from the concavity of square root. ■*Claim 1*

Deterministic *entropy* condensers from very strong extractors. It is not hard to see that a very strong extractor is not necessarily a good deterministic min-entropy condenser: while we require from very strong extractors to behave nearly-uniformly only *on average*, bounding the *min-entropy* of the condenser's output requires that the probability of a most likely (i.e., *worst-case*) output value is not too high.

On the other hand, we have seen that deterministic *entropy* condensers are very strong extractors (Claim 1). It turns out that the connection between *very strong extractors* and *deterministic entropy* (as opposed to min-entropy) *condensers* is two-way:

Claim 2. *Let $\delta \in [0,1]$, $\mathcal{F}$ be a family of distributions over $\{0,1\}^n$ and $h : \{0,1\}^n \times [D] \to \{0,1\}^m$ be a very strong δ-extractor for $\mathcal{F}$. Then*

$$\mathop{H}_{X\sim\mu}(h(X,1),\ldots,h(X,D)) \geq D\cdot m - D\cdot m\cdot\delta - D\cdot\sqrt{4\log e\cdot m\cdot\delta}$$

for every $\mu \in \mathcal{F}$. That is, the entropy rate *of $(h(X,s))_{s=1}^{D}$ is at least $1-\delta-\sqrt{4\log e\cdot\frac{\delta}{m}}$.*

Proof. Similarly to Definition 5, for every $s \in [D]$ and $y_1,\dots,y_{s-1} \in \{0,1\}^m$ let $\nu^{\mu}_{s,y_1,\dots,y_{s-1}}$ be the distribution of $h(X,S)$ when $X \sim \mu$, conditioned on $[S=s, h(X,1)=y_1,\dots,h(X,s-1)=y_{s-1}]$ (let it be undefined if the conditioning is inconsistent).

Then

$$\begin{aligned}
&\underset{X\sim\mu}{H}\left(h(X,1),\dots,h(X,D)\right)\\
&\qquad=\sum_{t=1}^{D}\underset{X\sim\mu}{H}\left(h(X,t)\middle|h(X,1),\dots,h(X,t-1)\right)\\
&\qquad=D\cdot\underset{T\sim\mathcal{U}_{[D]}}{\boldsymbol{E}}\left[\underset{X\sim\mu}{H}\left(h(X,T)\middle|h(X,1),\dots,h(X,T-1)\right)\right]\\
&\qquad=D\cdot\underset{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}{\boldsymbol{E}}\left[H\left(\nu^{\mu}_{T,h(Z,1),\dots,h(Z,T-1)}\right)\right]\\
&\qquad=Dm-D\cdot\underset{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}{\boldsymbol{E}}\left[m-H\left(\nu^{\mu}_{T,h(Z,1),\dots,h(Z,T-1)}\right)\right]. \qquad (19)
\end{aligned}$$

By Claim 1,

$$\begin{aligned}
&\underset{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}{\boldsymbol{E}}\left[m-H\left(\nu^{\mu}_{T,h(Z,1),\dots,h(Z,T-1)}\right)\right]\\
&\quad\le\frac{m}{2}\cdot\underset{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}{\boldsymbol{E}}\left[\left\|\nu^{\mu}_{T,h(Z,1),\dots,h(Z,T-1)}-\mathcal{U}_{\{0,1\}^m}\right\|_1\right]\\
&\qquad+\sqrt{2\log e\cdot m}\cdot\underset{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}{\boldsymbol{E}}\left[\sqrt{\left\|\nu^{\mu}_{T,h(Z,1),\dots,h(Z,T-1)}-\mathcal{U}_{\{0,1\}^m}\right\|_1}\right]\\
&\quad\le\frac{m}{2}\cdot\underset{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}{\boldsymbol{E}}\left[\left\|\nu^{\mu}_{T,h(Z,1),\dots,h(Z,T-1)}-\mathcal{U}_{\{0,1\}^m}\right\|_1\right]\\
&\qquad+\sqrt{2\log e\cdot m\cdot\underset{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}{\boldsymbol{E}}\left[\left\|\nu^{\mu}_{T,h(Z,1),\dots,h(Z,T-1)}-\mathcal{U}_{\{0,1\}^m}\right\|_1\right]},
\end{aligned}$$

where the latter inequality follows from the concavity of square root. As h is a very strong δ-extractor,

$$\underset{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}{\boldsymbol{E}}\left[\left\|\nu^{\mu}_{T,h(Z,1),\dots,h(Z,T-1)}-\mathcal{U}_{\{0,1\}^m}\right\|_1\right]\le 2\delta$$

and

$$\boldsymbol{E}_{\substack{Z\sim\mu;\\ T\sim\mathcal{U}_{[D]}}}\left[m - H\left(\nu^{\mu}_{T,h(Z,1),\ldots,h(Z,T-1)}\right)\right] \leq m\cdot\delta + \sqrt{4\log e\cdot m\cdot\delta}.$$

From (19),

$$\underset{X\sim\mu}{H}\left(h(X,1),\ldots,h(X,D)\right) \geq D\cdot m - D\cdot m\cdot\delta - D\cdot\sqrt{4\log e\cdot m\cdot\delta},$$

and the result follows. ■

Conclusion. We have established the following.

Theorem 3. *Let $\mathcal{F}$ be a family of distributions over $\{0,1\}^n$.*

If $h : \{0,1\}^n \to \{0,1\}^m$ is a deterministic entropy condenser for $\mathcal{F}$ of entropy rate at least $1-\varepsilon$, where $\varepsilon \leq \frac{2D}{\ln 2\cdot m}$ for some $D|m$, then $g_h : \{0,1\}^n \times [D] \to \{0,1\}^{m/D}$, defined as

$$g_h(x,s) = h(x)\restriction_{\frac{(s-1)m}{D}+1,\ldots,\frac{sm}{D}},$$

is a very strong $\sqrt{\frac{\ln 2}{2}\varepsilon\cdot\frac{m}{D}}$-extractor for $\mathcal{F}$.

If $\delta\in[0,1]$ and $g : \{0,1\}^n \times [D] \to \{0,1\}^m$ is a very strong δ-extractor for $\mathcal{F}$, then $h_g : \{0,1\}^n \to \{0,1\}^{D\cdot m}$, defined as

$$h_g(x) = h(X,1)\circ\cdots\circ h(X,D)$$

is a deterministic entropy condenser for $\mathcal{F}$ of entropy rate at least $1-\delta-\sqrt{4\log e\cdot\frac{\delta}{m}}$.

6 Conclusions

We conclude our article with an open problem. We have shown that there is an essential difference between SV-sources and strong SV-sources: for the former, there are no non-trivial condensers, while for the latter we have constructed them. But this only concerns *min-entropy,* so we pose as an open problem the following.

Problem 1. *Does there exist non-trivial* entropy *condensers for SV_δ sources? If so, how much can one condense the entropy of these sources?*

Acknowledgements

We are grateful to anonymous reviewers for a number of very useful suggestions.

References

[Ham50] R. W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal 29*, pages 147–160, 1950.

[MS77] F. J. MacWilliams and N. J. A. Sloane. The Theory of Error-Correcting Codes. *Elsevier-North-Holland*, 1977.

[Pin60] M. S. Pinsker. Информационная устойчивость гауссовских случайных величин и процессов. *Докл. АН СССР 133(1)*, pages 28–30, 1960.

[RVW04] O. Reingold, S. Vadhan, and A. Wigderson. A Note on Extracting Randomness from Santha-Vazirani Sources. *Unpublished*, 2004.

[SV86] M. Santha and U. V. Vazirani. Generating Quasi-Random Sequences from Slightly-Random Sources. *Journal of Computer and System Sciences 33(1)*, pages 75–87, 1986.