Electronic Colloquium on Computational Complexity, Report No. 92 (2021)

A Note on One-way Functions and Sparse Languages

Yanyi Liu
Cornell University
yl2866@cornell.edu

Rafael Pass*
Cornell Tech
rafael@cs.cornell.edu

June 28, 2021

Abstract

We show equivalence between the existence of one-way functions and the existence of a *sparse* language that is hard-on-average w.r.t. some efficiently samplable "high-entropy" distribution. In more detail, the following are equivalent:

- The existentence of a $S(\cdot)$-sparse language L that is hard-on-average with respect to some samplable distribution with Shannon entropy $h(\cdot)$ such that $h(n) - \log(S(n)) \geq 4 \log n$;
- The existentence of a $S(\cdot)$-sparse language $L \in \mathsf{NP}$, that is hard-on-average with respect to some samplable distribution with Shannon entropy $h(\cdot)$ such that $h(n) - \log(S(n)) \geq n/3$;
- The existence of one-way functions.

Our results are insipired by, and generalize, the recent elegant paper by Ilango, Ren and Santhanam (ECCC'21), which presents similar characterizations for concrete sparse languages.

*Supported in part by NSF Award SATC-1704788, NSF Award RI-1703846, AFOSR Award FA9550-18-1-0267, and a JP Morgan Faculty Award. This material is based upon work supported by DARPA under Agreement No. HR00110C0086. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Government or DARPA.

ISSN 1433-8092

1 Introduction

In this note, we consider the long-standing open problem of basing one-way functions (OWF) on the assumption that NP contains a language that is average-case hard with respect to some efficiently samplable distribution.

We take a step towards achieving this goal and demonstrate that the existence of OWF is equivalent to the existence of a *sparse* language that is hard-on-average w.r.t. some efficiently samplable "high-entropy" distribution. In more details, the Shannon entropy of the sampler needs to be just slightly bigger than the logarithm of the density of the language.

Our results are insipired by, and generalize, the recent elegant paper by Ilango, Ren and Santhanam [IRS21], which presents similar characterizations for concrete (sparse) languages; we observe that these results in [IRS21] follow as direct corollaries from our characterization.

Preliminaries We say that a language $L \subset \{0,1\}^*$ is $S(\cdot)$-*sparse* if for all $n \in \mathbb{N}$, $|L_n| \leq S(n)$, where $L_n = |L \cap \{0,1\}^n|$. Given a language L, we abuse of notation and let $L(x) = 1$ iff $x \in L$. For a random variable X, let $H(X) = \mathsf{E}[\log \frac{1}{\Pr[X=x]}]$ denote the Shannon entropy of X.

We say that $\mathcal{D} = \{D_n\}_{n\in\mathbb{N}}$ is an *ensemble* if for all $n \in \mathbb{N}$, D_n is a probability distribution over $\{0,1\}^n$. We say that an ensemble $\mathcal{D} = \{D_n\}_{n\in\mathbb{N}}$ is *samplable* if there exists a probabilistic polynomial-time Turing machine S such that $S(1^n)$ samples D_n; we use the notation $S(1^n; r)$ to denote the algorithm S with randomness fixed to r. We say that an ensemble $\mathcal{D}$ has entropy $h(\cdot)$ if for all sufficiently large $n \in \mathbb{N}$, $H(D_n) \geq h(n)$.

We say that a language $L \subset \{0,1\}^*$ is $\alpha(\cdot)$ *hard-on-average (α-HoA) on an ensemble* $\mathcal{D} = \{D_n\}_{n\in\mathbb{N}}$ if for all probabilistic polynomial-time heuristics $\mathcal{H}$, for all sufficiently large $n \in \mathbb{N}$,

$$\Pr[x \leftarrow D_n : \mathcal{H}(x) = L(x)] < 1 - \alpha(n).$$

We simply say that L is *hard-on-average (HoA)* on $\mathcal{D}$ if for every c, $\alpha(n) = 1/2 - n^{-c}$, L is α-HoA.

Let $f : \{0,1\}^* \rightarrow \{0,1\}^*$ be a polynomial-time computable function. f is said to be a *one-way function (OWF)* if for every PPT algorithm $\mathcal{A}$, there exists a negligible function μ such that for all $n \in \mathbb{N}$,

$$\Pr[x \leftarrow \{0,1\}^n; y = f(x) : A(1^n, y) \in f^{-1}(f(x))] \leq \mu(n)$$

Main Theorem We are now ready to state our main theorem.

Theorem 1.1. *The following are equivalent:*

1. *The existentence of a $S(\cdot)$-sparse language L that is $(\frac{1}{2} - \frac{1}{4n})$-HoA with respect to some samplable distribution with Shannon entropy $h(\cdot)$ such that $h(n) - \log(S(n)) \geq 4 \log n$;*

2. *The existentence of a $S(\cdot)$-sparse language $L \in \mathsf{NP}$, that is HoA with respect to some samplable distribution with Shannon entropy $h(\cdot)$ such that $h(n) - \log(S(n)) \geq n/3$;*

3. *the existence of one-way functions.*

Theorem 1.1 is proven by, in Section 2 showing that (1) implies (3), and in Section 3 showing that (3) implies (2); the fact that (2) implies (1) is trivial. We finally present some corollaries of Theorem 1.1 in Section 4.

2 OWFs from Avg-case Hardness of Sparse Languages

Theorem 2.1. *Let $S(\cdot)$ be a function, let $h(n) \geq \log S(n) + 4\log n$, and let L be a $S(\cdot)$-sparse language. Assume there exists some samplable ensemble $\mathcal{D}$ with entropy $h(\cdot)$ such that L is $(\frac{1}{2} - \frac{1}{4n})$-HoA on $\mathcal{D}$. Then, one-way functions exist.*

Before proving the theorem, we will state some useful lemmas.

Lemma 2.1 (Implicit in [LP20, IRS21])**.** *Let D_n be a distribution over $\{0,1\}^n$ with entropy at least h. Then, with probability at least $\frac{1}{n}$ over $x \leftarrow D_n$, it holds that*

$$\Pr[D_n = x] \leq 2^{-h+3}$$

Proof: Assume for contradiction that with probability less than $\frac{1}{n}$ over $x \leftarrow D_n$, $\Pr[D_n = x] \leq 2^{-h+3}$. Let Freq denote the set of strings $x \subseteq \{0,1\}^n$ such that $\Pr[D_n = x] > 2^{-h+3}$, and let Rare denote the set of strings $\subseteq \{0,1\}^n$ such that $\Pr[D_n = x] \leq 2^{-h+3}$. Let flag be a binary random variable such that $\mathsf{flag} = 0$ if $D_n \in \mathsf{Freq}$ and 1 otherwise (i.e. if $D_n \in \mathsf{Rare}$). Let p_{Freq} be the probability that $D_n \in \mathsf{Freq}$ and p_{Rare} be the probability that $D_n \in \mathsf{Rare}$. By the chain rule for entropy, it holds that

$$H(D_n) \leq H(D_n, \mathsf{flag}) = H(\mathsf{flag}) + p_{\mathsf{Freq}} H(D_n \mid D_n \in \mathsf{Freq}) + p_{\mathsf{Rare}} H(D_n \mid D_n \in \mathsf{Rare})$$

In the RHS, the first term is at most 1 (since flag is a binary variable). The second term is at most $h-3$ since $|\mathsf{Freq}| \leq 2^{h-3}$. Recall that by assumption, we have that $p_{\mathsf{Rare}} < \frac{1}{n}$; furthermore, $H(D_n \mid D_n \in \mathsf{Rare}) \leq n$ (since $|\mathsf{Rare}| \leq 2^n$) and thus the last term of the RHS is at most 1. Therefore, $H(D_n) \leq 1 + (h-3) + 1 < h$, which is a contradiction. ■

Lemma 2.2. *Let $L_n \subset \{0,1\}^n$ be a set of strings such that $|L_n| \leq S(n)$. Let D_n be a distribution over $\{0,1\}^n$. Let ε be a constant such that $\varepsilon \leq \frac{1}{S(n)n^2}$. Then, the following holds:*

$$\Pr_{x \leftarrow D_n}[L_n(x) = 1 \wedge \Pr[D_n = x] \leq \varepsilon] \leq \frac{1}{n^2}$$

Proof: By the union bound, it follows that $\Pr_{x \leftarrow D_n}[L_n(x) = 1 \wedge \Pr[D_n = x] \leq \varepsilon]$ is bounded by $S(n) \times \frac{1}{S(n)n^2} = \frac{1}{n^2}$. ■

We will rely on the following important lemma showing that approximate counting can be efficiently done if one-way functions do not exist.

Lemma 2.3 ([IL90, IL89, IRS21])**.** *Assume that one-way functions do not exist. Then, for any samplable ensemble $\mathcal{D} = \{D_n\}_{n \in \mathbb{N}}$ and any constant $q \geq 1$, there exist a PPT algorithm $\mathcal{A}$ and a constant δ such that for infinitely many n,*

$$\Pr_{x \leftarrow D_n}[\delta \cdot p_x \leq \mathcal{A}(x) \leq p_x] \geq 1 - \frac{1}{n^q}$$

where $p_x = \Pr[D_n = x]$.

In addition, we observe that if approximate counting can be done, the Shannon entropy of any samplable distribution $\mathcal{D}$ can be estimated efficiently.

Lemma 2.4. *Let $\mathcal{D} = \{D_n\}_{n\in\mathbb{N}}$ be a samplable ensemble, let* Samp *be the corresponding sampler, and let $m(\cdot)$ be a polynomial such that $m(n)$ is greater than the number of random coins used by* $\mathsf{Samp}(1^n)$*. Assume that there exist a* PPT *algorithm $\mathcal{A}$, a constant δ, and an infinite set $I \subseteq \mathbb{N}$ such that for all $n \in I$,*

$$\Pr_{x\leftarrow D_n}[\delta \cdot p_x \leq \mathcal{A}(x) \leq p_x] \geq 1 - \frac{1}{m(n)}$$

where $p_x = \Pr[D_n = x]$. Then, there exist a PPT *algorithm* est *and a constant δ' such that for all $n \in I$, with probability at least $1 - \frac{1}{n^2}$,*

$$|\mathsf{est}(1^n) - H(D_n)| \leq \delta'$$

Proof: Let $n \in I$ be a sufficiently large input length on which $\mathcal{A}$ succeeds. Let p_x denote $\Pr[D_n = x]$. Let $\mathcal{A}'$ be an algorithm such that $\mathcal{A}'(x) = \max(2^{-m}, \min(1, \mathcal{A}(x)))$. $\mathcal{A}'$ will have the same property that $\mathcal{A}$ has in the assumption since for all x in the support of D_n, it holds that $2^{-m} \leq p_x \leq 1$. We first claim that

$$|\mathsf{E}_{x\leftarrow D_n}[-\log \mathcal{A}'(x)] - H(D_n)| \leq -\log \delta + 1 \tag{1}$$

If this holds, note that $\mathcal{D}$ is samplable and $\mathcal{A}'$ runs in PPT, it follows that we can empirically estimate $\mathsf{E}_{x\leftarrow D_n}[-\log \mathcal{A}'(x)]$ in polynomial time by sampling at least n^6 samples and taking the average. By Hoeffding's inequality, the difference between this estimation and the real expectation value is at most 1 with very high probability ($\geq 1 - \frac{1}{n^2}$).

Thus, it remains to show that inequality 1 holds. Notice that

$$\begin{aligned}
&|\mathsf{E}_{x\leftarrow D_n}[-\log \mathcal{A}'(x)] - H(D_n)| \\
=&|\mathsf{E}_{x\leftarrow D_n}[-\log \mathcal{A}'(x)] - \mathsf{E}_{x\leftarrow D_n}[-\log p_x]| \\
\leq&\mathsf{E}_{x\leftarrow D_n}[|-\log \mathcal{A}'(x) - (-\log p_x)|] \\
=&\Pr_{x\leftarrow D_n}[\mathcal{A}' \text{ succeeds}] \cdot \mathsf{E}_{x\leftarrow D_n}[|-\log \mathcal{A}'(x) - (-\log p_x)| \mid \mathcal{A}' \text{ succeeds}] \\
&+ \Pr_{x\leftarrow D_n}[\mathcal{A}' \text{ fails}] \cdot \mathsf{E}_{x\leftarrow D_n}[|\log \mathcal{A}'(x) - (-\log p_x)| \mid \mathcal{A}' \text{ fails}] \\
\leq&\mathsf{E}_{x\leftarrow D_n}[|\log \frac{p_x}{\mathcal{A}'(x)}| \mid \mathcal{A}' \text{ succeeds}] + \frac{1}{m} \cdot m \\
\leq&\mathsf{E}_{x\leftarrow D_n}[-\log \delta \mid \mathcal{A}' \text{ succeeds}] + 1 \\
\leq&-\log \delta + 1
\end{aligned}$$

■

Now we are ready to prove Theorem 2.1.

Proof: [Proof of Theorem 2.1] Assume for contradiction that one-way functions do not exist. Then, by Lemma 2.3, there exist a PPT algorithm $\mathcal{A}$ and a constant δ such that for infinitely many n,

$$\Pr_{x\leftarrow D_n}[\delta \cdot p_x \leq \mathcal{A}(x) \leq p_x] \geq 1 - \frac{1}{n^2}$$

where $p_x = \Pr[D_n = x]$. By Lemma 2.4, there exist a PPT algorithm est and a constant δ' such that for all n on which $\mathcal{A}$ succeeds, with probability at least $1 - \frac{1}{n^2}$,

$$|\mathsf{est}(1^n) - H(D_n)| \leq \delta' \tag{2}$$

Consider some sufficiently large input length n on which $\mathcal{A}$ succeeds. Let

$$\varepsilon = 2^{-\mathsf{est}(1^n)+\log n}$$

We are now ready to describe our heuristic $\mathcal{H}$ for L. On input $x \leftarrow D_n$, $\mathcal{H}$ computes ε and outputs 0 if $\mathcal{A}(x) \leq \varepsilon$; otherwise, $\mathcal{H}$ outputs a random guess $b \in \{0,1\}$. We will show that $\mathcal{H}$ solves L with probability $\frac{1}{2} + \frac{1}{4n}$ on the input length n (whenever n is sufficiently large).

Towards this, let us first assume we have access to a "perfect" approximate-counter algorithm $\mathcal{O}$ such that $\delta \cdot p_x \leq \mathcal{O}(x) \leq p_x$ with probability 1 when x sampled from D_n; let us also assume we have access to a "perfect" entropy-estimator algorithm est^* such that $|\mathsf{est}^*(1^n) - H(D_n)| \leq \delta'$ with probability 1; consider the heuristic $\mathcal{H}'$ that behaves just as $\mathcal{H}$ except that $\mathcal{H}'$ uses $\mathcal{O}$ and est^* instead of $\mathcal{A}$ and est.

We first show that $\mathcal{H}'$ solves L with high probability on D_n. Note that

$$\begin{aligned}
&\Pr_{x \leftarrow D_n}[\mathcal{H}'(x) = L(x)] \\
=&\Pr_{x \leftarrow D_n}[\mathcal{H}'(x) = L(x) \mid \mathcal{O}(x) > \varepsilon] \Pr[\mathcal{O}(x) > \varepsilon] + \Pr_{x \leftarrow D_n}[\mathcal{H}'(x) = L(x) \mid \mathcal{O}(x) \leq \varepsilon] \Pr[\mathcal{O}(x) \leq \varepsilon] \\
=&\frac{1}{2}\left(1 - \Pr[\mathcal{O}(x) \leq \varepsilon]\right) + \left(1 - \Pr_{x \leftarrow D_n}[\mathcal{H}'(x) \neq L(x) \mid \mathcal{O}(x) \leq \varepsilon]\right) \Pr[\mathcal{O}(x) \leq \varepsilon] \\
=&\frac{1}{2}\left(1 - \Pr[\mathcal{O}(x) \leq \varepsilon]\right) + \left(1 - \Pr_{x \leftarrow D_n}[L(x) = 1 \mid \mathcal{O}(x) \leq \varepsilon]\right) \Pr[\mathcal{O}(x) \leq \varepsilon] \\
=&\frac{1}{2} + \frac{1}{2}\Pr[\mathcal{O}(x) \leq \varepsilon] - \Pr_{x \leftarrow D_n}[L(x) = 1 \mid \mathcal{O}(x) \leq \varepsilon] \Pr[\mathcal{O}(x) \leq \varepsilon] \\
=&\frac{1}{2} + \frac{1}{2}\Pr[\mathcal{O}(x) \leq \varepsilon] - \Pr_{x \leftarrow D_n}[L(x) = 1 \wedge \mathcal{O}(x) \leq \varepsilon]
\end{aligned}$$

Note that $p_x \leq \varepsilon$ implies $\mathcal{O}(x) \leq \varepsilon$ (since $\mathcal{O}$ is a prefect approximate-counter). In addition, for sufficiently large n, $p_x \leq 2^{-H(D_n)+3}$ implies $p_x \leq \varepsilon$ since

$$2^{-H(D_n)+3} \leq 2^{-\mathsf{est}^*(1^n)+\delta'+3} \leq 2^{-\mathsf{est}^*(1^n)+\log n} = \varepsilon.$$

Thus,

$$\Pr[\mathcal{O}(x) \leq \varepsilon] \geq \Pr_{x \leftarrow D_n}[p_x \leq \varepsilon] \geq \Pr_{x \leftarrow D_n}[p_x \leq 2^{-H(D_n)+3}] \geq \frac{1}{n}$$

where the last inequality follows from by Lemma 2.1.

Next, observe that $\varepsilon \leq \frac{1}{S(n)n^2}$ (for sufficiently large n). This follows since if n is sufficiently large, we have:

$$\begin{aligned}
\varepsilon = 2^{-\mathsf{est}^*(1^n)+\log n} \leq 2^{-H(D_n)+\delta'+\log n} &\leq 2^{-H(D_n)+2\log n} \leq 2^{-h(n)+2\log n} \\
&\leq 2^{-\log S(n) - 4\log n + 2\log n} = \frac{1}{S(n)n^2}
\end{aligned}$$

Finally, since $L(x) = 1 \wedge p_x \leq \varepsilon$ implies $L(x) = 1 \wedge \mathcal{O}(x) \leq \varepsilon$, we have that

$$\Pr_{x \leftarrow D_n}[L(x) = 1 \wedge \mathcal{O}(x) \leq \varepsilon] \leq \Pr_{x \leftarrow D_n}[L(x) = 1 \wedge p_x \leq \varepsilon] \leq \frac{1}{n^2}$$

where the last inequality follows from Lemma 2.2 and the fact that $\varepsilon \leq \frac{1}{S(n)n^2}$. Thus, we conclude that

$$\Pr_{x \leftarrow D_n}[\mathcal{H}'(x) = L(x)] \geq \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{n} - \frac{1}{n^2}$$

We now turn to analyzing $\mathcal{H}$ as opposed to $\mathcal{H}'$ and note that $\mathcal{H}$ and $\mathcal{H}'$ work identically the same except when either A or est "fail". Observe that the probability that $\mathcal{A}(x) \neq \mathcal{O}(x)$ on x sampled from D_n is at most $\frac{1}{n^2}$. Additionally, the probability that $|\mathsf{est}(1^n) - H(D_n)| > \delta'$ is at most $\frac{1}{n^2}$. Thus, by a union bound, we have that

$$\Pr_{x \leftarrow D_n}[\mathcal{H}(x) = L(x)] \geq \frac{1}{2} + \frac{1}{2n} - \frac{3}{n^2} \geq \frac{1}{2} + \frac{1}{4n}$$

on infinitely many $n \in \mathbb{N}$, which is a contradiction. ■

3 Avg-case Hardness of Sparse Languages from OWFs

Theorem 3.1. *Assume the existence of one-way functions. Let $S(n) = 2^{n/10}$ and $h(n) = n/2$. Then there exists a $S(\cdot)$-sparse language $L \in \mathsf{NP}$ and a samplable ensemble $\mathcal{D}$ with entropy $h(\cdot)$ such that L is HoA on $\mathcal{D}$.*

Proof: Assume the existence of OWFs. By [HILL99], there exists some pseudorandom generator $g : \{0,1\}^{n/10} \to \{0,1\}^n$. Consider the NP-language $L = \{g(s) \mid s \in \{0,1\}^*\}$. Note that L is $S(\cdot)$-sparse for $S(n) = 2^{n/10}$. Let $\mathcal{D} = \{D_n\}_{n \in \mathbb{N}}$ be an ensemble such that D_n samples from $g(\mathcal{U}_{n/10})$ with probability 1/2 and from $\mathcal{U}_n$ with probability 1/2. Note that $\mathcal{D}$ has entropy at least $h(n) = n/2$ (since with probability 1/2, we sample from U_n). Finally, it follows from the pseudorandomness property of g (using a standard argument) that L is HoA over $\mathcal{D}$. ■

4 Corollaries

In this section, we present some direct corollaries that follow by applying our main theorem to known sparse languages. For convenience of the reader, we recall the (standard) proofs that these languages are sparse.

4.1 Kolmogorov Complexity

The Kolmogorov complexity of a string $x \in \{0,1\}^*$ is defined to be the length of the shortest program Π that outputs the string x. More formally, let U be a fixed Universal Turing machine, for any string $x \in \{0,1\}^*$, we define $K(x) = \min_{\Pi \in \{0,1\}^*}\{|\Pi| : U(\Pi) = x\}$. Let $\mathsf{MINK}[s]$ denote the language of strings x having the property that $K(x) \leq s(|x|)$. We observes that $\mathsf{MINK}[s]$ is a sparse language when $s(n)$ is slightly below n.

Lemma 4.1. *For all $n \in \mathbb{N}$, $|\mathsf{MINK}[s] \cap \{0,1\}^n| \leq 2^{s(n)+1}$.*

Proof: The lemma directly follows from the fact that the number of strings with length $\leq s(n)$ is at most $2^{s(n)+1}$. ■

Combing Lemma 4.1, we get:

Corollary 4.1. *Let $s(n) \leq n - 4\log n - 1$ be a function. Assume that there exists some samplable ensemble $\mathcal{D}$ with entropy $h(n) \geq s(n) + 4\log n + 1$ such that $\mathsf{MINK}[s]$ is $(\frac{1}{2} - \frac{1}{4n})$-HoA on $\mathcal{D}$. Then, one-way functions exist.*

Proof: By Lemma 4.1, the number of n-bit YES instances is at most $S(n) = 2^{s(n)+1}$. Since D_n has entropy $h(n) \geq s(n) + 1 + 4\log n$, the corollary follows directly from Theorem 1.1. ■

4.2 k-SAT

We then show that one-way functions can be based on some average-case Let k, c be two positive integers. The language $k\text{-}\mathsf{SAT}(m, cm)$ is defined to consist of all satisfiable k-CNF formulas on m variables with cm clauses. We recall the well-known fact that $k\text{-}\mathsf{SAT}(m, cm)$ is a sparse language when $c \geq 2^{k+1}$.

Lemma 4.2. *The number of satisfiable k-CNF formulas on m variables with cm clauses is at most $2^m \left((2^k - 1)\binom{m}{k}\right)^{cm}$, and the number of all such k-CNF formulas is $\left((2^k)\binom{m}{k}\right)^{cm}$.*

Proof: We first show that there are $((2^k)\binom{m}{k})^{cm}$ k-CNF formulas on m variables with cm clauses. Note that are are $2^k\binom{m}{k}$ choices for a single k-clause; therefore, the number of cm k-clauses is $((2^k)\binom{m}{k})^{cm}$.

We then show that there are at most $2^m((2^k - 1)\binom{m}{k})^{cm}$ satisfiable k-CNF formulas on m variables with cm clauses. Consider any possible assignment x; the number of k-clauses that is satisfied by x is at most $(2^k - 1)\binom{m}{k}$ since given the choice of k variables, there are at most $2^k - 1$ possible choices of the polarities. Finally, since there are cm such k-clauses with m variables, we have that the total number of satisfiable formulas is at most $2^m((2^k - 1)\binom{m}{k})^{cm}$ ■

To consider average-case hardness of this problem, we need to have a way to encode formulas as strings. We use the following standard encoding scheme for k-SAT from [IRS21]: a m-variable cm-clause k-CNF is represented by using $n(m, k, c) = cm(k\lceil \log m \rceil + k)$ bits to describe a sequence of cm clauses. In each clause, we specify k literals one-by-one, and each of them takes $\lceil \log m \rceil$ bits to specify the index of a variable and 1 bit to fix the polarity. When n is not of the form $n(m, k, c)$, for an input of length n, we ignore as few bits as possible in the end of the input such that the prefix the input is of length $n(m, k, c)$ for some m. Following [IRS21], let the *entropy deficiency* of a distribution D_n over n bits denote the difference between n and $H(D_n)$. The follow corollary implies [IRS21, Theorem 4, Term 1].

Corollary 4.2. *Let k, c be two integers such that $c \geq 2^{k+2}$. Let $m = m(n)$ be a polynomial. Assume that there exists some samplable ensemble $\mathcal{D} = \{D_n\}_{n \in \mathbb{N}}$ with entropy deficiency at most $cm(n)/2^{k+1}$ distributed over k-CNF formulas on $m(n)$ variables and $cm(n)$ clauses such that k-SAT is $(\frac{1}{2} - \frac{1}{4n})$-HoA on $\mathcal{D}$. Then, one-way functions exist.*

Proof: Recall that k-CNF formulas are represented by binary strings using the standard encoding scheme. Let $n' = n(m(n), k, c)$ (be the length of the input without padding); by the encoding scheme, it follows that every $m(n)$-variable $cm(n)$-clause k-CNF formula will be encoded by $2^{n-n'}$ n-bit strings. By Lemma 4.2, it follows that n' is at least

$$\log \left(\left((2^k) \binom{m}{k} \right)^{cm} \right) = cm \log 2^k + cm \log \binom{m}{k}$$

Since D_n has entropy deficiency at most $cm/2^{k+1}$, it follows that D_n has entropy lower bounded by:

$$n' + (n - n') - cm/2^{k+1} \geq cm \left(\log 2^k - \frac{1}{2^{k+1}} + \log \binom{m}{k} \right) + (n - n')$$

By Lemma 4.2, the number of n-bit YES instances is at most

$$S(n) = 2^m \left((2^k - 1) \binom{m}{k} \right)^{cm} \times 2^{n-n'}$$

It follows that

$$\begin{aligned}
H(D_n) - \log S(n) &\geq cm\left(\log 2^k - \frac{1}{2^{k+1}} + \log\binom{m}{k}\right) + (n-n') - \log\left(2^m\left((2^k-1)\binom{m}{k}\right)^{cm} \times 2^{n-n'}\right)\\
&= m(c\log 2^k - c\log(2^k-1) - \frac{c}{2^{k+1}} - 1)\\
&\geq m(\frac{c}{2^k} - \frac{c}{2^{k+1}} - 1)\\
&\geq m\\
&\geq 4\log n.
\end{aligned}$$

where the second inequality follows from the standard inequalty that $\log x - \log(x-1) \geq \frac{1}{x}$ for all $x \geq 2$, the third one from the fact that, by assumption, $c \geq 2^{k+2}$, and the fourth one inequality follows from the fact that due to the encoding scheme, $m \geq \Omega(\sqrt{n})$. ■

4.3 t-Clique

Let $t : \mathbb{N} \to \mathbb{N}$ be a function and let t-Clique(m) be the set of graphs on m vertices having a clique of size at least $t(m)$. We recall the well-known fact that t-Clique(m) is sparse when $t(\cdot)$ is large enough.

Lemma 4.3. *The number of m-vertex graphs with at least a t-size clique is at most $\binom{m}{t}2^{\binom{m}{2}-\binom{t}{2}}$. However, the number of m-vertex graphs is $2^{\binom{m}{2}}$.*

Proof: There are $\binom{m}{2}$ edges in a m-vertex graph, and thus the number of possible graphs is $2^{\binom{m}{2}}$. There are $\binom{m}{t}$ choices of cliques in a graph, and after fixing a clique, there are $\binom{m}{2} - \binom{t}{2}$ edges in the rest of the graph and therefore the number of graphs with at least 1 clique is at most $\binom{m}{t}2^{\binom{m}{2}-\binom{t}{2}}$. ■

We use the standard encoding scheme for t-Clique from [IRS21]. A m-vertex graph is encoded by a $(n = n(m) = \binom{m}{2})$-bit string where the i-th bit is 1 iff the i-th edge appears in the graph. For input lengths n that are not of the form $n(m)$, we ignore as few bits as possible at the end of the input such that the prefix of the input is of length $n(m)$ for some m.

Corollary 4.3. *Let $m(n), t(n) \in \omega(\log m)$ be two polynomials. Assume that there exists some samplable ensemble $\mathcal{D} = \{D_n\}_{n\in\mathbb{N}}$ with entropy deficiency at most $0.99\binom{t(n)}{2}$ distributed over $m(n)$-vertex graphs such that t-Clique(m) is $(\frac{1}{2} - \frac{1}{4n})$-HoA on $\mathcal{D}$. Then, one-way functions exist.*

Proof: Recall that graphs are represented by binary strings using the standard encoding scheme. Let $n' = n(m(n))$ (be the length of the input without padding); by the encoding scheme, it follows that every $m(n)$-vertex graph will be encoded by at least $2^{n-n'}$ n-bit strings. By Lemma 4.3, it follows that n' is at least

$$\log 2^{\binom{m}{2}} = \binom{m}{2}$$

Since D_n has entropy deficiency $0.99\binom{t}{2}$, it follows that D_n has entropy lower bounded by:

$$n' + (n-n') - 0.99\binom{t}{2} \geq \binom{m}{2} - 0.99\binom{t}{2} + (n-n')$$

By Lemma 4.3, the number of n-bit YES instances is at most

$$S(n) = \binom{m}{t}2^{\binom{m}{2}-\binom{t}{2}} \times 2^{n-n'}$$

It follows that

$$\begin{aligned} H(D_n) - \log S(n) &\geq \binom{m}{2} - 0.99\binom{t}{2} + (n-n') - \log\left(\binom{m}{t} 2^{\binom{m}{2}-\binom{t}{2}} \times 2^{n-n'}\right) \\ &\geq \binom{m}{2} - 0.99\binom{t}{2} - \log\binom{m}{t} - \left(\binom{m}{2} - \binom{t}{2}\right) \\ &\geq \binom{t}{2} - 0.99\binom{t}{2} - t\log m \\ &\geq 4\log n \end{aligned}$$

since $t(n) = \omega(\log m)$. ■

References

[HILL99] Johan Håstad, Russell Impagliazzo, Leonid A. Levin, and Michael Luby. A pseudorandom generator from any one-way function. *SIAM J. Comput.*, 28(4):1364–1396, 1999.

[IL89] Russell Impagliazzo and Michael Luby. One-way functions are essential for complexity based cryptography (extended abstract). In *30th Annual Symposium on Foundations of Computer Science, Research Triangle Park, North Carolina, USA, 30 October - 1 November 1989*, pages 230–235, 1989.

[IL90] Russell Impagliazzo and Leonid A. Levin. No better ways to generate hard NP instances than picking uniformly at random. In *31st Annual Symposium on Foundations of Computer Science, St. Louis, Missouri, USA, October 22-24, 1990, Volume II*, pages 812–821, 1990.

[IRS21] Rahul Ilango, Hanlin Ren, and Rahul Santhanam. Hardness on any samplable distribution suffices: New characterizations of one-way functions by meta-complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, (082), 2021.

[LP20] Yanyi Liu and Rafael Pass. On one-way functions and kolmogorov complexity. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pages 1243–1254. IEEE, 2020.