Computational Analogues of Entropy

Boaz Barak* Ronen Shaltiel† Avi Wigderson‡

December 5, 2003

Abstract

Min-entropy is a statistical measure of the amount of randomness that a particular distribution contains. In this paper we investigate the notion of *computational min-entropy* which is the computational analog of statistical min-entropy. We consider three possible definitions for this notion, and show equivalence and separation results for these definitions in various computational models.

We also study whether or not certain properties of statistical min-entropy have a computational analog. In particular, we consider the following questions:

1. Let X be a distribution with high computational min-entropy. Does one get a pseudorandom distribution when applying a "randomness extractor" on X?
2. Let X and Y be (possibly dependent) random variables. Is the computational min-entropy of (X, Y) at least as large as the computational min-entropy of X?
3. Let X be a distribution over $\{0,1\}^n$ that is "weakly unpredictable" in the sense that it is hard to predict a constant fraction of the coordinates of X with a constant bias. Does X have computational min-entropy $\Omega(n)$?

We show that the answers to these questions depend on the computational model considered. In some natural models the answer is false and in others the answer is true. Our positive results for the third question exhibit models in which the "hybrid argument bottleneck" in "moving from a distinguisher to a predictor" can be avoided.

Keywords: Min-Entropy, Pseudorandomness

*Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. Email: boaz@wisdom.weizmann.ac.il

†Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. Email: ronens@wisdom.weizmann.ac.il

‡School of Mathematics, Institute for Advanced Study, Princeton, NJ and Hebrew University, Jerusalem, Israel. Email: avi@ias.edu

Contents

1 Introduction

One of the most fundamental notions in theoretical computer science is that of *computaional indistinuishability* [GM84, Yao82]. Two probability distributions are deemed close if no *efficient*[1] test can tell them apart - this comes in stark contrast to the information theoretic view which allows *any* test whatsoever. The discovery [BM82, Yao82, HILL99] that simple computational assumptions (namely the existance of one-way functions) make the computational and information theoretic notions completely different has been one of the most fruitful in CS history, with impact on cryptography, complexity theory and computational learning theory.

The most striking result of these studies has been the efficient construction of nontrivial *pseudorandom* distributions, namely ones which are information theoretically very far from the uniform distribution, but are nevertheless indistinguishable from it. Two of the founding papers [Yao82, HILL99] found it natural to extend information theory more generally to the computational setting, and attempt to define its most fundamental notion of entropy[2]. The basic question is the following: when should we say that a distribution has (or is close to having) computational entropy (or pseudoentropy) k?. Interestingly, these two papers give two very different definitions! This point may be overlooked, since for the most interesting special case, the case of pseudorandomness (i.e., when the distributions are over n-bit strings and $k = n$), the two definitions coincide. This paper is concerned with the other cases, namely $k < n$, attempting to continue the project of building a computational analog of information theory.

1.1 Definitions of pseudoentropy

To start, let us consider the two original definitions. Let X be a probability distribution over a set S.

A definition using "compression". Yao's definition of pseudoentropy [Yao82] is based on compression. He cites Shannon's definition [Sha48], defining $H(X)$ to be the minimum number of bits needed to describe a typical element of X. More precisely, one imagines the situation of Alice having to send Bob (a large number of) samples from X, and is trying to save on communication. Then $H(X)$ is the smallest k for which there are a compression algorithm A (for Alice) from S into k-bit strings, and a decompression algorithm B (for Bob) from k-bit strings into S, such that $B(A(x)) = x$ (in the limit, for typical x from X). Yao take this definition verbatim, adding the crucial computational constraint that both compression and decompression algorithms must be efficient. This notion of efficient compression is further studied in [GS91].

A definition using indistinguishability. Hastad et al's definition of pseudoentropy [HILL99] extends the definition of pseudorandomness syntactically. As a distribution is said to be pseudorandom if it is indistinguishable from a distribution of maximum entropy (which is unique), they define a distribution to have pseudoentropy k is it is indistinguishable from a distribution of Sahnnon entropy k (for which there are many possibilities).

It turns out that the two definitions of pseudoentropy above can be very different in natural computational settings, despite the fact that in the information theoretic setting they are identical

[1] What is meant by "efficient" can naturally vary by specifying machine models and resource bounds on them

[2] While we will first mainly talk about Shannon's entropy, we later switch to min-entropy and stay with it throughout the paper. However the whole introduction may be read when regarding the term "entropy" with any other of its many formal variants, or just as well as the informal notion of "information content" or "uncertainty"

for any k. Which definition, then, is the "natural one" to choose from? This question is actually more complex, as another natural point of view lead to yet another definition.

A definition using a natural metric space. The computational viewpoint of randomness may be thought of as endowing the space of *all* probability distributions with new, interesting metrics.

For every event (=test) T in our probability space we define: $d_T(X,Y) = |\Pr_X[T] - \Pr_Y[T]|$. In words, the distance between X and Y is the difference (in absolute value) of the probabilities they assign to T.[3]

Note that given a family of metrics, their maximum is also a metric. An information theoretic metric on distributions, the *statistical distance*[4] (which is basically $\frac{1}{2}L_1$-distance) is obtained by taking the maximum over the T-metrics above for *all* possible tests T. A natural computational metric, is given by taking the maximum over any class $\mathcal{C}$ of efficient tests. When should we say that a distribution X is indistinguishable from having Shannon entropy k? Distance to a set is the distance to the closest point in it, so X has to be close in this metric to *some* Y with Shannon entropy k.

A different order of quantifiers. At first sight this may look identical to the "indistinguishability" definition in [HILL99]. However let us parse them to see the difference. The [HILL99] definition say that X has pseudoentropy k if *there exists* a distribution Y of Shannon entropy k, such that *for all* tests T in $\mathcal{C}$, T has roughly the same probability under both X and Y. The metric definition above reverses the quantifiers: X has pseudoentropy k if *for every* a distribution Y of Shannon entropy k, *there exists* a test T in $\mathcal{C}$, which has roughly the same probability under both X and Y. It is easy to see that the metric definition is more liberal - it allows for at least as many distributions to have pseudoentropy k. Are they really different?

Relations between the three definitions. As all these definitions are natural and well-motivated, it makes sense to study their relationship. In the information theoretic world (when ignoring the "efficiency" constraints) all definitions are equivalent. It is easy to verify that regardless of the choice of a class $\mathcal{C}$ of "efficient" tests, they are ordered in permisiveness (allowing more distributions to have pseudoentropy k). The "indistinguishability" definition of [HILL99] is the most stringent, then the "metric definition", and then the "compression" definition of [Yao82]. What is more interesting is that we can prove collapses and separations for different computational settings and assumptions. For example, we show that the first two definitions drastically differ for logspace observers, but coincide for polynomial time observers (both in the uniform and nonuniform settings). The proof of the latter statement uses the "min-max" Theorem of [vN28] to "switch" the order of quantifiers. We can show some weak form of equivalence between all three definitions for circuits. We show that the "metric" coincides with the "compression" definition if $\mathbf{NP} \subseteq \mathbf{BPP}$. More precisely, we give a *non-deterministic* reduction showing the equivalence of the two definitions. This reduction guarantees high min-entropy according to the "metric" definition if the distribution has high min-entropy according to the "compression" distribution with respect to an **NP** oracle. A clean way to state this is that all three definitions are equivalent for $\mathbf{PH}/\mathrm{poly}$. We refer to this class as the class of poly-size **PH**-circuits. Such circuits are poly-size circuits which are allowed to compute an arbitrary function in the polynomial-hierarchy (**PH**). We remark

[3] This isn't precisely a metric as there may be different X and Y such that $d_T(X,Y) = 0$. However it is symmetric and satisfies the triangle inequality.

[4] Another basic distance measure is the so called KL-divergence, but for our purposes, which concern very close distributions, is not much different than statistical distance

that similar circuits (for various levels of the **PH** hierarchy) arise in related contexts in the study of "computational randomness": They come up in conditional "derandomization" results of **AM** [KvM02, MV99, SU01] and "extractors for samplable distributions" [TV00].

1.2 Pseudoentropy versus information theoretic entropy

We now move to another important part of our project. As these definitions are supposed to help establish a computational version of information theory, we attempt to see which of them respect some natural properties of information-theoretic entropy.

Using randomness extractors. In the information theoretic setting, there are *randomness extractors* which convert a high entropy[5] distribution into one which is statistically close to uniform. The theory of extracting the randomness from such distributions is by now quite developed (see surveys [Nis96, NT99, Sha02]). It is natural to expect that applying these randomness extractors on high pseudoentropy distributions produces a pseudorandom distribution. In fact, this is the motivation for pseudoentropy in some previous works [ILL89, HILL99, STV99].

It is easy to see that the the "indistinguishability" definition of [HILL99] has this property. This also holds for the "metric" definition by the equivalence above. Interestingly, we do not know whether this holds for the "compression" definition. Nevertheless, we show that some extractor constructions in the literature (the ones based on Trevisan's technique [Tre99, RRV99, ISW00, TSZS01, SU01]) do produce a pseudorandom distribution when working with the "compression" definition.

The information in two dependent distributions. One basic principle in information theory is that two (possibly dependent) random variables have at least as much entropy as any one individually, e.g. $H(X,Y) \geq H(X)$. A natural question is whether this holds when we replace information-theoretic entropy with pseudoentropy. We show that the answer depends on the model of computation. If there exist one-way functions, then the answer is *no* for the standard model of polynomial-time distinguishers. On the other hand, if $\mathbf{NP} \subseteq \mathbf{BPP}$, then the answer is *yes*. Very roughly speaking, the negative part follows from the existence of pseudorandom generators, while the positive part follows from giving a *nondeterministic* reduction which relies on nondeterminism to perform approximate counting. Once again, this result can be also stated as saying that the answer is positive for poly-size **PH**-circuits. We remark that the positive result holds for (nonuniform) online space-bounded computation as well.

Entropy and unpredictability. A deeper and interesting connection is the one between entropy and unpredictability. In the information theoretic world, a distribution which is unpredictable has high entropy.[6] Does this relation between entropy and unpredictability holds in the computational world?

Let us restrict ourselves here for a while to the metric definition of pseudoentropy. Two main results we prove is that this connection indeed holds in two natural computational notions of efficient observers. One is for logspace observers. The second is for **PH**-circuits. Both results

[5] It turns out that a different variant of entropy called "min-entropy" is the correct measure for this application. The min-entropy of a distribution X is $\log_2(\min_x 1/\Pr[X = x])$. This should be compared with Shannon's entropy in which the minimum is replaced by averaging.

[6] We consider two different forms of prediction tests: The first called "next bit predictor" attempts to predict a bit from previous bits, whereas the second called "complement predictor" has access to all the other bits, both previous and latter.

use one mechanism - a different characterization of the metric definition, in which distinuguishers accept very few inputs (less than 2^k when the pseudoentropy is k). We show that predictors for the accepted set are also good for any distribution "caught" by such a distinguisher. This direction is promising as it suggests a way to "bypass" the weakness of the "hybrid argument".

The weakness of the hybrid argument. Almost all pseudorandom generators (whether conditional such as the ones for small circuits or unconditional such as the ones for logspace) use the hybrid argument in their proof of correctness. The idea is that if the output distribution can be efficiently distinguished from random, some bit can be efficiently predicted with nontrivial advantage. Thus, pseudorandomness is established by showing unpredictability.

However, in standard form, if the distinughishability advantage is ϵ, the prediction advantage is only ϵ/n. In the results above, we manage (for these two computational models) to avoid this loss and make the prediction advantage $\Omega(\epsilon)$ (just as information theory suggests).

While we have no concrete applications, this seem to have potential to improve various constructions of pseudorandom generators. To see this, it suffices to observe the consequences of the hybrid argument loss. It requires every output bit of the generator to be very unpredictable, for which a direct cost is paid in the seed length (and complexity) of the generator. For generators against circuits, a long sequence of works [Yao82, BFNW91, IW97, STV99] resolved it optimally using efficient *hardness amplification*. These results allow constructing distributions which are unpredictable even with advantage $1/\text{poly}(n)$. The above suggests that sometimes this amplification may not be needed. One may hope to construct a pseudorandom distribution by constructing an unpredictable distribution which is only unpredictable with constant advantage, and then use a randomness extractor to obtain a pseudorandom distribution.[7]

This problem is even more significant when constructing generators against logspace machines [Nis90, INW94]. The high unpredictability required seems to be the bottleneck for reducing the seed length in Nisan's generator [Nis90] and its refinements from $O((\log n)^2)$ bits to the optimal $O(\log n)$ bits (that will result in $BPL = L$). The argument above gives some hope that for fooling logspace machines (or even just constant-width oblivious branching programs) the suggested approach may yield substantial improvements. However, in this setup there is another hurdle: In [BYRST02] it was shown that randomness extraction cannot be done by one pass log-space machines. Thus, in this setup it is not clear how to move from pseudoentropy to pseudorandomness.

1.3 Organization of the paper

In Section 2 we give some basic notation. Section 3 formally defines our three basic notions of pseudoentropy, and proves a useful characterization of the metric definition. In Sections 5 and 6 we prove equivalence and separations results between the various definitions in several natural computational models. Section 7 is devoted to our results about computational analogs of information theory for concatenation and unpredictability of random variables.

2 Preliminaries

Let X be a random variable over some set S. We say that X has *(statistical) min-entropy* at least k, denoted $H^\infty(X) \geq k$, if for every $x \in S$, $\Pr[X = x] \leq 2^{-k}$. We use U_n to denote the uniform

[7] This approach was used in [STV99]. They show that even "weak" hardness amplification suffices to construct a high pseudoentropy distribution using the pseudo-random generator construction of [NW94]. However, their technique relies on the properties of the specific generator and cannot be applied in general.

distribution on $\{0,1\}^n$.

Let X, Y be two random variables over a set S. Let $f : S \to \{0,1\}$ be some function. The *bias* of X and Y with respect to f, denoted $\mathsf{bias}_f(X,Y)$, is defined by $\left|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\right|$. Since it is sometimes convenient to omit the absolute value, we denote $\mathsf{bias}^*_f(X,Y) = \mathbb{E}[f(X)] - \mathbb{E}[f(Y)]$.

The *statistical distance* of X and Y, denoted $\mathsf{dist}(X,Y)$, is defined to be the maximum of $\mathsf{bias}_f(X,Y)$ over all functions f. Let $\mathcal{C}$ be a class of functions from S to $\{0,1\}$ (e.g., the class of functions computed by circuits of size m for some integer m). The *computational distance* of X and Y w.r.t. $\mathcal{C}$, denoted $\mathsf{comp\text{-}dist}_{\mathcal{C}}(X,Y)$, is defined to be the maximum of $\mathsf{bias}_f(X,Y)$ over all $f \in \mathcal{C}$. We will sometimes drop the subscript $\mathcal{C}$ when it can be inferred from the context.

Computational models. In addition to the standard model of uniform and non-uniform polynomial-time algorithms, we consider two additional computational models. The first is the model of **PH**-*circuits*. A **PH**-circuit is a boolean circuit that allows queries to a language in the polynomial hierarchy as a basic gate.[8] The second model is the model of *bounded-width read-once oblivious branching programs*. A width-S read once oblivious branching program P is a directed graph with Sn vertices, where the graph is divided into n layers, with S vertices in each layer. The edges of the graph are only between from one layer to the next one, and each edge is labelled by a bit $b \in \{0,1\}$ which is thought of as a variable. Each vertex has two outgoing edges, one labelled 0 and the other labelled 1. One of the vertices in the first layer is called the *source* vertex, and some of the vertices in the last layer are called the **accepting vertices**. A computation of the program P on input $x \in \{0,1\}^n$ consists of walking the graph for n steps, starting from the source vertex, and in step i taking the edge labelled by x_i. The output of $P(x)$ is 1 iff the end vertex is accepting. Note that variables are read in the natural order and thus width-S read once oblivious branching programs are the non-uniform analog of one-pass (or online) space-log S algorithms.

3 Defining computational min-entropy

In this section we give three definitions for the notion of computational (or "pseudo") min-entropy. In all these definitions, we fix $\mathcal{C}$ to be a class of functions which we consider to be efficiently computable. Our standard choice for this class will be the class of functions computed by a boolean circuit of size $p(n)$, where n is the circuit's input length and $p(\cdot)$ is some fixed polynomial. However, we will also be interested in instantiations of these definitions with respect to different classes $\mathcal{C}$. We will also sometimes treat $\mathcal{C}$ as a class of *sets* rather then functions, where we say that a set D is in $\mathcal{C}$ iff its characteristic function is in $\mathcal{C}$. We will assume that the class $\mathcal{C}$ is closed under complement.

3.1 HILL-type pseudoentropy: using indistinguishability

We start with the standard definition of computational (or "pseudo") min-entropy, as given by [HILL99]. We call this definition *HILL-type pseudoentropy*.

Definition 3.1. Let X be a random variable over a set S. Let $\epsilon \geq 0$. We say that X has **ϵ-HILL-type pseudoentropy** at least k, denoted $H^{\mathbf{HILL}}_{\epsilon}(X) \geq k$, if there exists a random variable Y with (statistical) min-entropy at least k such that the computational distance (w.r.t. $\mathcal{C}$) of X and Y is at most ϵ.

[8] Equivalently, the class languages accepted by poly-size **PH**-circuits is **PH**/poly.

We will usually be interested in ϵ-pseudoentroy for ϵ that is a small constant. In these cases we will sometimes drop ϵ and simply say that X has (HILL-type) pseudoentropy at least k (denoted $H^{\mathbf{HILL}}(X) \geq k$).

3.2 Metric-type pseudoentropy: using a metric space

In Definition 3.1 the distribution X has high pseudoentropy if there *exists* a high min-entropy Y such that X and Y are indistinguishable. As explained in the introduction, it is also natural to reverse the order of quantifiers: Here we allow Y to be a function of the "distinguishing test" f.

Definition 3.2. Let X be a random variable over a set S. Let $\epsilon \geq 0$. We say that X has **ϵ-metric-type pseudoentropy** at least k, denoted $H^{\mathbf{Metric}}_{\epsilon}(X) \geq k$, if for every test f on S there exists a Y which has (statistical) min-entropy at least k and $\mathsf{bias}_f(X,Y) < \epsilon$.

It turns out that metric-pseudoentropy is equivalent to a different formulation. (Note that the condition below is only meaningful for D such that $|D| < 2^k$).

Lemma 3.3. *For every class $\mathcal{C}$ which is closed under complement and for every $k \leq \log|S| - 1$ and ϵ, $H^{\mathbf{Metric}}_{\epsilon}(X) \geq k$ if and only if for every set $D \in \mathcal{C}$, $\Pr[X \in D] \leq \frac{|D|}{2^k} + \epsilon$*

Proof. An equivalent way to state the condition above is that there exists a distinguisher $D \in \mathcal{C}$ such that $\mathsf{bias}_D(X,Y) > \epsilon$ for every Y with $H^{\infty}(Y) \geq k$. Indeed, suppose that there exists a distinguisher $D \in \mathcal{C}$ such that $\Pr[X \in D] > \frac{|D|}{2^k} + \epsilon$. Yet, for *every* Y with $H^{\infty}(Y) \geq k$ it holds that $\Pr[Y \in D] \leq \frac{|D|}{2^k}$. Thus, it holds that for any such Y, $\mathsf{bias}_D(X,Y) > \epsilon$. For the other direction, suppose that there exists a D such that $\mathsf{bias}_D(X,Y) > \epsilon$ for every Y with $H^{\infty}(Y) \geq k$. We assume without loss of generality that $|D| < |S|/2$ (otherwise we take D's complement). We need to prove that $\Pr[X \in D] > \frac{|D|}{2^k} + \epsilon$. Indeed, suppose otherwise that $\Pr[X \in D] \leq \frac{|D|}{2^k} + \epsilon$. We construct a distribution $Y = Y_D$ in the following way: Y will be uniform on the set D with probability $\Pr[X \in D] - \epsilon$, and otherwise it is uniform on the set $S \setminus D$. By the construction it is clear that $\mathsf{bias}_D(X,Y) = \epsilon$, and so we can get a contradiction if we show that $H^{\infty}(Y) \geq k$. Indeed, let $y \in S$. If $y \in D$ then $\Pr[Y = y] = (\Pr[X \in D] - \epsilon)/|D| \leq 2^{-k}$. If $y \notin D$ then $\Pr[Y = y] \leq 1/(|S| - |D|) \leq \frac{2}{|S|} = 2^{-(\log|S|-1)} \leq 2^{-k}$. □

3.3 Yao-type pseudoentropy: using compression

Let $\mathcal{C}$ be a class of functions which we consider to efficiently computable. Recall that we said that a set D is a member of $\mathcal{C}$ if its characteristic function was in $\mathcal{C}$. That is, a set D is in $\mathcal{C}$ if it is *efficiently decidable*. We now define a family $\mathcal{C}_{\mathsf{compress}}$ of sets that are **efficiently compressible**. That is, we say that a set $D \subseteq S$ is in $\mathcal{C}_{\mathsf{compress}}(\ell)$ if there exist functions $c, d \in \mathcal{C}$ ($c : S \to \{0,1\}^{\ell}$ stands for *compress* and $d : \{0,1\}^{\ell} \to S$ for *decompress*) such that $D = \{x | d(c(x)) = x\}$. Note that every efficiently compressible set is also efficiently decidable (assuming the class $\mathcal{C}$ is closed under composition). Yao-type pseudoentropy is defined by replacing the quantification over $D \in \mathcal{C}$ in the alternative characterization of metric-type pseudoentropy (Lemma 3.3) by a quantification over $D \in \mathcal{C}_{\mathsf{compress}}(\ell)$ for all $\ell < k$. The resulting definition is the following:

Definition 3.4. Let X be a random variable over a set S. X has *ϵ-Yao-type pseudoentropy* at least k, denoted $H^{\mathbf{Yao}}_{\epsilon}(X) \geq k$, if for every $\ell < k$ and every set $D \in \mathcal{C}_{\mathsf{compress}}(\ell)$,

$$\Pr[X \in D] \leq 2^{l-k} + \epsilon$$

4 Using randomness extractors

An extractor uses a short seed of truly random bits to extract many bits which are (close to) uniform.

Definition 4.1 ([NZ93]). A function $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k,ϵ)-extractor if for every distribution X on $\{0,1\}^n$ with $H^\infty(x) \geq k$, the distribution $E(X,U_d)$ has statistical distance at most ϵ from U_m.

We remark that there are explicit (polynomial time computable) extractors with seed length polylog(n/ϵ) and $m = k$. The reader is referred to survey papers on extractors [Nis96, NT99, Sha02]. It is easy to see that if a distribution X has HILL-type pseudoentropy at least k, then for every (k,ϵ)-randomness extractor the distribution $E(X,U_d)$ is ϵ-computationally indistinguishable pseudorandom.

Lemma 4.2. *Let $\mathcal{C}$ be the class of polynomial size circuits. Let X be a distribution with $H^{\mathbf{HILL}}_\epsilon(X) \geq k$ and let E be a (k,ϵ)-extractor computable in time* poly(n) *then* $\mathsf{comp\text{-}dist}_{\mathcal{C}}(E(X,U_d),U_m) < 2\epsilon$.

Proof. Let Y be a distribution with $H^\infty(Y) \geq k$ and $\mathsf{comp\text{-}dist}_{\mathcal{C}}(X,Y) < \epsilon$. If the claim does not hold then there is an $f \in \mathcal{C}$ such that $\mathsf{bias}_f(E(X,U_D),U_m) \geq 2\epsilon$. However, $\mathsf{bias}_f(E(Y,U_d),U_m) < \epsilon$ and thus, $\mathsf{bias}_f(E(X,U_d),E(Y,U_d)) \geq \epsilon$. It follows that there exists $s \in \{0,1\}^d$ such that $\mathsf{bias}_{f(E(\cdot,s))}(X,Y) > \epsilon$. This is a contradiction as $f(E(\cdot,s)) \in \mathcal{C}$. □

In Theorem 5.2 we show equivalence between HILL-type pseudoentropy and metric-type pseudoentropy and thus we get that $E(X,U_d)$ is also pseudorandom when X has metric-type pseudoentropy. Interestingly, we do not know whether this holds for Yao-type pseudoentropy. We can however show that this holds for some extractors, namely ones with a "reconstruction procedure".

Definition 4.3 (reconstruction procedure). An (ℓ,ϵ)-reconstruction for a function $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a pair of machines C and R where $C : \{0,1\}^n \to \{0,1\}^\ell$ is a randomized Turing machine, and $R : \{0,1\}^\ell \to \{0,1\}^n$ is a randomized oracle Turing machine which runs in time polynomial in n. Furthermore, for every x and f with $\mathsf{bias}_f(E(x,U_d),U_m) \geq \epsilon$, $\Pr[R^f(C(x)) = x] > 1/2$ (the probability is over the random choices of C and R).

It was shown by Trevisan [Tre99] that every function E which has a reconstruction is an extractor.[9]

Theorem 4.4 (implicit in [Tre99]). *If E has an (ℓ,ϵ)-reconstruction then E is a $(\ell+\log(1/\epsilon),3\epsilon)$-extractor.*

We include the proof for completeness.

Proof. Assume for the purpose of contradiction that there is a distribution Y with $H^\infty(Y) \geq k$ and a test f such that $\mathsf{bias}_f(E(Y,U_d),U_m) \geq 3\epsilon$. It follows that there is a subset $B \subseteq S$ with $\Pr_Y[B] \geq 2\epsilon$ such that for every $y \in B$, $\mathsf{bias}_f(E(y,U_d),U_m) \geq \epsilon$. For every $y \in B$, $\Pr[R^f(C(y)) = y] > 1/2$. This probability is over the random choices of R and C. Thus, there exist fixings to the random coins of the machines R and C such that $\Pr_Y[R^f(C(y)) = y] > \epsilon$. We use c to denote the machine C with such fixing. We use d to denote the machine R^f with the such fixing. The set $D = \{y | d(c(y)) = y\}$ is of size at most 2^ℓ, as Y has min-entropy at least k, this means that $\Pr[Y \in D] \leq 2^{\ell-k} = \epsilon$. However we just showed that $\Pr[Y \in D] > \epsilon$. □

[9] Reconstruction procedures (with stronger efficiency requirements) were previously designed to construct pseudorandom generators from hard functions [NW94, BFNW91, IW97]. In [Tre99], Trevisan observed that these constructions also yield extractors.

The reader is referred to a survey [Sha02] for a detailed coverage of the "reconstruction proof procedure". Interestingly, such extractors can be used with Yao-type pseudoentropy.[10] Loosely speaking, this is because the failure of such an extractor implies an efficient compression of a noticeable fraction of the high min-entropy distribution.

Lemma 4.5. *Let $\mathcal{C}$ be the class of polynomial size circuits. Let X be a distribution with $H^{\mathbf{Yao}}_{\epsilon}(X) \geq k$ and let E be an extractor with a $(k - \log(1/\epsilon), \epsilon)$-reconstruction which is computable in time* $\mathrm{poly}(n)$ *then* $\mathsf{comp\text{-}dist}_{\mathcal{C}}(E(X, U_d), U_m) < 5\epsilon$.

Proof. Assume for the purpose of contradiction that there is an $f \in \mathcal{C}$ such that $\mathsf{bias}_f(E(X, U_d), U_m) \geq 5\epsilon$. It follows that there is a subset $B \subseteq S$ with $\Pr_X[B] \geq 4\epsilon$ such that for every $x \in B$, $\mathsf{bias}_f(E(x, U_d), U_m) \geq \epsilon$. For every $x \in B$, $\Pr[R^f(C(x)) = x] > 1/2$. This probability is over the random choices of R and C. Thus, there exist fixings to the random coins of the machines R and C such that $\Pr_X[R^f(C(x)) = x] > 2\epsilon$. We use c to denote the machine C with such fixing. We use d to denote the machine R^f with such fixing. Note that $c, d \in \mathcal{C}$ and that set $D = \{x | d(c(x)) = x\}$ has $\Pr[X \in D] > 2\epsilon \geq \frac{|D|}{2^k} + \epsilon = 2^{\ell - k} + \epsilon$. □

5 Relationships between definitions

In this section we study relationships between the various definitions. It is important to note that if computational issues are removed (if the class $\mathcal{C}$ is the class of all functions) the three definitions are essentially equivalent to having statistical distance ϵ from a distribution Y with $H^{\infty}(Y) \geq k$. We also note that all definitions result in a pseudo-random distribution for $k = n$. For $k < n$, we show that the HILL-type and metric-type definitions are essentially equivalent for polynomial circuits and Turing machines. However, things are somewhat different with the Yao-type definition: We are only able to show equivalence to the other types for the much stronger model of PH-circuits.

5.1 Equivalence in the case of pseudo-randomness ($k = n$)

The case of $k = n$ is the one usually studied in the theory of pseudo-randomness. In this case the HILL-type definition coincides with the standard definition of pseudo-randomness. This is because there's only one distribution Y over $\{0,1\}^n$ with $H^{\infty}(Y) \geq n$, that is the uniform distribution. It also follows that the HILL-type and metric-type definition coincide for every class $\mathcal{C}$. The equivalence of Yao-type pseudoentropy to pseudo-randomness follows from the hybrid argument of [Yao82, GM84].

Lemma 5.1. *If $H^{\mathbf{Yao}}_{\epsilon/n}(X) = n$ then $H^{\mathbf{HILL}}_{\epsilon}(X) = n$.*

Proof. By the hybrid argument of [Yao82, GM84], if $H^{\mathbf{HILL}}_{\epsilon}(X) < n$ (with respect to circuits of size s) then there is an $1 \leq i < n$ and a "predictor circuit" P of size $s + O(n)$ such that

$$\Pr_X[P(X_1, \cdots, X_{i-1} = X_i] > 1/2 + \epsilon/n$$

Let $\ell = n - 1$, we define $c : \{0,1\}^n \to \{0,1\}^{\ell}$ by $c(x_1, \cdots, x_n) = x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$. The "decompressor" function $d : \{0,1\}^{\ell} \to \{0,1\}^n$ only needs to evaluate x_i. This is done by running $P(x_1, \cdots, x_{i-1})$. We conclude that $\Pr_X[d(c(x)) = x] > 1/2 + \epsilon/n = 2^{\ell - k} + \epsilon/n$ and thus, $H^{\mathbf{Yao}}_{\epsilon/n}(X) < n$. □

[10] For completeness we mention that Trevisan's extractor [Tre99] can achieve $d = O(\log n)$, $m = \sqrt{k}$ for $k = n^{\Omega(1)}$ and $\epsilon = 1/k$. The correctness of Trevisan's extractor follows from an (ℓ, ϵ)-reconstruction for $\ell = mn^c$ for a constant c. (This constant depends on the constant hidden in the O, Ω-notation above.)

5.2 Equivalence between HILL-type and metric-type

The difference between HILL-type and metric-type pseudoentropy is in the order of quantifiers. HILL-type requires that there exist a unique "reference distribution" Y with $H^{\infty}(Y) \geq k$ such that for every D, $\mathsf{bias}_D(X,Y) < \epsilon$, whereas metric-type allows Y to depend on D, and only requires that for every D there exists such a Y. It immediately follows that for every class $\mathcal{C}$ and every X, $H^{\mathbf{Metric}}(X) \geq H^{\mathbf{HILL}}(X)$. In this section we show that the other direction also applies (with small losses in ϵ and time/size) for small circuits.

Theorem 5.2 (Equivalence of HILL-type and metric-type for circuits). *Let X be a distribution over $\{0,1\}^n$. For every $\epsilon, \delta > 0$ and k, if $H^{\mathbf{Metric}}_{\epsilon-\delta}(X) \geq k$ (with respect to circuits of size $O(ns/\delta^2)$ then $H^{\mathbf{HILL}}_{\epsilon}(X) \geq k$ (with respect to circuits of size s)*

The proof of Theorem 5.2 relies on the "min-max" theorem of [vN28], which is used to "switch" the order of the quantifiers. We explain this technique in Section 5.3, and prove the Theorem in Section 5.4.

5.3 Switching quantifiers using the "min-max" theorem

We want to show that if X has $H^{\mathbf{Metric}}(X) \geq k$ then $H^{\mathbf{HILL}}(X) \geq k$. Our strategy will be to show that if $H^{\mathbf{HILL}}(X) < k$ then then $H^{\mathbf{Metric}}(X) < k$. Thus, our assumption gives that for every Y with $H^{\infty}(Y) \geq k$ there is a $D \in \mathcal{C}$ such that $\mathsf{bias}_D(X,Y) \geq \epsilon$. The following lemma allows us to switch the order of quantifiers, the cost is that we get a "distribution" over D's instead of a single D.

Lemma 5.3. *Let X be a distribution over S. Let $\mathcal{C}$ be a class that is closed under complement. If for every Y with $H^{\infty}(Y) \geq k$ there exists a $D \in \mathcal{C}$ such that $\mathsf{bias}_D(X,Y) \geq \epsilon$ then there is a distribution $\hat{D}$ over $\mathcal{C}$ such that for every Y with $H^{\infty}(Y) \geq k$*

$$\mathbb{E}_{D \leftarrow \hat{D}}[\mathsf{bias}^*_D(X,Y)] \geq \epsilon$$

The proof of Lemma 5.3 use von-Neuman's "min-max" theorem for finite 2-player zero-sum games.

Definition 5.4 (zero-sum games). Let A and B be finite sets. A game is a function $g : A \times B \to R$. Let $\hat{A}$ and $\hat{B}$ denote the set of distributions over A and B: We define $\hat{g} : \hat{A} \times \hat{B} \to R$.

$$\hat{g}(\hat{a}, \hat{b}) = \mathbb{E}_{a \leftarrow \hat{a}, b \leftarrow \hat{b}} g(a,b)$$

We use $a \in A$ to also denote the distribution $\hat{a} \in \hat{A}$ which gives probability one to a.

Theorem 5.5 ([vN28]). *For every game g there is a value v such that*

$$max_{\hat{a} \in \hat{A}} min_{b \in B} \hat{g}(\hat{a}, b) = v = min_{\hat{b} \in \hat{B}} max_{a \in A} \hat{g}(a, \hat{b})$$

Proof. (of Lemma 5.3) We define the following game. Let $A = \mathcal{C}$ and B be the set of all "flat" distributions Y with $H^{\infty}(Y) \geq k$. That is all distributions Y which are uniform over a subset T of size 2^k. We define $g(D,Y) = \mathsf{bias}^*_D(X,Y)$. Let v be the value of the game g. A nice feature of this game is that every $\hat{b} \in \hat{B}$ is a distribution over S which has $H^{\infty}(\hat{b}) \geq k$. It is standard that all distributions Y with $H^{\infty}(Y) \geq k$ are convex combinations of "flat" distributions. In other words, $\hat{B}$ is the set of all distributions Y with $H^{\infty}(Y) \geq k$. By our assumption for every $Y \in \hat{B}$ there exists a D in A such that $\mathsf{bias}_D(X,Y) \geq \epsilon$. The same holds for $\mathsf{bias}^*_D(X,Y)$ because if

$\mathsf{bias}_D(X,Y) \leq (-\epsilon)$ then $\mathsf{bias}_{\neg D}(X,Y) \geq \epsilon$ for $\neg D(x) = 1 - D(x)$. As $B = \hat{B}$ this means that $v = min_{\hat{b}\in\hat{B}} max_{a\in A} g(a,\hat{b}) \geq \epsilon$. By the min-max theorem, it follows that $max_{\hat{a}\in\hat{A}} min_{b\in B} \hat{g}(\hat{a},b) \geq \epsilon$. In other words, there exists a distribution $\hat{D}$ over $D \in \mathcal{C}$ such that for all $R \in B$, $\mathbb{E}_{D\leftarrow\hat{D}}[\mathsf{bias}^*_D(X,R)] \geq \epsilon$. As every distribution Y with $H^\infty(Y) \geq k$ is a convex combination of distributions R from B, we get that for every such Y, $\mathbb{E}_{D\leftarrow\hat{D}}[\mathsf{bias}^*_D(X,Y)] \geq \epsilon$. □

5.4 Proof of Theorem 5.2

The proof of Theorem 5.2 relies on the Definitions and Lemmas from Section 5.3.

Proof. (of Theorem 5.2) Let X be a distribution on $\{0,1\}^n$ with $H^{\mathbf{HILL}}_\epsilon(X) < k$ (with respect to circuits of size s) we will show that $H^{\mathbf{HILL}}_\epsilon(X) < k$ (with respect to circuits of size s). Let $\mathcal{C}$ be the class of circuits of size s. By our assumption for every Y with $H^\infty(Y) \geq k$ there is a $D \in \mathcal{C}$ such that $\mathsf{bias}_D(X,Y) \geq \epsilon$. By Lemma 5.3 there is a distribution $\hat{D}$ over $\mathcal{C}$ such that for every Y, $\mathbb{E}_{D\leftarrow\hat{D}}[\mathsf{bias}^*_D(X,Y)] \geq \epsilon$. We define $\bar{D}(x) = \mathbb{E}_{D\in\hat{D}}[D(x)]$ it follows that

$$\mathsf{bias}_{\bar{D}}(X,Y) \geq \mathsf{bias}^*_{\bar{D}}(X,Y) = \mathbb{E}_{D\leftarrow\hat{D}}[\mathsf{bias}^*_D(X,Y)] \geq \epsilon$$

To conclude, we approximate $\bar{D}$ by a small circuit. We choose $t = 8n/\delta^2$ samples $D_1, \cdots, D_t$ from $\hat{D}$ and define

$$D'_{D_1,\cdots,D_t}(x) = \frac{1}{t}\sum_{1\leq i\leq t} D_i(x)$$

By Chernoff's inequality, for every $x \in \{0,1\}^n$, $\Pr_{D_1,\cdots,D_t\leftarrow\hat{D}}[|D'_{D_1,\cdots,D_t}(x) - \bar{D}(x)| \geq \delta/2] \leq 2^{-2n}$. Thus, there exists $D_1, \cdots, D_t$ such that for all x, $|D'_{D_1,cdots,D_t}(x) - \bar{D}(x)| \leq \delta/2$. It follows that for every Y, $\mathsf{bias}_{D'_{D_1,\cdots,D_t}}(X,Y) \geq \epsilon - \delta$. Note that $D'_{D_1,\ldots,D_t}$ is of size $O(ts) = O(ns/\delta^2)$. □

5.5 Equivalence of metric and Hill for uniform polynomial-time Turing machines.

It is somewhat surprising that we can use the argument of Section 5.4 for *uniform* Turing machines. This is because the argument seem to exploit the non-uniformity of circuits: The "min-max theorem" works only for *finite* games and is non-constructive - it only shows existence of a distribution $\hat{D}$ and gives no clue to its complexity. The key idea is to consider Turing machines with bounded *description size*.

We now adapt definitions given in Section 3 to uniform machines. Let $\mathcal{M}$ be some class of Turing Machines (e.g., poly-time machines, probabilistic poly-time machines).

Definition 5.6 (pseudo-entropy for uniform machines). Let $X = \{X_n\}$ be a collection of distributions where X_n is on $\{0,1\}^n$. Let $k = k(n)$ and $\epsilon = \epsilon(n)$ be some functions. Let Δ_k denote the set of collection $\{Y_n\}$ such that for every n, $H^\infty(Y_n) \geq k(n)$.

- $H^{\mathbf{HILL}}_\epsilon(X) \geq k$ if $\exists\{Y_n\} \in \Delta_k, \forall M \in \mathcal{M}, \mathsf{bias}_M(X_n,Y_n) < \epsilon$, a.e.
- $H^{\mathbf{Metric}}_\epsilon(X) \geq k$ if $\forall M \in \mathcal{M}, \exists\{Y_n\} \in \Delta_k, \mathsf{bias}_M(X_n,Y_n) < \epsilon$, a.e.

Definition 5.7 (Description size). We use $\mathcal{M}$ to denote some class of Turing machines. (e.g., polynomial time machines). Fix some encoding of Turing machines.[11] We identify a Turing machine

[11] For technical reasons we assume that if $M \in \mathcal{M}$ then $1 - M \in \mathcal{M}$ an that both machines have descriptions of the same length.

M with its description. We use $|M|$ to denote the length of the description of M. We use $\mathcal{M}(s)$ to denote all machines $M \in \mathcal{M}$ with $|M| < s$.

Consider for example HILL-type pseudoentropy. For every M there is a input length from which point on the bias of M is small. We define $h(n)$ to be the largest number s such that for all machines $M \in \mathcal{M}$ with $|M| \leq s$, $\mathsf{bias}_M(X_n, Y_n) < \epsilon$. The definition says that $h(n) \to \infty$. We can rewrite the definitions with this view in mind. We use $\omega(1)$ to denote all functions which go to infinity.

Lemma 5.8 (pseudoentropy with description size). *The following holds:*

- $H^{\mathbf{HILL}}_{\epsilon}(X) \geq k$ *iff* $\exists h \in \omega(1), \forall n, \exists Y_n, H^{\infty}(Y_n) \geq k, \forall M \in \mathcal{M}(h(n)), \mathsf{bias}_M(X_n, Y_n) < \epsilon$.
- $H^{\mathbf{Metric}}_{\epsilon}(X) \geq k$ *iff* $\exists h \in \omega(1), \forall n, \forall M \in \mathcal{M}(h(n)), \exists Y_n, H^{\infty}(Y_n) \geq k, \mathsf{bias}_M(X_n, Y_n) < \epsilon$.

The proof of Lemma 5.8 uses the following trivial lemma.

Lemma 5.9. *Let $\{f_m\}$ be a family of boolean functions over the integers. The following conditions are equivalent:*

- *For every m, f_m outputs 1 a.e.*
- *There exists a function $h \in \omega(1)$ such that for every n and every $m < h(n)$ we have $f_m(n) = 1$.*

Proof. (of Lemma 5.8) We enumerate the machines $M \in \mathcal{M}$ by their descriptions m as an integer. For HILL-type pseudoentropy, both formulations fix some distribution $\{Y_n\}$ as a function of $\{X_n\}$. We define $f_M(n) = 1$ iff $\mathsf{bias}_M(X_n, Y_n) < \epsilon$. The lemma follows from Lemma 5.9. For metric-type pseudoentropy, $\{Y_n\}$ depends on M we denote it by $\{Y^M_n\}$ and define $f_M(n) = 1$ iff $\mathsf{bias}_M(X_n, Y^M_n) < \epsilon$. Again the lemma follows from Lemma 5.9. □

The following Theorem shows that for every constant ϵ if $H^{\mathbf{Metric}}_{\epsilon}(X) \geq k$ with respect to Turing machines with then $H^{\mathbf{HILL}}_{2\epsilon}(X) \geq k$ with small losses in running time.

Theorem 5.10. *[Equivalence of HILL-type and metric-type for uniform machines] For every constant ϵ and $w \in \omega(1)$. If $H^{\mathbf{Metric}}_{\epsilon/2}(X) \geq k$ (with respect to machines M which run in time $T(n) \log T(n) w(n)$) then $H^{\mathbf{HILL}}_{\epsilon}(X) \geq k$ (with respect to machines M which run in time $T(n)$).*

Proof. We will show that if $H^{\mathbf{HILL}}_{\epsilon}(X) < k$ (with respect to machines M which run in time $T(n)$) then $H^{\mathbf{Metric}}_{\epsilon/2}(X) < k$ (with respect to machines M which run in time $T(n) \log T(n) w(n)$). Let $\mathcal{M}'$ be the set of Turing machines which run in time $T(n) \log T(n) w(n)$. By Lemma 5.8 it is sufficient to show that:

$$\forall h' \in \omega(1), \exists n, \exists M \in \mathcal{M}'(h'(n)), \forall Y_n, H^{\infty}(Y_n) \geq k, \mathsf{bias}_M(X_n, Y_n) \geq \epsilon$$

Let h' be some function in $\omega(1)$. Let $\mathcal{M}$ be the set of Turing machines which run in time $T(n)$. By Lemma 5.8 our starting assumption is that:

$$\forall h \in \omega(1), \exists n, \forall Y_n, H^{\infty}(Y_n) \geq k, \exists M \in \mathcal{M}(h(n)), \mathsf{bias}_M(X_n, Y_n) < \epsilon$$

The key observation is that these statements involves the behavior of the machine on a fixed n. On this fixed n the quantification is over finitely many machines (those in $\mathcal{M}(h(n))$). We can think of $\mathcal{M}(h(n))$ as a new *non-uniform* circuit class and use Lemma 5.3 as in the proof of Theorem 5.2.

Let $h(n)$ be the largest function that $2^{h(n)^2+2\log(1/\epsilon)} \leq h'(n)$. Note that $h \in \omega(1)$. We can assume wlog that $2^{h(n)^2+2\log(1/\epsilon)} < w(n)$ by restricting ourselves to small enough h'. Let n be a number which existence is guaranteed above. We define $\mathcal{C} = \mathcal{M}(h(n))$.

By our assumption for every Y_n with $H^\infty(Y_n) \geq k$ there is a $D \in \mathcal{C}$ such that $\mathsf{bias}_D(X_n, Y_n) \geq \epsilon$. By Lemma 5.3 there is a distribution $\hat{D}$ over $\mathcal{C}$ such that for every Y_n, $\mathbb{E}_{D \leftarrow \hat{D}}[\mathsf{bias}^*_D(X_n, Y_n)] \geq \epsilon$. We define $\bar{D}(x) = \mathbb{E}_{D \in \hat{D}}[D(x)]$ it follows that

$$\mathsf{bias}_{\bar{D}}(X_n, Y_n) \geq \mathsf{bias}^*_{\bar{D}}(X_n, Y_n) = \mathbb{E}_{D \leftarrow \hat{D}}[\mathsf{bias}^*_D(X_n, Y_n)] \geq \epsilon$$

As in the proof of Theorem 5.2 we now approximate $\bar{D}$ by a "small machine".[12] Note that $\hat{D}$ is a distribution over only $2^{h(n)}$ elements. Let $t = \log(1/\epsilon) + 1$. We round the distribution $\hat{D}$ to a distribution $\hat{P}$ such that every element D in the range of $\hat{P}$ has probability $\frac{i}{2^{h(n)+t}}$ for some $0 \leq i \leq 2^{h(n)+t}$. We define $\bar{P}(x) = \mathbb{E}_{D \in \hat{P}}[D(x)]$. It follows that $\mathsf{bias}_{\bar{P}}(X_n, Y_n) \geq \mathsf{bias}_{\bar{D}}(X_n, Y_n) - 2^{-t} \geq \epsilon/2$. To describe $\bar{P}$ we need to describe $\hat{P}$ (this takes $2^{h(n)}(h(n)+t)$ bits) and all the machines in $\mathcal{M}(h(n))$ (this takes $2^{h(n)}h(n)$ bits. Thus, $\bar{P}$ has description size $2^{O(h(n))+\log(1/\epsilon)} \leq h'(n)$. Simulating a (multi-tape) Turing machine which runs in time $T(n)$ can be done in time $O(T(n)\log T(n)$ on a (2-tape) Turing machine, and thus $\bar{P}$ runs in time $O(T(n)\log T(n))\text{poly}(2^{h(n)+\log(1/\epsilon)}) \leq T(n)\log T(n)w(n)$. We have indeed shown that:

$$\forall h' \in \omega(1), \exists n, \exists M \in \mathcal{M}'(h'(n)), \forall Y_n, H^\infty(Y_n) \geq k, \mathsf{bias}_M(X_n, Y_n) \geq \epsilon$$

□

5.6 Equivalence between all types for PH-circuits.

We do not know whether the assumption that $H^{\mathbf{Yao}}_\epsilon(X) \geq k$ for circuits implies that $H^{\mathbf{Metric}}_\epsilon(X) \geq k'$ for slightly smaller k' and circuit size (and in fact, we conjecture that it's false). However, we can prove it assuming the circuits for the Yao-type definition have access to an NP-oracle.

Theorem 5.11. *Let $k' = k + 1$ There is a constant c so that if $H^{\mathbf{Yao}}_\epsilon(X) \geq k'$ (with respect to circuits of size $max(s, n^c)$ that use an NP-oracle) then $H^{\mathbf{Metric}}_\epsilon(X) \geq k$ (with respect to circuits of size s).*

Proof. We start with a proof for a weaker result with $k' = 2k$. We then sketch how to get $k' = k + 1$. Let X be a distribution with $H^{\mathbf{Metric}}_\epsilon(X) < k$ (for circuits of size s). We will show that $H^{\mathbf{Yao}}_\epsilon(X) < 2k$ with respect to circuits with NP-oracle. By Lemma 3.3 there exists a circuit C of size s with $Pr_X[C(X) = 1] > \frac{|D|}{2^k} + \epsilon$ where $D = \{x | C(x) = 1\}$. We define $t = \log|D|$. Let H be a 2-universal family of hash functions $h : \{0,1\}^n \to \{0,1\}^{2t}$.[13] There are such families such that each h can be computed by an n^c size circuit for some constant c [CW79]. The expected number of collisions (pairs $x_1 \neq x_2$ s.t. $x_1, x_2 \in D$ and $h(x_1) = h(x_2)$) is bounded from above by $\binom{2^t}{2}2^{-2t} \leq 1/2$ and therefore there exists an $h \in H$ such that h is one to one on D. We set $\ell = 2t$ and define the "compressor circuit" $c(x) = h(x)$. We now define the "de-compressor circuit" d which will use an NP-oracle. When given $z \in \{0,1\}^{2t}$, d uses its NP-oracle to find the unique

[12] There is however a major difference. In the non-uniform case we sampled $t > n$ elements from $\hat{D}$ and took their average to get one circuit. Intuitively, sampling was necessary because $\hat{D}$ could be over a lot of circuits. In our setup $\hat{D}$ is over only $2^{h(n)}$ elements. We can assume that h grows so slowly that, $2^{h(n)} << n$. Thus, computing $\bar{D}$ is just as economic as sampling. However, we need to be careful that the description size of computing $\bar{D}$ depends only on $h(n)$ and not on n or the choice of h.

[13] By that we mean that for every $x_1 \neq x_2$ $\Pr_{h \in_R H}[h(x_1) = h(x_2)] = 1/2^{2t}$.

$x \in D$ such that $h(x) = z$. The circuit d then outputs x. We set $k' = k + t \leq 2k$. It follows that $\Pr_X[d(c(x)) \geq 2^{t-k} + \epsilon \geq 2^{2t-k'} + \epsilon = 2^{\ell-k'} + \epsilon$. We conclude that $H_\epsilon^{\mathbf{Yao}}(X) < k' \leq 2k$ with respect to circuits of size $max(s, n^c)$ which use an NP-oracle. □

Remark 5.12. In the proof above we chose $\ell = 2t$ which could be as large as $2k$. As a result we only got that $H_\epsilon^{\mathbf{Yao}}(X) \leq 2k$ instead of $k + 1$. One simple way to decrease ℓ is to allow the circuit c to also use an NP-oracle. Here's a rough sketch: We will choose h from a n^2-wise independent family of hash functions from n bits to $t - \log n$ bits. We expect that n elements of D are mapped to each output string of h. We can use k-wise independent tail inequalities [BR94] to do a union bound over all "bins" and argue that there exists an h such that for each output $z \in \{0,1\}^{t-\log n}$ the number of "pre-images" in D is at most $2n$. The circuit c on input x will use an NP-oracle to find these $2n$ pre-images of $h(x)$ and will output $h(x)$ and the index of x amongst the pre-images. This is a one to one mapping into $t + 1$ bits (instead of $2t$ bits) and the argument can continue with $k' = k + 1$ instead of $k' = 2k$.

The reduction in the proof of Theorem 5.11 uses an NP-oracle. The class of polynomial size **PH**-circuits are closed under the use of NP-oracles ($\mathbf{PH}^{NP}/poly = \mathbf{PH}/poly$). Applying the argument of Theorem 5.11 give the following corollary.

Corollary 5.13. *Let $\mathcal{C}$ be the class of polynomial size* **PH***-circuits. If $H_\epsilon^{\mathbf{Yao}}(X) \geq 2k$ then $H_\epsilon^{\mathbf{Metric}}(X) \geq k$.*

6 Separation between types

Given the results of the previous section it is natural to ask if HILL-type and metric-type pseudoentropy are equivalent in **all** natural computational models? We give a negative answer and prove that there's large gap between HILL-type and metric-type pseudoentropy in the model of bounded-width read-once oblivious branching programs.

Theorem 6.1. *For every constant $\epsilon > 0$ and sufficiently large $n \in \mathbb{N}$, and , there exists a random X variable over $\{0,1\}^n$ such that $H_\epsilon^{\mathbf{Metric}} X \geq (1-\epsilon)n$ with respect to width-S read once oblivious branching programs, but $H_{1-\epsilon}^{\mathbf{HILL}}(X) \leq \text{polylog}(n, S)$ with respect to width-4 oblivious branching programs.*

Theorem 6.1 follows from the following two lemmas:

Lemma 6.2 (Based on [Sak96]). *Let $\epsilon > 0$ be some constant and $S \in \mathbb{N}$ such that $S > \frac{1}{\epsilon}$. Let $l = \frac{10}{\epsilon} \log S$ and consider the distribution $X = (U_l, U_l, \ldots, U_l)$ over $\{0,1\}^n$ for some $n < S$ which is a multiple of l. Then, $H_\epsilon^{\mathbf{Metric}}(X) \geq (1-\epsilon)n$ with respect to width-S oblivious branching programs.*

Proof. The proof is based on an extension of a theorem by Saks [Sak96]. Suppose, for the sake of contradiction, that $H_\epsilon^{\mathbf{Metric}}(X) < (1-\epsilon)n$. Then, there exists a width-S oblivious branching program D such that $\Pr[D(X) = 1] \geq \epsilon$ but $|D^{-1}(1)| \leq 2^{(1-\epsilon)n}$. The program D is a graph with n layers, where at each layer there are S vertices. The edges of the graph are only between consecutive layers and each edge is labelled with a bit $b \in \{0,1\}$. We consider a "contracted" graph that has n/l layers, where again the edges of the graph are only between consecutive layers. However, this time each edge (u, v) is labelled with a subset of $\{0,1\}^l$ that corresponds to all possible labels of paths between (u, v) in the original graph. Clearly the contracted graph computes the same language as the original graph (when again a string is accepted if the corresponding walk on the graph edges ends in an accepting vertex).

We say that an edge is "bad" if its corresponding set of labels is of size at most $S^{-4}2^l$. Note that, if $r \leftarrow_{\mathrm{R}} \{0,1\}^l$, then the probability that when performing the walk $(r, r, \ldots, r)$ on the graph we traverse a "bad" edge is at most $\frac{n}{l}S^2S^{-4} < S^{-1} < \epsilon$ (by a union bound over the at most $\frac{n}{l}S^2$ edges). Because $\Pr_{r\leftarrow_{\mathrm{R}}\{0,1\}^l}[D(r,r,\ldots,r)=1] > \epsilon$, there must exist an accepting path on the graph that consists only of good edges. Let S_i, where $1 \leq i \leq \frac{n}{l}$ denote the set of labels of the i^{th} edge on this path. Then, D accepts the set $S_1 \times S_2 \times \cdots \times S_{n/l}$. But this set is of size at least $(S^{-4}2^l)^{n/l} = (2^{l-4\log S})^{n/l} \geq 2^{(1-\epsilon)n}$ (since $l = \frac{10}{\epsilon}\log S$), and so we've reached a contradiction. □

Lemma 6.3. *Let $\epsilon > 0$ be some constant, and X be the random variable $(U_l, U_l, \ldots, U_l)$ over $\{0,1\}^n$ (where $l > \log n$). Then, $H^{\mathbf{HILL}}_{(1-\epsilon)}(X) \leq \frac{100}{\log(1/\epsilon)}l^3$ with respect to width-4 oblivious branching programs.*

Proof. Let $\mathcal{I}$ be a family of subsets of $[n]$ such that $I \in \mathcal{I}$ iff $\left|I \cap \big[jl, jl+l\big)\right| = 1$ for all $1 \leq j \leq n/l$ (where $\big[jl, jl+l\big) = \{jl, jl+1, \ldots, jl+l-1\}$). For every $I \in \mathcal{I}$, we define $D_I(x) = 1$ iff for every $i \in I$, $x_i = x_{i-l}$. Note that $D_I(\cdot)$ can be computed by a width-4 oblivious branching program. Note that $\Pr[D_I(X)=1] = 1$ for every $I \in \mathcal{I}$. We suppose, for the sake of contradiction, that $H^{\mathbf{HILL}}_{(1-\epsilon)}(X) > \frac{100}{\log(1/\epsilon)}l^3$. This means in particular that there exists a distribution Y such that $H^{\infty}(Y) \geq \frac{100}{\log(1/\epsilon)}l^3$ but $\Pr[D_I(Y)=1] > \epsilon$ for every $I \in \mathcal{I}$.

For a string $x \in \{0,1\}^n$, we define $S(x) \subseteq [l+1, n]$ to be the set of all indices i such that $x_i \neq x_{i-l}$. The number of strings x such that $|S(x)| \leq \frac{10}{\log(1/\epsilon)}l^2$ is at most $2^l\binom{n}{(10/\log(1/\epsilon))l^2}2^{l^2} \leq 2^{(15/\log(1/\epsilon))l^3}$ (since $l > \log n$). Therefore, $\Pr[|S(Y)| \leq \frac{10}{\log(1/\epsilon)}l^2] \ll \frac{\epsilon}{2}$ (since $H^{\infty}(Y) > \frac{100}{\log(1/\epsilon)}l^3$). We let Y' be the distribution Y conditioned on $|S(Y)| > \frac{10}{\log(1/\epsilon)}l^2$. We note that $\Pr[D_I(Y')=1] > \frac{\epsilon}{2}$ for every $I \in \mathcal{I}$.

We will now show that

$$Ex_{I\leftarrow_{\mathrm{R}}\mathcal{I}, y\leftarrow_{\mathrm{R}}Y'}[D_I(y)] < \tfrac{\epsilon}{2}$$

This will provide us with the desired contradiction, because it implies that in particular there exists $I \in \mathcal{I}$ such that $\Pr[D_I(Y')=1] < \frac{\epsilon}{2}$. We remark that choosing $I \leftarrow_{\mathrm{R}} \mathcal{I}$ can be thought as choosing independently a random index from each block $\big[jl, jl+l\big)$.

Indeed, let $y \leftarrow_{\mathrm{R}} Y'$. We need to prove that $\Pr_{I\leftarrow_{\mathrm{R}}\mathcal{I}}[D_I(y)=1] < \frac{\epsilon}{2}$. Indeed, $D_I(y) = 1$ iff $I \cap S(y) = \emptyset$. Yet, let $S'(y)$ be the a subset of $S(y)$ chosen such that $S'(y)$ contains a single element in each block $[jl, (j+1)l)$ (e.g., $S'(y)$ can be chosen to contain the first element of $S(y)$ in each block). Then, $|S'(y)| \geq \frac{|S(y)|}{l} \geq \frac{10}{\log(1/\epsilon)}l$. Since $S'(y) \subseteq S(y)$, it is enough to prove that $\Pr_{I\leftarrow_{\mathrm{R}}\mathcal{I}}[S'(y) \cap I \neq \emptyset] > 1 - \frac{\epsilon}{2}$.

Yet, for each $i \in S'(y)$, the probability that $i \in I$ (when I is chosen at random from $\mathcal{I}$) is $\frac{1}{l}$ and this probability is independent of the probability that $j \in I$ for every other $j \in S'(y)$ (since $S'(y)$ contains at most a single element in each block). Thus, there is a probability of at least $(1-\frac{1}{l})^{(10/\log(1/\epsilon))l} > 1 - \frac{\epsilon}{2}$ that $S'(y) \cap I \neq \emptyset$. □

7 Analogs of information-theoretic inequalities

7.1 Concatenation lemma

A basic fact in information theory is that for every (possibly correlated) random variables X and Y, the entropy of (X,Y) is at least as large as the entropy of X. We show that if one-way-functions exist then this does not hold for all types of pseudoentropy with respect to polynomial time circuits.

On the other hand, we show that the fact above does hold for polynomial-sized **PH**-circuits and for bounded-width oblivious branching programs.[14]

Negative result for standard model. Our negative result is the following easy lemma:

Lemma 7.1. *Let $G : \{0,1\}^l \to \{0,1\}^n$ be a (poly-time computable) pseudorandom generator.*[15] *Let (X,Y) be the random variables $(G(U_l), U_l)$. Then $H_\epsilon^{\mathbf{HILL}}(X) = n$ (for a negligible ϵ) but $H_{1/2}^{\mathbf{Yao}}(X,Y) \leq l+1$.*

Proof Sketch: $H_\epsilon^{\mathbf{HILL}}(X) = n$ from the definition of pseudorandomness. On the other hand, it is possible to reconstruct (X,Y) from Y alone with probability 1, where $|Y| = l$. □

Positive result for PH-circuits. Our positive result for **PH**-circuits is stated in the following lemma:

Lemma 7.2. *Let X be a random variable over $\{0,1\}^n$ and Y be a random variable over $\{0,1\}^m$. Suppose that $H_\epsilon^{\mathbf{Yao}}(X) \geq k$ with respect to s-sized* **PH***-circuits. Then $H_\epsilon^{\mathbf{Yao}}(X,Y) \geq k$ with respect to $O(s)$-sized* **PH***-circuits.*

Proof. Suppose that $H_\epsilon^{\mathbf{Yao}}(X,Y) < k$. This means that there exist $l \in [k]$ and s-sized **PH**-circuits C, D, where $C : \{0,1\}^{n+m} \to \{0,1\}^\ell$, $D : \{0,1\}^\ell \to \{0,1\}^{n+m}$ such that

$$\Pr_{(x,y) \leftarrow_{\mathrm{R}} (X,Y)}[D(C(x,y)) = (x,y)] > \tfrac{2^l}{2^k} + \epsilon$$

We define $D' : \{0,1\}^\ell \to \{0,1\}^n$ to be the following **PH**-circuit: on input $a \in \{0,1\}^\ell$, compute (x,y) to be $D(a)$ and output x. We define $C' : \{0,1\}^n \to \{0,1\}^\ell$ to be the following **PH**-circuit: on input $x \in \{0,1\}^n$, non-deterministically guess $y \in \{0,1\}^m$ such that $D'(C(x,y)) = x$. If such y is found then output $C(x,y)$. Otherwise, output 0^ℓ. Clearly,

$$\Pr_{x \leftarrow_{\mathrm{R}} X}[D'(C'(x)) = x] \geq \Pr_{(x,y) \leftarrow_{\mathrm{R}} (X,Y)}[D(C(x,y)) = (x,y)] > \tfrac{2^l}{2^k} + \epsilon$$

and thus $H_\epsilon^{\mathbf{Yao}}(X) < k$. □

Applying the results of Section 5.6, we obtain that with respect to **PH**-circuit, the concatenation property is satisfied also for HILL-type and Metric-type pseudoentropy.

Positive result for bounded-width oblivious branching programs. We also show that the concatenation property holds also for metric-type pseudoentropy with respect to bounded-width read-once oblivious branching programs. This is stated in Lemma 7.3. Note that the quality of this statement depends on the order of the concatenation (i.e., whether we consider (X,Y) or (Y,X)).

Lemma 7.3. *Let X be a random variable over $\{0,1\}^n$ and Y be a random variable over $\{0,1\}^m$. Suppose that $H_\epsilon^{\mathbf{Metric}}(X) \geq k$ with respect to width-S read-once oblivious branching programs. Then, $H_\epsilon^{\mathbf{Metric}}(X,Y) \geq k$ and $H_{2\epsilon S}^{\mathbf{Metric}}(Y,X) \geq k - \log(1/\epsilon)$ with respect to such algorithms.*

[14] With respect to the latter, we only prove that concatenation holds for metric-type pseudoentropy.

[15] We mean here a pseudorandom generator in the "cryptographic" sense of Blum, Micali and Yao [BM82, Yao82]. That is, we require that G is polynomial time computable.

Proof Sketch: Suppose that $H_{\epsilon}^{\mathbf{Metric}}(X,Y) < k$. This means that there exists an width-S branching program D such that $\Pr[D(X,Y)=1] \geq \frac{|D^{-1}(1)|}{2^k} + \epsilon$. We consider the following branching program D': On input x, run $D(x)$ and then accept if there exists a possible continuation y such that $D(x,y)=1$. It is not hard to see that $|D'^{-1}(1)| \leq |D^{-1}(1)|$ and $\Pr[D'(X)=1] \geq \Pr[D(X,Y)=1]$.

Suppose now that $H_{2\epsilon S}^{\mathbf{Metric}}(Y,X) < k - \log(1/\epsilon)$. Then there exists an width-S branching program D such that $\Pr[D(Y,X)=1] \geq \frac{|D^{-1}(1)|}{2^k \epsilon} + 2\epsilon S$. In particular, it holds that $\frac{|D^{-1}(1)|}{2^k} \leq \epsilon$. Let α be state of D after seeing $y \leftarrow_{\mathrm{R}} Y$ that maximizes the probability that $D(y, X|Y=y)=1$. We let D' denote the following branching program: on input x, run D on x starting from state α. Again, it is not hard to see that $|D'^{-1}(1)| \leq |D^{-1}(1)|$ and so $\frac{|D'^{-1}(1)|}{2^k} \leq \epsilon$. On the other hand $\Pr[D'(X)=1] \geq \frac{1}{S}\Pr[D(X,Y)=1] \geq 2\epsilon$. Thus $\Pr[D'(X)=1] \geq \frac{|D'^{-1}(1)|}{2^k} + \epsilon$. □

7.2 Unpredictability and entropy

Loosely speaking, a random variable X over $\{0,1\}^n$ is **δ-unpredictable** is for every index i, it is hard to predict X_i from $X_{[1,i-1]}$ (which denotes $X_1,\ldots,X_{i-1}$) with probability better than $\frac{1}{2}+\delta$.

Definition 7.4. Let X be a random variable over $\{0,1\}^n$. We say that X is *δ-unpredictable in index i* with respect to a class of algorithms $\mathcal{C}$ if for every $P \in \mathcal{C}$, $\Pr[P(X_{[1,i-1]}) = X_i] < \frac{1}{2}+\delta$. X is *δ-unpredictable* if for every $P \in \mathcal{C}$ $\Pr[P(i, X_{[1,i-1]}) = X_i] < \frac{1}{2}+\delta$ where this probability is over the choice of X and over the choice of $i \leftarrow_{\mathrm{R}} [n]$. We also define *complement* unpredictability by changing $X_{[1,i-1]}$ to $X_{[n]\setminus\{i\}}$ in the definition above.

Yao's Theorem [Yao82] says that if X is δ-unpredictable in all indices by polynomial-time (uniform or non-uniform) algorithms, then it is $n\delta$-indistinguishable from the uniform distribution. Note that this theorem can't be used for a constant $\delta > 0$. This loss of a factor of n comes from the use of the "hybrid argument" [GM84, Yao82]. In contrast, in the context of information theory it is known that if a random variable X is δ-unpredictable (w.r.t. to all possible algorithms) for a small constant δ and for a constant fraction of the indices, then $H^{\infty}(X) \geq \Omega(n)$. Thus, in this context it is possible to extract $\Omega(n)$ bits of randomness even from δ-unpredictable distributions where δ is a *constant* [TSZS01].

In this section we consider the question of whether or not there exists a computational analog to this information-theoretic statement.

Negative result in standard model. We observe that if one-way functions exist, then the distribution $(G(U_l), U_l)$ where $|G(U_l)| = \omega(l)$) used in Lemma 7.1 is also a counterexample (when considering polynomial-time distinguishers). That is, this is a distribution that is δ-unpredictable for a negligible δ in almost all the indices, but has low pseudoentropy. We do not know whether or not there exists a distribution that is δ-unpredictable for a *constant* δ for *all* the indices, and has sublinear pseudoentropy.

Positive results. We also show some computational settings in which the information theoretic intuition *does* holds. We show this for **PH**-circuits, and for bounded-width oblivious branching programs using the metric definition of pseudoentropy. We start by considering a special case in which the distinguisher has distinguishing probability 1 (or very close to 1).[16]

[16] Intuitively, this corresponds to applications that use the high entoropy distribution for hitting a set (like a disperser) rather than for approximation of a set (like an extractor).

Theorem 7.5. *Let X be a random variable over $\{0,1\}^n$. Suppose there exists a size-s* **PH***-circuit (width-S oblivious branching program) D such that $|D^{-1}(1)| \leq 2^k$ and $\Pr[D(X)=1]=1$. Then there exists a size-$O(s)$* **PH***-circuit (width-S oblivious branching program) P such that*

$$\Pr_{i\in[n],x\leftarrow_R X}[P(x_{[1,i]}) = x_i] \geq 1 - O(\tfrac{k}{n})$$

The main step in the proof of Theorem 7.5 is the following lemma:

Lemma 7.6. *Let $D \subseteq \{0,1\}^n$ be a set such that $|D| < 2^k$. Let $x = x_1 \ldots x_{i-1} \in \{0,1\}^{i-1}$, we define N_x to be the number of continuations of x in D (i.e., $N_x = |\{x' \in \{0,1\}^{n-i} \mid xx' \in D\}|$). We define $P(x)$ as follows:*

$$P(x) = \begin{cases} 1 & \frac{N_{x1}}{N_x} > \frac{2}{3} \\ \\ 0 & \frac{N_{x1}}{N_x} < \frac{1}{3} \end{cases}$$

, where $P(x)$ is undefined otherwise. Then, for every random variable X such that $X \subseteq D$,

$$\Pr_{i\in[n],x\leftarrow_R X}\Big[P(x_{[1,i-1]}) \text{ is defined and equal to } x_i\Big] \geq 1 - O\left(\tfrac{k}{n}\right)$$

Proof. For $x \in \{0,1\}^n$, we let $bad(x) \subseteq [n]$ denote the set of indices $i \in [n]$ such that $P(x_{[1,i-1]})$ is either undefined or different from x_i. We will prove the lemma by showing that $|bad(x)| \leq O(k)$ for *every* string $x \in D$. Note that an equivalent condition is that $|D| \geq 2^{-\Omega(|bad(x)|)}$. Indeed, we will prove that $|D| \geq (1+\frac{1}{2})^{|bad(x)|}$. Let N_i denote the number of continuations of $x_{[1,i]}$ in D (i.e., $N_i = N_{x_{[1,i]}}$). We define $N_n = 1$. We claim that for every $i \in bad(x)$, $N_{i-1} \geq (1+\frac{1}{2})N_i$. (Note that this is sufficient to prove the lemma). Indeed, $N_{i-1} = N_{x_{[1,i-1]}0} + N_{x_{[1,i-1]}1}$, or in other words, $N_{i-1} = N_i + N_{x_{[1,i-1]}\overline{x_i}}$ (where $\overline{x_i} \stackrel{def}{=} 1-x_i$). Yet, if $i \in bad(x)$ then $N_{x_{[1,i-1]}\overline{x_i}} \geq \frac{1}{3}(N_i + N_{x_{[1,i-1]}\overline{x_i}}) \geq \frac{1}{2}N_i$. □ □

We obtain Theorem 7.5 from Lemma 7.6 for the case of **PH**-circuits by observing that deciding whether $P(x)$ is equal to 1 or 0 (in the cases that it is defined) can be done in the polynomial-hierarchy (using approximate counting [JVV86]). The case of bounded-width oblivious branching programs is obtained by observing that the state of the width-S oblivious branching program D after seeing $x_1, \ldots, x_{i-1}$ completely determines the value $P(x_1, \ldots, x_{i-1})$ and so $P(x_1, \ldots, x_{i-1})$ can be computed (non-uniformly) from this state.[17]

We now consider the case that $\Pr_{x\leftarrow_R X}[x \in D] = \epsilon$ for an arbitrary constant ϵ (that may be smaller than $\frac{1}{2}$). In this case we are not able to use standard unpredictability and use *complement unpredictability*.

Theorem 7.7. *Suppose that X is δ-complement-unpredictable for a random index with respect to s-sized* **PH***-circuits, where $\frac{1}{2} > \delta > 0$ is some constant. Let $\epsilon > \delta$ be some constant, then $H_\epsilon^{\mathbf{Metric}}(X) \geq \Omega(n)$ with respect to $O(s)$-sized* **PH***-circuits.*

Proof. We prove the theorem by the contrapositive. Let $\epsilon > \delta$ and suppose that $H_\epsilon^{\mathbf{Metric}}(X) < k$ where $k = \epsilon' n$ (for a constant $\epsilon' > 0$ that will be chosen later). This means that there exists a set $D \in \mathcal{C}$ such that $\Pr_{x\leftarrow_R X}[x \in D] \geq \frac{|D|}{2^k} + \epsilon$. In particular, this means that $|D| < 2^k$ and $\Pr_{x\leftarrow_R X}[x \in D] \geq$

[17] Lemma 7.6 only gives a predictor given a distinguisher D such that $\Pr_{x\leftarrow_R X}[x \in D] = 1$. However, the proof of Lemma 7.6 will still yield a predictor with constant bias even if 1 is replaced by $\frac{9}{10}$ (or any constant greater than $\frac{1}{2}$).

ϵ. We consider the following predictor P': On input $i \in [n]$ and $x = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \in \{0,1\}^{n-1}$, P' considers the strings x^0, x^1 where $x^b = x_1, \ldots, x_{i-1}, b, x_{i+1}, \ldots, x_n$. If both x^0 and x^1 are not in D, then P' outputs a random bit. If $x^b \in D$ and $x^{\overline{b}} \notin D$ then P' outputs b. Otherwise (if $x^0, x^1 \in D$), P' outputs $P(x_1, \ldots, x_{i-1})$, where P is the predictor constructed from D in the proof of Lemma 7.6. Let $\Gamma(D)$ denote the set of all strings x such that $x \notin D$ but x is of Hamming distance 1 from D (i.e., there is $i \in [n]$ such that $x_1, \ldots, x_{i-1}, \overline{x_i}, x_{i+1}, \ldots, x_n \in D$). If $S \subseteq \{0,1\}^n$, then let $X_{\restriction S}$ denote the random variable $X|X \in S$. By Lemma 7.6 $\Pr_{i \in [i], x \leftarrow_R X_{\restriction D}}[P'(x_{[n]\setminus\{i\}}) = x_i] \geq 1 - O(\frac{k}{n})$ while it is clear that $\Pr_{i \in [i], x \leftarrow_R X_{\restriction \{0,1\}^n \setminus (D \cup \Gamma(D))}}[P'(x_{[n]\setminus\{i\}}) = x_i] = \frac{1}{2}$. Thus if it holds that $\Pr[X \in \Gamma(D)] < \epsilon'$ and $k < \epsilon' n$, where ϵ' is some small constant (depending on ϵ and δ) then $\Pr_{i \in [i], x \leftarrow_R X}[P'(x_{[n]\setminus\{i\}}) = x_i] \geq \frac{1}{2} + \delta$ and the proof is finished.

However, it may be the case that $\Pr[X \in \Gamma(D)] \geq \epsilon'$. In this case, we will consider the distinguisher $D^{(1)} = D \cup \Gamma(D)$, and use $D^{(1)}$ to obtain a predictor $P^{(1)\prime}$ in the same way we obtained P' from D. Note that $|D^{(1)}| \leq n|D|$ and that, using non-determinism, the circuit size of $D^{(1)}$ is larger than the circuit size of D by at most a $O(\log n)$ additive factor.[18] We will need to repeat this process for at most $\frac{1}{\epsilon'}$ steps,[19] to obtain a distinguisher $D^{(c)}$ (where $c \leq \frac{1}{\epsilon'}$) such that $|D^{(c)}| \leq n^{O(1/\epsilon')}|D| \leq 2^{k+O(\log n(1/\epsilon'))}$, $\Pr[X \in D^{(c)}] \geq \epsilon$ and $\Pr[X \in \Gamma(D^{(c)})] < \epsilon'$. The corresponding predictor $P^{(c)\prime}$ will satisfy that $\Pr_{i \in [i], x \leftarrow_R X}[P^{(c)\prime}(x_{[n]\setminus\{i\}}) = x_i] \geq \frac{1}{2} + \delta$ thus proving the theorem. □ □

Acknowledgements We thank Oded Goldreich and the RANDOM 2003 referees for helpful comments.

References

[BFNW91] L. Babai, L. Fortnow, N. Nisan, and A. Wigderson. BPP Has Subexponential Time Simulations Unless EXPTIME has Publishable Proofs. *Computational Complexity*, 3(4):307–318, 1993. Preliminary version in Structures in Complexity Theory 1991.

[BYRST02] Z. Bar-Yossef, O. Reingold, R. Shaltiel, and L. Trevisan. Streaming Computation of Combinatorial Objects. In *Conference on Computational Complexity (CCC)*. ACM, 2002.

[BR94] M. Bellare and J. Rompel. Randomness-efficient oblivious sampling. In *Proc. 35th FOCS*. IEEE, 1994.

[BM82] M. Blum and S. Micali. How to Generate Cryptographically Strong Sequences of Pseudo-Random Bits. *SIAM J. Comput.*, 13(4):850–864, Nov. 1984.

[CW79] J. L. Carter and M. N. Wegman. Universal Classes of Hash Functions. *J. Comput. Syst. Sci.*, 18(2):143–154, Apr. 1979.

[GS91] A. V. Goldberg and M. Sipser. Compression and Ranking. *SIAM J. Comput.*, 20(3):524–536, June 1991.

[GM84] S. Goldwasser and S. Micali. Probabilistic Encryption. *Journal of Computer and System Sciences*, 28(2):270–299, Apr. 1984.

[18] To compute $D^{(1)}(x)$, guess $i \in [n], b \in \{0,1\}$ and compute $D(x')$ where x' is obtained from x by changing x_i to b.
[19] Actually, a tighter analysis will show that we only need $O(\log \frac{1}{\epsilon'})$ steps.

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

[ILL89] R. Impagliazzo, L. A. Levin, and M. Luby. Pseudo-Random Generation from One-Way Functions. In *Proc. 21st STOC*, pages 12–24. ACM, 1989.

[INW94] R. Impagliazzo, N. Nisan, and A. Wigderson. Pseudorandomness for network algorithms. In *Proc. 26th STOC*, pages 356–364. ACM, 1994.

[ISW00] R. Impagliazzo, R. Shaltiel, and A. Wigderson. Extractors and pseudo-random generators with optimal seed length. In *Proc. 31st STOC*, pages 1–10. ACM, 2000.

[IW97] R. Impagliazzo and A. Wigderson. **P** = **BPP** if **E** Requires Exponential Circuits: Derandomizing the XOR Lemma. In *Proc. 29th STOC*, pages 220–229. ACM, 1997.

[JVV86] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Comput. Sci.*, 43(2-3):169–188, 1986.

[KvM02] A. R. Klivans and D. van Melkebeek. Graph nonisomorphism has subexponential size proofs unless the polynomial-time hierarchy collapses. *SIAM J. Comput.*, 31(5):1501–1526, 2002.

[MV99] P. B. Miltersen and N. V. Vinodchandran. Derandomizing Arthur-Merlin games using hitting sets. In *Proc. 40th FOCS*, pages 71–80. IEEE, 1999.

[Nis90] N. Nisan. Pseudorandom generators for space-bounded computations. In *Proc. 22nd STOC*, pages 204–212. ACM, 1990.

[Nis96] N. Nisan. Extracting Randomness: How and Why: A Survey. In *Conference on Computational Complexity*, pages 44–58, 1996.

[NT99] N. Nisan and A. Ta-Shma. Extracting Randomness: A Survey and New Constructions. *J. Comput. Syst. Sci.*, 58, 1999.

[NW94] N. Nisan and A. Wigderson. Hardness vs Randomness. *J. Comput. Syst. Sci.*, 49(2):149–167, Oct. 1994.

[NZ93] N. Nisan and D. Zuckerman. Randomness is Linear in Space. *J. Comput. Syst. Sci.*, 52(1):43–52, Feb. 1996. Preliminary version in STOC' 93.

[RRV99] R. Raz, O. Reingold, and S. Vadhan. Extracting all the Randomness and Reducing the Error in Trevisan's Extractors. *J. Comput. Syst. Sci.*, 65, 2002. Preliminary version in STOC' 99.

[Sak96] M. Saks. Randomization and Derandomization in Space-Bounded Computation. In *Conference on Computational Complexity (CCC)*, pages 128–149. ACM, 1996.

[Sha02] R. Shaltiel. Recent developments in extractors. *Bulletin of the European Association for Theoretical Computer Science*, 2002. Available from http://www.wisodm.weizmann.ac.il/~ronens.

[SU01] R. Shaltiel and C. Umans. Simple extractors for all min-entropies and a new pseudo-random generator. In *Proc.* 42*nd FOCS*, pages 648–657. IEEE, 2001.

[Sha48] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 623–656, July , Oct. 1948.

[STV99] M. Sudan, L. Trevisan, and S. Vadhan. Pseudorandom Generators without the XOR Lemma. *J. Comput. Syst. Sci.*, 62, 2001. Preliminary version in STOC' 99.

[TSZS01] A. Ta-Shma, D. Zuckerman, and S. Safra. Extractors from Reed-Muller codes. In *Proc.* 42*nd FOCS*, pages 638–647. IEEE, 2001.

[Tre99] L. Trevisan. Construction of Extractors Using Pseudo-Random Generators. In *Proc.* 31*st STOC*, pages 141–148. ACM, 1999.

[TV00] L. Trevisan and S. Vadhan. Extracting Randomness from Samplable Distributions. In *Proc.* 41*st FOCS*, pages 32–42. IEEE, 2000.

[vN28] J. von Neumann. Zur Theorie der Gesellschaftsspiele. *Math. Ann.*, 100:295–320, 1928.

[Yao82] A. C. Yao. Theory and applications of trapdoor functions. In *Proc.* 23*rd FOCS*, pages 80–91. IEEE, 1982.