comput. complex.

DOI 10.1007/s00037-019-00177-4

computational complexity

PREDICTION FROM PARTIAL INFORMATION AND HINDSIGHT, WITH APPLICATION TO CIRCUIT LOWER BOUNDS

Or Meir and Avi Wigderson

Abstract. Consider a random sequence of n bits that has entropy at least $n - k$, where $k \ll n$. A commonly used observation is that an average coordinate of this random sequence is close to being uniformly distributed, that is, the coordinate "looks random." In this work, we prove a stronger result that says, roughly, that the average coordinate looks random to an adversary that is allowed to query $\approx \frac{n}{k}$ other coordinates of the sequence, even if the adversary is non-deterministic. This implies corresponding results for decision trees and certificates for Boolean functions.

As an application of this result, we prove a new result on depth-3 circuits, which recovers as a direct corollary the known lower bounds for the parity and majority functions, as well as a lower bound on sensitive functions due to Boppana (Circuits Inf Process Lett 63(5):257–261, 1997). An interesting feature of this proof is that it works in the framework of Karchmer and Wigderson (SIAM J Discrete Math 3(2):255–265, 1990), and, in particular, it is a "top-down" proof (Håstad *et al.* in Computat Complex 5(2):99–112, 1995). Finally, it yields a new kind of a random restriction lemma for non-product distributions, which may be of independent interest.

Keywords. Certificate complexity, Circuit complexity, Circuit complexity lower bounds, Decision tree complexity, Information theoretic, Query complexity, Sensitivity

Subject classification. 68Q15

Published online: 06 April 2019

Birkhäuser

1. Introduction

1.1. Background and main result. Let $X \in \{0,1\}^n$ be a random variable such that $H(X) \geq n - k$, where $k \ll n$ and $H(X)$ is the Shannon entropy of X. By the subadditivity of entropy, we know that an average coordinate X_i of X has entropy close to 1, which means that it is close to being uniformly distributed. Indeed, the average value of $H(X_i)$ for a uniformly chosen coordinate $i \in [n]$) is at least $1 - k/n$. Putting it differently, in terms of prediction, an adversary, who knows the distribution of X as well as the value of the index i chosen uniformly, has only negligible advantage in guessing the value of X_i.

This simple observation (and its generalization to strings over larger alphabets) turns out to be extremely useful and is a crucial ingredient in the proof of many important results such as the parallel repetition theorem (Raz 1998), lower bounds on the communication complexity of set-disjointness (Bar-Yossef *et al.* 2004; Razborov 1992b), lower bounds on the round complexity of communication protocols (e.g., Duris *et al.* 1987; Mcgeoch 1986; Nisan & Wigderson 1993; Papadimitriou & Sipser 1984), composition theorems for communication protocols (Dinur & Meir 2016; Edmonds *et al.* 2001), lower bounds on interactive coding and interactive compression (e.g., Ganor *et al.* 2014; Kol & Raz 2013) and the construction of extractors (Nisan & Zuckerman 1996).

In this work, we prove a generalization of this observation. Specifically, we consider the setting in which the adversary is stronger in the following way: Besides knowing the distribution of X and the randomly chosen index i, the adversary is allowed to query q other coordinates of X before it tries to guess X_i. Our main result says, roughly, that the adversary cannot guess X_i with non-negligible advantage even after querying $q = O(n/k)$ coordinates of X. Moreover, this holds even if the adversary is allowed to choose the queries non-deterministically. We note that while our adversary model is non-standard, it generalizes the two standard models of decision trees and certificates (see Section 1.1.1).

More specifically, our prediction model is the following. The adversary is given the distribution X and the random coordinate i, and a parameter $\varepsilon > 0$ (here ε is the *bias parameter*). The adver-

sary first makes q non-deterministic queries to *other* coordinates in the sequence.[1] The adversary is *successful* on coordinate i if *some* choice of q queries result in answers to the queries which enable it to guess X_i with advantage ε, namely with success probability at least $\frac{1}{2} + \frac{1}{2} \cdot \varepsilon$. We prove that for the average coordinate and for a random sample from the distribution X, the adversary is successful in having such advantage ε only with very small probability. In particular, for any fixed $\varepsilon > 0$, this success probability goes to 0 as long as $q = o(n/k)$.

One way to understand the non-determinism of the adversary is by defining, for each coordinate i, a set of "witnesses" for good prediction, each over q coordinates in $[n]$. Conditioned on the event that *at least one* of these witnesses occurs in the given sample of X, the distribution of X_i has a bias of ε. We proceed to give the formal definition and result.

Definition 1.1. *A* witness *for a coordinate $i \in [n]$ is a pair (Q, a) where $Q \subseteq [n] - \{i\}$ and $a \in \{0,1\}^{|Q|}$. The witness* appears *in a string $x \in \{0,1\}^n$ if $x|_Q = a$. The* length *of the witness is $|Q|$.*

Definition 1.2. *A q-*family of witnesses *F for a coordinate $i \in [n]$ is a set of witnesses for i of length at most q. We say that a string $x \in \{0,1\}^n$* satisfies *F if at least one of the witnesses in F appears in x. For a random string $X \in \{0,1\}^n$, a bit $b \in \{0,1\}$ and $0 \leq \varepsilon \leq 1$, we say that F ε-predicts $X_i = b$ if*

$$\Pr\left[X_i = b | X \text{ satisfies } F\right] \geq \frac{1}{2} + \frac{1}{2} \cdot \varepsilon.$$

Using the above definitions, an adversary is simply a pair (F^0, F^1) such that F^b is a q-family of witnesses that ε-predicts $X_i = b$. Our main theorem says that for the average coordinate i, the probability that X satisfies either F^0 or F^1 is small.

Theorem 1.3 (Main theorem). *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, and let $q \in \mathbb{N}$,*

[1] Being non-deterministic, it does not matter if these queries are adaptive or not.

$0 \leq \varepsilon \leq 1$. *Suppose for every coordinate* $i \in [n]$ *there is a pair* (F_i^0, F_i^1) *such that* F_i^b *is a* q*-family of witnesses for* i *that* ε*-predicts* $X_i = b$*, and let* δ_i *denote the probability that a string drawn from* X *satisfies either* F_i^0 *or* F_i^1*. Then, the average value of* δ_i *over* $i \in [n]$ *is at most* $\frac{300 \cdot k \cdot q}{\varepsilon^3 \cdot n}$.

We note that this result is almost tight, as is demonstrated by the following example. We partition the string X to k blocks of length $\frac{n}{k}$. Now, suppose that X is a uniformly distributed string such that the parity of each block is 0. Then, the adversary can guess every coordinate X_i with probability 1 by querying $\frac{n}{k}-1$ other coordinates: The adversary will simply query all the other coordinates in the block of X_i, and output their parity. Note that in this example, the adversary does not need to use non-determinism, and does not even need to be adaptive.

Remark 1.4. *We note that a* q*-family of witnesses* F *can be viewed alternatively as a DNF formula of width at most* q*, where a string* x *satisfies* F *if the formula outputs* 1 *on* x*. Taking this view, the adversary defines a pair of DNF formulas* (ϕ_0, ϕ_1)*, and guesses that* $X_i = b$ *if* $\phi_b(X) = 1$*. It is an interesting open problem to generalize this result to adversaries that use constant-depth circuits rather than DNFs. Ajtai (1992) proved a result in a similar spirit in the special case where* X *is distributed uniformly over all strings of some fixed Hamming weight* $n^{\Omega(1)}$.

Follow-up works. Following an early version of this work, a recent work of Smal & Talebanfard (2018) has claimed to improve the bound in Theorem 1.3 to $\frac{k \cdot (q+1)}{(1-H(\frac{1}{2}+\frac{\varepsilon}{2})) \cdot n}$. However, this claim has later been retracted. The current version of this paper proves this bound for a special case of Theorem 1.3, which in turn implies this bound for our Corollary 1.5. The paper also improves the constants in our Corollary 1.6.

In addition, a recent work of Grinberg *et al.* (2018) proved a lemma about the ability of decision trees to distinguish certain distributions. They observe that their lemma also follows from our Corollary 1.5.

Very recently, Viola (2018) generalized our Theorem 1.3 to constant-depth circuits for the special case where every δ_i is either 1 or 0, thus solving almost completely the open problem posed in Remark 1.4.

1.1.1. Applications to decision trees and certificates. While our model of adversary is somewhat non-standard, our main theorem has immediate consequences for two standard models, namely, decision trees and certificates.

We start by discussing the application to decision trees, which correspond to deterministic adaptive adversaries. Given a random string X and a coordinate i, we say that a decision tree ε-*predicts* X_i if the decision tree makes queries to the coordinates in $[n] - \{i\}$ and outputs the value of X_i correctly with probability at least $\frac{1}{2} + \frac{1}{2} \cdot \varepsilon$. We prove the following direct corollary of Theorem 1.3.

Corollary 1.5. *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, and let $q \in \mathbb{N}$, $0 \leq \varepsilon \leq 1$. Then, the number of coordinates $i \in [n]$ that are ε-predicted by some decision tree that makes at most q queries is at most $\frac{300 \cdot k \cdot q}{\varepsilon^3}$.*

We turn to discuss certificates, which correspond to a non-deterministic adversary that predicts coordinates with perfect accuracy (i.e., $\varepsilon = 1$). Given a random string $X \in \{0,1\}^n$, a coordinate $i \in [n]$ and a bit $b \in \{0,1\}$, a b-certificate for i is a witness (Q, a) such that

$$\Pr\left[X_i = b | X|_Q = a\right] = 1.$$

In the context of certificates, we do not need to discuss families of witnesses, since it easy to see that the best strategy for the adversary is to take F_i^b to be the family of all b-certificates for X_i. We prove the following direct corollary of Theorem 1.3.

Corollary 1.6. *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, and let $q \in \mathbb{N}$. For every coordinate $i \in [n]$, we denote by δ_i the probability that any certificate for X_i of length at most q appears in X. Then, the average value of δ_i over $i \in [n]$ is at most $\frac{300 \cdot k \cdot q}{n}$.*

Connection to the satisfiability coding lemma. Following an earlier version of this work, it was brought to our attention that Corollary 1.6 could also be derived from the satisfiability coding lemma of Paturi *et al.* (1999). The aforementioned recent work of Smal & Talebanfard (2018) has made the connection apparent by giving a simpler and stronger proof of Corollary 1.6 by combining our ideas with those of Paturi *et al.* (1999).

1.1.2. Observation on random restrictions. In most random restriction arguments (which are most typically apply to their effect on DNF formulae), a random subset of the coordinates is chosen to be fixed, and then each is fixed independently at random. Some generalizations of this were found useful, in which the values to the fixed variables are not independent, but are still quite structured (see, e.g., the primer on random restrictions Beame 1994 and the recent lower bounds Chen *et al.* 2016; Pitassi *et al.* 2016 for such examples). Here we consider a rather general form of a random restriction argument, in which the values to the coordinates to be fixed are chosen from an arbitrarily correlated random variable X, according to its marginals. The following result follows from our proof of Theorem 1.3, and may be interesting in its own right.

PROPOSITION 1.7. *Let ϕ be a DNF formula over n variables of width at most w, and let X be a random variable that is distributed arbitrarily in $\{0,1\}^n$ such that $\phi(X) = 1$ with probability δ. Let ρ be a random restriction that fixes each variable with probability at least p independently, and that chooses the values of the fixed variables according to the marginal distribution of X on those variables. Then, $\phi|_\rho$ is fixed to 1 with probability at least $p^w \cdot \delta$.*

See Section 3.3 for the proof of this proposition.

1.2. Application to circuit lower bounds. Proving circuit lower bounds is a central challenge of complexity theory. Unfortunately, proving even super-linear lower bounds for general circuits seems to be beyond our reach at this stage. In order to make progress and develop new proof techniques, much of the current research focuses on proving lower bounds for restricted models of circuits.

One of the simplest restricted models that are not yet fully understood is circuits of constant depth, and in particular, circuits of depth 3. By a standard counting argument, we know that there exists a non-explicit function that requires such circuits of size $\Omega(2^n/n)$. On the other hand, the strongest lower bound we have for an explicit function (Ajtai 1983; Furst *et al.* 1984; Håstad 1986) says that circuits of depth d computing the parity of n bits must be of size $2^{\Omega(n^{1/(d-1)})}$ (and in particular, depth-3 circuits must be of size $2^{\Omega(\sqrt{n})}$). Hence, while strong lower bounds are known in this model, there is still a significant gap in our understanding. It is therefore important to develop new techniques for analyzing such circuits.

An important insight about constant-depth circuits is that such circuits cannot compute sensitive functions. Given a function $f : \{0,1\}^n \to \{0,1\}$ and an input $x \in \{0,1\}^n$, the *sensitivity of f at x* is the number of coordinates $i \in [n]$ such that flipping the i-th bit of x changes the value of f. The *average sensitivity of f* is the average of the sensitivities of f over all inputs. The following theorem of Boppana (1997), which improves on a result of Linial *et al.* (1993), gives a lower bound on functions in terms of their average sensitivity.

THEOREM 1.8 (Boppana 1997). *There exists a constant $\gamma > 0$ such that every function $f : \{0,1\}^n \to \{0,1\}$ with average sensitivity s requires depth-d circuits of size $2^{\gamma \cdot s^{1/(d-1)}}$.*

This theorem of Boppana (1997) can be viewed as a powerful generalization of the aforementioned lower bound on the parity function. In particular, note that it implies that lower bound as a special case, since it is easy to see that the average sensitivity of the parity function is n. However, there are some functions whose hardness for constant-depth circuits is not captured by this theorem. For example, it is known that the majority function requires depth-3 circuits of size $2^{\Omega(\sqrt{n})}$ (Håstad 1986), but Theorem 1.8 only gives a lower bound of $2^{\Omega(n^{1/4})}$ for majority, since its average sensitivity is $\theta(\sqrt{n})$.

In this work, we show that Theorem 1.3 can be used rather easily to prove a generalization of the theorem of Boppana (1997)

for depth 3 that also captures the latter lower bound for majority. This generalization proves a lower bound on a function based on the condition it has a significant fraction of sensitive inputs, even if the average input is not very sensitive.

THEOREM 1.9. *There exists a constant $\gamma > 0$ such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a function has sensitivity at least s on at least $\alpha \cdot 2^n$ inputs in $f^{-1}(0)$ for some $0 < \alpha < 1$ (respectively, $f^{-1}(1)$). Then every depth-3 circuit that computes f whose top gate is an AND gate (respectively, OR gate) must be of size at least $\frac{\alpha}{n} \cdot 2^{\gamma \cdot \sqrt{s}}$.*

It is easy to see that Theorem 1.9 shows a depth-3 lower bound of $2^{\Omega(\sqrt{n})}$ for majority: For the majority function, all the inputs whose Hamming weight is about $\frac{n}{2}$ have sensitivity $s = \frac{n}{2}$, and there is about $\alpha = \frac{1}{\sqrt{n}}$ fraction of such inputs. Furthermore, it implies Theorem 1.8 for the special case of depth-3 circuits, under the mild condition that the average sensitivity of f is at least $O(\log^2 n)$: To see why, observe that a function $f : \{0,1\}^n \to \{0,1\}$ with average sensitivity s must have sensitivity at least $s/2$ on at least $\frac{s}{2n}$ fraction of the inputs. Since $\frac{s}{2n} > \frac{1}{n}$, we can apply Theorem 1.9 with sensitivity $s/2$ and $\alpha = \frac{1}{n}$ and get a lower bound of $2^{\Omega(\sqrt{s})}$.

We note that Ajtai proved the following similar (but incomparable) result, which works for every constant depth:

THEOREM 1.10 (Ajtai 1993). *For every natural number d, there exists $\beta > 0$ such that for every sufficiently large $n \in \mathbb{N}$ the following holds. If a function $f : \{0,1\}^n \to \{0,1\}$ has sensitivity at least $n^{1-\beta}$ on at least $2^{-n^\beta} \cdot 2^n$ inputs, then f requires depth-d circuits of size at least 2^{n^β}.*

Theorem 1.10 is stronger than our Theorem 1.9 in the sense that it works for every constant depth, but is weaker in the sense that it works only for a very large sensitivity. It is an interesting question whether our Theorem 1.9 could be extended to larger depths. This would give a more refined understanding of the connection between sensitivity and constant-depth lower bounds.

Remark 1.11. *In fact, in order to prove Theorem 1.9, we do not need the full power of Theorem 1.3—the corollary for certificates (Corollary 1.6) is sufficient.*

On Karchmer–Wigderson relations. An interesting feature of our proof of Theorem 1.9 is that it uses the framework of Karchmer–Wigderson relations. Karchmer & Wigderson (1990) observed that for every function $f : \{0,1\}^n \to \{0,1\}$ there is a corresponding communication problem R_f such that the depth complexity of f is tightly related to the deterministic communication complexity of R_f. This correspondence allows us to attack questions about circuits using tools from communication complexity.

While this framework has been very successful in proving lower bounds for monotone circuits (Grigni & Sipser 1991; Karchmer *et al.* 1995b; Karchmer & Wigderson 1990; Klawe *et al.* 1984; Raz & Wigderson 1992), so far had less success in the non-monotone setting. One reason is that in the non–monotone setting it is impossible to prove lower bounds better than n^2 on R_f using techniques that work against *randomized* protocols (Gavinsky *et al.* 2014; Raz & Wigderson 1989), and for constant-depth circuits, it is impossible to prove super-polynomial lower bounds using such techniques Jowhari *et al.* 2011; Meir 2017. Indeed, this barrier was bypassed only recently in the context of formula lower bounds Dinur & Meir 2016. This work is the first time that the framework of Karchmer and Wigderson is used to prove lower bounds for constant-depth circuits in the non-monotone setting (although it is related to the top-down technique for the same purpose described next).

On top-down vs. bottom-up techniques. Håstad *et al.* (1995) proposed to distinguish between two types of techniques for proving circuit lower bounds. "Bottom-up techniques" are techniques that start by analyzing the bottom layer of the circuit (the inputs layer) and then proceed to analyze higher layers—the canonical example of such techniques is the switching lemma and the proofs that are based on it (Håstad 1986). "Top-down techniques," on the other hand, are techniques that start by analyzing the top layer and then proceed to analyze lower layers—two canonical examples of

such techniques are the Karchmer–Wigderson framework and techniques that are based on formal complexity measures (Razborov 1992a) (e.g., the method of Khrapchenko (Khrapchenko 1972)).

Håstad *et al.* (1995) observed that all the techniques that were used to prove constant-depth lower bounds until that time were bottom-up techniques. They argued that it would be valuable to develop top-down approaches for constant-depth lower bounds in order to deepen our understanding and extend our array of techniques. They then showed how to prove the depth-3 lower bounds of $2^{\Omega(\sqrt{n})}$ for parity and majority using such a top-down proof. Their approach bears much similarity to the approach of Karchmer and Wigderson, but there are some differences.

Our proof of Theorem 1.9 provides a second example for a top-down proof of constant-depth lower bounds. One notable difference between our work and Håstad *et al.* (1995) is that Håstad *et al.* (1995) give two separate proofs for the lower bounds for parity and majority. While these two proofs share a common framework, each of them still requires some different non-trivial ideas. Our Theorem 1.9, on the other hand, is powerful enough to imply both these results as well as Theorem 1.8 of Boppana (1997).

Connection to the satisfiability coding lemma. An anonymous referee has pointed out that our Theorem 1.9 could also be proved directly using the satisfiability coding lemma of Paturi *et al.* (1999) (rather than going through Corollary 1.6). The satisfiability coding lemma implies that a CNF formula of width q can have at most $2^{n-\frac{s}{q}}$ satisfying assignments of sensitivity s. This could be used to prove Theorem 1.9 as follows: Let $f : \{0,1\}^n \to \{0,1\}$ be a function that has sensitivity at least s on at least $\alpha \cdot 2^n$ inputs in $f^{-1}(1)$, and suppose that it can be computed by a depth-3 circuit C of size T whose top gate is an OR gate. Without loss of generality, we may assume that the circuit C has bottom fan-in $\approx \log T$ (this can be guaranteed by applying a random restriction that fixes every variable with probability $\frac{1}{2}$).

Now, C is a disjunction of at most T CNF formulas, so at least one of these formulas contributes at least $\alpha \cdot 2^{n-\log T}$ inputs in $f^{-1}(1)$ that have sensitivity s. On the other hand, this formula has width

at most $\log T$, and therefore by the satisfiability coding lemma, it can contribute at most $2^{n-\frac{s}{\log T}}$ such inputs. It therefore holds that $2^{n-\frac{s}{\log T}} \geq \alpha \cdot 2^{n-\log T}$, which implies $T \geq 2^{\Omega(\sqrt{s})}$, as required.

1.3. Certificates for sets of coordinates. In Section 1.1.1, we discussed an application of our main theorem to certificates, which correspond to an adversary that predicts a coordinate with perfect accuracy. In this section, we discuss an extension of this result to adversaries that attempt to predict a *set of coordinates*. In addition to being interesting in its own right, we believe that this extension might be useful for generalizing our lower bound for depth-3 circuits to higher depths.

In order to explain this extension, we take a slightly different view of certificates. Recall that a certificate is a witness (Q, a) such that conditioned on $X|_Q = a$, the value of X_i is known with certainty. A different way to phrase this definition is to say that conditioned on $X|_Q = a$, the random variable X_i does not have full support. This leads to the following generalization of certificates to sets of coordinates.

Definition 1.12. *Let X be a random variable taking values from $\{0,1\}^n$, let $R \subseteq [n]$ be a set of coordinates. A* certificate for R *(*with respect to X*) is a pair (Q, a) where $Q \subseteq [n] - R$ and $a \in \{0,1\}^{|Q|}$, such that conditioned on $X|_Q = a$, the random variable $X|_R$ does not have full support. The* length *of the certificate is $|Q|$, and we say that a string $x \in \{0,1\}^n$* satisfies the certificate *if $x|_Q = a$.*

Our corollary for certificates (Corollary 1.6) said that for an average coordinate $i \in [n]$, the string X does not satisfy any certificate for X_i *with high probability*. Our result for sets of coordinates is not as strong: It only says that for an average set of coordinates $R \subseteq [n]$, the string X does not satisfy any certificate for R with probability that is *non-trivial* (but is exponentially vanishing in $|R|$). Still, this result could be useful in certain applications—for example, our Theorem 1.9 could be proved even using a theorem that suffers from such a limitation. We have the following result.

THEOREM 1.13. *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, let $r, q \in \mathbb{N}$, and assume that $(q+r)\cdot(2k+r+1) \leq \frac{1}{4000}\cdot n$. For every set of coordinates $R \subseteq [n]$ of size r, we denote by p_R the probability that a string drawn from X does not satisfy* any *certificate for R of length at most q. Then, the average value of p_R over $R \subseteq [n]$ is at least 2^{-r-1}.*

Observe that the certificates of Definition 1.12 correspond to a very strong adversary: The adversary makes at most q queries to the coordinates in $[n] - R$ non-deterministically, and then it is considered successful even if it only managed to rule out the possibility that $X|_R = b$ for a single string $b \in \{0,1\}^{|R|}$. Theorem 1.13 establishes limits even against such powerful adversaries.

Open problems. As mentioned above, it would be interesting to generalize both our main theorem and our Theorem 1.9 to circuits of higher depth (again, Viola 2018 achieved it for a special case of our main theorem). An additional interesting question is whether the improved bound of Smal & Talebanfard (2018) for our main theorem is tight for every $\varepsilon > 0$. Moreover, since our main theorem could be viewed as a generalization of the satisfiability coding lemma, it would also be interesting to find applications for this theorem that cannot be derived from that lemma.

Organization of the paper. We cover the required preliminaries in Section 2. We prove our main result (Theorem 1.3) and its corollaries in Section 3 and present the application to circuit lower bounds (Theorem 1.9) in Section 4. Finally, we prove the result on certificates for sets of coordinates in Section 5.

2. Preliminaries

For $n \in \mathbb{N}$, we denote $[n] \stackrel{\text{def}}{=} \{1, \ldots n\}$. Given a string $x \in \{0,1\}^n$ and a set of coordinates $I \subseteq [n]$, we denote by $x|_I$ the projection of x to the coordinates in I.

2.1. Information theory. We use basic concepts from information theory, see Cover & Thomas (1991) for more details.

Definition 2.1 (Entropy). *The* entropy *of a random variable X is*

$$H(X) \stackrel{\text{def}}{=} \mathbb{E}_{x \leftarrow X}\left[\log \frac{1}{\Pr[X=x]}\right] = \sum_x \Pr[X=x] \cdot \log \frac{1}{\Pr[X=x]}.$$

Given a random variable Y, the conditional entropy *$H(X|Y)$ is defined to be $\mathbb{E}_{y \leftarrow Y}[H(X|Y=y)]$.*

Fact 2.2. *$H(X)$ is lower bounded by 0 and is upper bounded by the logarithm of the size of the support of X. The lower bound is achieved when X is a fixed value, and the upper bound is achieved when X is uniformly distributed.*

The conditional entropy $H(X|Y)$ is lower bounded by 0 and is upper bounded by $H(X)$. The lower bound is achieved when X is a function of Y, and the upper bound is achieved when X is independent of Y.

The following useful fact is a special case of the data processing inequality. Intuitively, it says that if X, Y, Z are random variables and Z is a function of Y, then Z cannot give more information on X than Y.

Fact 2.3. *Let X, Y, Z be random variables, such that Z is determined by Y. Then $H(X|Y) \leq H(X|Z)$.*

Fact 2.4 (The chain rule). *Let X, Y be random variables. Then $H(X,Y) = H(X|Y) + H(Y)$.*

Facts 2.2 and 2.4 imply that entropy is subadditive.

Corollary 2.5 (The subadditivity of entropy). *Let X, Y be random variables. Then $H(X,Y) \leq H(X) + H(Y)$.*

We also define the binary entropy function, which will be useful in the proof of our main theorem.

Definition 2.6 (Binary entropy function). *The* binary entropy function $H : [0,1] \to [0,1]$ *is the function defined by*

$$H(x) = x \cdot \log \frac{1}{x} + (1-x) \cdot \log \frac{1}{1-x},$$

and by $H(0) = H(1) = 0$. *In other words,* $H(p)$ *is the entropy of a binary random variable that takes one value with probability* p *and the other value with probability* $1-p$.

The following approximation of the binary entropy function, which follows from its Taylor expansion, is useful.

Fact 2.7. *Let* $0 \leq \varepsilon \leq 1$. *Then* $H(\frac{1}{2} - \frac{1}{2} \cdot \varepsilon) = H(\frac{1}{2} + \frac{1}{2} \cdot \varepsilon) \geq 1 - \frac{1}{2} \cdot \varepsilon^2$.

We also define the notion of "min-entropy," which will be used in the proof of the result on certificates for sets of coordinates.

Definition 2.8. *The* min-entropy *of a random variable* X *is*

$$H_\infty(X) = \min_x \left\{ \log \frac{1}{\Pr[X = x]} \right\}.$$

In other words, $H_\infty(X)$ *is the smallest number* h *such that* $\Pr[X = x] = 2^{-h}$ *for some* x.

One useful feature of min-entropy is that it behaves nicely under conditioning:

Fact 2.9. *Let* X *be a random variable, and let* E *be an event. Then* $H_\infty(X|E) \geq H_\infty(X) - \log \frac{1}{\Pr[E]}$.

Proof. For every value x, it holds that

$$\Pr[X = x|E] = \frac{\Pr[X = x \wedge E]}{\Pr[E]} \leq \frac{\Pr[X = x]}{\Pr[E]} \leq 2^{-H_\infty(X) + \log \frac{1}{\Pr[E]}}.$$

It therefore follows that $H_\infty(X|E) \geq H_\infty(X) - \log \frac{1}{\Pr[E]}$, as required. □

The following fact allows us to transform a random variable that has high entropy into one that has high min-entropy.

FACT 2.10. *Let X be a random variable taking values from a set $\mathcal{X}$ such that $H(X) \geq \log|\mathcal{X}| - k$. Then there is an event E of probability at least $\frac{1}{2}$ such that $H_\infty(X|E) \geq \log|\mathcal{X}| - 2k - 1$.*

PROOF. Let E be the event that X takes a value x that satisfies $\Pr[X = x] \geq 2^{-(\log|\mathcal{X}|-2k)}$. We claim that E has probability at least $\frac{1}{2}$: To see why, observe by Markov's inequality and the fact that $H(X) \leq \log|\mathcal{X}|$, it holds with probability at least $\frac{1}{2}$ that

$$\log|\mathcal{X}| - \log\frac{1}{\Pr[X = x]} \leq 2k$$

or in other words $\Pr[X = x] \leq 2^{-(\log|\mathcal{X}|-2k)}$. Next, for every value x in the support of $X|E$ it holds that

$$\Pr[X = x|E] \leq \frac{\Pr[X = x]}{\Pr[E]} \leq 2^{-(\log|\mathcal{X}|-2k-1)}.$$

It follows that $H_\infty(X|E) \geq \log|\mathcal{X}| - 2k - 1$, as required. □

2.2. Karchmer–Wigderson relations. Let $f : \{0,1\}^n \to \{0,1\}$ be a non-constant function. The *Karchmer–Wigderson relation of f*, denoted R_f, is the following communication problem: Alice gets a string $x \in f^{-1}(0)$, Bob gets a string $y \in f^{-1}(1)$, and they wish to find a coordinate $j \in [n]$ such that $x_j \neq y_j$. There is a tight connection between protocols for R_f and formulas that compute f (Karchmer & Wigderson 1990) (see also Gavinsky *et al.* 2014; Karchmer *et al.* 1995a; Razborov 1990). The following proposition is a direct corollary of this connection.

PROPOSITION 2.11. *Let $f : \{0,1\}^n \to \{0,1\}$ be a non-constant function. If there is a depth-3 circuit of size S that computes f whose top gate is an AND gate, then there is a 3-round protocol that solves R_f of the following form:*

- *In the first round, Alice sends to Bob a message of at most $\log S$ bits.*
- *In the second round, Bob sends to Alice a message of at most $\log S$ bits.*

- *In the third round, Alice sends to Bob the solution $j \in [n]$.*

If there is such a circuit whose top gate is an OR gate, then there is such a protocol with the roles of Alice and Bob being reversed.

Therefore, if we wish to prove lower bounds on depth-3 circuits computing f, it suffices to prove lower bounds on the communication complexity of protocols of the foregoing form.

3. The main theorem and its corollaries

In this section, we prove our main theorem, restated next, and its corollaries regarding decision trees and certificates.

Definition (1.1, restated). *A* witness *for a coordinate $i \in [n]$ is a pair (Q, a) where $Q \subseteq [n] - \{i\}$ and $a \in \{0,1\}^{|Q|}$. The witness* appears *in a string $x \in \{0,1\}^n$ if $x|_Q = a$. The* length *of the witness is $|Q|$.*

Definition (1.2, restated). *A q-*family of witnesses *F for a coordinate $i \in [n]$ is a set of witnesses for i of length at most q. We say that a string $x \in \{0,1\}^n$* satisfies *F if at least one of the witnesses in F appears in x. For a random string $X \in \{0,1\}^n$, a bit $b \in \{0,1\}$ and $0 \leq \varepsilon \leq 1$, we say that F ε-predicts $X_i = b$ if*

$$\Pr\left[X_i = b | X \text{ satisfies } F\right] \geq \frac{1}{2} + \frac{1}{2} \cdot \varepsilon.$$

Theorem (1.3, restated). *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n-k$, and let $q \in \mathbb{N}$, $0 \leq \varepsilon \leq 1$. Suppose for every coordinate $i \in [n]$ there is a pair (F_i^0, F_i^1) such that F_i^b is a q-family of witnesses for i that ε-predicts $X_i = b$, and let δ_i denote the probability that a string drawn from X satisfies either F_i^0 or F_i^1. Then, the average value of δ_i over $i \in [n]$ is at most $\frac{300 \cdot k \cdot q}{\varepsilon^3 \cdot n}$.*

The rest of this section is organized as follows: In Section 3.1, we describe the high-level idea of the proof of the main theorem. Then, in Section 3.2, we give the full proof of the theorem. Next,

in Section 3.3, we derive our observation on random restrictions (Proposition 1.7). Finally, in Section 3.4, we derive the applications to decision trees and certificates.

3.1. Proof idea. As a warm-up, let us consider the simpler problem of proving limitations of a deterministic, non-adaptive adversary. Such an adversary is defined as follows: In order to predict the coordinate X_i, the adversary chooses a priori a set of queries $Q_i \subseteq [n] - \{i\}$ of size q. The adversary then gets to see $X|_{Q_i}$ and makes a guess for X_i based on this string. For simplicity, we assume that the adversary may not err, that is, if the adversary was told that $X|_{Q_i} = a$ and then guessed that $X_i = b$, then it must be the case that

$$\Pr\left[X_i = b \,|X|_{Q_i} = a\right] = 1.$$

In other words, this means that the coordinate X_i must be a deterministic function of $X|_{Q_i}$. Now, suppose we wish to prove that if the number of queries q is less than $\frac{n}{k} - 1$, there exists a coordinate that such an adversary cannot guess using q queries.

Suppose for the sake of contradiction that such an adversary can predict *every* coordinate $i \in [n]$ of X. We prove that in such case, the entropy $H(X)$ must be smaller than $n - k$, contradicting our assumption. To this end, we choose a sequence of sets of coordinates, and use them to upper bound the entropy of X. Consider the following process: Let i_1 be an arbitrary coordinate, and let Q_{i_1} be the corresponding set that the adversary chooses. Let i_2 be an arbitrary coordinate in $[n] - (\{i_1\} \cup Q_{i_1})$, and let Q_{i_2} be the corresponding set. Let i_3 be an arbitrary coordinate in $[n] - (\{i_1\} \cup Q_{i_1} \cup \{i_2\} \cup Q_{i_2})$, and let Q_{i_3} be the corresponding set. We continue in this manner until there are no more coordinates that are left to choose, that is, until $\{i_1\} \cup Q_{i_1} \cup \ldots$ covers all the coordinates, and let $i_1, \ldots, i_t$ the coordinates that were chosen in this process. In each iteration, we removed at most $q + 1$ coordinates, and therefore the number of iterations is $t \geq \frac{n}{q+1}$. By the chain rule (Fact 2.4), it holds that

$$\begin{aligned} H(X) = H\left(X|_{Q_{i_1}}, X|_{Q_{i_2}}, \ldots, X|_{Q_{i_t}}\right) \\ + H\left(X_{i_1}, X_{i_2}, \ldots, X_{i_t} \,\middle|\, X|_{Q_{i_1}}, X|_{Q_{i_2}}, \ldots, X|_{Q_{i_t}}\right) \end{aligned}$$

Now, by assumption, each coordinate X_{i_j} is completely determined by $X|_{Q_{i_j}}$, so the second term is 0. Therefore

$$H(X) = H\left(X|_{Q_{i_1}}, X|_{Q_{i_2}}, \ldots, X|_{Q_{i_t}}\right) \leq |Q_{i_1} \cup Q_{i_2} \cup \ldots \cup Q_{i_t}| \leq n - t$$

where the first inequality is due to by Fact 2.2, and the second inequality is because the set $Q_{i_1} \cup Q_{i_2} \cup \ldots \cup Q_{i_t}$ does not include $i_1, \ldots, i_t$. It follows that

$$H(X) \leq n - t \leq n - \frac{n}{q+1} < n - \frac{n}{\left(\frac{n}{k} - 1\right) + 1} = n - k,$$

and this contradicts the assumption that $H(X) \geq n - k$.

Now let us consider again the harder case of a non-deterministic adversary. In this setting, matters are more complicated: The set of queries Q_{i_1}that predicts $X|_{i_1}$ is not chosen a priori before seeing X, but is rather chosen from a family of witnesses F_{i_1} based on the value of X. Therefore, we cannot use the foregoing simple process to choose the coordinates $i_1, \ldots i_t$ and the sets $Q_{i_1}, \ldots, Q_{i_t}$. Instead, we choose the coordinates $i_1, \ldots, i_t$ at random. We then show that the entropy

$$H\left(X_{i_1}, X_{i_2}, \ldots, X_{i_t} \left| X|_{[n]-\{i_1,\ldots,i_t\}} \right.\right)$$

is small. From this point, an analysis along the same lines as above shows that $H(X) < n - k$, as required. We note that our actual proof is not a proof by contradiction, but rather uses the assumption $H(X) \geq n - k$ to derive a bound on the average probability that a coordinate is predicted by a certificate.

The reason that the latter entropy is small is that for every coordinate i_j, the string $X|_{[n]-\{i_1,\ldots,i_t\}}$ satisfies some family of witnesses for X_{i_j} with significant probability, and X_{i_j} is biased and has less than full entropy. Intuitively, the reason for $X|_{[n]-\{i_1,\ldots,i_t\}}$ satisfies a family of witnesses with significant probability is the following; Recall that satisfying a family of witnesses can be viewed as satisfying a small-width DNF formula. Conditioning on a random set of coordinates $[n]-\{i_1, \ldots, i_t\}$ is equivalent to subjecting the formula to a random restriction, which causes the formula to be fixed to 1 with high probability. We note that the latter implication does

not follow from the switching lemma (Håstad 1986), but from a simpler and more general observation on random restrictions (see Section 3.3).

3.2. The proof. Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, and let $q \in \mathbb{N}$, $0 \leq \varepsilon \leq 1$. For every coordinate $i \in [n]$, let (F_i^0, F_i^1) be a pair such that F_i^b is a q-family of witnesses for i that ε-predicts $X_i = b$, and let δ_i denote the probability that X satisfies either F_i^0 or F_i^1. We wish to prove that the average value of the δ_i's is at most $\frac{300 \cdot k \cdot q}{\varepsilon^3 \cdot n}$. To this end, for $b \in \{0,1\}$ let $\delta_{i,b}$ denote the probability that X satisfies F_i^b. We prove our claim for the $\delta_{i,0}$'s and the $\delta_{i,1}$'s separately: We will prove that the average of the $\delta_{i,0}$'s is at most $\frac{150 \cdot k \cdot q}{\varepsilon^3 \cdot n}$, and the same holds for the $\delta_{i,1}$'s, and the upper bound on the average of the δ_i's will follow by the union bound.

Specifically, we prove the upper bound on the average of the $\delta_{i,1}$'s, and the upper bound for the $\delta_{i,0}$'s can be proved similarly. Let $\overline{\delta}_1$ denote the average of the $\delta_{i,1}$'s. We prove the following result.

LEMMA 3.1. *Let $T \subseteq [n]$ be a uniformly distributed set of coordinates of size $t = \frac{\varepsilon \cdot n}{8 \cdot q}$. Then*

$$\mathbb{E}_T \left[H\left(X|_T \,\middle|\, X|_{[n]-T}\right) \right] \leq t - \frac{\varepsilon^2}{16} \cdot \overline{\delta}_1 \cdot t.$$

Observe that Lemma 3.1 implies the desired upper bound on $\overline{\delta}_1$: To see why, observe that assuming the latter inequality, it holds by the chain rule (Fact 2.4) that

$$\begin{aligned} H(X) &= \mathbb{E}_T \left[H(X) \right] \\ &= \mathbb{E}_T \left[H(X|_{[n]-T}) + H\left(X|_T \,\middle|\, X|_{[n]-T}\right) \right] \\ &\leq n - t + t - \frac{\varepsilon^2}{16} \cdot \overline{\delta^1} \cdot t \\ &= n - \frac{\varepsilon^2}{16} \cdot \overline{\delta}_1 \cdot t, \end{aligned}$$

By combining the latter inequality with the assumption that $H(X) \geq n - k$, we get

$$n - k \leq n - \frac{\varepsilon^2}{16} \cdot \overline{\delta}_1 \cdot t$$
$$\overline{\delta}_1 \leq \frac{k}{\frac{\varepsilon^2}{16} \cdot t}.$$

By substituting $t = \frac{\varepsilon \cdot n}{8 \cdot q}$, we get

$$\overline{\delta^1} \leq \frac{150 \cdot k \cdot q}{\varepsilon^3 \cdot n},$$

as required.

In the rest of this section, we prove Lemma 3.1. To this end, we will prove an upper bound on the entropy of a single coordinate in T, and then use the subadditivity of entropy to prove the upper bound on the entropy of $X|_T$. The following claim provides an upper bound on the entropy of a single coordinate.

Claim 3.2. *For every $i \in [n]$ it holds that*

$$\mathbb{E}_T \left[H\left(X_i \left| X|_{[n]-T}\right.\right) | i \in T \right] \leq 1 - \frac{\varepsilon^2}{16} \cdot \delta_{i,1}$$

Proof. Let $i \in [n]$, let E be the event that X satisfies F_i^1, and let E' be the event that $X|_{[n]-T}$ satisfies F_i^1 (formally, E' is the event that there is a witness $(Q, a) \in F_i^1$ that appears in X such that $Q \subseteq [n] - T$). The idea of the proof is that the probability of E' is close to $\delta_{i,1}$, and when this event occurs, the coordinate X_i is biased and therefore its entropy is low.

Observe that for every string x that satisfies F_i^1, it holds that

$$\Pr_{X,T} [E' | X = x, i \in T] \geq 1 - \frac{\varepsilon}{8}$$

To see it, let (Q, a) be the first witness in F_i^1 that appears in x. Then, each coordinate in Q has probability at most $\frac{t}{n}$ to belong to T, and by the union bound, the probability that any coordinate in Q belongs to T is at most $q \cdot \frac{t}{n} = \frac{\varepsilon}{8}$. Hence, it follows that $\Pr_{X,T} [E' | E, i \in T] \geq 1 - \frac{\varepsilon}{8}$. Since the events E and $i \in T$ are independent, it follows that the probability of E' is

$$\begin{aligned}\Pr_{X,T}\left[E'|i\in T\right] &\geq \left(1-\frac{\varepsilon}{8}\right)\cdot \Pr_{X,T}\left[E|i\in T\right] = \left(1-\frac{\varepsilon}{8}\right)\cdot \Pr_{X}\left[E\right]\\ &\geq \left(1-\frac{\varepsilon}{8}\right)\cdot \delta_{i,1}.\end{aligned}$$

We now show that if the event E' occurs, the coordinate X_i is biased. The reason that this holds is that the coordinate is biased conditioned on the event E, and the event E' has high probability conditioned E. Formally, it holds that

$$\begin{aligned}\Pr_{X,T}\left[X_i=0|E',i\in T\right] &\leq \frac{\Pr_{X,T}\left[X_i=0|E,i\in T\right]}{\Pr_{X,T}\left[E'|E,i\in T\right]}\\ &\leq \frac{\Pr_{X,T}\left[X_i=0|E,i\in T\right]}{1-\frac{\varepsilon}{8}}\\ &\leq \Pr_{X,T}\left[X_i=0|E,i\in T\right]+\frac{\varepsilon}{4}\\ &= \Pr_{X,T}\left[X_i=0|E\right]+\frac{\varepsilon}{4}\\ &\leq \frac{1}{2}-\frac{\varepsilon}{2}+\frac{\varepsilon}{4}\\ &\leq \frac{1}{2}-\frac{\varepsilon}{4},\end{aligned}$$

and therefore $\Pr_{X,T}\left[X_i=1|E',i\in T\right]\geq \frac{1}{2}+\frac{\varepsilon}{4}$. It follows that

$$\begin{aligned}&\mathbb{E}_T\left[H(X_i|E')|E',i\in T\right]\\ &= \mathbb{E}_T\left[H(\Pr_X\left[X_i=1|E'\right])\middle| E',i\in T\right]\\ \text{(binary entropy is concave)}\quad &= H\left(\mathbb{E}_T\left[\Pr_X\left[X_i=1|E'\right]|E',i\in T\right]\right)\\ &= H\Big(\Pr_{X,T}\left[X_i=1|E',i\in T\right]\Big)\\ &\leq H\Big(\frac{1}{2}+\frac{\varepsilon}{4}\Big)\\ &\leq 1-\frac{1}{8}\cdot\varepsilon^2.\end{aligned}$$

We finally turn to upper bound the expectation

$$\mathbb{E}_T\left[H\left(X_i\left|X\right|_{[n]-T}\right)|i\in T\right].$$

To this end, we use the fact that this expectation can be written as the conditional entropy $H\left(X_i \left| X|_{[n]-T}, T, i \in T\right.\right)$. Now, let $1_{E'}$ be the indicator random variable of E'. Since the value of $1_{E'}$ is determined by the random variables T and $X|_{[n]-T}$, it follows from Fact 2.3 that

$$H\left(X_i \left| X|_{[n]-T}, T, i \in T\right.\right) \leq H\left(X_i \left| 1_{E'}, i \in T\right.\right).$$

Therefore,

$$\begin{aligned}
H\left(X_i \left| X|_{[n]-T}, T, i \in T\right.\right) &\leq H\left(X_i \left| 1_{E'}, T, i \in T\right.\right) \\
&= H(X_i|E', T, i \in T) \cdot \Pr\left[E'|i \in T\right] \\
&\quad + H(X_i|\neg E', T, i \in T) \cdot \Pr\left[\neg E'|i \in T\right] \\
&\leq \left(1 - \frac{1}{8} \cdot \varepsilon^2\right) \cdot \Pr\left[E'|i \in T\right] \\
&\quad + 1 \cdot \left(1 - \Pr\left[E'|i \in T\right]\right) \\
&= 1 - \frac{1}{8} \cdot \varepsilon^2 \cdot \Pr\left[E'|i \in T\right] \\
&\leq 1 - \frac{1}{8} \cdot \varepsilon^2 \cdot \left(1 - \frac{1}{8} \cdot \varepsilon\right) \cdot \delta_i^1 \\
&\leq 1 - \frac{1}{16} \cdot \varepsilon^2 \cdot \delta_{i,1},
\end{aligned}$$

as required. □

Finally, we use the subadditivity of min-entropy to derive an upper bound on $\mathbb{E}_T\left[H\left(X|_T \left| X|_{[n]-T}\right.\right)\right]$. To this end, it will be convenient to view T as if it is chosen by choosing a sequence of uniformly distributed distinct coordinates $i_1, \ldots, i_t$. Then, we can write the latter expectation as

$$\mathbb{E}_T\left[H\left(X|_T \left| X|_{[n]-T}\right.\right)\right] = \mathbb{E}_{i_1,\ldots,i_t}\left[H\left(X_{i_1}, \ldots, X_{i_t} \left| X|_{[n]-\{i_1,\ldots,i_t\}}\right.\right)\right],$$

and therefore it suffices to upper bound the right-hand side. By the subadditivity of entropy, it holds that

$$\begin{aligned}
&\mathbb{E}_{i_1,\ldots,i_t}\left[H\left(X_{i_1}, \ldots, X_{i_t} \left| X|_{[n]-\{i_1,\ldots,i_t\}}\right.\right)\right] \\
&\leq \sum_{j=1}^{t} \mathbb{E}_{i_1,\ldots,i_t}\left[H\left(X_{i_j} \left| X|_{[n]-\{i_1,\ldots,i_t\}}\right.\right)\right].
\end{aligned} \tag{3.3}$$

For each $j \in [t]$, it holds that

$$\begin{aligned}
&\mathbb{E}_{i_1,\ldots,i_t}\left[H\left(X_{i_j}\left|X|_{[n]-\{i_1,\ldots,i_t\}}\right.\right)\right] \\
&= \frac{1}{n}\sum_{i_j=1}^{n}\mathbb{E}_{i_1,\ldots,i_{j-1},i_{j+1},\ldots i_t}\left[H\left(X_{i_j}\left|X|_{[n]-\{i_1,\ldots,i_t\}}\right.\right)\right] \\
&= \frac{1}{n}\sum_{i_j=1}^{n}\mathbb{E}_T\left[H(X_{i_j}|X_{[n]-T})|i_j \in T\right] \\
&= \frac{1}{n}\sum_{i_j=1}^{n}H(X_{i_j}|X_{[n]-T}, T, i_j \in T) \\
&\leq \frac{1}{n}\sum_{i_j=1}^{n}1 - \frac{\varepsilon^2}{16}\cdot\delta_{i_j,1} \\
&= 1 - \frac{\varepsilon^2}{16}\cdot\overline{\delta}_1.
\end{aligned}$$

Together with Inequality (3.3), the last inequality implies Lemma 3.1. This concludes the proof of Theorem 1.3.

3.3. Connection to random restrictions. As discussed above, one way to view the proof of Theorem 1.3 is to view the families (F_i^0, F_i^1) as DNF formulas, and to view the conditioning on $X|_{[n]-T}$ as applying a random restriction that simplifies these formulas. In particular, the following simple proposition about random restrictions is implicit in the proof of Claim 3.2. Since we believe this proposition is interesting in its own right, we make its proof explicit below.

PROPOSITION (1.7, restated). *Let ϕ be a DNF formula over n variables of width at most w, and let X be a random variable that is distributed arbitrarily in $\{0,1\}^n$ such that $\phi(X) = 1$ with probability δ. Let ρ be a random restriction that fixes each variable with probability at least p independently, and that chooses the values of the fixed variables according to the marginal distribution of X on those variables. Then, $\phi|_\rho$ is fixed to 1 with probability at least $p^w \cdot \delta$.*

PROOF. Let ϕ, X, ρ be as in the proposition. Observe that we can view ρ as if it is sampled as follows: first sample a string x from the distribution of X, and then for every $i \in [n]$ set $\rho(i) = x_i$ with probability p and set $\rho(i) = \star$ otherwise. Now, conditioned on any specific choice of x such that $\phi(x) = 1$, the probability that $\phi|_\rho$ is fixed to 1 is at least p^w, since this is a lower bound on the probability that ρ fixes the variables of the first term that is satisfied by x. By summing over all the strings x for which $\phi(x) = 1$, we get that the total probability that $\phi|_\rho$ is fixed to 1 is at least $p^w \cdot \delta$. □

3.4. Applications to decision trees and certificates. We now show how to derive the applications of Theorem 1.3 to decision trees and certificates.

3.4.1. Decision trees. We prove the application of the theorem to decision trees, restated next. Recall that we say that a decision tree ε-*predicts* X_i if the decision tree makes queries to the coordinates in $[n] - \{i\}$ and outputs the value of X_i correctly with probability at least $\frac{1}{2} + \frac{1}{2} \cdot \varepsilon$.

COROLLARY (1.5, restated). *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, and let $q \in \mathbb{N}$, $0 \leq \varepsilon \leq 1$. Then, the number of coordinates $i \in [n]$ that are ε-predicted by some decision tree that makes at most q queries is at most $\frac{300 \cdot k \cdot q}{\varepsilon^3}$.*

Let X be a random variable as in the corollary. In order to apply the theorem, we define for each coordinate a pair of families (F_i^0, F_i^1). For every coordinate $i \in [n]$ that is ε-predicted by a decision tree, and each $b \in \{0,1\}$, we construct the family of witnesses F_i^b that ε-predicts $X_i = b$ by taking the collection of all the paths in the tree that lead to a leaf that is labeled b. It can be seen that a string $x \in \{0,1\}^n$ satisfies F_i^b if and only if the tree outputs b on x. For every coordinate that is *not* predicted by a decision tree, we take F_i^0 and F_i^1 to be empty.

Now, for every $i \in [n]$, if the coordinate i is predicted by a decision tree, then the probability that it satisfies either F_i^0 or F_i^1 is 1, and otherwise the probability is 0. On the other hand,

Theorem 1.3 tells us that the average of those probabilities is at most $\frac{300 \cdot k \cdot q}{\varepsilon^3 \cdot n}$. It follows that the number of coordinates that are predicted by decision trees is at most $\frac{300 \cdot k \cdot q}{\varepsilon^3}$, as required.

3.4.2. Certificates. We prove the application of the theorem to certificates. Recall that a b-certificate for a coordinate $i \in [n]$ is a witness (Q, a) such that

$$\Pr\left[X_i = b | X|_Q = a\right] = 1.$$

Then, we have the following result.

Corollary (1.6, restated). *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, and let $q \in \mathbb{N}$, $0 \leq \varepsilon \leq 1$. For every coordinate $i \in [n]$, we denote by δ_i the probability that any certificate for X_i of length at most q appears in X. Then, the average value of δ_i over $i \in [n]$ is at most $\frac{300 \cdot k \cdot q}{n}$.*

Let X be a random variable as in the corollary. In order to apply the theorem, we define for each coordinate a pair of families (F_i^0, F_i^1). For every coordinate $i \in [n]$ and each $b \in \{0,1\}$, we define the F_i^b to be the family of all b-certificates for i of length at most q. It is easy to see that this family 1-predicts that $X_i = b$. Moreover, the probability δ_i that X satisfies F_i^b is exactly the probability that any certificate for i of length at most q appears in X. Then, Theorem 1.3 tells us that the average of those probabilities is at most $\frac{300 \cdot k \cdot q}{n}$, as required.

4. Depth-3 lower bounds

In this section, we use Corollary 1.6 to prove our result on depth-3 circuits, restated next. Recall that this result says that if a function has a noticeable fraction of sensitive inputs then it is hard for depth-3 circuits, thus extending Boppana's theorem (Theorem 1.8) for depth-3 circuits.

Theorem (1.9, restated). *There exists a constant $\gamma > 0$ such that the following holds. Let $f : \{0,1\}^n \to \{0,1\}$ be a function has sensitivity at least s on at least $\alpha \cdot 2^n$ inputs in $f^{-1}(0)$ for some*

$0 < \alpha < 1$ *(respectively,* $f^{-1}(1)$*). Then every depth-3 circuit that computes* f *whose top gate is an AND gate (respectively, OR gate) must be of size at least* $\frac{\alpha}{n} \cdot 2^{\gamma \cdot \sqrt{s}}$.

Let $\gamma > 0$ be a sufficiently small constant to be fixed later. Let $f : \{0,1\}^n \to \{0,1\}$ be a function that has sensitivity at least s on at least $\alpha \cdot 2^n$ inputs in $f^{-1}(0)$ for some $0 < \alpha < 1$. We prove every depth-3 circuit that computes f whose top gate is an AND gate must be of size at least $\frac{\alpha}{n} \cdot 2^{\gamma \cdot \sqrt{s}}$. By the Karchmer–Wigderson connection (Proposition 2.11), it suffices to prove a lower bound on the communication complexity of 3-round protocols that solve the Karchmer–Wigderson relation R_f. Specifically, fix a protocol for R_f that behaves as follows:

- Alice gets a string $x \in f^{-1}(0)$ and Bob gets a string $y \in f^{-1}(1)$.
- In the first round, Alice sends at most $\gamma \cdot \sqrt{s} - \log \frac{n}{\alpha}$ bits.
- In the second round, Bob sends at most $\gamma \cdot \sqrt{s} - \log \frac{n}{\alpha}$ bits.
- In the third round, Alice sends a coordinate $j \in [n]$ that is supposed to satisfy $x_j \neq y_j$.

We will prove that the protocol must err on some pair of inputs (x, y).

Proof sketch. We start by making some observations on how any such protocol must behave. First, observe that when the second round ends, Alice must know a coordinate $j \in [n]$ for which $x_j \neq y_j$, since she has to send it in the third round. For a given coordinate $j \in [n]$, Alice can be sure that $x_j \neq y_j$ only if she knows the value of y_j. Hence, the only valuable information that Bob can send in the second round is the values of bits of y. We can therefore assume[2] without loss of generality that in the second round, Bob chooses some set of coordinates $F \subseteq [n]$ of size at most $\gamma \cdot \sqrt{s}$ and sends to Alice the string $y|_F$. Moreover, since

[2] See Remark 4.1 for a formal justification of this intuition.

Bob has to be sure that Alice will be able to extract a correct coordinate $j \in [n]$ from $y|_F$, Bob can only choose a set $F \subseteq [n]$ for which he knows for sure that $x|_F \neq y|_F$.

Therefore, the proof of the lower bound boils down to showing that after Alice sent her first message, Bob cannot know that $x|_F \neq y|_F$ for any set of coordinates $F \subseteq [n]$ of size $\gamma \cdot \sqrt{s}$. To this end, we use Corollary 1.6. Suppose that Alice's input is a random string which is uniformly distributed over the set of inputs in $f^{-1}(0)$ with sensitivity s. This random string has entropy at least $n - \log \frac{1}{\alpha}$. Then after Alice sends her first message, Alice's input has entropy at least $n - \gamma \cdot \sqrt{s}$ conditioned on this message—let us denote this random string by X. By Corollary 1.6 and our choice of parameters, we can show there exists some coordinate $i \in [n]$ such that with constant probability, the coordinate i is sensitive and the string X does not satisfy any certificate of length $\gamma \cdot \sqrt{s}$ for X_i—let us denote this event by E_i.

Now, suppose that we sample an input for Bob from the following distribution: We first sample a random string X' from the distribution of X conditioned on E_i (but X' is not necessarily equal to Alice's input X). Then, we choose the input Y of Bob to be the string obtained by flipping the i-th coordinate of X'. Note that Y is indeed an input in $f^{-1}(1)$. We claim that for every $F \subseteq [n]$ of size at most $\gamma \cdot \sqrt{s}$, it holds that $X|_F = Y|_F$ with nonzero probability.

Let $F \subseteq [n]$ be such a set, and let $F' = F - \{i\}$. Then, due to the way we sampled Y, the marginal of Y on $[n] - \{i\}$ is identical to the marginal of X'. This implies that with nonzero probability it holds that $X'|_{F'} = Y|_{F'}$, and the same holds for $X|_{F'} = Y|_{F'}$. If $F' = F$, we are done. Otherwise, we note that because X' is conditioned on the event E_i, the string X' does not satisfy any certificate for i, and therefore $(F', Y|_{F'})$ cannot be a certificate for i (since $X'|_{F'} = Y|_{F'}$ with nonzero probability). This implies that X_i has nonzero probability to be either 0 or 1 conditioned on $X|_{F'} = Y|_{F'}$. Hence, it holds that $X|_F = Y|_F$ with nonzero probability. This concludes the proof.

Proof of Theorem 1.9. We prove that the protocol errs using an adversary argument. Let $A_0 \subseteq f^{-1}(0)$ be the set of inputs in $f^{-1}(0)$ at which f has sensitivity at least s, so $|A_0| \geq \alpha \cdot 2^n$. On each input

in A_0, Alice sends some message in the first round. Let π_A be the message that corresponds to the largest number of inputs in A_0, and let A_1 be the set of those inputs, so

$$|A_1| \geq 2^{-(\gamma \cdot \sqrt{s} - \log \frac{n}{\alpha})} \cdot |A_0| \geq n \cdot 2^{n - \gamma \cdot \sqrt{s}}.$$

Let X be a random variable that is uniformly distributed in A_1, so $H(X) \geq n - \gamma \cdot \sqrt{s}$.

For every $i \in [n]$, let δ_i be the probability that any certificate for X_i of length at most $2 \cdot \gamma \cdot \sqrt{s}$ appears in X. We now choose γ to be sufficiently small such that it would follow from Corollary 1.6 that the average value of the δ_i's is at most $\frac{s}{2n}$. Next, observe that the average probability that a coordinate $i \in [n]$ is sensitive (in the sense that flipping X_i would result in a string in $f^{-1}(1)$) is at least $\frac{s}{n}$ since $X \in A_0$. Therefore, there exists some coordinate $i \in [n]$ such that with probability at least $\frac{s}{2n}$, the coordinate i is sensitive and no certificate for X_i of length at most $2 \cdot \gamma \cdot \sqrt{s}$ appears in X. Let X' be a random variable that is distributed like X conditioned on the latter event, and let A_2 be the support of X', so $|A_2| \geq \frac{s}{2n} \cdot |A_1|$.

Let $B_0 \subseteq f^{-1}(1)$ be a set obtained from the set A_2 by flipping the i-th coordinate of every string in A_2. Since flipping the i-th coordinate is a bijection, it holds that

$$|B_0| = |A_2| \geq \frac{s}{2n} \cdot n \cdot 2^{n - \gamma \cdot \sqrt{s}} \geq 2^{n - \gamma \cdot \sqrt{s}},$$

where in the last inequality we assumed that $s \geq 2$ since otherwise the theorem holds trivially. On each input in B_0, Bob sends some message in the second round (given that Alice sent π_A in the first round). Let π_B be the message the corresponds to the largest number of inputs in B_0, and let B_1 be the set of those inputs, so $|B_1| \geq 2^{n - 2\gamma \cdot \sqrt{n}}$.

Now, let $F \subseteq [n]$ be the set of coordinates that are fixed in B_1, i.e., it is the set of coordinates j such that all strings in B_1 have the same value at j. It is not hard to see that $|F| \leq 2 \cdot \gamma \cdot \sqrt{s}$. Let $y_F \in \{0,1\}^F$ be the unique string in the projection of B_1 to the set F. We show that there exists a string $x \in A_1$ such that[3]

[3] Note that if F is empty, then this claim holds vacuously.

$x|_F = y_F$. Intuitively, this means that Bob cannot know for sure that Alice's input differs from his input on F.

Let $F' = F - \{i\}$, and let $y_{F'}$ be the projection of y_F to F'. Since $B_1 \subseteq B_0$, it holds that $y_{F'} \in B_0|_{F'}$. Furthermore, due to the way we constructed B_0, it holds that $B_0|_{F'} = A_2|_{F'}$ and thus $y_{F'} \in A_2|_{F'}$. If $F' = F$ (i.e., $i \notin F$), then the fact that $y_{F'} \in A_2|_{F'}$ implies that there exists $x \in A_2 \subseteq A_1$ such that $x|_F = y_F$ and we are done. Suppose otherwise, i.e., $i \in F$. Then, the fact that $y_{F'} \in A_2|_{F'} \subseteq A_1|_{F'}$ implies that

$$\Pr[X|_{F'} = y_{F'}] \geq \Pr[X'|_{F'} = y_{F'}] > 0.$$

Moreover, by the definition of X', no certificate for the coordinate i of at most length $2\gamma \cdot \sqrt{n}$ appears in the string X', and therefore $(F', y_{F'})$ is not a certificate for i (since $(F', y_{F'})$ appears in X' with nonzero probability by the last inequality). This implies that

$$\Pr[X_i = (y_F)_i \,|\, X|_{F'} = y|_{F'}] > 0.$$

It follows that

$$\Pr[X|_F = y_F] = \Pr[X_i = (y_F)_i \,|\, X|_{F'} = y_{F'}] \cdot \Pr[X|_{F'} = y_{F'}] > 0,$$

and therefore there exists a string $x \in A_1$ such that $x|_F = y_F$.

Finally, let j be the coordinate that Alice sends in the third round, provided that she gets the input x and that the messages π_A and π_B were sent in the first and second rounds, respectively. We consider two cases, based on whether $j \in F$ or not, and show that in both cases we can choose an input $y \in B_1$ for Bob such that $x_j = y_j$ (and hence the protocol errs):

- **The case where $j \in F$:** In this case, we know that $x_j = (y_F)_j$. Moreover, by the definition of y_F, every string $y \in B_1$ satisfies $y|_F = y_F$ and hence $x_j = y_j$. It follows that we can choose any string $y \in B_1$ to be the input of Bob.

- **The case where $j \notin F$:** Recall that the set F was defined to be the set of coordinates that are fixed in B_1. Therefore, the coordinate j is not fixed in B_1, so for any bit $b \in \{0, 1\}$ there is a string y in B_1 such that $y_j = b$. In particular, there is a string $y \in B_1$ such that $x_j = y_j$, and we can choose this string to be the input of Bob.

We showed that in both cases there exist a string $y \in B_1$ such that $x_j = y_j$. Now, observe that when Alice and Bob get as inputs the strings x and y, the transcripts of the protocol is indeed (π_A, π_B, j). In particular, the protocol errs on those inputs, which is what we wanted to show. □

REMARK 4.1. *In the beginning of the proof sketch above, we argued that we can assume without loss of generality that all Bob does it to choose a set of coordinates F and send $y|_F$ to Alice. We can now justify this intuition formally: Given any message of Bob, let us define the set $F \subseteq [n]$ as it was defined in our proof. Note that due to the way that F was defined, we may view Bob's message as conveying the content of $y|_F$ to Alice (perhaps along with other information). Next, note that our proof shows that Alice can only output a coordinate j that belongs to F (see "the case where $j \notin F$" above). Finally, observe that if Alice knows $y|_F$, and she has to output an index j that belongs to F, she does not need any additional information—the value $y|_F$ is all Alice needs to find a coordinate $j \in F$ where $x_j \neq y_j$ (if such a coordinate exists). Hence, any additional information in Bob's message can be safely ignored.*

Using this line of reasoning, it is not hard to prove the following claim: Any 3-round protocol can be transformed into one in which Bob chooses a set of coordinates $F \subseteq [n]$ and sends $F, y|_F$ to Alice, and Alice replies with an index $j \in F$.

5. Certificates for sets of coordinates

Recall the application of the main theorem to certificates.

COROLLARY (1.6, restated). *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, and let $q \in \mathbb{N}$. For every coordinate $i \in [n]$, we denote by δ_i the probability that any certificate for X_i of length at most q appears in X. Then, the average value of δ_i over $i \in [n]$ is at most $\frac{300 \cdot k \cdot q}{n}$.*

In this section, we extend our result on certificates to certificates for sets of coordinates. Recall that such certificates are defined as follows.

Definition (1.12, restated). *Let X be a random variable taking values from $\{0,1\}^n$, let $R \subseteq [n]$ be a set of coordinates. A* certificate for R *(*with respect to X*) is a pair (Q, a) where $Q \subseteq [n]-R$ and $a \in \{0,1\}^{|Q|}$, such that conditioned on $X|_Q = a$, the random variable $X|_R$ does not have full support. The* length *of the certificate is $|Q|$, and we say that a string $x \in \{0,1\}^n$* satisfies the certificate *if $x|_Q = a$.*

We prove the following result.

Theorem (1.13, restated). *Let X be a random variable taking values from $\{0,1\}^n$ such that $H(X) \geq n - k$, let $r, q \in \mathbb{N}$, and assume that $(q + r) \cdot (2k + r + 1) \leq \frac{1}{4000} \cdot n$. For every set of coordinates $R \subseteq [n]$ of size r, we denote by p_R the probability that a string drawn from X does not satisfy* any *certificate for R of length at most q. Then, the average value of p_R over $R \subseteq [n]$ is at least 2^{-r-1}.*

The basic idea of the proof is the following: Let R be a random set of size r. We lower bound the probability that X satisfies any certificate for R over the choice of both X and R, and this is equivalent to lower bounding the average value of p_R. Suppose that we choose the set $R \subseteq [n]$ by choosing a sequence of random distinct coordinates $i_1, \ldots, i_r$ We first observe that by our choice of parameters, with probability at least $\frac{1}{2}$, the coordinate i_1 is "good," in the sense that X does not satisfy any certificate for i_1 of length at most $q + r$. Moreover, with probability at least $\frac{1}{2}$ the coordinate i_2 is good even conditioned on i_1 being good. Continuing in this manner, we get that with probability at least 2^{-r}, every coordinate i_j is good even conditioned on all the previous coordinates being good. Finally, we observe that if the latter event occurs, then X does not satisfy any certificate for R; Otherwise, we could have used this certificate and the string that is missing from the support of $X|_R$ to construct a certificate for some coordinate i_j. Details follow.

Let X be a random string in $\{0,1\}^n$ such that $H(X) \geq n - k$. Since we are going to analyze X conditioned on several events, it would be easier to work with min-entropy instead of entropy. By Fact 2.10, there exists an event E of probability at least $\frac{1}{2}$ such that $H_\infty(X|E) \geq n - 2k - 1$. For ease of notation, let X' denote

the random variable X conditioned on the event E. In the rest of this proof, we will work with X' instead of X.

Let $i_1, \ldots, i_r \in [n]$ be uniformly distributed distinct coordinates, and let $R = \{i_1, \ldots, i_r\}$. We prove that the probability, over X' and R, that X' does not satisfy any certificate for R of length at most q is at least 2^{-r}. This will imply that the probability X does not satisfy any such certificate is at least 2^{-r-1}, and the required result will follow.

We define a sequence of events $E_1, \ldots, E_r$ as follows: The event E_j is the event that X' does not satisfy any certificate for i_j of length $q+r-j$ *with respect to the random variable* $X'|E_{j-1}$, conditioned on E_{j-1}. It is important to note that we refer to certificates that are with respect to $X'|E_{j-1}$ rather than X', that is, certificates that predict $(X'|E_{j-1})|_{i_j}$ from having full support. We will prove the following two claims, which say that $\Pr[E_r] \geq 2^{-r}$, and that conditioned on E_r, the string X' does not satisfy any certificate for R of length at most q. Together, these two claims imply the required result.

Claim 5.1. *For every $j \in [r]$ it holds that $\Pr[E_j] \geq 2^{-j}$.*

Proof. The proof is by induction on j. We prove the induction step, and the proof of the induction base is similar. Suppose the claim holds for $j \in [r-1]$. We prove the claim for $j+1$. By assumption, it holds that $\Pr[E_j] \geq 2^{-j}$. By Fact 2.9, this means that $H_\infty(X'|E_j) \geq n - 2k - 1 - j$. For every $i \in [n]$, let δ_i denote the probability that $X'|E_j$ satisfies any certificate for i of length $q+r-(j+1)$ with respect to $X'|E_j$. By Corollary 1.6, the average of the δ_i's is at most

$$\frac{300 \cdot (q+r-j-1) \cdot (2k+j+1)}{n} \leq \frac{300 \cdot (q+r) \cdot (2k+r)}{n} \leq \frac{300}{4000} = \frac{3}{40}.$$

Hence, by Markov's inequality it holds that $\delta_{i_{j+1}} \leq \frac{1}{4}$ with probability at least $\frac{2}{3}$ over the choice of i_{j+1}. Now, the probability of E_{j+1} conditioned on E_j is at least the probability that such i_{j+1} was chosen (which is at least $\frac{2}{3}$), times the probability that $X'|E_j$ does

not satisfy any certificate for i_{j+1} over the choice of $X'|E_j$ (which is at least $\frac{3}{4}$). Hence, this probability is at least $\frac{1}{2}$. It therefore follows that

$$\Pr[E_{j+1}] = \Pr[E_{j+1}|E_j] \cdot \Pr[E_j] \geq 2^{-(j+1)},$$

as required. □

Claim 5.2. *For every $j \in [r]$ and any specific choice of $i_1, \ldots, i_j$ the following holds: conditioned on the event E_j, then the string X' does not satisfy any certificate for $\{i_1, \ldots, i_j\}$ of length at most $q + r - j$ with respect to X'.*

Proof. We prove the induction step, and the proof of the induction base is similar. Suppose the claim holds for $j \in [r-1]$. We prove the claim for $j+1$. Fix an arbitrary choice of $i_1, \ldots, i_{j+1}$, and denote $R_j = \{i_1, \ldots, i_j\}$, $R_{j+1} = \{i_1, \ldots, i_{j+1}\}$. For ease of notation, we identify the event E_{j+1} with the set of strings x for which the event occurs, and the same for the event E_j (in other words, we identify E_{j+1} and E_j with the supports of $X'|E_{j+1}$ and $X'|E_j$). Let $x \in E_{j+1}$. We prove that x does not satisfy any certificate for R_{j+1} of length at most $q + r - j$ with respect to X'. Let (Q, a) be a witness of length at most $q+r-j$ that appears in x. We prove that the string $X'|_{R_{j+1}}$ has full support conditioned on $X'|_Q = a$. To this end, we prove that for every string $u \in \{0,1\}^{R_{j+1}}$ it holds that

$$\Pr\left[X'|_{R_{j+1}} = u \middle| X'|_Q = a\right] > 0. \tag{5.3}$$

By the assumption that $x \in E_{j+1}$, it follows that x does not satisfy any certificate for i_{j+1} of length at most $q + r - j - 1$ with respect to $X'|E_j$. In particular, this means that

$$\Pr\left[X'|_{i_{j+1}} = u_{i_{j+1}} \middle| X'|_Q = a, E_j\right] > 0. \tag{5.4}$$

It follows that there exists $y \in E_j$ such that $y|_Q = a$ and $y_{i_{j+1}} = u_{i_{j+1}}$. Furthermore, by the induction assumption, y does not satisfy any certificate for R_j of length at most $q+r-j$ with respect to X'. This implies in particular that the witness $(Q \cup \{i_{j+1}\}, a \cup u_{i_{j+1}})$

is not a certificate for R_j with respect to X', since it appears in y. Hence, the string $X'|_{R_j}$ conditioned on $X'|_Q = a$ and $X'_{i_{j+1}} = u_{i_{j+1}}$ has full support. It follows that

$$\Pr\left[X'|_{R_j} = u|_{R_j} \middle| X'|_Q = a, X'_{i_{j+1}} = u_{i_{j+1}}\right] > 0.$$

Combining the last inequality with Inequality (5.4), we get that

$$\begin{aligned}
&\Pr\left[X'|_{R_{j+1}} = u \middle| X'|_Q = a\right] \\
&\quad = \Pr\left[X'|_{R_j} = u|_{R_j} \middle| X'|_Q = a, X'_{i_{j+1}} = u_{i_{j+1}}\right] \\
&\qquad \cdot \Pr\left[X'|_{i_{j+1}} = u_{i_{j+1}} \middle| X'|_Q = a\right] \\
&\quad > 0,
\end{aligned}$$

as required. □

On extending our circuit lower bound to higher depths. As mentioned in the introduction, one motivation for the result proved in this section is that we believe it might be useful for extending our circuit lower bound (Theorem 1.9) to higher depths. We now explain how such an extension might be done, although we do not know how to fully realize this idea. Specifically, we explain how one might use Theorem 1.13 to prove a (suboptimal) lower bound of $2^{\Omega(n^{1/4})}$ on depth-4 circuits computing the parity function on n bits. In order to prove such a lower bound, we need to rule out the existence of a 4-round protocol that finds a coordinate $j \in [n]$ on which the inputs of the parties disagree with communication complexity $o(n^{1/4})$.

As in our proof in the depth-3 case, we consider a random variable X that is uniformly distributed over the inputs of Alice conditioned on her message in the first round. This random variable has min-entropy at least $n - \Omega(n^{1/4})$, and Theorem 1.13 tells us that there is some set of coordinates $R \subseteq [n]$ of size $\approx \sqrt{n}$ such that $X|_R$ does not have certificates of length $\approx \sqrt{n}$ with non-trivial probability. Now, we can make sure that the inputs of the parties have the same marginals over the coordinates in $[n] - R$, and therefore they cannot find the desired coordinate j in $[n] - R$ (just as in the depth-3 case they had the same marginals in $[n] - \{i\}$ and

thus could not find j there). Hence, they must find the solution j in the set R.

However, the random variable $X|_R$ has full support. Therefore, one would expect the task of finding the desired coordinate j in R to reduce to solving the Karchmer–Wigderson relation of parity on $|R|$ bits. Since the players have to perform the latter task using only three rounds, one would expect that they will have to transmit at least $\Omega(\sqrt{|R|}) = \Omega(n^{1/4})$ bits. Such an argument, if it could be carried out, would prove the desired lower bound of $\Omega(n^{1/4})$. The main challenge is that unlike the case of 3-round protocols, we can no longer assume that the only useful information that Bob can send is the values of coordinates F in his input.

Acknowledgements

We would like to thank Oded Goldreich and Benjamin Rossman for valuable discussions and ideas. We would also like to thank Roei Tell for pointing out an error in the introduction of an earlier version of this work. Finally, we thank anonymous referees for comments that improved the presentation of this work and for pointing out connections to the work of Paturi *et al.* (1999).

Or Meir is partially supported by the Israel Science Foundation (Grant No. 1445/16). Part of this research was done while Or Meir was partially supported by NSF Grant CCF-1412958. Avi Wigderson was partially supported from NSF Grant CCF-1412958.

Publisher's Note

References

Miklós Ajtai (1983). Σ^1_1-Formulae on finite structures. *Annals of Pure and Applied Logic* **24**(1), 1–48.

Miklós Ajtai (1992). *Boolean Complexity and Probabilistic Constructions*, 140–164. London Mathematical Society Lecture Note Series. Cambridge University Press.

Miklós Ajtai (1993). Geometric Properties of Sets Defined by Constant Depth Circuits. In *Combinatorics, Paul Erdős is eighty*. Budapest, Hungary : János Bolyai Mathematical Society. ISBN 9638022736 (set).

Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar & D. Sivakumar (2004). An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.* **68**(4), 702–732.

Paul Beame (1994). A switching lemma primer. Technical report, Technical Report UW-CSE-95-07-01, Department of Computer Science and Engineering, University of Washington.

Ravi B. Boppana (1997). The Average Sensitivity of Bounded-Depth Circuits. *Inf. Process. Lett.* **63**(5), 257–261.

Xi Chen, Igor Carboni Oliveira, Rocco A. Servedio & Li-Yang Tan (2016). Near-optimal small-depth lower bounds for small distance connectivity. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18–21, 2016*, 612–625.

Thomas M. Cover & Joy A. Thomas (1991). *Elements of information theory*. Wiley-Interscience. ISBN 0-471-06259-6.

Irit Dinur & Or Meir (2016). Toward the KRW Composition Conjecture: Cubic Formula Lower Bounds via Communication Complexity. In *31st Conference on Computational Complexity, CCC 2016, May 29 to June 1, 2016, Tokyo, Japan*, 3:1–3:51.

Pavol Duris, Zvi Galil & Georg Schnitger (1987). Lower Bounds on Communication Complexity. *Inf. Comput.* **73**(1), 1–22.

Jeff Edmonds, Russell Impagliazzo, Steven Rudich & Jiri Sgall (2001). Communication complexity towards lower bounds on circuit depth. *Computational Complexity* **10**(3), 210–246.

Merrick L. Furst, James B. Saxe & Michael Sipser (1984). Parity, Circuits, and the Polynomial-Time Hierarchy. *Mathematical Systems Theory* **17**(1), 13–27.

Anat Ganor, Gillat Kol & Ran Raz (2014). Exponential Separation of Information and Communication. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18–21, 2014*, 176–185.

Dmitry Gavinsky, Or Meir, Omri Weinstein & Avi Wigderson (2014). Toward better formula lower bounds: an information complexity approach to the KRW composition conjecture. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, 213–222.

Michelangelo Grigni & Michael Sipser (1991). Monotone Separation of Logspace from NC. In *Structure in Complexity Theory Conference*, 294–298.

Aryeh Grinberg, Ronen Shaltiel & Emanuele Viola (2018). Indistinguishability by adaptive procedures with advice, and lower bounds on hardness amplification proofs. *Electronic Colloquium on Computational Complexity (ECCC)* **25**, 61.

Johan Håstad (1986). Almost Optimal Lower Bounds for Small Depth Circuits. In *STOC*, 6–20.

Johan Håstad, Stasys Jukna & Pavel Pudlák (1995). Top-Down Lower Bounds for Depth-Three Circuits. *Computational Complexity* **5**(2), 99–112.

Hossein Jowhari, Mert Săglam & Gábor Tardos (2011). Tight bounds for Lp samplers, finding duplicates in streams, and related problems. In *Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2011, June 12–16, 2011, Athens, Greece*, 49–58.

Mauricio Karchmer, Eyal Kushilevitz & Noam Nisan (1995a). Fractional Covers and Communication Complexity. *SIAM J. Discrete Math.* **8**(1), 76–92.

Mauricio Karchmer, Ran Raz & Avi Wigderson (1995b). Super-Logarithmic Depth Lower Bounds Via the Direct Sum in Communication Complexity. *Computational Complexity* **5**(3/4), 191–204.

Mauricio Karchmer & Avi Wigderson (1990). Monotone Circuits for Connectivity Require Super-Logarithmic Depth. *SIAM J. Discrete Math.* **3**(2), 255–265.

V. M. Khrapchenko (1972). A method of obtaining lower bounds for the complexity of π-schemes. *Mathematical Notes Academy of Sciences USSR* **10**, 474–479.

Maria M. Klawe, Wolfgang J. Paul, Nicholas Pippenger & Mihalis Yannakakis (1984). On Monotone Formulae with Restricted Depth (Preliminary Version). In *Proceedings of the 16th Annual ACM Symposium on Theory of Computing, April 30 - May 2, 1984, Washington, DC, USA*, 480–487.

Gillat Kol & Ran Raz (2013). Interactive channel capacity. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1–4, 2013*, 715–724.

Nathan Linial, Yishay Mansour & Noam Nisan (1993). Constant Depth Circuits, Fourier Transform, and Learnability. *J. ACM* **40**(3), 607–620.

Lyle A. Mcgeoch (1986). A Strong Separation betweem k and $k-1$ Round Communication Complexity for a Constructive Language. Technical Report CMU-CS-86-157, Carnegie Mellon University.

Or Meir (2017). An Efficient Randomized Protocol for every Karchmer-Wigderson Relation with Three Rounds. *Electronic Colloquium on Computational Complexity (ECCC)* **24**, 129.

Noam Nisan & Avi Wigderson (1993). Rounds in Communication Complexity Revisited. *SIAM J. Comput.* **22**(1), 211–219.

Noam Nisan & David Zuckerman (1996). Randomness is Linear in Space. *J. Comput. Syst. Sci.* **52**(1), 43–52.

Christos H. Papadimitriou & Michael Sipser (1984). Communication Complexity. *J. Comput. Syst. Sci.* **28**(2), 260–269.

Ramamohan Paturi, Pavel Pudlák & Francis Zane (1999). Satisfiability Coding Lemma. *Chicago J. Theor. Comput. Sci.* **1999**.

Toniann Pitassi, Benjamin Rossman, Rocco A. Servedio & Li-Yang Tan (2016). Poly-logarithmic Frege depth lower bounds via an expander switching lemma. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18–21, 2016*, 644–657.

Ran Raz (1998). A Parallel Repetition Theorem. *SIAM J. Comput.* **27**(3), 763–803.

Ran Raz & Avi Wigderson (1989). Probabilistic Communication Complexity of Boolean Relations (Extended Abstract). In *FOCS*, 562–567.

Ran Raz & Avi Wigderson (1992). Monotone Circuits for Matching Require Linear Depth. *J. ACM* **39**(3), 736–744.

A. A. Razborov (1992a). On Submodular Complexity Measures. In *Poceedings of the London Mathematical Society Symposium on Boolean Function Complexity*, 76–83. Cambridge University Press, New York, NY, USA. ISBN 0-521-40826-1.

Alexander A. Razborov (1990). Applications of matrix methods to the theory of lower bounds in computational complexity. *Combinatorica* **10**(1), 81–93.

Alexander A. Razborov (1992b). On the Distributional Complexity of Disjointness. *Theor. Comput. Sci.* **106**(2), 385–390.

Alexander V. Smal & Navid Talebanfard (2018). Prediction from partial information and hindsight, an alternative proof. *Inf. Process. Lett.* **136**, 102–104.

Emanuele Viola (2018). AC0 unpredictability. *Electronic Colloquium on Computational Complexity (ECCC)* 209.

Manuscript received June 19, 2018

Or Meir
Department of Computer Science
University of Haifa
3498838 Haifa, Israel
ormeir@cs.haifa.ac.il
http://cs.haifa.ac.il/~ormeir

Avi Wigderson
Institute for Advanced Study
Princeton, NJ, USA
avi@ias.edu
https://www.math.ias.edu/avi/