Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 91 (2017)

Small bias requires large formulas

Andrej Bogdanov*

Abstract

A small-biased function is a randomized function whose distribution of truth-tables is small-biased. We demonstrate that known explicit lower bounds on (1) the size of general Boolean formulas, (2) the size of De Morgan formulas, and (3) correlation against small De Morgan formulas apply to small-biased functions. As a consequence, any strongly explicit small-biased generator is subject to the best-known explicit formula lower bounds in all these models.

On the other hand, we give a construction of a small-biased function that is tight with respect to lower bound (1) for the relevant range of parameters. We interpret this construction as a natural-type barrier against substantially stronger lower bounds for general formulas.

1 Introduction

Formula size is one of the most thoroughly studied complexity measures of Boolean functions. A formula is a circuit in which every internal gate has fan-out one. The power of formulas depends on the types of gates allowed. In this work we consider two models: *General formulas* in which any gate of some pre-specified fan-in c is allowed, and *De Morgan formulas* that only use NOT gates and AND/OR gates of fan-in two.

Explicit size lower bounds for general formulas were first proved by Nečiporuk [Nec66], who showed that the selector (addressing) function requires general constant fan-in formula size $\Omega(n^2/\log n \log\log n)$ over inputs of size n. Boppana and Sipser [BS90] applied a variant of this method to obtain an improved lower bound of $\Omega(n^2/\log n)$ for the element distinctness function by a related but different method.

The case of De Morgan formulas had been studied even earlier. Subbotovskaya [Sub61] proved that computing parity on n bits requires formula size $\Omega(n^{3/2})$. Andreev [And87] combined the ideas of Nečiporuk and Subbotovskaya to obtain a $n^{5/2-o(1)}$ De Morgan formula size lower bound for an explicit family of functions from $\{0,1\}^n$ to $\{0,1\}$. Following partial improvements (Impagliazzo and Nisan [IN93], Paterson and Zwick [PZ93]), Håstad [Hås98]

*Department of Computer Science and Engineering and Institute for Theoretical Science and Communications, Chinese University of Hong Kong. Work supported by HK RGC GRF grant no. CUHK14208215.

ISSN 1433-8092

showed that Andreev's function requires formula size $n^{3-o(1)}$, which is optimal in the exponent.[1] The same lower bound was reproved by Dinur and Meir [DM16] using different methods.

More recently, Tal gave two lower-order improvements to Håstad's result. First, in [Tal14] he showed that Andreev's function requires De Morgan formulas of size $\Omega(n^3/(\log n)^2 \log\log n)$, which is optimal for this function up to the doubly logarithmic term. Later, in [Tal16] he showed that another function introduced by Komargodski and Raz [KR13] requires De Morgan formula size $\Omega(n^3/(\log n)(\log\log n)^2)$.

In a related line of works, Komargodski, Raz, and Tal [KR13, KRT13, Tal14] study correlation lower bounds against small formulas. For every $k \leq n^{1/3}$, they construct two variants of an explicit function that has correlation at most 2^{-k} with any De Morgan formula of size $n^3/(\log n)^{O(1)}k^2$. Their hard functions make use of error-correcting codes with good list-decodable properties and extractors for bit-fixing sources. Weaker correlation bounds for the parity function were proved by Santhanam [San10] and, as observed in [KRT13], also follow implicitly from bounds on the approximate degree of De Morgan formulas [Rei11, BBC+01].

Razborov and Rudich [RR94] observed that all formula size lower bounds (known at the time) are natural, meaning that the formulas to which the bounds apply cannot compute cryptographically pseudorandom functions. On the other hand, the class NC^1 of polynomial-size logarithmic-depth bounded fan-in circuit families, which are equivalent in power to polynomial-size formula families, is believed to contain pseudorandom functions. Naor and Reingold [NR99, NRR02] and Banerjee, Peikert, and Rosen [BPR12] proposed such candidate families based on the Decisional Diffie Hellman, hardness of factoring, and Learning With Errors hardness assumptions, respectively. These constructions suggest that explicit size n^C lower bounds for formulas is out of reach for current techniques for sufficiently large values of the exponent C. The values of C in these constructions (for the requisite levels of hardness) are apparently rather large, so they are unlikely to explain the perceived barriers of n^2 and n^3 for general and De Morgan formula size, respectively.

Our results Our main conceptual contribution is the realization that all known formula size lower bound techniques also apply to small-biased functions. A randomized function is (K, ε)-biased if the induced distribution over truth-tables is a (K, ε)-biased distribution (it satisfies (1) below).

From the perspective of natural proofs, the known properties that distinguish small formulas from random functions are local in the sense that they only make a bounded number of non-adaptive queries to the function.[2] It is therefore reasonable to expect that the largeness condition of the relevant natural properties should continue to hold for random functions that only exhibit bounded independence. We show that these properties, in fact, merely require small bias [NN93], which is closely related to *approximate* bounded independence. As a direct consequence, we show that the best-known explicit formula lower bounds hold against any implicitly specified small-biased generator (the precise definition is given below).

[1] Our discussion of formula lower bounds is based on Chapter 6 of Jukna's book [Juk10].

[2] The correlation lower bounds [KR13, KRT13] in fact apply adaptive queries of a restricted type.

Theorem 1. *Any small-biased generator* $SB_{n,2^{-15n}} \colon \{0,1\}^{O(n)} \to \{0,1\}$

1. *requires fan-in c formulas of size* $\Omega(n^2/2^c \log n)$,
2. *requires De Morgan formulas of size* $\Omega(n^3/\log n\ (\log\log n)^2)$,
3. *has correlation at most* $2^{-\Omega(k)}$ *with De Morgan formulas of size at most* $n^3/(\log n)^{O(1)}k^2$ *for any* k *such that* $\omega(\log n) \le k \le n$.

Together with the existence of strongly explicit small-biased generators (see definition and discussion below), Theorem 1 reproves the best-known formula lower bounds in a unified manner and even gives a minor improvement in one case. Item 1 matches the explicit formula size lower bound of Nečiporuk. Item 2 matches the lower bound of Tal [Tal16]. Item 3 is a minor improvement over the lower bound of Tal [Tal14]. His proof requires the additional assumption $k \le n^{1/3}$.

Like previous formula size lower bounds with the exception of [DM16], the proof of Theorem 1 relies on shrinkage. It is therefore not surprising that it merely matches but fails to improve the state of the art in explicit lower bounds. The value of Theorem 1 is in explaining hardness against formulas by a single natural property, namely small bias. In contrast, shrinkage proofs are tailored to the model in question. The proofs of [Nec66, Hås98, Tal16] rely on shrinkage of a random restriction, the one of [BS90] on simultaneous shrinkage of multiple restrictions, while [KR13, KRT13, Tal14] use high probability shrinkage. While the role of small bias in the shrinkage arguments is more or less self-evident in certain proofs (Propositions 1 and 3), it is less obvious properties of small bias (Lemmas 1 and 2) that enables the others (Propositions 2, 4, and 5).

From a wider perspective, the utility of circuit lower bounds extends far beyond separating complexity classes, which is merely a motivating purpose. It is just as important to identify which natural (both in the common and technical sense) properties of functions make them intractable in specific computational models. In this sense Theorem 1 provides a new criterion for pseudorandomness of a cryptographic function against a restricted class of distinguishers.

On the other hand, in Theorem 2 we construct a (K,ε)-biased function F with fan-in two formula size $O(n(\log K)^2(\log 1/\varepsilon))$. For $\varepsilon = 2^{-2K}$, this is a $(K, 2^{-K})$-wise independent function of formula size $O(nK(\log K)^2)$, which matches our lower bounds for general formulas in Propositions 1 and 2 up to terms polylogarithmic in K.

In the parameter regimes that yield lower bounds 1 and 2 in Theorem 1, the function F has formula size $O((n\log n)^2)$ and De Morgan formula size $O(n^4(\log n)^3)$. We view this as a barrier to proving super-quadratic lower bounds for general formulas, and super-quartic ones for De Morgan formulas.

In the notation of Razborov and Rudich, our barriers are $\oplus$-natural, where $\oplus$ is the class of parity functions. However, they are not quasipolynomial-size-natural since our function F is not cryptographically pseudorandom: In addition to having small formula size, F is computable by polynomial-size, depth 3 circuit families with AND, OR, and PARITY

gates (the class $\mathrm{AC}^0[\oplus]$), which is known not to contain cryptographic pseudorandom functions [Raz87, Smo87, RR94]. It remains open whether our bounds can be matched (or even improved in the case of De Morgan formulas) by a different construction that is plausibly secure with respect to all subexponential-size circuits, of which linear tests are a very special case.

Theorems 1 and 2 suggest that small-biased functions should be studied as suitable candidates for formula size lower bounds. In the extreme setting of parameters $K = 2^n$, $\varepsilon = 2^{-\Theta(n)}$, known constructions of small-biased functions have seed lengths linear in n and may be plausible candidates for improved formula size lower bounds. In this regime, the general and De Morgan formula sizes of F in Theorem 2 are as large as $\tilde{\Theta}(n^4)$. Do there exist, say, $(2^n, 2^{-100n})$-biased functions of smaller formula size?

Bounded independence and small bias We will call a randomized function $F\colon \{0,1\}^n \to \{0,1\}$ (k,ε)-*wise independent* (in qualitative terms, *almost locally independent*) if for any k distinct inputs $x_1, \ldots, x_k$, the distribution $(F(x_1), \ldots, F(x_k))$ is within statistical distance ε of the uniform distribution over $\{0,1\}^k$. A random function $F\colon \{0,1\}^n \to \{0,1\}$ is (K,ε)-*biased* (*locally small-biased*) if for any nonempty set X of at most K distinct inputs,

$$\left|\mathbb{E}\left[\prod\nolimits_{x\in X}(-1)^{F(x)}\right]\right| \leq \varepsilon. \tag{1}$$

When $K = 2^n$ the family is called ε-*biased* (*small-biased*). Small bias implies bounded independence by the following claim [NN93, Corollary 2.1].

Claim 1. *Every (K,ε)-biased function is $(K, 2^{K/2}\varepsilon)$-wise independent.*

Small-biased generators A family of functions $SB_{n,\varepsilon}\colon \{0,1\}^{s(n,\varepsilon)} \times \{0,1\}^n \to \{0,1\}$ is a *small-biased generator* if the random function $F_r(x) = SB_{n,\varepsilon}(r,x)$ (with r uniformly random) is ε-biased for all n and ε. If we view $SB_{n,\varepsilon}$ as a function from $\{0,1\}^{s(n,\varepsilon)}$ to the set of truth-tables of functions $\{0,1\}^n \to \{0,1\}$, we recover the usual representation of a pseudorandom generator as a function of its seed.

The generator is *strongly explicit* if $s(n,\varepsilon) = O(n + \log 1/\varepsilon)$ and $SB_{n,\varepsilon}$ is uniformly polynomial-time computable. Known constructions of small-biased sets [NN93, AGHP92, ABN+92, BT13, TS17] are strongly explicit.

2 Small bias requires large formulas

We are aware of two techniques for proving general formula size lower bounds, the one of Nečiporuk [Nec66] and the variant due to Boppana and Sipser [BS90]. We show that both imply lower bounds on the formula size of almost locally independent functions. While the second technique yields a stronger lower bound, we find the first one instructive as the role of almost-independence is more transparent.

In the case of De Morgan formulas, we study three proof techniques. The first one, based on average-case shrinkage, underlies the lower bound of Andreev including improvements by

Impagliazzo and Nisan, Paterson and Zwick, Håstad, and Tal. We show that this method also bounds the formula size of almost independent functions.

The second method for De Morgan formula lower bounds is due to Tal, who applies a correlation-to-computation reduction in addition to bounds on average-case shrinkage. The third method, due to Komargodski and Raz and improvements by these authors and Tal, applies a high-probability shrinkage lemma to derive strong correlation lower bounds. We show that these two methods give lower bounds on the size of small-biased functions.

Arbitrary formulas

A *restriction* $f|_\rho$ of a function f under a partial assignment ρ of its inputs is the function on the unassigned inputs obtained by fixing all the assigned variables to their values. A random $\overline{k}$-restriction of f is the distribution of restrictions of f under a uniform random assignment that leaves exactly k inputs unassigned.

The *size* of a formula is the number of leaves in the formula tree, namely the number of variables occurring in the formula. The following shrinkage property of formulas follows immediately from linearity of expectation:

Claim 2. *Assume $f\colon \{0,1\}^n \to \{0,1\}$ has formula size s. Then the expected formula size of a random $\overline{k}$-restriction of f is at most $(k/n) \cdot s$.*

We say a random function F has formula size at most s if every function in the support of F can be computed by a formula of size at most s.

Proposition 1. *Assuming $c \leq \log\log k$, any $(2^k, 1/4)$-wise independent function $F\colon \{0,1\}^n \to \{0,1\}$ requires fan-in c formulas of size $\Omega(n \cdot 2^k/k \log k)$.*

Proof. Suppose F has formula size s. By Claim 2 and averaging, there exists a partial assignment ρ with k unassigned variables under which the expected formula size of $F|_\rho$ is at most ks/n. By Markov's inequality,

$$\Pr_F[\text{size}(F|_\rho) \leq 2ks/n] \geq \frac{1}{2} \tag{2}$$

for any distribution of functions F, where size denotes formula size.

A formula of size $\tilde{s}$ can be specified by listing its at most $2\tilde{s}$ gates in depth-first order. For a formula of fan-in c on k inputs, there are 2^{2^c} possible internal gates and k possible input gates, so the number of such formulas is at most $(2^{2^c} + k)^{2\tilde{s}} \leq (2k)^{2\tilde{s}}$. Therefore, setting $\tilde{s} = 2ks/n$, for a uniformly random function R it holds that

$$\Pr_R[\text{size}(R|_\rho) \leq 2ks/n] \leq \frac{(2k)^{4ks/n}}{2^{2^k}}. \tag{3}$$

The event "$\text{size}(F|_\rho) \leq 2ks/n$" depends on at most 2^k values of F, so if F is $(2^k, 1/4)$-wise independent, then

$$\Pr_F[\text{size}(F|_\rho) \leq 2ks/n] \leq \Pr_R[\text{size}(R|_\rho) \leq 2ks/n] + \frac{1}{4}. \tag{4}$$

Combining (2), (3), and (4), we obtain that $(2k)^{4ks/n}/2^{2^k} \geq 1/4$, from where the desired lower bound on s follows. □

An improved lower bound We now discuss the other proof of Nečiporuk, which gives a slightly stronger lower bound in the regime of $k < \log n$ and for exponentially small error.

Proposition 2. *For $k \leq \log n - 1$, any $(2 \cdot 2^k, 2^{-2^k})$-wise independent function $F\colon \{0,1\}^n \to \{0,1\}$ requires fan-in c formulas of size $\Omega(n \cdot 2^{k-c}/k)$.*

The proposition is proved by showing that the number of possible restrictions of a small formula that leave the least frequently occurring inputs unrestricted is small. On the other hand, the following lemma shows that the number of distinct restrictions of an almost locally independent function is large, even when the set of unrestricted variables is fixed. A $\overline{U}$-restriction is a restriction under any assignment in which U is the set of free variables.

Lemma 1. *Assume F is $(2 \cdot 2^k, 2^{-2^k})$-wise independent. For any set U of k variables, the number of distinct $\overline{U}$-restrictions of F is at least $\min\{2^{n-k-2}, 2^{2^k-3}\}$ with probability more than half.*

In particular, when $k \leq \log n - 1$, a $(2 \cdot 2^k, 2^{-2^k})$-wise independent function family has at least $\frac{1}{8} \cdot 2^{2^k}$ distinct $\overline{U}$-restrictions with probability more than half.

Proof. Let ρ, ρ' be independent random partial assignments to the variables in $\overline{U}$. Then

$$\Pr_{F,\rho,\rho'}[F|_\rho = F|_{\rho'}] \leq \Pr[\rho = \rho'] + \Pr[F|_\rho = F|_{\rho'} \mid \rho \neq \rho']. \tag{5}$$

The first term equals 2^{-n+k}. To bound the second term, fix an arbitrary pair of distinct ρ, ρ'. The event that the restricted functions $F|_\rho$ and $F|_{\rho'}$ are identical depends on at most $2 \cdot 2^k$ values of F. By the almost local independence of F,

$$\Pr_F[F|_\rho = F|_{\rho'} \mid \rho \neq \rho'] \leq \Pr[R = R'] + 2^{-2^k},$$

where $R, R'\colon \{0,1\}^k \to \{0,1\}$ are independent uniformly random functions. Such functions are equal with probability at most 2^{-2^k}, and so the second term in (5) at most 2^{-2^k+1}. Therefore

$$\Pr_{F,\rho,\rho'}[F|_\rho = F|_{\rho'}] \leq 2^{-n+k} + 2^{-2^k+1}.$$

Now assume the support size of $F|_\rho$ over random ρ is less than S for at least half the functions F. Then the collision probability $\Pr_{\rho,\rho'}[F|_\rho = F|_{\rho'}]$ is at least $1/S$ for at least half the functions F and so

$$2^{-n+k} + 2^{-2^k+1} \geq \frac{1}{2S},$$

from where it follows that the larger of 2^{-n+k} and 2^{-2^k+1} is at least $1/4S$. It follows that $S \geq \min\{2^{n-k-2}, 2^{2^k-3}\}$. □

Proof of Proposition 2. Let s be the size of F. By Claim 2 and averaging, there is a set U of size k so that on average, F has at most $(k/n) \cdot s$ occurrences of variables from U. By Markov's inequality, at least half of the formulas in F have no more than $\tilde{s} = 2ks/n$ occurrences of variables from U.

We now upper bound the number of $\overline{U}$-restrictions of ϕ (for fixed ϕ and U). Under each partial assignment ρ to this inputs in $\overline{U}$, ϕ reduces to a formula $\phi|_\rho$ of size at most $\tilde{s}$. This formula can be simplified by propagating the restricted inputs and subsuming all gates of fan-in one into their parents or children in some canonical way. The simplified formula can then be described by specifying, say in depth first order, the truth-tables of its gates (of fan-in at least two). As there are at most $\tilde{s}$ such gates and each can compute one of at most 2^{2^c} possible functions, the desired number of restrictions can be at most $2^{2^c \tilde{s}}$.

By Lemma 1, there must then exist a formula in the support of F whose number of $\overline{U}$-restrictions is at most $2^{2^c \tilde{s}} = 2^{2^{c+1} ks/n}$ and at least $\frac{1}{8} \cdot 2^{2^k}$. It follows that $s = \Omega(n \cdot 2^{k-c}/k)$. □

Computation by De Morgan formulas

In this section we show that known proofs for De Morgan formula size also apply to small-biased functions. The following proof relies on expected shrinkage of De Morgan formulas under random restrictions [And87, IN93, PZ93, Hås98].

Proposition 3. *Assuming $k \leq n/2$, any $(2^k, 1/4)$-wise independent function $F\colon \{0,1\}^n \to \{0,1\}$ requires De Morgan formula size $\Omega(n^2 \cdot 2^k/k^2 \log k)$.*

Proof. In a $\overline{p}$-random restriction, the unrestricted variables are sampled from the binomial distribution with parameter p. Tal [Tal14] showed that if f has a De Morgan formula of size s then the expected formula size of a $\overline{p}$-random restriction of f is $\tilde{s} = O(p^2 s + \sqrt{p^2 s})$. Set $p = 2k/n$. By deviation bounds, for every f in the support of F, the event that ρ has fewer than $k = \frac{1}{2}pn$ unassigned inputs or $f|_\rho$ has formula size more than $4\tilde{s}$ has probability at most $\frac{1}{2}$.

By averaging, there exists a partial assignment ρ with k unassigned inputs under which $F|_\rho$ has formula size at most $4\tilde{s}$ for at least half the functions F. Since F is $(2^k, 1/4)$-wise independent, the same is true for at least a quarter of truly random functions R. The number of size $4\tilde{s}$ De Morgan formulas on k inputs is at most $(2k)^{8\tilde{s}}$, and these must compute at least $\frac{1}{4} \cdot 2^{2^k}$ distinct functions. It follows that $\tilde{s} = \Omega(2^k/\log k)$. As $\tilde{s} = O(p^2 s + \sqrt{p^2 s})$ it follows that $p^2 s = \Omega(2^k/\log k)$. Using the constraint $k \geq \frac{1}{2}pn$ we obtain the desired bound. □

Tal [Tal14] recently obtained a slight improvement to the aforementioned bounds. His method also applies to small-biased functions as demonstrated in the following proposition:

Proposition 4. *Assume that $k \leq n/2$ and $2^{-\frac{7}{16}k2^{k/8}} < \varepsilon \leq 2^{-2k}$. Then every $(2^k, \varepsilon)$-biased F requires De Morgan formula size $\Omega(n^2 \log(1/\varepsilon)/k(\log k)^2)$.*

The proof relies on the large deviation bound for small-bias distributions of Naor and Naor [NN93, Section 5]. We rework it here in more convenient notation. We say a random variable X over $\{-1,1\}^K$ is ε-biased if $|\mathbb{E}[\prod_{i \in S} X_i]| \leq \varepsilon$ for every subset S of indices.

Lemma 2. *Let t be even and X be a (t,ε)-biased random variable over $\{-1,1\}^K$. The probability that $|\sum X_i|$ exceeds δK is at most $\delta^{-t}\cdot(2(t/K)^{t/2}+\varepsilon)$.*

Proof. We apply a t-th moment calculation. By Markov's inequality,

$$\begin{aligned}\Pr\Big[\Big|\sum\nolimits_{i=1}^{K} X_i\Big| \geq \delta K\Big] &\leq \frac{1}{(\delta K)^t}\,\mathbb{E}\Big[\Big(\sum\nolimits_{i=1}^{K} X_i\Big)^t\Big] \\ &= \frac{1}{(\delta K)^t}\Big(\sum\nolimits_{S\in\mathcal{E}}\mathbb{E}\Big[\prod\nolimits_{i\in S} X_i\Big] + \sum\nolimits_{S\in\overline{\mathcal{E}}}\mathbb{E}\Big[\prod\nolimits_{i\in S} X_i\Big]\Big),\end{aligned}$$

where $\mathcal{E}$ is the set of ordered terms of size t in which every index appears an even number of times. The first expectation is upper bounded by the number of such terms, which is at most $K^{t/2}\cdot t!/(t/2)! \leq 2\cdot(tK)^{t/2}$. The second expectation is upper bounded by the number of terms times the maximum bias of each term, namely $K^t\cdot\varepsilon$. The desired bound follows. □

The following consequence of the lemma is far from tight but will be of use in the proof of Proposition 4. The *correlation* of two functions $f,\phi\colon\{0,1\}^k\to\{0,1\}$ is $\langle f,\phi\rangle = \mathbb{E}_x[(-1)^{f(x)}\cdot(-1)^{g(x)}]$, where x is uniform in $\{0,1\}^n$.

Corollary 1. *Assuming $2^{-\frac{7}{16}k2^{k/8}} < \varepsilon \leq 4\cdot 2^{-2k}$ and $F\colon\{0,1\}^k\to\{0,1\}$ is $(2^k,\varepsilon)$-biased, for every $\phi\colon\{0,1\}^k\to\{0,1\}$, the probability that $|\langle F,\phi\rangle|$ is greater than $2^{-k/4}$ is at most $3\cdot\varepsilon^{1/4}$.*

Proof. Assuming $4\leq t\leq 2^{k/8}$, the expression $(t\cdot 2^{-k})^{t/2}$ is non-increasing as a function of t. By our assumption on ε, there must exist even value $4\leq t<2^{k/8}$ for which

$$\big((t+2)2^{-k}\big)^{(t+2)/2} < \varepsilon \leq (t2^{-k})^{t/2}. \tag{6}$$

Applying Lemma 2 with parameters $K=2^k$, $\delta=2^{-k/4}$ to the truth-table X of the (K,ε)-biased function $(-1)^{F(x)\oplus\phi(x)}$, we obtain that the desired probability is at most

$$(2^{-k/4})^t\cdot\big(2(t2^{-k})^{t/2}+\varepsilon\big) = 3\cdot t^{t/2}\cdot 2^{-kt/4}.$$

To derive the corollary, it remains to show that $t^{t/2}\cdot 2^{-kt/4}\leq((t+2)2^{-k})^{(t+2)/8}$. This follows from $k\geq(4t\log t)/(t-2)$, which is true in the regime $4\leq t<2^{k/8}$. □

Proof of Proposition 4. Initially we proceed as in the proof of Proposition 3 to obtain a partial assignment ρ with k unassigned inputs under which $F|_\rho$ has formula size $\tilde{s} = O((k/n)^2 s + \sqrt{(k/n)^2 s})$ for at least half the functions F. Let $\mathcal{S}$ (for shrinkage) denote this event so that $\Pr[\mathcal{S}]\geq\frac{1}{2}$.

Tal [Tal16] showed that every formula of size $\tilde{s}$ has correlation at least $\delta=2^{-k/4}$ (7) with some formula of size $\tilde{s}' = O(\sqrt{\tilde{s}}+\tilde{s}\log k/k)$. Let Φ be the set of all such formulas over inputs in U. Then we have

$$\mathbb{E}[|\langle F|_\rho,\Phi\rangle|] \geq \mathbb{E}[|\langle F|_\rho,\Phi\rangle| \mid \mathcal{S}]\cdot\Pr[\mathcal{S}] \geq \frac{\delta}{2},$$

where $|\langle f, \Phi\rangle|$ denotes the maximum value of $|\langle f, \phi\rangle|$ over all $\phi \in \Phi$. By Markov's inequality,

$$\Pr[|\langle F|_\rho, \Phi\rangle| \geq \delta/4] \geq \frac{\delta}{4}.$$

On the other hand, by a union bound and Corollary 1,

$$\Pr[|\langle F|_\rho, \Phi\rangle| \geq \delta/4] \leq 3|\Phi|\varepsilon^{1/4}.$$

From these two inequalities we obtain that

$$|\Phi| \geq \frac{1}{12} \cdot \varepsilon^{-1/4} \cdot \delta \geq \frac{\varepsilon^{-1/8}}{12}$$

by (7) and the assumption $\varepsilon \leq 2^{-2k}$. Since $|\Phi| \leq (9k)^{\tilde{s}'}$, it follows that $\tilde{s}' = \Omega(\log(1/\varepsilon)/\log k)$. A calculation shows that $s = \Omega(n^2 \log(1/\varepsilon)/k(\log k)^2)$ as desired. □

Correlation with De Morgan formulas

Komargodski, Raz, and Tal [KR13, KRT13, Tal14] proved a correlation lower bound for small De Morgan formulas. Their main technical ingredient is the following high-probability shrinkage lemma for De Morgan formulas [KRT13, Tal14].

Lemma 3 and the proof of Proposition 5 use the following notation. Given a depth-$(n-k)$ decision tree Δ in n variables and a seed $\pi \in \{0,1\}^{n-k}$, the partial assignment $\Delta(\pi)$ is the one obtained by assigning values to the variables in Δ from root to leaf according to the sequence π (the first bit π_1 is assigned to the root variable, the second bit π_2 is assigned to its child, and so on).

Lemma 3 (High-probablity shrinkage)**.** *For every constant $c > 0$ there exists a constant $c' > 0$ such that for every $c' \log n \leq k \leq n$ the following holds. For every formula f on n variables of size $s \leq n^c$ there exists a decision tree Δ over its variables of depth at most $n-k$ so that $f|_{\Delta(\pi)}$ has formula size $\tilde{s} = (\log n)^{O(1)} \cdot (k/n)^2 \cdot s$ except with probability $\delta = 2^{-\Omega(k)}$ over the choice of π.*

Without loss of generality we will assume that Δ is a complete decision tree of depth exactly $n-k$ so that $\Delta(\pi)$ has exactly k unrestricted variables for every π.

Proposition 5. *Assuming $2^{-\frac{7}{16}k2^{k/8}} \leq \varepsilon \leq 3^{-8n} \cdot 2^{-2k}$ and $\omega(\log n) \leq k \leq n$, for every $(2^k, \varepsilon)$-biased F, at most a $2^{-\Omega(k)}$-fraction of F has correlation more than $2^{-\Omega(k)}$ with formulas of size at most $n^2 \log(1/\varepsilon)/(\log n)^{O(1)}k^2$.*

Proof of Proposition 5. Let $\mathcal{C}$ (for correlating) be the event that F has correlation at least 2δ in absolute value with some formula $\hat{F}$ of size at most s (which may depend on F) so that

$$\mathbb{E}_F[|\langle F, \hat{F}\rangle| \mid \mathcal{C}] \geq 2\delta.$$

We set $\delta = 2^{-\Omega(k)}$ and assume that $\delta \geq 2^{-k/8}$ (8). For every complete decision tree Δ (which may depend on $\hat{F}$) of depth $n-k$,

$$\mathbb{E}_{F,\pi}[|\langle F|_{\Delta(\pi)}, \hat{F}|_{\Delta(\pi)}\rangle| \mid \mathcal{C}] \geq \mathbb{E}_F\big[|\mathbb{E}_\rho[\langle F|_{\Delta(\pi)}, \hat{F}|_{\Delta(\pi)}\rangle]| \;\big|\; \mathcal{C}\big] = \mathbb{E}_F[|\langle F, \hat{F}\rangle| \mid \mathcal{C}] \geq 2\delta, \qquad (9)$$

where $\pi \sim \{0,1\}^{n-k}$ is a random seed. Let $\mathcal{S}$ be the event that $\hat{F}|_{\Delta(\pi)}$ has formula size at most $\tilde{s} = (\log n)^C \cdot (k/n)^2 \cdot s$ (10). By Lemma 3,

$$\Pr_{F,\pi}[\overline{\mathcal{S}} \mid \mathcal{C}] \leq \delta. \qquad (11)$$

By the formula for conditional expectations,

$$\begin{aligned}\mathbb{E}[|\langle F|_{\Delta(\pi)}, \hat{F}|_{\Delta(\pi)}\rangle| \mid \mathcal{C}] &= \mathbb{E}[|\langle F|_{\Delta(\pi)}, \hat{F}|_{\Delta(\pi)}\rangle| \mid \mathcal{C}\mathcal{S}] \cdot \Pr[\mathcal{S} \mid \mathcal{C}] \\ &\quad + \mathbb{E}[|\langle F|_{\Delta(\pi)}, \hat{F}|_{\Delta(\pi)}\rangle| \mid \mathcal{C}\overline{\mathcal{S}}] \cdot \Pr[\overline{\mathcal{S}} \mid \mathcal{C}] \\ &\leq \mathbb{E}[|\langle F|_{\Delta(\pi)}, \hat{F}|_{\Delta(\pi)}\rangle| \mid \mathcal{C}\mathcal{S}] + \Pr[\overline{\mathcal{S}} \mid \mathcal{C}],\end{aligned}$$

so (9) and (11) imply that

$$\mathbb{E}_{F,\pi}[|\langle F|_{\Delta(\pi)}, \hat{F}|_{\Delta(\pi)}\rangle| \mid \mathcal{C}\mathcal{S}] \geq \delta.$$

Let Φ be the set of all size-$\tilde{s}$ formulas over k variables. Then $|\Phi| \leq (9k)^{\tilde{s}}$ (12). Since conditioned on $\mathcal{S}$ all formulas $\hat{F}|_{\Delta(\pi)}$ are in Φ, it must be the case that

$$\mathbb{E}_{F,\pi}[|\langle F|_{\Delta(\pi)}, \Phi\rangle| \mid \mathcal{C}\mathcal{S}] \geq \delta,$$

where $\langle f, \Phi\rangle$ denotes the maximum of $\langle f, \phi\rangle$ over all $\phi \in \Phi$. By the formula for conditional expectations, $\mathbb{E}_{F,\pi}[|\langle F|_{\Delta(\pi)}, \Phi\rangle|]$ must be at least $\delta \cdot \Pr[\mathcal{C}\mathcal{S}]$. We can then bound $\Pr[\mathcal{C}\mathcal{S}]$ by

$$\Pr[\mathcal{C}\mathcal{S}] \leq \frac{1}{\delta} \cdot \mathbb{E}_{F,\pi}[|\langle F|_{\Delta(\pi)}, \Phi\rangle|] \leq \frac{1}{\delta}\Big(\frac{\delta^2}{4} + \Pr_{F,\pi}[|\langle F|_{\Delta(\pi)}, \Phi\rangle| \geq \delta^2/4]\Big). \qquad (13)$$

Let $\mathcal{P}_k^n$ denote the set of partial assignments that leave k inputs unassigned. As each input can take value 0, take value 1, or be unassigned, $\mathcal{P}_k^n$ has size at most 3^n. By Corollary 1, (8), and a union bound,

$$\begin{aligned}\Pr_{F,\pi}[|\langle F|_{\Delta(\pi)}, \Phi\rangle| \geq \delta^2/4] &\leq \Pr_F[|\langle F|_\rho, \phi\rangle| \geq \delta^2/4 \text{ for some } \rho \in \mathcal{P}_k^n \text{ and } \phi \in \Phi] \\ &\leq 3^n \cdot |\Phi| \cdot 3\varepsilon^{1/4}.\end{aligned}$$

Using (8) and the assumption $\varepsilon \leq 3^{-8n} \cdot 2^{-2k}$, the right hand side is at most $(\delta^2/4) \cdot 12|\Phi|\varepsilon^{1/8}$. By (12) and (10), this quantity is at most $\delta^2/4$ as long as $s \leq n^2 \log(1/\varepsilon)/(\log n)^C k^2$. Plugging into (13), we conclude that $\Pr[\mathcal{C}\mathcal{S}]$ is at most $\delta/2$ for formulas of the desired size.

Finally, applying (11) again, we have

$$\Pr[\mathcal{C}] = \frac{\Pr[\mathcal{C}\mathcal{S}]}{1 - \Pr[\overline{\mathcal{S}} \mid \mathcal{C}]} \leq \frac{\delta/2}{1-\delta} \leq \delta. \qquad \square$$

3 Main results

Proof of Theorem 1

Let F be the random function $F(x) = SB_{n,2^{-15n}}(s,x)$ for uniformly random s. To obtain item 1, we apply Proposition 2 with $k = \log n - 1$ and Claim 1. (Proposition 1 gives the weaker bound $\Omega(n^2/\log n \log\log n)$ for fan-in up to $c = \log\log\log n$.)

For item 2, we apply Proposition 4 with $k = 3\log n$ and $\varepsilon = n^9 e^{-n}$. (Proposition 3 with $k = \log n$ gives the weaker bound $\Omega(n^3/(\log n)^2 \log\log n)$.)

For item 3, we apply Proposition 5 with $\varepsilon = 3^{-8n} \cdot 2^{-2n}$. The conclusion is that at most a $2^{-\Omega(k)}$-fraction of F can have correlation more than $2^{-\Omega(k)}$ with formulas of size s. Therefore the correlation between $SB_{n,2^{-15n}}$ and size s formulas can be at most $2^{-\Omega(k)}$. □

Moderate formulas for small bias

Theorem 2. *For every n, k, and ε, there exists a $(2^k,\varepsilon)$-biased $F\colon \{0,1\}^n \to \{0,1\}$ of fan-in two formula size $O(nk^2 \cdot \log 1/\varepsilon)$.*

Applying Claim 1 and a suitable change of parameters we obtain the following corollary to Theorem 2:

Corollary 2. *For every n, K, and ε there exist (K,ε)-wise independent functions with formula size $O(n \cdot (\log K)^2 \cdot (K + \log 1/\varepsilon))$.*

Proof of Theorem 2. Let $H_t\colon \{0,1\}^n \to \{0,1\}$ be the random function

$$H_t(x) = \begin{cases} \text{a random bit}, & \text{if } Ax = b, \\ 0, & \text{if not,} \end{cases}$$

where A and b are a uniformly random $t \times n$ matrix and t-dimensional boolean vector, respectively, and all algebra is over $\mathbb{F}_2$. We let

$$F = F_1 \oplus F_2 \oplus \cdots \oplus F_{k+2},$$

where the F_t are independent XORs of $6\log 1/\varepsilon$ independent copies of H_t. Since H_t has formula size $O(tn)$, F has formula size $O(nk^2 \cdot \log 1/\varepsilon)$.

We now prove that F is $(2^k,\varepsilon)$-biased. Let X be any nonempty set of at most 2^k distinct inputs. Set $t = \lfloor \log|X| \rfloor + 2$ and let $\mathcal{U}$ (for unique) be the event that exactly one x in X satisfies $Ax = b$ for a random $t \times n$ matrix A and t-dimensional vector b. By the isolation lemma of Valiant and Vazirani [VV86], $\mathcal{U}$ has probability at least $1/8$ (see for example [AB09, Lemma 17.19]). By the rule of conditional expectations,

$$\begin{aligned} \left|\mathbb{E}\Big[\prod\nolimits_{x\in X}(-1)^{H_t(x)}\Big]\right| &\le \left|\mathbb{E}\Big[\prod\nolimits_{x\in X}(-1)^{H_t(x)}\Big|\mathcal{U}\Big]\right| \cdot \Pr[\mathcal{U}] + \left|\mathbb{E}\Big[\prod\nolimits_{x\in X}(-1)^{H_t(x)}\Big|\overline{\mathcal{U}}\Big]\right| \cdot \Pr[\overline{\mathcal{U}}] \\ &\le |\mathbb{E}[(-1)^{H_t(u)} \mid \mathcal{U}]| \cdot \Pr[\mathcal{U}] + 1 \cdot \Pr[\overline{\mathcal{U}}] \\ &= 0 \cdot \Pr[\mathcal{U}] + 1 \cdot \Pr[\overline{\mathcal{U}}] \\ &\le 7/8. \end{aligned}$$

By independence, it follows that

$$\left|\mathbb{E}\left[\prod_{x\in X}(-1)^{F_t(x)}\right]\right| = \left|\mathbb{E}\left[\prod_{x\in X}(-1)^{H_t(x)}\right]\right|^{6\log 1/\varepsilon} \leq \left(\frac{7}{8}\right)^{6\log 1/\varepsilon} \leq \varepsilon,$$

so $|\mathbb{E}[\prod_{x\in X}(-1)^{F(x)}]| = \prod_{t=1}^{k+2}|\mathbb{E}[\prod_{x\in X}(-1)^{F_t(x)}]|$ is also upper bounded by ε. □

Our small-biased function can be viewed as a simplified variant of a construction of Naor and Naor [NN93, Section 3.1.1]. The simplifications can be partly explained by a difference in objectives: Naor and Naor (and other constructions) aim to optimize the seed length, while we are interested in minimizing formula size.

By the standard simulation of fan-in two formulas by De Morgan formulas, F has De Morgan formula size at most $O((nk^2 \log 1/\varepsilon)^2)$. The De Morgan formula size analysis can be slightly improved to $O(n^2k^3(\log 1/\varepsilon)^2)$ by observing that the middle layer of AND gates does not suffer from the quadratic blow-up.

Specifically, in the parameter settings used in the proof of items 1 and 2 in Theorem 1, the function F has fan-in two formula size $O((n \log n)^2)$ and De Morgan formula size $O(n^4(\log n)^3)$.

For item 3, plugging in $\varepsilon = 2^{-k}$ gives the (De Morgan) formula size upper bound $O(n^4k^3)$. This can be improved to $O(n^4(\log(n/k))^3)$: In the proof of Corollary 1 (and Proposition 5) it is sufficient that F be (t, ε)-biased where t is the unique even integer satisfying (6). For our choice of parameters t is on the order of n/k.

Acknowledgments

Part of this work was done at the Simons Institute for the Theory of Computing at UC Berkeley. I thank Yuval Ishai for raising the question of natural proofs against formula size, Arbel Peled and Alon Rosen for insightful discussions, and anonymous referees for pointing out inaccuracies in previous versions and other useful comments.

References

[AB09] Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[ABN+92] Noga Alon, Jehoshua Bruck, Joseph Naor, Moni Naor, and Ron M. Roth. Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs. *IEEE Trans. Information Theory*, 38(2):509–516, 1992.

[AGHP92] Noga Alon, Oded Goldreich, Johan Hastad, and René Peralta. Simple construction of almost k-wise independent random variables. *Random Struct. Algorithms*, 3(3):289–304, 1992.

[And87] A. E. Andreev. On a method for obtaining more than quadratic effective lower bounds for the complexity of π-schemes. *Moscow Univ. Math. Bull.*, 42(1):63–66, 1987.

[BBC+01] Robert Beals, Harry Buhrman, Richard Cleve, Michele Mosca, and Ronald de Wolf. Quantum lower bounds by polynomials. *J. ACM*, 48(4):778–797, July 2001.

[BPR12] Abhishek Banerjee, Chris Peikert, and Alon Rosen. Pseudorandom functions and lattices. In *EUROCRYPT*, pages 719–737, 2012.

[BS90] Ravi B. Boppana and Michael Sipser. The complexity of finite functions. In *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity (A)*, pages 757–804. 1990.

[BT13] Avraham Ben-Aroya and Amnon Ta-Shma. Constructing small-bias sets from algebraic-geometric codes. *Theory of Computing*, 9:253–272, 2013.

[DM16] Irit Dinur and Or Meir. 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*, pages 3:1–3:51, 2016.

[Hås98] Johan Håstad. The shrinkage exponent of de Morgan formulas is 2. *SIAM J. Comput.*, 27(1):48–64, 1998.

[IN93] Russell Impagliazzo and Noam Nisan. The effect of random restrictions on formula size. *Random Struct. Algorithms*, 4(2):121–134, 1993.

[Juk10] Stasys Jukna. *Extremal Combinatorics: With Applications in Computer Science*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[KR13] Ilan Komargodski and Ran Raz. Average-case lower bounds for formula size. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 171–180, 2013.

[KRT13] Ilan Komargodski, Ran Raz, and Avishay Tal. Improved average-case lower bounds for de Morgan formula size. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 588–597, 2013.

[Nec66] E. I. Nechiporuk. On a Boolean function. *Soviet Math. Dokl.*, 7(4):999–1000, 1966.

[NN93] Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM J. Comput*, 22:838–856, 1993.

[NR99] Moni Naor and Omer Reingold. On the construction of pseudorandom permutations: Luby-Rackoff revisited. *J. Cryptology*, 12(1):29–66, 1999.

[NRR02] Moni Naor, Omer Reingold, and Alon Rosen. Pseudorandom functions and factoring. *SIAM J. Comput.*, 31(5):1383–1404, 2002.

[PZ93] Mike Paterson and Uri Zwick. Shrinkage of de Morgan formulae under restriction. *Random Struct. Algorithms*, 4(2):135–150, 1993.

[Raz87] Alexander A. Razborov. Lower bounds on the size of bounded depth circuits over a complete basis with logical addition. *Matematicheskie Zametki*, 41(4):598–607, 1987.

[Rei11] Ben W. Reichardt. Reflections for quantum query algorithms. In *Proceedings of the Twenty-second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, pages 560–569, Philadelphia, PA, USA, 2011. Society for Industrial and Applied Mathematics.

[RR94] Alexander A. Razborov and Steven Rudich. Natural proofs. In *STOC*, pages 204–213, 1994.

[San10] Rahul Santhanam. Fighting perebor: New and improved algorithms for formula and QBF satisfiability. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 183–192, 2010.

[Smo87] Roman Smolensky. Algebraic methods in the theory of lower bounds for boolean circuit complexity. In *STOC*, pages 77–82, 1987.

[Sub61] B. A. Subbotovskaya. Realizations of linear functions by formulas using $+$, $\cdot$, $-$. *Soviet Math. Dokl.*, 2:110–112, 1961.

[Tal14] Avishay Tal. Shrinkage of de Morgan formulae by spectral techniques. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 551–560, 2014.

[Tal16] Avishay Tal. Computing requires larger formulas than approximating. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:179, 2016.

[TS17] Amnon Ta-Shma. Explicit, almost optimal, epsilon-balanced codes. *Electronic Colloquium on Computational Complexity (ECCC)*, 24:41, 2017.

[VV86] Leslie G. Valiant and Vijay V. Vazirani. NP is as easy as detecting unique solutions. *Theor. Comput. Sci.*, 47(3):85–93, 1986.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il