Electronic Colloquium on Computational Complexity, Report No. 65 (2015)

An average-case depth hierarchy theorem for Boolean circuits

Benjamin Rossman
NII, Simons Institute
rossman@nii.ac.jp

Rocco A. Servedio[*]
Columbia University
rocco@cs.columbia.edu

Li-Yang Tan[†]
Simons Institute
liyang@cs.columbia.edu

April 17, 2015

Abstract

We prove an average-case depth hierarchy theorem for Boolean circuits over the standard basis of AND, OR, and NOT gates. Our hierarchy theorem says that for every $d \geq 2$, there is an explicit n-variable Boolean function f, computed by a linear-size depth-d formula, which is such that any depth-$(d-1)$ circuit that agrees with f on $(1/2 + o_n(1))$ fraction of all inputs must have size $\exp(n^{\Omega(1/d)})$. This answers an open question posed by Håstad in his Ph.D. thesis [Hås86b].

Our average-case depth hierarchy theorem implies that the polynomial hierarchy is infinite relative to a random oracle with probability 1, confirming a conjecture of Håstad [Hås86a], Cai [Cai86], and Babai [Bab87]. We also use our result to show that there is no "approximate converse" to the results of Linial, Mansour, Nisan [LMN93] and Boppana [Bop97] on the total influence of small-depth circuits, thus answering a question posed by O'Donnell [O'D07], Kalai [Kal12], and Hatami [Hat14].

A key ingredient in our proof is a notion of *random projections* which generalize random restrictions.

[*] Supported by NSF grants CCF-1319788 and CCF-1420349.

[†] Part of this research was done while visiting Columbia University.

ISSN 1433-8092

Contents

1 Introduction

The study of small-depth Boolean circuits is one of the great success stories of complexity theory. The exponential lower bounds against constant-depth $\mathsf{AND\text{-}OR\text{-}NOT}$ circuits [Yao85, Hås86a, Raz87, Smo87] remain among our strongest unconditional lower bounds against concrete models of computation, and the techniques developed to prove these results have led to significant advances in computational learning theory [LMN93, Man95], pseudorandomness [Nis91, Baz09, Raz09, Bra10], proof complexity [PBI93, Ajt94, KPW95], structural complexity [Yao85, Hås86a, Cai86], and even algorithm design [Wil14a, Wil14b, AWY15].

In addition to *worst-case* lower bounds against small-depth circuits, *average-case* lower bounds, or *correlation bounds*, have also received significant attention. As one recent example, Impagliazzo, Matthews, Paturi [IMP12] and Håstad [Hås14] independently obtained optimal bounds on the correlation of the parity function with small-depth circuits, capping off a long line of work on the problem [Ajt83, Yao85, Hås86a, Cai86, Bab87, BIS12]. These results establish strong limits on the computational power of constant-depth circuits, showing that their agreement with the parity function can only be an exponentially small fraction better than that of a constant function.

In this paper we will be concerned with average-case complexity *within* the class of small-depth circuits: our goal is to understand the computational power of depth-d circuits relative to those of strictly smaller depth. Our main result is an *average-case depth hierarchy theorem* for small-depth circuits:

Theorem 1. *Let $2 \le d \le \frac{c\sqrt{\log n}}{\log\log n}$, where $c > 0$ is an absolute constant, and Sipser_d be the explicit n-variable read-once monotone depth-d formula described in Section 6. Then any circuit C of depth at most $d-1$ and size at most $S = 2^{n^{\frac{1}{6(d-1)}}}$ over $\{0,1\}^n$ agrees with Sipser_d on at most $(\frac{1}{2} + n^{-\Omega(1/d)}) \cdot 2^n$ inputs.*

(We actually prove two incomparable lower bounds, each of which implies Theorem 1 as a special case. Roughly speaking, the first of these says that Sipser_d cannot be approximated by size-S, depth-d circuits which have significantly smaller bottom fan-in than Sipser_d, and the second of these says that Sipser_d cannot be approximated by size-S, depth-d circuits with a different top-level output gate than Sipser_d.)

Theorem 1 is an average-case extension of the worst-case depth hierarchy theorems of Sipser, Yao, and Håstad [Sip83, Yao85, Hås86a], and answers an open problem of Håstad [Hås86a] (which also appears in [Hås86b, Hås89]). We discuss the background and context for Theorem 1 in Section 1.1, and state our two main lower bounds more precisely in Section 1.2.

Applications. We give two applications of our main result, one in structural complexity and the other in the analysis of Boolean functions. First, via a classical connection between small-depth computation and the polynomial hierarchy [FSS81, Sip83], Theorem 1 implies that the polynomial hierarchy is infinite relative to a random oracle:

Theorem 2. *With probability 1, a random oracle A satisfies $\Sigma_d^{\mathrm{P},A} \subsetneq \Sigma_{d+1}^{\mathrm{P},A}$ for all $d \in \mathbb{N}$.*

This resolves a well-known conjecture in structural complexity, which first appeared in [Hås86a, Cai86, Bab87] and has subsequently been discussed in a wide range of surveys [Joh86, Hem94, ST95, HRZ95, VW97, Aar], textbooks [DK00, HO02], and research papers [Hås86b, Hås89, Tar89, For99, Aar10a]. (Indeed, the results of [Hås86a, Cai86, Bab87], along with much of the pioneering work on lower bounds against small-depth circuits in the 1980's, were largely motivated by the aforementioned connection to the polynomial hierarchy.) See Section 2 for details.

Our second application is a strong negative answer to questions of Kalai, Hatami, and O'Donnell in the analysis of Boolean functions. Seeking an *approximate converse* to the fundamental results of Linial, Mansour, Nisan [LMN93] and Boppana [Bop97] on the total influence of small-depth circuits, Kalai asked whether every Boolean function with total influence polylog(n) can be approximated by a constant-depth circuit of quasipolynomial size [Kal10, Kal12, Hat14]. O'Donnell posed a variant of the same question with a more specific quantitative bound on how the size of the approximating circuit depends on its influence and depth [O'D07]. As a consequence of Theorem 1 we obtain the following:

Theorem 3. *There are functions $d(n) = \omega_n(1)$ and $S(n) = \exp((\log n)^{\omega_n(1)})$ such that there is a monotone $f : \{0,1\}^n \to \{0,1\}$ with total influence $\mathbf{Inf}(f) = O(\log n)$, but any circuit C that has depth $d(n)$ and agrees with f on at least $(\frac{1}{2} + o_n(1)) \cdot 2^n$ inputs in $\{0,1\}^n$ must have size greater than $S(n)$.*

Theorem 3 significantly strengthens O'Donnell and Wimmer's counterexample [OW07] to a conjecture of Benjamini, Kalai, and Schramm [BKS99], and shows that the total influence bound of [LMN93, Bop97] does not admit even a very weak approximate converse. See Section 3 for details.

1.1 Previous work

In this subsection we discuss previous work related to our average-case depth hierarchy theorem. We discuss the background and context for our applications, Theorems 2 and 3, in Sections 2 and 3 respectively.

Sipser was the first to prove a worst-case depth hierarchy theorem for small-depth circuits [Sip83]. He showed that for every $d \in \mathbb{N}$, there exists a Boolean function $F_d : \{0,1\}^n \to \{0,1\}$ such that F_d is computed by a linear-size depth-d circuit, but any depth-$(d-1)$ circuit computing F_d has size $\Omega(n^{\log^{(3d)} n})$, where $\log^{(i)} n$ denotes the i-th iterated logarithm. The family of functions $\{F_d\}_{d\in\mathbb{N}}$ witnessing this separation are depth-d read-once monotone formulas with alternating layers of AND and OR gates with fan-in $n^{1/d}$ — these came to be known as the *Sipser functions.* Following Sipser's work, Yao claimed an improvement of Sipser's lower bound to $\exp(n^{c_d})$ for some constant $c_d > 0$ [Yao85]. Shortly thereafter Håstad proved a near-optimal separation for (a slight variant of) the Sipser functions:

Theorem 4 (Depth hierarchy of small-depth circuits [Hås86a]; see also [Hås86b, Hås89])**.** *For every $d \in \mathbb{N}$, there exists a Boolean function $F_d : \{0,1\}^n \to \{0,1\}$ such that F_d is computed by a linear-size depth-d circuit, but any depth-$(d-1)$ circuit computing F_d has size $\exp(n^{\Omega(1/d)})$.*

The parameters of Håstad's theorem were subsequently refined by Cai, Chen, and Håstad [CCH98], and Segerlind, Buss, and Impagliazzo [SBI04]. Prior to the work of Yao and Håstad, Klawe, Paul, Pippenger, and Yannakakis [KPPY84] proved a depth hierarchy theorem for small-depth *monotone* circuits, showing that for every $d \in \mathbb{N}$, depth-$(d-1)$ *monotone* circuits require size $\exp(\Omega(n^{1/(d-1)}))$ to compute the depth-d Sipser function. Klawe et al. also gave an upper bound, showing that every linear-size monotone formula — in particular, the depth-d Sipser function for all $d \in \mathbb{N}$ — can be computed by a depth-k monotone formula of size $\exp(O(k\, n^{1/(k-1)}))$ for all $k \in \mathbb{N}$.

To the best of our knowledge, the first progress towards an *average-case* depth hierarchy theorem for small-depth circuits was made by O'Donnell and Wimmer [OW07]. They constructed a linear-size depth-3 circuit F and proved that any depth-2 circuit that approximates F must have size $2^{\Omega(n/\log n)}$:

Theorem 5 (Theorem 1.9 of [OW07])**.** *For $w \in \mathbb{N}$ and $n := w2^w$, let $\mathsf{Tribes} : \{0,1\}^n \to \{0,1\}$ be the function computed by a 2^w-term read-once monotone DNF formula where every term has width exactly w. Let $\mathsf{Tribes}^\dagger$ denote its Boolean dual, the function computed by a 2^w-clause read-once monotone CNF formula where every clause has width exactly w, and define the $2n$-variable function $F : \{0,1\}^{2n} \to \{0,1\}$ as*

$$F(x) = \mathsf{Tribes}(x_1, \ldots, x_n) \vee \mathsf{Tribes}^\dagger(x_{n+1}, \ldots, x_{2n}).$$

Then any depth-2 circuit C on $2n$ variables that has size $2^{O(n/\log n)}$ agrees with F on at most a 0.99-fraction of the 2^{2n} inputs. (Note that F is computed by a linear-size depth-3 circuit.)

Our Theorem 1 gives an analogous separation between depth-d and depth-$(d+1)$ for all $d \geq 2$, with $(1/2 - o_n(1))$-inapproximability rather than 0.01-inapproximability. The [OW07] size lower bound of $2^{\Omega(n/\log n)}$ is much larger, in the case $d = 2$, than our $\exp(n^{\Omega(1/d)})$ size bound. However, we recall that achieving a $\exp(\omega(n^{1/(d-1)}))$ lower bound against depth-d circuits for an explicit function, even for worst-case computation, is a well-known and major open problem in complexity theory (see e.g. Chapter §11 of [Juk12] and [Val83, GW13, Vio13]). In particular, an extension of the $2^{\Omega(n/\mathrm{polylog}(n))}$-type lower bound of [OW07] to depth 3, even for worst-case computation, would constitute a significant breakthrough.

1.2 Our main lower bounds

We close this section with precise statements of our two main lower bound results, a discussion of the (near)-optimality of our correlation bounds, and a very high-level overview of our techniques.

Theorem 6 (First main lower bound)**.** *For $2 \leq d \leq \frac{c\sqrt{\log n}}{\log\log n}$, the n-variable Sipser_d function has the following property: Any depth-d circuit $C : \{0,1\}^n \to \{0,1\}$ of size at most $S = 2^{n^{\frac{1}{6(d-1)}}}$ and bottom fan-in $\frac{\log n}{10(d-1)}$ agrees with Sipser_d on at most $(\frac{1}{2} + n^{-\Omega(1/d)}) \cdot 2^n$ inputs.*

Theorem 7 (Second main lower bound)**.** *For $2 \leq d \leq \frac{c\sqrt{\log n}}{\log\log n}$, the n-variable Sipser_d function has the following property: Any depth-d circuit $C : \{0,1\}^n \to \{0,1\}$ of size at most $S = 2^{n^{\frac{1}{6(d-1)}}}$ and the opposite alternation pattern to Sipser_d (i.e. its top-level output gate is OR if Sipser_d's is AND and vice versa) agrees with Sipser_d on at most $(\frac{1}{2} + n^{-\Omega(1/d)}) \cdot 2^n$ inputs.*

Clearly both these results imply Theorem 1 as a special case, since any size-S depth-$(d-1)$ circuit may be viewed as a size-S depth-d circuit satisfying the assumptions of Theorems 6 and 7.

(Near)-optimality of our correlation bounds. For constant d, our main result shows that the depth-d Sipser_d function has correlation at most $(1/2 + n^{-\Omega(1)})$ with any subexponential-size circuit of depth $d-1$. Since Sipser_d is a monotone function, well-known results [BT96] imply that its correlation with some input variable x_i or one of the constant functions 0,1 (trivial approximators of depth at most one) must be at least $(1/2 + \Omega(1/n))$; thus significant improvements on our correlation bound cannot be achieved for this (or for any monotone) function.

What about non-monotone functions? If $\{f_d\}_{d\geq 2}$ is any family of n-variable functions computed by poly(n)-size, depth-d circuits, the "discriminator lemma" of Hajnal et al. [HMP$^+$93] implies that f_d must have correlation at least $(1/2+n^{-O(1)})$ with one of the depth-$(d-1)$ circuits feeding into its topmost gate. Therefore a "d versus $d-1$" depth hierarchy theorem for correlation $(1/2+n^{-\omega(1)})$ does not hold.

Our techniques. Our approach is based on *random projections*, a generalization of random restrictions. At a high level, we design a carefully chosen (adaptively chosen) sequence of random projections, and argue that with high probability under this sequence of random projections, (i) any circuit C of the type specified in Theorem 6 or Theorem 7 "collapses," while (ii) the Sipser_d function "retains structure," and (iii) moreover this happens in such a way as to imply that the circuit C must have originally been a very poor approximator for Sipser_d (before the random projections). Each of (i)–(iii) above requires significant work; see Section 4 for a much more detailed explanation of our techniques (and of why previous approaches were unable to successfully establish the result).

2 Application #1: Random oracles separate the polynomial hierarchy

2.1 Background: $\mathsf{PSPACE} \neq \mathsf{PH}$ relative to a random oracle

The pioneering work on lower bounds against small-depth circuits in the 1980's was largely motivated by a connection between small-depth computation and the polynomial hierarchy shown by Furst, Saxe, and Sipser [FSS81]. They gave a super-polynomial size lower bound for constant-depth circuits, proving that depth-d circuits computing the n-variable parity function must have size $\Omega(n^{\log^{(3d-6)} n})$, where $\log^{(i)} n$ denotes the i-th iterated logarithm. They also showed that an improvement of this lower bound to super-quasipolynomial for constant-depth circuits (i.e. $\Omega_d\big(2^{(\log n)^k}\big)$ for all constants k) would yield an oracle A such that $\mathsf{PSPACE}^A \neq \mathsf{PH}^A$. Ajtai independently proved a stronger lower bound of $n^{\Omega_d(\log n)}$ [Ajt83]; his motivation came from finite model theory. Yao gave the first super-quasipolynomial lower bounds on the size of constant-depth circuits computing the parity function [Yao85], and shortly after Håstad proved the optimal lower bound of $\exp(\Omega(n^{1/(d-1)}))$ via his influential Switching Lemma [Hås86a].

Yao's relativized separation of PSPACE from PH was improved qualitatively by Cai, who showed that the separation holds even relative to a *random* oracle [Cai86]. Leveraging the connection made by [FSS81], Cai accomplished this by proving *correlation bounds* against constant-depth circuits, showing that constant-depth circuits of sub-exponential size agree with the parity function only on a $(1/2 + o_n(1))$ fraction of inputs. (Independent work of Babai [Bab87] gave a simpler proof of the same relativized separation.)

2.2 Background: The polynomial hierarchy is infinite relative to some oracle

Together, these results paint a fairly complete picture of the status of the PSPACE versus PH question in relativized worlds: not only does there exist an oracle A such that $\mathsf{PSPACE}^A \neq \mathsf{PH}^A$, this separation holds relative to almost all oracles. A natural next step is to seek analogous results showing that the relativized polynomial hierarchy is infinite; we recall that the polynomial hierarchy being infinite implies $\mathsf{PSPACE} \neq \mathsf{PH}$, and furthermore, this implication relativizes. We begin with the following question, attributed to Albert Meyer in [BGS75]:

Meyer's Question. *Is there a relativized world within which the polynomial hierarchy is infinite? Equivalently, does there exist an oracle A such that $\Sigma_d^{\mathrm{P},A} \subsetneq \Sigma_{d+1}^{\mathrm{P},A}$ for all $d \in \mathbb{N}$?*

Early work on Meyer's question predates [FSS81]. It was first considered by Baker, Gill, and Solovay in their paper introducing the notion of relativization [BGS75], in which they prove the existence of an oracle A such that $\mathsf{P}^A \neq \mathsf{NP}^A \neq \mathsf{coNP}^A$, answering Meyer's question in the affirmative for $d \in \{0, 1\}$. Subsequent work of Baker and Selman proved the $d = 2$ case [BS79]. Following [FSS81], Sipser noted the analogous connection between Meyer's question and circuit lower bounds [Sip83]: to answer Meyer's question in the affirmative, it suffices to exhibit, for every constant $d \in \mathbb{N}$, a Boolean function F_d computable by a depth-d AC^0 circuit such that any depth-$(d-1)$ circuit computing F_d requires super-quasipolynomial size. (This is a significantly more delicate task than proving super-quasipolynomial size lower bounds for the parity function; see Section 4 for a detailed discussion.) Sipser also constructed a family of Boolean functions for which he proved an n versus $\Omega(n^{\log^{(3d)} n})$ separation — these came to be known as the *Sipser functions*, and they play the same central role in Meyer's question as the parity function does in the relativized PSPACE versus PH problem.

As discussed in the introduction (see Theorem 4), Håstad gave the first proof of a near-optimal n versus $\exp(n^{\Omega(1/d)})$ separation for the Sipser functions [Hås86a], obtaining a strong depth hierarchy theorem for small-depth circuits and answering Meyer's question in the affirmative for all $d \in \mathbb{N}$.

2.3 This work: The polynomial hierarchy is infinite relative to a random oracle

Given Håstad's result, a natural goal is to complete our understanding of Meyer's question by showing that the polynomial hierarchy is not just infinite with respect to *some* oracle, but in fact with respect to *almost all* oracles. Indeed, in [Hås86a, Hås86b, Hås89], Håstad poses the problem of extending his result to show this as an open question:

Question 1 (Meyer's Question for Random Oracles [Hås86a, Hås86b, Hås89])**.** *Is the polynomial hierarchy infinite relative to a random oracle? Equivalently, does a random oracle A satisfy $\Sigma_d^{\mathrm{P},A} \subsetneq \Sigma_{d+1}^{\mathrm{P},A}$ for all $d \in \mathbb{N}$?*

Question 1 also appears as the main open problem in [Cai86, Bab87]; as mentioned above, an affirmative answer to Question 1 would imply Cai and Babai's result showing that $\mathsf{PSPACE}^A \neq \mathsf{PH}^A$ relative to a random oracle A. Further motivation for studying Question 1 comes from a surprising result of Book, who proved that the *unrelativized* polynomial hierarchy collapses if it collapses relative to a random oracle [Boo94]. Over the years Question 1 has been discussed in a wide range of surveys [Joh86, Hem94, ST95, HRZ95, VW97, Aar], textbooks [DK00, HO02], and research papers [Hås86b, Hås89, Tar89, For99, Aar10a].

Our work. As a corollary of our main result (Theorem 1) — an *average-case* depth hierarchy theorem for small-depth circuits — we answer Question 1 in the affirmative for all $d \in \mathbb{N}$:

Theorem 2. *The polynomial hierarchy is infinite relative to a random oracle: with probability 1, a random oracle A satisfies $\Sigma_d^{\mathrm{P},A} \subsetneq \Sigma_{d+1}^{\mathrm{P},A}$ for all $d \in \mathbb{N}$.*

Prior to our work, the $d \in \{0,1\}$ cases were proved by Bennett and Gill in their paper initiating the study of random oracles [BG81]. Motivated by the problem of obtaining relativized separations in quantum structural complexity, Aaronson recently showed that a random oracle A separates Π_2^{P} from P^{NP} [Aar10b, Aar10a]; he conjectures in [Aar10a] that his techniques can be extended to resolve the $d = 2$ case of Theorem 2. We observe that O'Donnell and Wimmer's techniques (Theorem 5 in our introduction) can be used to prove the $d = 2$ case [OW07], though the authors of [OW07] do not discuss this connection to the relativized polynomial hierarchy in their paper.

	$\mathsf{PSPACE}^A \neq \mathsf{PH}^A$	$\Sigma_d^{\mathrm{P},A} \subsetneq \Sigma_{d+1}^{\mathrm{P},A}$ for all $d \in \mathbb{N}$
Connection to lower bounds for constant-depth circuits	[FSS81]	[Sip83]
Hard function(s)	Parity	Sipser functions
Relative to *some* oracle A	[Yao85, Hås86a]	[Yao85, Hås86a]
Relative to *random* oracle A	[Cai86, Bab87]	**This work**

Table 1: Previous work and our result on the relativized polynomial hierarchy

We refer the reader to Chapter §7 of Håstad's thesis [Hås86b] for a detailed exposition (and complete proofs) of the aforementioned connections between small-depth circuits and the polynomial hierarchy (in particular, for the proof of how Theorem 2 follows from Theorem 1).

3 Application #2: No approximate converse to Boppana–Linial–Mansour–Nisan

The famous result of Linial, Mansour, and Nisan gives strong bounds on Fourier concentration of small-depth circuits [LMN93]. As a corollary, they derive an upper bound on the total influence of small-depth circuits, showing that depth-d size-S circuits have total influence $(O(\log S))^d$. (We remind the reader that the total influence of an n-variable Boolean function f is $\mathbf{Inf}(f) := \sum_{i=1}^n \mathbf{Inf}_i(f)$, where $\mathbf{Inf}_i(f)$ is the probability that flipping coordinate $i \in [n]$ of a uniform random input from $\{0,1\}^n$ causes the value of f to change.) This was subsequently sharpened by Boppana via a simpler and more direct proof [Bop97]:

Theorem 8 (Boppana, Linial–Mansour–Nisan)**.** *Let $f : \{0,1\}^n \to \{0,1\}$ be a computed by a size-S depth-d circuit. Then $\mathbf{Inf}(f) = (O(\log S))^{d-1}$.*

(We note that Boppana's bound is asymptotically tight by considering the parity function.) Several researchers have asked whether an *approximate converse* of some sort holds for Theorem 8:

> *If $f : \{0,1\}^n \to \{0,1\}$ has low total influence, is it the case that f can be approximated to high accuracy by a small constant-depth circuit?*

A result of this flavor, taken together with Theorem 8, would yield an elegant characterization of Boolean functions with low total influence. In this section we formulate a very weak approximate converse to Theorem 8 and show, as a consequence of our main result (Theorem 1), that even this weak converse does not hold.

3.1 Background: BKS conjecture and O'Donnell–Wimmer's counterexample

An approximate converse to Theorem 8 was first conjectured by Benjamini, Kalai, and Schramm, with a very specific quantitative bound on how the size of the approximating circuit depends on its influence and depth [BKS99] (the conjecture also appears in the surveys [Kal00, KS05]). They posed the following:

Benjamini–Kalai–Schramm (BKS) Conjecture. *For every $\varepsilon > 0$ there is a constant $K = K(\varepsilon)$ such that the following holds: Every monotone $f : \{0,1\}^n \to \{0,1\}$ can be ε-approximated by a depth-d circuit of size at most*

$$\exp\left((K \cdot \mathbf{Inf}(f))^{1/(d-1)}\right)$$

for some $d \geq 2$.

(We associate a circuit with the Boolean function that it computes, and we say that a circuit *ε-approximates* a Boolean function f if it agrees with f on all but an ε-fraction of all inputs.) If true, the BKS conjecture would give a quantitatively strong converse to Theorem 8 for monotone functions.[1] In addition, it would have important implications for the study of threshold phenomena in Erdös–Rényi random graphs, which is the context in which Benjamini, Kalai, and Schramm made their conjecture; we refer the reader to [BKS99] and Section 1.4 of [OW07] for a detailed discussion of this connection. However, the BKS conjecture was disproved by O'Donnell and Wimmer [OW07]. Their result (Theorem 5 in our introduction) disproves the case $d = 2$ of the BKS conjecture, and the case $d > 2$ is disproved by an easy argument which [OW07] give.

3.2 This work: Disproving a weak variant of the BKS conjecture

A significantly weaker variant of the BKS conjecture is the following:

Conjecture 1. *For every $\varepsilon > 0$ there is a $d = d(\varepsilon)$ and $K_1 = K_1(\varepsilon), K_2 = K_2(\varepsilon)$ such that the following holds: Every monotone $f : \{0,1\}^n \to \{0,1\}$ can be ε-approximated by a depth-d circuit of size at most*

$$\exp\left((K_1 \cdot \mathbf{Inf}(f))^{K_2}\right).$$

The [OW07] counterexample to the BKS conjecture does not disprove Conjecture 1; indeed, the function f that [OW07] construct and analyze is computed by a depth-3 circuit of size $O(n)$.[2] Observe that Conjecture 1, if true, would yield the following rather appealing consequence: every monotone $f : \{0,1\}^n \to \{0,1\}$ with total influence at most polylog(n) can be approximated to any constant accuracy by a quasipolynomial-size, constant-depth circuit (where both the constant in the quasipolynomial size bound and the constant depth of the circuit may depend on the desired accuracy).

Following O'Donnell and Wimmer's disproof of the BKS conjecture, several researchers have posed questions similar in spirit to Conjecture 1. O'Donnell asked if the BKS conjecture is true if the bound on the size of the approximating circuit is allowed to be $\exp\left((K \cdot \mathbf{Inf}(f))^{1/d}\right)$ instead

[1] We remark that although the BKS conjecture was stated for monotone Boolean functions, it seems that (a priori) it could have been true for all Boolean functions: prior to [OW07], we are not aware of any counterexample to the BKS conjecture even if f is allowed to be non-monotone.

[2] As with the BKS conjecture, prior to our work we are not aware of any counterexample to Conjecture 1 even if f is allowed to be non-monotone.

of $\exp\left((K \cdot \mathbf{Inf}(f))^{1/(d-1)}\right)$ [O'D07]. This is a weaker statement than the original BKS conjecture (in particular, it is not ruled out by the counterexample of [OW07]), but still significantly stronger than Conjecture 1. Subsequently Kalai asked if Boolean functions with total influence $\mathrm{polylog}(n)$ (resp. $O(\log n)$) can be approximated by constant-depth circuits of quasipolynomial size (resp. AC^0) [Kal12] (see also [Kal10] where he states a qualitative version). Kalai's question is a variant of Conjecture 1 in which f is allowed to be non-monotone, but $\mathbf{Inf}(f)$ is only allowed to be $\mathrm{polylog}(n)$; furthermore, $K_2(\varepsilon)$ is only allowed to be 1 if $\mathbf{Inf}(f) = O(\log n)$. Finally, H. Hatami recently restated the $\mathbf{Inf}(f) = O(\log n)$ case of Kalai's question:

Problem 4.6.3 of [Hat14]. *Is it the case that for every $\varepsilon, C > 0$, there are constants d, k such that for every $f : \{0,1\}^n \to \{0,1\}$ with $\mathbf{Inf}(f) \le C \log n$, there is a size-$n^k$, depth-$d$ circuit which ε-approximates f?*

Our work. As a corollary of our main result (Theorem 1), we show that Conjecture 1 is false even for (suitable choices of) $\varepsilon = \frac{1}{2} - o_n(1)$. Our counterexample also provides a strong negative answer to O'Donnell's and Kalai–Hatami's versions of Conjecture 1. We prove the following:

Theorem 3. *Conjecture 1 is false. More precisely, there is a monotone $f : \{0,1\}^n \to \{0,1\}$ and a $\delta(n) = o_n(1)$ such that $\mathbf{Inf}(f) = O(\log n)$ but any circuit of depth $d(n) = \sqrt{\log\log n}$ that agrees with f on $(\frac{1}{2} + \delta(n))$ fraction of all inputs must have size at least $S(n) = 2^{2^{\tilde{\Omega}\left(2^{\sqrt{\log\log n}}\right)}}$.*

Proof of Theorem 3 assuming Theorem 1. Consider the monotone Boolean function $f : \{0,1\}^n \to \{0,1\}$ corresponding to Sipser_d of Theorem 1 defined over the first $m = 2^{2^{\lfloor\sqrt{\log\log n}\rfloor}}$ variables, and of depth $d = \lfloor \log\log m \rfloor + 1 = \lfloor \sqrt{\log\log n} \rfloor + 1$. By Boppana's theorem (Theorem 8), we have that

$$\mathbf{Inf}(f) = O(\log m)^{d-1} = O\left(2^{\lfloor\sqrt{\log\log n}\rfloor}\right)^{\lfloor\sqrt{\log\log n}\rfloor} = O(\log n).$$

On the other hand, our main theorem (Theorem 1) implies that even circuits of depth $d - 1 = \lfloor\sqrt{\log\log n}\rfloor$ which agree with f on $(\frac{1}{2}+\delta(n))$ fraction of all inputs, where $\delta(n) = 2^{-\Omega(2^{\lfloor\sqrt{\log\log n}\rfloor}/\lfloor\sqrt{\log\log n}\rfloor)}$, must have size at least

$$S(n) = 2^{m^{\Omega(1/d)}} = 2^{\left(2^{2^{\sqrt{\log\log n}}}\right)^{\Omega(1/\sqrt{\log\log n})}} = 2^{2^{\tilde{\Omega}\left(2^{\sqrt{\log\log n}}\right)}}. \qquad \square$$

4 Our techniques

The method of random restrictions dates back to Subbotovskaya [Sub61] and continues to be an indispensable technique in circuit complexity. Focusing only on small-depth circuits, we mention that the random restriction method is the common essential ingredient underlying the landmark lower bounds discussed in the previous sections [FSS81, Ajt83, Sip83, Yao85, Hås86a, Cai86, Bab87, IMP12, Hås14].

We begin in Section 4.1 by describing the general framework for proving worst- and average-case lower bounds against small-depth circuits via the random restriction method. Within this framework, we sketch the now-standard proof of correlation bounds for the parity function based on Håstad's Switching Lemma. We also recall why the lemma is not well-suited for proving a depth hierarchy theorem for small-depth circuits, hence necessitating the "blockwise variant" of the lemma that Håstad developed and applied to prove his (worst-case) depth hierarchy theorem. In Section 4.2 we highlight the difficulties that arise in extending Håstad's depth hierarchy theorem to the average-case, and how our techniques — specifically, the notion of random *projections* — allow us to overcome these difficulties.

4.1 Background: Lower bounds via random restrictions

Suppose we would like to show that a *target function* $f : \{0,1\}^n \to \{0,1\}$ has small correlation with any size-S depth-d *approximating circuit* C under the uniform distribution $\mathcal{U}$ over $\{0,1\}^n$. A standard approach is to construct a series of random restrictions $\{\mathcal{R}_k\}_{k\in\{2,\ldots,d\}}$ satisfying three properties:

- **Property 1: Approximator C simplifies.** The randomly-restricted circuit $C \upharpoonright \boldsymbol{\rho}^{(d)} \cdots \boldsymbol{\rho}^{(2)}$, where $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}_k$ for $2 \le k \le d$, should "collapse to a simple function" with high probability. This is typically shown via iterative applications of an appropriate "Switching Lemma for the $\mathcal{R}_k$'s", which shows that each random restriction $\boldsymbol{\rho}^{(k)}$ decreases the depth of the circuit $C \upharpoonright \boldsymbol{\rho}^{(d)} \cdots \boldsymbol{\rho}^{(k-1)}$ by one with high probability. The upshot is that while C is a depth-d size-S circuit, $C \upharpoonright \boldsymbol{\rho}^{(d)} \cdots \boldsymbol{\rho}^{(2)}$ will be a small-depth decision tree, a "simple function", with high probability.
- **Property 2: Target f retains structure.** In contrast with the approximating circuit, the target function f should (roughly speaking) be resilient against the random restrictions $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}_k$. While the precise meaning of "resilient" depends on the specific application, the key property we need is that $f \upharpoonright \boldsymbol{\rho}^{(d)} \cdots \boldsymbol{\rho}^{(2)}$ will with high probability be a "well-structured" function that is uncorrelated with any small-depth decision tree.

Together, these two properties imply that random restrictions of f and C are uncorrelated with high probability. Note that this already yields *worst-case* lower bounds, showing that $f : \{0,1\}^n \to \{0,1\}$ cannot be computed exactly by C. To obtain correlation bounds, we need to translate such a statement into the fact that f and C *themselves* are uncorrelated. For this we need the third key property of the random restrictions:

- **Property 3: Composition of $\mathcal{R}_k$'s completes to $\mathcal{U}$.** Evaluating a Boolean function $h : \{0,1\}^n \to \{0,1\}$ on a random input $\mathbf{X} \leftarrow \mathcal{U}$ is equivalent to first applying random restrictions $\boldsymbol{\rho}^{(d)}, \ldots, \boldsymbol{\rho}^{(2)}$ to h, and then evaluating the randomly-restricted function $h \upharpoonright \boldsymbol{\rho}^{(d)} \cdots \boldsymbol{\rho}^{(2)}$ on $\mathbf{X}' \leftarrow \mathcal{U}$.

Correlation bounds for parity. For uniform-distribution correlation bounds against constant-depth circuits computing the parity function, the random restrictions are all drawn from $\mathcal{R}(p)$, the "standard" random restriction which independently sets each free variable to 0 with probability $\frac{1}{2}(1-p)$, to 1 with probability $\frac{1}{2}(1-p)$, and keeps it free with probability p. The main technical challenge arises in proving that Property 1 holds — this is precisely Håstad's Switching Lemma — whereas Properties 2 and 3 are straightforward to show. For the second property, we note that

$$\mathsf{Parity}_n \upharpoonright \rho \equiv \pm\,\mathsf{Parity}(\rho^{-1}(*)) \quad \text{for all restrictions } \rho \in \{0,1,*\}^n,$$

and so $\mathsf{Parity}_n \upharpoonright \boldsymbol{\rho}^{(d)} \cdots \boldsymbol{\rho}^{(2)}$ computes the parity of a random subset $\mathbf{S} \subseteq [n]$ of coordinates (or its negation). With an appropriate choice of the $*$-probability p we have that $|\mathbf{S}|$ is large with high probability; recall that $\pm\,\mathsf{Parity}_k$ (the k-variable parity function or its negation) has zero correlation with any decision tree of depth at most $k-1$. For the third property, we note that for all values of $p \in (0,1)$, a random restriction $\boldsymbol{\rho} \leftarrow \mathcal{R}(p)$ specifies a uniform random subcube of $\{0,1\}^n$ (of dimension $|\boldsymbol{\rho}^{-1}(*)|$). Therefore, the third property is a consequence of the simple fact that a uniform random point within a uniform random subcube is itself a uniform random point from $\{0,1\}^n$.

Håstad's blockwise random restrictions. With the above framework in mind, we notice a conceptual challenge in proving AC^0 depth hierarchy theorems via the random restriction method: even focusing only on the worst-case (i.e. ignoring Property 3), the random restrictions $\mathcal{R}_k$ will have to satisfy Properties 1 and 2 with the target function f being *computable in* AC^0. This is a significantly more delicate task than (say) proving $\mathsf{Parity} \notin \mathsf{AC}^0$ since, roughly speaking, in the latter case the target function $f \equiv \mathsf{Parity}$ is "much more complex" than the circuit $C \in \mathsf{AC}^0$ to begin with. In an AC^0 depth hierarchy theorem, *both* the target f and the approximating circuit C are constant-depth circuits; the target f is "more complex" than C in the sense that it has larger circuit depth, but this is offset by the fact that the circuit size of C is allowed to be exponentially larger than that of f (as is the case in both Håstad's and our theorem). We refer the reader to Chapter §6.2 of Hastad's thesis [Hås86b] which contains a discussion of this very issue.

Håstad overcomes this difficulty by replacing the "standard" random restrictions $\mathcal{R}(p)$ with random restrictions *specifically suited to Sipser functions being the target*: his "blockwise" random restrictions are designed so that (1) they reduce the depth of the formula computing the Sipser function by one, but otherwise essentially preserve the rest of its structure, and yet (2) a switching lemma still holds for any circuit with sufficiently small bottom fan-in. These correspond to Properties 2 and 1 respectively. However, unlike $\mathcal{R}(p)$, Håstad's blockwise random restrictions are not independent across coordinates and do not satisfy Property 3: their composition does not complete to the uniform distribution $\mathcal{U}$ (and indeed it does not complete to any product distribution). This is why Håstad's construction establishes a worst-case rather than average-case depth hierarchy theorem.

4.2 Our main technique: Random projections

The crux of the difficulty in proving an average-case AC^0 depth hierarchy theorem therefore lies in designing random restrictions that satisfy Properties 1, 2, and 3 simultaneously, for a target f in AC^0 and an arbitrary approximating circuit C of smaller depth but possibly exponentially larger size. To recall, the "standard" random restrictions $\mathcal{R}(p)$ satisfy Properties 1 and 3 but not 2, and Håstad's blockwise variant satisfies Properties 1 and 2 but not 3.

In this paper we overcome this difficulty with *projections*, a generalization of restrictions. Given a set of formal variables $\mathcal{X} = \{x_1, \ldots, x_n\}$, a restriction ρ either fixes a variable x_i (i.e. $\rho(x_i) \in \{0,1\}$) or keeps it alive (i.e. $\rho(x_i) = x_i$, often denoted by $*$). A *projection*, on the other hand, either fixes x_i or maps it to a variable y_j from a possibly different space of formal variables $\mathcal{Y} = \{y_1, \ldots, y_{n'}\}$. Restrictions are therefore a special case of projections where $\mathcal{Y} \equiv \mathcal{X}$, and each x_i can only be fixed or mapped to itself. (See Definition 4 for precise definitions.) Our arguments crucially employ projections in which $\mathcal{Y}$ is smaller than $\mathcal{X}$, and where moreover each x_i is only mapped to a specific element y_j where j depends on i in a carefully designed way that depends on the structure of the formula computing the Sipser function. Such "collisions", where blocks of distinct formal variables in $\mathcal{X}$ are mapped to the same new formal variable $y_i \in \mathcal{Y}$, play a crucial role in our approach. (We remark that ours is not the first work to consider such a generalization of restrictions. Random projections are also used in the work of Impagliazzo and Segerlind, which establishes lower bounds against constant-depth Frege systems with counting axioms in proof complexity [IS01].)

At a high level, our overall approach is structured around a sequence $\mathbf{\Psi}$ of (*adaptively chosen*) random projections satisfying Properties 1, 2, and 3 simultaneously, with the target f being Sipser, a slight variant of the Sipser function which we define in Section 6. We briefly outline how we establish each of the three properties (it will be more natural for us to prove them in a slightly different order from the way they are listed in Section 4.1):

– **Property 3: $\mathbf{\Psi}$ completes to the uniform distribution.** Like Håstad's blockwise random restrictions (and unlike the "standard" random restrictions $\mathcal{R}(p)$), the distributions of our random projections are not independent across coordinates: they are carefully correlated in a way that depends on the structure of the formula computing Sipser. As discussed above, there is an inherent tension between the need for such correlations on one hand (to ensure that Sipser "retains structure"), and the requirement that their composition completes to the uniform distribution on the other hand (to yield uniform-distribution correlation bounds). We overcome this difficulty with our notion of projections: in Section 8 we prove that the composition $\mathbf{\Psi}$ of our sequence of random projections completes to the uniform distribution (despite the fact that every one of the individual random projections comprising $\mathbf{\Psi}$ is highly-correlated among coordinates.)

– **Property 1: Approximator C simplifies.** Next we prove that approximating circuits C of the types specified in our main lower bounds (Theorems 6 and 7) "collapse to a simple function" with high probability under our sequence $\mathbf{\Psi}$ of random projections. Following the standard "bottom-up" approach to proving lower bounds against small-depth circuits, we establish this by arguing that each of the individual random projections comprising $\mathbf{\Psi}$ "contributes to the simplification" of C by reducing its depth by (at least) one.

  More precisely, in Section 9 we prove a *projection switching lemma*, showing that a small-width DNF or CNF "switches" to a small-depth decision tree with high probability under our random projections. (The depth reduction of C follows by applying this lemma to every one of its bottom-level depth-2 subcircuits.) Recall that the random projection of a depth-2 circuit over a set of formal variables $\mathcal{X}$ yields a function over a new set of formal variables $\mathcal{Y}$, and in our case $\mathcal{Y}$ is significantly smaller than $\mathcal{X}$. In addition to the structural simplification that results from setting variables to constants (as in Håstad's Switching Lemma for random *restrictions*), the proof of our projection switching lemma also crucially exploits the additional structural simplification that results from distinct variables in $\mathcal{X}$ being mapped to the same variable in $\mathcal{Y}$.

– **Property 2: Target Sipser retains structure.** Like Håstad's blockwise random restrictions, our random projections are defined with the target function Sipser in mind; in particular, they are carefully designed so as to ensure that Sipser "retains structure" with high probability under their composition $\mathbf{\Psi}$.

  In Section 10.1 we define the notion of a "typical" outcome of our random projections, and prove that with high probability *all* the individual projections comprising $\mathbf{\Psi}$ are typical. (Since our sequence of random projections is chosen adaptively, this requires a careful definition of typicality to facilitate an inductive argument showing that our definition "bootstraps" itself.) Next, in Section 10.2 we show that typical projections have a "very limited and well-controlled" effect on the structure of Sipser; equivalently, Sipser is resilient against typical projections. Together, the results of Section 10.1 and 10.2 show that with high probability, Sipser reduces under $\boldsymbol{\Psi}$ to a "well-structured" formula, in sharp contrast with our results from Section 9 showing that the approximator "collapses to a simple function" with high probability under $\boldsymbol{\Psi}$.

We remark that the notion of random projections plays a key role in ensuring all three properties above. (We give a more detailed overview of our proof in Section 7.3 after setting up the necessary terminology and definitions in the next two sections.)

5 Preliminaries

5.1 Basic mathematical tools

Fact 5.1 (Chernoff bounds)**.** *Let* $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *be independent random variables satisfying* $0 \le \mathbf{Z}_i \le 1$ *for all* $i \in [n]$. *Let* $\mathbf{S} = \mathbf{Z}_1 + \cdots + \mathbf{Z}_n$, *and* $\mu = \mathbf{E}[\mathbf{S}]$. *Then for all* $\gamma \ge 0$,

$$\mathbf{Pr}[\mathbf{S} \ge (1+\gamma)\mu] \le \exp\left(-\frac{\gamma^2}{2+\gamma}\cdot\mu\right)$$
$$\mathbf{Pr}[\mathbf{S} \le (1-\gamma)\mu] \le \exp\left(-\frac{\gamma^2}{2}\cdot\mu\right).$$

We will use the following fact implicitly in many of our calculations:

Fact 5.2. *Let* $\delta = \delta(n) > 0$ *and* $n \in \mathbb{N}$, *and suppose* $\delta n = o_n(1)$. *The following inequalities hold for sufficiently large* n:

$$1 - \delta n \le (1-\delta)^n \le 1 - \tfrac{1}{2}\delta n.$$

Finally, the following standard approximations will be useful:

Fact 5.3. *For* $x \ge 2$, *we have*

$$e^{-1}\left(1 - \frac{1}{x}\right) \le \left(1 - \frac{1}{x}\right)^x \le e^{-1}, \qquad \textit{or equivalently,} \qquad \left(1 - \frac{1}{x}\right)^x \le e^{-1} \le \left(1 - \frac{1}{x}\right)^{x-1},$$

and for $0 \le x \le 1$, *we have* $1 + x \le e^x \le 1 + 2x$.

We write log to denote logarithm base 2 and ln to denote natural log.

5.2 Notation

A DNF is an OR of ANDs (terms) and a CNF is an AND of ORs (clauses). The *width* of a DNF (respectively, CNF) is the maximum number of variables that occur in any one of its terms (respectively, clauses). We will assume throughout that our circuits are *alternating*, meaning that every root-to-leaf path alternates between AND gates and OR gates, and *layered*, meaning that for every gate G, every root-to-G path has the same length. By a standard conversion, every depth-d circuit is equivalent to a depth-d alternating layered circuit with only a modest increase in size (which is negligible given the slack on our analysis). The size of a circuit is its number of gates, and the depth of a circuit is the length of its longest root-to-leaf path.

For $p \in [0,1]$ and symbols $\bullet, \circ$, we write "$\{\bullet_p, \circ_{1-p}\}$" to denote the distribution over $\{\bullet, \circ\}$ which outputs $\bullet$ with probability p and $\circ$ with probability $1-p$. We write "$\{\bullet_p, \circ_{1-p}\}^k$" to denote the product distribution over $\{\bullet,\circ\}^k$ in which each coordinate is distributed independently according to $\{\bullet_p,\circ_{1-p}\}$. We write "$\{\bullet_p,\circ_{1-p}\}^k\setminus\{\bullet\}^k$" to denote the product distribution conditioned on not outputting $\{\bullet\}^k$.

Given $\tau\in\{0,1,*\}^{A\times[\ell]}$ and $a\in A$, we write τ_a to denote the ℓ-character string $(\tau_{a,i})_{i\in[\ell]}\in\{0,1,*\}^{[\ell]}$, and we sometimes refer to this as the "a-th block of τ."

Throughout the paper we use boldfaced characters such as $\boldsymbol{\rho}$, $\mathbf{X}$, etc. to denote random variables. We write "$a=b\pm c$" as shorthand to denote that $a\in[b-c,b+c]$, and similarly $a\neq b\pm c$ to denote that $a\notin[b-c,b+c]$. For a positive integer k we write "$[k]$" to denote the set $\{1,\ldots,k\}$.

The *bias* of a Boolean function f under an input distribution $\mathbf{Z}$ is defined as

$$\mathrm{bias}(f,\mathbf{Z}):=\min\left\{\Pr_{\mathbf{Z}}[f(\mathbf{Z})=0],\Pr_{\mathbf{Z}}[f(\mathbf{Z})=1]\right\}.$$

5.3 Restrictions and random restrictions

Definition 1 (Restriction). *A* restriction *ρ of a finite base set $\{x_\alpha\}_{\alpha\in\Omega}$ of Boolean variables is a string $\rho\in\{0,1,*\}^\Omega$. (We sometimes equivalently view a restriction ρ as a function $\rho:\Omega\to\{0,1,*\}$.) Given a function $f:\{0,1\}^\Omega\to\{0,1\}$ and restriction $\rho\in\{0,1,*\}^\Omega$, the ρ-*restriction *of f is the function $(f\upharpoonright\rho):\{0,1\}^\Omega\to\{0,1\}$ where*

$$(f\upharpoonright\rho)(x)=f(x\upharpoonright\rho),\quad \textit{and } (x\upharpoonright\rho)_\alpha:=\begin{cases}x_\alpha & \textit{if } \rho_\alpha=*\\ \rho_\alpha & \textit{otherwise}\end{cases}\quad \textit{for all } \alpha\in\Omega.$$

*Given a distribution $\mathcal{R}$ over restrictions $\{0,1,*\}^\Omega$ the $\mathcal{R}$-*random restriction *of f is the random function $f\upharpoonright\boldsymbol{\rho}$ where $\boldsymbol{\rho}\leftarrow\mathcal{R}$.*

Definition 2 (Refinement). *Let $\rho,\tau\in\{0,1,*\}^\Omega$ be two restrictions. We say that τ is a* refinement *of ρ if $\rho^{-1}(1)\subseteq\tau^{-1}(1)$ and $\rho^{-1}(0)\subseteq\tau^{-1}(0)$, i.e. every variable x_α that is set to 0 or 1 by ρ is set in the same way by τ (and τ may set additional variables to 0 or 1 that ρ does not set).*

Definition 3 (Composition). *Let $\rho,\rho'\in\{0,1,*\}^\Omega$ be two restrictions. Their* composition*, denoted $\rho\rho'\in\{0,1,*\}^\Omega$, is the restriction defined by*

$$(\rho\rho')_\alpha=\begin{cases}\rho_\alpha & \textit{if } \rho_\alpha\in\{0,1\}\\ \rho'_\alpha & \textit{otherwise.}\end{cases}$$

Note that $\rho\rho'$ is a refinement of ρ.

5.4 Projections and random projections

A key ingredient in this work is the notion of *random projections* which generalize random restrictions. Throughout the paper we will be working with functions over spaces of formal variables that are partitioned into disjoint blocks of some length ℓ (see Section 6 for a precise description of these spaces). In other words, our functions will be over spaces of formal variables that can be described as $\mathcal{X}=\{x_{a,i}\colon a\in A,i\in[\ell]\}$, where we refer to $x_{a,i}$ as the i-th variable in the a-th block. We associate with each such space $\mathcal{X}$ a smaller space $\mathcal{Y}=\{y_a\colon a\in A\}$ containing a new formal variable for each block of $\mathcal{X}$. Given a function f over $\mathcal{X}$, the *projection* of f yields a function over $\mathcal{Y}$, and the *random projection* of f is the projection of a random restriction of f (which again is a function over $\mathcal{Y}$). Formally, we have the following definition:

Definition 4 (Projection)**.** *The* projection operator proj *acts on functions* $f : \{0,1\}^{A\times[\ell]} \to \{0,1\}$ *as follows. The* projection of f *is the function* $(\mathrm{proj}\, f) : \{0,1\}^A \to \{0,1\}$ *defined by*

$$(\mathrm{proj}\, f)(y) = f(x) \quad \textit{where } x_{a,i} = y_a \textit{ for all } a \in A \textit{ and } i \in [\ell].$$

Given a restriction $\rho \in \{0,1,*\}^{A\times[\ell]}$*, the* ρ-projection of f *is the function* $(\mathrm{proj}_\rho f) : \{0,1\}^A \to \{0,1\}$ *defined by*

$$(\mathrm{proj}_\rho f)(y) = f(x) \quad \textit{where } x_{a,i} = \begin{cases} y_a & \textit{if } \rho_{a,i} = * \\ \rho_{a,i} & \textit{otherwise} \end{cases} \quad \textit{for all } a \in A \textit{ and } i \in [\ell].$$

Equivalently, $(\mathrm{proj}_\rho f) \equiv (\mathrm{proj}\,(f \upharpoonright \rho))$*. Given a distribution* $\mathcal{R}$ *over restrictions in* $\{0,1,*\}^{A\times[\ell]}$*, the associated random projection operator is* $\mathrm{proj}_{\boldsymbol{\rho}}$ *where* $\boldsymbol{\rho} \leftarrow \mathcal{R}$*, and for* $f : \{0,1\}^{A\times[\ell]} \to \{0,1\}$ *we call* $\mathrm{proj}_{\boldsymbol{\rho}} f$ *its* $\mathcal{R}$-random projection*.*

Note that when $\ell = 1$, the spaces $\mathcal{X}$ and $\mathcal{Y}$ are identical and our definitions of a ρ-projection and $\mathcal{R}$-random projection coincide exactly with that of a ρ-restriction and $\mathcal{R}$-random restriction in Definition 1 (in this case the projection operator proj is simply the identity operator).

Remark 9. The following interpretation of the projection operator will be useful for us. Let f be a function over $\mathcal{X}$, and consider its representation as a circuit C (or decision tree) accessing the formal variables $x_{a,i}$ in $\mathcal{X}$. The projection of f is the function computed by the circuit C', where C' is obtained from C by replacing every occurrence of $x_{a,i}$ in C by y_a for all $a \in A$ and $i \in [\ell]$. Note that this may result in a significant simplification of the circuit: for example, an AND gate (OR gate, respectively) in C that access both $x_{a,i}$ and $\overline{x}_{a,j}$ for some $a \in A$ and $i, j \in [\ell]$ will access both y_a and $\overline{y}_a$ in C', and therefore can be simplified and replaced by the constant 0 (1, respectively). This is a fact we will exploit in the proof of our projection switching lemma in Section 9.1.

6 The Sipser function and its basic properties

For $2 \leq d \in \mathbb{N}$, in this subsection we define the depth-d monotone n-variable read-once Boolean formula Sipser_d and establish some of its basic properties. The Sipser_d function is very similar to the depth-d formula considered by Håstad [Hås86b]; the only difference is that the fan-ins of the gates in the top and bottom layers have been slightly adjusted, essentially so as to ensure that the formula is very close to balanced between the two output values 0 and 1 (note that such balancedness is a prerequisite for any $(1/2 - o_n(1))$-inapproximability result.) The Sipser_d formula is defined in terms of an integer parameter m; in all our results this is an asymptotic parameter that approaches $+\infty$, so m should be thought of as "sufficiently large" throughout the paper.

Every leaf of Sipser_d occurs at the same depth (distance from the root) d; there are exactly n leaves (n will be defined below) and each variable occurs at precisely one leaf. The formula is *alternating*, meaning that every root-to-leaf path alternates between AND gates and OR gates; all of the gates that are adjacent to input variables (i.e. the depth-$(d-1)$ gates) are AND gates, so the root is an OR gate if d is even and is an AND gate if d is odd. The formula is also *depth-regular*, meaning that for each depth (distance from the root) $0 \leq k \leq d-1$, all of the depth-k gates have the same fan-in. Hence to completely specify the Sipser_d formula it remains only to specify the fan-in sequence $w_0, \ldots, w_{d-1}$, where w_k is the fan-in of every gate at depth k. These fan-ins are as follows:

– The bottommost fan-in is

$$w_{d-1} := m. \tag{1}$$

We define

$$p := 2^{-w_{d-1}} = 2^{-m}, \tag{2}$$

and we observe that p is the probability that a depth-$(d-1)$ AND gate is satisfied by a uniform random choice of $\mathbf{X} \leftarrow \{0_{1/2}, 1_{1/2}\}^n$.

– For each value $1 \le k \le d-2$, the value of w_k is $w_k = w$ where

$$w := \lfloor m2^m / \log(e) \rfloor. \tag{3}$$

– The value w_0 is defined to be

$$w_0 := \text{the smallest integer such that } (1-t_1)^{qw_0} \text{ is at most } \frac{1}{2}, \tag{4}$$

where t_1 and q will be defined in Section 7.1, see specifically Equations (8) and (7). Roughly speaking, w_0 is chosen so that the overall formula is essentially balanced under the uniform distribution (i.e. Sipser_d satisfies (6) below); see (9) and the discussion thereafter.

The number of input variables n for Sipser_d is $n = \prod_{k=0}^{d-1} w_k = w^{d-2} w_{d-1} w_0$. The estimates for t_1 and q given in (10) imply that $w_0 = 2^m \ln(2) \cdot (1 \pm o_m(1))$, so we have that

$$n = \frac{1 \pm o_m(1)}{\log e} \cdot \left(\frac{m2^m}{\log e} \right)^{d-1}. \tag{5}$$

We note that for the range of values $2 \le d \le \frac{c\sqrt{\log n}}{\log \log n}$ that we consider in this paper, a direct (but somewhat tedious) analysis implies that the Sipser_d function is indeed essentially balanced, or more precisely, that it satisfies

$$\mathop{\mathbf{Pr}}_{\mathbf{X} \leftarrow \{0_{1,2}, 1_{1,2}\}^n} [\mathsf{Sipser}_d(\mathbf{X}) = 1] = \frac{1}{2} \pm o_n(1). \tag{6}$$

However, since this fact is a direct byproduct of our main theorem (which shows that Sipser_d cannot be $(1/2 - o_n(1))$-approximated by any depth-$(d-1)$ formula, let alone by a constant function), we omit the tedious direct analysis here.

We specify an addressing scheme for the gates and input variables of our Sipser_d formula which will be heavily used throughout the paper. Let $A_0 = \{\mathsf{output}\}$, and for $1 \le k \le d$, let $A_k = A_{k-1} \times [w_{k-1}]$. An element of A_k specifies the address of a gate at depth (distance from the output node) k in Sipser_d in the obvious way; so $A_d = \{\mathsf{output}\} \times [w_0] \times \cdots \times [w_{d-1}]$ is the set of addresses of the input variables and $|A_d| = n$.

We close this section by introducing notation for the following family of formulas related to Sipser_d:

Definition 5. *For $1 \le k \le d$, we write $\mathsf{Sipser}_d^{(k)} : \{0,1\}^{A_k} \to \{0,1\}$ to denote the depth-k formula obtained from Sipser_d by discarding all gates at depths $k+1$ through $d-1$, and replacing every depth-k gate at address $a \in A_k$ with a fresh formal variable y_a.*

Note that $\mathsf{Sipser}_d^{(1)}$ is the top gate of Sipser_d; in particular, $\mathsf{Sipser}_d^{(1)}$ is an w_0-way OR if d is even, and an w_0-way AND if d is odd. Note also that $\mathsf{Sipser}_d^{(d)}$ is simply Sipser_d itself, although we stress that $\mathsf{Sipser}_d^{(k)}$ is not the same as Sipser_k for $1 \le k \le d-1$.

7 Setup for and overview of our proof

7.1 Key parameter settings

The starting point for our parameter settings is the pair of fixed values

$$\lambda := \frac{(\log w)^{3/2}}{w^{5/4}} \quad \text{and} \quad q := \sqrt{p} = 2^{-m/2}. \tag{7}$$

Given these fixed values of λ and q, we define a sequence of parameters $t_{d-1}, \dots, t_1$ as

$$t_{d-1} := \frac{p - \lambda}{q}, \qquad t_{k-1} := \frac{(1 - t_k)^{qw} - \lambda}{q} \quad \text{for } k = d-1, \dots, 2. \tag{8}$$

Each of our $d-1$ random projections will be defined with respect to an underlying product distribution. Our first random projection $\mathrm{proj}_{\boldsymbol{\rho}^{(d)}}$ will be associated with the uniform distribution over $\{0,1\}^n$; this is because our ultimate goal is to establish uniform-distribution correlation bounds. For $k \in \{2, \dots, d-1\}$ the subsequent random projections $\mathrm{proj}_{\boldsymbol{\rho}^{(k)}}$ will be associated with either the t_k-biased or $(1-t_k)$-biased product distribution (depending on whether $d-k$ is even or odd). Recalling our discussion in Section 4 of the framework for proving correlation bounds — in particular, the three key properties our random projections have to satisfy — the values for $t_1, \dots, t_{d-1}$ are chosen carefully so that the compositions of our $d-1$ random projections complete to the uniform distribution, satisfying Property 3 (we prove this in Section 8).

The next lemma gives bounds on $t_{d-1}, \dots, t_1$ which show that these values "stay under control". By our definitions of λ, p and q in (7), we have that $t_{d-1} = q - o(q)$, and we will need the fact that the values of t_k for $k = d-1, \dots, 2$ remain in the range $q \pm o(q)$. Roughly speaking, since each t_{k-1} is defined inductively in terms of t_k from $k = d-1$ down to 1, we have to argue that these values do not "drift" significantly from the initial value of $t_{d-1} = q - o(q)$. We need to keep these values under control for two reasons: first, the magnitude of these values directly affects the strength of our Projection Switching Lemma — as we will see in Section 9.1, our error bounds depend on the magnitude of these t_k's. Second, since the top fan-in w_0 of our Sipser_d function is directly determined by t_1 (recall (4)), we need a bound on t_1 to control the structure of this function.

Lemma 7.1. *There is a universal constant $c > 0$ such that for $2 \le d \le \frac{cm}{\log m}$, we have that $t_k = q \pm q^{1.1}$ for all $k \in [d-1]$.*

We defer the proof of Lemma 7.1 to Appendix A. The $k = 1$ case of Lemma 7.1 along with our definition of w_0 (recall (4)) give us the bounds

$$\frac{1}{2} \ge (1 - t_1)^{qw_0} \ge \frac{1}{2}(1 - tq) = \frac{1}{2}\left(1 - \frac{\Theta(\log w)}{w}\right) = \frac{1}{2}\left(1 - \Theta(2^{-m})\right). \tag{9}$$

These bounds (showing that $(1-t_1)^{qw_0}$ is very close to 1/2) will be useful for our proof in Section 10.2 that Sipser_d remains essentially unbiased (i.e. it remains "structured") under our random projections, which in turn implies our claim (6) that Sipser_d is essentially balanced (see Remark 17).

We close this subsection with the following estimates of our key parameters in terms of w for later reference:

$$p = \Theta\left(\frac{\log w}{w}\right), \quad q = \Theta\left(\sqrt{\frac{\log w}{w}}\right), \quad t_k = \Theta\left(\sqrt{\frac{\log w}{w}}\right) \quad \text{for all } k \in [d-1]. \tag{10}$$

7.2 The initial and subsequent random projections

As described in Section 4, our overall approach is structured around a sequence of random projections which we will apply to both the target function Sipser_d and the approximating circuit C. Both are functions over $\{0,1\}^n \equiv \{0,1\}^{A_d}$, and our $d-1$ random projections will sequentially transform them from being over $\{0,1\}^{A_k}$ to being over $\{0,1\}^{A_{k-1}}$ for $k = d$ down to $k = 1$. Thus, at the end of the overall process both the randomly projected target and the randomly projected approximator are functions over $\{0,1\}^{A_1} \equiv \{0,1\}^{w_0}$.

We now formally define this sequence of random projections; recalling Definition 4, to define a random projection operator it suffices to specify a distribution over random restrictions, and this is what we will do. We begin with the initial random projection:

Definition 6 (Initial random projection)**.** *The distribution $\mathcal{R}_{\mathrm{init}}$ over restrictions ρ in $\{0,1,*\}^{A_{d-1}\times[m]} \equiv \{0,1,*\}^n$ (recall that $w_{d-1} = m$) is defined as follows: independently for each $a \in A_{d-1}$,*

$$\boldsymbol{\rho}_b \leftarrow \begin{cases} \{1\}^m & \textit{with probability } \lambda \\ \{*_{1/2}, 1_{1/2}\}^m \setminus \{1\}^m & \textit{with probability } q \\ \{0_{1/2}, 1_{1/2}\}^m \setminus \{1\}^m & \textit{with probability } 1-\lambda-q. \end{cases} \tag{11}$$

Remark 10. The description of $\mathcal{R}_{\mathrm{init}}$ given in Definition 6 will be most convenient for our arguments, but we note here the following equivalent view of an $\mathcal{R}_{\mathrm{init}}$-random projection. Let $\mathcal{R}'_{\mathrm{init}}$ be the distribution over restrictions ρ' in $\{0,1,*\}^{A_{d-1}\times[m]} \equiv \{0,1,*\}^n$ where

$$\boldsymbol{\rho}'_a \leftarrow \{*_{1/2}, 1_{1/2}\}^m \setminus \{1\}^m \quad \text{independently for each } a \in A_{d-1},$$

and $\mathcal{R}''_{\mathrm{init}}$ be the distribution of restrictions ρ'' in $\{0,1,*\}^{A_{d-1}}$ where

$$\boldsymbol{\rho}''_a \leftarrow \begin{cases} 1 & \text{with probability } \lambda \\ * & \text{with probability } q \\ 0 & \text{with probability } 1-\lambda-q \end{cases} \qquad \text{independently for each } a \in A_{d-1}.$$

Then for all $f : \{0,1\}^n \to \{0,1\}$ we have that $\mathrm{proj}_{\boldsymbol{\rho}} f$, where $\boldsymbol{\rho} \leftarrow \mathcal{R}_{\mathrm{init}}$, is distributed identically to

$$(\mathrm{proj}_{\boldsymbol{\rho}'} f) \upharpoonright \boldsymbol{\rho}'' \quad \text{where } \boldsymbol{\rho}' \leftarrow \mathcal{R}'_{\mathrm{init}} \text{ and } \boldsymbol{\rho}'' \leftarrow \mathcal{R}''_{\mathrm{init}}.$$

7.2.1 Subsequent random projections

Our subsequent random projections will alternate between two types, depending on whether $d-k$ is even or odd. These types are dual to each other in the sense that their distributions are completely identical, except with the roles of 1 and 0 swapped; in other words, the bitwise complement of a draw from the first type yields a draw from the second type. To avoid redundancy in our definitions we introduce the notation in Table 2: we represent $\{0,1\}^{A_k}$ as $\{\bullet,\circ\}^{A_k}$, where a $\circ$-value corresponds to either 1 or 0 depending on whether $d-k$ is even or odd, and the $\bullet$-value is simply the complement of the $\circ$-value. For example, the string $(\circ,\circ,\bullet,\circ)$ translates to $(1,1,0,1)$ if $d-k$ is even, and $(0,0,1,0)$ if $d-k$ is odd.

In an interesting contrast with Håstad's proofs of the worst-case depth hierarchy theorem (Theorem 4) and of $\mathsf{Parity} \notin \mathsf{AC^0}$, our stage-wise random projection process is *adaptive*: apart from the initial $\mathcal{R}_{\mathrm{init}}$-random projection, the distribution of each random projection depends on the outcome of the previous. We will need the following notion of the "lift" of a restriction to describe this dependence:

	Gates of Sipser_d at depth $k-1$	$\circ$	$\bullet$
$d-k \equiv 0 \mod 2$	AND	1	0
$d-k \equiv 1 \mod 2$	OR	0	1

Table 2: Conversion table for $\tau \in \{\bullet, \circ, *\}^{A_k}$ where $1 \le k \le d$.

Definition 7 (Lift)**.** *Let $2 \le k \le d$ and $\tau \in \{\bullet, \circ, *\}^{A_{k-1} \times [w_{k-1}]} \equiv \{\bullet, \circ, *\}^{A_k}$. The* lift *of τ is the string $\widehat{\tau} \in \{\bullet, \circ, *\}^{A_{k-1}}$ defined as follows: for each $a \in A_{k-1}$, the coordinate $\widehat{\tau}_a$ of $\widehat{\tau}$ is*

$$\widehat{\tau}_a = \begin{cases} \circ & \textit{if } \tau_{a,i} = \bullet \textit{ for any } i \in [w_{k-1}] \\ \bullet & \textit{if } \tau_a = \{\circ\}^{w_{k-1}} \\ * & \textit{if } \tau_a \in \{*, \circ\}^{w_{k-1}} \setminus \{\circ\}^{w_{k-1}}. \end{cases}$$

We remind the reader that $\tau \in \{\bullet, \circ, *\}^{A_k}$ and $\widehat{\tau} \in \{\bullet, \circ, *\}^{A_{k-1}}$ belong to adjacent levels (i.e. they fall under different rows in Table 2). Consequently, for example, if 1 corresponds to $\bullet$ as a symbol in τ then it corresponds to $\circ$ as a symbol in $\widehat{\tau}$, and vice versa.

Later this notion of the "lift" of a restriction will also be handy when we describe the effect of our random projections on the target function Sipser_d. The high-level rationale behind it is that $\widehat{\tau} \in \{\bullet, \circ, *\}^{A_{k-1}}$ denotes the values that the bottom-layer gates of $\mathsf{Sipser}_d^{(k)}$ take on when its input variables are set according to $\tau \in \{\bullet, \circ, *\}^{A_k}$. As a concrete example, suppose $d-k \equiv 0 \mod 2$ and let $\tau \in \{0, 1, *\}^{A_k}$ be a restriction. Since $d-k \equiv 0 \mod 2$, recalling Table 2 we have that the bottom-layer gates of $\mathsf{Sipser}_d^{(k)}$ (or equivalently, the gates of Sipser_d at depth $k-1$) are AND gates. For every block $a \in A_{k-1}$,

- If $\tau_{a,i} = 0$ for some $i \in [w_{k-1}]$, the AND gate at address a is falsified and has value 0.
- If $\tau_{a,i} = \{1\}^{w_{k-1}}$, the AND gate at address a is satisfied and has value 1.
- If $\tau_a \in \{*, 1\} \setminus \{1\}^{w_{k-1}}$, the value of the AND gate at address a remains undetermined (which we denote as having value $*$).

These three cases correspond exactly to the three branches in Definition 7, and so indeed $\widehat{\tau}_a \in \{0, 1, *\}$ represents the value that the AND gate at address a takes when its input variables are set according to $\tau_a \in \{0, 1, *\}^{w_{k-1}}$.

We shall require the following technical definition:

Definition 8 (k-acceptable)**.** *For $2 \le k \le d-1$ and a set $S \subseteq [w_{k-1}]$, we say that S is k-*acceptable *if*

$$|S| = qw \pm w^{\beta(k,d)}, \quad \textit{where } \beta(k,d) := \frac{1}{3} + \frac{d-k-1}{12d}.$$

Note that $\frac{1}{3} \le \beta(k,d) \le \frac{5}{12} < \frac{1}{2}$ for all $d \in \mathbb{N}$ and $2 \le k \le d-1$.

For intuition, in the above definition S should be thought of as specifying those children of a particular depth-$(k-1)$ gate of Sipser_d that take the value $*$ under certain restrictions (defined below). We want the size of this set to be essentially qw, and as k gets smaller (closer to the root), for technical reasons we allow more and more — but never too much — deviation from this desired value. See Section 10.1 for a detailed discussion.

We are now ready to give the key definition for our subsequent random projections:

Definition 9 (Subsequent random projections)**.** *Let $\tau \in \{\bullet, \circ, *\}^{A_k}$ where $2 \leq k \leq d-1$. We define a distribution $\mathcal{R}(\tau)$ over refinements $\boldsymbol{\rho} \in \{\bullet, \circ, *\}^{A_k}$ of τ as follows. Independently for each $a \in A_{k-1}$, writing $S_a = S_a(\tau)$ to denote $\tau_a^{-1}(*) = \{i \in [w_{k-1}] \colon \tau_{a,i} = *\}$ and $\boldsymbol{\rho}(S_a)$ to denote the substring of $\boldsymbol{\rho}_a$ with coordinates in S_a,*

- *If $\widehat{\tau}_a = \circ$ (i.e. if $\tau_{a,i} = \bullet$ for some $i \in [w_{k-1}]$) or if S_a is not k-acceptable, then*

$$\boldsymbol{\rho}(S_a) \leftarrow \{\bullet_{t_k}, \circ_{1-t_k}\}^{S_a}.$$

- *If $\widehat{\tau}_a = *$ (i.e. if $\tau_{a,i} \in \{*, \circ\}^{w_{k-1}} \setminus \{\circ\}^{w_{k-1}}$) and S_a is k-acceptable, then*

$$\boldsymbol{\rho}(S_a) \leftarrow \begin{cases} \circ^{S_a} & \text{with probability } \lambda \\ \{*_{t_k}, \circ_{1-t_k}\}^{S_a} \setminus \{\circ\}^{S_a} & \text{with probability } q_a \\ \{\bullet_{t_k}, \circ_{1-t_k}\}^{S_a} \setminus \{\circ\}^{S_a} & \text{with probability } 1-\lambda-q_a, \end{cases} \tag{12}$$

where

$$q_a := \frac{(1-t_k)^{|S_a|} - \lambda}{t_{k-1}} \quad \text{is chosen to satisfy } (1-t_k)^{|S_a|} = \lambda + q_a t_{k-1}. \tag{13}$$

(Note that if $\widehat{\tau}_a = \bullet$ then $\tau_{a,i} = \circ$ for all $i \in [w_{k-1}]$, and so τ_a cannot be refined further.)

For all $a \in A_{k-1}$ and $i \in [w_{k-1}]$ such that $\tau_{a,i} \in \{\bullet, \circ\}$, we set $\boldsymbol{\rho}_{a,i} = \tau_{a,i}$ and so $\boldsymbol{\rho}$ is indeed a refinement of τ.

Remark 11. We remark that q_a as defined in (13) is indeed a well-defined quantity in $[0,1]$ if S_a is k-acceptable. We omit the straightforward verification here since our analysis in Section 10.1 will in fact establish a stronger statement showing that $q_a = q \pm o(q)$; see Lemma 10.5.

Remark 12. By inspecting Definition 6, we see that for all $\rho \in \mathrm{supp}(\mathcal{R}_{\mathrm{init}})$ and blocks $a \in A_{d-1}$

$$\begin{aligned} &\rho_{a,i} = * \text{ for some } i \in [m] \quad \text{iff} \quad \rho_a \in \{*, 1\}^m \setminus \{1\}^m, \quad \text{or equivalently,} \\ &\rho_{a,i} = * \text{ for some } i \in [m] \quad \text{iff} \quad \widehat{\rho}_a = *, \end{aligned}$$

and hence for all $h : \{0,1\}^n \to \{0,1\}$ the projection $\mathrm{proj}_\rho h : \{0,1\}^{A_{d-1}} \to \{0,1\}$ depends only on the coordinates in $(\widehat{\rho})^{-1}(*) \subseteq A_{d-1}$. Likewise, by inspecting Definition 9 we have that for all $\tau \in \{\bullet, \circ, *\}^{A_k}, \rho \in \mathrm{supp}(\mathcal{R}(\tau))$, and blocks $a \in A_{k-1}$,

$$\begin{aligned} &\rho_{a,i} = * \text{ for some } i \in [w_{k-1}] \quad \text{iff} \quad \rho_a \in \{*, \circ\}^{w_{k-1}} \setminus \{\circ\}^{w_{k-1}}, \quad \text{or equivalently,} \\ &\rho_{a,i} = * \text{ for some } i \in [w_{k-1}] \quad \text{iff} \quad \widehat{\rho}_a = *, \end{aligned}$$

and hence for all $h : \{0,1\}^{A_k} \to \{0,1\}$ the projection $\mathrm{proj}_\rho h : \{0,1\}^{A_{k-1}} \to \{0,1\}$ depends only on the coordinates in $(\widehat{\rho})^{-1}(*) \subseteq A_{k-1}$. Our proof that our sequence of random projections (based on Definitions 6 and 9 as described in Definition 4) completes to the uniform distribution will rely on these properties; see Section 8.

7.3 Overview of our proof

With the definitions from Section 7.2 in hand, we are (finally) in a position to give a detailed overview of our proof. Let C be a depth-d approximating circuit for Sipser_d, where C either has significantly smaller bottom fan-in than Sipser_d (in the case of Theorem 6) or the opposite alternation pattern to Sipser_d (in the case of Theorem 7), and C satisfies the size bounds given in the respective theorem statements. In both cases our goal is to show that C has small correlation with Sipser_d, i.e. to prove that

$$\mathbf{Pr}[\mathsf{Sipser}_d(\mathbf{X}) \neq C(\mathbf{X})] \geq \frac{1}{2} - o_n(1) \tag{14}$$

for a uniform random input $\mathbf{X} \leftarrow \{0_{1/2}, 1_{1/2}\}^n$. At a high level, we do this by analyzing the effect of $d-1$ random projections on the target and the approximator: we begin with an $\mathcal{R}_{\text{init}}$-random projection $\text{proj}_{\boldsymbol{\rho}^{(d)}}$ where $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\text{init}}$, followed by $\text{proj}_{\boldsymbol{\rho}^{(d-1)}}$ where $\boldsymbol{\rho}^{(d-1)} \leftarrow \mathcal{R}(\widehat{\boldsymbol{\rho}^{(d)}})$, and then $\text{proj}_{\boldsymbol{\rho}^{(d-2)}}$ where $\boldsymbol{\rho}^{(d-2)} \leftarrow \mathcal{R}(\widehat{\boldsymbol{\rho}^{(d-1)}})$, and so on. It is interesting to note that unlike Håstad's proofs of the worst-case depth hierarchy theorem (Theorem 4) and of $\mathsf{Parity} \notin \mathsf{AC}^0$, the distribution of our k-th random projection is defined *adaptively* depending on the outcome of the $(k-1)$-st. For notational concision we introduce the following definition for this overall $(d-1)$-stage projection:

Definition 10. *Given a function $f : \{0,1\}^n \to \{0,1\}$, we write $\mathbf{\Psi}(f) : \{0,1\}^{w_0} \to \{0,1\}$ to denote the following random projection of f:*

$$\mathbf{\Psi}(f) \equiv \text{proj}_{\boldsymbol{\rho}^{(2)}} \text{proj}_{\boldsymbol{\rho}^{(3)}} \cdots \text{proj}_{\boldsymbol{\rho}^{(d-1)}} \text{proj}_{\boldsymbol{\rho}^{(d)}} f,$$

where $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\text{init}}$ and $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}(\widehat{\boldsymbol{\rho}^{(k+1)}})$ for all $2 \leq k \leq d-1$. We will sometimes refer to the overall process as a $\mathbf{\Psi}$-random projection, and $\mathbf{\Psi}(f)$ as the $\mathbf{\Psi}$-random projection of f. (We remind the reader that the projection of a function over $\{0,1\}^{A_k}$ yields a function over $\{0,1\}^{A_{k-1}}$ for all $2 \leq k \leq d$, and in particular $\mathbf{\Psi}(f)$ is indeed a function over $\{0,1\}^{A_1} \equiv \{0,1\}^{w_0}$.)

Recalling the framework for proving correlation bounds discussed in Section 4, the rest of the paper is structured around showing that a $\mathbf{\Psi}$-random projection satisfies the three key properties outlined in Section 4:

Property 1. The approximating circuit C simplifies under a $\mathbf{\Psi}$-random projection.

Property 2. The target Sipser_d remains structured under a $\mathbf{\Psi}$-random projection.

Property 3. $\mathbf{\Psi}$ completes to the uniform distribution.

Section 8. We begin in Section 8 with Property 3. We show that

$$\mathbf{Pr}[\mathsf{Sipser}_d(\mathbf{X}) \neq C(\mathbf{X})] = \mathbf{Pr}[(\mathbf{\Psi}(\mathsf{Sipser}_d))(\mathbf{Y}) \neq (\mathbf{\Psi}(C))(\mathbf{Y})] \tag{15}$$

where $\mathbf{Y}$ is drawn from an appropriate product distribution $\mathcal{D}$ over $\{0,1\}^{w_0}$ ($\mathcal{D}$ is the t_1-biased product distribution if d is even, and $(1-t_1)$-biased product distribution if d is odd). This reduces our goal of bounding the correlation between Sipser_d and C (i.e. (14)) under the uniform distribution, to the task of bounding the correlation between their $\mathbf{\Psi}$-random projections $\mathbf{\Psi}(\mathsf{Sipser}_d)$ and $\mathbf{\Psi}(C)$ with respect to $\mathcal{D}$.

Section 9. With the reduction (15) in hand, we turn our attention to Property 1, showing that the approximating circuit C of the type specified in either Theorems 6 or 7 "collapses to a simple function" under a $\boldsymbol{\Psi}$-random projection. More precisely, for the case that the depth-d circuit C has significantly smaller bottom fan-in than Sipser_d we show that C collapses to a shallow decision tree, and for the case that C has the opposite alternation pattern to Sipser_d we show that C collapses to a small-width depth-two circuit with top gate opposite to that of $\boldsymbol{\Psi}(\mathsf{Sipser}_d)$. (In both cases these statements are with high probability under a $\boldsymbol{\Psi}$-random projection.)

In close parallel with Håstad's "bottom-up" proof of $\mathsf{Parity} \notin \mathsf{AC}^0$, the main technical ingredient in this section is a *projection switching lemma* showing that the random projection $\mathrm{proj}_{\boldsymbol{\rho}^{(k)}}$ of a small-width DNF or CNF "switches" to a small-depth decision tree with high probability. Applying this lemma to every bottom-level depth-2 subcircuit of C, we are able to argue that each of the $d-1$ random projections comprising $\boldsymbol{\Psi}$ reduces the depth of C by one with high probability, and thus $\boldsymbol{\Psi}(C)$ collapses to a small-depth decision tree or small-width depth-two circuit as claimed.

Section 10. It remains to argue that the target Sipser_d — in contrast with the approximating circuit C — "retains structure" with high probability under a $\boldsymbol{\Psi}$-random restriction. This is a high-probability statement because there is a nonzero failure probability introduced by each of the $d-1$ individual random projections $\mathrm{proj}_{\boldsymbol{\rho}^{(k)}}$ that comprise $\boldsymbol{\Psi} \equiv \{\boldsymbol{\rho}^{(k)}\}_{k\in\{2,\ldots,d\}}$ (see Footnote 3 for an example of a possible "failure event" for one of these restrictions). To reason about and bound these failure probabilities, in Section 10.1 we introduce the notion of a "typical" restriction. The parameters of our definition of typicality are chosen carefully to ensure that

(i) $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\mathrm{init}}$ is typical with high probability, and

(ii) if $\rho^{(k+1)}$ is typical, then $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}(\widehat{\rho^{(k+1)}})$ is also typical with high probability.

We establish (i) and (ii) in Section 10.1. Together, (i) and (ii) imply that with high probability $\boldsymbol{\Psi} \equiv \{\boldsymbol{\rho}^{(k)}\}_{k\in\{2,\ldots,d\}}$ is such that $\boldsymbol{\rho}^{(d)},\ldots,\boldsymbol{\rho}^{(2)}$ are *all* typical; we use this in Section 10.2.

With the notion of typical restrictions in hand, in Section 10.2 we establish Property 2 showing that Sipser_d "survives" a $\boldsymbol{\Psi}$-random projection (i.e. it "retains structure") with high probability. More formally, for outcomes $\Psi \equiv \{\rho^{(k)}\}_{k\in\{2,\ldots,d\}}$ of $\boldsymbol{\Psi}$ such that $\rho^{(d)},\ldots,\rho^{(2)}$ are all typical, we prove that the Ψ-projected target $\Psi(\mathsf{Sipser}_d)$ is "well-structured" in the following sense:

(i) $\Psi(\mathsf{Sipser}_d)$ is a depth-one formula: an OR if d is even, an AND if d is odd.

(ii) The bias of $\Psi(\mathsf{Sipser}_d)$ under $\mathcal{D}$ is close to $1/2$; that is,

$$\mathrm{bias}(\Psi(\mathsf{Sipser}_d), \mathbf{Y}) = \frac{1}{2} - o_n(1).$$

Recall that we have shown in Subsection 10.1 that with high probability $\boldsymbol{\Psi} \equiv \{\boldsymbol{\rho}^{(k)}\}_{k\in\{2,\ldots,d\}}$ is such that $\boldsymbol{\rho}^{(d)},\ldots,\boldsymbol{\rho}^{(2)}$ are all typical. Therefore, the results of these two subsections together imply that the randomly projected target $\boldsymbol{\Psi}(\mathsf{Sipser}_d)$ satisfies both (i) and (ii) with high probability.

Section 11. Having established Properties 1, 2, and 3, it remains to bound the correlation between a depth-one formula with bias essentially $1/2$ and a small-width CNF formula of opposite alternation with respect to the product distribution $\mathcal{D}$ over $\{0,1\}^{w_0}$. (Recall that our results from Section 10.2 show that $\boldsymbol{\Psi}(\mathsf{Sipser}_d)$ collapses to the former with high probability, and our results from Section 9 shows that $\mathbf{\Psi}(C)$ collapses to the latter with high probability — this holds in both cases since a shallow decision tree is a small-width CNF.) We prove this correlation bound using a slight extension of an argument in [OW07], and with this final piece in hand our main theorems follow from straightforward arguments putting the pieces together.

8 Composition of projections complete to uniform

Our goal in this section is to establish the following lemma:

Proposition 8.1. *Consider $f, g : \{0,1\}^n \to \{0,1\}$. Let $\mathbf{X} \leftarrow \{0_{1/2}, 1_{1/2}\}^n$. Let $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$ if d is even, and $\mathbf{Y} \leftarrow \{0_{t_1}, 1_{1-t_1}\}^{w_0}$ if d is odd. Then*

$$\mathbf{Pr}[f(\mathbf{X}) \neq g(\mathbf{X})] = \mathbf{Pr}[(\mathbf{\Psi}(f))(\mathbf{Y}) \neq (\mathbf{\Psi}(g))(\mathbf{Y})].$$

As discussed in Section 7.3 we will ultimately apply Proposition 8.1 with f being our target function Sipser_d and g being the approximating circuit C. This allows us to translate the inapproximability of $\mathbf{\Psi}(\mathsf{Sipser}_d)$ by $\mathbf{\Psi}(C)$ (either with respect to the t_1-biased or $(1-t_1)$-based product distribution, depending on whether d is even or odd) into the uniform-distribution inapproximability of Sipser_d by C.

Overview of proof. We will actually derive Proposition 8.1 as a consequence of a stronger claim, which, roughly speaking, states that we can generate a uniformly random input $\mathbf{X} \leftarrow \{0_{1/2}, 1_{1/2}\}^n$ via $\mathbf{\Psi}$ and $\mathbf{Y}$ in a stage-wise manner. In more detail, given $\mathbf{\Psi} \equiv \{\boldsymbol{\rho}^{(k)}\}_{k\in\{2,\ldots,d\}}$ and $\mathbf{Y}$ we consider the following random $\{0,1,*\}$-valued labeling $\boldsymbol{\ell}$ of the leaves and non-root nodes of the depth-d depth-regular tree corresponding to the depth-d formula computing Sipser_d:

- The $|A_d| = n$ leaves of the tree are each labeled $\{0,1,*\}$ according to $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\text{init}}$.
- For $2 \leq k \leq d-1$, the $|A_k|$ nodes at depth k are each labeled $\{0,1,*\}$ according to $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}(\widehat{\boldsymbol{\rho}^{(k+1)}})$.
- Finally, for each $i \in [w_0] = [|A_1|]$, if $\widehat{\boldsymbol{\rho}^{(2)}}_i = *$ then the i-th node at depth 1 is labeled $\mathbf{Y}_i \in \{0,1\}$, and otherwise it is labeled $\widehat{\boldsymbol{\rho}^{(2)}}_i \in \{0,1\}$. (The root of the tree is left unlabeled.)

Next, we let the $\{0,1\}$-valued labels of $\boldsymbol{\ell}$ "percolate down the tree" as follows: every node or leaf that is labeled $*$ by $\boldsymbol{\ell}$ inherits the ($\{0,1\}$-valued) label from its closest ancestor that is not labeled $*$. Note that this "percolation step" ensures that every leaf and non-root node of the tree is labeled either 0 or 1, since every depth-1 node is assigned a $\{0,1\}$-valued label by $\boldsymbol{\ell}$.

Let $\boldsymbol{\ell}^{\downarrow}$ denote this $\{0,1\}$-valued random labeling of the leaves and non-root nodes. Our main result in this section, Proposition 8.4, can be viewed as stating that the random string $\mathbf{X} \in \{0,1\}^n$ defined by $\boldsymbol{\ell}^{\downarrow}$'s labeling of the n leaves is distributed uniformly at random; Proposition 8.1 follows as a straightforward consequence of this claim along with our definition of projections.

We begin with the following lemma, which explains our choice of t_{d-1} in (8) in the definition of $\mathcal{R}_{\text{init}}$ (Definition 6). (Note that in the lemma each coordinate of $\mathbf{Y}$ is distributed as $\{0_{1-t_{d-1}}, 1_{t_{d-1}}\}$ regardless of whether d is even or odd; this is because of our convention that the bottom-layer gates of Sipser_d are always AND gates.)

Lemma 8.2. *Let $\boldsymbol{\rho} \leftarrow \mathcal{R}_{\text{init}}$ and $\mathbf{Y} \leftarrow \{0_{1-t_{d-1}}, 1_{t_{d-1}}\}^{(\widehat{\boldsymbol{\rho}})^{-1}(*)}$, and consider the string $\mathbf{X} \in \{0,1\}^n \equiv \{0,1\}^{A_{d-1}\times[m]}$ defined as follows:*

$$\mathbf{X}_{a,i} = \begin{cases} \mathbf{Y}_a & \textit{if } \boldsymbol{\rho}_{a,i} = * \\ \boldsymbol{\rho}_{a,i} & \textit{otherwise} \end{cases} \quad \textit{for all } a \in A_{d-1} \textit{ and } i \in [m].$$

The string $\mathbf{X}$ is distributed according to the uniform distribution $\{0_{1/2}, 1_{1/2}\}^n$. (Recalling Remark 12 we have that $\boldsymbol{\rho}_{a,i} = *$ if and only if $\widehat{\boldsymbol{\rho}}_a = *$, and so $\mathbf{Y}_a$ in the equation above is indeed well-defined.)

Proof. Since the blocks of $\boldsymbol{\rho}$ are independent across $a \in A_{d-1}$ and the coordinates of $\mathbf{Y}$ are independent across $a \in (\widehat{\boldsymbol{\rho}})^{-1}(*) \subseteq A_{d-1}$, it suffices to prove that $\mathbf{X}_a$ is distributed according to $\{0_{1/2}, 1_{1/2}\}^m$ for a fixed $a \in A_{d-1}$. We first observe that

$$\mathbf{Pr}[\mathbf{X}_a = 1^m] = \lambda + q\,\mathbf{Pr}[\mathbf{Y}_a = 1] = \lambda + qt_{d-1} = p = 2^{-m},$$

where the λ is from the first line of (11), the $q\,\mathbf{Pr}[\mathbf{Y}_a = 1]$ is from the second line of (11), and the penultimate equality is by our choice of t_{d-1} in (8). Next, for any string $\mathrm{Z} \in \{0,1\}^m \setminus \{1\}^m$, we have that

$$\begin{aligned} \mathbf{Pr}[\mathbf{X}_a = \mathrm{Z}] &= (1-\lambda-q)\cdot\frac{2^{-m}}{1-2^{-m}} + q\,\mathbf{Pr}[\mathbf{Y}_a = 0]\cdot\frac{2^{-m}}{1-2^{-m}} \qquad (16) \\ &= \frac{p}{1-p}\cdot((1-\lambda-q) + q(1-t_{d-1})) \\ &= \frac{p}{1-p}\cdot(1-\lambda-qt_{d-1}) = \frac{p}{1-p}\cdot(1-p) = p = 2^{-m}, \qquad (17) \end{aligned}$$

where the first summand on the RHS of (16) is by the third line of (11), the second summand is by the second line of (11), and (17) again uses our choice of t_{d-1} in (8). Since this is exactly the probability mass function of the uniform distribution $\{0_{1/2}, 1_{1/2}\}^m$, the proof is complete. □

The following lemma, the analogue of Lemma 8.2 for $\mathcal{R}(\tau)$, explains our choice of q_a in terms of t_k and t_{k-1} in (13):

Lemma 8.3. *For $2 \leq k \leq d-1$ let $\tau \in \{0,1,*\}^{A_k}$, $\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)$, and*

$$\begin{cases} \mathbf{Y} \leftarrow \{0_{1-t_{k-1}}, 1_{t_{k-1}}\}^{(\widehat{\boldsymbol{\rho}})^{-1}(*)} & \textit{if } d-k \equiv 0 \mod 2 \\ \mathbf{Y} \leftarrow \{0_{t_{k-1}}, 1_{1-t_{k-1}}\}^{(\widehat{\boldsymbol{\rho}})^{-1}(*)} & \textit{if } d-k \equiv 1 \mod 2. \end{cases}$$

For each $a \in A_{k-1}$, writing $S_a = S_a(\tau)$ to denote $\tau_a^{-1}(*) = \{i \in [w_{k-1}] \colon \tau_{a,i} = *\}$ and $\boldsymbol{\rho}(S_a)$ to denote the substring of $\boldsymbol{\rho}_a$ with coordinates in S_a, we consider the string $\mathbf{Z}_a \in \{0,1\}^{S_a}$ defined as follows:

$$\mathbf{Z}_{a,i} = \begin{cases} \mathbf{Y}_a & \textit{if } \boldsymbol{\rho}_{a,i} = * \\ \boldsymbol{\rho}_{a,i} & \textit{otherwise} \end{cases} \quad \textit{for all } i \in S_a.$$

The string $\mathbf{Z}_a$ is distributed according to

$$\begin{cases} \{0_{t_k}, 1_{1-t_k}\}^{S_a} & \textit{if } d-k \equiv 0 \mod 2 \\ \{0_{1-t_k}, 1_{t_k}\}^{S_a} & \textit{if } d-k \equiv 1 \mod 2, \end{cases}$$

and furthermore, $\mathbf{Z}_a$ and $\mathbf{Z}_{a'}$ are independent for any two distinct $a, a' \in A_{k-1}$. (Again, recalling Remark 12 we have that $\boldsymbol{\rho}_{a,i} = *$ if and only if $\widehat{\boldsymbol{\rho}}_a = *$, and so $\mathbf{Y}_a$ in the equation above is indeed well-defined.)

Proof. We prove the $d - k \equiv 0 \mod 2$ case (the other case follows by a symmetric argument). If $\widehat{\tau}_a$ falls in the first case of Definition 9 (i.e. if $\widehat{\tau}_a = 0$ or if S_a is not k-acceptable) then the claim is true since $\mathbf{Z}_a \equiv \boldsymbol{\rho}(S_a) \leftarrow \{0_{t_k}, 1_{1-t_k}\}^{S_a}$. Otherwise, if $\widehat{\tau}_a$ falls in the second case of Definition 9 (i.e. if $\widehat{\tau}_a = *$ and S_a is k-acceptable) we first observe that

$$\begin{aligned}\mathbf{Pr}[\mathbf{Z}_a = 1^{S_a}] &= \lambda + q_a \, \mathbf{Pr}[\mathbf{Y}_a = 1] \\ &= \lambda + q_a t_{k-1} \\ &= (1 - t_k)^{|S_a|},\end{aligned}$$

where as before the λ is from the first line of (12), the $q_a \, \mathbf{Pr}[\mathbf{Y}_a = 1]$ is from the second line of (12), and the final equality is by our definition of q_a in (13). Next, for any string $\mathrm{Z} \in \{0,1\}^{S_a} \setminus \{1\}^{S_a}$ and $u := |\mathrm{Z}^{-1}(0)| \in \{1, \ldots, |S_a|\}$, we have that

$$\begin{aligned}\mathbf{Pr}[\mathbf{Z}_a = \mathrm{Z}] &= (1 - \lambda - q_a) \cdot \frac{t_k^u (1-t_k)^{|S_a|-u}}{1 - (1-t_k)^{|S_a|}} + q_a \, \mathbf{Pr}[\mathbf{Y}_a = 0] \cdot \frac{t_k^u (1-t_k)^{|S_a|-u}}{1 - (1-t_k)^{|S_a|}} \qquad (18) \\ &= \frac{t_k^u (1-t_k)^{|S_a|-u}}{1 - (1-t_k)^{|S_a|}} \cdot (1 - \lambda - q_a + q_a(1 - t_{k-1})) \\ &= \frac{t_k^u (1-t_k)^{|S_a|-u}}{1 - (1-t_k)^{|S_a|}} \cdot (1 - \lambda - q_a t_{k-1}) \\ &= \frac{t_k^u (1-t_k)^{|S_a|-u}}{1 - (1-t_k)^{|S_a|}} \cdot \left(1 - (1-t_k)^{|S_a|}\right) = t_k^u (1-t_k)^{|S_a|-u}, \qquad (19)\end{aligned}$$

where as before the first summand on the RHS of (18) is by the third line of (12), the second summand is by the second line of (12), and (19) again uses our definition of q_a. Therefore indeed, the resulting string is distributed according to $\{0_{t_k}, 1_{1-t_k}\}^{S_a}$. Finally, since the blocks of $\boldsymbol{\rho}$ are independent across $a \in A_{k-1}$ and the coordinates of $\mathbf{Y}$ are independent across $a \in (\widehat{\boldsymbol{\rho}})^{-1}(*) \subseteq A_{k-1}$, we have that $\mathbf{Z}_a$ and $\mathbf{Z}_{a'}$ are independent for any two distinct $a, a' \in A_{k-1}$. □

Together Lemmas 8.2 and 8.3 give us the following proposition, which in turn yields Proposition 8.1, our main result in this section.

Proposition 8.4. *Let $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\mathrm{init}}$ and $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}(\widehat{\boldsymbol{\rho}^{(k+1)}})$ for $2 \leq k \leq d-1$. Let*

$$\mathbf{Y}^{(1)} \leftarrow \begin{cases} \{0_{1-t_1}, 1_{t_1}\}^{(\widehat{\boldsymbol{\rho}^{(2)}})^{-1}(*)} & \textit{if } d \textit{ is even} \\ \{0_{t_1}, 1_{1-t_1}\}^{(\widehat{\boldsymbol{\rho}^{(2)}})^{-1}(*)} & \textit{if } d \textit{ is odd,} \end{cases}$$

and for $2 \leq k \leq d-1$ consider random strings $\mathbf{Y}^{(k)} \in \{0,1\}^{(\widehat{\boldsymbol{\rho}^{(k+1)}})^{-1}(*)}$ defined inductively from $k = 2$ up to $d-1$ as follows:

$$\mathbf{Y}^{(k)}_{a,i} = \begin{cases} \mathbf{Y}^{(k-1)}_a & \textit{if } \boldsymbol{\rho}^{(k)}_{a,i} = * \\ \boldsymbol{\rho}^{(k)}_{a,i} & \textit{otherwise} \end{cases} \quad \textit{for all } a \in A_{k-1} \textit{ and } i \in [w_{k-1}] \textit{ s.t. } \widehat{\boldsymbol{\rho}^{(k+1)}}_{a,i} = *. \qquad (20)$$

Then the string $\mathbf{X} \in \{0,1\}^n \equiv \{0,1\}^{A_{d-1} \times [m]}$ defined by

$$\mathbf{X}_{a,i} = \begin{cases} \mathbf{Y}^{(d-1)}_a & \textit{if } \boldsymbol{\rho}_{a,i} = * \\ \boldsymbol{\rho}^{(d)}_{a,i} & \textit{otherwise} \end{cases} \quad \textit{for all } a \in A_{d-1} \textit{ and } i \in [m]$$

is distributed according to the uniform distribution $\{0_{1/2}, 1_{1/2}\}^n$.

Proof. By the $k = 2$ case of Lemma 8.3, for all possible outcomes $\rho^{(k)}$ of $\boldsymbol{\rho}^{(k)}$ for $3 \leq k \leq d$, conditioned on such an outcome the random string $\mathbf{Y}^{(2)}$ is distributed according to $\{0_{t_2}, 1_{1-t_2}\}^{\widehat{\rho^{(3)}}}$ if d is even and according to $\{0_{1-t_2}, 1_{t_2}\}^{\widehat{\rho^{(3)}}}$ if d is odd. Applying this argument repeatedly and arguing inductively from $k = 2$ up to $k = d-1$, we have that conditioned on any outcome $\rho^{(d)}$ of $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\text{init}}$, the random string $\mathbf{Y}^{(d-1)}$ is distributed according to $\{0_{1-t_{d-1}}, 1_{t_{d-1}}\}^{\widehat{\rho^{(d)}}}$. The claim then follows by Lemma 8.2. □

Proof of Proposition 8.1. Recall that $\mathbf{X} \leftarrow \{0_{1/2}, 1_{1/2}\}^n$ and $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$ if d is even, $\mathbf{Y} \leftarrow \{0_{t_1}, 1_{1-t_1}\}^{w_0}$ if d is odd. Let $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\text{init}}$ and $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}(\widehat{\boldsymbol{\rho}^{(k+1)}})$ for $2 \leq k \leq d-1$. For $1 \leq k \leq d-1$ let $\mathbf{Y}^{(k)} \in \{0,1\}^{(\widehat{\boldsymbol{\rho}^{(k+1)}})^{-1}(*)}$ be defined as in Proposition 8.4. Recalling Remark 12, for all functions $h : \{0,1\}^n \to \{0,1\}$ and $1 \leq k \leq d-1$, the random projection

$$(\mathrm{proj}_{\boldsymbol{\rho}^{(k+1)}} \cdots \mathrm{proj}_{\boldsymbol{\rho}^{(d)}} h) : \{0,1\}^{A_k} \to \{0,1\}$$

depends only on the coordinates in $(\widehat{\boldsymbol{\rho}^{(k+1)}})^{-1}(*) \subseteq A_k$, and so we may equivalently view it as a function $\{0,1\}^{(\widehat{\boldsymbol{\rho}^{(k+1)}})^{-1}(*)} \to \{0,1\}$. By Proposition 8.4, the definition of the $\mathbf{Y}^{(k)}$'s, and the definition of projections, we see that

$$\begin{aligned}
\mathbf{Pr}[f(\mathbf{X}) \neq g(\mathbf{X})] &= \mathbf{Pr}[(\mathrm{proj}_{\boldsymbol{\rho}^{(d)}} f)(\mathbf{Y}^{(d-1)}) \neq (\mathrm{proj}_{\boldsymbol{\rho}^{(d)}} g)(\mathbf{Y}^{(d-1)})] \\
&= \mathbf{Pr}[(\mathrm{proj}_{\boldsymbol{\rho}^{(d-1)}} \mathrm{proj}_{\boldsymbol{\rho}^{(d)}} f)(\mathbf{Y}^{(d-2)}) \neq (\mathrm{proj}_{\boldsymbol{\rho}^{(d-1)}} \mathrm{proj}_{\boldsymbol{\rho}^{(d)}} g)(\mathbf{Y}^{(d-2)})] \\
&= \cdots \\
&= \mathbf{Pr}[(\mathrm{proj}_{\boldsymbol{\rho}^{(2)}} \cdots \mathrm{proj}_{\boldsymbol{\rho}^{(d)}} f)(\mathbf{Y}^{(1)}) \neq (\mathrm{proj}_{\boldsymbol{\rho}^{(2)}} \cdots \mathrm{proj}_{\boldsymbol{\rho}^{(d)}} g)(\mathbf{Y}^{(1)})] \\
&= \mathbf{Pr}[(\mathrm{proj}_{\boldsymbol{\rho}^{(2)}} \cdots \mathrm{proj}_{\boldsymbol{\rho}^{(d)}} f)(\mathbf{Y}) \neq (\mathrm{proj}_{\boldsymbol{\rho}^{(2)}} \cdots \mathrm{proj}_{\boldsymbol{\rho}^{(d)}} g)(\mathbf{Y})] \\
&= \mathbf{Pr}[(\boldsymbol{\Psi}(f))(\mathbf{Y}) \neq (\boldsymbol{\Psi}(g))(\mathbf{Y})]
\end{aligned}$$

where the final inequality is by the definition of $\boldsymbol{\Psi}$ (Definition 10). □

9 Approximator simplifies under random projections

With Proposition 8.1 in hand we next prove that the approximating circuit C of the type specified in either Theorems 6 or 7 "collapses to a simple function" with high probability under a $\boldsymbol{\Psi}$-random restriction. For the case that the depth-d circuit C has significantly smaller bottom fan-in than Sipser_d we show that C collapses to a shallow decision tree with high probability, and for the case that C has the opposite alternation pattern to Sipser_d we show that C collapses to a small-width depth-two circuit with top gate opposite to that of $\boldsymbol{\Psi}(\mathsf{Sipser}_d)$ with high probability.

We do so via a *projection switching lemma*, showing that each of the $d-1$ individual random projections $\mathrm{proj}_{\boldsymbol{\rho}^{(k)}}$ comprising $\boldsymbol{\Psi}$ "contribute to the simplification" of C with high probability. We state and prove our projection switching lemma in Sections 9.1 through 9.5, and in Section 9.6 we show how the lemma can be applied iteratively to prove our structural claims about $\boldsymbol{\Psi}(C)$.

9.1 The projection switching lemma and its proof

Proposition 9.1 (Projection switching lemma for $\mathcal{R}_{\text{init}}$)**.** *Let $F : \{0,1\}^n \to \{0,1\}$ be a depth-2 circuit with bottom fan-in r. Then for all $s \geq 1$,*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}_{\text{init}}}[\text{proj}_{\boldsymbol{\rho}} F \textit{ is not a depth-}s\textit{ decision tree}] = \Big(O\Big(r2^r \cdot w^{-1/4}\Big)\Big)^s.$$

Proposition 9.2 (Projection switching lemma for $\mathcal{R}(\tau)$)**.** *Let $2 \leq k \leq d-1$ and $F : \{0,1\}^{A_k} \to \{0,1\}$ be a depth-2 circuit with bottom fan-in r. Then for all $\tau \in \{0,1,*\}^{A_k}$ and $s \geq 1$,*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)}[\text{proj}_{\boldsymbol{\rho}} F \textit{ is not a depth-}s\textit{ decision tree}] = \Big(O\Big(re^{rt_k/(1-t_k)} \cdot w^{-1/4}\Big)\Big)^s.$$

The proofs of Propositions 9.1 and 9.2 have the same overall structure, and they share many of the same ingredients. We will only prove (the slightly more involved) Proposition 9.2, and at the end of this section we point out the essential differences in the proof of Proposition 9.1.

Furthermore, we will prove Proposition 9.2 assuming that F is a DNF and $d - k \equiv 0 \mod 2$. Both assumptions are without loss of generality. (For the first, we recall that F is a width-r DNF if and only if its Boolean dual $F^\dagger$ is a width-r CNF, and that a Boolean function is computed by a depth-s decision tree if and only if its Boolean dual is as well, and we observe that $(\text{proj}_\rho F)^\dagger = \text{proj}_\rho (F^\dagger)$ for all ρ and all F. For the second we note that the definition of $\mathcal{R}(\tau)$ when $d - k \equiv 0 \mod 2$ is dual to that of $\mathcal{R}(\tau)$ when $d - k \equiv 1 \mod 2$, and so applying the former to $F(x)$ is equivalent to applying the latter to $F(\overline{x})$.)

Overview of proof. At a high level, we adopt Razborov's strategy in his alternative proof [Raz95] of Håstad's Switching Lemma. We briefly recall the overall structure of Razborov's argument. Given a DNF $F : \{0,1\}^n \to \{0,1\}$ and a distribution $\mathcal{R}$ over restrictions in $\{0,1,*\}^n$, we let $\mathcal{B} \subseteq \{0,1,*\}^n$ denote the set of all *bad* restrictions, namely the ones such that $F \upharpoonright \rho$ is not computed by a small-depth decision tree. Our goal in a switching lemma is to bound $\mathbf{Pr}_{\boldsymbol{\rho} \leftarrow \mathcal{R}}[\boldsymbol{\rho} \in \mathcal{B}]$, the weight of $\mathcal{B}$ under $\mathcal{R}$. To do so, we define an *encoding* of each bad restriction $\rho \in \mathcal{B}$ as a different restriction $\rho' \in \{0,1,*\}^n$ and a small amount (say at most ℓ bits) of "auxiliary information":

$$\text{encode} : \mathcal{B} \to \{0,1,*\}^n \times \{0,1\}^\ell$$
$$\text{encode}(\rho) = (\rho', \text{auxiliary information}).$$

This encoding should satisfy two key properties. First, it should be uniquely decodable, meaning that one is always able to recover ρ given ρ' and the auxiliary information; equivalently, the function encode$(\cdot)$ is an injection. Second, the weight $\mathbf{Pr}_{\boldsymbol{\rho} \leftarrow \mathcal{R}}[\boldsymbol{\rho} = \rho']$ of ρ' under $\mathcal{R}$ should be larger than that of ρ by a significant multiplicative factor (say by a factor of Γ). It is not hard to see that together, these two properties imply that total weight of all bad restrictions with the *same* auxiliary information is at most $1/\Gamma$. To complete the proof of the switching lemma, we then bound the overall weight of $\mathcal{B}$ via a union bound over all 2^ℓ possible strings of auxiliary information. (For a detailed exposition of Razborov's proof technique see [Bea94, Tha09] and Chapter §14 of [AB09].)

The proof of our projection switching lemma follows this high-level strategy quite closely; specifically, we build off of a reformulation (due to Thapen [Tha09]) of Håstad's proof of the blockwise variant of his Switching Lemma in Razborov's framework. In Section 9.3 we define our encoding, specifying the restriction ρ' and auxiliary information that is associated with every bad restriction

ρ; in Section 9.4 we prove that our encoding is an injection by describing a procedure for unique decoding; in Section 9.5 we verify that every bad ρ is indeed paired with a ρ' whose weight under $\mathcal{R}(\tau)$ is much larger, and show how this completes the proof of our projection switching lemma.

One important aspect in which we differ from Håstad's and Razborov–Thapen's proof — and indeed, this is the key distinction between our projection switching lemma and previous switching lemmas — is that we will be concerned with the complexity of the randomly *projected* DNF $\mathrm{proj}_{\boldsymbol{\rho}} F \equiv \mathrm{proj}\,(F \upharpoonright \boldsymbol{\rho})$, rather than the randomly *restricted* DNF $F \upharpoonright \boldsymbol{\rho}$. Recalling our definition of projections (Definition 4) and Remark 9 in particular, we see that the decision tree depth of $\mathrm{proj}\,(F \upharpoonright \boldsymbol{\rho})$ can in general be significantly smaller than that of $F \upharpoonright \boldsymbol{\rho}$, since groups of distinct formal variables $\{x_{a,i} \colon i \in [w]\}$ of $F \upharpoonright \boldsymbol{\rho}$ get mapped to the same formal variable y_a under the projection operator. As we will see, the proof of our projection switching lemma crucially exploits this fact.

9.2 Canonical projection decision tree

To emphasize the fact that the DNF F and its random projection $\mathrm{proj}_{\boldsymbol{\rho}} F$ are over two different spaces of formal variables, we will let $\mathcal{X} = \{x_{a,i} \colon a \in A_{k-1}, i \in [w_{k-1}]\}$ denote the formal variables of F, and $\mathcal{Y} = \{y_a \colon a \in A_{k-1}\}$ denote the formal variables of $\mathrm{proj}_{\boldsymbol{\rho}} F$. For notational clarity, from this section through Section 9.4 we omit the subscripts on A_{k-1} and w_{k-1} and simply write A and w.

Definition 11. *Let $G : \{0,1\}^{A\times[w]} \to \{0,1\}$ be a DNF over $\mathcal{X}$ and T be a term in G. We say that a variable $x_{a,i}$* occurs positively in T *if T contains the unnegated literal $x_{a,i}$, and that it* occurs negatively in T *if T contains the negated literal $\overline{x}_{a,i}$. We say that $x_{a,i}$* occurs in T *if it either occurs positively or negatively in T.*

Definition 12. *For any $\eta \subseteq \mathcal{Y}$ and assignment $\pi \in \{0,1\}^{\eta}$, the restriction $(\eta \mapsto \pi) \in \{0,1,*\}^{A\times[w]}$ to the variables in $\mathcal{X}$ is defined as follows: for all $a \in A$ and $i \in [w]$,*

$$(\eta \mapsto \pi)_{a,i} = \begin{cases} \pi(y_a) & \textit{if } y_a \in \eta \\ * & \textit{otherwise.} \end{cases}$$

We stress that for a given a, the value of $(\eta \mapsto \pi)_{a,i}$ is independent of the value of $i \in [w]$.

Next, we define a procedure which, given any DNF G over $\mathcal{X}$, returns a "canonical" decision tree $\mathsf{ProjDT}(G)$ over $\mathcal{Y}$ computing its projection $\mathrm{proj}\, G$. The proof of our switching lemma will establish that the depth of $\mathsf{ProjDT}(F \upharpoonright \boldsymbol{\rho})$ is small with high probability; this clearly implies that the decision tree depth of $\mathrm{proj}_{\boldsymbol{\rho}} F \equiv \mathrm{proj}\,(F \upharpoonright \boldsymbol{\rho})$ is small with high probability. (We remark that both Håstad's and Razborov's proofs of Håstad's Switching Lemma consider an analogous notion of a canonical decision tree whose depth they bound; in their context, however, the canonical decision tree computes the DNF itself, whereas the canonical decision tree we now define computes the *projection* of the DNF.)

Definition 13 (Canonical projection decision tree)**.** *Let $G : \{0,1\}^{A\times[w]} \to \{0,1\}$ be a DNF over $\mathcal{X}$, where we assume a fixed but arbitrary ordering on its terms, and likewise on the literals within each term. The* canonical projection decision tree $\mathsf{ProjDT}(G) : \{0,1\}^{A} \to \{0,1\}$ *associated with G is defined recursively as follows:*

1. *If $G \equiv 1$ (i.e. if $G(X) = 1$ for all $X \in \{0,1\}^{A\times[w]}$) output the trivial decision tree* $\mathsf{ProjDT}(G) \equiv 1$*, and likewise, if $G \equiv 0$ output* $\mathsf{ProjDT}(G) \equiv 0$.

2. *Otherwise, let T be the first term in G such that $T \not\equiv 0$, and let*
$$\eta = \{y_a \colon x_{a,i} \text{ occurs in } T \text{ for some } i \in [w]\} \subseteq \mathcal{Y}$$
3. $\mathsf{ProjDT}(G)$ *queries all the variables in η in its first $|\eta|$ levels.*
4. *For each path $\pi \in \{0,1\}^{\eta}$, recurse on $G \upharpoonright (\eta \mapsto \pi)$.*

We stress that while G is a DNF over the variables in $\mathcal{X}$, the canonical projection decision tree $\mathsf{ProjDT}(G)$ queries variables in $\mathcal{Y}$. The following fact is a straightforward consequence of Definition 13:

Fact 9.3. $\mathsf{ProjDT}(G)$ *computes* $\operatorname{proj} G$.

9.3 Encoding bad restrictions

Fix $\tau \in \{0,1,*\}^{A\times[w]}$, and consider
$$\mathcal{B} = \{\rho \in \{0,1,*\}^{A\times[w]} \colon \rho \text{ refines } \tau \text{ and } \operatorname{proj}_\rho F \text{ is not a depth-}s\text{ decision tree}\},$$
We call these restrictions $\rho \in \mathcal{B}$ *bad*, and recall that our goal is to bound $\mathbf{Pr}[\boldsymbol{\rho} \in \mathcal{B}]$ for $\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)$. Fix a bad restriction $\rho \in \mathcal{B}$. It will be convenient for us to adopt the equivalent view of $\operatorname{proj}_\rho F$ as $\operatorname{proj}(F \upharpoonright \rho)$ in this section. Since $\operatorname{proj}(F \upharpoonright \rho)$ is not computed by a depth-s DT over $\mathcal{Y}$, this in particular implies that the canonical projection decision tree $\mathsf{ProjDT}(F \upharpoonright \rho)$ has depth at least s (recall by Fact 9.3 that $\mathsf{ProjDT}(F \upharpoonright \rho)$ computes $\operatorname{proj}(F \upharpoonright \rho)$), and so we may let $\pi \in \{0,1\}^{\geq s}$ be the leftmost root-to-leaf path of length at least s in $\mathsf{ProjDT}(F \upharpoonright \rho)$.

We now define a few objects associated with ρ and π: for some $1 \leq j \leq s$, we define

- A collection of terms $T_1, \ldots, T_j$ in F.
- Disjoint sets of variables $\eta_1, \ldots, \eta_j \subseteq \mathcal{Y}$, and for each such η_ℓ, a bit string $\operatorname{encode}(\eta_\ell) \in \{0,1\}^{|\eta_\ell|(\log r+1)}$.
- A restriction $\sigma = \sigma^1\sigma^2\cdots\sigma^j \in \{0,1,*\}^{A\times[w]}$ such that $\sigma^{-1}(\{0,1\}) \subseteq \rho^{-1}(*)$ (i.e. σ only sets to constants variables left free by ρ).
- Disjoint sets of variables $\gamma_1, \ldots, \gamma_j \subseteq \mathcal{X}$, and for each such γ_ℓ, a bit string $\operatorname{encode}(\gamma_\ell)$ of Hamming weight $|\gamma_\ell|$ and length r.
- A decomposition of the length-s prefix $\pi' = \pi^1\pi^2\cdots\pi^j \in \{0,1\}^s$ of π.

These objects are defined inductively starting from $\ell = 1$ up to $\ell = j$, where $j \in [s]$ is the smallest integer such that the η_ℓ's as defined below satisfy $|\eta_1 \cup \cdots \cup \eta_j| \geq s$. For $\ell \in [j]$,

- T_ℓ is the first term in F such that $T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1}) \not\equiv 0$ and
$$\eta_\ell = \{y_a \colon x_{a,i} \text{ occurs in } T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1}) \text{ for some } i \in [w]\} \subseteq \mathcal{Y}.$$

  We define $\operatorname{encode}(\eta_\ell) \in \{0,1\}^{|\eta_\ell|(\log r+1)}$ as follows: for each $y_a \in \eta_\ell$, we use $\log|T_\ell| \leq \log r$ bits to encode the location of x_{a,i_1} in T_ℓ, where
$$i_1 := \min\{i \in [w] \colon x_{a,i} \text{ occurs in } T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})\},$$
  along with a single bit to indicate whether y_a is the last variable in η_ℓ.

– Let $\sigma^\ell \in \{0,1,*\}^{A\times[w]}$ be defined as follows: for each $y_a \in \eta_\ell$ and $i \in [w]$,

$$\sigma^\ell_{a,i} = \begin{cases} 1 & \text{if } x_{a,i} \text{ occurs positively in } T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1}), \\ 0 & \text{if } x_{a,i} \text{ occurs negatively in } T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1}), \\ 0 & \text{if } \rho_{a,i} = * \text{ and } x_{a,i} \text{ does not occur in } T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1}). \end{cases}$$

(Note that if $x_{a,i}$ occurs in $T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})$, then certainly $\rho_{a,i} = *$.) All remaining entries of σ^ℓ not specified above have value $*$.

We make a few observations that will be useful for us later. First observe that for every $y_a \in \eta_\ell$,

(i) $(\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1}))_a \equiv \rho_a$

since $y_a \notin \eta_1 \cup \cdots \cup \eta_{\ell-1}$. Furthermore, writing $S_a = S_a(\tau)$ to denote $\tau_a^{-1}(*) = \{i \in [w] \colon \tau_{a,i} = *\}$ and $\rho(S_a)$ to denote the substring of ρ_a with coordinates in S_a, we claim that for every $y_a \in \eta_\ell$,

(ii) $\tau_a \in \{*,1\}^w \setminus \{1\}^w$,

(iii) the set S_a is k-acceptable,

(iv) $\rho(S_a) \in \{*,1\}^{S_a} \setminus \{1\}^{S_a}$ (and hence $\rho_a \in \{*,1\}^w \setminus \{1\}^w$ by (ii)),

(v) $(\rho\sigma^\ell)(S_a) \in \{0,1\}^{S_a}$,

(vi) $(\rho(S_a))^{-1}(1) \subseteq ((\rho\sigma^\ell)(S_a))^{-1}(1)$.

To see this, first note that since $y_a \in \eta_\ell$ it must be the case that $\rho_{a,i} = *$ for at least one $i \in S_a$, and by inspecting (12) of Definition 9 we have that indeed (ii), (iii), and (iv) hold. Claims (v) and (vi) follow from the fact that σ^ℓ is defined so that $\sigma^\ell_{a,i} \in \{0,1\}$ iff $\rho_{a,i} = *$. These claims will be useful for us later in the proof of Lemma 9.7.

Second, we claim that

$$T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})\sigma^\ell \equiv 1. \tag{21}$$

To see this, we note that every variable that occurs in term $T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})$ is fixed by σ^ℓ, and furthermore, each is fixed in the unique way so as to satisfy the term. This will be useful for us later in the proof of Proposition 9.4.

$$\gamma_\ell = \{x_{a,i} : \sigma^\ell_{a,i} = 1\} \subseteq \mathcal{X},$$

and let encode(γ_ℓ) be the string encode$(\gamma_\ell) \in \{0,1\}^r$ of Hamming weight $|\gamma_\ell|$ and length r indicating the location of the elements of γ_ℓ within T_ℓ.

– Let π^ℓ be the length-$|\eta_\ell|$ substring of π from index $|\eta_1 \cup \cdots \cup \eta_{\ell-1}| + 1$ through $|\eta_1 \cup \cdots \cup \eta_\ell|$ inclusive.

After the final iteration $\ell = j$, if necessary, we trim η_j and π^j so that $|\eta_1 \cup \cdots \cup \eta_j| = |\pi^1 \cdots \pi^j|$ is *exactly* s, and redefine σ^j and γ_j appropriately. We refer the reader to Figure 1 and its caption for a concrete example and explanation of our encoding procedure.

$$T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})$$
$$\equiv\ \overline{x}_{a,1} \wedge x_{a,8} \wedge x_{a',4} \wedge \overline{x}_{a'',2}$$

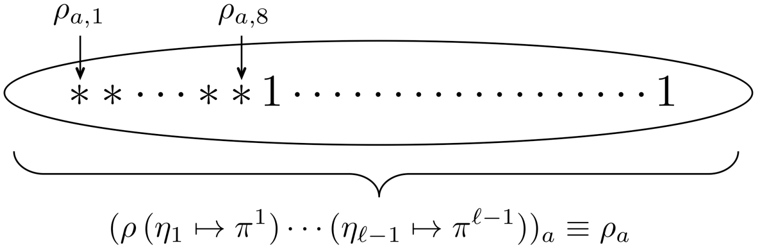
$\rho_{a,1}$
$\rho_{a,8}$
$* * \cdots * * 1 \cdots\cdots\cdots\cdots\cdots\cdots 1$
$(\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1}))_a \equiv \rho_a$

Figure 1: Let T_ℓ be the first term not falsified by $\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})$, and suppose it evaluates to $\overline{x}_{a,1} \wedge x_{a,8} \wedge x_{a',4} \wedge \overline{x}_{a'',2}$. In this example η_ℓ will be the set $\{y_a, y_{a'}, y_{a''}\} \subseteq \mathcal{Y}$. Focusing on variables from the a-th block, we first recall our observation earlier that $(\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1}))_a \equiv \rho_a$ since $y_a \notin \eta_1 \cup \cdots \cup \eta_{\ell-1}$ (Claim (i) in Section 9.4). Furthermore, as illustrated above, we have that $\rho_a \in \{*, 1\}^w \setminus \{1\}^w$ and ρ_a refines $\tau_a \in \{*, 1\}^w \setminus \{1\}^w$ (Claims (ii) and (iv) of Section 9.4).

Since $x_{a,1}$ and $x_{a,8}$ occur in $T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})$ it certainly must be the case that $\rho_{a,1} = \rho_{a,8} = *$; there may also be other coordinates $i \in [w]$ such that $\rho_{a,i} = *$ and $x_{a,i}$ does not occur in $T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})$ (coordinates 2 through 7 in our example above). For $i \in [w]$ such that $\rho_{a,i} = *$ and $x_{a,i}$ occurs in $T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})$, the restriction σ^ℓ fixes $x_{a,i}$ so as to partially satisfy $T_\ell \upharpoonright \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})$: in our example above, $\sigma^\ell_{a,1} = 0$ (since $x_{a,1}$ occurs negatively) whereas $\sigma^\ell_{a,8} = 1$ (since $x_{a,8}$ occurs positively). The remaining variables $x_{a,2}, \ldots, x_{a,7}$ are set to 0 by σ^ℓ, yielding a completely fixed block $(\rho\sigma^\ell)_a \in \{0, 1\}^w$ (Claim (v) in Section 9.4). Intuitively, we "break symmetry" and set these variables to 0 (rather than 1) so that the decoder will be able to "undo" them in $(\rho\sigma^\ell)_a$ without any auxiliary information: since $\rho_a \in \{*, 1\}^w \setminus \{1\}^w$, the decoder readily infers $\rho_{a,i} = *$ for all $i \in [w]$ such that $(\rho\sigma^\ell)_{a,i} = 0$. And indeed, for the set $\gamma_\ell \subseteq \mathcal{X}$ of variables $x_{a,i}$ that are set to 1 by σ^ℓ, we provide the decoder with the auxiliary information encode(γ_ℓ) so that she is able to "undo" them in $(\rho\sigma^\ell)_a$.

9.4 Decodability

Let $\eta = \eta_1 \cup \cdots \cup \eta_j$, $\mathrm{encode}(\eta) = \mathrm{encode}(\eta_1) \cdots \mathrm{encode}(\eta_j) \in \{0,1\}^{s(1+\log r)}$, $\sigma = \sigma^1 \cdots \sigma^j$, $\gamma = \gamma_1 \cup \cdots \cup \gamma_j$, $\mathrm{encode}(\gamma) = \mathrm{encode}(\gamma_1) \cdots \mathrm{encode}(\gamma_j) \in \{0,1\}^{rs}$, and $\pi' = \pi^1 \cdots \pi^j \in \{0,1\}^s$. Our main result in this subsection is the following proposition:

Proposition 9.4. *The map* $\theta : \mathcal{B} \to \{0,1,*\}^{A\times[w]} \times \{0,1\}^s \times \{0,1\}^{s(1+\log r)} \times \{0,1\}^{rs}$,

$$\theta(\rho) = (\rho\sigma, \pi', \mathrm{encode}(\eta), \mathrm{encode}(\gamma)),$$

is an injection.

Before proving Proposition 9.4, we state a slight extension of an observation made above in the definition of σ^ℓ:

Lemma 9.5. *For all* $1 \le \ell \le j-1$ *we have*

$$T_\ell \restriction \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})\,\sigma^\ell \cdots \sigma^j \equiv 1,$$

and when $\ell = j$ *we have* $T_j \restriction \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{j-1} \mapsto \pi^{j-1})\,\sigma^j \not\equiv 0$.

Proof. As we observed in the definition of σ^ℓ above (c.f. (21)), we have that σ^ℓ is designed so that

$$T_\ell \restriction \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})\,\sigma^\ell \equiv 1,$$

and certainly this remains true when the restriction is further extended by $\sigma^{\ell+1}\cdots\sigma^j$. We do not necessarily have this property for $\ell = j$ due to our possible trimming of η_j so that $\eta_1 \cup \cdots \cup \eta_j$ has cardinality exactly s; this results in a redefinition of σ^j where some of its coordinates are set from $\{0,1\}$ back to $*$. However it is still the case that σ^j partially satisfies $T_j \restriction \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{j-1} \mapsto \pi^{j-1})$, and hence $T_j \restriction \rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{j-1} \mapsto \pi^{j-1})\,\sigma^j \not\equiv 0$. □

Proof of Proposition 9.4. We prove the proposition by describing a procedure that allows a "decoder" to uniquely obtain ρ given $(\rho\sigma, \pi', \mathrm{encode}(\eta), \mathrm{encode}(\gamma))$. Recall that T_1 is defined to be the first term in F not falsified by ρ. By Lemma 9.5, this remains true when ρ is extended by σ: that is, the first term T_1' in F such that $T_1' \restriction \rho\sigma \not\equiv 0$ is precisely T_1 itself. Therefore, given $\rho\sigma$ the decoder is able to identify T_1 in F, and with T_1 in hand she is able to then use $\mathrm{encode}(\eta_1)$ and $\mathrm{encode}(\gamma_1)$ to recover η_1 and γ_1 respectively. Next, she "undoes" σ^1 in $\rho\sigma = \rho\sigma^1\sigma^2\cdots\sigma^j$ and obtains $\rho\sigma^2\cdots\sigma^j$ as follows: for every $y_a \in \eta_1$, she sets $(\rho\sigma)_{a,i}$ back to $*$ for all $i \in U_a$, where

$$U_a = \{i \in [w] \colon (\rho\sigma)_{a,i} = 0 \text{ or } x_{a,i} \in \gamma_1\}.$$

To see that this indeed "undoes" σ^1, first recall that for every $y_a \in \eta_1$, the restriction σ^1 is defined so that $\sigma^1_{a,i} \in \{0,1\}$ iff $\rho_{a,i} = *$, and furthermore, $\sigma^1_{a,i} = 1$ iff $x_{a,i} \in \gamma_1$. (Recall the example in Figure 1.) Therefore, to obtain $\rho\sigma^2\cdots\sigma^j$ from $\rho\sigma^1\sigma^2\cdots\sigma^j$, for every $y_a \in \eta_1$ and $i \in [w]$ the decoder sets $(\rho\sigma)_{a,i}$ back to $*$ if either $(\rho\sigma)_{a,i} = 0$ or $x_{a,i} \in \gamma_1$. Finally, using $\pi^1 \in \{0,1\}^{\eta_1}$ she constructs the hybrid restriction $\rho\,(\eta_1 \mapsto \pi^1)\,\sigma^2\cdots\sigma^j$.

By the same reasoning, for every $2 \le \ell \le j$ the decoder is able to iteratively recover $T_\ell, \eta_\ell, \gamma_\ell$, and π^ℓ from the hybrid restriction

$$\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})\,\sigma^\ell\cdots\sigma^j.$$

With this information she "undoes" σ^ℓ within $\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_{\ell-1} \mapsto \pi^{\ell-1})\,\sigma^\ell \cdots \sigma^j$, and constructs the next hybrid restriction

$$\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_\ell \mapsto \pi^\ell)\,\sigma^{\ell+1}\cdots\sigma^j.$$

Finally, having recovered $\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_j \mapsto \pi^j)$ and $\eta = \eta_1 \cup \cdots \cup \eta_j$, the decoder will have all the information she needs to recover the actual restriction ρ: she sets $(\rho\,(\eta_1 \mapsto \pi^1)\cdots(\eta_j \mapsto \pi^j))_{a,i}$ back to $*$ for every $y_a \in \eta$ and $i \in U_a$. □

9.5 Proof of Proposition 9.2

For all possible outcomes $\vartheta_2, \vartheta_3, \vartheta_4$ of the second, third, and fourth coordinates of the map θ defined in Proposition 9.4, we define

$$\begin{aligned} \mathcal{B}_{\vartheta_2,\vartheta_3} &= \{\rho \in \mathcal{B} \colon \theta_2(\rho) = \vartheta_2, \theta_3(\rho) = \vartheta_3\} \subseteq \mathcal{B}. \\ \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4} &= \{\rho \in \mathcal{B} \colon \theta_2(\rho) = \vartheta_2, \theta_3(\rho) = \vartheta_3, \theta_4(\rho) = \vartheta_4\} \subseteq \mathcal{B}_{\vartheta_2,\vartheta_3}. \end{aligned}$$

We begin by bounding the probability that $\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)$ belongs to $\mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}$ for a fixed tuple $(\vartheta_2, \vartheta_3, \vartheta_4)$. The following fact, giving the probability mass function of $\mathcal{R}(\tau)$, will be useful for us (its proof is by inspection of Definition 9):

Fact 9.6. *Fix $\tau \in \{0,1,*\}^{A_k}$, and write $S_a = S_a(\tau)$ to denote $\tau_a^{-1}(*) = \{i \in [w_{k-1}] \colon \tau_{a,i} = *\}$. Then $\mathbf{Pr}_{\boldsymbol{\rho}\leftarrow\mathcal{R}(\tau)}[\boldsymbol{\rho} = \rho] = \xi(\rho)$ for all $\rho \in \{0,1,*\}^{A_k}$, where $\xi : \{0,1,*\}^{A_k} \to [0,1]$ is the probability mass function:*

$$\xi(\rho) = \prod_{\substack{a \in A_{k-1} \\ S_a \neq \emptyset}} \zeta_a(\rho(S_a)),$$

and $\rho(S_a)$ denotes the substring of ρ_a with coordinates in S_a, and $\zeta_a : \{0,1,*\}^{S_a} \to [0,1]$ is the probability mass function:

$$\zeta_a(\varrho) = \begin{cases} \lambda & \text{if } \varrho = \{1\}^{S_a}, \\ q_a \cdot \dfrac{t_k{}^{|\varrho^{-1}(*)|}(1-t_k)^{|\varrho^{-1}(1)|}}{1-(1-t_k)^{|S_a|}} & \text{if } \varrho \in \{*,1\}^{S_a} \setminus \{1\}^{S_a}, \\ (1-\lambda-q_a)\cdot \dfrac{t_k{}^{|\varrho^{-1}(0)|}(1-t_k)^{|\varrho^{-1}(1)|}}{1-(1-t_k)^{|S_a|}} & \text{if } \varrho \in \{0,1\}^{S_a} \setminus \{1\}^{S_a}. \end{cases}$$

Lemma 9.7. *For all $\vartheta_2, \vartheta_3, \vartheta_4$,*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho}\leftarrow\mathcal{R}(\tau)}\left[\boldsymbol{\rho} \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}\right] = \Big(O\big(w^{-1/4}\big)\Big)^s \left(\frac{t_k}{1-t_k}\right)^{\|\vartheta_4\|},$$

where $\|\vartheta_4\|$ denotes $|\vartheta_4^{-1}(1)|$, the Hamming weight of ϑ_4.

Proof. Fix $\rho \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}$. The restrictions ρ and $\theta_1(\rho) = \rho\sigma$ differ in exactly s blocks: these are the blocks $a \in A_{k-1}$ such that $y_a \in \eta$. Consider any such $a \in A_{k-1}$, and recall (as observed in the definition of σ) that S_a is k-acceptable and $\rho(S_a) \in \{*,1\}^{S_a} \setminus \{1\}^{S_a}$ whereas $(\rho\sigma)(S_a) \in \{0,1\}^{S_a}$.

Let Δ_a denote $|(\rho\sigma)_a^{-1}(1)| - |\rho_a^{-1}(1)|$, the number of "new 1's" that σ introduces into block a (note that as observed earlier we have that $\Delta_a \geq 0$). By Fact 9.6, we have that

$$\frac{\zeta_a((\rho\sigma)(S_a))}{\zeta_a(\rho(S_a))} = \begin{cases} \dfrac{\lambda}{q_a} \cdot \dfrac{1-(1-t_k)^{|S_a|}}{t_k^{\Delta_a}(1-t_k)^{|S_a|-\Delta_a}} & \text{if } (\rho\sigma)(S_a) = \{1\}^{S_a} \\ \dfrac{1-\lambda-q_a}{q_a}\left(\dfrac{1-t_k}{t_k}\right)^{\Delta_a} & \text{if } (\rho\sigma)(S_a) \in \{0,1\}^{S_b} \setminus \{1^{S_a}\}. \end{cases} \tag{22}$$

Since S_a is k-acceptable, we have that $|S_a| = qw \pm w^{\beta(k,d)}$ and therefore

$$\begin{aligned} (1-t_k)^{|S_a|} &\leq \frac{(1-t_k)^{qw}}{(1-t_k)^{w^{\beta(k,d)}}} \\ &= \frac{qt_{k-1}+\lambda}{(1-t_k)^{w^{\beta(k,d)}}} \\ &\leq \frac{qt_{k-1}+\lambda}{1-t_k w^{\beta(k,d)}} \ \leq \ 2q^2, \end{aligned}$$

where the equality is by (8) and the final inequality uses Lemma 7.1, (7) and (10). Since $q_a \leq 2q$ by Lemma 10.5, we may lower bound the quantity in the first line of (22) by

$$\frac{\lambda}{8q^3}\left(\frac{1-t_k}{t_k}\right)^{\Delta_a} = \Omega(w^{1/4})\left(\frac{1-t_k}{t_k}\right)^{\Delta_a},$$

where we have used our choice of λ in (7) and the estimates (10). Similarly, for the second quantity in the second line of (22) we have the lower bound

$$\frac{1-\lambda-q_a}{q_a}\left(\frac{1-t_k}{t_k}\right)^{\Delta_a} = \Omega\left(\sqrt{\frac{w}{\log w}}\right)\left(\frac{1-t_k}{t_k}\right)^{\Delta_a}$$

and so in both cases we may lower bound the ratio in (22) by

$$\frac{\zeta_a((\rho\sigma)(S_a))}{\zeta_a(\rho(S_a))} = \Omega(w^{1/4})\left(\frac{1-t_k}{t_k}\right)^{\Delta_a}.$$

Since $\sum_{a:\ \rho_a \neq (\rho\sigma)_a} \Delta_a = \|\vartheta_4\|$, it follows from Fact 9.6 that

$$\frac{\xi(\theta_1(\rho))}{\xi(\rho)} = \frac{\xi(\rho\sigma)}{\xi(\rho)} = \prod_{\substack{a \in A_{k-1} \\ S_a \neq \emptyset}} \frac{\zeta_a((\rho\sigma)(S_a))}{\zeta_a(\rho(S_a))} = \left(\Omega(w^{1/4})\right)^s \left(\frac{1-t_k}{t_k}\right)^{\|\vartheta_4\|}. \tag{23}$$

Finally, summing over all $\rho \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}$ we conclude that

$$\begin{aligned} \mathop{\mathbf{Pr}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)}\left[\boldsymbol{\rho} \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}\right] = \sum_{\rho \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}} \xi(\rho) &= \left(O(w^{-1/4})\right)^s \left(\frac{t_k}{1-t_k}\right)^{\|\vartheta_4\|} \sum_{\rho \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}} \xi(\theta_1(\rho)) \\ &= \left(O(w^{-1/4})\right)^s \left(\frac{t_k}{1-t_k}\right)^{\|\vartheta_4\|}. \end{aligned}$$

Here the first inequality is by (23), and the second uses the fact that θ is an injection (Proposition 9.4), and hence any two distinct $\rho, \rho' \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}$ map to distinct $\theta_1(\rho), \theta_1(\rho') \in \{0,1,*\}^{A\times[w]}$, so $\sum_{\rho \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}} \xi(\theta_1(\rho))$ is at most 1 since ξ is a probability mass function. □

Proposition 9.2 follows as a straightforward consequence of Lemma 9.7:

Proof of Proposition 9.2. Summing over all $\vartheta_4 \in \{0,1\}^{rs}$ and stratifying according to Hamming weight, we have that

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\boldsymbol{\rho}\leftarrow\mathcal{R}(\tau)}[\boldsymbol{\rho} \in \mathcal{B}_{\vartheta_2,\vartheta_3}] &= \sum_{i=0}^{rs} \sum_{\substack{\vartheta_4\in\{0,1\}^{rs} \\ \|\vartheta_4\|=i}} \mathop{\mathbf{Pr}}_{\boldsymbol{\rho}\leftarrow\mathcal{R}(\tau)}[\boldsymbol{\rho} \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}] \\
&\le \sum_{i=0}^{rs} \binom{rs}{i} \left(\frac{t_k}{1-t_k}\right)^i \Big(O\big(w^{-1/4}\big)\Big)^s \\
&= \left(1+\frac{t_k}{1-t_k}\right)^{rs} \Big(O\big(w^{-1/4}\big)\Big)^s = \Big(O\Big(e^{rt_k/(1-t_k)}\cdot w^{-1/4}\Big)\Big)^s.
\end{aligned}$$

Taking a union bound over all 2^s possible $\vartheta_2 \in \{0,1\}^s$ and $(2r)^s$ possible $\vartheta_3 \in \{0,1\}^{s(1+\log r)}$ completes the proof. □

Proof of Proposition 9.1. For Proposition 9.1, we first observe that Proposition 9.4 also holds for $\boldsymbol{\rho} \leftarrow \mathcal{R}_{\text{init}}$ (the proof is completely identical, with τ being the trivial restriction $\{*\}^n$). Proposition 9.1 then follows as a consequence of Proposition 9.4 in a very similar manner (the calculations are in fact significantly simpler); we point out the essential differences in this section. We begin with the following analogue of Fact 9.6, specifying the probability mass function of $\mathcal{R}_{\text{init}}$ (like Fact 9.6, its proof is by inspection of Definition 6):

Fact 9.8. *$\mathbf{Pr}_{\boldsymbol{\rho}\leftarrow\mathcal{R}_{\text{init}}}[\boldsymbol{\rho} = \rho] = \xi(\rho)$ for all $\rho \in \{0,1,*\}^{A_{d-1}\times[m]}$ (recall that $w_{d-1} = m$), where $\xi : \{0,1,*\}^{A_{d-1}\times[m]} \to [0,1]$ is the probability mass function:*

$$\xi(\rho) = \prod_{a\in A_{d-1}} \zeta(\rho_a),$$

and $\zeta : \{0,1,*\}^m \to [0,1]$ is the probability mass function:

$$\zeta(\varrho) = \begin{cases} \lambda & \textit{if } \varrho = \{1\}^m, \\ q \cdot \dfrac{p}{1-p} & \textit{if } \varrho \in \{*,1\}^m \setminus \{1\}^m, \\ (1-\lambda-q)\cdot\dfrac{p}{1-p} & \textit{if } \varrho \in \{0,1\}^m \setminus \{1\}^m. \end{cases}$$

Fact 9.8 gives us the following analogue of (22):

$$\frac{\zeta((\rho\sigma)_a)}{\zeta(\rho_a)} = \begin{cases} \dfrac{\lambda(1-p)}{qp} & \text{if } (\rho\sigma)_a = \{1\}^m \\ \dfrac{1-\lambda-q}{q} & \text{if } (\rho\sigma)_a \in \{0,1\}^m \setminus \{1\}^m, \end{cases}$$

and so by our choice of λ in (7) and our estimates (10) this ratio is always at least $\Omega\big(w^{1/4}\big)$. (Unlike the proof of Lemma 9.7, our lower bound here does not depend on $\Delta_a = |(\rho\sigma)_a^{-1}(1)| - |\rho_a^{-1}(1)|$.) By the same calculations as in the proof of Lemma 9.7, we have the following analogue of Lemma 9.7:

Lemma 9.9. *For all $\vartheta_2, \vartheta_3, \vartheta_4$, we have that* $\mathbf{Pr}_{\boldsymbol{\rho} \leftarrow \mathcal{R}_{\text{init}}}\left[\boldsymbol{\rho} \in \mathcal{B}_{\vartheta_2,\vartheta_3,\vartheta_4}\right] = \Big(O\big(w^{-1/4}\big)\Big)^s$.

Proposition 9.1 follows by a union bound over all 2^s possible $\vartheta_2 \in \{0,1\}^s$, $(2r)^s$ possible $\vartheta_3 \in \{0,1\}^{s(1+\log r)}$, and 2^{rs} possible $\vartheta_4 \in \{0,1\}^{rs}$ (unlike in the proof of Proposition 9.2 we do not have to stratify the union bound over $\vartheta_4 \in \{0,1\}^{rs}$ according to Hamming weight).

9.6 Approximator simplifies under random projections

The main results of this section are Theorems 13 and 14. The first of these theorems says that any depth-d circuit whose size is not too large and whose bottom fan-in is significantly smaller than that of Sipser_d will collapse to a shallow decision tree with high probability under the random projection $\boldsymbol{\Psi}$ from Definition 10:

Theorem 13. *For $2 \le d \le \frac{c\sqrt{\log n}}{\log\log n}$, let $C : \{0,1\}^n \to \{0,1\}$ be a depth-d circuit with bottom fan-in at most $\frac{\log n}{10(d-1)}$ and size $S \le 2^{n^{\frac{1}{6(d-1)}}}$. Then $\boldsymbol{\Psi}(C)$ is computed by a decision tree of depth $n^{\frac{1}{4(d-1)}}$ with probability $1 - \exp\big(-\Omega\big(n^{\frac{1}{6(d-1)}}\big)\big)$.*

The second theorem is quite similar; it says that under the random projection $\boldsymbol{\Psi}$, any depth-d circuit C that is not too large, regardless of its bottom fan-in, will collapse to a depth-2 circuit with bounded bottom fan-in and with top gate matching that of C:

Theorem 14. *For $2 \le d \le \frac{c\sqrt{\log n}}{\log\log n}$, let $C : \{0,1\}^n \to \{0,1\}$ be a depth-d circuit of size $S \le 2^{\frac{1}{2}n^{\frac{1}{6(d-1)}}}$ and unbounded bottom fan-in.*

1. *If the top gate of C is an AND, then $\boldsymbol{\Psi}(C)$ is $(1/S)$-close (with respect to the uniform distribution on $\{0,1\}^n$) to a width-$n^{\frac{1}{4(d-1)}}$ CNF with probability $1 - \exp\big(-\Omega\big(n^{\frac{1}{6(d-1)}}\big)\big)$.*

2. *If the top gate of C is an OR, then $\boldsymbol{\Psi}(C)$ is $(1/S)$-close to a width-$n^{\frac{1}{4(d-1)}}$ DNF with probability $1 - \exp\big(-\Omega\big(n^{\frac{1}{6(d-1)}}\big)\big)$.*

We first prove Theorem 13, which deals with depth-d circuits with bounded bottom fan-in. We state the following simple lemma explicitly for convenience of later reference:

Lemma 9.10. *Suppose that $3 \le d \le \frac{c\log w}{\log\log w}$. For $2 \le k \le d-1$ and $\ell \in \mathbb{N}$, let $C : \{0,1\}^{A_{k+1}} \to \{0,1\}$ be a size-S depth-ℓ circuit with bottom fan-in $w^{1/5}$. For any $\tau \in \{\bullet, \circ, *\}^{A_{k+1}}$, with probability at least $1 - S \cdot 4^{-w^{1/5}}$ over $\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)$, we have that $\mathrm{proj}_{\boldsymbol{\rho}}\, C$ is a depth-$(\ell-1)$ circuit with bottom fan-in $w^{1/5}$, and has the same number of gates at distance at least two from the input variables as C.*

Proof. The lemma follows from applying Proposition 9.2 with $r = s = w^{1/5}$ and a union bound over all gates of C (at most S many) that are at distance 2 from the input variables. □

The following proposition directly implies Theorem 13 by straightforward translation of parameters, recalling (5):

Proposition 9.11. *For $2 \le d \le \frac{c\sqrt{\log n}}{\log\log n}$, let $C : \{0,1\}^{A_d} \to \{0,1\}$ be a depth-d circuit with bottom fan-in $\frac{1}{5}m$ and size $S \le 2^{w^{1/5}}$. Then $\mathbf{\Psi}(C)$ is computed by a depth-$(w^{1/5})$ decision tree with probability $1 - e^{-\Omega(w^{1/5})}$.*

Proof. Applying Proposition 9.1 with $r = \frac{1}{5}m$ and $s = w^{1/5}$ to each of the bottom-layer gates of C, we have that $\mathrm{proj}_{\boldsymbol{\rho}^{(d)}} C$ is a depth-$(d-1)$ circuit with bottom fan-in $w^{1/5}$ with probability at least $1 - S \cdot 4^{-w^{1/5}} \ge 1 - 2^{-w^{1/5}}$ over $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\mathrm{init}}$. If $d = 2$, we observe that in fact Proposition 9.1 gives us that $\mathrm{proj}_{\boldsymbol{\rho}^{(d)}} C$ is a decision tree of the desired depth, and we are done. If $d \ge 3$, the claim follows by a union bound over $d-2$ applications of Lemma 9.10 (where we observe from the proof of Lemma 9.10 that in the last application of Lemma 9.10 we may conclude that $\mathbf{\Psi}(C)$ is in fact a decision tree of depth $w^{1/5}$). □

Next we turn to Theorem 14. We require the following standard lemma showing that any circuit can be "trimmed" to reduce its bottom fan-in while changing its value on only a few inputs:

Lemma 9.12. *Let $C : \{0,1\}^n \to \{0,1\}$ be a circuit and let $\varepsilon > 0$. There exists a circuit $C' : \{0,1\}^n \to \{0,1\}$ such that*

1. *The size and depth of C' are both at most that of C;*
2. *The bottom fan-in of C' is at most $\log(S/\varepsilon)$;*
3. *C and C' are ε-close with respect to the uniform distribution.*

Proof. C' is obtained from C by replacing each bottom-level AND (OR, respectively) gate whose fan-in is too large with 0 (1, respectively). Each such gate originally takes its minority value on at most an ε/S fraction of all inputs so the lemma follows from a union bound. □

The following proposition directly implies Theorem 14 (by straightforward translation of parameters):

Proposition 9.13. *For $2 \le d \le \frac{c\sqrt{\log n}}{\log\log n}$, let $C : \{0,1\}^{A_d} \to \{0,1\}$ be a depth-d circuit of size $S \le 2^{\frac{1}{2}w^{1/5}}$ and unbounded bottom fan-in.*

1. *If the top gate of C is an AND, then $\mathbf{\Psi}(C)$ is $(1/S)$-close to a width-$(w^{1/5})$ CNF with probability $1 - e^{-\Omega(w^{1/5})}$.*
2. *If the top gate of C is an OR, then $\mathbf{\Psi}(C)$ is $(1/S)$-close to a width-$(w^{1/5})$ DNF with probability $1 - e^{-\Omega(w^{1/5})}$.*

Proof. By symmetry it suffices to prove the first claim. Applying Lemma 9.12 with $\varepsilon = 1/S$, we have that C is $(1/S)$-close to a circuit $C' : \{0,1\}^{A_d} \to \{0,1\}$ of size and depth at most that of C, and with bottom fan-in $\log(S/\varepsilon) = 2\log(S) \le w^{1/5}$. Certainly the size, depth, and bottom fan-in of $\mathrm{proj}_{\boldsymbol{\rho}^{(d)}} C'$ is at most that of C' with probability 1 over the randomness of $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\mathrm{init}}$ (note that unlike in the proof of Proposition 9.11, we do not argue that the depth of C' decreases by one under an $\mathcal{R}_{init}$-random projection; the bottom fan-in of C' is too large for us to apply Proposition 9.1). If $d = 2$ then this already gives the result (in fact with no failure probability). If $d \ge 3$, the proposition then follows by a union bound over $d-2$ applications of Proposition 9.10. □

10 Sipser retains structure under random projections

Now we turn our attention to the randomly projected target $\mathbf{\Psi}(\mathsf{Sipser}_d)$. As discussed in Section 7.3, we would like to establish Property 2 showing that Sipser_d "retains structure" under a $\mathbf{\Psi}$-random projection: with high probability over $\mathbf{\Psi}$, the randomly projected target $\mathbf{\Psi}(\mathsf{Sipser}_d)$ is a depth-one formula whose bias remains very close to 1/2 (with respect to an appropriate product distribution over $\{0,1\}^{w_0}$). This is necessarily a high-probability statement; to establish it, we must account for the failure probabilities introduced by each of the $d-1$ individual random projections $\mathrm{proj}_{\boldsymbol{\rho}^{(k)}}$ that comprise $\mathbf{\Psi} \equiv \{\boldsymbol{\rho}^{(k)}\}_{k\in\{2,\ldots,d\}}$.[3] To reason about these failure probabilities and carefully account for them, in Section 10.1 we introduce the notion of a "typical" restriction and prove some useful properties about how typicality interacts with our random projections. In Section 10.2 we use these properties to establish the main results of this section, that Sipser_d "retains structure" when it is hit with the random projection $\mathbf{\Psi}$.

10.1 Typical restrictions

Recalling the $\bullet, \circ$ notation from Table 2, we begin with the following definition:

Definition 14. *Let $\tau \in \{\bullet, \circ, *\}^{A_k}$ where $2 \le k \le d-1$. We say that τ is* typical *if it satisfies:*

1. *For every $a \in A_{k-1}$ the set $\tau_a^{-1}(*) \subseteq [w_{k-1}]$ is k-acceptable, where we recall from Definition 8 that this means*

$$|\tau_a^{-1}(*)| = qw \pm w^{\beta(k,d)} \quad \text{where } \beta(k,d) := \frac{1}{3} + \frac{d-k-1}{12d}.$$

(Note that $\frac{1}{3} \le \beta(k,d) \le \frac{5}{12} < \frac{1}{2}$ for all $d \in \mathbb{N}$ and $2 \le k \le d-1$.) We observe that by Definition 7, this condition implies that for every $\alpha \in A_{k-2}$, we have

$$\widehat{\tau}_\alpha \in \{*, \circ\}^{w_{k-2}}. \tag{24}$$

2. *For every $\alpha \in A_{k-2}$,*

$$|(\widehat{\tau}_\alpha)^{-1}(*)| \ge w_{k-2} - w^{4/5}.$$

We note that (24) and Condition (2) together imply that

$$\widehat{\widehat{\tau}}_\alpha = * \quad \text{for all } \alpha \in A_{k-2}.$$

See Figure 2 on the next page for an illustration of a typical τ. The rationale behind Definition 14 is that projections proj_ρ such that $\widehat{\rho}$ is typical have a very limited (and well-controlled) effect on the target Sipser_d: roughly speaking, these projections "wipe out" the bottom-level gates of the formula (reducing its depth by one), "trim" the fan-ins of the next-to-bottom-level gates from w to approximately $qw = \tilde{\Theta}(\sqrt{w})$, but otherwise essentially preserves the rest of the structure of the formula. We give a precise description in Section 10.2; see Remark 16.

[3] As a concrete example of a failure event, consider an outcome $\rho^{(d)} \in \mathrm{supp}(\mathcal{R}_{\mathrm{init}}) \equiv \{0,1,*\}^{A_{d-1}\times[m]}$ which is such that $(\rho_b^{(d)})^{-1}(0)$ is nonempty for all $b \in A_{d-1}$. In this case

$$\mathrm{proj}_{\rho^{(d)}} \mathsf{Sipser}_d \equiv \mathrm{proj}\,(\mathsf{Sipser}_d \upharpoonright \boldsymbol{\rho}^{(d)}) \equiv 0$$

(recall that the bottom-level gates of Sipser_d are AND gates), and our target function is set to the constant 0 already after the first $\mathcal{R}_{\mathrm{init}}$-random projection.

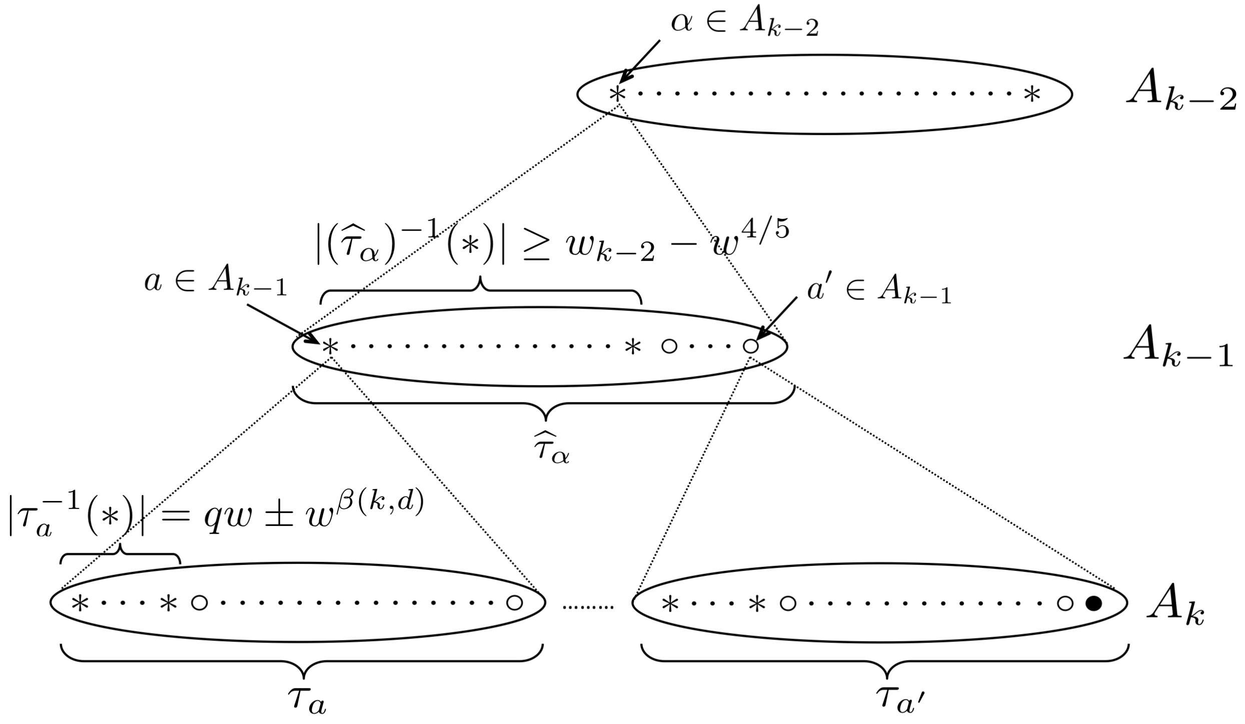

Figure 2: The figure illustrates a typical $\tau \in \{\bullet, \circ, *\}^{A_k}$. For $a \in A_{k-1}$, τ_a is a block of length w_{k-1}, i.e. a string in $\{\bullet, \circ, *\}^{w_{k-1}}$. We may think of the block τ_a as being located at level k. By Condition (1) of Definition 14, for every $a \in A_{k-1}$ we have that $|\tau_a^{-1}(*)|$, the number of $*$'s in τ_a, is roughly $qw = \tilde{\Theta}(\sqrt{w})$. The lift $\widehat{\tau}$ of τ is a string in $\{\bullet, \circ, *\}^{A_{k-1}}$, and for $\alpha \in A_{k-2}$, $\widehat{\tau}_\alpha$ is a block of length w_{k-2}. We may think of the block $\widehat{\tau}_\alpha$ as being located at level $k-1$. As stipulated by (24), for every $\alpha \in A_{k-2}$, the string $\widehat{\tau}_\alpha$ belongs to $\{*, \circ\}^{w_{k-2}}$. By Condition (2) of Definition 14, for every $\alpha \in A_{k-2}$, we have that $|(\widehat{\tau}_\alpha)^{-1}(*)|$, the number of $*$'s in $\widehat{\tau}_\alpha$, is at least $w_{k-2} - w^{4/5} = w_{k-2}(1 - o(1))$. Finally, we observe that (24) and Condition (2) of Definition 14 imply that $\widehat{\widehat{\tau}}_\alpha = *$ for every $\alpha \in A_{k-2}$.

To prove that $\mathbf{\Psi}(\mathsf{Sipser}_d)$ is a well-structured formula with high probability over the random choice of $\mathbf{\Psi} \equiv \{\boldsymbol{\rho}^{(k)}\}_{k\in\{2,\ldots,d\}}$, we will in fact establish the stronger statement showing that with high probability, every single one of the individual random projections $\mathrm{proj}_{\boldsymbol{\rho}^{(k)}}$ only has a limited and well-controlled effect (in the sense described above) on the structure of Sipser_d. By Definition 14, this amounts to showing that the lifts $\widehat{\boldsymbol{\rho}^{(d)}}, \ldots, \widehat{\boldsymbol{\rho}^{(2)}}$ associated with the $d-1$ individual projections comprising $\mathbf{\Psi}$ are *all* typical with high probability. We prove this inductively: we first show that for $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\mathrm{init}}$ its lift $\widehat{\boldsymbol{\rho}^{(d)}}$ is typical with high probability (Proposition 10.1), and then argue that if $\rho^{(k+1)}$ is typical then the lift $\widehat{\boldsymbol{\rho}^{(k)}}$ of $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}(\widehat{\rho^{(k+1)}})$ is also typical with high probability (Proposition 10.2). The parameters of Definition 14 are chosen carefully so that it "bootstraps" in the sense of Proposition 10.2; in particular, this is the reason why we allow more and more deviation from qw in Condition 1 as k gets smaller (closer to the root).

Our two main results in this subsection are the following:

Proposition 10.1 (Establishing initial typicality)**.** *Suppose that* $3 \le d \le \frac{c\log w}{\log\log w}$ *for a sufficiently small absolute constant* $c > 0$*. Then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho}\leftarrow\mathcal{R}_{\mathrm{init}}}[\widehat{\boldsymbol{\rho}} \textit{ is typical}] \ge 1 - e^{\tilde{\Omega}(w^{1/6})}.$$

Proposition 10.2 (Preserving typicality)**.** *Suppose that* $3 \le d \le \frac{c\log w}{\log\log w}$ *for a sufficiently small absolute constant* $c > 0$*. Let* $2 \le k \le d-1$ *and let* $\tau \in \{\bullet, \circ, *\}^{A_{k+1}}$ *be typical. Then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho}\leftarrow\mathcal{R}(\tau)}[\widehat{\boldsymbol{\rho}} \textit{ is typical}] \ge 1 - e^{-\Omega(w^{1/6})}.$$

10.1.1 Establishing initial typicality: Proof of Proposition 10.1

For notational brevity, throughout this subsubsection we write $\boldsymbol{\tau}$ to denote $\widehat{\boldsymbol{\rho}} \in \{0,1,*\}^{A_{d-1}}$ where $\boldsymbol{\rho} \leftarrow \mathcal{R}_{\mathrm{init}}$. We proceed to establish the two conditions of Definition 14.

Lemma 10.3 (Condition (1) of typicality)**.** *Fix* $a \in A_{d-2}$*. Then*

$$\mathbf{Pr}\left[|\boldsymbol{\tau}_a^{-1}(*)| = qw \pm w^{1/3}\right] \ge 1 - e^{-\tilde{\Omega}(w^{1/6})}.$$

Proof. Recalling (11), we have that

$$\mathbf{Pr}[\boldsymbol{\tau}_{a,i} = *] = q \quad \text{independently for all } i \in [w].$$

We shall apply Fact 5.1 with

$$\mathbf{S} = \mathbf{Z}_1 + \cdots + \mathbf{Z}_w \quad \text{where } \mathbf{Z}_i \leftarrow \{0_{1-q}, 1_q\} \quad (\text{so } \mu = \mathbf{E}[\mathbf{S}] \text{ is } qw),$$

and γ such that $\gamma\mu = w^{1/3}$. Observe that since $\mu = qw = \Theta((w\log w)^{1/2})$, we have $\gamma = \Theta(w^{-1/6}(\log w)^{-1/2})$. Hence by Fact 5.1 we have that

$$\mathbf{Pr}\left[\big||\boldsymbol{\tau}_a^{-1}(*)| - qw\big| > w^{1/3}\right] \le \exp\big(-\Omega(\gamma^2\mu)\big) = \exp\big(-\tilde{\Omega}(w^{1/6})\big). \qquad \square$$

The following observations may help the reader follow the next proof: Recalling Table 2, since our $\boldsymbol{\tau}$ belongs to $\{0,1,*\}^{A_{d-1}}$, we see that $\boldsymbol{\tau}$ corresponds to the second row of the table: the gates at depth $d-2$ are OR gates, a $\circ$-value for a coordinate of $\boldsymbol{\tau}$ corresponds to 0, and a $\bullet$-value corresponds to 1. However, since $\widehat{\boldsymbol{\tau}}$, the lift of $\boldsymbol{\tau}$, is one level higher than $\boldsymbol{\tau}$ in the Sipser_d formula (see Figure 2), $\widehat{\boldsymbol{\tau}}$ corresponds to the first row of the table; so when Definition 7 specifies a coordinate $\widehat{\boldsymbol{\tau}}_{\alpha,i}$ of $\widehat{\boldsymbol{\tau}}$, a $\circ$-value for $\widehat{\boldsymbol{\tau}}_{\alpha,i}$ corresponds to 1 and a $\bullet$-value corresponds to 0.

Lemma 10.4 (Condition (2) of typicality)**.** *Fix $\alpha \in A_{d-3}$. Then*

$$\mathbf{Pr}\left[|(\widehat{\boldsymbol{\tau}}_\alpha)^{-1}(*)| < w_{d-3} - w^{4/5}\right] \leq e^{-\Omega(\sqrt{w})}.$$

Proof. Recall from Definition 7 that $\widehat{\boldsymbol{\tau}}_{\alpha,i} = 0$ iff $\boldsymbol{\tau}_{\alpha,i} = \{0\}^{w_{d-2}}$ (in order for an OR to be 0, all its inputs must be 0). In turn, each coordinate of $\boldsymbol{\tau}_{\alpha,i}$ (we emphasize that $\boldsymbol{\tau}_{\alpha,i}$ is a string of length w) is an AND of the w coordinates of some $\boldsymbol{\rho}_a$ from (11), and hence is 0 with probability $1-\lambda-q$. By independence we have that

$$\mathbf{Pr}[\widehat{\boldsymbol{\tau}}_{\alpha,i} = 0] = \delta := (1-\lambda-q)^w \leq (1-q)^w \leq e^{-qw} \tag{25}$$

holds independently for all $i \in [w_{d-3}]$.

We next give an expression for $\mathbf{Pr}\left[\widehat{\boldsymbol{\tau}}_{\alpha,i} = 1\right]$. From Definition 7 we have that $\widehat{\boldsymbol{\tau}}_{\alpha,i} = 1$ iff any of the w coordinates of $\boldsymbol{\tau}_{\alpha,i}$ is 1 (in order for an OR to be 1, we only need one input to be 1). As noted above, each coordinate of $\boldsymbol{\tau}_{\alpha,i}$ is an AND of the w coordinates of some $\boldsymbol{\rho}_a$ from (11); this AND is 1 iff its input string is $\{1\}^w$, so by (11) each coordinate of $\boldsymbol{\tau}_{\alpha,i}$ is not 1 with probability $1-\lambda$. Hence all w coordinates of $\boldsymbol{\tau}_{\alpha,i}$ are not 1 with probability $(1-\lambda)^w$, and $\widehat{\boldsymbol{\tau}}_{\alpha,i} = 1$ with probability $1-(1-\lambda)^w$.

We thus have that, independently for all $i \in [w_{d-3}]$,

$$\mathbf{Pr}\left[\widehat{\boldsymbol{\tau}}_{\alpha,i} \in \{0,1\}\right] = \delta + (1-(1-\lambda)^w) \leq \delta + (1-(1-\lambda w)) \leq 2\lambda w = \frac{2(\log w)^{3/2}}{w^{1/4}},$$

where the last inequality holds (with room to spare) by (25). Applying Fact 5.1, we have that

$$\mathbf{Pr}\left[|\widehat{\boldsymbol{\tau}}_\alpha^{-1}(\{0,1\})| > w^{4/5}\right] \leq e^{-\Omega(\sqrt{w})}$$

with room to spare. □

Proof of Proposition 10.1. The proposition follows immediately from Lemmas 10.3 and 10.4 and a union bound over all $a \in A_{d-2}$ and $\alpha \in A_{d-3}$, using the fact that $|A_{d-3}| \leq |A_{d-2}| \leq n \leq w^{O(d)}$ and the bound $d \leq \frac{c \log w}{\log\log w}$. □

10.1.2 Preserving typicality: Proof of Proposition 10.2

The following numerical lemma relates q_a as defined in (13) of Definition 9 to q as defined in (7):

Lemma 10.5. *Let $2 \leq k \leq d-1$ and $S \subseteq [w_{k-1}]$ be k-acceptable (i.e. $|S| = qw \pm w^{\beta(k,d)}$), and define*

$$q' = \frac{(1-t_k)^{|S|} - \lambda}{t_{k-1}}.$$

Then $q' = q \cdot (1 \pm 2t_k w^{\beta(k,d)})$. (And in particular, by our bounds on t_k in Lemma 7.1 and the definition of $\beta(k,d)$, we have that $q' = q \pm o(q)$ for all k.)

Proof. For the lower bound, we have the following:

$$\begin{aligned}
q' &\le \frac{(1-t_k)^{qw-w^{\beta(k,d)}}-\lambda}{t_{k-1}} \\
&= \frac{(1-t_k)^{qw}-\lambda(1-t_k)^{w^{\beta(k,d)}}}{t_{k-1}(1-t_k)^{w^{\beta(k,d)}}} \\
&\le \frac{(1-t_k)^{qw}-\lambda}{t_{k-1}(1-t_k)^{w^{\beta(k,d)}}} + \frac{\lambda t_k w^{\beta(k,d)}}{t_{k-1}(1-t_k)^{w^{\beta(k,d)}}} \\
&= \frac{t_{k-1}q}{t_{k-1}(1-t_k)^{w^{\beta(k,d)}}} + \frac{\lambda t_k w^{\beta(k,d)}}{t_{k-1}(1-t_k)^{w^{\beta(k,d)}}} && \text{(by (8))} \\
&\le \frac{q}{1-t_k w^{\beta(k,d)}} + \frac{1+3q^{0.1}}{1-t_k w^{\beta(k,d)}} \cdot \lambda w^{\beta(k,d)} && \text{(by Lemma 7.1)} \\
&\le q \cdot (1+2t_k w^{\beta(k,d)}),
\end{aligned}$$

where for the last inequality we have used the fact that $qt_k = \tilde{\Theta}(w^{-1})$ whereas $\lambda = \tilde{\Theta}(w^{-5/4})$. For the upper bound, we have

$$\begin{aligned}
q' &\ge \frac{(1-t_k)^{qw+w^{\beta(k,d)}}-\lambda}{t_{k-1}} \\
&\ge \frac{(1-t_k)^{qw}(1-t_k w^{\beta(k,d)})-\lambda}{t_{k-1}} \\
&\ge q \cdot (1-t_k w^{\beta(k,d)}) - \frac{\lambda}{t_{k-1}} && \text{(by (8))} \\
&\ge q \cdot (1-2t_k w^{\beta(k,d)}).
\end{aligned}$$

where the last inequality uses the definition of λ in (7) and our bound on t_{k-1} in Lemma 7.1. □

Similar to the proof of Proposition 10.1, Proposition 10.2 follows from Lemmas 10.6 and 10.8 (stated and proved below) and a union bound, again using the fact that each $|A_i| \le n$ and the bound $d \le \frac{c \log w}{\log\log w}$. Since Proposition 10.2 deals with general values of k which may correspond to either row of Table 2, to avoid redundancy we use $\circ, \bullet$ notation in the statements and proofs of the following lemmas.

Lemma 10.6 (Condition (1) of typicality)**.** *For $2 \le k \le d-2$ let $\tau \in \{\bullet, \circ, *\}^{A_{k+1}}$ be typical and fix $a \in A_{k-1}$. Then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)}\left[|(\widehat{\boldsymbol{\rho}}_a)^{-1}(*)| = qw \pm w^{\beta(k,d)}\right] \ge 1 - \exp(-\tilde{\Omega}(w^{2\beta(k,d)-\frac{1}{2}})) \ge 1 - e^{-\Omega(w^{1/6})}.$$

(Recall that from Definition 14 that $\beta(k,d) = \frac{1}{3} + \frac{d-k-1}{12d}$).

Proof. Since $\tau \in \{\bullet, \circ, *\}^{A_{k+1}}$ is typical, we have that

$$\widehat{\tau}_a \in \{*, \circ\}^w \quad \text{and} \quad |(\widehat{\tau}_a)^{-1}(*)| \ge w - w^{4/5} \tag{26}$$

by the second and third property of τ being typical. Furthermore, for every $i \in [w]$ such that $\widehat{\tau}_{a,i} = *$, we have that

$$\tau_{a,i} \in \{*, \circ\}^w \qquad \text{and} \qquad qw - w^{\beta(k+1,d)} \leq |(\tau_{a,i})^{-1}(*)| \leq qw + w^{\beta(k+1,d)}, \tag{27}$$

by the first property of τ being typical. Writing $S_{a,i}$ for $(\tau_{a,i})^{-1}(*)$ (a subset of $[w]$) and S_a for $(\widehat{\tau}_a)^{-1}(*)$ (a subset of $[w]$), it follows from the second branch of (12) and Definition 7 that every $i \in S_a$ satisfies

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)} \left[\widehat{\boldsymbol{\rho}}_{a,i} = *\right] = q_{a,i} = \frac{(1-t)^{|S_{a,i}|} - \lambda}{t}.$$

Since $S_{a,i}$ is $(k+1)$-acceptable, by the $k+1$ case of Lemma 10.5 we have that

$$q_{a,i} = q \cdot (1 \pm 2t_{k+1} w^{\beta(k+1,d)}).$$

Since $|S_a| \leq w$, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)} \left[|(\widehat{\boldsymbol{\rho}}_a)^{-1}(*)|\right] = \sum_{i \in S_a} q_{a,i} \leq w \cdot q(1 + 2t_{k+1} w^{\beta(k+1,d)}) \leq qw + \tilde{O}(w^{\beta(k+1,d)}),$$

where the $\tilde{O}$ comes from the fact that $wt_{k+1}q = \Theta(\log w)$ (recalling Lemma 7.1 we have that $t_{k+1} = q \pm o(q)$). On the other hand, by (26) and similar reasoning we also have the lower bound

$$\mathop{\mathbf{E}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)} \left[|(\widehat{\boldsymbol{\rho}}_a)^{-1}(*)|\right] \geq (w - w^{4/5}) \cdot q(1 - 2t_{k+1} w^{\beta(k+1,d)}) \geq qw - \tilde{O}(w^{\beta(k+1,d)}),$$

where we have taken advantage of the fact that $w^{4/5}q = \tilde{O}(w^{0.3}) = o(w^{\beta(k+1,d)})$. Since $w^{\beta(k,d)} = \omega(\mathrm{polylog}(w) \cdot w^{\beta(k+1,d)})$ (here is where we are using the fact that $d \leq \frac{c \log w}{\log \log w}$), it follows from Fact 5.1 that

$$\begin{aligned} \mathop{\mathbf{Pr}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)} \left[|(\widehat{\boldsymbol{\rho}}_a)^{-1}(*)| \neq qw \pm w^{\beta(k,d)}\right] &\leq \exp\bigl(-\Omega(w^{2\beta(k,d)}/qw)\bigr) \\ &\leq \exp(-\tilde{\Omega}(w^{2\beta(k,d)-\frac{1}{2}})). \end{aligned}$$

□

Lemma 10.7. *Fix $2 \leq k \leq d-2$ and let $\tau \in \{\bullet, \circ, *\}^{A_{k+1}}$ be typical. For each $a \in A_{k-1}$ we write $S_a = S_a(\tau)$ to denote $(\widehat{\tau}_a)^{-1}(*)$ (note that this is a subset of $[w]$). Then for $\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)$, we have that $\widehat{\boldsymbol{\rho}}_a$ (which is a string in $\{\bullet, \circ, *\}^w$) satisfies:*

$$\begin{cases} \widehat{\boldsymbol{\rho}}_a = \{\circ\}^w & \textit{with probability } \prod_{i \in S_a}(1 - \lambda - q_{a,i}) \\ (\widehat{\boldsymbol{\rho}}_a)^{-1}(\bullet) \neq \emptyset & \textit{with probability } 1 - (1-\lambda)^{|S_a|} \\ \widehat{\boldsymbol{\rho}}_a \in \{\circ, *\}^w \setminus \{\circ\}^w & \textit{otherwise,} \end{cases}$$

independently for all $a \in A_{k-1}$. (Recall that $\widehat{\tau}_a \in \{*, \circ\}^w \setminus \{\circ\}^w$ for all $a \in A_{k-1}$ since τ is typical.) This implies that

$$\widehat{\widehat{\boldsymbol{\rho}}}_a = \begin{cases} \bullet & \textit{with probability } \prod_{i \in S_a}(1 - \lambda - q_{a,i}) \\ \circ & \textit{with probability } 1 - (1-\lambda)^{|S_a|} \\ * & \textit{otherwise} \end{cases}$$

independently for all $a \in A_{k-1}$. (Recall that $\widehat{\widehat{\tau}}_a = *$ for all $a \in A_{k-1}$ since τ is typical.)

Proof. The value of $\widehat{\boldsymbol{\rho}}_{a,i}$ is independent across all $a \in A_{k-1}$ and $i \in [w]$ such that $\widehat{\tau}_{a,i} = *$. Fix such a $a \in A_{k-1}$ and $i \in [w]$, and recall that

$$\tau_{a,i} \in \{*, \circ\}^w \setminus \{\circ\}^w.$$

By (12) and Definition 7 (the definition of the lift operator), we have that

$$\widehat{\boldsymbol{\rho}}_{a,i} = \begin{cases} \bullet & \text{with probability } \lambda \\ * & \text{with probability } q_{a,i} \\ \circ & \text{otherwise, with probability } 1 - \lambda - q_{a,i}. \end{cases}$$

The lemma then follows by independence. □

Remark 15. If $\tau \in \{\bullet, \circ, *\}^{A_{k+1}}$ is typical then (recall that $S_a = (\widehat{\tau}_a)^{-1}(*)$ is a subset of $[w]$ and $S_{a,i} = (\tau_{a,i})^{-1}(*)$ is a subset of $[w]$) we have

$$|S_a| \geq w - w^{4/5} \qquad \text{and} \qquad qw - w^{\beta(k+1,d)} \leq |S_{a,i}| \leq qw + w^{\beta(k+1,d)} \text{ for all } i \in S_a.$$

Therefore we have the estimates

$$\mathbf{Pr}\left[\widehat{\widehat{\boldsymbol{\rho}}}_a = \bullet\right] = \prod_{i \in S_a} (1 - \lambda - q_{a,i}) \leq (1 - q_{a,i})^{w - w^{4/5}} \leq \left(1 - \tfrac{q}{2}\right)^{w - w^{4/5}} \leq e^{-qw/4} = e^{-\Omega(\sqrt{w \log w})},$$

where we have used Lemma 10.5 for the second inequality, and

$$\mathbf{Pr}\left[\widehat{\widehat{\boldsymbol{\rho}}}_a = \circ\right] = 1 - (1-\lambda)^{|S_a|} \leq 1 - (1-\lambda)^w \leq 1 - (1 - \lambda w) = \lambda w.$$

Lemma 10.8 (Condition (2) of typicality)**.** *For* $2 \leq k \leq d-2$ *let* $\tau \in \{\bullet, \circ, *\}^{A_{k+1}}$ *be typical and fix* $\alpha \in A_{k-2}$*. Then*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)}\left[\left|\left(\widehat{\widehat{\boldsymbol{\rho}}}_\alpha\right)^{-1}(*)\right| \geq w_{k-2} - w^{4/5}\right] = 1 - e^{-\Omega(\sqrt{w})}.$$

Proof. By Lemma 10.7 and the two estimates of Remark 15, each coordinate of $(\widehat{\widehat{\boldsymbol{\rho}}})_\alpha$ is independently in $\{\bullet, \circ\}$ with probability at most $e^{-\Omega(\sqrt{w})} + \lambda w = O\left(\frac{(\log w)^{3/2}}{w^{1/4}}\right)$. Hence the expected size of $\left|\left(\widehat{\widehat{\boldsymbol{\rho}}}_\alpha\right)^{-1}(\{\bullet, \circ\})\right|$ is $\tilde{O}(w^{3/4})$, and we may apply Fact 5.1 to get that

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)}\left[\left|\left(\widehat{\widehat{\boldsymbol{\rho}}}_\alpha\right)^{-1}(\{\bullet, \circ\})\right| > w^{4/5}\right] \leq e^{-\Omega(\sqrt{w})}$$

with room to spare. □

10.2 Sipser survives random projections

In this subsection we prove the main results of Section 10; these are two results which show, in different ways, that the Sipser_d function "retains structure" after being hit with the random projection $\boldsymbol{\Psi}$. The first of these results, Proposition 10.11, gives a useful characterization of $\boldsymbol{\Psi}(\mathsf{Sipser}_d)$ by showing that it is distributed identically to a (suitably randomly restricted) *depth-one* formula. The second of these results, Proposition 10.13, shows that this randomly restricted depth-one formula is very close to perfectly balanced in expectation. Our later arguments will use both these types of structure.

10.2.1 Sipser_d reduces under Ψ to a random restriction of $\mathsf{Sipser}_d^{(1)}$

Recalling the definitions of the depth-k $\mathsf{Sipser}_d^{(k)}$ formulas from Definition 5, we begin with the following observation regarding the effect of projections on the $\mathsf{Sipser}_d^{(k)}$ formulas:

Fact 10.9. *For $2 \leq k \leq d$ we have that*

$$\mathrm{proj}\,\mathsf{Sipser}_d^{(k)} \equiv \mathsf{Sipser}_d^{(k-1)}.$$

In words, Fact 10.9 says that the projection operator "wipes out" the bottom-layer gates of $\mathsf{Sipser}_d^{(k)}$, reducing its depth by exactly one. Fact 10.9 is a straightforward consequence of the definitions of projections and the $\mathsf{Sipser}_d^{(k)}$ formulas (Definitions 4 and 5 respectively), but is perhaps most easily seen to be true via the equivalently view of projections described in Remark 9: for every bottom-layer gate $a \in A_k$ of $\mathsf{Sipser}_d^{(k)}$, the projection operator simply replaces every one of its w_{k-1} formal input variables $x_{a,1}, \ldots, x_{a,w_{k-1}}$ with the same fresh formal variable y_a. Since $\mathsf{AND}(y_a, \ldots, y_a) \equiv \mathsf{OR}(y_a, \ldots, y_a) \equiv y_a$, the gate simplifies to the single variable y_a. (Indeed, we defined our projection operators precisely so that they sync up with $\mathsf{Sipser}_d^{(k)}$ this way.)

The same reasoning, along with the definition of lifts (see Definition 7 and the discussion after), yields the following extension of Fact 10.9:

Fact 10.10. *For $2 \leq k \leq d$ and $\rho \in \{0, 1, *\}^{A_k}$ we have*

$$\mathrm{proj}_\rho\,\mathsf{Sipser}_d^{(k)} \equiv \mathsf{Sipser}_d^{(k-1)} \upharpoonright \widehat{\rho}.$$

Remark 16. With Fact 10.10 in hand we now revisit our definition of typical restrictions (recall Definition 14 and the discussion thereafter). Recall that the high-level rationale behind this definition is that for ρ such that $\widehat{\rho}$ is typical, the projection proj_ρ has a "very limited and well-controlled effect" on the target Sipser_d. We now make this statement more precise (the reader may find it helpful to refer to the illustration in Figure 2).

Fix $\rho \in \mathrm{supp}(\mathcal{R}_{\mathrm{init}})$ such that $\widehat{\rho}$ is typical. By Fact 10.10, we have that

$$\mathrm{proj}_\rho\,\mathsf{Sipser}_d \equiv \mathsf{Sipser}_d^{(d-1)} \upharpoonright \widehat{\rho}.$$

Since $\widehat{\rho}$ is typical,

- The first condition of Definition 14 implies that $|(\widehat{\rho}_a)^{-1}(*)| = \Theta(qw) = \tilde{\Theta}(\sqrt{w})$ for all $a \in A_{d-2}$. Each such $a \in A_{d-2}$ is the address of an OR gate, and so if $(\widehat{\rho}_a)^{-1}(1) \neq \emptyset$ the gate is satisfied and evaluates to 1, and otherwise if $\widehat{\rho}_a \in \{*, 0\}^w$ the value of the gate remains undetermined (i.e. it "evaluates to $*$") and its fan-in becomes $|(\widehat{\rho}_a)^{-1}(*)| = \tilde{\Theta}(\sqrt{w})$.
- The second condition of Definition 14 tells us that between the two possibilities above, the latter is far more common: for every $\alpha \in A_{d-3}$ specifying a block of w_{d-3} many OR gates, at most $w^{4/5}$ of these gates evaluate to 1 and the remaining (vast majority) are undetermined. Equivalently, *all* the AND gates at level $d-3$ remain undetermined, and they all have fan-in at least $w_{d-3} - w^{4/5} = w_{d-3}\,(1 - o(1))$.

The same description holds for $\mathrm{proj}_{\rho^{(k)}}$ and $\mathsf{Sipser}_d^{(k)}$. For $\widehat{\rho^{(k)}}$'s that are typical the projection operator $\mathrm{proj}_{\rho^{(k)}}$:

- "wipes out" the bottom-level (level-k) gates of $\mathsf{Sipser}_d^{(k)}$,
- "trims" the fan-ins of the level-$(k-1)$ gates from w to $\tilde{\Theta}(\sqrt{w})$,
- keeps the fan-ins of all level-$(k-2)$ gates at least $w_{k-2} - w^{4/5} = w_{k-2}(1 - o(1))$.

Note in particular that the entire structure of the formula from levels 0 through $k-3$ is identical to that of $\mathsf{Sipser}_d^{(k)}$, and so $\mathrm{proj}_{\rho^{(k)}} \mathsf{Sipser}_d^{(k)}$ "contains a perfect copy of" $\mathsf{Sipser}_d^{(k-3)}$.

Repeated applications of Fact 10.10 gives us the following proposition. (The proposition is intuitively very useful since, it tells us that in order to understand the effect of the random projection $\boldsymbol{\Psi}$ on the (relatively complicated) Sipser_d function, it suffices to analyze the effect of the random restriction $\widehat{\boldsymbol{\rho}^{(2)}}$ on the (much simpler) $\mathsf{Sipser}_d^{(1)}$ function; we will apply it in the final proof of each of our main lower bounds.)

Proposition 10.11. *Consider* $\mathsf{Sipser}_d : \{0,1\}^n \to \{0,1\}$. *Then*

$$\boldsymbol{\Psi}(\mathsf{Sipser}_d) \equiv \mathsf{Sipser}_d^{(1)} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}}.$$

Proof. By Fact 10.10 we have that

$$\mathrm{proj}_{\rho^{(d)}} \mathsf{Sipser}_d \equiv \mathsf{Sipser}_d^{(d-1)} \upharpoonright \widehat{\rho^{(d)}} \tag{28}$$

for all $\rho^{(d)} \in \mathrm{supp}(\mathcal{R}_{\mathrm{init}}) \equiv \{0,1,*\}^n$. Furthermore for $\rho^{(k+1)} \in \{0,1,*\}^{A_{k+1}}$ and $\rho^{(k)} \in \mathrm{supp}(\mathcal{R}(\widehat{\rho^{(k+1)}})) \subseteq \{0,1,*\}^{A_k}$ we have

$$\begin{aligned}
\mathrm{proj}_{\rho^{(k)}} \left(\mathsf{Sipser}_d^{(k)} \upharpoonright \widehat{\rho^{(k+1)}}\right) &\equiv \mathrm{proj}\left(\left(\mathsf{Sipser}_d^{(k)} \upharpoonright \widehat{\rho^{(k+1)}}\right) \upharpoonright \rho^{(k)}\right) \\
&\equiv \mathrm{proj}\left(\mathsf{Sipser}_d^{(k)} \upharpoonright \rho^{(k)}\right) \\
&\equiv \mathsf{Sipser}_d^{(k-1)} \upharpoonright \widehat{\rho^{(k)}}, \qquad (29)
\end{aligned}$$

where the first equivalence is by the definition of ρ-projection (Definition 4), the second is by the fact that $\mathcal{R}(\widehat{\rho^{(k+1)}})$ is supported on refinements of $\widehat{\rho^{(k+1)}}$ (and in particular, $\rho^{(k)}$ refines $\widehat{\rho^{(k+1)}}$), and the last is Fact 10.10. The proposition follows from (28), repeated application of (29), and the definition of $\boldsymbol{\Psi}$ (Definition 10). □

10.2.2 Sipser_d remains unbiased after random projection by $\boldsymbol{\Psi}$

Recall that $\mathsf{Sipser}_d^{(1)}$ denotes the function computed by the top gate of Sipser_d, and in particular, $\mathsf{Sipser}_d^{(1)}$ is a w_0-way OR if d is even, and a w_0-way AND if d is odd (c.f. Definition 5). In this subsubsection we will assume that d is even; the argument for odd values of d follows via a symmetric argument.

To obtain our ultimate results we will need a lower bound on the bias of $\boldsymbol{\Psi}(\mathsf{Sipser}_d)$ under $\mathbf{Y}$ (or equivalently, by the preceding proposition, on the bias of $\mathsf{Sipser}_d^{(1)} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}}$ where $\boldsymbol{\rho}^{(2)}$ is distributed as described in Definition 10). The following lemma will help us establish such a lower bound:

Lemma 10.12. *Let $\tau \in \{0,1,*\}^{A_2}$ be typical. Then for $\boldsymbol{\rho} \leftarrow \mathcal{R}(\tau)$ and $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$ we have*

$$\mathop{\mathbf{E}}_{\boldsymbol{\rho}}\left[\mathrm{bias}(\mathsf{Sipser}_d^{(1)} \restriction \widehat{\boldsymbol{\rho}}, \mathbf{Y})\right] \geq \frac{1}{2} - \tilde{O}(w^{-1/12}).$$

Proof. By our assumption that d is even we may write OR_{w_0} in place of $\mathsf{Sipser}_d^{(1)}$. Since τ is typical, we have by Conditions (2) and (3) of Definition 14 that

$$\widehat{\tau} \in \{0,*\}^{w_0} \quad \text{and} \quad |(\widehat{\tau})^{-1}(*)| \geq w_0 - w^{4/5}.$$

Furthermore, by (12) of Definition 9 and Definition 7 (the definition of the lift operator), we have that

$$\widehat{\boldsymbol{\rho}}_i = \begin{cases} 1 & \text{with probability } \lambda \\ * & \text{with probability } q_i \\ 0 & \text{otherwise, with probability } 1-\lambda-q_i \end{cases} \tag{30}$$

independently for all $i \in (\widehat{\tau})^{-1}(*) \subseteq [w_0]$, where

$$q_i = \frac{(1-t_2)^{|S_i|} - \lambda}{t_1}$$

and $S_i = S_i(\tau) = \tau_i^{-1}(*) = \{j \in [w_1] : \tau_{i,j} = *\}$ satisfies $|S_i| = qw \pm w^{\beta(2,d)}$. By a calculation very similar to the one that was employed in the proof of Lemma 10.6, we have that

$$\mathbf{Pr}\left[|(\widehat{\boldsymbol{\rho}})^{-1}(*)| = qw_0 \pm w^{\beta(1,d)}\right] \geq 1 - e^{-\Omega(w^{1/6})}. \tag{31}$$

Furthermore, (30) also implies that

$$\mathbf{Pr}[\widehat{\boldsymbol{\rho}} \in \{0,*\}^{w_0}] = (1-\lambda)^{|\widehat{\tau}^{-1}(*)|} \geq (1-\lambda)^{w_0} \geq 1 - \lambda w_0 = 1 - \tilde{O}(w^{-1/4}). \tag{32}$$

Fix any $\rho \in \mathrm{supp}(\mathcal{R}(\tau))$ that satisfies the events of both (31) and (32). Writing $S(\widehat{\rho}) \subseteq [w_0]$ to denote the set $(\widehat{\rho})^{-1}(*)$, we have the bounds

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \restriction \widehat{\rho})(\mathbf{Y}) = 0] &= (1-t_1)^{|S(\widehat{\rho})|} \\
&\geq (1-t_1)^{qw_0 + w^{\beta(1,d)}} \\
&\geq \left(\frac{1}{2} - \Theta\left(\frac{\log w}{w}\right)\right)(1-t_1)^{w^{\beta(1,d)}} \\
&\geq \left(\frac{1}{2} - \Theta\left(\frac{\log w}{w}\right)\right)(1 - t_1 w^{\beta(1,d)}), \\
&\geq \frac{1}{2} - \tilde{O}(w^{-1/12}),
\end{aligned}$$

where the second inequality crucially uses the definition (4) of w_0 and its corollary (9). Similarly,

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \restriction \widehat{\rho})(\mathbf{Y}) = 0] &= (1-t_1)^{|S(\widehat{\rho})|} \\
&\leq (1-t_1)^{qw_0 - w^{\beta(1,d)}} \\
&\leq \frac{1}{2} \cdot (1-t_1)^{-w^{\beta(1,d)}} \\
&\leq \frac{1}{2} + \tilde{O}(w^{-1/12}),
\end{aligned}$$

which establishes the lemma. □

Now we are ready to lower bound the expected bias of $\mathbf{\Psi}(\mathsf{Sipser}_d)$ (or equivalently, of $\mathsf{Sipser}_d^{(1)} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}}$) under $\mathbf{Y}$:

Proposition 10.13. *For* $\mathbf{\Psi}$ *as defined in Definition 10,*

$$\mathbf{\Psi}(f) \equiv \mathrm{proj}_{\boldsymbol{\rho}^{(2)}}\, \mathrm{proj}_{\boldsymbol{\rho}^{(3)}} \cdots \mathrm{proj}_{\boldsymbol{\rho}^{(d-1)}}\, \mathrm{proj}_{\boldsymbol{\rho}^{(d)}}\, f$$

where $\boldsymbol{\rho}^{(d)} \leftarrow \mathcal{R}_{\mathrm{init}}$ *and* $\boldsymbol{\rho}^{(k)} \leftarrow \mathcal{R}(\widehat{\boldsymbol{\rho}^{(k+1)}})$ *for all* $2 \le k \le d-1$*, and for* $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$*, we have that*

$$\mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathrm{bias}(\mathsf{Sipser}_d^{(1)} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}}, \mathbf{Y})\right] \ge \frac{1}{2} - \tilde{O}(w^{-1/12}).$$

Proof. By Proposition 10.1 and $d-3$ successive applications of Proposition 10.2, we have that

$$\mathbf{Pr}\left[\widehat{\boldsymbol{\rho}^{(d)}}, \ldots, \widehat{\boldsymbol{\rho}^{(3)}} \text{ are all typical}\right] \ge 1 - d \cdot e^{-\tilde{\Omega}(w^{1/6})}.$$

For every typical $\widehat{\rho^{(3)}} \in \{0,1,*\}^{A_2}$, Lemma 10.12 gives that

$$\mathop{\mathbf{E}}_{\boldsymbol{\rho}^{(2)} \leftarrow \mathcal{R}(\widehat{\rho^{(3)}})}\left[\mathrm{bias}(\mathsf{Sipser}_d^{(1)} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}}, \mathbf{Y})\right] \ge \frac{1}{2} - \tilde{O}(w^{-1/12}),$$

which together with the preceding inequality gives the proposition. □

Remark 17. We note that combining Proposition 10.11 and Proposition 10.13, for $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$ we have that

$$\mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathrm{bias}(\mathbf{\Psi}(\mathsf{Sipser}_d), \mathbf{Y})\right] \ge \frac{1}{2} - \tilde{O}(w^{-1/12}),$$

which we may rewrite as

$$\mathbf{Pr}[(\mathbf{\Psi}(\mathsf{Sipser}_d))(\mathbf{Y}) = 0] = \mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathop{\mathbf{Pr}}_{\mathbf{Y}}(\mathbf{\Psi}(\mathsf{Sipser}_d))(\mathbf{Y}) = 0]\right] = \frac{1}{2} \pm \tilde{O}(w^{-1/12}).$$

Applying Proposition 8.1, we get that for $\mathbf{X} \leftarrow \{0_{1/2}, 1_{1/2}\}^n$ we have

$$\mathbf{Pr}[\mathsf{Sipser}_d(\mathbf{X}) = 1] = \frac{1}{2} \pm \tilde{O}(w^{-1/12}).$$

verifying (6) in Section 6: the Sipser_d function is indeed (essentially) balanced.

11 Proofs of main theorems

Recall that $\mathsf{Sipser}_d^{(1)}$ denotes the function computed by the top gate of Sipser_d, and in particular, $\mathsf{Sipser}_d^{(1)}$ is a w_0-way OR if d is even, and a w_0-way AND if d is odd (c.f. Definition 5). Throughout this section we will assume that d is even; the argument for odd values of d follows via a symmetric argument. For conciseness we will sometimes write OR_{w_0} in place of $\mathsf{Sipser}_d^{(1)}$ in the arguments below; we stress that these are the same function.

11.1 "Bottoming out" the argument

As we will see in the proofs of Theorems 6 and 7, the machinery we have developed enables us to relate the correlation between Sipser_d and the circuits C against which we are proving lower bounds, to the correlation between $\mathsf{Sipser}_d^{(1)} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}}$ (obtained by hitting Sipser_d with the random projection $\mathbf{\Psi}$) and bounded-width CNFs (that are similarly obtained by hitting C with $\mathbf{\Psi}$). To finish the argument, we need to bound the correlation between $\mathsf{Sipser}_d^{(1)} \upharpoonright \tau$ (for suitable restrictions τ) and such CNFs. The following proposition, which is a slight extension of Lemma 4.1 of [OW07], enables us to do this, by relating the correlation between $\mathsf{Sipser}_d^{(1)} \upharpoonright \tau$ and such CNFs to the bias of $\mathsf{Sipser}_d^{(1)} \upharpoonright \tau$.

Proposition 11.1. *Let $F : \{0,1\}^{w_0} \to \{0,1\}$ be a width-r CNF and $\tau \in \{0,*\}^{w_0} \setminus \{0\}^{w_0}$. Then for $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$,*

$$\mathbf{Pr}[(\mathsf{OR}_{w_0} \upharpoonright \tau)(\mathbf{Y}) \neq F(\mathbf{Y})]] \geq \mathrm{bias}(\mathsf{OR}_{w_0} \upharpoonright \tau, \mathbf{Y}) - rt_1.$$

Proof. Writing $S = S(\tau) \subseteq [w_0]$ to denote the set $\tau^{-1}(*)$, we have that $\mathsf{OR}_{w_0} \upharpoonright \tau$ computes the $|S|$-way OR of variables with indices in S (note that $S \neq \emptyset$ since $\tau \in \{0,*\}^{w_0} \setminus \{0\}^{w_0}$); for notational brevity we will write OR_S instead of $\mathsf{OR}_{w_0} \upharpoonright \tau$.

We begin with the claim that there exists a CNF $F' : \{0,1\}^{w_0} \to \{0,1\}$ of size and width at most that of F, depending only on the variables in S, such that

$$\mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) \neq F(\mathbf{Y})] \geq \mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) \neq F'(\mathbf{Y})]. \tag{33}$$

This holds because

$$\begin{aligned}\mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) \neq F(\mathbf{Y})] &= \mathop{\mathbf{E}}_{\boldsymbol{\rho} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{[w_0]\setminus S}} \Big[\mathbf{Pr}[(\mathsf{OR}_S \upharpoonright \boldsymbol{\rho})(\mathbf{Y}) \neq (F \upharpoonright \boldsymbol{\rho})(\mathbf{Y})]\Big] \\ &= \mathop{\mathbf{E}}_{\boldsymbol{\rho} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{[w_0]\setminus S}} \Big[\mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) \neq (F \upharpoonright \boldsymbol{\rho})(\mathbf{Y})]\Big],\end{aligned}$$

and so certainly there exists $\rho \in \{0,1\}^{[w_0]\setminus S}$ such that $F' := F \upharpoonright \rho$ satisfies (33). Next, writing $\{y_i\}_{i \in S}$ to denote the formal variables that both OR_S and F' depend on, we consider two possible cases:

1. For every clause T in F' there exists $i \in S$ such that $\overline{y}_i$ occurs in T. In this case we note that $F'(0^S) = 1$ (whereas $\mathsf{OR}_S(0^S) = 0$), and so

$$\mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) \neq F'(\mathbf{Y})] \geq \mathbf{Pr}[\mathbf{Y}_i = 0 \text{ for all } i \in S] = \mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) = 0].$$

2. Otherwise, there must exist a monotone clause T in F' (one containing only positive occurrences of variables) since F' depends only on the variables in S. In this case, since each unnegated literal is true with probability t_1 (recall that $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$) and T has width at most r, by a union bound we have that

$$\mathbf{Pr}[F'(\mathbf{Y}) = 1] \leq \mathbf{Pr}[T(\mathbf{Y}) = 1] \leq rt_1,$$

and so

$$\mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) \neq F'(\mathbf{Y})] \geq \mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) = 1] - \mathbf{Pr}[F'(\mathbf{Y}) = 1] \geq \mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) = 1] - rt_1.$$

Together, theses two cases give us the lower bound

$$\begin{aligned}\mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) \neq F'(\mathbf{Y})] &\geq \min\big\{\,\mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) = 1], \mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) = 0] - rt_1\big\}\\ &\geq \min\big\{\,\mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) = 1], \mathbf{Pr}[\mathsf{OR}_S(\mathbf{Y}) = 0]\big\} - rt_1,\end{aligned}$$

which along with (33) completes the proof. □

11.2 Approximators with small bottom fan-in

The pieces are in place to prove the first of our two main theorems, showing that Sipser_d cannot be approximated by depth-d size-S circuits with bounded bottom fan-in:

Theorem 6. *For $2 \leq d \leq \frac{c\sqrt{\log n}}{\log\log n}$, the n-variable Sipser_d function has the following property: Let $C : \{0,1\}^n \to \{0,1\}$ be any depth-d circuit of size $S = 2^{n^{\frac{1}{6(d-1)}}}$ and bottom fan-in $\frac{\log n}{10(d-1)}$. Then for a uniform random input $\mathbf{X} \leftarrow \{0_{1/2}, 1_{1/2}\}^n$, we have*

$$\mathbf{Pr}[\mathsf{Sipser}_d(\mathbf{X}) \neq C(\mathbf{X})] \geq \frac{1}{2} - \frac{1}{n^{\Omega(1/d)}}.$$

Proof. Let $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$. We successively apply Proposition 8.1 and Proposition 10.11 to obtain

$$\begin{aligned}\mathbf{Pr}[\mathsf{Sipser}_d(\mathbf{X}) \neq C(\mathbf{X})] &= \mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathbf{\Psi}(\mathsf{Sipser}_d))(\mathbf{Y}) \neq (\mathbf{\Psi}(C))(\mathbf{Y})]\right]\\ &= \mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}})(\mathbf{Y}) \neq (\mathbf{\Psi}(C))(\mathbf{Y})]\right]\end{aligned}$$

(for the second equality, recall that $\mathsf{Sipser}_d^{(1)}$ is simply OR_{w_0}, by our assumption from the start of the section that d is even). For every possible outcome Ψ of $\mathbf{\Psi}$ (corresponding to successive outcomes of $\rho^{(d)}$ for $\boldsymbol{\rho}^{(d)}$, ..., $\rho^{(2)}$ for $\boldsymbol{\rho}^{(2)}$) and every $r \in \mathbb{N}$, we have the bound

$$\begin{aligned}&\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \upharpoonright \widehat{\rho^{(2)}})(\mathbf{Y}) \neq (\Psi(C))(\mathbf{Y})]\\ &\geq \mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \upharpoonright \widehat{\rho^{(2)}})(\mathbf{Y}) \neq (\Psi(C))(\mathbf{Y}) \mid \Psi(C) \text{ is a depth-}r \text{ DT}] - \mathbf{1}[\Psi(C) \text{ is not a depth-}r\text{ DT}]\\ &\geq \mathrm{bias}(\mathsf{OR}_{w_0} \upharpoonright \widehat{\rho^{(2)}}, \mathbf{Y}) - rt_1 - \mathbf{1}[\Psi(C) \text{ is not a depth-}r\text{ DT}],\end{aligned}$$

where the final inequality is by Proposition 11.1 along with the fact that every depth-r DT can be expressed as either a width-r CNF or a width-r DNF. Setting $r = n^{\frac{1}{4(d-1)}}$ and taking expectation with respect to $\mathbf{\Psi}$, we conclude that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}})(\mathbf{Y}) \neq (\Psi(C))(\mathbf{Y})]\right] &\geq \mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathrm{bias}(\mathsf{OR}_{w_0} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}}, \mathbf{Y})\right] - rt_1 - \mathop{\mathbf{Pr}}_{\mathbf{\Psi}}[\mathbf{\Psi}(C) \text{ is not a depth-}r\text{ DT}]\\ &\geq \frac{1}{2} - \tilde{O}(w^{-1/12}) - rt_1 - \exp\left(-\Omega(n^{\frac{1}{6(d-1)}})\right)\\ &\geq \frac{1}{2} - \frac{1}{n^{\Omega(1/d)}},\end{aligned}$$

where the second-to-last inequality uses both Proposition 10.13 and Theorem 13, and the last claim follows by simple substitution, recalling the values of r, t_1 and w in terms of n and d. □

11.3 Approximators with the opposite alternation pattern

Our second main theorem states that Sipser_d cannot be approximated by depth-d size-S circuits with the opposite alternation pattern to Sipser_d:

Theorem 7. *For $2 \le d \le \frac{c\sqrt{\log n}}{\log\log n}$, the n-variable Sipser_d function has the following property: Let $C : \{0,1\}^n \to \{0,1\}$ be any depth-d circuit of size $S = 2^{n^{\frac{1}{6(d-1)}}}$ and the opposite alternation pattern to Sipser_d, (i.e. its top-level gate is OR if Sipser_d's is AND and vice versa). Then for a uniform random input $\mathbf{X} \leftarrow \{0_{1/2}, 1_{1/2}\}^n$, we have*

$$\mathbf{Pr}[\mathsf{Sipser}_d(\mathbf{X}) \neq C(\mathbf{X})] \ge \frac{1}{2} - \frac{1}{n^{\Omega(1/d)}}.$$

Proof. By our assumption that d is even, we have that the top gate of Sipser_d is a w_0-way OR, whereas the top gate of C is an AND. Let $\mathbf{Y} \leftarrow \{0_{1-t_1}, 1_{t_1}\}^{w_0}$. As in the proof of Theorem 6, we successively apply Proposition 8.1 and Proposition 10.11 to obtain

$$\begin{aligned}\mathbf{Pr}[\mathsf{Sipser}_d(\mathbf{X}) \neq C(\mathbf{X})] &= \mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathbf{\Psi}(\mathsf{Sipser}_d))(\mathbf{Y}) \neq (\mathbf{\Psi}(C))(\mathbf{Y})]\right] \\ &= \mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}})(\mathbf{Y}) \neq (\mathbf{\Psi}(C))(\mathbf{Y})]\right].\end{aligned}$$

For every possible outcome $\Psi = \rho^{(d)}, \ldots, \rho^{(2)}$ of $\mathbf{\Psi}$ and every $r \in \mathbb{N}$ we have the bound

$$\begin{aligned}&\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \upharpoonright \widehat{\rho^{(2)}})(\mathbf{Y}) \neq (\Psi(C))(\mathbf{Y})] \\ &\ge \mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \upharpoonright \widehat{\rho^{(2)}})(\mathbf{Y}) \neq (\Psi(C))(\mathbf{Y}) \mid \Psi(C) \text{ is } (1/S)\text{-close to a width-}r \text{ CNF}] \\ &\qquad - \mathbf{1}[\Psi(C) \text{ is not } (1/S)\text{-close to a width-}r \text{ CNF}] \\ &\ge \mathrm{bias}(\mathsf{OR}_{w_0} \upharpoonright \widehat{\rho^{(2)}}, \mathbf{Y}) - rt_1 - (1/S) - \mathbf{1}[\Psi(C) \text{ is not } (1/S)\text{-close to a width-}r \text{ CNF}],\end{aligned}$$

where the final inequality is by Proposition 11.1. As in the proof of Theorem 6, setting $r = n^{\frac{1}{4(d-1)}}$ and taking expectation with respect to $\mathbf{\Psi}$, we conclude that

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathop{\mathbf{Pr}}_{\mathbf{Y}}[(\mathsf{OR}_{w_0} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}})(\mathbf{Y}) \neq (\Psi(C))(\mathbf{Y})]\right] \\ &\ge \mathop{\mathbf{E}}_{\mathbf{\Psi}}\left[\mathrm{bias}(\mathsf{OR}_{w_0} \upharpoonright \widehat{\boldsymbol{\rho}^{(2)}}, \mathbf{Y})]\}\right] - (1/S) - rt_1 - \mathop{\mathbf{Pr}}_{\mathbf{\Psi}}[\mathbf{\Psi}(C) \text{ is not } (1/S)\text{-close to a width-}r \text{ CNF}] \\ &\ge \frac{1}{2} - \tilde{O}(w^{-1/12}) - (1/S) - rt_1 - \exp\left(-\Omega(n^{\frac{1}{6(d-1)}})\right) \\ &\ge \frac{1}{2} - \frac{1}{n^{\Omega(1/d)}},\end{aligned}$$

where the second-to-last inequality uses both Proposition 10.13 and Theorem 14, and the last claim follows by simple substitution, recalling the values of r, t_1, w and S in terms of n and d. □

References

[Aar] Scott Aaronson. The Complexity Zoo. Available at `http://cse.unl.edu/~cbourke/latex/ComplexityZoo.pdf`.

[Aar10a] Scott Aaronson. A counterexample to the generalized Linial-Nisan conjecture. *Electronic Colloquium on Computational Complexity*, 17:109, 2010.

[Aar10b] Scott Aaronson. BQP and the polynomial hierarchy. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, pages 141–150, 2010.

[AB09] Sanjeev Arora and Boaz Barak. *Computational Complexity: a modern approach*. Cambridge University Press, 2009.

[Ajt83] Miklós Ajtai. Σ^1_1-formulae on finite structures. *Annals of Pure and Applied Logic*, 24(1):1–48, 1983.

[Ajt94] Miklós Ajtai. The independence of the modulo p counting principles. In *Proceedings of the 26th Annual ACM Symposium on Theory of Computing*, pages 402–411, 1994.

[AWY15] Amir Abboud, Ryan Williams, and Huacheng Yu. More applications of the polynomial method to algorithm design. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2015.

[Bab87] László Babai. Random oracles separate PSPACE from the polynomial-time hierarchy. *Information Processing Letters*, 26(1):51–53, 1987.

[Baz09] Louay Bazzi. Polylogarithmic independence can fool DNF formulas. *SIAM Journal on Computing*, 38(6):2220–2272, 2009.

[Bea94] Paul Beame. A switching lemma primer. Technical Report UW-CSE-95-07-01, University of Washington, 1994.

[BG81] Charles Bennett and John Gill. Relative to a random oracle A, $\mathsf{P}^A \neq \mathsf{NP}^A \neq \mathsf{coNP}^A$ with probability 1. *SIAM Journal on Computing*, 10(1):96–113, 1981.

[BGS75] Theodore Baker, John Gill, and Robert Solovay. Relativizations of the $\mathsf{P}{=}?\mathsf{NP}$ question. *SIAM Journal on computing*, 4(4):431–442, 1975.

[BIS12] Paul Beame, Russell Impagliazzo, and Srikanth Srinivasan. Approximating $\mathsf{AC^0}$ by small height decision trees and a deterministic algorithm for $\#\mathsf{AC^0}$-SAT. In *Proceedings of the 27th Conference on Computational Complexity*, pages 117–125, 2012.

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, 90:5–43, 1999.

[Boo94] Ronald Book. On collapsing the polynomial-time hierarchy. *Information Processing Letters*, 52(5):235–237, 1994.

[Bop97] Ravi Boppana. The average sensitivity of bounded-depth circuits. *Information Processing Letters*, 63(5):257–261, 1997.

[Bra10] Mark Braverman. Polylogarithmic independence fools AC^0 circuits. *Journal of the ACM*, 57(5):28, 2010.

[BS79] Theodore Baker and Alan Selman. A second step toward the polynomial hierarchy. *Theoretical Computer Science*, 8(2):177–187, 1979.

[BT96] Nader Bshouty and Christino Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.

[Cai86] Jin-Yi Cai. With probability one, a random oracle separates PSPACE from the polynomial-time hierarchy. In *Proceedings of the 18th Annual ACM Symposium on Theory of Computing*, pages 21–29, 1986.

[CCH98] Liming Cai, Jianer Chen, and Johan Håstad. Circuit bottom fan-in and computational power. *SIAM Journal on Computing*, 27(2):341–355, 1998.

[DK00] Ding-Zhu Du and Ker-I Ko. *Theory of Computational Complexity*. John Wiley & Sons, Inc., 2000.

[For99] Lance Fortnow. Relativized worlds with an infinite hierarchy. *Information Processing Letters*, 69(6):309–313, 1999.

[FSS81] Merrick Furst, James Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. In *Proceedings of the 22nd IEEE Annual Symposium on Foundations of Computer Science*, pages 260–270, 1981.

[GW13] Oded Goldreich and Avi Wigderson. On the size of depth-three Boolean circuits for computing multilinear functions. *Electronic Colloquium on Computational Complexity*, 2013.

[Hås86a] Johan Håstad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the 18th Annual ACM Symposium on Theory of Computing*, pages 6–20, 1986.

[Hås86b] Johan Håstad. *Computational Limitations for Small Depth Circuits*. MIT Press, Cambridge, MA, 1986.

[Hås89] Johan Håstad. *Almost optimal lower bounds for small depth circuits*, pages 143–170. Advances in Computing Research, Vol. 5. JAI Press, 1989.

[Hås14] Johan Håstad. On the correlation of parity and small-depth circuits. *SIAM Journal on Computing*, 43(5):1699–1708, 2014.

[Hat14] Hamed Hatami. Scribe notes for the course *COMP760: Harmonic Analysis of Boolean Functions*, 2014. Available at `http://cs.mcgill.ca/~hatami/comp760-2014/lectures.pdf`.

[Hem94] Lane Hemaspaandra. Complexity theory column 5: the not-ready-for-prime-time conjectures. *ACM SIGACT News*, 25(2):5–10, 1994.

[HMP$^+$93] András Hajnal, Wolfgang Maass, Pavel Pudlák, Márió Szegedy, and György Turán. Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46:129–154, 1993.

[HO02] Lane Hemaspaandra and Mitsunori Ogihara. *The Complexity Theory Companion*. Springer, 2002.

[HRZ95] Lane Hemaspaandra, Ajit Ramachandran, and Marius Zimand. Worlds to die for. *ACM SIGACT News*, 26(4):5–15, 1995.

[IMP12] Russell Impagliazzo, William Matthews, and Ramamohan Paturi. A satisfiability algorithm for AC^0. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 961–972, 2012.

[IS01] Russell Impagliazzo and Nathan Segerlind. Counting axioms do not polynomially simulate counting gates. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pages 200–209, 2001.

[Joh86] David Johnson. The NP-completeness column: An ongoing guide. *Journal of Algorithms*, 7(2):289–305, 1986.

[Juk12] Stasys Jukna. *Boolean Function Complexity*. Springer, 2012.

[Kal00] Gil Kalai. Combinatorics with a geometric flavor: some examples, 2000. GAFA Special Volume 10, Birkhauser Verlag, Basel, 2000.

[Kal10] Gil Kalai. Noise Stability and Threshold Circuits. Blog post at *Combinatorics and more*, 2010. `https://gilkalai.wordpress.com/2010/02/10/noise-stability-and-threshold-circuits`.

[Kal12] Gil Kalai. Answer to the question: *Are all functions whose Fourier weight is concentrated on the small sized sets computed by* AC^0 *circuits?* Theoretical Computer Science StackExchange, 2012. `http://cstheory.stackexchange.com/questions/12769/are-all-the-functions-whose-fourier-weight-is-concentrated-on-the-small-sized-se`.

[KPPY84] Maria Klawe, Wolfgang Paul, Nicholas Pippenger, and Mihalis Yannakakis. On monotone formulae with restricted depth. In *Proceedings of the 16th Annual ACM Symposium on Theory of Computing*, pages 480–487, 1984.

[KPW95] Jan Krajíček, Pavel Pudlák, and Alan Woods. An exponential lower bound to the size of bounded depth frege proofs of the pigeonhole principle. *Random Structures & Algorithms*, 7(1):15–39, 1995.

[KS05] Gil Kalai and Shmuel Safra. Threshold phenomena and influence. In *Computational Complexity and Statistical Physics*, pages 25–60. Oxford University Press, 2005.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

[Man95] Yishay Mansour. An $O(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution. *Journal of Computer and System Sciences*, 50:543–550, 1995.

[Nis91] Noam Nisan. Pseudorandom bits for constant depth circuits. *Combinatorica*, 11(1):63–70, 1991.

[O'D07] Ryan O'Donnell. Lecture 29: Open Problems. Scribe notes for the course *CMU 18-859S: Analysis of Boolean Functions*, 2007. Available at `http://www.cs.cmu.edu/~odonnell/boolean-analysis`.

[OW07] Ryan O'Donnell and Karl Wimmer. Approximation by DNF: examples and counterexamples. In *34th International Colloquium on Automata, Languages and Programming*, pages 195–206, 2007.

[PBI93] Toniann Pitassi, Paul Beame, and Russell Impagliazzo. Exponential lower bounds for the pigeonhole principle. *Computational complexity*, 3(2):97–140, 1993.

[Raz87] Alexander Razborov. Lower bounds on the size of bounded depth circuits over a complete basis with logical addition. *Mathematical Notes of the Academy of Sciences of the USSR*, 41(4):333–338, 1987.

[Raz95] Alexander Razborov. Bounded arithmetic and lower bounds in Boolean complexity. In *Feasible Mathematics II*, pages 344–386. Springer, 1995.

[Raz09] Alexander Razborov. A simple proof of Bazzi's theorem. *ACM Transactions on Computation Theory*, 1(1):3, 2009.

[SBI04] Nathan Segerlind, Sam Buss, and Russell Impagliazzo. A switching lemma for small restrictions and lower bounds for k-DNF resolution. *SIAM Journal on Computing*, 33(5):1171–1200, 2004.

[Sip83] Michael Sipser. Borel sets and circuit complexity. In *Proceedings of the 15th Annual ACM Symposium on Theory of Computing*, pages 61–69, 1983.

[Smo87] Roman Smolensky. Algebraic methods in the theory of lower bounds for boolean circuit complexity. In *Proceedings of the 19th Annual ACM Symposium on Theory of Computing*, pages 77–82, 1987.

[ST95] David Shmoys and Éva Tardos. Computational Complexity. In *Handbook of Combinatorics (Ronald Graham, Martin Grötschel, and Lászlo Lovász, eds.)*, volume 2. North-Holland, 1995.

[Sub61] Bella Subbotovskaya. Realizations of linear functions by formulas using $\vee$, &, $\bar{}$. *Doklady Akademii Nauk SSSR*, 136(3):553–555, 1961.

[Tar89] Gábor Tardos. Query complexity, or why is it difficult to separate $\mathsf{NP}^A \cap \mathsf{coNP}^A$ from P^A by random oracles A? *Combinatorica*, 9(4):385–392, 1989.

[Tha09] Neil Thapen. Notes on switching lemmas, 2009. Available at `http://users.math.cas.cz/~thapen/switching.pdf`.

[Val83] Leslie Valiant. Exponential lower bounds for restricted monotone circuits. In *Proceedings of the 15th Annual ACM Symposium on Theory of Computing*, pages 110–117, 1983.

[Vio13] Emanuele Viola. Challenges in computational lower bounds. *Electronic Colloquium on Computational Complexity*, 2013.

[VW97] Heribert Vollmer and Klaus Wagner. *Measure One Results in Computational Complexity Theory*, pages 285–312. Advances in Algorithms, Languages, and Complexity. Springer, 1997.

[Wil14a] Ryan Williams. Faster all-pairs shortest paths via circuit complexity. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 664–673, 2014.

[Wil14b] Ryan Williams. The polynomial method in circuit complexity applied to algorithm design (invited survey). In *Proceedings of the 34th Foundations of Software Technology and Theoretical Computer Science Conference*, 2014.

[Yao85] Andrew Yao. Separating the polynomial-time hierarchy by oracles. In *Proceedings of the 26th Annual Symposium on Foundations of Computer Science*, pages 1–10, 1985.

A Proof of Lemma 7.1

Lemma 7.1. *There is a universal constant $c > 0$ such that for $2 \le d \le \frac{cm}{\log m}$, we have that $t_k = q \pm q^{1.1}$ for all $k \in [d-1]$.*

Proof. We shall establish the following bound, for $k = d-1, \ldots, 1$, by downward induction on k:

$$|t_k q - p| \le (2m)^{d-1-k}\lambda. \tag{34}$$

Lemma 7.1 follows directly from (34), using (7), (3) and the fact that $p = \Theta(\frac{\log w}{w})$.

The base case $k = d-1$ of (34) holds with equality since (8) gives us that $|t_{d-1}q - p| = \lambda$. For the inductive step suppose that (34) holds for some value $k = \ell + 1$. By (8) we have that $t_\ell q = (1 - t_{\ell+1})^{qw} - \lambda$, so our goal is to put upper and lower bounds on $(1 - t_{\ell+1})^{qw} - \lambda$ that are close to p. For the upper bound, we have

$$\begin{aligned}
(1 - t_{\ell+1})^{qw} - \lambda &= \left((1 - t_{\ell+1})^{\frac{1}{t_{\ell+1}}}\right)^{qwt_{\ell+1}} - \lambda \\
&\le \exp\left(-qwt_{\ell+1}\right) - \lambda && \text{(by Fact 5.3)} \\
&\le \exp\left(-w\left(p - (2m)^{d-\ell-2}\lambda\right)\right) - \lambda && \text{(by the inductive hypothesis)} \\
&\le \exp\left(-\left(\frac{m2^m}{\log e} - 1\right)\cdot\left(2^{-m} - (2m)^{d-\ell-2}\lambda\right)\right) - \lambda && \text{(by (3))} \\
&= 2^{-m}\cdot\exp\left(2^{-m} + \left(\frac{m2^m}{\log e} - 1\right)\cdot(2m)^{d-\ell-2}\lambda\right) - \lambda \\
&\le p\cdot\left(1 + 2^{-m+1} + \frac{m2^{m+1}}{\log e}\cdot(2m)^{d-\ell-2}\lambda\right) - \lambda && \text{(by Fact 5.3)} \\
&\le p + 2^{-2m+1} + \frac{2m}{\log e}(2m)^{d-\ell-2}\lambda - \lambda \\
&\le p + (2m)^{d-\ell-1}\lambda,
\end{aligned}$$

where in the last inequality we have used the fact that $\lambda = \tilde{\Theta}(2^{-5m/4})$.

For the lower bound we proceed similarly:

$$
\begin{aligned}
(1-t_{\ell+1})^{qw}-\lambda &= \left((1-t_{\ell+1})^{\frac{1}{t_{\ell+1}}}\right)^{qwt_{\ell+1}}-\lambda \\
&\geq \exp\left(-qwt_{\ell+1}\right)\cdot(1-t_{\ell+1})^{qwt_{\ell+1}}-\lambda && \text{(by Fact 5.3)} \\
&\geq \exp\left(-w\left(p+(2m)^{d-\ell-2}\lambda\right)\right)\cdot(1-qw(t_{\ell+1})^2)-\lambda && \text{(by the i.h. \& Fact 5.2)} \\
&\geq \exp\left(-\frac{m2^m}{\log e}\cdot\left(2^{-m}+(2m)^{d-\ell-2}\lambda\right)\right)\cdot(1-qw(t_{\ell+1})^2)-\lambda \\
&= 2^{-m}\cdot\exp\left(-\frac{m2^m}{\log e}\cdot(2m)^{d-\ell-2}\lambda\right)\cdot(1-qw(t_{\ell+1})^2)-\lambda \\
&\geq 2^{-m}\cdot\left(1-\frac{m2^m}{\log e}\cdot(2m)^{d-\ell-2}\lambda-qw(t_{\ell+1})^2\right)-\lambda && \text{(using Fact 5.2)} \\
&\geq 2^{-m}\cdot\left(1-\frac{m2^m}{\log e}\cdot(2m)^{d-\ell-2}\lambda-\frac{w}{q}\cdot\left(p+(2m)^{d-\ell-2}\lambda\right)^2\right)-\lambda && \text{(by the i.h.)} \\
&\geq 2^{-m}\cdot\left(1-\frac{m2^m}{\log e}\cdot(2m)^{d-\ell-2}\lambda-\frac{4wp^2}{q}\right)-\lambda && \text{(by the bound on } d\text{)} \\
&= p-\frac{m}{\log e}\cdot(2m)^{d-\ell-2}\lambda-\frac{4wp^3}{q}-\lambda \\
&\geq p-(2m)^{d-\ell-1}\lambda. && \square
\end{aligned}
$$