Electronic Colloquium on Computational Complexity, Report No. 311 (2026)

New Gate, Weight and Wire Complexity Bounds for Sigmoid Neural Networks

Jiatu Li*
MIT
jli@mit.edu

Daniel Reichman
WPI
dreichman@wpi.edu

Ryan Williams†
MIT
rrw@mit.edu

Tony Wu
MIT
tonyyzwu@mit.edu

Abstract

Feedforward neural networks with sigmoid activations are of fundamental importance in neural computation. However, our understanding of the circuit complexity of such networks for implementing Boolean functions remains limited. The analog output of the sigmoid, along with its transcendental nature, leads to unique mathematical challenges for proving lower bounds.

Introducing new techniques, we establish new upper and lower bounds on the complexity of sigmoidal networks computing Boolean functions:

- We prove nearly tight bounds for the wire complexity of computing PARITY with such networks of constant depth and polynomial-bounded weights.
- We extend the size lower bound of Kane and Williams [STOC'16] against depth-3 threshold circuits, to depth-3 sigmoidal networks with polynomial weights, proving a $\Omega(n^{1.5}/(\log n)^{O(1)})$ size lower bound for computing an explicit function in P.
- We prove that every symmetric function on n bits can be computed by a depth-3 sigmoidal network with $O(n)$ wires and polynomial weights. This stands in stark contrast with threshold circuits of $d = O(1)$ depth, where $n^{1+\alpha_d}$-wire lower bounds are known for $\alpha_d > 0$.

We also answer a few fundamental problems related to sigmoidal networks, including proving new simulations and separations between threshold circuits and sigmoidal networks.

We establish these results using a variety of techniques, including random restrictions for sigmoid gates, analyzing the average sensitivity of networks with sigmoids, simulation arguments, and using tools from approximation theory, complex analysis, and algebraic geometry.

*Supported by MIT Schwarzman College of Computing PhD Fellowship. Part of work is undertaken during the program Logical Foundations of Complexity Theory (LFC) organized by the Isaac Newton Institute for Mathematical Sciences, Cambridge; the LFC program was supported by EPSRC grant EP/Z000580/1.

†Work supported in part by NSF CCF-2420092 and the AI Security Institute (AISI).

ISSN 1433-8092

Contents

1 Introduction **1**
- 1.1 Our Results . . . 3

2 Intuition **6**
- 2.1 Wire Complexity for Symmetric Functions . . . 6
- 2.2 Lower Bounds from Random Restriction . . . 7
- 2.3 Unrestricted Weight Sigmoidal Circuits . . . 10
- 2.4 Equivalence between Sigmoidal and Threshold Circuits . . . 11
- 2.5 Related Work . . . 12
- 2.6 Discussion . . . 14

3 Preliminaries **14**

4 Wire Complexity for Sigmoidal Circuits **18**
- 4.1 Symmetric Functions using Linear Wires . . . 18
- 4.2 Wire Complexity for Depth-2 Sigmoidal Circuits . . . 19

5 Lower Bounds via Random Restriction **21**
- 5.1 Random Restriction Lemmas for Sigmoid Functions . . . 21
- 5.2 Lower Bounds for Generalized Andreev's Function . . . 24
- 5.3 Correlation Bounds with Parity via Average Sensitivity . . . 31

6 On Sigmoidal Circuits with Large Weight **37**
- 6.1 Weight Upper Bounds on Depth 2 . . . 37
- 6.2 Lower Bounds for Depth-2 Circuits . . . 39
- 6.3 A Sublinear Upper Bound for IP Mod 2 . . . 41

7 Further Simulations between Threshold and Sigmoidal Circuits **46**
- 7.1 Connection to Threshold-of-Majority . . . 46
- 7.2 Connection to Threshold Circuits . . . 48

A Further Related Work **57**

B Missing Proofs **58**
- B.1 Proof of the Parameter Compression Lemma . . . 58
- B.2 Proof of the Restriction Lemma . . . 58
- B.3 Rational Approximation of Sigmoid . . . 60

1 Introduction

For decades, researchers have studied the expressive power of neural networks [MP88, MP43] with respect to natural parameters, such as depth, size (number of neurons or number of connections), and the activation functions used. Our focus here is on neural networks computing Boolean functions (i.e., binary classification problems on discrete inputs).

The basic building block of a neural network is a *gate* (or computation unit, or neuron). A fan-in m gate has m incoming wires, where the i-th wire has a real *weight* $w_i \in \mathbb{R}$. Fix a real function $\phi : \mathbb{R} \to \mathbb{R}$, commonly referred to as the *activation function* or simply activation. The output of the gate is $\phi(\sum_{i=1}^{m} w_i x_i + \theta)$ where $\theta \in \mathbb{R}$ is called *bias*. From now on, the number of input variables in the network is denoted as n.

There are many possible functional forms the activation function can have, such as polynomial, piecewise linear and sigmoidal. Two of the most studied activations are the *threshold* (or *Heaviside*) activation $\tau(x)$ and the *sigmoid* activation

$$\tau(x) = \mathbb{I}[x \geq 0], \qquad \mathsf{Sigmoid}(x) := \frac{1}{1+e^{-x}}.$$

Given an activation function σ, a *σ-circuit* is defined as a circuit with gates computing functions of form $\vec{x} \mapsto \sigma(\langle \vec{w}, \vec{x} \rangle + \theta)$. In particular, the ordinary linear threshold circuit (see [Juk12, Section 11]) is a σ-circuit with $\sigma(x)$ being a threshold function. For a σ-circuit $C : \{0,1\}^n \to \mathbb{R}$ and a Boolean function $f : \{0,1\}^n \to \{0,1\}$, we say C computes f with ε separation, if there exists $\tau \in [0,1]$ such that

$$f(x) = 0 \implies C(x) \leq \tau \qquad \text{and} \qquad f(x) = 1 \implies C(x) > \tau + \varepsilon. \tag{1}$$

In particular, C computes f *with no separation* if $\varepsilon = 0$, and it is said to compute f *with noticeable separation* if $\varepsilon = \Omega(n^{-k})$ for some $k \geq 1$. Our focus in this work is proving upper and lower bounds for various complexity measures (number of gates, wires and magnitude of weights) for neural networks with sigmoid activations.

Beyond their foundational role in machine learning, sigmoidal networks serve as the canonical testbed for a central question in neural computation:

> *When do continuous, smooth analog neurons confer a provable advantage over discrete neurons?* [MSS94, DS96, Koi96, KS97]

Unlike threshold gates (which are discrete) or ReLU and polynomial activations (which are piecewise-linear or algebraic), the sigmoid combines local curvature with the exponential function e^{-x}. On the one hand, this makes sigmoidal gates a core primitive across modern machine learning architectures and verification problems—underpinning Boltzmann machines [HS+86, MCPZ13], gating mechanisms, softmax and attention layers [AA22, MSS22, DYZ+26], graph neural networks [Gro24, DBS25], and formal DNN verification [IZBK23]. On the other hand, from a circuit complexity perspective, passing continuous real values across wires and introducing exponential non-linearities fundamentally alters the landscape of bounded-depth computation, rendering classical tools for threshold circuits (TC^0) challenging to apply and raising basic questions about whether smooth analog networks can bypass long-standing Boolean circuit barriers [SRK94, MSS94, Koi96].

Circuit Lower Bounds for Neural Networks. While we have built up a decent (yet still very incomplete) understanding of circuits with threshold gates (see, e.g., [HG91, GHR92, HMP+93,

ROS94, Hås94, PS94, IPS97, HP10, KW16, KKL17, CSS18, HHTT23, FLY22]), sigmoidal networks have been a far more challenging representation to understand [Koi96, SRK94]. This is perhaps not surprising, as sigmoidal networks are at least as powerful as threshold circuits (see Lemma 3.13). Indeed, it is also known that for certain Boolean functions, sigmoidal networks can compute these functions with *asymptotically few* gates compared to threshold circuits implementing the same functions [MSS94, DS96, Koi96].

Additionally, lower bound techniques for threshold circuits simply do not apply in the sigmoidal setting. At a high level, the Boolean output of threshold gates makes them amenable to a variety of methods from computational complexity such as communication complexity [Nis93, ROS94], the discriminator lemma [HMP+93], and random restrictions [IPS97, CSS18, HHTT23, FLY22, KKL17]. There seems to be no portable analog of these in the sigmoidal case.

Example 1. An important component in establishing lower bounds for threshold circuits is *upper bounding* the magnitude of weights of threshold circuits: it is known that if a Boolean function is computed by a threshold circuit C then without loss of generality the size of every weight in C is at most $2^{O(n \log n)}$ (and weights of order $2^{\Omega(n \log n)}$ are sometimes necessary); see, e.g., [Hås94]. Devising weight upper bounds for sigmoidal networks in terms of the input dimension is not known, limiting the applicability of techniques to prove lower bounds that rely on the boundedness of the weights, e.g., Williams' algorithmic approach [Tam16, ACW16, CTW26].

In the zero separation case, the transcendental nature of the sigmoid function is an additional complicating factor. While simulation results of neural networks with piecewise linear (such as ReLUs) or piecewise polynomial activation of constant depth using threshold circuits with a polynomial blowup are known [Maa97a, MSS94], such results are unknown for sigmoidal networks in the zero separation case.

Motivating Questions. These challenges translate into noticeable gaps in terms of known lower bounds for sigmoidal networks versus threshold circuits. While the study of the latter has seen progress in recent years, the former has not seen progress since the 90's.

- While superlinear gate lower bounds have been established for *threshold circuits* of depth 2 (with arbitrary weights) [Tam16, ACW16, CTW26] and depth 3 (with polynomial weights) [KW16], comparable lower bounds are not known even for sigmoidal networks with polynomial weights computing Boolean functions with a noticeable separation.
- Several works have successfully applied the random restriction method to prove lower bounds on the wire complexity of well studied functions such as parity for threshold circuits both in the worst and average case [PS94, IPS97, CSS18] as well as the more general case of circuits with polynomial threshold functions [KKL17]. However, we are not aware of any lower bounds on the wire complexity of sigmoidal networks for computing parity or other Boolean function.
- Finally, a linear lower bound is known on the gate complexity of threshold circuits (with arbitrary depth and weights) computing the IP Mod 2 function using communication complexity [ROS94]. For sigmoidal networks computing with zero separation the best size lower bound is $\Omega(n^{1/4})$ for the Boolean selection function relying on upper bounds of the VC dimension of such networks [KM97] (which relies on results from algebraic geometry [Kho91]) [Koi96].

We establish new lower and upper bounds for sigmoidal networks computing Boolean functions related to some (though not all) of the open questions above.

1.1 Our Results

We are interested in proving upper and lower bounds for the number of gates (i.e. neurons), number of wires (i.e. non-zero parameters), and magnitude of weights in the following parameter regimes:

Robust sigmoidal circuits: Sigmoidal circuits with *polynomially bounded weights and biases* that compute Boolean functions with *noticeable separation* (see [MSS94, PS94, SRK94, DS96]).

Unrestricted-weight sigmoidal circuits: Sigmoidal circuits with arbitrary weights computing Boolean functions with *no separation*[1] (see [Koi96]), which is a generalization of THR circuits.

Small-weight sigmoidal circuits: Sigmoidal circuits with *polynomially bounded weights* that compute Boolean functions with *no separation*, which is a generalization of MAJ circuits.

Robust sigmoidal circuits are indeed equivalent to MAJ circuits of the same depth up to a polynomial overhead (see Lemmas 3.10 and 3.13). Therefore, existing lower bounds against depth-2 MAJ circuits [HMP+93, For02, FKL+01] applies directly. The more interesting question is whether we can prove lower bounds for depth three and beyond given, e.g., the $\tilde{\Omega}(n^{1.5})$ lower bound[2] against depth-3 MAJ circuits from [KW16] and the super-linear wire lower bound [IPS97, CSS18] against TC^0. These lower bounds do not translate directly due to the polynomial overhead.

The unrestricted-weight model might be surprisingly strong: it is not immediately clear whether we can simulate them by even $\mathsf{P/poly}$. On the other hand, however, it is also unclear whether it is much stronger than threshold circuits.

Wire Complexity of Parity. A classical question in circuit complexity and also machine learning is the complexity of computing the parity function $\oplus_n(x_1, \ldots, x_n) := x_1 \oplus \cdots \oplus x_n$ by neural networks. We fully resolved (up to poly-log factors) the wire complexity of polynomial weight sigmoidal circuits computing the parity function with noticeable separation.

Theorem 1.1 (see Theorem 4.1, Proposition 4.3, and Corollary 4.6)**.** *Every symmetric Boolean function (including parity) is computable by a depth-3 robust sigmoidal circuit with $O(n)$ wires.*

Any symmetric function (including parity) is computable by a depth-2 robust sigmoidal circuit with $O(n^2)$ wires, and least $\Omega(n^2/\log^2 n)$ wires are required for parity in the same model.

The linear wire upper bound for $d \geq 3$ stands in sharp contrast with the known super-linear wire complexity lower bound for parity against threshold circuits for *any* $d \geq 1$ (see, e.g., [IPS97, KKL17, CSS18]). It shows that there is a provable advantage for real-valued activation functions (compared to Boolean-valued activation) in wire complexity measure.

Lower Bounds for Constant-Depth Polynomial Weight Sigmoidal Circuits. Given the separation between sigmoidal and threshold circuits in wire complexity, it is natural to consider whether known threshold circuit lower bounds can be adapted to sigmoidal circuits.

Average-case Lower Bound for Parity. The number of gates and wires required to compute the parity function is a classical question in circuit complexity. For worst-case computation, almost

[1] While computing without separation allows arbitrarily small margins that may be infeasible under finite-precision hardware, we view this as a natural theoretical regime worthy of study. Indeed, the same convention is standard in unrestricted threshold circuits, where gate margins can be exponentially small in the input dimension.

[2] The notation $\tilde{\Omega}(\cdot)$ means that we ignore poly log factors

optimal wire lower bounds are known for both threshold circuits [IPS97] (following [PS94]), and almost optimal gate lower bounds are known for polynomial-weight sigmoidal circuits [SRK94]. Chen, Santhanam, and Srinivasan [CSS18] extends [IPS97] to show a near-optimal average-case lower bound on the number of wires of threshold circuits computing parity.

We prove an average-case lower bound against polynomial weight sigmoidal circuits.[3]

Theorem 1.2 (see Theorem 5.19)**.** *For every $d \geq 2$, and any function $f : \{0,1\}^n \to \{0,1\}$ that is computable by a depth-d robust sigmoidal circuit using $O(n^{1/(2(d-1))-\Omega(1)})$ gates, the correlation between f and $\oplus_n$ is at most $O(n^{-\Omega(1)})$.*

The lower bound matches that in [CSS18] and is thus also near optimal. Indeed, it is a corollary of an upper bound on the *average sensitivity* of functions computable by sigmoidal circuits, which is of independent interests. It is worth noting that the results of [CSS18] are obtained to analyze the *noise sensitivity* of threshold circuits; we suspect both techniques (with appropriate adaption) would work for both settings.

Theorem 1.3 (see Theorem 5.20)**.** *For any $d \geq 1$, and any function $f : \{0,1\}^n \to \{0,1\}$ that is computable by a depth-d robust sigmoidal circuit using s gates, the average sensitivity of f is at most $O(s^{1-1/d} \cdot n^{1-1/(2d)+o(1)})$.*

To derive this lower bound, we need to analyze an extension of sigmoidal circuits, where the bias of each gate is replaced by a low-degree polynomial. This is possible due to a strong random restriction lemma for *polynomial threshold circuits* [KKL17]; see Section 2 and the technical sections.

Explicit Lower Bounds for Depth-3 Circuits. Recall that Kane and Williams [KW16] proved that there is a polynomial-time computable function that requires at least $\Omega(n^{1.5}/(\log n)^{O(1)})$ gates and $\Omega(n^{2.5}/(\log n)^{O(1)})$ wires to compute by $\mathsf{THR} \circ \mathsf{THR}$. In this work, we show that this lower bound can be adapted to sigmoidal circuits with *polynomial weights* and *noticeable separation*.

Theorem 1.4 (see Theorems 5.10 and 5.18)**.** *There is a polynomial-time computable function $f : \{0,1\}^n \to \{0,1\}$ that requires $\Omega(n^{1.5}/(\log n)^{O(1)})$ gates and $\Omega(n^{2.5}/(\log n)^{O(1)})$ wires to compute by depth-three robust sigmoidal circuits.*

We note that the lower bound is nontrivial in that it cannot be obtained by simply simulating sigmoidal circuits with threshold circuits: the simulation involves a gate overhead corresponding to weights (see Lemma 3.10). The wire lower bound is particularly interesting because of the surprising upper bound in Theorem 1.1.

We prove the lower bound by generalizing the technique in [KW16]. Instead of directly using their results, we have to re-engineer the entire proof [KW16] in white box as most parts of the proof are not applicable to sigmoidal circuits. The main issue is that the key random restriction lemma does not work for sigmoidal circuits: unlike threshold gates, sigmoidal gates have continuous output and cannot be well approximated by constants after random restrictions. To circumvent this problem, we prove a random restriction lemma for sigmoidal gates that further uses *Taylor approximation*, and shows that it suffices for the framework in [KW16].

[3] Average case lower bounds for computing Boolean functions by neural networks are desirable as they can be used to achieve lower bounds for approximating functions in the ℓ^2-norm that is common in machine learning applications (see, e.g., [VS20]).

Sigmoidal Circuits with Unbounded Parameters. The results above work only for robust sigmoidal circuits. A natural question then is to study the nature of unrestricted-weight sigmoidal circuits. This is indeed a surprisingly tough model to analyze: very few upper and lower bounds are known. In this work, we answer a few fundamental questions related to this model.

A Weight Upper Bounds for Sigmoidal Circuits. For threshold circuits, it is known that THR gates with unbounded parameters can be simulated by THR gates with integer weights at most $2^{O(n \log n)}$ (see [Hås94] and references therein). As THR is Boolean valued, it further implies that any threshold circuit can be simulated by one with the same topology but weights bounded by $2^{O(n \log n)}$. This serves an fundamental tool to analyze threshold circuits (see, e.g., [GHR92, HP10]).

However, this result does not easily generalize to sigmoidal circuits as it is real-valued and we thus cannot deal with each gates individually. We prove a weight upper bound for depth-two sigmoidal circuits using tools from computational algebraic geometry [BPR96, BPR06].

Theorem 1.5 (see Theorem 6.1)**.** *Suppose that $f : \{0,1\}^n \to \{0,1\}$ is computable by a depth-two unrestricted-weight sigmoidal circuit with s gates. It is computable by a circuit that has the same architecture with weights and biases having absolute values at most $2^{O(sn \log(sn))}$.*

Explicit Lower Bounds for Depth-2 Sigmoidal Circuits. Proving explicit lower bound against unbounded weight sigmoidal circuits appears to be a surprisingly hard problem. Prior to this work, the best known lower bound against constant-depth sigmoidal circuits with unbounded weights is only $\Omega(n^{1/4})$ [Koi96]. Following Koiran's approach, we observe that there is a near-linear lower bound[4] for depth-two sigmoidal circuits.

Theorem 1.6 (see Theorem 6.5)**.** *The selection function $f(x, y) := y_x$ requires $\Omega(n/\log n)$ gates to compute by any depth-two unrestricted-weight sigmoidal circuits.*

Advantage of Sigmoids. Sigmoidal circuits are known to be asymptotically stronger than threshold circuits in various settings [MSS94, DS96, Koi96]. For circuits with unbounded weight, it is known that the Selection function $f(x, y) = y_x$ requires $\Omega(n/\log n)$ threshold gates, while there is an $O(\sqrt{n})$ upper bound by Sigmoidal circuits [Koi96]. We show that there is a similar upper bound for another well-studied function $\mathsf{IP2}(\vec{x}, \vec{y}) := \langle \vec{x}, \vec{y} \rangle \bmod 2$.

Theorem 1.7 (see Section 6.3)**.** $\mathsf{IP2}$ *can be computed by an unrestricted-weight sigmoidal circuit with $O(\sqrt{n \log n})$ gates.*

This upper bound is particularly interesting as an $\Omega(n)$ lower bound is known for threshold circuits [ROS94], and a $\Omega(n/\log n)$ lower bound is known for ReLU networks [VRPS21]; in contrast, the Selection function can be computed by ReLU networks of size $O(\sqrt{n})$ [Koi96]. It therefore demonstrates an asymptotic separation between sigmoidal and ReLU networks. Finally we note that the improved upper bound above for the Inner Product function is the only result in this paper concerning a sigmoidal circuit whose depth grows with n.

An Equivalence between Sigmoidal and Threshold Circuits. We are also interested in the intermediate small-weight model that generalizes MAJ circuits: sigmoidal circuits with small weights that computes Boolean functions with no separation. It is tempting to ask whether this model is strictly stronger than its discrete counterpart, i.e., $\mathsf{MAJ} \circ \mathsf{MAJ}$.

[4] This is nearly optimal as there is a depth-two sigmoidal circuit of linear size computing selection [Koi96].

We show that depth-two circuits in this model is at least as strong as $\mathsf{THR} \circ \mathsf{MAJ}$, and indeed, a further restriction to integer weights make it exactly equivalent to $\mathsf{THR} \circ \mathsf{MAJ}$.

Theorem 1.8. *Let $n \in \mathbb{N}$ be sufficiently large and $s = s(n), W = W(n) \geq 1$ be functions, and $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function. The following two statements are equivalent:*

(1) f can be computed by a depth-two sigmoidal circuit of size $s \cdot \mathsf{poly}(n, W)$ with no separation, where all weights are integers in $[-W, W]$ and all biases are real numbers in $(-1, 1)$.

(2) f can be computed by a $\mathsf{THR} \circ \mathsf{MAJ}$ circuit of size $s \cdot \mathsf{poly}(n, W)$, where the bottom MAJ gates have weight taken from $\{-W, \ldots, W\}$.

Moreover, the output gate of the circuit in (1) can be assumed to have weights in $\{\pm 1\}$.

The implication (2) $\Rightarrow$ (1) gives an interesting structural characterization of $\mathsf{THR} \circ \mathsf{MAJ}$. It shows that sigmoidal circuits can be surprisingly strong even with small integer weights and small real biases. In particular, as exponential separations between $\mathsf{THR} \circ \mathsf{MAJ}$ and $\mathsf{MAJ} \circ \mathsf{MAJ}$ is known [GHR92], this also indicates that real-valued activation functions are exponentially stronger than threshold activation if we restrict to the polynomial weight regime.

Following a similar approach, we also prove a structural upper bound for simulating $\mathsf{THR} \circ \mathsf{THR}$ with sigmoidal circuits with top weights in $\{\pm 1\}$. Since $\mathsf{MAJ} \circ \mathsf{THR} \subsetneq \mathsf{THR} \circ \mathsf{MAJ} \subsetneq \mathsf{THR} \circ \mathsf{THR}$ [GHR92, CM22], sigmoidal circuit with small top weights is stronger than its counterpart for threshold sigmoidal circuits ($\mathsf{MAJ} \circ \mathsf{THR}$) and also the restricted sigmoidal circuits in Theorem 1.8.

Theorem 1.9. *Suppose that $f : \{0,1\}^n \to \{0,1\}$ is computed by a $\mathsf{THR} \circ \mathsf{THR}$ circuit of size s. Then f is computed with no separation by a depth-two sigmoidal circuit of size $s \cdot \mathsf{poly}(n)$ whose output gate has weights in $\{\pm 1\}$.*

We also note that the simulation of threshold circuits using sigmoidal circuits in Theorems 1.8 and 1.9 also generalize to larger depth; see Section 7 for more details.

2 Intuition

2.1 Wire Complexity for Symmetric Functions

Recall that Parity requires $\widetilde{\Theta}(n^2)$ wires for $\mathsf{MAJ} \circ \mathsf{MAJ}$ [PS94], and super-linear wires for constant depth threshold circuits [IPS97, CSS18]. Assume that n is even for simplicity. The depth-two construction is as follows: for each $k \in \{1, 3, 5, \ldots, n-1\}$, create a gate $G_k(\vec{x})$ that outputs 1 iff $|\vec{x}| = k$, and take an OR of all such gates. Each gate needs $O(n)$ wires, and therefore there are $O(n^2)$ wires in total.

The depth-three sigmoidal circuit for Parity (and indeed any symmetric function) is built on the same construction. The key observation is, instead of letting $G_1, G_3, \ldots, G_{n-1}$ read all inputs by themselves, we can create a gate G_{sum} that outputs $\mathsf{Sigmoid}(\frac{1}{n}\sum_{i=1}^{n} x_i)$, so $G_1, G_3, \ldots, G_{n-1}$ only need to read the output of G_{sum} and collect $|\vec{x}|$ by appropriate scaling. The total number of wires is $O(n)$ as $G_1, G_3, \ldots, G_{n-1}$ have fan-in 1, and the weights and biases of all gates are $\mathsf{poly}(n)$.

The wire complexity upper bound for depth-two sigmoidal circuits follows by simulating the $\mathsf{MAJ} \circ \mathsf{MAJ}$ circuit gate by gate. The matching lower bound follows by generalizing the *rational approximation* technique in [PS94].

In more detail, the proof of [PS94] shows that if the activation function can be well approximated by a low-degree rational function, then every function computed by a depth-two network with

$o(n^2/\log^2 n)$ wires and polynomial weights and biases with noticeable separation can be sign-represented by a rational function of degree less than $n/2$ (see Theorem 4.4). The lower bound then follows from the fact that a rational function P/Q sign-representing Parity has degree at least $n/2$ [PS94]. Subsequently, it suffices to show that $\mathsf{Sigmoid}(x)$ can be well approximated by a low-degree rational function. Fortunately, this has been studied in [SRK94], and there is an approximation by combining Taylor expansion and an approximation of the sign function in [PS94].

2.2 Lower Bounds from Random Restriction

We now explain the intuition behind Theorems 1.2 to 1.4, which all use the random restriction method (but in slightly different manners).

Random Restriction Method, and Why it is Nontrivial for Sigmoid. One of our main technical contributions is to show that the random restriction method, which has been widely used to prove lower bounds against threshold circuits (see, e.g., [IPS97, KW16, CSS18, KKL17, HHTT23, FLY22]), can be adapted to analyze sigmoidal circuits.

A random restriction lemma (see, e.g., the one in [KW16]) for THR states that for any threshold gate $\mathsf{THR}(\vec{x}) := \mathsf{sgn}(\langle \vec{w}, \vec{x} \rangle + \beta)$, after applying a random assignment to all but a few random bits, the function likely becomes a constant function on all (or almost all) assignments on free variables. It is often used to simplify a circuit by removing the bottom layer, as many gates become constants and can be absorbed into the gates above it.

In our setting, gates involve real-valued activation functions (e.g. $\mathsf{Sigmoid}$), so we cannot hope for them to become a constant under random restriction. A natural attempt is to say, instead, that a gate will likely become close to a constant. This is indeed possible by approximating $\mathsf{Sigmoid}(x)$ by a piece-wise constant function; specifically, letting t and δ be two parameters, we write:

$$\mathsf{Sigmoid}(x) \approx \mathbb{I}[x \geq (t+1)\delta] + \sum_{-t \leq j \leq t} \mathsf{Sigmoid}(j\delta) \cdot \mathbb{I}[j\delta \leq x < (j+1)\delta], \tag{2}$$

and subsequently approximate a sigmoid gate by a linear combination of $O(t)$ THR gates. Then we can apply the random restriction lemma for THR gates, and use a *union bound* to say that it also works for sigmoid gates: if every threshold gate (used to test $\mathbb{I}[j\delta \leq x < (j+1)\delta]$) is trivialized, the sigmoidal gate will be close to a constant up to an error δ.

However, this appears to be insufficient to prove lower bounds: The random restriction lemmas for threshold circuits have an error probability $n^{-\Omega(1)}$ if we want to leave $n^{\Omega(1)}$ variables free (which is usually required for lower bound techniques), so we could afford at most $t = n^{\varepsilon}$ for a constant $\varepsilon < 1$, and subsequently $\delta \geq n^{-\varepsilon}$ (as otherwise the error for $x < -t\delta$ or $x \geq (t+1)\delta$ will be too large). An approximation of a gate in the bottom layer with error $n^{-\varepsilon}$ can *not* be ignored, as the error might be amplified by higher layers of the circuit and affect the final output.

Random Restriction Lemma. To prevent the aforementioned issue, we indeed need a random restriction lemma for $\mathsf{Sigmoid}$ with an approximation error at most n^{-c} for *any* c. This appears to be impossible if we approximate it by a constant after random restriction: $\mathsf{Sigmoid}(\sum_{i=1}^{n} x_i/n^2)$ is not close to any constant after any partial restriction.

We therefore use approximation by *polynomials*. Instead of using the constant $\mathsf{Sigmoid}(j\delta)$ in Equation (2), we use a degree-k Taylor approximation of $\mathsf{Sigmoid}(x)$ at $x_0 := j\delta$. For any $x \in$

$[j\delta, (j+1)\delta)$, this estimation has an error about δ^k; if we choose $t = \mathsf{polylog}(n)$, $\delta := 1/\mathsf{polylog}(n)$, and $k = O(\log n / \log\log n)$, the error can be boosted to n^{-c} for any fixed $c \geq 1$. Subsequently, we can show that for any sigmoidal gate, after a random restriction, it becomes close to a low-degree polynomial with error small enough to avoid affecting the final output.

Lower Bounds for Depth-3. Recall that Theorem 1.4 shows lower bounds comparable to results of Kane and Williams [KW16] for $\mathsf{MAJ} \circ \mathsf{THR} \circ \mathsf{THR}$ circuits. Indeed, our main contribution is an adaption of their proof to work for sigmoidal circuits.

Kane-Williams Lower Bounds. We first review the result in [KW16], focusing on the $n^{1.5}/(\log n)^{O(1)}$ gate lower bound. The hard function F is called the Generalized Andreev's function. It takes $(\vec{x}, \vec{y}) \in \{0,1\}^n \times \{0,1\}^m$ as input and satisfies the following property: For every set $\mathscr{L}$ of Boolean functions on m variables, if $|\mathscr{L}| \leq 2^{o(n)}$, there exists an $\vec{x} \in \{0,1\}^n$ such that for every $f \in \mathscr{L}$, the correlation between f and $F(\vec{x}, \cdot)$ is extremely small.

Suppose for contradiction that a small $\mathsf{MAJ} \circ \mathsf{THR} \circ \mathsf{THR}$ circuit C computes $F(\vec{x}, \vec{y})$. Then by a random restriction ρ to some variables in $\vec{y}$, many gates in the first layer will be trivialized. In that case, $C|_\rho$ can be computed by a $\mathsf{MAJ} \circ \mathsf{THR} \circ \mathsf{THR}$ circuit C' where there are only *a few bottom layer gates*. We then apply a standard tool called the discriminator lemma [HMP+93] to show that there exists a second layer gate G of C', which is a $\mathsf{THR} \circ \mathsf{THR}$ circuit with a few bottom gates that has nontrivial correlation with $F(\vec{x}, \cdot)$ for every $\vec{x}$. We will show that this is impossible.

Let $\mathscr{L}$ be the set of all depth-two circuits with a few bottom gates. By counting the total number of parameters, we can argue that $|\mathscr{L}| \leq 2^{o(n)}$, and subsequently there exists $\vec{x}$ such that $F(\vec{x}, \cdot)$ has very low correlation with every function in $\mathscr{L}$. Fix this $\vec{x}$. Then in particular, the gate G, which is a $\mathsf{THR} \circ \mathsf{THR}$ circuit with a few bottom gates, has very low correlation with $F(\vec{x}, \cdot)$. This leads to a contradiction.

Adaption to Sigmoidal Circuits. There are two main challenges to adapt the Kane-Williams lower bound to sigmoidal circuits. First, the discriminator lemma [HMP+93] is used to remove the top MAJ gate, while we need to remove a $\mathsf{Sigmoid}$ gate. To address this, we observe that as we assume polynomial weights for all $\mathsf{Sigmoid}$ gates, it is more similar to MAJ compared to THR, so we can discretize the top $\mathsf{Sigmoid}$ gate and apply the same proof idea.

The second challenge is that the random restriction lemma for $\mathsf{Sigmoid}$ gates simplifies them only to low-degree polynomials rather than constants, and in particular, each of the low-degree polynomials may have large weights. We call the subcircuits corresponding to the second-layer gates (in the simplified circuit) $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$ circuits, as the bottom layer consists of both $\mathsf{Sigmoid}$ gates and low-degree polynomials.

We may not use the same counting argument in [KW16] as there are too many polynomials (one for each simplified gate). Nevertheless, as the top $\mathsf{Sigmoid}$ gates compute a linear combination of bottom gates, and polynomials are *closed* under linear combination, we can merge all polynomials into one. Therefore, the total number of such polynomials can be bounded in terms of the number of free inputs and the degree, both of which are independent of the number of gates in the original circuit. We use this to conclude that the total number of possible $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$ circuits with certain parameters is small enough to for the argument in [KW16].

Remark 1. We note that another technique often used in analyzing neural networks is rational approximation (see, e.g., [PS94, SRK94]). Within the same degree, rational approximation could have

exponentially better precision (see, e.g., [New64] and Section B.3). However, the proof above crucially relies on the fact that polynomials are closed under linear combination *without degree blow up*, which is not satisfied by rational approximation; therefore it is unclear whether we can use rational approximation to improve this lower bound.

Remark 2. It can be observed that the random restriction lemma works for any sigmoid gate (i.e. with unbounded weights). The reason that the lower bounds work only for bounded weights is that otherwise (1) we cannot control the propagation of error (caused by the approximation of the bottom layer by polynomials) to the final output, and (2) we cannot apply the discriminator lemma. Indeed, the lower bound in Theorem 1.4 works even if the bottom layer has unbounded weights.

Correlation Bound with Parity. For threshold circuits, Chen, Santhanam, and Srinivasan [CSS18] proved an almost-tight average-case lower bound on both the number of gates and wires. We show that the gate lower bound can be adapted to sigmoidal circuits with polynomial weights.

Our proof follows from a somewhat different approach compared to that in [CSS18]. The lower bound in [CSS18] is proved by analyzing the *noise sensitivity* of threshold circuits with random restriction; we also use the random restriction method, while we instead prove an *average sensitivity* upper bound for functions computable by sigmoidal circuits. (Nevertheless, we suspect both ours and the approach in [CSS18] work in both cases.)

We first recall some basic concepts from Boolean function analysis. Let $f : \{0,1\}^n \to \{0,1\}$ be a function and $x \in \{0,1\}^n$. The sensitivity of f at x is defined as the number of i such that flipping the i-th bit of x will change $f(x)$. The average sensitivity of a function $f : \{0,1\}^n \to \{0,1\}$, denoted $\mathsf{AS}(f)$, is defined as the expected sensitivity at a random input $x \leftarrow \{0,1\}^n$. The relative average sensitivity of f is defined as $\mathsf{AS}(f)/n$. We need an observation of Boppana [Bop97] that over a random restriction ρ, the expected relative average sensitivity of $f|_\rho$ is exactly the relative average sensitivity of f.

We prove an average sensitivity upper bound for functions computable by depth-d, size-s sigmoidal circuits by induction on d. The base case $d = 1$ can be derived from the average sensitivity upper bound for linear threshold functions. Consider the induction case. Suppose we are given a circuit C of depth $d + 1$. By the observation of Boppana [Bop97], we only need to upper bound the expected average sensitivity of $C|_\rho$ over a random restriction ρ. We hope to simplify C by the random restriction lemma and apply the induction hypothesis for depth d. The only problem is that the sigmoidal gates simplify only to low-degree polynomials, and we cannot convert the resulting circuit into a depth d sigmoidal circuit.

To circumvent this issue, we strengthen the induction hypothesis to work for extended sigmoidal circuits, in which the bias of each gate is replaced by a low-degree polynomial in input variables (rather than a constant). It can then absorb any polynomials attached to the bottom layer. As a result, the base case should be changed to work with polynomial threshold functions (PTF), and the random restriction lemma must work for $\mathsf{PSigmoid}$ gates. Fortunately, the average sensitivity of PTF has been resolved by Kane [Kan14], and the random restriction lemma for $\mathsf{PSigmoid}$ can be proved by a random restriction lemma for PTF in [KKL17] (in place of [KW16]).

2.3 Unrestricted Weight Sigmoidal Circuits

The lower bounds above all require small weights (and possibly also biases) for various technical reasons. Now we explain our results on depth-two sigmoidal circuits with *unbounded parameters*.

Parameter Upper Bound for Depth-2. The main difficulty to generalize the weight bounds (see [Hås94]) for threshold functions to sigmoidal circuits is that

1. the gates are real-valued, so we cannot hope to replace them by bounded weight gates with bounded parameters without incurring an approximation error;
2. even for depth-two circuits, approximation error in the bottom layer will be amplified by the next layer, so it may change the function computed by the circuit.

Therefore, we cannot hope to prove an upper bound for parameters by adjusting the parameters gate-by-gate, or even layer-by-layer.

The key idea is to consider the depth-two circuits as a whole. Suppose that a function $f : \{0,1\}^n \to \{0,1\}$ is computed by a depth-two sigmoidal circuit with no separation. As the $\mathsf{Sigmoid}$ function is monotone, we can remove the output gate so that

$$f(\vec{x}) = \mathbb{I}\left[c + \sum_{i=1}^{h} v_i \mathsf{Sigmoid}\left(b_i + \sum_{j=1}^{n} w_{ij} x_j\right) > 0\right]. \tag{3}$$

As $\mathsf{Sigmoid}(x) = 1/(1+e^{-x})$, we can observe that if we write $W_{ij} = e^{-w_{ij}}$ and $B_i = e^{-b_i}$, for every *fixed* input $\vec{x} \in \{0,1\}^n$, the RHS of (3) is the indicator of a *polynomial* inequality in $c, \{v_i\}, \{B_i\}, \{W_{ij}\}$. We can then write a system of 2^n polynomial inequalities with small integer weights (by enumerating all 2^n inputs) that defines the feasible set of parameters such that the circuit computes f.[5]

Our theorem then follows from a generic result due to Basu, Pollack, and Roy [BPR96] that upper bounds the bit complexity of solutions for systems of polynomial inequalities (see also [BPR06] for a detailed explanation). Remarkably, their upper bound depends only on the number of variables, the weights and degree of polynomials, and is independent of the number of inequalities (which is 2^n); for this reason, we can upper bound all weights and biases by a single exponential in n and s (rather than exponential in 2^n).

Lower Bound for Selection. The only technique for proving lower bounds against unrestricted sigmoidal circuits is due to Koiran [Koi96]. Fix a topology of a sigmoidal circuit. Let $N_{\vec{w}}(x;\vec{v})$ be the circuit where some of the parameters are fixed as $\vec{w}$, while other parameters (called *programmable parameters*) are left unspecified. We can define the family of real functions computed by the $N_{\vec{w}}$ as the set of functions computed by $N_{\vec{w}}(x;\vec{v})$ for any possible $\vec{v}$.

Koiran [Koi96] observed that if a sigmoidal circuit C of size s computes the Selection function $f(x,y) := y_x$, where $x \in \{0,1\}^{\log n}$ and $y \in \{0,1\}^n$, by making $\leq s$ parameters (corresponding to y) programmable, the family of functions could have VC dimension at least n. Combining this with an $O(s^2m^2)$ VC dimension upper bound for sigmoidal circuits with s gates and m programmable parameters, he proved an $\Omega(n^{1/4})$ lower bound for computing Selection.

[5] Indeed, we also need $B_i, W_{ij} > 0$ due to the change of variables.

Our lower bound is proved by showing a stronger VC dimension upper bound for depth-two Sigmoidal circuits. Specifically, for any depth-two sigmoidal circuit of size s, the family of functions computed by making the biases programmable is at most $O(s \log s)$. This is proved using a VC dimension bound for polynomial threshold functions in [AB02, BW96]. Plugging the upper bound to Koiran's approach, we can obtain the $\Omega(n/\log n)$ lower bound.

Upper Bound for IP2. Recall that $\mathsf{IP2}(\vec{x}, \vec{y}) := \langle \vec{x}, \vec{y} \rangle \bmod 2$. We explain the upper bound with $O(\sqrt{n \log n})$ gates; the other one follows from a similar idea. Let the input length be $2n$. The upper bound relies on two tricks:

- Multiplication of two real numbers x, y can be approximated by a linear combination of four functions of form $\mathsf{Sigmoid}(ax + by + c)$, where $a, b, c \in \mathbb{R}$ [DS96].
- Let $C(\vec{x})$ be any sigmoidal circuits and $\varepsilon \in (0, 1)$, then there is a sigmoidal circuit D with one input and has k gates such that $|(D \circ C)(\vec{x}) - b_k| \leq \varepsilon$, where b_k denotes the k-th bit in the binary expansion of $C(\vec{x}) \in (0, 1)$ (see, e.g., [Koi96]).

Let $\vec{x} = (\vec{x}_1, \ldots, \vec{x}_k)$ and $\vec{y} = (\vec{y}_1, \ldots, \vec{y}_k)$ be a partition of the input into k blocks of size $r = n/k$. The key observation is that we can compute $\langle \vec{x}_i, \vec{y}_i \rangle$ by looking into the binary expansion of

$$p_i := \left(\sum_{j=0}^{r-1} 2^{jl} \vec{x}_{i,j} \right) \left(\sum_{j=0}^{r-1} 2^{(r-1-j)l} \vec{y}_{i,j} \right)$$

where $\vec{x}_i = (\vec{x}_{i,0}, \ldots, \vec{x}_{i,r-1})$ and $\vec{y}_i$ likewise. In more detail, the coefficient for the term $2^{(r-1)l}$ is exactly $\langle \vec{x}_i, \vec{y}_i \rangle$. When $l = O(\log n)$ is large enough, $\langle \vec{x}_i, \vec{y}_i \rangle \bmod 2$ is exactly the $(r-1)l$-th bit of p_i; indeed, we can even read out $\sum_{i=1}^{k} \langle \vec{x}_i, \vec{y}_i \rangle \bmod 2 = \mathsf{IP2}(\vec{x}, \vec{y})$ from the $(r-1)l$-th bit of $\sum_{i=1}^{k} p_i$.

Therefore, we can compute $\mathsf{IP2}(\vec{x}, \vec{y})$ by having a layer of $O(k)$ sigmoidal gates computing p_i, taking the sum of them, and then read out the $(r-1)l$-th bit using $O(rl)$ gates. By taking $k = \sqrt{n \log n}$, the circuit is of size $O(\sqrt{n \log n})$.

2.4 Equivalence between Sigmoidal and Threshold Circuits

We first explain the intuition behind Theorem 1.8. For simplicity, we assume $s(n), W(n) = n^{O(1)}$; the theorem then states that depth-two sigmoidal circuits with polynomial integer weights and real biases in $(-1, 1)$ compute, with no separation, the same functions as $\mathsf{THR} \circ \mathsf{MAJ}$ of polynomial size.

Simulating Sigmoid with $\mathsf{THR} \circ \mathsf{MAJ}$. The simulation of sigmoidal circuits follows from the observation that each gate in the bottom layer has $\mathsf{poly}(n)$ possible pre-activation values. Therefore, we can replace each sigmoidal gate in the bottom layer by a linear combination of MAJ gates (with no error). The top sigmoidal gate can be replaced by a threshold gate as $\mathsf{Sigmoid}$ is monotone, so the circuit computes the same Boolean function with no separation even before activation.

Simulating $\mathsf{THR} \circ \mathsf{MAJ}$ *with Sigmoid.* The main nontrivial direction is to simulate $\mathsf{THR} \circ \mathsf{MAJ}$ with small integer weight sigmoidal circuits. Without the constraint to have small integer weights, we can simulate $\mathsf{THR} \circ \mathsf{MAJ}$ as follows. Let $T \subseteq \mathbb{R}$ be a finite set of size t. For any distinct $\beta_1, \ldots, \beta_t \in (-1, 1)$, it can be proved that the matrix $(\mathsf{Sigmoid}(u + \beta_j))_{u \in T, j \in [t]}$ is invertible. Subsequently, every function $\phi : T \to \mathbb{R}$ can be represented as a linear combination of $f_j(u) := \mathsf{Sigmoid}(u + \beta_j)$. For each MAJ gate $G_i(\vec{x})$ in the bottom layer, letting $w_i \in \mathbb{R}$ be its weight in the top threshold gate, we can replace $w_i \cdot G_i(\vec{x})$ by a linear combination of at most t sigmoidal gates (by taking T as a

$\mathsf{poly}(n)$-sized set). The top threshold gate can thus be replaced by a sigmoidal gate applied to the sum of all sigmoidal gates.

This simulation does not work directly as the real coefficients to represent $w_i \cdot G_i(\vec{x})$ as linear combinations of sigmoidal gates are not guaranteed to be small integers; we can make the weights small by normalization, but they may no longer be integers. The key idea is to apply the above idea with the *derivative* of $\mathsf{Sigmoid}$. With essentially the same argument, we can represent $w_i \cdot G_i(\vec{x})$ as a linear combination of $g_j(u) := \mathsf{Sigmoid}'(u + \beta_j)$. We can further represent

$$\delta \cdot \mathsf{Sigmoid}'(u + \beta_j) \approx \mathsf{Sigmoid}(u + \beta_j + \delta) - \mathsf{Sigmoid}(u + \beta_j) \tag{4}$$

when δ is small. We then can encode the coefficients into the biases of sigmoidal gates, so the weights of the top gate are only ± 1. (With a more careful argument via the inverse function theorem, we can obtain an exact version of (4), see Lemma 7.1.)

Extension to $\mathsf{THR} \circ \mathsf{THR}$. The main obstacle to extend this argument to simulate $\mathsf{THR} \circ \mathsf{THR}$ is that the bottom gates may have large weights and thus the pre-activation values no longer come from a small set. To resolve this issue, we use a structural result that $\mathsf{THR} \subseteq \mathsf{DOR} \circ \mathsf{ETHR}$ [HP10], where DOR (disjoint OR) denotes an OR with a promise that at most one of the inputs is 1. Using a slightly different construction, we can simulate each of the ETHR gates approximately as a linear combination of $\mathsf{Sigmoid}'$ functions. By applying the construction as in (4), we can use differences of sigmoidal gates to simulate $\mathsf{Sigmoid}'$, which leads to the final simulation.

2.5 Related Work

Complexity of Threshold Circuits. The rising popularity of neural networks in the 80's and 90's have led researchers to study the complexity of threshold circuits. This study has proven challenging, and several basic questions remain wide open despite much attention. For depth-2 threshold circuits with polynomial integer weights either for gates in the first or second layer, exponential size lower bounds are known for the IP Mod 2 function [HMP+93, Nis93, For02, FKL+01]. However, the best lower bound against depth-2 threshold circuits without weight restriction is only $\Omega(n^{2.5}/\log^{O(1)}(n))$ [CTW26], which improves upon the near-quadratic lower bound [Tam16, ACW16]; these are built on Williams' algorithmic method (see [Wil14]).

For depth-3, no super polynomial size lower bounds are known, even if *all* gates have polynomial integer weights. In fact, establishing super linear lower bounds for depth-3 circuits with polynomial weights was proved relatively recently [KW16]; their $\Omega(n^{3/2}/\log^{O(1)} n)$ gate lower bound remains the best known to date.

Threshold vs Sigmoid networks. Both [Maa97b, DS96] pioneered the study of how threshold and signoid circuits compare in computing Boolean functions. With respect to sigmoid circuits, their focus is on *upper bounds* and they do not consider lower bounds for such networks. For their upper bounds they rely on the following: While threshold gates only depend on the sign of their pre-activation, a smooth activation $\gamma(x)$ possesses a curved "middle zone" around points x_0 where $\gamma''(x_0) \neq 0$, allowing a constant number of gates to locally approximate degree-2 polynomials. This, as well as the analog output of sigmoid gate has led researchers to believe that the use of sigmoids in neural architectures should lead in some cases to more efficient representations [KS97].

Several works have supported this belief. Maass, Schnitger, and Sontag [MSS94] used this property to show that $f(x, y) = \mathsf{Majority}(x) \oplus \mathsf{Majority}(y)$ can be computed with noticeable separation

by a depth-2 sigmoidal circuit with 5 gates, whereas depth-2 threshold circuits with noticeable separation require $\omega(1)$ gates, by exploiting the middle zone of Sigmoid function. Their lower bound holds for $\mathsf{THR} \circ \mathsf{THR}$ circuits and is proved using random restrictions for threshold gates. DasGupta and Schnitger [DS96] consider stronger separation against arbitrary threshold circuits (no depth or weight restriction): they showed that unary squaring, $\mathrm{SQ}(x, y) = \mathbb{I}[|x|^2 \leq |y|]$ for $x \in \{0,1\}^n$ and $y \in \{0,1\}^{n^2}$ (where $|z|$ is the Hamming weight of z), is computable by a depth-2 sigmoidal circuit of constant size, whereas TC^0 requires $\Omega(\log n)$ gates (which is nearly tight).

The "middle zone" idea is used (along with additional arguments) to establish the improved upper bound on the Inner Product function (see Theorem 1.7). On the other hand, our proofs or equivalence between depth-2 sigmoidal circuits computing with zero separation and $\mathsf{THR} \circ \mathsf{MAJ}$ circuits (e.g., Theorem 1.8) rely on different ideas than those in [MSS94, DS96].

Wire and gate complexity of Parity and Andreev's function. We are not aware of previous work studying the wire complexity of sigmoidal networks implementing Boolean functions (neither for upper or lower bounds). We are also not aware of previous study of the study of random restrictions in the context of sigmoidal networks with Boolean inputs and outputs.

For depth-2 majority circuits computing Parity, Paturi and Saks [PS94] used Newman's rational approximation of $|x|$ [New64, PP87] to prove an $\Omega(n^2/\log^2 n)$ wire lower bound and $\Omega(n)$ gate lower bound. Building on [PS94], Siu, Roychowdhury, and Kailath [SRK92, SRK94] proved nearly matching $\Theta(dn^{1/(d-1)})$ worst-case and $\Theta(n^{1/(2(d-1))})$ average-case *gate* lower bounds for depth-d circuits computing Parity (see also [Bru90, ABFR94, BBGL08, SÖ15, AT19] for related polynomial sign-representation bounds). Because the lower bounds of [SRK92, SRK94] rely on rational approximation, they apply only when weights and biases are polynomially bounded.

For general (unbounded-weight) threshold circuits of depth d, worst-case gate $(\Omega(n^{1/(2(d-1))}))$ and super-linear wire $(n^{1+\exp(-O(d))})$ lower bounds for Parity were later proved by Impagliazzo, Paturi, and Saks [IPS97] via random restrictions. Chen, Santhanam, and Srinivasan [CSS18] strengthened [IPS97] to the average case under the uniform distribution: they proved $n^{-\Omega(1)}$ correlation bounds for Parity against depth-d threshold circuits with $o(n^{1/(2(d-1))})$ gates or $n^{1+\epsilon_d}$ wires, and $\exp(-n^{\Omega(1)})$ correlation bounds for the Generalized Andreev function [And87, IN93, PZ93, Hås98, IMZ12, KR13]. Concurrently, Kane and Williams [KW16] proved $n^{1.5-o(1)}$ gate and $n^{2.5-o(1)}$ wire lower bounds against $\mathsf{MAJ} \circ \mathsf{THR} \circ \mathsf{THR}$ circuits. Kabanets, Kane, and Lu [KKL17] extended super-linear wire lower bounds to depth-d PTF circuits.

Our work shows that while the gate and depth-3 wire lower bounds of [KW16, CSS18] extend to polynomial-weight sigmoidal circuits with noticeable separation, the super-linear wire lower bound of [IPS97, CSS18] for Parity fails for sigmoidal circuits of depth $d \geq 3$ (Theorem 1.1).

VC dimension. When weights are unrestricted and separation is zero, VC dimension and related shattering bounds are the primary known tools for proving gate lower bounds [Koi96, KS97, Maa97b, AB02]. Using Khovanskii's fewnomial bounds [Kho91], Karpinski and Macintyre [KM97] upper bounded the VC dimension of sigmoidal and Pfaffian networks with s gates and w programmable parameters by $O(s^2w^2)$ (recent works have extended such VC dimension and expressivity analyses to Graph Neural Networks [Gro24, DBS25, DYZ+26]).

Koiran [Koi96] applied [KM97] to prove an $\Omega(n^{1/4})$ gate lower bound for the Boolean Selection (indexing) function on $\log_2 n + n$ bits against unrestricted sigmoidal circuits of arbitrary depth, while showing that Selection has depth-3 sigmoidal circuits of size $O(\sqrt{n})$ but requires $\Omega(n/\log n)$ gates

for unrestricted threshold circuits (Maass [Maa97b] proved an analogous $\Omega(n^{1/4})$ sigmoidal and $\Omega(n/\log n)$ threshold gate lower bound for the Boolean Coincidence Detection function $\mathrm{CD}_n(x,y) = \bigvee_{i=1}^{n}(x_i \wedge y_i)$).

2.6 Discussion

Our results reveal an interesting landscape for expressing Boolean functions using neural networks varying activations: for certain complexity measures and Boolean functions, sigmoidal and threshold circuits have similar circuit complexity, while for others, sigmoidal networks lead to significant savings. Some of our results apply to additional activation functions (e.g., ReLU activations), demonstrating noticeable differences of lower and upper bounds in terms of implementation of Boolean functions depending on whether the activation function used is threshold, sigmoid or ReLU. For example, our upper bound for representing the inner product function using sigmoid networks is provably better than what can be achieved by neural networks with ReLUs [VRPS21].

Our investigation of the complexity of sigmoidal circuits reveals several intriguing open problems:

1. Prove super-linear wire complexity lower bound for sigmoidal circuits with polynomial weights, large constant depth that computes a function with noticeable separation. The function is necessarily not a symmetric function by Theorem 1.1.
2. Prove a weight upper bound for depth-three unrestricted sigmoidal circuits and beyond; our weight upper bound uses change-of-variable trick that do not seem to work for depth three. Such weight upper bounds could be a first step in extending our lower bounds for depth 3 networks with polynomial weights (Theorem 1.4) to depth 3 networks with arbitrary weights. Alternatively, it might be that sigmoidal networks with unbounded weights and bounded depth (say, depth 3) in the zero separation model are significantly stronger than threshold circuits and can compute functions that are believed to be outside TC_0. Obtaining new results in this regard is an interesting future direction.

3 Preliminaries

Notation. Let $\mathsf{Sigmoid}(x) := (1+e^{-x})^{-1}$ and $\mathsf{Sigmoid}_{\vec{w},\theta}(\vec{x}) := \mathsf{Sigmoid}(\langle \vec{x}, \vec{w}\rangle + \theta)$. Let $\sigma : \mathbb{R} \to [0,1]$ be a function, a σ-circuit is a circuit with unbounded fan-in gates of form $\sigma_{\vec{w},\theta}(\vec{x}) := \sigma(\langle \vec{w}, \vec{x}\rangle + \theta)$, where $\vec{w} = (w_1, \ldots, w_k) \in \mathbb{R}^k$ is said to be the *weight* of the gate, and $\theta \in \mathbb{R}$ is said to be the *bias* of the gate. In particular, a Sigmoidal circuit is a σ-circuit with $\sigma := \mathsf{Sigmoid}$.

We also use $\mathsf{Sigmoid}$ to denotes a layer of Sigmoid gates, and thus $\mathsf{Sigmoid} \circ \mathsf{Sigmoid}$ denotes depth-two sigmoidal circuits.

For $x \in \{0,1\}^n$, we use $|x|$ to denote the Hamming weight of x. Similarly, for $x, y \in \{0,1\}^n$, we use $|x-y|$ to denote the Hamming distance between x and y. Let $\mathsf{sgn} : \mathbb{R} \to \{0,1\}$ where $\mathsf{sgn}(x) = 1$ iff $x > 0$.

Definition 3.1 (separation)**.** Let $\varepsilon = \varepsilon(n)$. We say that a σ-circuit C computes $f : \{0,1\}^n \to \{0,1\}$ with ε-separation, if there exists $\theta \in [0,1]$ such that for every $x \in \{0,1\}^n$,

$$\begin{aligned} f(x) = 0 \quad &\Longrightarrow \quad C(x) \le \theta \\ f(x) = 1 \quad &\Longrightarrow \quad C(x) > \theta + \varepsilon \end{aligned}$$

It is said to compute $f : \{0,1\}^n \to \{0,1\}$ with *noticeable separation* if it has separation $\Omega(n^{-k})$ for some $k \geq 1$. It is said to compute f with *no separation* if it has separation 0. It is said to compute f *exactly* if $C(x) = f(x)$ for every $x \in \{0,1\}^n$.

3.1 Lipschitz Property of σ-circuits

Definition 3.2. Let $f_1, f_2 : \{0,1\}^n \to \mathbb{R}$ be real valued functions on Boolean inputs. They are δ-almost ε-close to each other, if

$$\Pr_{\vec{x} \in \{0,1\}^n}[|f_1(x) - f_2(x)| \leq \varepsilon] \geq 1 - \delta,$$

and are said to be ε-close to each other if they are 0-almost ε-close to each other.

Definition 3.3 (Lipschitz functions)**.** A function $\sigma : \mathbb{R}^n \to [0,1]$ is said to be κ-Lipschitz if for every $\vec{x}, \vec{y} \in \mathbb{R}^n$,

$$|\sigma(\vec{x}) - \sigma(\vec{y})| \leq \kappa \cdot \|\vec{x} - \vec{y}\|_1.$$

Lemma 3.4 (Lipschitz property of σ-circuits)**.** *Let $\sigma : \mathbb{R} \to [0,1]$ be a κ-Lipschitz function. For every depth-d σ-circuit C with n inputs, s wires, and the maximum absolute value of weights is $W \in \mathbb{R}$ such that $\kappa s W > 1$, it computes a $(\kappa s W)^d$-Lipschitz function.*

Proof. Let $\kappa_d := (\kappa s W)^d$, we prove this lemma by induction on d. The base case $d = 0$ is trivial, so it suffices to consider the induction case.

Let G be the output gate and $C_1, \ldots, C_k$ be the sub-circuits that feed G, $k \leq s$, each of size at most s. Let $\vec{x}_1, \vec{x}_2 \in \mathbb{R}^n$ be inputs and $\delta = \|\vec{x}_1 - \vec{x}_2\|_1$, we need to prove that

$$|G(\vec{c}_1) - G(\vec{c}_2)| \leq \kappa_d \cdot \delta,$$

where $\vec{c}_i := (C_1(\vec{x}_i), \ldots, C_k(\vec{x}_i))$ for $i \in \{1,2\}$. By the induction hypothesis, each sub-circuit C_j computes a κ_{d-1}-Lipschitz function, and thus $\|\vec{c}_1 - \vec{c}_2\|_1 \leq \kappa_{d-1} \cdot \delta k \leq \kappa_{d-1} \cdot \delta s$.

It then suffices to prove that G is a (κW)-Lipschitz function. Recall that each gate of a σ-circuit is of form $\sigma(\langle \vec{w}, \vec{x} \rangle + \theta)$, where $\vec{w} = (w_1, \ldots, w_k) \in [-W, W]^k$. Then for $\vec{c}_1, \vec{c}_2 \in \mathbb{R}^k$,

$$\begin{aligned}
|G(\vec{c}_1) - G(\vec{c}_2)| &\leq |\sigma(\langle \vec{w}, \vec{c}_1 \rangle + \theta) - \sigma(\langle \vec{w}, \vec{c}_2 \rangle + \theta)| \\
&\leq \kappa \cdot |\langle \vec{w}, \vec{c}_1 \rangle - \langle \vec{w}, \vec{c}_2 \rangle| && (\sigma \text{ is } \kappa\text{-Lipschitz}) \\
&= \kappa \cdot |\langle w, \vec{c}_1 - \vec{c}_2 \rangle| \\
&\leq \kappa W \cdot \|\vec{c}_1 - \vec{c}_2\|_1. && (\vec{w} \in [-W, W]^k)
\end{aligned}$$

This completes the proof. □

3.2 Analysis of Boolean Functions

Definition 3.5 (correlation)**.** The *correlation* between Boolean functions $f, g : \{0,1\}^n \to \{0,1\}$ is the bias of their agreement probability over a uniformly random input:

$$\mathsf{Cor}(f, g) := \left| \Pr_{\vec{x} \leftarrow \{0,1\}^n}[f(\vec{x}) = g(\vec{x})] - \tfrac{1}{2} \right|.$$

Definition 3.6 (ε-balanced)**.** A Boolean function $f : \{0,1\}^n \to \{0,1\}$ is said to be *ε-balanced* if $\left|\Pr_{\vec{x}\leftarrow\{0,1\}^n}[f(\vec{x}) = 1] - \frac{1}{2}\right| \le \varepsilon$.

We will need the following lemma that converts a discriminator-type lower bound into a lower bound on $\mathsf{Cor}(f, g)$.

Lemma 3.7. *Let $f : \{0,1\}^n \to \{0,1\}$ be ε-balanced, and suppose $g : \{0,1\}^n \to \{0,1\}$ satisfies*

$$\left|\Pr_{\vec{x}\leftarrow\{0,1\}^n}[g(\vec{x}) = 1 \mid f(\vec{x}) = 1] - \Pr_{\vec{x}\leftarrow\{0,1\}^n}[g(\vec{x}) = 1 \mid f(\vec{x}) = 0]\right| \ge \delta.$$

Then $\mathsf{Cor}(f, g) \ge \delta/2 - \varepsilon$.

Proof. Define $p_b := \Pr_{\vec{x}\leftarrow\{0,1\}^n}[g(\vec{x}) = 1 \mid f(\vec{x}) = b]$ for $b \in \{0,1\}$, and $\eta := \Pr_{\vec{x}\leftarrow\{0,1\}^n}[f(\vec{x}) = 1] - \frac{1}{2}$; by the assumption, $|p_1 - p_0| \ge \delta$ and $|\eta| \le \varepsilon$. We can see that

$$\begin{aligned}\Pr_{\vec{x}\leftarrow\{0,1\}^n}[f(\vec{x}) = g(\vec{x})] &= \Pr_{\vec{x}\leftarrow\{0,1\}^n}[f(\vec{x}) = 1] \cdot p_1 + \Pr_{\vec{x}\leftarrow\{0,1\}^n}[f(\vec{x}) = 0] \cdot (1 - p_0) \\ &= \left(\tfrac{1}{2} + \eta\right)p_1 + \left(\tfrac{1}{2} - \eta\right)(1 - p_0) \\ &= \tfrac{1}{2} + \tfrac{1}{2}(p_1 - p_0) + \eta\,(p_1 + p_0 - 1).\end{aligned}$$

Subsequently, by $|p_1 + p_0 - 1| \le 1$, we have

$$\mathsf{Cor}(f, g) \ge \tfrac{1}{2}|p_1 - p_0| - |\eta| \cdot |p_1 + p_0 - 1| \ge \tfrac{\delta}{2} - \varepsilon. \qquad \square$$

We need the definition of average sensitivity of Boolean functions, and a theorem from [Kan14] on the average sensitivity of polynomial threshold functions.

Definition 3.8 (average sensitivity)**.** Let $f : \{0,1\}^n \to \{0,1\}$. The *sensitivity* of f at the input x is defined as

$$\mathsf{S}(f, x) := \sum_{i=1}^{n} |f(\vec{x}) - f(\vec{x}^{(i)})|,$$

where $\vec{x}^{(i)}$ denotes the input with the i-th bit flipped. The *average sensitivity* of f is defined as $\mathsf{AS}(f) := \mathbb{E}_{\vec{x}}[\mathsf{S}(f, x)]$, where $\vec{x}$ is sampled uniformly from $\{0,1\}^n$.

Theorem 3.9 ([Kan14])**.** *Let $p(\vec{x})$ be a degree-d multilinear polynomial and $f(\vec{x}) := \mathsf{sgn}(p(\vec{x}))$, then $\mathsf{AS}(f) \le \sqrt{n}(\log n)^{O(d \log d)} 2^{O(d^2 \log d)}$.*

3.3 Relationship with TC^0 Circuits

We now consider the relationship between Sigmoidal circuits and TC^0 circuits, i.e., linear threshold circuits. It is known that these two circuit model (assuming polynomial weights) simulates each other within polynomial overhead. However, from a more fine-grained view, simulation of Sigmoidal circuits with TC^0 incurs an substantial size blowup, while the other direction is almost for free.

Simulating Sigmoidal circuits with TC^0. The result of Maass, Schnitger, and Sontag [MSS94] shows that $\mathsf{poly}(n)$-weight depth-d Sigmoidal circuits can be simulated by depth-d threshold circuits with $\mathsf{poly}(n)$ size overhead.

This simulation is obtained by replacing Sigmoid gates in depth $1, 2, \ldots, d$, where depth 1 is the layer fed by input variables, by a step function (using a linear combination of threshold gates) that approximates Sigmoid. The correctness of the simulation is derived from the sensitivity of Sigmoidal circuits, see, Lemma 3.4. By a close inspection of the proof, we can obtain:

Lemma 3.10 (implicit in [MSS94])**.** *Let $C(x)$ be a Sigmoidal circuit with depth d, size s, and weight within $[-W, W]$ that computes the Boolean function $f : \{0,1\}^n \to \{0,1\}$ with separation ε. Then there exists a TC^0 circuit C' that computes the same Boolean function f of depth d, size $\widetilde{O}((Ws)^d \cdot \varepsilon^{-1})$, and with integer weights of $O_d(\log((n+s)W/\varepsilon))$ bits.*

As a corollary, sigmoidal circuits with polynomial weights (i.e. $W \le \mathsf{poly}(n)$) can be simulated by Majority circuits with polynomial size overhead. In particular:

Corollary 3.11. *Any function $f : \{0,1\}^n \to \{0,1\}$ computable by depth-two, polynomial weights, and polynomial sized Sigmoidal circuits with noticeable separation can also be computed by polynomial-sized $\mathsf{MAJ} \circ \mathsf{MAJ}$ circuits.*

Corollary 3.12 ([HMP+93], see also [Raz92])**.** *Inner Product function $\mathsf{IP}(x, y) := \langle x, y \rangle \bmod 2$ is not computable by polynomial sized $\mathsf{Sigmoid} \circ \mathsf{Sigmoid}$ circuits with polynomial weights and noticeable separation.*

> **Remark 3.** Note that the simulation incurs a size overhead that is polynomial in the weight W (i.e. exponential in the bit length of weights). As a result, more fine-grained lower bounds, such as the gate and wire complexity lower bounds against threshold circuits due to Impagliazzo, Paturi, and Saks [IPS97], do not immediately yield nontrivial lower bounds against Sigmoidal circuits.

Simulating TC^0 with Sigmoidal circuits. We can also simulate TC^0 with Sigmoidal circuits at almost no cost. Formally:

Lemma 3.13. *Let $C(x)$ be a TC^0 circuit with depth d, size s, and integer weight at most W that computes Boolean function $f : \{0,1\}^n \to \{0,1\}$. Then there is a Sigmoidal circuit with depth d, size s, and weight at most $\widetilde{O}(sW)$ that computes f with noticeable separation.*

Proof Sketch. We replace each threshold gate $G(x) := \mathsf{sgn}(\langle \vec{w}, \vec{x} \rangle + \theta)$ by the following sigmoid gate $G'(x) := \mathsf{Sigmoid}(\langle \vec{w}', \vec{x} \rangle + \theta')$, where

$$\vec{w}' = (\log^2(sW)) \cdot \vec{w}, \quad \theta' = \log^2(sW) \cdot (\theta - 1/2),$$

shifting the bias by $1/2$ to account for $\langle \vec{w}, \vec{x} \rangle + \theta = 0$.

Fix any input $\vec{x}$ to the circuit. We can prove by induction on the topological order of gates that for each gate G' in the new Sigmoidal circuit, the output of G' differs from the output of G by at most $1/(10sW)$, and in particular, the output is within $b \pm 1/(10sW)$ for some $b \in \{0,1\}$. Therefore, the output of C' is within $C(x) \pm 1/(10sW)$, which implies that it computes f with noticeable separation. □

Note that the proof above utilizes the property that $\mathsf{Sigmoid}(x)$ converges exponentially to 0 (resp. 1) for $x \to -\infty$ (resp. $x \to +\infty$).

Remark 4. Subsequently, super-polynomial lower bounds for depth-two Sigmoidal circuits with unbounded weights would imply super-polynomial lower bounds for depth-two TC^0 circuits, a major open problem in circuit complexity.

Another corollary is that the size blowup in Lemma 3.10 is likely unavoidable if we require the threshold circuits to have small weights. Recall that there is an explicit function in $\mathsf{THR} \circ \mathsf{THR}$ that requires $2^{\Omega(n^{1/4})}$-size $\mathsf{THR} \circ \mathsf{MAJ}$ circuits [CM22]. As $\mathsf{THR} \circ \mathsf{THR}$ can be simulated by $\mathsf{Sigmoid} \circ \mathsf{Sigmoid}$, we have that:

Corollary 3.14. *There is a function $f : \{0,1\}^n \to \{0,1\}$ computable by depth-two polynomial-sized Sigmoidal circuits that requires $2^{\Omega(n^{1/4})}$-size to compute by $\mathsf{THR} \circ \mathsf{MAJ}$ circuits.*

4 Wire Complexity for Sigmoidal Circuits

In this section, we explore interesting *wire complexity* upper and lower bounds for sigmoidal circuits. Note that sparse sigmoidal circuits (i.e. circuits with few wires) models *sparse* neural networks, where some of the weights are fixed to be 0.

4.1 Symmetric Functions using Linear Wires

We prove a wire-complexity upper bound for *every* symmetric Boolean function. Recall that $f : \{0,1\}^n \to \{0,1\}$ is *symmetric* if $f(\vec{x})$ depends only on the Hamming weight $|\vec{x}|$, i.e., $f(\vec{x}) = F(|\vec{x}|)$ for some $F : \{0,1,\dots,n\} \to \{0,1\}$.

Theorem 4.1. *For every symmetric Boolean function $f : \{0,1\}^n \to \{0,1\}$, there is a depth-three sigmoidal circuit with $O(n)$ wires and polynomial weights that computes f with separation $\Omega(1)$.*

Proof. Let $\gamma = 1/6$. We can see that $\mathsf{Sigmoid}'(\xi) \geq \gamma$ for all $\xi \in [0,1]$, as $\mathsf{Sigmoid}'(1) = e/(1+e)^2 > 1/6$ and $\mathsf{Sigmoid}'$ is decreasing on $[0,\infty)$. By the mean value theorem, $\mathsf{Sigmoid}(v) - \mathsf{Sigmoid}(u) \geq \gamma(v-u)$ for all $0 \leq u \leq v \leq 1$. We construct the circuit layer by layer.

Bottom layer. A single gate $\mathsf{Sigmoid}_{\mathsf{sum}}(\vec{x}) := \mathsf{Sigmoid}\left(\frac{1}{n}\sum_{i=1}^n x_i\right) = \mathsf{Sigmoid}(|\vec{x}|/n)$ of fan-in n. Its output is a real-valued digest of the Hamming weight: it equals $\mathsf{Sigmoid}(m/n)$ when $|\vec{x}| = m$, and consecutive weights are separated by $\mathsf{Sigmoid}((m+1)/n) - \mathsf{Sigmoid}(m/n) \geq \gamma/n$.

Middle layer. Let $A := \lceil (2n/\gamma)\ln(10n) \rceil = O(n \log n)$, and for each $k \in [n]$ let $\tau_k := \mathsf{Sigmoid}((k - \frac{1}{2})/n)$ and

$$V_k(\vec{x}) := \mathsf{Sigmoid}\big(A \cdot (\mathsf{Sigmoid}_{\mathsf{sum}}(\vec{x}) - \tau_k)\big),$$

a gate of fan-in *one* that reads only the output of the gate $\mathsf{Sigmoid}_{\mathsf{sum}}$. We first show that $|V_k(\vec{x}) - \mathbb{I}[\,|\vec{x}| \geq k\,]| \leq 1/(10n)$:

- If $|\vec{x}| \geq k$ then $\mathsf{Sigmoid}_{\mathsf{sum}}(\vec{x}) - \tau_k \geq \mathsf{Sigmoid}(k/n) - \mathsf{Sigmoid}((k - \frac{1}{2})/n) \geq \gamma/(2n)$. Then $V_k(\vec{x}) \geq \mathsf{Sigmoid}(\ln(10n)) = 1 - \frac{1}{10n+1} \geq 1 - \frac{1}{10n}$.
- Otherwise, symmetrically, $V_k(\vec{x}) \leq \mathsf{Sigmoid}(-\ln(10n)) \leq 1/(10n)$.

Output gate. Since f is symmetric, $f(\vec{x}) = F(|\vec{x}|)$ for some function $F : \{0,1,\dots,n\} \to \{0,1\}$. Notice that $F(m) = F(0) + \sum_{k=1}^n c_k\, \mathbb{I}[m \geq k]$ with $c_k := F(k) - F(k-1) \in \{-1,0,1\}$. Let

$$C(\vec{x}) := \mathsf{Sigmoid}\Big(4\big(F(0) + \textstyle\sum_{k=1}^n c_k V_k(\vec{x}) - \frac{1}{2}\big)\Big),$$

a gate of fan-in n reading the middle layer gates. Let $L(\vec{x}) := F(0)+\sum_k c_k V_k(\vec{x})$. Since $\sum_k |c_k| \le n$, we have

$$|L(\vec{x}) - F(|\vec{x}|)| = \left|\sum_k c_k\big(V_k(\vec{x}) - \mathbb{I}[\,|\vec{x}| \ge k\,]\big)\right| \le n \cdot \tfrac{1}{10n} = \tfrac{1}{10}.$$

We can verify that the circuit computes f with separation at least $1/4$: if $f(\vec{x}) = 1$ then $L(\vec{x}) \ge \frac{9}{10}$ and $C(\vec{x}) \ge \mathsf{Sigmoid}(8/5) > \frac{3}{4}$; if $f(\vec{x}) = 0$ then $L(\vec{x}) \le \frac{1}{10}$ and $C(\vec{x}) \le \mathsf{Sigmoid}(-8/5) < \frac{1}{2}$. The circuit has depth three and $n + n + n = O(n)$ wires, and all weights and biases are at most $A = O(n \log n)$ in absolute value. □

Since the Parity function is symmetric, Theorem 4.1 applies to it.

Corollary 4.2. *There is a depth-three sigmoidal circuit with $O(n)$ wires and polynomial weights that computes $\oplus_n$ with separation $\Omega(1)$.*

The construction illustrates a concrete advantage of sigmoidal gates: It is known that the parity function requires super-linear wires for threshold circuits [IPS97] when the depth is an arbitrary constant and weights are arbitrary, and even for average-case computation [CSS18].

4.2 Wire Complexity for Depth-2 Sigmoidal Circuits

To fully close the question of wire complexity for sigmoidal circuits, we prove tight upper and lower bounds for depth-two sigmoidal circuits (up to polylogarithmic factors).

Proposition 4.3. *For every symmetric function $f : \{0,1\}^n \to \{0,1\}$, there is a depth-two polynomial-weight sigmoidal circuit with $O(n^2)$ wires that computes f with separation $\Omega(1)$.*

Proof Sketch. For each $k \in \{0, 1, \ldots, n\}$ such that f accepts strings with Hamming weight k, create a gate $V_k(\vec{x})$ that outputs ≈ 1 when $|\vec{x}| = k$, and ≈ 0 otherwise. The output gate is a top sigmoidal gate that outputs ≈ 1 when the sum of $V_k \gg 0$, and ≈ 0 otherwise. □

Recall that Paturi and Saks [PS94] proved an $\Omega(n^2/\log^2 n)$ wire complexity lower bound for depth-two threshold circuits with polynomially bounded integer weights computing Parity. They indeed show that depth-two threshold circuits with wire complexity E and integer weights within $[-W, W]$ can be represented as the sign of a *rational function* of degree $O(\sqrt{E}\log(nW))$.

It can be observed that the result generalizes to sigmoidal circuits by combining [PS94] with a rational approximation of $\mathsf{Sigmoid}$. We first state a generalized version of [PS94, Theorem 3]. Recall that the degree of a rational function p/q is defined as $\max\{\deg p, \deg q\}$.

Theorem 4.4 (Adapted from [PS94, Theorem 3])**.** *Let $c > 0, c_0 \ge 1$ be constants, and $\sigma : \mathbb{R} \to [0,1]$ be any non-decreasing κ-Lipschitz function satisfying the following conditions: For every $k \ge 1$, $a \ge 1$, there is a rational function $r_k(x)$ of degree at most k with no pole in $[-a, a]$ such that for every $x \in [-a, a]$, $|r_k(x) - \sigma(x)| \le c_0 a^{c_0} \cdot e^{-k^{1/c}/c_0}$.*

Then for any depth-two σ-circuit $C : \{0,1\}^n \to [0,1]$ with weights and biases within $[-W, W]$ and has at most E wires, if C computes a Boolean function $f : \{0,1\}^n \to \{0,1\}$ with separation δ, there is a rational function $r : \{0,1\}^n \to \mathbb{R}$ of degree $O(\sqrt{E}\log^{c/2}(n\kappa EW/\delta))$ that sign-represent f, i.e., $f(x) = \mathsf{sgn}(r(x))$ for every $x \in \{0,1\}^n$.

Proof Sketch. We follow the proof of [PS94, Theorem 3]. Fix the circuit C and let $\sigma_1, \dots, \sigma_t$ be the gates in the bottom layer, w_i be the weight for σ_i in the top σ-gate, and $\ell(\vec{x})$ be the affine function consisting of the bias of the top σ-gate and its wires from the input variables. As σ is non-decreasing and κ-Lipschitz, there exists ζ_0, ζ_1 such that $\zeta_1 - \zeta_0 \geq \delta/\kappa$, and $f(\vec{x}) = 1$ (resp. $f(\vec{x}) = 0$) iff $\sum_i w_i \sigma_i(\vec{x}) + \ell(\vec{x}) \geq \zeta_1$ (resp. $\sum_i w_i \sigma_i(\vec{x}) + \ell(\vec{x}) \leq \zeta_0$).

Let $k, d \geq 1$ be parameters to be determined. We consider gates with large and small fan-in:

- (*Large Gates*). For each gate $\sigma_i(\vec{x})$ with fan-in larger than d (called *large gates*), we approximate it using the provided degree-k rational function $\hat{\sigma}_i$ with degree k. Note that $\sigma_i(\vec{x}) = \sigma(\langle \vec{w}, \vec{x} \rangle + \beta)$, where $|\langle \vec{w}, \vec{x} \rangle + \beta| \leq (n+1)W$. It then suffices to instantiate the rational approximation with $a := (n+1)W$, so the error is $O((nW)^{c_0} \cdot e^{-k^{1/c}/c_0})$.
- (*Small Gates*). For each gate $\sigma_j(\vec{x})$ with fan-in at most d (called small gates), we use a degree-d polynomial $\hat{\sigma}_j(\vec{x})$ to compute it exactly over $\{0,1\}^n$.

Let $L := \{i \in [t] \mid \sigma_i \text{ is large}\}$. Since there are at most E/d large gates, $\sum_{i \in L} w_i \hat{\sigma}_i(\vec{x})$ is a rational function with degree at most kE/d. Meanwhile, we can observe that

$$\left| \sum_{i \in L} w_i \sigma_i(\vec{x}) - \sum_{i \in L} w_i \hat{\sigma}_i(\vec{x}) \right| = O\left(EW \cdot (nW)^{c_0} \cdot e^{-k^{1/c}/c_0} \right) = (nEW)^{O(1)} \cdot e^{-k^{1/c}/c_0}.$$

By setting $k := \Theta(\log^c(n\kappa EW/\delta))$, the error is at most $\delta/(3\kappa)$. This implies that $\sum_i w_i \hat{\sigma}_i(\vec{x})$ approximates $\sum_i w_i \sigma_i(\vec{x})$ within error $\delta/(3\kappa)$, and subsequently $R(x) := \sum_i w_i \hat{\sigma}_i(\vec{x}) + \ell(\vec{x}) - (\zeta_1 + \zeta_0)/2$ sign-represent f, as desired. Note that the degree of $R(\vec{x})$ is $kE/d + d$, which is minimized when we set $d := \sqrt{kE}$. □

The following lemma gives the rational approximation of Sigmoid that we need. Its proof follows [SRK94] and is based on Newman's theorem on the rational approximation of $|x|$ [New64]; we defer it to Section B.3.

Lemma 4.5 (Rational approximation of Sigmoid)**.** *There is a constant C_0 such that for every $a \geq 1$ and $\varepsilon \in (0, 1/2]$, there is a rational function r of degree at most $C_0 \ln^2(a/\varepsilon)$, with no pole in $[-a, a]$, such that $|r(x) - \mathsf{Sigmoid}(x)| \leq \varepsilon$ for every $x \in [-a, a]$.*

Corollary 4.6. *Any function that can be computed by a depth-two sigmoidal circuit containing E wires and using polynomial weights and biases with noticeable separation, can be sign-represented by a rational function of degree $O(\sqrt{E} \log n)$. Consequently, any depth-two sigmoidal circuit with polynomial weights and biases that compute Parity with noticeable separation requires $\Omega(n^2/\log^2 n)$ wires.*

Proof. Let $\sigma := \mathsf{Sigmoid}$, which is non-decreasing and κ-Lipschitz for $\kappa = 1/4$, and by Lemma 4.5 it satisfies the condition in Theorem 4.4 with $c = 2$. Now let C be a depth-two sigmoidal circuit with E wires and weights within $[-W, W]$, where $1 \leq W \leq n^{O(1)}$, that computes $f : \{0,1\}^n \to \{0,1\}$ with separation $\delta \geq n^{-O(1)}$. If $E \geq n^2$, then f is sign-represented by a polynomial of degree at most $n \leq \sqrt{E}$ (e.g., the multilinear extension of f minus $1/2$). Otherwise, Theorem 4.4 yields a rational function of degree $O(\sqrt{E} \log(n\kappa EW/\delta)) = O(\sqrt{E} \log n)$ that sign-represents f.

For the second part, suppose that a rational function P/Q sign-represents Parity, where Q has no zero on $\{0,1\}^n$. Then the polynomial $P \cdot Q$ also sign-represents Parity, so $\deg P + \deg Q \geq n$ by [PS94, Lemma 3] (see also [PS94, Corollary 1]). Hence the degree of P/Q is at least $n/2$, and subsequently $O(\sqrt{E} \log n) \geq n/2$, i.e., $E = \Omega(n^2/\log^2 n)$. □

5 Lower Bounds via Random Restriction

In this section, we prove two lower bounds against sigmoidal circuits with noticeable separation via random restriction methods.

5.1 Random Restriction Lemmas for Sigmoid Functions

We first develop lemmas that analyze the behavior of linear and sigmoidal functions under random restrictions, using random restriction lemmas for threshold functions [KW16, KKL17] and the analytic property of the Sigmoid function.

5.1.1 Analytic Properties of Sigmoid Functions

Theorem 5.1 (Cauchy's coefficient formula, see, e.g., [FS09, Theorem IV.7]). *Let λ be a positively oriented simple loop around 0, and $f(z)$ be analytic on and inside λ, then*

$$\frac{f^{(k)}(0)}{k!} = \frac{1}{2i\pi} \int_{\lambda} \frac{f(z)}{z^{k+1}} \, \mathrm{d}z.$$

Lemma 5.2 (see, e.g., [How03, Section 5.4]). *Let λ be a contour, and f be a continuous complex function such that $|f(z)| \leq M$ for all $z \in \lambda$, then*

$$\left| \int_{\lambda} f(z) \, \mathrm{d}z \right| \leq M \cdot \ell(\lambda),$$

where $\ell(\lambda)$ denotes the length of λ.

Corollary 5.3. *For any $k \geq 0$ and $\xi \in \mathbb{R}$, $|\mathsf{Sigmoid}^{(k)}(\xi)| \leq k!/(\pi/2)^k$.*

Proof. Let $g(z) = \mathsf{Sigmoid}(z + \xi)$ be the shifted function defined over $\mathbb{C}$; clearly, $g^{(k)}(0) = \mathsf{Sigmoid}^{(k)}(\xi)$. The function $g(z)$ is analytic at 0. Moreover, the singular points of $g(z)$ are exactly $z = -(\xi + i\pi(2n+1))$ for $n \in \mathbb{Z}$, $|z| \geq \sqrt{\xi^2 + (2n+1)^2\pi^2} \geq \pi$, and thus $g(z)$ is analytic in $\{z \in \mathbb{C} \mid |z| < \pi\}$.

Let λ be the positively oriented circle with radius $\pi/2$. Note that $|g(z)| \leq 1$ within the circle. By Cauchy's coefficient formula, we can see that

$$\left| \frac{g^{(k)}(0)}{k!} \right| \leq \frac{1}{2\pi} \left| \int_{\lambda} \frac{g(z)}{z^{k+1}} \, \mathrm{d}z \right| \leq \frac{1}{2\pi} \left(\frac{\pi}{2}\right)^{-(k+1)} \pi^2 \leq \left(\frac{\pi}{2}\right)^{-k},$$

where the second inequality follows from Lemma 5.2. □

5.1.2 Random Restriction on Linear and Sigmoidal Functions

We start by defining a few concepts. A *restriction* is a pair $\rho = (I, \vec{y})$, where $I \subseteq [n]$ is the set of free variables and $\vec{y} \in \{0,1\}^{[n]\setminus I}$. We use $f|_\rho$ to denote the function on the free variables obtained from f by fixing x_i to y_i for every $i \notin I$. For any fixed partition $\mathscr{P}$ of $[n]$ into m blocks, we use $\mathcal{R}_{\mathscr{P}}$ to denote the distribution of restrictions in which I contains exactly one uniformly random variable of each block, chosen independently for different blocks, and $\vec{y}$ is uniformly random.

Lemma 5.4 ([KW16])**.** *Let $f : \{0,1\}^n \to \mathbb{R}$ be any linear function, and $g(\vec{x}) := \mathsf{sgn}(f(\vec{x}))$ be a linear threshold function. Then for any partition $\mathscr{P}$ of $[n]$ into m blocks of equal size,*

$$\Pr_{\rho \leftarrow \mathcal{R}_{\mathscr{P}}}[g|_\rho \text{ is not a constant function}] \le O(m/\sqrt{n}).$$

For a linear function $f(\vec{x}) = \langle \vec{w}, \vec{x} \rangle + \theta$ and $a \le b$, δ, we say that $f(\vec{x})$ is δ-concentrated within $[a,b]$ if there exists $z \in [a,b]$ such that for every $\vec{x} \in \{0,1\}^n$, $|f(\vec{x}) - z| \le \delta$. The following lemma is a simple corollary of the restriction lemma above.

Lemma 5.5. *For any linear function $f(\vec{x})$ with n variables and any partition $\mathscr{P}$ of $[n]$ into m blocks of equal size, $m \le n/16$, $\delta > 0$, and a, b such that $b - a \ge \delta$, over a block random restriction $\rho \leftarrow \mathcal{R}_{\mathscr{P}}$, with probability $1 - O(m(b-a) \cdot \delta^{-1}/\sqrt{n})$, at least one of the following two conditions holds:*

1. *$f|_\rho(\vec{z}) \le a$ for every $\vec{z} \in \{0,1\}^m$, or $f|_\rho(\vec{z}) \ge b$ for every $\vec{z} \in \{0,1\}^m$;*
2. *$f|_\rho$ is δ-concentrated within $[a,b]$.*

Proof. Let $N \in \mathbb{N}$ such that $\delta/2 \le (b-a)/N \le \delta$, let $\delta' := (b-a)/N$, and $\tau_i := a + i\delta'$ for $i = 0, 1, \ldots, N$, so $\tau_0 = a$ and $\tau_N = b$. For each i, let $g_i(\vec{x}) := \mathsf{sgn}(f(\vec{x}) - \tau_i)$, which is a linear threshold function. Applying Lemma 5.4 to each g_i and by union bound over i, with probability $1 - O(Nm/\sqrt{n}) = 1 - O(m(b-a)\delta^{-1}/\sqrt{n})$ every $g_i|_\rho$ is constant.

Condition on this event. By monotonicity in i, the tuple $(g_0|_\rho, \ldots, g_N|_\rho) \in \{0,1\}^{N+1}$ has the form $(1, \ldots, 1, 0, \ldots, 0)$; let $i^* \in \{-1, \ldots, N\}$ be the index of the last 1 (with $i^* = -1$ if all entries are 0). If $i^* = -1$ (resp. $i^* = N$), then $f|_\rho \le \tau_0 = a$ (resp. $f|_\rho > \tau_N = b$) on $\{0,1\}^m$, yielding condition 1. Otherwise $\tau_{i^*} < f|_\rho \le \tau_{i^*+1} = \tau_{i^*} + \delta'$ on $\{0,1\}^m$, which satisfies condition 2 with $z := \tau_{i^*} + \delta'/2$. □

By a case analysis, we can also show that after a (block) random restriction, a sigmoid-like gate will likely be approximable by a low-degree polynomial.

Corollary 5.6. *Let $n, t \ge 20$, $m \le n/16$, $k \ge 0$, $\vec{w} \in \mathbb{R}^n$, $\theta \in \mathbb{R}$. Then for any partition $\mathscr{P}$ of $[n]$ into m blocks of equal size, over $\rho \leftarrow \mathcal{R}_{\mathscr{P}}$, with probability at least $1 - O(mt(k+1)\ln t/\sqrt{n})$, there is a degree-$k$ multilinear polynomial $p : \mathbb{R}^m \to \mathbb{R}$ such that for every $\vec{x} \in \{0,1\}^m$,*

$$\left|\mathsf{Sigmoid}_{\vec{w},\theta}|_\rho(\vec{x}) - p(\vec{x})\right| \le t^{-(k+1)},$$

where $O(\cdot)$ hides an absolute constant. Moreover, every coefficient of p has absolute value at most 2^{k+1}.

Proof. Let $f(\vec{x}) := \langle \vec{w}, \vec{x} \rangle + \theta$, and $a > 0$ be determined later. By Lemma 5.5, we know that over $\rho \leftarrow \mathcal{R}_{\mathscr{P}}$, with probability at least $1 - O(mta/\sqrt{n})$, $f|_\rho$ is either $(1/t)$-concentrated within $[-a, a]$, or $f|_\rho \le -a$ on $\{0,1\}^m$, or $f|_\rho \ge a$ on $\{0,1\}^m$. We show that in either case, it can be approximated by a degree-k polynomial.

- (*Case 1*). Suppose that $f|_\rho \le -a$ on $\{0,1\}^m$ (the case $f|_\rho \ge a$ is symmetric). Then $0 < \mathsf{Sigmoid}_{\vec{w},\theta}|_\rho(\vec{z}) \le e^{-a}$ for every $\vec{z} \in \{0,1\}^m$. We set $a := (k+1)\ln t$, so it suffices to choose $p \equiv 0$ (resp. $p \equiv 1$ in the symmetric case).

- (*Case 2*). Suppose that $f|_\rho$ is $(1/t)$-concentrated at $\mu \in [-a, a]$, i.e., $|f|_\rho(\vec{z}) - \mu| \le 1/t$ for every $\vec{z} \in \{0,1\}^m$. Let q be the degree-k Taylor approximation of $\mathsf{Sigmoid}$ at μ. Taylor's theorem shows that for every y with $|y - \mu| \le 1/t$, there exists a $\xi \in \mathbb{R}$ such that

$$|\mathsf{Sigmoid}(y) - q(y)| \le \frac{|\mathsf{Sigmoid}^{(k+1)}(\xi)|}{(k+1)!} \cdot |y - \mu|^{(k+1)} \le t^{-(k+1)},$$

  where the last inequality follows from Corollary 5.3. It then suffices to take $p(\vec{x}) := q(f|_\rho(\vec{x}))$, which is a degree-$k$ polynomial in $\vec{x}$ as $f|_\rho$ is a linear function.

In both cases, we may assume p is multilinear as the input is Boolean. For the moreover part, note that $|p(\vec{z})| \le 1 + t^{-(k+1)} \le 2$ for any $\vec{z} \in \{0,1\}^m$, as $\mathsf{Sigmoid}$ has range $(0,1)$. For every $S \subseteq [m]$ with $|S| \le k$, by the inclusion–exclusion principle, the coefficient of $\prod_{i \in S} z_i$ in p is $\sum_{T \subseteq S} (-1)^{|S|-|T|} p(\vec{1}_T)$, where $\vec{1}_T \in \{0,1\}^m$ is the indicator vector of T. Therefore, its absolute value is at most $2^{|S|} \cdot 2 \le 2^{k+1}$. □

5.1.3 Random Restriction for Bounded Fan-in Gates

We also need the following restriction lemma from [KW16] that works for sigmoidal gates with bounded fan-in. A Boolean function $g : \{0,1\}^n \to \{0,1\}$ is said to be *degenerated* if it depends on at most one of its input variables.

Lemma 5.7 ([KW16, Lemma 3.1])**.** *Let $f : \{0,1\}^n \to \mathbb{R}$ be any linear function that depends on at most ℓ of its input variables, and $g(\vec{x}) := \mathsf{sgn}(f(\vec{x}))$ be a linear threshold function. Then for any partition $\mathscr{P}$ of $[n]$ into m blocks of equal size,*

$$\Pr_{\rho \leftarrow \mathcal{R}_{\mathscr{P}}} [g|_\rho \text{ is not degenerated}] \le O(m^{3/2}\ell/n^{3/2}).$$

Lemma 5.8. *Let $f(\vec{x}) : \{0,1\}^n \to \mathbb{R}$ be a linear function that depends on at most ℓ of its input variables, $\mathscr{P}$ be a partition of $[n]$ into m blocks of equal size, $m \le n/16$, $\delta > 0$, and a, b such that $b - a \ge \delta$. Then over a block random restriction $\rho \leftarrow \mathcal{R}_{\mathscr{P}}$ with probability $1 - O(m^{3/2} \cdot \ell\delta^{-1}(b - a)/n^{3/2})$, there is a free variable such that for each restriction σ fixing it to 0 or 1, at least one of the following two conditions holds:*

1. *$f|_{\rho\circ\sigma} \le a$ on all inputs, or $f|_{\rho\circ\sigma} \ge b$ on all inputs;*
2. *$f|_{\rho\circ\sigma}$ is δ-concentrated within $[a, b]$.*

Proof Sketch. We follow the proof of Lemma 5.5. The key observation is that if a non-constant $\mathsf{sgn}(f|_\rho - \tau)$ depends on only one variable, then this variable has the strictly largest absolute weight in $f|_\rho$. Subsequently if that variable is fixed arbitrarily, all threshold functions must be constant simultaneously. □

Corollary 5.9. *Let $n, t \ge 20$, $m \le n/16$, $k \ge 0$, $\vec{w} \in \mathbb{R}^n$, $\theta \in \mathbb{R}$. Suppose that $\vec{w}$ has at most ℓ non-zero entries. Then for any partition $\mathscr{P}$ of $[n]$ into m blocks of equal size, over $\rho \leftarrow \mathcal{R}_{\mathscr{P}}$, with probability at least $1 - O(m^{3/2}t\ell(k+1)\ln t/n^{3/2})$, there is a degree-$(k+1)$ multilinear polynomial $p : \mathbb{R}^m \to \mathbb{R}$ such that for every $\vec{x} \in \{0,1\}^m$,*

$$\left|\mathsf{Sigmoid}_{\vec{w},\theta}|_\rho(\vec{x}) - p(\vec{x})\right| \le t^{-(k+1)}.$$

Moreover, every coefficient of p has absolute value at most 2^{k+2}.

Proof Sketch. We follow the proof of Corollary 5.6 using Lemma 5.8. The only difference is that after the restriction ρ, we find the free variable x provided by Lemma 5.8 and construct two degree-k polynomials p_0, p_1 that approximate $\mathsf{Sigmoid}_{\vec{w},\theta}|_\rho$ when $x = 0$ and $x = 1$, respectively. It then suffices to define $p(\vec{x}) := (1 - x) \cdot p_0(\vec{x}) + x \cdot p_1(\vec{x})$. As in Corollary 5.6, we may assume that p is multilinear, and the moreover part follows in the same way since p has degree at most $k + 1$. □

5.2 Lower Bounds for Generalized Andreev's Function

We now prove a super-linear lower bound against depth-three sigmoidal circuits with a hard function, known as generalized Andreev's function, that is computable in polynomial time. Formally:

Theorem 5.10. *There exists an absolute constant $c_0 \geq 1$ such that the following holds. Let $a \geq 1$, let $n \in \mathbb{N}$ be sufficiently large (depending on a), and $d, W, \delta^{-1} \leq n^a$. Suppose that C is a $(2n)$-input depth-3 sigmoidal circuit, where*

- *the output gate is a sigmoidal gate with fan-in d and weights within $[-W, W]$;*
- *the middle layer consists of d sigmoidal gates with weights and biases within $[-W, W]$;*
- *the bottom layer (i.e. closest to input variables) has s sigmoidal gates in total.*

Then if C computes $\mathsf{GA}_{n,m}$ with separation $\delta \in (0, 1)$, $s \geq n^{3/2}/(\log n)^{c_0 \cdot a}$.

5.2.1 Counting Depth-2 Augmented Sigmoidal Circuits

For the purpose of proving Andreev's function lower bound, we need to count the number of possible depth-2 sigmoidal circuits augmented with a low-degree polynomial.

Definition 5.11 ($\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$)**.** A depth-two augmented Sigmoidal circuit, denoted $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$, is of the following form:

- the bottom layer consists of s sigmoidal gates $\sigma_1, \ldots, \sigma_s$ and a real-valued multilinear polynomial $p : \mathbb{R}^n \to \mathbb{R}$, both fed by input variables;
- the top layer is a sigmoidal gate σ fed by $\sigma_1, \ldots, \sigma_s$ and p, where the wire from p has weight 1 and the constant term of p serves as the bias of σ, so it outputs $\mathsf{Sigmoid}(\sum_{i=1}^{s} v_i \sigma_i(\vec{x}) + p(\vec{x}))$ for some $v_1, \ldots, v_s \in \mathbb{R}$.

The *size* of the circuit is defined as $s + 1$, the *wire complexity* is defined as the total number of wires of $\sigma_1, \ldots, \sigma_s, \sigma$, and the *degree* is defined as the degree of p. The *weights* of the circuit consist of weights of $\sigma_1, \ldots, \sigma_s, \sigma$ (including the biases of $\sigma_1, \ldots, \sigma_s$) as well as the weights of monomials in p.

Lemma 5.12. *Let $n, d, s, w \geq 1$, $W \geq 1$, $\delta \in (0, 1/2]$. Then there is a list $\mathscr{L}$ of $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$ circuits such that for every $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$ circuit C with size at most $s + 1$, wire complexity at most w, degree at most d, and weights within $[-W, W]$, there is a circuit $C' \in \mathscr{L}$ such that C and C' are δ-close to each other. Moreover,*

$$\log_2 |\mathscr{L}| = O\big(w \log(nsW/\delta) + p(n, d) \log(p(n, d)W/\delta)\big),$$

where $p(n, d) = \sum_{i=0}^{d} \binom{n}{i}$.

Proof. Let C be such a circuit, where bottom gates are $\sigma_j(\vec{x}) = \mathsf{Sigmoid}(\langle \vec{u}_j, \vec{x}\rangle + b_j)$, $1 \le j \le s$, and it computes $\mathsf{Sigmoid}(\sum_j v_j \sigma_j(\vec{x}) + p(\vec{x}))$. Let C' be the circuit with the same topology obtained by rounding each weight and bias of the sigmoidal gates to the nearest integer multiple of $\varepsilon := \delta/(4(n+1)sW)$, and each coefficient of p to the nearest integer multiple of $\varepsilon' := \delta/(4p(n,d))$. The list $\mathscr{L}$ consists of all circuits obtained in this way. Such a circuit is determined by

- the number of bottom gates, the $\le w$ pairs (j,i) such that σ_j reads x_i, and $\le 2w$ rounded parameters of sigmoidal gates (one per wire and one bias per bottom gate; note that $s \le w$), each taking $O(nsW^2/\delta)$ values;
- At most $p(n,d)$ rounded coefficients of p, each taking $O(p(n,d)W/\delta)$ values.

This gives the bound on $|\mathscr{L}|$.

We show that C and C' are δ-close. Fix $\vec{x} \in \{0,1\}^n$. As $\mathsf{Sigmoid}$ is $(1/4)$-Lipschitz and each bottom gate has at most $n+1$ parameters, $|\sigma_j(\vec{x}) - \sigma'_j(\vec{x})| \le (n+1)\varepsilon/4 \le \delta/(16sW)$. Since $\sigma_j(\vec{x}) \in [0,1]$, $|v'_j| \le 2W$, and every multilinear monomial takes values in $\{0,1\}$, the preactivation of the output gate changes by at most

$$\sum_j \left(|v_j - v'_j| \cdot \sigma_j(\vec{x}) + |v'_j| \cdot |\sigma_j(\vec{x}) - \sigma'_j(\vec{x})|\right) + p(n,d) \cdot \varepsilon' \;\le\; \frac{\delta}{4} + \frac{\delta}{8} + \frac{\delta}{4} \;<\; \delta,$$

so the output changes by less than $\delta/4$. □

Lemma 5.12 only counts circuits with bounded weights, while the bottom layer of the circuits in our lower bounds may have unbounded weights. After a random restriction, however, these gates depend on few variables, and the next lemma bounds their parameters up to a small error.

Lemma 5.13 (Parameter compression)**.** *Let $m \ge 1$, $\varepsilon \in (0,1)$, $\vec{w} \in \mathbb{R}^m$, and $b \in \mathbb{R}$. Then there are $\vec{w}' \in \mathbb{R}^m$ and $b' \in \mathbb{R}$, whose entries have absolute values at most $(m+1)! \cdot \ln(3/\varepsilon)$, such that for every $\vec{z} \in \{0,1\}^m$, $\left|\mathsf{Sigmoid}(\langle \vec{w}, \vec{z}\rangle + b) - \mathsf{Sigmoid}(\langle \vec{w}', \vec{z}\rangle + b')\right| \le \varepsilon$.*

Lemma 5.13 is an approximate analogue of the classical fact that every threshold function on m variables can be realized with integer weights of absolute value $2^{O(m \log m)}$, and we prove it by the classical argument (see, e.g., [Hås94, Section 3]). The proof is deferred to Section B.1.

5.2.2 A Discriminator Lemma for Sigmoidal Functions

We generalize the standard discriminator lemma (see, e.g., [HMP+93]) to sigmoidal gates that compute a Boolean function with noticeable separation.

Lemma 5.14. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-decreasing and κ-Lipschitz function. Let $g_1, \dots, g_m : \{0,1\}^n \to [0,1]$, $\vec{w} \in [-W, W]^m, \theta, \delta_0, \delta_1 \in \mathbb{R}$, $f : \{0,1\}^n \to \{0,1,*\}$ be the partial function defined as*

$$f(\vec{x}) := \begin{cases} 1 & \sigma_{\vec{w},\theta}(g_1(\vec{x}), \dots, g_m(\vec{x})) \ge \delta_1 \\ 0 & \sigma_{\vec{w},\theta}(g_1(\vec{x}), \dots, g_m(\vec{x})) \le \delta_0 \\ * & \textit{otherwise} \end{cases}$$

Assume that $\delta_1 > \delta_0$ and $f^{-1}(0), f^{-1}(1) \ne \varnothing$. Then for every integer $B \ge 1$, there exist $t \in \{1, \dots, B\}$ and $i \in [m]$ such that for $g'_{i,t}(\vec{x}) := \mathsf{sgn}(g_i(\vec{x}) - t/B) \in \{0,1\}$,

$$\left| \Pr_{\vec{x} \leftarrow \{0,1\}^n}[g'_{i,t}(\vec{x}) = 1 \mid f(\vec{x}) = 1] - \Pr_{\vec{x} \leftarrow \{0,1\}^n}[g'_{i,t}(\vec{x}) = 1 \mid f(\vec{x}) = 0] \right| \ge \frac{\delta_1 - \delta_0}{\kappa m W} - \frac{2}{B}.$$

In particular, for $B = \lceil 4\kappa mW/(\delta_1 - \delta_0) \rceil$, the lower bound becomes at least $(\delta_1 - \delta_0)/(2\kappa mW)$.

Proof. Throughout the proof, all expectations and probabilities are taken over a uniform $\vec{x} \leftarrow \{0,1\}^n$; we omit the subscript for brevity. By definition, $g'_{i,t}(\vec{x}) = 1$ iff $g_i(\vec{x}) > t/B$.

Recall that $g_i(\vec{x}) \in [0,1]$ for every $i \in [m]$ and $\vec{x} \in \{0,1\}^n$. Subsequently, for each $i \in [m]$, we have

$$\left| g_i(\vec{x}) - \frac{1}{B} \sum_{t=1}^{B} g'_{i,t}(\vec{x}) \right| \le \frac{1}{B}.$$

Let $\tilde{g}_i(\vec{x}) := \frac{1}{B} \sum_{t=1}^{B} g'_{i,t}(\vec{x})$, $L(\vec{x}) := \sum_{i=1}^{m} w_i g_i(\vec{x}) + \theta$, and $\tilde{L}(\vec{x}) := \sum_{i=1}^{m} w_i \tilde{g}_i(\vec{x}) + \theta$. Since $|w_i| \le W$,

$$|L(\vec{x}) - \tilde{L}(\vec{x})| \le \sum_{i=1}^{m} |w_i| \cdot |g_i(\vec{x}) - \tilde{g}_i(\vec{x})| \le \frac{mW}{B}. \tag{5}$$

Let $a_1 := \min\{L(\vec{x}) \in \mathbb{R} \mid \vec{x} \in \{0,1\}^n,\ \sigma(L(\vec{x})) \ge \delta_1\}$ and $a_0 := \max\{L(\vec{x}) \in \mathbb{R} \mid \vec{x} \in \{0,1\}^n,\ \sigma(L(\vec{x})) \le \delta_0\}$; the existence of a_0, a_1 follows from $f^{-1}(0), f^{-1}(1) \neq \varnothing$. Note that by the definition of f, L, and $\tilde{L}$,

$$\begin{aligned} f(\vec{x}) = 1 \ &\Rightarrow\ L(\vec{x}) \ge a_1 \ \Rightarrow\ \tilde{L}(\vec{x}) \ge a_1 - \frac{mW}{B}; \\ f(\vec{x}) = 0 \ &\Rightarrow\ L(\vec{x}) \le a_0 \ \Rightarrow\ \tilde{L}(\vec{x}) \le a_0 + \frac{mW}{B}. \end{aligned}$$

Moreover, as σ is non-decreasing and κ-Lipschitz, we have $a_1 - a_0 \ge (\delta_1 - \delta_0)/\kappa$. Subsequently, we have

$$\mathbb{E}[\tilde{L}(\vec{x}) \mid f(\vec{x}) = 1] - \mathbb{E}[\tilde{L}(\vec{x}) \mid f(\vec{x}) = 0] \ge \frac{\delta_1 - \delta_0}{\kappa} - \frac{2mW}{B}.$$

Finally, by linearity of expectation and the fact that $g'_{i,t}(\vec{x}) \in \{0,1\}$,

$$\mathbb{E}[\tilde{L}(\vec{x}) \mid f(\vec{x}) = 1] - \mathbb{E}[\tilde{L}(\vec{x}) \mid f(\vec{x}) = 0] = \frac{1}{B} \sum_{i=1}^{m} \sum_{t=1}^{B} w_i \cdot \Delta_{i,t},$$

where $\Delta_{i,t} := \Pr[g'_{i,t}(\vec{x}) = 1 \mid f(\vec{x}) = 1] - \Pr[g'_{i,t}(\vec{x}) = 1 \mid f(\vec{x}) = 0]$. Since $|w_i| \le W$, the absolute value of the RHS is at most $mW \cdot \max_{i \in [m], t \in [B]} |\Delta_{i,t}|$, hence

$$\max_{i \in [m],\, t \in [B]} |\Delta_{i,t}| \ge \frac{1}{mW} \left(\frac{\delta_1 - \delta_0}{\kappa} - \frac{2mW}{B} \right) = \frac{\delta_1 - \delta_0}{\kappa mW} - \frac{2}{B},$$

as required. The "in particular" statement follows from $B \ge 4\kappa mW/(\delta_1 - \delta_0)$, which gives $2/B \le (\delta_1 - \delta_0)/(2\kappa mW)$. □

5.2.3 Proof of the Size Lower Bound

We need the hard function $F : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}$ constructed in [KW16], with a simple modification so it runs in $\mathsf{poly}(n, m)$ time (rather than $\mathsf{poly}(n, 2^m)$ as stated in [KW16]). We first recall a lemma from [KW16].

Lemma 5.15 ([KW16, Lemma 6.1]). *There is an absolute constant $C \geq 1$ such that the following holds. Let $\varepsilon > 0$, and let $\mathcal{S}$ be a set of Boolean functions on m variables such that $\mathsf{Cor}(g, g') \leq \varepsilon$ for any two distinct $g, g' \in \mathcal{S}$. Then for every Boolean function f on m variables, at most C/ε functions $g \in \mathcal{S}$ satisfy $\mathsf{Cor}(f, g) > C\sqrt{\varepsilon}$.*

Lemma 5.16 (implicit in [KW16, Corollary 6.2]). *Let n be sufficiently large and m be an even integer such that $n^2 \leq 2^m \leq 2^n$. There is a function $F : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}$ computable in $\mathsf{poly}(n)$ time such that given any class of Boolean functions $\mathscr{L}$ on m variables with $|\mathscr{L}| \leq 2^{n/2}$, there is an $\vec{x} \in \{0,1\}^n$ such that for any $f \in \mathscr{L}$, we have $\mathsf{Cor}(f, F(\vec{x}, \cdot)) \leq O(\sqrt{n})/2^{m/4}$.*

Proof Sketch. Let $\ell := m/2$, and identify $\{0,1\}^\ell$ with the field $\mathbb{F}_{2^\ell}$ using an irreducible polynomial of degree ℓ over $\mathbb{F}_2$, which can be found deterministically in $\mathsf{poly}(\ell)$ time [Sho90]. For $\vec{x} \in \{0,1\}^n$ and $\vec{z} = (u, v) \in \mathbb{F}_{2^\ell}^2$, let

$$F(\vec{x}, \vec{z}) := \left\langle \sum_{i=1}^{n} x_i u^{i-1},\ v \right\rangle \bmod 2,$$

where $\langle \cdot, \cdot \rangle$ is the inner product of the binary representations; it can be computed with $O(n)$ field operations. This is indeed the powering construction of small-biased sets from [AGHP92].

For $\vec{x} \neq \vec{x}'$, the polynomial $q(u) := \sum_{i=1}^{n} (x_i \oplus x'_i) u^{i-1}$ is nonzero and has fewer than n roots, and $F(\vec{x}, (u, v)) \oplus F(\vec{x}', (u, v)) = \langle q(u), v \rangle \bmod 2$ is a uniformly random bit over v when $q(u) \neq 0$. Hence $F(\vec{x}, \cdot)$ and $F(\vec{x}', \cdot)$ agree on a $1/2 + O(n/2^{m/2})$ fraction of inputs. By Lemma 5.15 with $\varepsilon := n/2^{m/2}$, each $f \in \mathscr{L}$ has correlation more than $O(\sqrt{n})/2^{m/4}$ with $F(\vec{x}, \cdot)$ for at most $O(2^{m/2}/n)$ strings $\vec{x}$. Since $|\mathscr{L}| \cdot O(2^{m/2}/n) \leq O(2^n/n) < 2^n$, some $\vec{x} \in \{0,1\}^n$ satisfies the lemma. □

Fix $m := 16\lceil \log_2 n \cdot \log_2 \log_2 n \rceil$. For simplicity of presentation, we assume n/m is an integer; the general case follows by a simple padding argument. Let $F(\vec{x}, \vec{z})$ be the polynomial time computable function above. We will use the following function $\mathsf{GA}_{n,m}(\vec{x}, \vec{y})$ with $n + (n/m) \cdot m = 2n$ input variables, where $\vec{x} \in \{0,1\}^n$ and $\vec{y} \in \{0,1\}^n$. The string $\vec{y}$ is parsed as $\vec{y}_1, \ldots, \vec{y}_m \in \{0,1\}^{n/m}$, and we define

$$\mathsf{GA}_{n,m}(\vec{x}, \vec{y}) := F(\vec{x}, \vec{z}),$$

where $z_i := \oplus(\vec{y}_i)$ is the parity of bits in $\vec{y}_i$. By Lemma 5.16, $\mathsf{GA}_{n,m}$ is computable in polynomial time. The function is introduced by Kane and Williams [KW16] and is a generalization of Andreev's function [And87]; the variants of Andreev's function have found many applications in proving circuit lower bounds, see, e.g., [IN93, PZ93, Hås98, IMZ12, KR13, KW16, KKL17].

Theorem 5.10. *There exists an absolute constant $c_0 \geq 1$ such that the following holds. Let $a \geq 1$, let $n \in \mathbb{N}$ be sufficiently large (depending on a), and $d, W, \delta^{-1} \leq n^a$. Suppose that C is a $(2n)$-input depth-3 sigmoidal circuit, where*

- *the output gate is a sigmoidal gate with fan-in d and weights within $[-W, W]$;*
- *the middle layer consists of d sigmoidal gates with weights and biases within $[-W, W]$;*
- *the bottom layer (i.e. closest to input variables) has s sigmoidal gates in total.*

Then if C computes $\mathsf{GA}_{n,m}$ with separation $\delta \in (0,1)$, $s \geq n^{3/2}/(\log n)^{c_0 \cdot a}$.

Proof. Suppose that C computes $\mathsf{GA}_{n,m}$ with separation δ. Let $\vec{x} \in \{0,1\}^n$ be a string and $t, k \in \mathbb{N}$ be parameters to be determined later (⋄), $k \leq n/2$, and $C_{\vec{x}}(\vec{y}) := C(\vec{x}, \vec{y})$ be the circuit after fixing the first n input bits to be $\vec{x}$. As mentioned above, we assume for simplicity of presentation that n/m is an integer.

Behavior under random restriction. Parse $\vec{y} = (\vec{y}_1, \ldots, \vec{y}_m)$, where $\vec{y}_i \in \{0,1\}^{n/m}$ for every $i \in [m]$, and let $\mathscr{P}$ be the partition of $[n]$ into m blocks so the i-th block corresponds to $\vec{y}_i$. Then by Corollary 5.6 and Markov inequality, there is a restriction ρ over $\mathcal{R}_{\mathscr{P}}$ such that all but

$$s_1 := O\left(\frac{s|\mathscr{P}| \cdot t(k+1)\ln t}{\sqrt{n}}\right) = O\left(\frac{smt(k+1)\ln t}{\sqrt{n}}\right) \tag{6}$$

sigmoidal gates in the first layer are $t^{-(k+1)}$-close to a degree-k real coefficient polynomial. Fix any restriction that satisfies this condition, and let $n_1 := m$ be the number of remaining input variables of $C_{\vec{x}}|_\rho$. Let $C_1 := C_{\vec{x}}|_\rho$. Note that C_1 computes the function $\vec{z} \mapsto F(\vec{x}, \vec{z} \oplus \vec{\eta})$ with separation δ, where $\eta_i \in \{0,1\}$ is the parity of the bits of $\vec{y}_i$ fixed by ρ.

Proposition 5.17. *Let*

$$W_1 := \max\{(2^{k+1}s + 2n + 1)W,\ (n_1+1)! \cdot \ln(3t^{k+1})\}. \tag{7}$$

For each subcircuit G in the middle layer of C_1, there exists a $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid}\|\mathsf{Poly})$ *circuit G' of size at most $s_1 + 1$, degree k, and weights within $[-W_1, W_1]$ such that $|G(\vec{z}) - G'(\vec{z})| \le sW \cdot t^{-(k+1)}$ on any $\vec{z} \in \{0,1\}^{n_1}$.*

The proof is straightforward, and is deferred to the end of the proof.

Applying the discriminator lemma. Let $C_1'(\vec{z})$ be the circuit obtained from C_1 by replacing each subcircuit G in the middle layer with G' in Proposition 5.17. Then we know that $|C_1'(\vec{z}) - C_1(\vec{z})| \le (dW) \cdot sW \cdot t^{-(k+1)} \le sdW^2/t^{k+1}$. Assume that $sdW^2/t^{k+1} \le \delta/10$ (which we will verify later ($\star$)), $C_1'(\vec{z})$ computes $F(\vec{x}, \vec{z} \oplus \vec{\eta})$ with separation at least $\delta - 2sdW^2/t^{k+1} \ge 4\delta/5$ (an additive error of e can increase the outputs on 0-inputs and decrease the outputs on 1-inputs by e each, so it decreases the separation by at most $2e$).

Note that $C_1'(\vec{z})$ is a sigmoidal gate taking at most d inputs (each of which is a $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid}\|\mathsf{Poly})$ circuit of size $s_1 + 1$, degree k, and weights within $[-W_1, W_1]$). We now define another circuit $C_1''(\vec{z})$ obtained from $C_1'(\vec{z})$.

Let $\mathscr{L}$ be the list of circuits in Lemma 5.12 for approximating $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid}\|\mathsf{Poly})$ circuit of size $s_1 + 1$, degree k, and weights within $[-W_1, W_1]$ up to precision $\delta/(10dW)$. We replace each $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid}\|\mathsf{Poly})$ subcircuit that feeds the output sigmoidal gate of $C_1'(\vec{z})$ with its best approximation in $\mathscr{L}$, so that

$$|C_1''(\vec{z}) - C_1'(\vec{z})| \ \le\ (dW) \cdot \frac{\delta}{10dW} \ \le\ \frac{\delta}{10}.$$

Therefore, $C_1''(\vec{z})$ still computes $F(\vec{x}, \vec{z} \oplus \vec{\eta})$ with separation at least $4\delta/5 - 2 \cdot \delta/10 = 3\delta/5$. Moreover, we know that each subcircuit in $\mathscr{L}$ can be completely specified with

$$\log_2 |\mathscr{L}| = O\left(n_1 s_1 \log(dn_1 s_1 \delta^{-1} W_1) + k^2 \left(\frac{n_1 e}{k}\right)^k \log(kdn_1 \delta^{-1} W_1)\right) \tag{8}$$

bits, where we use the approximation $\binom{n}{k} \le (ne/k)^k$.

Then by Lemma 5.14 with $\sigma = \mathsf{Sigmoid}$ (which is 1/4-Lipschitz) and $B = \lceil 4\kappa dW/(3\delta/5) \rceil = O(dW/\delta)$, which applies since $F(\vec{x}, \cdot)$ is not constant for the string $\vec{x}$ we fix below, there exist

a $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid}\|\mathsf{Poly})$ circuit $G' \in \mathscr{L}$ and an index $\tau \in [B]$ such that, defining $G'_\tau(\vec{z}) := \mathsf{sgn}(G'(\vec{z}) - \tau/B)$,

$$\left| \Pr_{\vec{z} \leftarrow \{0,1\}^{n_1}} [G'_\tau(\vec{z}) = 1 \mid F(\vec{x}, \vec{z} \oplus \vec{\eta}) = 1] - \Pr_{\vec{z} \leftarrow \{0,1\}^{n_1}} [G'_\tau(\vec{z}) = 1 \mid F(\vec{x}, \vec{z} \oplus \vec{\eta}) = 0] \right| \geq \frac{3\delta/5}{2\kappa dW} = \frac{6\delta}{5dW} \geq \frac{\delta}{dW}. \tag{9}$$

Note that $G'_\tau(\vec{z})$ can be completely specified with $O(\log_2 |\mathscr{L}| + \log_2 B)$ bits.

Counting argument. We now set parameters ($\diamond$) and prove that $\log_2(|\mathscr{L}| \cdot B + 1) = o(n)$, where:

$$t := \lceil (\log_2 n)^{6a} \rceil \qquad \text{and} \qquad k := \lceil \log_2 n / \log_2 \log_2 n \rceil.$$

By assumption we have $s \leq n^{3/2}$. Combined with $d, W, 1/\delta \leq n^a$ we have $sdW^2/\delta \leq n^{4a+3/2}$. Note that

$$t^{k+1} \geq t^k \geq 2^{k \cdot 6a \log_2 \log_2 n} \geq 2^{6a \log_2 n} = n^{6a},$$

which is at least $10 \cdot sdW^2/\delta$ for every $a \geq 1$ and n sufficiently large, as required by ($\star$).

We now instantiate Equation (8) with the chosen parameters. Recall that $n_1 = m = \Theta(\log n \cdot \log\log n)$, and let $\delta_{\mathscr{L}} := \delta/(10dW)$, so $\log_2(1/\delta_{\mathscr{L}}) = O(a \log n)$. Subsequently,

- By Equation (6), $s_1 = O\big(smt(k+1) \ln t/\sqrt{n}\big) = O\big(a\, s\, (\log n)^{6a+2} \log\log n/\sqrt{n}\big)$.
- By Equation (7), $\log_2 W_1 = O(a \log n + n_1 \log n_1) = O\big(a(\log n)^2/\log\log n\big)$.
- The binomial term $(n_1 e/k)^k$ in Equation (8) satisfies

$$\big(n_1 e/k\big)^k \leq \big(17e(\log_2 \log_2 n)^2\big)^{\log_2 n/\log_2 \log_2 n + 1} = n^{o(1)}.$$

The bounds above give $\log_2(dn_1 s_1 W_1/\delta_{\mathscr{L}}) = O(a(\log n)^2/\log\log n)$. Plugging them into Equation (8),

$$\log_2 |\mathscr{L}| = O\left(\frac{a^2\, s\, (\log n)^{6a+5} \log\log n}{\sqrt{n}}\right) + n^{o(1)}, \qquad \log_2 B = O(\log(dW/\delta)) = O(a \log n).$$

Set $c_0 := 12$, and suppose toward contradiction that $s \leq n^{3/2}/(\log n)^{c_0 \cdot a}$. Since $c_0 a - (6a+5) = 6a - 5 \geq 1$ for every $a \geq 1$,

$$\frac{a^2\, s\, (\log n)^{6a+5} \log\log n}{\sqrt{n}} \leq \frac{a^2\, n \log\log n}{(\log n)^{c_0 a - (6a+5)}} = \frac{a^2\, n \log\log n}{(\log n)^{6a-5}} \leq \frac{a^2\, n \log\log n}{\log n} = o(n),$$

and so $\log_2(|\mathscr{L}| \cdot B + 1) = o(n)$, as desired.

Applying Lemma 5.16. Define the augmented list $\mathscr{L}' := \big\{ \vec{z} \mapsto G'_\tau(\vec{z} \oplus \vec{\eta}') : G' \in \mathscr{L}, \tau \in [B], \vec{\eta}' \in \{0,1\}^m \big\} \cup \{\mathbf{0}\}$, where $\mathbf{0} : \{0,1\}^{n_1} \to \{0,1\}$ is the constant zero function. Then $|\mathscr{L}'| \leq 2^m \cdot |\mathscr{L}| \cdot B + 1 \leq 2^{o(n)} \leq 2^{n/2}$, as $m = o(n)$. Note also that $n^2 \leq 2^m \leq 2^n$ since $16 \log_2 n \leq m \leq n$. Subsequently, by Lemma 5.16 applied to $\mathscr{L}'$, there exists $\vec{x}^* \in \{0,1\}^n$ such that for every $f \in \mathscr{L}'$,

$$\mathsf{Cor}\big(f,\, F(\vec{x}^*, \cdot)\big) \leq \frac{O(\sqrt{n})}{2^{m/4}} \leq O(\sqrt{n}) \cdot n^{-4 \log_2 \log_2 n}. \tag{10}$$

Let ε be the largest correlation between $f \in \mathscr{L}'$ and $F(\vec{x}^*, \cdot)$ and set $\vec{x} := \vec{x}^*$.

Recall that the discriminator lemma yields (for some $G' \in \mathscr{L}$ and $\tau \in [B]$) the inequality (9). Replacing $\vec{z}$ by $\vec{z} \oplus \vec{\eta}$ there, the function $g(\vec{z}) := G'_\tau(\vec{z} \oplus \vec{\eta})$ in $\mathscr{L}'$ satisfies

$$\big|p_1 - p_0\big| \;\geq\; \frac{\delta}{dW}, \qquad \text{where } p_b := \Pr_{\vec{z} \leftarrow \{0,1\}^{n_1}}[g(\vec{z}) = 1 \mid F(\vec{x}^*, \vec{z}) = b]. \tag{11}$$

Since $\mathbf{0} \in \mathscr{L}'$, (10) also holds for $\mathbf{0}$, which implies that $F(\vec{x}^*, \cdot)$ is ε-balanced; as $\varepsilon < 1/2$, $F(\vec{x}^*, \cdot)$ is not constant, so the discriminator lemma indeed applies. Combining (11) with Lemma 3.7, we have

$$\mathsf{Cor}\big(g, F(\vec{x}^*, \cdot)\big) \;\geq\; \frac{\delta}{2dW} - \varepsilon.$$

On the other hand, $g \in \mathscr{L}'$ gives $\mathsf{Cor}(g, F(\vec{x}^*, \cdot)) \leq \varepsilon$. Subsequently, $\delta/(dW) \leq 4\varepsilon = O(\sqrt{n}) \cdot n^{-4\log_2\log_2 n}$. However, $\delta/(dW) \geq n^{-a} \cdot n^{-2a} = n^{-3a}$, which is much larger than $O(\sqrt{n}) \cdot n^{-4\log_2\log_2 n}$ for n sufficiently large. This leads to a contradiction. □

Proof of Proposition 5.17. Let G be a subcircuit in the middle layer of C_1, and $F_1, \ldots, F_s$ be the gates in the first layer, so that

$$G(\vec{z}) = \mathsf{Sigmoid}\left(\sum_{i=1}^{s} w_i \cdot F_i(\vec{z}) + \ell(\vec{z})\right)$$

for some $w_1, \ldots, w_s \in [-W, W]$ and an affine function ℓ that collects the bias of G and its wires from the (restricted) input variables. As G has at most $2n$ input wires, the absolute values of the coefficients of ℓ sum to at most $(2n+1)W$.

Let $\mathscr{A}$ be the set of indices of the (at most s_1) first-layer gates that are not approximated by polynomials. For $i \notin \mathscr{A}$, let p_i be the multilinear polynomial given by Corollary 5.6, whose coefficients have absolute values at most 2^{k+1}. For $i \in \mathscr{A}$, let F'_i be the sigmoid gate given by Lemma 5.13 for F_i (as a function of the n_1 free variables) and $\varepsilon := t^{-(k+1)}$, whose weights and bias have absolute values bounded by $(n_1+1)! \cdot \ln(3t^{k+1}) \leq W_1$. Let

$$G'(\vec{z}) := \mathsf{Sigmoid}\left(\sum_{i \in \mathscr{A}} w_i \cdot F'_i(\vec{z}) + P(\vec{z})\right), \qquad \text{where } P := \sum_{i \notin \mathscr{A}} w_i p_i + \ell.$$

Then G' is a $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$ circuit of size at most $s_1 + 1$ and degree k, and each coefficient of P has absolute value at most $2^{k+1}sW + (2n+1)W \leq W_1$. Moreover, every first-layer gate F_i is replaced by a function that differs at most $t^{-(k+1)}$ from F_i on $\{0,1\}^{n_1}$; as $\mathsf{Sigmoid}$ is $(1/4)$-Lipschitz, $|G(\vec{z}) - G'(\vec{z})| \leq \frac{1}{4}\sum_i |w_i| \cdot t^{-(k+1)} \leq sW \cdot t^{-(k+1)}$. □

5.2.4 Proof of the Wire Lower Bound

Now we prove a similar lower bound on the total number of wires in the first layer. It follows closely the proof of Theorem 5.10; we describe only the changes (marked in blue).

Theorem 5.18. *There exists an absolute constant $c_0 \geq 1$ such that the following holds. Let $a \geq 1$, let $n \in \mathbb{N}$ be sufficiently large (depending on a), and $d, W, 1/\delta \leq n^a$. Suppose that C is a $(2n)$-input depth-3 sigmoidal circuit, where*

- *the output gate is a sigmoidal gate with fan-in d and weights within $[-W, W]$;*
- *the middle layer consists of d sigmoidal gates with weights and biases within $[-W, W]$;*
- *the bottom layer consists of sigmoidal gates, each with fan-in at least one, whose total fan-in is w.*

If C computes $\mathsf{GA}_{n,m}$ with separation $\delta \in (0,1)$, then $w \geq n^{5/2}/(\log n)^{c_0 \cdot a}$.

Proof Sketch. The crucial step is to replace Corollary 5.6 with the wire-complexity variant Corollary 5.9: the failure probability of the j-th first-layer gate with fan-in w_j under random restriction $\rho \leftarrow \mathcal{R}_{\mathscr{p}}$ becomes $O(m^{3/2}\, t\, w_j (k+1) \ln t / n^{3/2})$. Since $w = \sum_j w_j$, by linearity of expectation, the expected number of bad gates is at most $O(m^{3/2}\, t\, w (k+1) \ln t / n^{3/2})$, and Markov's inequality then yields a restriction $\rho \in \mathcal{R}_{\mathscr{p}}$ such that all but

$$s_1 \;:=\; O\left(\frac{m^{3/2}\, t\, w\, (k+1) \ln t}{n^{3/2}}\right) \tag{12}$$

first-layer gates are $t^{-(k+1)}$-close to a degree-$(k+1)$ polynomial (rather than degree k, by Corollary 5.9). Analogously to Proposition 5.17, for each middle-layer subcircuit G, there is a $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$ circuit G' of size $s_1 + 1$, degree $k+1$, and weights within $[-W_1, W_1]$ for $W_1 := \max\{(2^{k+2} s + 2n + 1) W,\, (n_1+1)! \cdot \ln(3t^{k+1})\}$, with the same per-gate error $|G - G'| \leq sW \cdot t^{-(k+1)}$. We can then apply the discriminator lemma to obtain a $\mathsf{Sigmoid} \circ (\mathsf{Sigmoid} \| \mathsf{Poly})$ discriminator for $F(\vec{x}, \cdot \oplus \vec{\eta})$, provided that $t^{k+1} \geq 10 \cdot sdW^2/\delta$ $(\star)$.

Assume for contradiction that we have $w \leq n^{5/2}/(\log n)^{c_0 \cdot a}$; in particular, let s be the number of bottom-layer gates, then $s \leq w \leq n^{5/2}$. Combined with $d, W, 1/\delta \leq n^a$, this gives $sdW^2/\delta \leq n^{4a+5/2}$. We can set

$$t := \lceil (\log_2 n)^{7a} \rceil \qquad \text{and} \qquad k := \lceil \log_2 n / \log_2 \log_2 n \rceil$$

so that $t^{k+1} \geq 10 \cdot sdW^2/\delta$ for every $a \geq 1$ and n sufficiently large (the condition $(\star)$).

Following the counting argument of Theorem 5.10, the dominant term $n_1 s_1 \log_2 W_1$ becomes $O(w\, a^2 (\log n)^{7a+11/2} (\log\log n)^{3/2} / n^{3/2})$. By setting $c_0 := 14$, it follows that $\log_2(|\mathscr{L}| \cdot B + 1) = o(n)$. We can then apply Lemma 5.16 with the list $\mathscr{L}'$ defined as in the proof of Theorem 5.10 to conclude a contradiction. □

5.3 Correlation Bounds with Parity via Average Sensitivity

In this section, we prove a correlation bound for polynomial-weight small sigmoidal circuits and the parity function. Our main result is the following lower bound that matches the correlation bound for linear threshold circuits (see, e.g., [CSS18]).

Theorem 5.19. *For any $a, d \geq 1$ and sufficiently large n, let $C : \{0,1\}^n \to [0,1]$ be any depth-d sigmoidal circuit with s gates and weights within $[-n^a, n^a]$. Then for any $0 \leq \delta_0 < \delta_1 \leq 1$ with $\delta_1 - \delta_0 \geq n^{-a}$, we have*

$$\Pr_{\vec{x} \leftarrow \{0,1\}^n}[C(\vec{x}) > \delta_1 \wedge \oplus_n(\vec{x}) = 1] + \Pr_{\vec{x} \leftarrow \{0,1\}^n}[C(\vec{x}) < \delta_0 \wedge \oplus_n(\vec{x}) = 0] \leq \frac{1}{2} + s^{1-1/d} \cdot n^{-1/(2d)+o(1)}.$$

In particular, for $d \geq 2$ and $s \leq n^{1/(2(d-1))-\Omega(1)}$, if C computes a Boolean function f with separation at least n^{-a}, then $\mathsf{Cor}(f, \oplus_n) \leq n^{-\Omega(1)}$.

Indeed, we also prove an average sensitivity upper bound for any Boolean function computed by polynomial-weight small sigmoidal circuits with noticeable separation. Formally:

Theorem 5.20. *For any $a, d \geq 1$ and sufficiently large n, let $C : \{0,1\}^n \to [0,1]$ be any depth-d sigmoidal circuit with s gates and weights within $[-n^a, n^a]$. Suppose that it computes a Boolean function $f : \{0,1\}^n \to \{0,1\}$ with separation $\delta \geq n^{-a}$, then $\mathsf{AS}(f) \leq s^{1-1/d} \cdot n^{1-1/(2d)+o(1)}$.*

5.3.1 Random Restriction on Polynomial Sigmoid Functions

To prove the correlation bound for parity, we need to analyze the behavior of *polynomial sigmoid functions* under random restriction.

Definition 5.21 (Polynomial Sigmoid Functions)**.** A *polynomial sigmoid function* is a function of the form

$$\mathsf{PSigmoid}_p(\vec{x}) = \mathsf{Sigmoid}(p(\vec{x})),$$

where p is a real-weighted multi-linear polynomial. The *degree* of the function is defined as the degree of the polynomial p. The function is said to have *weight* at most W if each coefficient of monomials in p has absolute value at most W.

Recall the definition of restrictions $\rho = (I, \vec{y})$ in Section 5.1. We use $\mathcal{R}_{n,m}$ to denote the distribution of restrictions in which I is a uniformly random subset of $[n]$ of size m and $\vec{y}$ is uniformly random. When the input length n is clear from the context, we may drop it and simply use $\mathcal{R}_m$ to denote the distribution.

Lemma 5.22. *For any degree-d multilinear polynomial $p(\vec{x})$ with n variables, $t \geq 2$, $k \geq 1$, $m \leq n/16$, $\delta \in (0, 1/16)$ with $\delta \leq \sqrt{m/n}$, over a random restriction $\rho \leftarrow \mathcal{R}_m$, with probability at least*

$$1 - O\left(\sqrt{m/n} \cdot kt \cdot \left(\log(n/m) \cdot \log(tk/\delta)\right)^{O(d^2)}\right)$$

there is a degree-(dk) polynomial $q : \mathbb{R}^m \to \mathbb{R}$ such that $\mathsf{PSigmoid}_p|_\rho, q : \{0,1\}^m \to \mathbb{R}$ are δ-almost $t^{-(k+1)}$-close to each other.

The proof of the lemma is close to its linear version (see Corollary 5.6), with a crucial difference that we need to apply the random restriction lemma for polynomials in [KKL17]. We defer the proof to the appendix (see Section B.2).

5.3.2 Average Gap Sensitivity

We will prove the correlation bound by considering the *average sensitivity* of functions computed by Sigmoidal circuits. We start with the definition of average gap sensitivity for real-valued Boolean functions $f : \{0,1\}^n \to [0,1]$.

Definition 5.23 (average gap sensitivity)**.** Let $f : \{0,1\}^n \to [0,1]$ be a function, its (δ_0, δ_1)-*gap sensitivity* at $x \in \{0,1\}^n$, denoted $\mathsf{S}_{\delta_0,\delta_1}(f, x)$, is defined as

$$\mathsf{S}_{\delta_0,\delta_1}(f, x) := \sum_{i \in [n]} \mathbb{I}[\max\{f(\vec{x}), f(\vec{x}^{(i)})\} > \delta_1 \wedge \min\{f(\vec{x}), f(\vec{x}^{(i)})\} < \delta_0],$$

where $\vec{x}^{(i)}$ denotes the input with the i-th bit flipped. The *average* (δ_0, δ_1)-*gap sensitivity* of f, denoted $\mathsf{AS}_{\delta_0,\delta_1}(f)$, is

$$\mathsf{AS}_{\delta_0,\delta_1}(f) := \mathop{\mathbb{E}}_{x \leftarrow \{0,1\}^n}[\mathsf{S}_{\delta_0,\delta_1}(f, x)].$$

The following lemmas show that a function $f : \{0,1\}^n \to [0,1]$ with small gap sensitivity, when we view it as a binary classifier with gap δ, has small correlation with the parity function over n bits.

Lemma 5.24. *Let $\delta_0, \delta_1 \in \mathbb{R}$, $\delta_0 < \delta_1$, $s \in \mathbb{R}$, and $f : \{0,1\}^n \to [0,1]$ be any function with $\mathsf{AS}_{\delta_0,\delta_1}(f) \le s$. Then*

$$\Pr_{x \leftarrow \{0,1\}^n}[f(\vec{x}) > \delta_1 \wedge \oplus_n(\vec{x}) = 1] + \Pr_{x \leftarrow \{0,1\}^n}[f(\vec{x}) < \delta_0 \wedge \oplus_n(\vec{x}) = 0] \le \frac{1}{2} + \frac{s}{2n}$$

Proof. By the definition of average (δ_0, δ_1)-gap sensitivity, we know that

$$\frac{\mathsf{AS}_{\delta_0,\delta_1}(f)}{n} := \Pr_{\vec{x},i}[\max\{f(\vec{x}), f(\vec{x}^{(i)})\} > \delta_1 \wedge \min\{f(\vec{x}), f(\vec{x}^{(i)})\} < \delta_0], \tag{13}$$

where $x \leftarrow \{0,1\}^n$ and $i \leftarrow [n]$. Let $\mathcal{G}$ be the event in $\Pr[\cdot]$ above, $\mathcal{E}$ be the event that either $f(\vec{x}) > \delta_1 \wedge \oplus_n(\vec{x}) = 1$ or $f(\vec{x}) < \delta_0 \wedge \oplus_n(\vec{x}) = 0$, and $\mathcal{E}'$ be the event that either $f(\vec{x}^{(i)}) > \delta_1 \wedge \oplus_n(\vec{x}^{(i)}) = 1$ or $f(\vec{x}^{(i)}) < \delta_0 \wedge \oplus_n(\vec{x}^{(i)}) = 0$. By the definition of the events, $\mathcal{E} \wedge \mathcal{E}'$ implies $\mathcal{G}$, since $\oplus_n(\vec{x}) \neq \oplus_n(\vec{x}^{(i)})$. Moreover, as the marginal distribution of both $\vec{x}$ and $\vec{x}^{(i)}$ are uniformly random, we know that $\Pr[\mathcal{E}] = \Pr[\mathcal{E}']$, and subsequently

$$\frac{s}{n} \ge \frac{\mathsf{AS}_{\delta_0,\delta_1}(f)}{n} = \Pr_{\vec{x},i}[\mathcal{G}] \ge \Pr_{\vec{x},i}[\mathcal{E} \wedge \mathcal{E}'] \ge 1 - 2\Pr_{\vec{x},i}[\neg\mathcal{E}].$$

This implies that $\Pr_{\vec{x},i}[\mathcal{E}] = 1 - \Pr_{\vec{x},i}[\neg\mathcal{E}] \le 1/2 + s/(2n)$. □

Next, we prove two useful lemmas on average sensitivity. The first lemma is an adaptation of [Bop97, Lemma 3].

Lemma 5.25. *Let $f : \{0,1\}^n \to [0,1]$ be a function and $1 \le m \le n$. Then*

$$\frac{\mathsf{AS}_{\delta_0,\delta_1}(f)}{n} = \frac{\mathbb{E}_{\rho \leftarrow \mathcal{R}_m}[\mathsf{AS}_{\delta_0,\delta_1}(f|_\rho)]}{m}.$$

Proof. For $\vec{x} \in \{0,1\}^n$ and $i \in [n]$, we say $(\vec{x}, i)$ is good if $\max\{f(\vec{x}), f(\vec{x}^{(i)})\} > \delta_1$ and $\min\{f(\vec{x}), f(\vec{x}^{(i)})\} < \delta_0$. For a random restriction $\rho \leftarrow \mathcal{R}_m$, we can consider it as a random subset $I \subseteq [n]$ of size m (the free variables) and a random assignment $\vec{y} \in \{0,1\}^{n-m}$ to variables outside of I. It can be verified that for each $I \subseteq [n]$,

$$\begin{aligned}\mathop{\mathbb{E}}_{\vec{y} \leftarrow \{0,1\}^{n-m}}[\mathsf{AS}_{\delta_0,\delta_1}(f|_\rho); \rho := (I, \vec{y})] &= \sum_{i \in I} \mathop{\mathbb{E}}_{\vec{x} \leftarrow \{0,1\}^n}[\mathbb{I}[(\vec{x}, i) \text{ is a good pair}]] \\ &= \frac{1}{2^n} \sum_{i \in I} \sum_{\vec{x} \in \{0,1\}^n} \mathbb{I}[(\vec{x}, i) \text{ is a good pair}].\end{aligned}$$

Subsequently,

$$\begin{aligned}\mathop{\mathbb{E}}_{\rho\leftarrow\mathcal{R}_m}[\mathsf{AS}_{\delta_0,\delta_1}(f|_\rho)] &= \mathop{\mathbb{E}}_{I\subseteq[n],|I|=m}\left[\frac{1}{2^n}\sum_{i\in I}\sum_{\vec{x}\in\{0,1\}^n}\mathbb{I}[(\vec{x},i)\text{ is a good pair}]\right]\\ &= \mathop{\mathbb{E}}_{\vec{x}\leftarrow\{0,1\}^n}\left[\sum_{i\in[n],(\vec{x},i)\text{ good}}\mathop{\mathbb{E}}_{I\subseteq[n],|I|=m}[i\in I]\right]\\ &= \mathop{\mathbb{E}}_{\vec{x}\leftarrow\{0,1\}^n}\left[\sum_{i\in[n],(\vec{x},i)\text{ good}}\frac{m}{n}\right] = \frac{m}{n}\cdot\mathsf{AS}_{\delta_0,\delta_1}(f).\end{aligned}$$

□

Lemma 5.26. *Let $f, g : \{0,1\}^n \to [0,1]$ be γ-almost δ-close to each other. If $\mathsf{AS}_{\delta_0+\delta,\delta_1-\delta}(f) \le s$, where $\delta_1 - \delta_0 > \delta$, then $\mathsf{AS}_{\delta_0,\delta_1}(g) \le s + 2\gamma n$.*

Proof. Let $\delta_0' := \delta_0 + \delta$ and $\delta_1' := \delta_1 - \delta$. Let $B \subseteq \{0,1\}^n$ be the set of inputs on which f and g differ by more than δ, and thus $|B| \le \gamma \cdot 2^n$. Call a pair $(\vec{x}, i) \in \{0,1\}^n \times [n]$ *good* if both $\vec{x}, \vec{x}^{(i)} \notin B$, and *bad* otherwise; the number of bad pairs is at most $2|B| \cdot n \le 2\gamma \cdot n \cdot 2^n$.

For a good pair $(\vec{x}, i)$ contributing to $\mathsf{S}_{\delta_0,\delta_1}(g, \vec{x})$, i.e.,

$$\max\{g(\vec{x}), g(\vec{x}^{(i)})\} > \delta_1 \text{ and } \min\{g(\vec{x}), g(\vec{x}^{(i)})\} < \delta_0,$$

we can see that $(\vec{x}, i)$ also contributes to $\mathsf{S}_{\delta_0',\delta_1'}(f, \vec{x})$. Subsequently, $\mathsf{AS}_{\delta_0,\delta_1}(g) \le \mathsf{AS}_{\delta_0',\delta_1'}(f) + 2\gamma n \le s + 2\gamma n$. □

5.3.3 An Upper Bound on Average Sensitivity

By Lemma 5.24, it suffices to analyze the average gap sensitivity of a sigmoidal circuit to prove a correlation lower bound with the parity function. We consider a strengthened model called *polynomial-adjusted sigmoidal circuits*.

Definition 5.27 (polynomial-adjusted sigmoidal circuit)**.** A *polynomial-adjusted sigmoidal circuit* of degree r on n input variables is a circuit with a single output gate, where each gate v computes

$$G_v(\vec{x}) = \mathsf{Sigmoid}\Big(p_v(\vec{x}) + \sum_{u\to v} w_{uv}\cdot G_u(\vec{x})\Big)$$

for a multilinear polynomial p_v of degree at most r in the input variables (with arbitrary real coefficients), and the sum is over the gates u that feed v.

The *size* of the circuit is the number of its gates, and its *weights* are the coefficients w_{uv} of the wires between gates. The *depth* of a gate is the maximum number of gates on a path from an input variable to it (so the gates of depth 1 are fed only by input variables), and the depth of the circuit is the depth of its output gate.

In particular, a depth-d sigmoidal circuit whose weights are within $[-W, W]$ is a depth-d polynomial-adjusted sigmoidal circuit of degree 1 whose weights are within $[-W, W]$.

Lemma 5.28. *Fix any $d \geq 1$, there exists $c = c(d) \geq 1$ such that the following is true for any $a \geq 1$ and any sufficiently large n and any $1 \leq r \leq n$.*

Suppose that C is an n-input polynomial-adjusted sigmoidal circuit of depth d, degree r, and size at most s, whose weights are within $[-W, W]$ for $W \leq n^a$. Then for every $\delta \geq n^{-a}$ and every $\delta_0, \delta_1 \in [0,1]$ such that $\delta_1 - \delta_0 \geq \delta$,

$$\mathsf{AS}_{\delta_0,\delta_1}(C) \leq s^{1-1/d} \cdot n^{1-1/(2d)} \cdot n^{2/\log\log n} \cdot 2^{r^3 \cdot (\log\log n)^c}.$$

Proof. Let $\Phi_d(s,n,r)$ denote the right-hand side of the bound. If $\Phi_d(s,n,r) \geq n$, the bound holds trivially as $\mathsf{AS}_{\delta_0,\delta_1}(C) \leq n$, so we may assume that $\Phi_d(s,n,r) < n$. Let $c = c(d) \geq 1$ be determined later. We prove the lemma by induction on the depth $d \geq 1$.

Base Case. In the base case, i.e., $d = 1$, the circuit is simply a $\mathsf{PSigmoid}_p(\vec{x})$ gate, where $p(\vec{x})$ is a multilinear polynomial of degree at most r (so $s = 1$).

Fix $\delta_0, \delta_1 \in \mathbb{R}$. By the monotonicity of $\mathsf{Sigmoid}$, there is $\theta \in \mathbb{R}$ such that $\mathsf{PSigmoid}_p(\vec{x}) > \delta_1 \Rightarrow p(\vec{x}) > \theta$, and $\mathsf{PSigmoid}_p(\vec{x}) < \delta_0 \Rightarrow p(\vec{x}) < \theta$. Let $f(\vec{x}) = \mathsf{sgn}(p(\vec{x}) - \theta)$, then by Theorem 3.9,[6] when c is larger than an absolute constant and n is sufficiently large, we have

$$\mathsf{AS}_{\delta_0,\delta_1}(C) \leq \mathsf{AS}(f) \leq \sqrt{n}(\log n)^{O(r\log r)} 2^{O(r^2 \log r)} \leq n^{1/2} \cdot 2^{r^3 (\log\log n)^c}, \tag{14}$$

where the first inequality follows from the definition of average (gap) sensitivity.

Applying Random Restriction. Now we consider the induction case: assume that the lemma holds for d, we will prove that it also holds for $d+1$. Let C be a circuit of depth $d+1$, and $\mathcal{V}$ be the set of gates of depth 1 in C; as they are not fed by any gate, each $u \in \mathcal{V}$ computes $\mathsf{PSigmoid}_{p_u}$ for a polynomial p_u of degree at most r. Let $m \leq n/16$ be determined later (♡), and we apply a random restriction $\rho \leftarrow \mathcal{R}_m$ to the circuit C. Let $t \geq 2, k \geq 1, \gamma \in (0, 1/16)$ with $\gamma \leq \sqrt{m/n}$ be parameters to be determined (⋄). By Lemma 5.22 and union bound, we know that with probability at least $1 - \beta$, where

$$\beta := s\sqrt{m/n} \cdot kt \left(\log(n/m) \cdot \log(tk/\gamma)\right)^{O(r^2)}, \tag{15}$$

every $G_u|_\rho$ with $u \in \mathcal{V}$ is γ-almost $t^{-(k+1)}$-close to a polynomial q_u of degree at most rk; where $O(\cdot)$ hides an absolute constant.

Let $\mathcal{E}$ be the above event, and fix any ρ such that $\mathcal{E}$ happens. Let C' be the circuit on $\{0,1\}^m$ obtained from $C|_\rho$ by removing the gates in $\mathcal{V}$ and, for each remaining gate v, replacing p_v by $p_v|_\rho + \sum_{u \in \mathcal{V},\, u \to v} w_{uv} \cdot q_u$, which has degree at most $r' := \min\{rk, m\}$. We may remove duplicated variables in each monomial so it is multilinear. The depth of every remaining gate decreases by exactly one, so C' is a polynomial-adjusted sigmoidal circuit on m inputs of depth d, degree r', and size at most s, whose weights are within $[-W, W]$.

Let $S \subseteq \{0,1\}^m$ be the set of $\vec{z}$ such that $|G_u|_\rho(\vec{z}) - q_u(\vec{z})| \leq t^{-(k+1)}$ for every $u \in \mathcal{V}$; by union bound, $|S| \geq (1 - s\gamma) \cdot 2^m$. Fix $\vec{z} \in S$. At $\vec{z}$, the preactivations of a remaining gate G_v in $C|_\rho$ and in C' share the term $p_v|_\rho(\vec{z})$ and differ only through the gates feeding v, where each $u \in \mathcal{V}$ is replaced by q_u. As $\mathsf{Sigmoid}$ is $(1/4)$-Lipschitz, the error of v is at most $sW/4$ times the maximum error of the gates feeding v, and the error of each $u \in \mathcal{V}$ is at most $t^{-(k+1)}$. By induction on the depth, $|C|_\rho(\vec{z}) - C'(\vec{z})| \leq \delta_\rho := (sW+1)^d \cdot t^{-(k+1)}$. As a result, $C|_\rho$ and C' are $s\gamma$-almost δ_ρ-close.

[6] It is worth noting that there is a claimed improvement to this upper bound by OpenAI, which may slightly improve our results in certain parameter regimes; see https://github.com/openai/math/blob/main/preprints/Average-Sensitivity-of-Polynomial-Threshold-Functions-September-25-2026/main.pdf.

Applying Induction Hypothesis. We instantiate the deferred parameters (♡), (⋄) as

$$\gamma := n^{-(a+2)}, \quad t := \lceil n^{1/\log_2 \log_2 n} \rceil, \quad k := \lceil 10ad \log_2 \log_2 n \rceil, \quad m := \lceil n^{d/(d+1)}/s^{2/(d+1)} \rceil,$$

taking n sufficiently large compared to a, d throughout. We can verify that:

- $m \le n/16$ and $m \ge n^{(d-1)/d}$: As $\Phi_{d+1}(s,n,r) \ge s^{d/(d+1)} \cdot n^{1-1/(2(d+1))}$ and $\Phi_{d+1}(s,n,r) < n$, we have $s < n^{1/(2d)}$, and thus $m \ge n^{d/(d+1)} \cdot n^{-1/(d(d+1))} = n^{(d-1)/d}$.
- $\delta_\rho \le \delta/4$. Since $t^{k+1} \ge n^{10ad}$, $s < n$, $W \le n^a$, and $\delta \ge n^{-a}$, we have

$$\delta_\rho \le (2n^{a+1})^d / n^{10ad} \le n^{-a-2} \le \delta/4.$$

- $\beta \le s^{1-1/(d+1)} \cdot n^{-1/(2(d+1))} \cdot n^{2/\log\log n} \cdot (\log n)^{O(r^2)}$ for sufficiently large n; see Equation (15) for the definition of β. To see this, notice that $\sqrt{m/n} \le 2n^{-1/(2(d+1))}/s^{1/(d+1)}$, $kt \le n^{2/\log\log n}$, and $\log(n/m) \cdot \log(tk/\gamma) = (\log n)^{O(1)}$.
- $\gamma \le \sqrt{m/n}$ *and* $\gamma < 1/16$. Since $m \ge 1$, $a \ge 1$, we have $\gamma = n^{-(a+2)} \le n^{-1/2} \le \sqrt{m/n}$.

First assume that $d \ge 2$, so that $m \ge \sqrt{n}$. Set $\delta_0' := \delta_0 + \delta_\rho$, $\delta_1' := \delta_1 - \delta_\rho$, $a' := 2(a+1)$, $\delta' := \delta/2$. It follows that $nW \le n^{a+1} \le m^{a'}$, $\delta' = \delta/2 \ge n^{-a}/2 \ge m^{-a'}$, and $\delta_1' - \delta_0' \ge \delta'$. Hence by the induction hypothesis,

$$\mathsf{AS}_{\delta_0',\delta_1'}(C') \le s^{1-1/d} \cdot m^{1-1/(2d)} \cdot m^{2/\log\log m} \cdot 2^{(rk)^3 (\log\log m)^{c'}}, \tag{16}$$

where $c' := c'(d)$ is the constant from IH. As $k = \Theta(ad \log\log n)$, provided that $c \ge c' + 5$ and n is sufficiently large, we can obtain that

$$\mathsf{AS}_{\delta_0',\delta_1'}(C') \le s^{1-1/d} \cdot m^{1-1/(2d)} \cdot n^{2/\log\log n} \cdot 2^{r^3 (\log\log n)^c}. \tag{17}$$

Note that conditioned on $\mathcal{E}$, C' and $C|_\rho$ are $(s\gamma)$-almost δ_ρ-close, and thus Lemma 5.26 gives $\mathsf{AS}_{\delta_0,\delta_1}(C|_\rho) \le \mathsf{AS}_{\delta_0',\delta_1'}(C') + 2s\gamma m$. Subsequently we have

$$\begin{aligned}
\mathsf{AS}_{\delta_0,\delta_1}(C) &= \frac{n}{m} \cdot \mathop{\mathbb{E}}_{\rho \leftarrow \mathcal{R}_m} [\mathsf{AS}_{\delta_0,\delta_1}(C|_\rho)] && \text{(Lemma 5.25)} \\
&\le \frac{n}{m} \left(\mathop{\mathbb{E}}_{\rho \leftarrow \mathcal{R}_m} [\mathsf{AS}_{\delta_0,\delta_1}(C|_\rho) \mid \mathcal{E}] + m \cdot \Pr[\neg \mathcal{E}] \right) \\
&\le \frac{n}{m} \cdot \mathsf{AS}_{\delta_0',\delta_1'}(C') + 2s\gamma n + \beta n. && \text{(Lemma 5.26)}
\end{aligned}$$

Plugging Equation (17) into the first term and using $m \ge n^{d/(d+1)}/s^{2/(d+1)}$,

$$\begin{aligned}
\frac{n}{m} \cdot \mathsf{AS}_{\delta_0',\delta_1'}(C') &\le s^{1-1/d} \cdot n \cdot m^{-1/(2d)} \cdot n^{2/\log\log n} \cdot 2^{r^3 (\log\log n)^c} \\
&\le s^{1-1/(d+1)} \cdot n^{1-1/(2(d+1))} \cdot n^{2/\log\log n} \cdot 2^{r^3 (\log\log n)^c},
\end{aligned}$$

where the last step uses the fact that $1 - 1/d + 1/(d(d+1)) = 1 - 1/(d+1)$. The remaining terms satisfy $2s\gamma n \le 2n^{-a} \le 2$ and $\beta n \le s^{1-1/(d+1)} \cdot n^{1-1/(2(d+1))} \cdot n^{2/\log\log n} \cdot (\log n)^{O(r^2)}$. Subsequently, we can conclude that for some $c = c(d)$ and sufficiently large n, $\mathsf{AS}_{\delta_0,\delta_1}(C) \le s^{1-1/(d+1)} \cdot n^{1-1/(2(d+1))} \cdot n^{2/\log\log n} \cdot 2^{r^3 (\log\log n)^c}$.

For $d = 1$, the circuit C' is a single $\mathsf{PSigmoid}_Q$ gate with $\deg Q \le r'$. We therefore apply the base-case argument directly, which requires no bounds on the weights or separation. Using $r' \le rk = O_a(r \log\log n)$ in the PTF estimate of (14), we obtain

$$\mathsf{AS}_{\delta_0',\delta_1'}(C') \le \sqrt{m}\, 2^{O_a(r^3(\log\log n)^3)}.$$

Here δ_0' and δ_1' are defined as above. For $c > 3$ and sufficiently large n, this also implies (17), since its leading factor $s^{1-1/d} m^{1-1/(2d)}$ equals $\sqrt{m}$ when $d = 1$. □

By instantiating $r = 1$ to Lemma 5.28, we can obtain the average-sensitivity bounds for Boolean functions computed by sigmoidal circuits.

Theorem 5.20. *For any $a, d \ge 1$ and sufficiently large n, let $C : \{0,1\}^n \to [0,1]$ be any depth-d sigmoidal circuit with s gates and weights within $[-n^a, n^a]$. Suppose that it computes a Boolean function $f : \{0,1\}^n \to \{0,1\}$ with separation $\delta \ge n^{-a}$, then $\mathsf{AS}(f) \le s^{1-1/d} \cdot n^{1-1/(2d)+o(1)}$.*

By combining Lemma 5.24 and Lemma 5.28 (for $r = 1$), we can immediately derive the correlation bound with parity for sigmoidal circuits.

Theorem 5.19. *For any $a, d \ge 1$ and sufficiently large n, let $C : \{0,1\}^n \to [0,1]$ be any depth-d sigmoidal circuit with s gates and weights within $[-n^a, n^a]$. Then for any $0 \le \delta_0 < \delta_1 \le 1$ with $\delta_1 - \delta_0 \ge n^{-a}$, we have*

$$\Pr_{\vec{x} \leftarrow \{0,1\}^n}[C(\vec{x}) > \delta_1 \wedge \oplus_n(\vec{x}) = 1] + \Pr_{\vec{x} \leftarrow \{0,1\}^n}[C(\vec{x}) < \delta_0 \wedge \oplus_n(\vec{x}) = 0] \le \frac{1}{2} + s^{1-1/d} \cdot n^{-1/(2d)+o(1)}.$$

In particular, for $d \ge 2$ and $s \le n^{1/(2(d-1))-\Omega(1)}$, if C computes a Boolean function f with separation at least n^{-a}, then $\mathsf{Cor}(f, \oplus_n) \le n^{-\Omega(1)}$.

6 On Sigmoidal Circuits with Large Weight

In this section, we study sigmoidal circuits with large or even unbounded weight. As the simplest case to attack, we first consider depth-two circuits: $\mathsf{Sigmoid} \circ \mathsf{Sigmoid}$. We establish an architecture-preserving weight bound and a gate lower bound for depth-two circuits. We then allow the depth to grow and show that unrestricted weight sigmoidal circuits can compute inner product mod 2 with sublinear size.

6.1 Weight Upper Bounds on Depth 2

Recall that for linear threshold gates (i.e. $\mathsf{THR}(\vec{x}) := \mathbb{I}[\langle \vec{w}, \vec{x} \rangle + \beta \ge 0]$), it is known that integer weights $\vec{w}$ and bias β of $O(n \log n)$-bits suffices (see, e.g., [Hås94]). This upper bound naturally generalizes to polynomial threshold gates; see, e.g., [Sak93]. However, it does not apply to sigmoidal gates (or other standard activation functions) as the output is non-Boolean, so arbitrary changes on the current gate could alter the output of downstream gates.

In this section, we prove a weight bound for depth-2 Sigmoid circuits to compute a Boolean function with no separation. Specifically, every function computed by depth-two sigmoidal circuits

with no separation can be computed by a depth-two sigmoidal circuit of the same architecture and bounded weights.[7]

Theorem 6.1. *Suppose that $f : \{0,1\}^n \to \{0,1\}$ is computed by a depth-2 sigmoidal circuit with $h \geq 1$ gates in the bottom layer and no separation. Then f is computed by a circuit with the same architecture whose weights and biases have absolute values at most*

$$(h(n+1)+1)^{O(h(n+2)+1)} = 2^{O(hn\log(hn))}$$

with no separation.

On Boolean inputs, exponentiating the bottom-layer parameters turns sigmoid outputs into rational functions of the new exponentiated parameters. Clearing positive denominators expresses correct classification as polynomial inequalities. Recall that a subset of $\mathbb{R}^k$ is semi-algebraic if it is defined by a finite Boolean combination of polynomial equalities and inequalities; see [BPR06, Section 2.3].

To prove Theorem 6.1, we will use a standard result due to Basu, Pollack, and Roy [BPR96] that shows the existence of rational points contained in semi-algebraic sets.

Theorem 6.2 ([BPR96, Theorem 4.1.2], see also [BPR06, Theorem 13.16])**.** *Let $\mathcal{S} \subseteq \mathbb{R}^k$ be a nonempty semi-algebraic set defined by only strict inequalities in integer polynomials $\{P_i\}$. If each P_i has degree at most $d \geq 2$ and coefficient bit size at most τ, then $\mathcal{S}$ contains a point in $\mathbb{Q}^k$ whose coordinates have bit-size at most $\tau d^{O(k)}$, where $O(\cdot)$ hides a constant independent of the number of equalities and inequalities.*

Proof of Theorem 6.1. Let $\sigma(x) := \mathsf{Sigmoid}(x)$ for simplicity. Suppose our depth 2 circuit denoted as $N : \{0,1\}^n \to \mathbb{R}$ has h gates in the first layer. Write $c(x) = c_0 + \sum_{j=1}^n u_j x_j$, where u_j is the weight of the direct wire from x_j to the output, or zero if that wire is absent. By definition, any Boolean function computed by the circuit (with no separation) may be parameterized by

$$f(x) = \mathbb{I}\left[c(x) + \sum_{i=1}^{h} v_i \sigma\left(b_i + \sum_{j=1}^{n} w_{ij}x_j\right) > 0\right].$$

(The top sigmoidal gate is removed by the monotonicity of $\mathsf{Sigmoid}(x)$.) By shifting c_0 slightly, we may assume that the output of the inner function in $f(x)$ is always non-zero.

Rewriting variables $B_i = \exp(-b_i)$ and $W_{ij} = \exp(-w_{ij})$, we can express

$$\begin{aligned}\sigma\left(b_i + \sum_{j=1}^{n} w_{ij}x_j\right) &= \left(1 + \exp\left(-b_i - \sum_{j=1}^{n} w_{ij}x_j\right)\right)^{-1} \\ &= \left(1 + B_i \prod_{j\in[n]} W_{ij}^{x_j}\right)^{-1}.\end{aligned}$$

[7] Note that every Boolean function on n inputs can be computed by a depth-two sigmoidal circuit with $O(2^n)$ gates and polynomially bounded weights and biases, by simulating its disjunctive normal form. This construction requires exponentially many gates, whereas our theorem bounds the parameters for much smaller circuit while maintaining the gate count.

Note that $M_i(x) = B_i \prod_{j\in[n]} W_{ij}^{x_j}$ is a monomial term in the variables B_i and W_{ij}. Then

$$\begin{aligned} f(x) &= \mathbb{I}\left[c(x) + \sum_{i=1}^{h} v_i \sigma\left(b_i + \sum_{j=1}^{n} w_{ij}x_j\right) > 0\right] \\ &= \mathbb{I}\left[c(x) + \sum_{i=1}^{h} v_i (1 + M_i(x))^{-1} > 0\right] \\ &= \mathbb{I}\left[c(x) \prod_{i\in[h]} (1 + M_i(x)) + \sum_{i=1}^{h} v_i \prod_{j\neq i} (1 + M_j(x)) > 0\right]. \end{aligned}$$

Let $P_x := c(x)\prod_{i\in[h]}(1 + M_i(x)) + \sum_{i=1}^{h} v_i \prod_{j\neq i}(1 + M_j(x))$ be a polynomial in variables $B_i, W_{ij}, c_0, u_j, v_i$ with $\deg(P_x) \le h(n+1)+1$. To preserve the architecture, substitute $W_{ij} = 1$ for absent bottom wires and $u_j = 0$ or $v_i = 0$ for absent output wires, we don't treat these entries as free variables.

Let $\epsilon_x = 2f(x) - 1$, and let $\mathcal{S}_f$ be the set of parameters where the model classifies the input correctly. Thus

$$\mathcal{S}_f = \{(c_0, u_j, v_i, B_i, W_{ij}) : \epsilon_x P_x > 0 \text{ for every } x \in \{0,1\}^n,\ B_i, W_{ij} > 0\}.$$

We further restrict $\mathcal{S}_f$ by fixing the signs of $W_{ij} - 1$, u_j, and v_i on the nonzero wires to their original signs. The parameters obtained after shifting c_0 satisfy these inequalities, so $\mathcal{S}_f \neq \varnothing$. Subsequently, this is a nonempty open semi-algebraic set in $k \le h(n+2) + n + 1$ variables, with each defining polynomial of degree at most $d = h(n+1)+1$. After expanding P_x, every coefficient is either 0 or 1. Hence the signed constraints have coefficients in $\{-1, 0, 1\}$ and coefficient bit size $O(1)$.

By Theorem 6.2, $\mathcal{S}_f$ contains a rational point $(\hat{c}_0, \hat{u}_j, \hat{v}_i, \hat{B}_i, \hat{W}_{ij})$ whose coordinates have numerator and denominator bit lengths at most $L \le d^{O(k)}$. Since B_i, W_{ij} are positive, it implies that $2^{-L} \le \hat{B}_i, \hat{W}_{ij} \le 2^L$. Then we know that the circuit $\hat{N}$ with the architecture of N and parameters $\hat{c}_0, \hat{u}_j, \hat{v}_i$, $\hat{b}_i = -\ln \hat{B}_i, \hat{w}_{ij} = -\ln \hat{W}_{ij}$ computes f with output threshold $1/2$, where

$$|\hat{b}_i|, |\hat{w}_{ij}| \le L \ln 2 = d^{O(k)}.$$

That is, the bottom layer weights are bounded by $d^{O(k)}$.

We may normalize the top layer weights $\hat{c}_0, \hat{u}_j, \hat{v}_i$ by dividing them by their maximum magnitude $M := \max\big(|\hat{c}_0|, \max_{j\in[n]} |\hat{u}_j|, \max_{i\in[h]} |\hat{v}_i|\big)$. As P_x is linear in $(\hat{c}_0, \hat{u}_j, \hat{v}_i)$, dividing by $M > 0$ preserves every sign, and the top layer weights become at most 1 in magnitude. Hence every parameter is bounded by

$$d^{O(k)} = \big(h(n+1)+1\big)^{O(h(n+2)+1)} = 2^{O(hn\log(hn))}.$$

This completes the proof. □

6.2 Lower Bounds for Depth-2 Circuits

We next prove an $\Omega(n/\log n)$ gate lower bound for depth-two sigmoidal circuits computing the selection function. For depth-2 circuits with sigmoid gates and discrete inputs, better VC dimension upper bounds are known. We use these to obtain a lower bound on the gate complexity of

neural networks with sigmoids computing an explicit Boolean function using ideas from [Koi96]. Throughout this section, we follow the notation and terminology of [Koi96], take logarithms in base 2, and assume for simplicity that n is a power of 2.

We define the *selection Boolean function* $f_n(\vec{x}, \vec{y})$ on inputs $\vec{x} \in \{0,1\}^{\log n}$ and $\vec{y} \in \{0,1\}^n$, where $f_n(\vec{x}, \vec{y}) = y_{\vec{x}}$ is simply the $\vec{x}$-th entry of $\vec{y}$. Here, we identify $\vec{x}$ with the integer in $\{0, \ldots, n-1\}$ whose binary representation is $\vec{x}$, and index the coordinates of $\vec{y}$ from 0 to $n-1$. For brevity, we write f instead of f_n from now on.

The selection function defines a family $\mathcal{F}$ of 2^n Boolean functions on $\{0,1\}^{\log n}$ parameterized by $\vec{y}^* \in \{0,1\}^n$, where each $f_{\vec{y}^*}$ is given by $f_{\vec{y}^*}(\vec{x}) = f(\vec{x}, \vec{y}^*)$. As observed in [Koi96], every Boolean function g on $\log n$ inputs belongs to $\mathcal{F}$, since g is identical to $f_{\vec{y}^*}$ for $\vec{y}^* = (g(0), g(1), \ldots, g(n-1))$. Thus $\mathcal{F}$ shatters $\{0,1\}^{\log n}$ and $\mathrm{VCdim}(\mathcal{F}) = n$.

For a fixed network architecture, we can view some of its parameters (weights and biases) as indeterminate *programmable parameters*, which gives rise to a family of Boolean functions as those parameters vary over $\mathbb{R}$ [Koi96, KS97]. To bound the VC dimension of our restricted architecture, we use a special case of [AB02, Theorem 8.3] (originally due to Goldberg and Jerrum [GJ93]). Recall that $\mathrm{sgn}(\alpha) = 1$ if $\alpha \geq 0$ and $\mathrm{sgn}(\alpha) = 0$ otherwise.

Theorem 6.3 (Special case of [AB02, Theorem 8.3])**.** *Let X be a set and let $p : \mathbb{R}^d \times X \to \mathbb{R}$ be such that, for every fixed $\vec{x} \in X$, the map $\vec{a} \mapsto p(\vec{a}, \vec{x})$ is a polynomial of degree at most l. Then the class $H = \{\vec{x} \mapsto \mathbb{I}[p(\vec{a}, \vec{x}) > 0] : \vec{a} \in \mathbb{R}^d\}$ satisfies $\mathrm{VCdim}(H) \leq 2d \log(12l)$.*

[AB02, Theorem 8.3] is stated for $\mathrm{sgn}(p(\vec{a}, \vec{x}))$ (and more generally for fixed Boolean combinations of k such classes, with bound $2d \log(12kl)$). The strict inequality version above follows immediately by applying their theorem to $-p$, since $\mathbb{I}[p > 0] = 1 - \mathrm{sgn}(-p)$ and replacing every function $h \in H$ with $1 - h$ does not change the VC dimension of the class.

We now upper bound the VC dimension of depth-2 sigmoid architectures whose programmable parameters are restricted to the biases.

Lemma 6.4. *Let C be a depth-2 sigmoid circuit with s gates in which all weights are fixed real numbers and every gate other than the output gate takes only input coordinates as inputs. For each bias vector $\vec{t} = (t_1, \ldots, t_s) \in \mathbb{R}^s$, let $C_{\vec{t}}$ denote the circuit obtained by assigning the bias t_i to the i-th gate. For any fixed threshold $\theta \in (0,1)$, the class $H = \{h_{\vec{t}} : \vec{t} \in \mathbb{R}^s\}$ of Boolean functions $h_{\vec{t}}(\vec{x}) = \mathbb{1}[C_{\vec{t}}(\vec{x}) > \theta]$ satisfies*

$$\mathrm{VCdim}(H) \leq 2s \log(12s).$$

The same bound holds if the output gate is a linear threshold gate instead of a sigmoid gate.

Proof. Number the gates so that gates $1, \ldots, s-1$ are in the first layer and gate s is the output gate. Let $\vec{w}_i$ be the weight vector of gate $i < s$, let $v_1, \ldots, v_{s-1} \in \mathbb{R}$ be the weights on the wires from gates $1, \ldots, s-1$ to the output gate, and let $\vec{u}$ be the vector of weights on the direct wires from the input coordinates to the output gate (wires not present in C are assigned weight zero).

Fix an input $\vec{x}$. For each $i < s$, the output of gate i is

$$\sigma(\vec{w}_i \cdot \vec{x} + t_i) = \frac{1}{1 + e^{-\vec{w}_i \cdot \vec{x}} e^{-t_i}} = \frac{1}{1 + c_i a_i}, \quad \text{where } c_i := e^{-\vec{w}_i \cdot \vec{x}} > 0 \text{ and } a_i := e^{-t_i} > 0,$$

so that c_i depends on $\vec{x}$ but not on $\vec{t}$. The preactivation of the output gate is therefore

$$z(\vec{x}, \vec{t}) := \sum_{i<s} \frac{v_i}{1 + c_i a_i} + \vec{u} \cdot \vec{x} + t_s. \tag{18}$$

Since σ is strictly increasing, $C_{\vec{t}}(\vec{x}) = \sigma(z(\vec{x},\vec{t})) > \theta$ holds if and only if $z(\vec{x},\vec{t}) > \tau$, where $\tau := \sigma^{-1}(\theta) = \ln(\theta/(1-\theta))$. (For a linear threshold output gate, simply take τ to be its threshold.) Multiplying (18) by the strictly positive quantity $\prod_{j<s}(1+c_j a_j) > 0$, the condition $z(\vec{x},\vec{t}) > \tau$ is equivalent to $p(a_1,\ldots,a_{s-1},t_s;\vec{x}) > 0$, where

$$p(a_1,\ldots,a_{s-1},t_s;\vec{x}) := \sum_{i<s} v_i \prod_{\substack{j<s \\ j\neq i}} (1+c_j a_j) + (\vec{u}\cdot\vec{x} + t_s - \tau)\prod_{j<s}(1+c_j a_j).$$

For each fixed $\vec{x}$, $p(\vec{b};\vec{x})$ is a polynomial of degree at most s in the s real variables $\vec{b} = (a_1,\ldots,a_{s-1},t_s)$, with coefficients depending only on $\vec{x}$ and the fixed weights of C. Setting $\vec{b} = (e^{-t_1},\ldots,e^{-t_{s-1}},t_s) \in \mathbb{R}^s$ shows that every $h_{\vec{t}} \in H$ is of the form $\vec{x} \mapsto \mathbb{I}[p(\vec{b};\vec{x}] > 0]$. Thus H is a subclass of the class in Theorem 6.3 with $d = l = s$, which yields $\mathrm{VCdim}(H) \le 2s\log(12s)$. □

We now combine Lemma 6.4 with Koiran's method [Koi96] to derive the gate complexity lower bound for computing f_n.

Theorem 6.5. *Let C be an arbitrary depth-2 sigmoid circuit (with no restriction on the weights) that computes the selection Boolean function f_n with zero separation. Then the number m of gates in C satisfies $n \le 2m\log(12m)$; in particular, $m = \Omega(n/\log n)$.*

Proof. Gates that lie on no directed path to the output gate can be deleted without changing the function computed, so we may assume C has none; since C has depth 2, every gate other than the output gate takes only input coordinates as inputs. Zero separation means that there is a threshold $\theta \in [0,1]$ such that $C(\vec{x},\vec{y}) > \theta$ whenever $f_n(\vec{x},\vec{y}) = 1$ and $C(\vec{x},\vec{y}) \le \theta$ whenever $f_n(\vec{x},\vec{y}) = 0$. Since σ takes values in $(0,1)$ and f_n is not constant, we necessarily have $\theta \in (0,1)$.

Let gates $1,\ldots,m-1$ be the first-layer gates and gate m be the output gate. In $C(\vec{x},\vec{y})$, each first-layer gate $i < m$ computes $\sigma(\vec{w}_i\cdot\vec{x} + \vec{\beta}_i\cdot\vec{y} + \gamma_i)$, and the output gate m computes $\sigma\big(\sum_{i<m} v_i g_i + \vec{u}\cdot\vec{x} + \vec{\beta}_m\cdot\vec{y} + \gamma_m\big)$, where g_i is the output of gate i. Fixing $\vec{y} = \vec{y}^* \in \{0,1\}^n$ simply replaces each bias γ_i with $t_i = \gamma_i + \vec{\beta}_i\cdot\vec{y}^*$ while keeping all weights $\vec{w}_i, v_i, \vec{u}$ fixed. Thus, the resulting circuit $C^{\vec{y}^*}$ on input $\vec{x} \in \{0,1\}^{\log n}$ is precisely the circuit $C_{\vec{t}}$ of Lemma 6.4 (with $s = m$).

Consequently, every function $f_{\vec{y}^*} \in \mathcal{F}$ belongs to the class H of Lemma 6.4 (with $s = m$ and threshold θ), meaning $\mathcal{F} \subseteq H$. Since $\mathrm{VCdim}(\mathcal{F}) = n$, Lemma 6.4 gives

$$n = \mathrm{VCdim}(\mathcal{F}) \le \mathrm{VCdim}(H) \le 2m\log(12m),$$

which implies $m = \Omega(n/\log n)$. □

6.3 A Sublinear Upper Bound for IP Mod 2

Inner product mod 2 (we refer to as IP2) is a classical circuit lower bound target. Given inputs $x, y \in \{0,1\}^n$, IP_2 computes $\sum_{i\in[n]} x_i y_i$ mod 2. For both Threshold gate circuits and ReLU gate circuits, IP2 requires $\Omega(n)$ gates to compute, under unrestricted weights (Due to [ROS94] and [VRPS21] respectively).

However, we prove a surprising *upper bound* for Sigmoid circuits computing the Inner Product Mod 2 function: $\mathsf{IP2}(\vec{x},\vec{y}) := \langle\vec{x},\vec{y}\rangle \bmod 2$. Let n be the length of $\vec{x}$ (and $\vec{y}$), i.e., $2n$ be the input length.

Theorem 1.7 (see Section 6.3)**.** IP2 *can be computed by an unrestricted-weight sigmoidal circuit with* $O(\sqrt{n \log n})$ *gates.*

The construction built on a bit-decoding technique due to Koiran [Koi96], and an observation from [DS96] that multiplication can be approximately computed from a linear combination of Sigmoid. We first explain these two tricks.

Let $\sigma(x) := \mathsf{Sigmoid}(x)$. The following lemma (implicit in [DS96]) shows how to compute multiplication of two real numbers approximately using a linear combination of four σ-units.

Lemma 6.6. *Fix* $t \in \mathbb{R}$*. Suppose* $|x|, |y| < \varepsilon$*, then the* 4 *gate construction*

$$p' = \sigma(t + x + y) - \sigma(t + x) - \sigma(t + y) + \sigma(t)$$

satisfies $p' - \sigma''(t)xy = O(\varepsilon^3)$.

Proof. Writing $\sigma(x + t) = \sigma(t) + x\sigma'(t) + \frac{x^2}{2}\sigma''(t) + O(x^3)$. Calculating directly:

$$\begin{aligned} p' &= \sigma(t + x + y) - \sigma(t + x) - \sigma(t + y) + \sigma(t) \\ &= \big((x + y) - x - y\big)\sigma'(t) + \frac{\sigma''(t)}{2}\big((x + y)^2 - x^2 - y^2\big) + O(\varepsilon^3) \\ &= \frac{\sigma''(t)}{2} \cdot 2xy + O(\varepsilon^3) \\ &= \sigma''(t)xy + O(\varepsilon^3). \end{aligned}$$

Thus taking the difference we obtain $p' - \sigma''(t)xy = O(\varepsilon^3)$. □

Applying the lemma after scaling arbitrarily large x, y results in the following corollary showing that for any pair of reals x, y, we can approximate xy up to arbitrary accuracy using Sigmoid gates.

Corollary 6.7. *Suppose* $x, y \in \mathbb{R}$*,* $\max(|x|, |y|) \le M$*,* $h > 0$*, the* 4 *gate construction*

$$Q_{t,h}(x, y) = \frac{\sigma(t + hx + hy) - \sigma(t + hx) - \sigma(t + hy) + \sigma(t)}{\sigma''(t)h^2}$$

satisfies $|Q_{t,h}(x, y) - xy| \le O(hM^3)$*, where the constant scales with* $1/\sigma''(t)$.

Proof. Writing $p = \sigma(t)$, we have

$$\sigma'''(u) = p(1 - p)(1 - 6p + 6p^2),$$

we can see that $\sigma'''(t) \le \frac{1}{4}$ for all t.

Expanding the difference as an integral we get

$$\sigma(t + a + b) - \sigma(t + a) - \sigma(t + b) + \sigma(t) = \int_0^a \int_0^b \sigma''(t + u + v)\, dv\, du.$$

So,

$$Q_{t,h}(x, y) - xy = \frac{1}{\sigma''(t)h^2} \int_0^{hx} \int_0^{hy} \big(\sigma''(t + u + v) - \sigma''(t)\big)\, dv\, du.$$

Thus, for $h > 0$,

$$\begin{aligned}|Q_{t,h}(x,y) - xy| &= \frac{1}{|\sigma''(t)|h^2}\left|\int_0^{hx}\int_0^{hy}\int_0^1 \sigma'''\big(t + s(u+v)\big)(u+v)\,ds\,dv\,du\right| \\ &\le \frac{1}{4|\sigma''(t)|h^2}\int_0^{h|x|}\int_0^{h|y|}(u+v)\,dv\,du. \\ &\le \frac{h|x||y||x+y|}{4|\sigma''(t)|} = O\left(\frac{hM^3}{\sigma''(t)}\right)\end{aligned}$$

since $|u+v| \le h|x+y|$, thus showing our corollary. □

The above derivation only works because of the accuracy of the degree-2 Taylor approximation to $\mathsf{Sigmoid}(x)$; indeed, this technique generalizes to any function $\gamma(x)$ that is 3 times continuously differentiable in a neighborhood of $a \in \mathbb{R}$ satisfying $\gamma''(a) \neq 0$ (see, e.g., [DS96, Section 2]). This approximation is impossible for ReLU and THR, which is why Theorem 1.7 does not generalize to ReLU and THR circuits (in fact, near-linear lower bounds hold for both classes [ROS94, VRPS21]).

For a threshold circuit or sigmoidal circuit $C(\vec{x})$ with n-input and m-output, we can naturally the function $f : \mathbb{R}^n \to \mathbb{R}^m$ by allowing the input to take values from $\mathbb{R}$ (rather than $\{0,1\}$). The following lemma explain a bit decoding trick in [KS97].

Lemma 6.8. *For every $n \ge 1$, there is a threshold circuit $C : \mathbb{R} \to \{0,1\}^n$ with n gates such that, for every $x \in [0,1)$ with binary expansion $x = 0.x_1x_2\ldots$, its i-th gate outputs x_i for every $i \in [n]$. Trailing digits are treated as 0s.*

Proof. We construct the gates in order. Suppose $G_i(x) = x_i$ for all $i < d$, and let

$$z_d = 2^d x - \sum_{i=1}^{d-1} 2^{d-i} G_i(x) = x_d + \sum_{j>d} 2^{d-j} x_j.$$

The last sum lies in $[0,1)$, so the threshold gate $G_d(x) = \mathbb{I}[z_d \ge 1]$ outputs x_d. For $d = 1$ the first sum is empty, giving the base case. We used 1 gate for each digit, totaling n gates. □

By choosing large enough weights, we may simulate threshold gates by sigmoidal gates within an arbitrarily small error (see Lemma 3.13). Thus we expect a similar statement to hold for sigmoidal circuits.

Note that binary digits cannot be uniformly approximated on $[0,1)$ within error less than $1/2$ by continuous sigmoid circuits. We therefore restrict to a fixed finite input set, as arises from a circuit on Boolean inputs, and formulate the corollary below.

Corollary 6.9. *Fix a finite set $T \subseteq [0,1)$, $\varepsilon > 0$, and $n \ge 1$. There is an n-gate sigmoidal circuit such that, for every $x \in T$, its i-th gate approximates the i-th binary digit of x to within ε, for every $i \in [n]$.*

This follows by applying the finite-input simulation argument of Lemma 3.13 to the threshold bit decoder on T.

Now we explain the upper bound. At a high level, we group the input into k groups of equal size. By *encoding* the input appropriately, we can compute $\mathsf{IP2}$ in each group using the multiplication

gadget in Lemma 6.6 and then decode the t-bit from the result p_i of the multiplication for some $t = \Theta(n \log n/k)$. The key observation is that, instead of decoding for each group and then compute the XOR, we can simply add $\sum_i p_i$ and decode the t-th bit of that number. This leads to a square-root saving in total circuit size by choosing $k = \sqrt{n \log n}$.

Proof of Theorem 1.7. Suppose our input bits are $(x_1, x_2 \dots x_n, y_1, y_2 \dots y_n)$. Define

$$l = \lceil \log_2(n+1) \rceil, \qquad k = \lceil \sqrt{n \log n} \rceil$$

We group the bits into k buckets, each containing at most $\lceil \frac{n}{k} \rceil$ bits. Write $r = \lceil n/k \rceil$ and pad each bucket with zeroes to length r.

Consider any such bucket with input bits $(x_{m_0}, x_{m_1} \dots x_{m_{r-1}})$ and $(y_{m_0}, y_{m_1} \dots y_{m_{r-1}})$. Let

$$X := \sum_{i=0}^{r-1} 2^{il} x_{m_i}, \; Y := \sum_{i=0}^{r-1} 2^{(r-1-i)l} y_{m_i} \tag{19}$$

We define $p = XY$ be the product of X and Y, so

$$p = XY = \sum_{j=0}^{2r-2} 2^{jl} \left(\sum_{\substack{0 \le a,c < r \\ a+r-1-c=j}} x_{m_a} y_{m_c} \right).$$

The coefficient for the term $2^{(r-1)l}$ is directly $\sum_{i=0}^{r-1} x_{m_i} y_{m_i}$. Moreover, if $2^l > r$, this would imply in the binary representation of p, the bits in the region $[(r-1)l, rl-1]$ is directly the binary representation of $\sum_{i=0}^{r-1} x_{m_i} y_{m_i}$, and in particular, the parity of it is the bit of p of weight $2^{(r-1)l}$.

We construct the first layer using four gates for each of the k buckets, padding each bucket with zeros to length r. Let X_b, Y_b, p_b be (19) for the basket b. Fix arbitrary t with $\sigma''(t) > 0$ and put $M = 2^{rl}$, we write $\tilde{p}_b$ following Lemma 6.6:

$$\tilde{p}_b = \frac{\sigma(t + hX_b + hY_b) - \sigma(t + hX_b) - \sigma(t + hY_b) + \sigma(t)}{\sigma''(t) h^2} = p_b + O_t(hM^3).$$

Choose $h > 0$ such that $|\tilde{p}_b - p_b| \le \frac{1}{8k}$ for every bucket and input. Summing up the product from each group, we see that

$$P = \sum_{b=1}^{k} p_b = \sum_{j=0}^{2r-2} C_j 2^{jl}, \qquad C_{r-1} = \sum_{i=1}^{n} x_i y_i, \qquad 0 \le C_j \le n.$$

Choose $l = \lceil \log_2(n+1) \rceil$. Since $0 \le C_j \le n < 2^l$, no carries occur between the base-2^l digits of P. Thus the digits from $2^{(r-1)l}$ to 2^{rl-1} are exactly the digits of the binary representation of C_{r-1}. And note by our choice that $|\tilde{p}_d - p_d| \le \frac{1}{8k}$ we have that

$$|\tilde{P} - P| = |\sum_d \tilde{p}_d - p_d| \le \sum_d |\tilde{p}_d - p_d| \le \frac{1}{8}.$$

Let $L = (2r-1)l$. Since P has at most $2r-1$ base-2^l digits, $0 \le P < 2^L$. Set $z = (\tilde{P} + \frac{1}{2})/2^L$. Also by the error bound on $\tilde{P}$, we have that $\frac{P+3/8}{2^L} \le z \le \frac{P+5/8}{2^L}$. Thus the first L fractional bits

of z are the L-bit binary representation of P. The bit of P in position $(r-1)l$ is therefore the $L-(r-1)l = rl$-th fractional bit of z. By Corollary 6.9, applied to z with error $1/8$, thresholding the last of these rl gates at $1/2$ gives $\mathsf{IP2}(x, y)$.

The circuit requires a total of

$$rl + 4k = \left\lceil \frac{n}{k} \right\rceil \lceil \log_2(n+1) \rceil + 4k$$

gates for a depth of $\lceil \frac{n}{k} \rceil \lceil \log_2(n+1) \rceil + 1$. Choosing $k = \lceil \sqrt{n \log n} \rceil$ obtains an asymptotic gate count of $O(\sqrt{n \log n})$, concluding our construction. □

It is worth noting that ReLU circuits computing $\mathsf{IP2}$ require a minimum of $\Omega(n/\log n)$ gates [VRPS21], and threshold circuits require $\Omega(n)$ gates [ROS94]; thus our result shows a separation between Sigmoid circuits and ReLU circuits.

Upper bound for slightly super-constant depth. It is worth noting that if we adjust our block sizes according to our allocated depth and decode in parallel, we obtain the following theorem.

Theorem 6.10 (Sublinear gate for growing depth)**.** *For any unbounded increasing $d : \mathbb{N} \to \mathbb{N}$, there exists a set of Sigmoid circuits S computing $\mathsf{IP2}$ on n coordinates with depth at most $d(n)$ such that the asymptotic gates of S_i is $o(n)$.*

Proof. Write $D = d(n) \geq 4$ and choose

$$m = \left\lfloor \sqrt{\min(n, D-3)} \right\rfloor, \qquad B = \left\lceil \frac{n}{m} \right\rceil$$

Divide the input into B blocks of size m, padding the final block, we run theorem 1.7 algorithm on each block b in parallel, denoted as the *decoder algorithm* of block b. A block b requires $l \geq \lceil \log_2(m+1) \rceil$, where l denotes the exponent of 2 in the multiplication expansion. Divide the bits in this block into $r = \lceil \sqrt{\frac{m}{l}} \rceil$ groups for multiplication. This means that this block requires a total gate count of

$$G_b = 3\lceil m/r \rceil + rl = O(m/r + rl)$$

gates, and a depth of $1 + rl \leq 1 + \sqrt{ml} + l$.

Let p_b be the parity of $\mathsf{IP2}$ on the inputs to that block. As the algorithm in *theorem* 1.7 can compute p_b up to arbitrary accuracy, we assume the output F_b of the decoder on block b has

$$|F_b - p_b| \leq \frac{1}{16B}$$

Now say $P = \sum_{b \in [B]} p_b$ and $F = \sum_{b \in [B]} F_b$, we have that

$$|P - F| \leq \frac{1}{16}$$

Now we simply compute the parity of $P = round(F)$ using the depth 3 circuit from theorem 4.1 thresholding at half integers instead of integers. Since F is between 0 and B, requires a total of $B+2$ gates to output PF whose distance to the parity of P, denoted P_F has

$$|PF - P_F| \leq \frac{1}{16}.$$

thus we obtain a constant separation for calculating IP_2.

The total gate count is

$$O(B(m/r + rl)) + 3 = O(n \log m/\sqrt{m}) = o(n),$$

since $m \to \infty$ as $D \to \infty$. And the depth is

$$4 + \sqrt{ml} + l = O(\sqrt{\min(n, D)} \log(\min(n, D))$$

showing the theorem. □

7 Further Simulations between Threshold and Sigmoidal Circuits

Given that unrestricted sigmoidal circuits are very hard to analyze, we may consider an intermediate model between polynomial weights sigmoidal circuits with noticeable separation and unrestricted sigmoidal circuits. We now study the connection between this model and threshold circuits.

7.1 Connection to Threshold-of-Majority

One natural candidate is to assume that all weights are small, but it is allowed to compute a function with no separation. We show that this model is strong enough to simulate $\mathsf{THR} \circ \mathsf{MAJ}$; indeed, a further restricted model is equivalent to $\mathsf{THR} \circ \mathsf{MAJ}$. Formally:

Theorem 1.8. *Let $n \in \mathbb{N}$ be sufficiently large and $s = s(n), W = W(n) \geq 1$ be functions, and $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function. The following two statements are equivalent:*

(1) f can be computed by a depth-two sigmoidal circuit of size $s \cdot \mathsf{poly}(n, W)$ with no separation, where all weights are integers in $[-W, W]$ and all biases are real numbers in $(-1, 1)$.

(2) f can be computed by a $\mathsf{THR} \circ \mathsf{MAJ}$ circuit of size $s \cdot \mathsf{poly}(n, W)$, where the bottom MAJ gates have weight taken from $\{-W, \ldots, W\}$.

Moreover, the output gate of the circuit in (1) can be assumed to have weights in $\{\pm 1\}$.

The key observation underlying (2) ⇒ (1) is that for any function $f : T \to \mathbb{R}$, where $T \subseteq \mathbb{R}$ is a finite set, we can represent a small multiple of f as a sum of differences of pairs of sigmoids with nearby biases. Formally:

Lemma 7.1. *Let $T \subseteq \mathbb{R}$ be a finite set. There are $\beta_1, \ldots, \beta_{|T|} \in (-1/2, 1/2)$ such that for every $\psi : T \to \mathbb{R}$ and every sufficiently small $\delta > 0$ (depending on T and ψ), there are $\gamma_1, \ldots, \gamma_{|T|} \in (-1, 1)$ satisfying*

$$\sum_{j=1}^{|T|} \Big(\mathsf{Sigmoid}(u + \gamma_j) - \mathsf{Sigmoid}(u + \beta_j) \Big) = \delta \cdot \psi(u) \quad \text{for every } u \in T.$$

Proof. Recall that $\mathsf{Sigmoid}(u + t) = e^u/(e^u + e^{-t})$. The matrix $(\mathsf{Sigmoid}(u + t))_{u,t \in T} \in \mathbb{R}^{|T| \times |T|}$ is invertible as it is the product of two matrices:

- The diagonal matrix with entries e^u, which is invertible.
- The matrix $(1/(e^u + e^{-t}))_{u,t \in T}$. This is a Cauchy matrix over $\mathbb{R}$ (see, e.g., [HJ13, Section 0.9]) with lists $(e^u)_{u \in T}$ and $(e^{-t})_{t \in T}$ as parameters, where both lists contain distinct elements, and $e^u + e^{-t} \neq 0$. It is known that Cauchy matrices with distinct elements are invertible.

Let $\vec{v}_\beta := (\mathsf{Sigmoid}'(u+\beta))_{u\in T} \in \mathbb{R}^T$, we show that the vectors $\{\vec{v}_\beta\}_{\beta\in(-1/2,1/2)}$ span $\mathbb{R}^T$. Suppose for contradiction that $\vec{w}\in\mathbb{R}^T$ is orthogonal to every $\vec{v}_\beta$ for $\beta\in(-1/2,1/2)$. Let $g(\beta) := \sum_{u\in T} w_u \mathsf{Sigmoid}(u+\beta)$. Then $g'(\beta) = \langle \vec{w}, \vec{v}_\beta\rangle$ vanishes for $\beta\in(-1/2,1/2)$. As $g'(\beta)$ is a rational function of e^β, it vanishes for all $\beta\in\mathbb{R}$. Therefore, g is a constant, which must be 0 as $g(\beta)\to 0$ when $\beta\to-\infty$. In particular, $g(t)=0$ for every $t\in T$, and thus $\vec{w}=\vec{0}$ by the invertibility of the matrix $(\mathsf{Sigmoid}(u+t))_{u,t\in T}\in\mathbb{R}^{|T|\times|T|}$.

Fix $\beta_1,\dots,\beta_{|T|}$ such that the vectors for $\beta=\beta_j$ form a basis. Consider the continuously differentiable vector-valued function $F:\mathbb{R}^{|T|}\to\mathbb{R}^T$ defined by

$$F(\vec{\eta})_u := \sum_{j=1}^{|T|}(\mathsf{Sigmoid}(u+\beta_j+\eta_j)-\mathsf{Sigmoid}(u+\beta_j)).$$

Then $F(\vec{0})=\vec{0}$, and the Jacobian matrix of F at $\vec{0}$ is the matrix $(\mathsf{Sigmoid}'(u+\beta_j))_{u\in T, j\in[|T|]}$, which we have proved to be invertible.

Subsequently, by the inverse function theorem (see, e.g., [Rud76, Theorem 9.24]), there are neighborhoods V, W of $\vec{0}$ such that $F:V\to W$ has a continuous inverse $F^{-1}:W\to V$. Then for every sufficiently small $\delta>0$, there is $\vec{\eta}$ satisfying that $F(\vec{\eta})=\delta\psi$ and $\|\vec{\eta}\|_\infty<1/2$. The lemma then holds by setting $\gamma_j := \beta_j+\eta_j$. □

Proof of Theorem 1.8. (1) ⇒ (2): Assume that f can be computed by a depth-two sigmoidal circuit of size $h := s\cdot\mathsf{poly}(n,W)$ with no separation, where weights are integers from $[-W,W]$. Each gate in the bottom layer is of form $\mathsf{Sigmoid}_{\vec{w},\beta}(\vec{x}) := \mathsf{Sigmoid}(\langle\vec{w},\vec{x}\rangle+\beta)$. As $\vec{x}\in\{0,1\}^n$, $\langle\vec{w},\vec{x}\rangle$ takes values in $T := \{-nW,\dots,nW\}$, and subsequently $\mathsf{Sigmoid}_{\vec{w},\beta}$ can be exactly simulated by a linear combination of the $2nW+1$ functions $\mathbb{I}[\langle\vec{w},\vec{x}\rangle=t]$, $t\in T$, each of which is the difference of the two MAJ gates $\mathbb{I}[\langle\vec{w},\vec{x}\rangle\ge t]$ and $\mathbb{I}[\langle\vec{w},\vec{x}\rangle\ge t+1]$.

We can also replace the top $\mathsf{Sigmoid}$ gate by a THR gate, since $\mathsf{Sigmoid}$ is monotone, and the original circuit computes f with no separation. The resulting $\mathsf{THR}\circ\mathsf{MAJ}$ circuit also computes f, and is of size at most $2(2nW+1)\cdot h = s\cdot\mathsf{poly}(n,W)$.

(2) ⇒ (1): Let $K := nW$ and $T := \{-K,\dots,K\}$. Now assume that $f(\vec{x}) = \mathbb{I}[\sum_{i=1}^m \alpha_i G_i(\vec{x})\ge\theta]$, $m\le s\cdot\mathsf{poly}(n,W)$, where each $G_i(\vec{x})=\mathbb{I}[\langle\vec{a}_i,\vec{x}\rangle\ge b_i]$ is a MAJ gate with $\vec{a}_i\in\{-W,\dots,W\}^n$. By slightly decreasing $\theta\in\mathbb{R}$, we may assume that $\sum_i\alpha_i G_i(\vec{x})\ne\theta$ for every $\vec{x}$. As $\langle\vec{a}_i,\vec{x}\rangle\in T$, we can apply Lemma 7.1 to the function $\psi_i(u) := \alpha_i\cdot\mathbb{I}[u\ge b_i]$ to obtain biases $\beta_{i,1},\dots,\beta_{i,|T|}\in(-1,1)$ such that for every sufficiently small $\delta>0$, there are $\gamma^{(\delta)}_{i,1},\dots,\gamma^{(\delta)}_{i,|T|}\in(-1,1)$ such that

$$\delta\alpha_i\cdot G_i(\vec{x}) = \sum_{j=1}^{|T|}\left(\mathsf{Sigmoid}(\langle\vec{a}_i,\vec{x}\rangle+\gamma^{(\delta)}_{i,j})-\mathsf{Sigmoid}(\langle\vec{a}_i,\vec{x}\rangle+\beta_{i,j})\right)$$

for every $\vec{x}\in\{0,1\}^n$. For $\delta>0$, let

$$S_\delta(\vec{x}) := \sum_{i=1}^m\sum_{j=1}^{|T|}\left(\mathsf{Sigmoid}(\langle\vec{a}_i,\vec{x}\rangle+\gamma^{(\delta)}_{i,j})-\mathsf{Sigmoid}(\langle\vec{a}_i,\vec{x}\rangle+\beta_{i,j})\right) = \sum_{i=1}^m\delta\alpha_i\cdot G_i(\vec{x}).$$

As $\{0,1\}^n$ is finite and $\sum_i\alpha_i G_i(\vec{x})\ne\theta$ for every $\vec{x}$, we can fix a small $\delta>0$ such that $|\delta\theta|<1$ and for every $\vec{x}\in\{0,1\}^n$, $S_\delta(\vec{x})-\delta\theta>0$ if and only if $f(\vec{x})=1$. Hence the depth-two sigmoidal circuit $C(\vec{x}) := \mathsf{Sigmoid}(S_\delta(\vec{x})-\delta\theta)$ satisfies $C(\vec{x})>1/2$ if $f(\vec{x})=1$ and $C(\vec{x})<1/2$ otherwise.

Finally, C has $2m(2nW+1)+1 = s \cdot \mathsf{poly}(n, W)$ gates. Its bottom gates have weights $\vec{a}_i \in \{-W, \ldots, W\}^n$ and biases in $(-1, 1)$; its output gate has weights ± 1 and bias in $(-1, 1)$. □

By applying the same argument, we may simulate $\mathsf{THR} \circ \mathsf{MAJ}^d$ by depth-$(d+1)$ sigmoidal circuits with small integer weights and small biases for any $d \geq 1$.

Theorem 7.2. *Let $d \geq 1$ and $W \in \mathbb{N}$. Suppose that $f : \{0,1\}^n \to \{0,1\}$ is computed by a $\mathsf{THR} \circ \mathsf{MAJ}^d$ circuit C of size s in which the MAJ gates have integer weights in $[-W, W]$. Then f is computed with no separation by a depth-$(d+1)$ sigmoidal circuit of size $\mathsf{poly}(n, s, W)$ in which all gates have integer weights in $[-W, W]$ and biases in $(-1, 1)$.*

Proof. Without loss of generality, we assume C is layered, i.e., every gate reads from the previous layer. For a gate H in layer $\ell \in [d]$ of C, write $H(\vec{x}) = \mathbb{I}[\langle \vec{a}, \vec{x} \rangle \geq b]$ if $\ell = 1$ and $H(\vec{x}) = \mathbb{I}[\sum_i c_i H_i(\vec{x}) \geq b]$ if $\ell \geq 2$, where $H_1, \ldots, H_r$ are the inputs of H. Let $v_H(\vec{x})$ be the integer $\langle \vec{a}, \vec{x} \rangle$ for $\ell = 1$, or $\sum_i c_i H_i(\vec{x})$ for $\ell \geq 2$. Then $H = \phi_H \circ v_H$, where $\phi_H(v) := \mathbb{I}[v \geq b]$ is a Boolean function on $T_\ell := \{-K_\ell, \ldots, K_\ell\}$, with $K_\ell = (n+s)W$.

Let $\delta_1 := 1$ and $\delta_2, \ldots, \delta_{d+1} > 0$ be sufficiently small parameters. For every gate H in layer ℓ and every $\beta \in (-1, 1)$, we will construct a sigmoid gate $S_{H,\beta}$ in layer ℓ of the sigmoidal circuit that outputs *exactly* $\mathsf{Sigmoid}(\beta + \delta_\ell v_H(\vec{x}))$. For $\ell = 1$, let $S_{H,\beta} := \mathsf{Sigmoid}(\langle \vec{a}, \vec{x} \rangle + \beta)$. For $\ell \geq 2$, assume that such gates have been constructed for depth $\ell - 1$, we apply Lemma 7.1 to the set $\delta_{\ell-1} T_{\ell-1}$ and the functions $\psi_i(\delta_{\ell-1} v) := c_i \phi_{H_i}(v), i \in [r]$. Then for any sufficiently small δ_ℓ, there are $\beta_j, \gamma_{i,j} \in (-1, 1)$ such that

$$\sum_{i=1}^{r} \sum_{j=1}^{|T_{\ell-1}|} \Big(S_{H_i,\gamma_{i,j}}(\vec{x}) - S_{H_i,\beta_j}(\vec{x}) \Big) = \delta_\ell \sum_{i=1}^{r} \psi_i\big(\delta_{\ell-1} v_{H_i}(\vec{x})\big) = \delta_\ell v_H(\vec{x}).$$

We define $S_{H,\beta}(\vec{x})$ as the sigmoid of the LHS plus β. Note that we may assume δ_ℓ is small enough so such $\beta_j, \gamma_{i,j} \in (-1, 1)$ exist for all gates in layer ℓ. Let G be the output gate of C that outputs $G(\vec{x}) = \mathbb{I}[\sum_k \alpha_k H_k(\vec{x}) \geq \theta]$ with real α_k, θ. The final sigmoidal circuit outputs $S_{G,0}(\vec{x}) = \mathsf{Sigmoid}(\delta_{d+1}(\sum_k \alpha_k H_k(\vec{x}) - \theta))$. Assuming without loss of generality that $\sum_k \alpha_k H_k(\vec{x}) \neq \theta$ for every $\vec{x}$, the circuit computes f with no separation.

The weights and biases are as claimed by construction. Let $w = \mathsf{poly}(n, s)$ be the total number of wires in C. For each wire that connects the gate H in layer $\ell \leq d$ to the upper layer, we use $2|T_\ell|$ sigmoidal gates of form S_{H,β_j} and $S_{H,\gamma_{i,j}}$. So the final circuit is of size at most $O(\max_\ell |T_\ell| \cdot (s+w)) = \mathsf{poly}(n, s, W)$. □

7.2 Connection to Threshold Circuits

Using a slightly different approach, we can simulate $\mathsf{THR} \circ \mathsf{THR}$ using $\mathsf{Sigmoid} \circ \mathsf{Sigmoid}$ with the restriction that the top $\mathsf{Sigmoid}$ gate has small weights.

Theorem 1.9. *Suppose that $f : \{0,1\}^n \to \{0,1\}$ is computed by a $\mathsf{THR} \circ \mathsf{THR}$ circuit of size s. Then f is computed with no separation by a depth-two sigmoidal circuit of size $s \cdot \mathsf{poly}(n)$ whose output gate has weights in $\{\pm 1\}$.*

Proof. Assume without loss of generality that the output gate is not fed directly by input variables. Write $f(\vec{x}) = \mathbb{I}[\sum_{i=1}^m \alpha_i G_i(\vec{x}) \geq \theta]$, where $G_1, \ldots, G_m$ $(m \leq n+s)$ are the THR gates. By slightly

decreasing θ, we may assume that $\sum_i \alpha_i G_i(\vec{x}) \neq \theta$ for every $\vec{x}$. Each gate G_i can be written as a disjoint OR, and therefore a sum, of at most $p = \mathsf{poly}(n)$ ETHR gates [HP10, Theorem 7](by padding with the constant ETHR gate $\mathbb{I}[1 = 0]$, we may assume that there are exactly p of them). Denote $G_i(\vec{x}) = \sum_{j=1}^{p} \mathbb{I}[L_{i,j}(\vec{x}) = 0]$, where each $L_{i,j}$ is an affine function with integer coefficients. Hence $f(\vec{x}) = \mathbb{I}[P(\vec{x}) > 0]$, where

$$P(\vec{x}) := \sum_{i=1}^{m} \alpha_i \sum_{j=1}^{p} \mathbb{I}[L_{i,j}(\vec{x}) = 0] - \theta$$

satisfies that $P(\vec{x}) \neq 0$ for every $\vec{x} \in \{0,1\}^n$.

Recall that $\mathsf{Sigmoid}'(u) = \mathsf{Sigmoid}(u)\mathsf{Sigmoid}(-u) \leq e^{-|u|}$. Then the function $\mathsf{Sigmoid}'(\lambda L_{i,j}(\vec{x}))$ equals $\mathsf{Sigmoid}'(0) = 1/4$ if $L_{i,j}(\vec{x}) = 0$, and is at most $e^{-\lambda}$ otherwise (since $L_{i,j}$ has integer coefficients). In other words, $4 \cdot \mathsf{Sigmoid}'(\lambda L_{i,j}(\vec{x}))$ converges to the ETHR gate $\mathbb{I}[L_{i,j}(\vec{x}) = 0]$ as $\lambda \to \infty$. For $\lambda, \delta > 0$, let

$$D_{\lambda,\delta}(\vec{x}) := \sum_{i=1}^{m} \sum_{j=1}^{p} \Big(\mathsf{Sigmoid}(\lambda L_{i,j}(\vec{x}) + 4\delta\alpha_i) - \mathsf{Sigmoid}(\lambda L_{i,j}(\vec{x})) \Big) - \delta\theta \tag{20}$$

$$P_\lambda(\vec{x}) := 4 \sum_{i=1}^{m} \alpha_i \sum_{j=1}^{p} \mathsf{Sigmoid}'(\lambda L_{i,j}(\vec{x})) - \theta. \tag{21}$$

By the definition of the derivative, $\lim_{\delta \to 0^+} D_{\lambda,\delta}(\vec{x})/\delta = P_\lambda(\vec{x})$ for every λ and $\vec{x}$. By the discussion above, $\lim_{\lambda \to \infty} P_\lambda(\vec{x}) = P(\vec{x})$ for every $\vec{x}$.

As $\{0,1\}^n$ is finite and $P(\vec{x}) \neq 0$, we can fix a large enough λ such that $P_\lambda(\vec{x})$ has the sign of $P(\vec{x})$ for every $\vec{x}$, then a small enough δ such that $D_{\lambda,\delta}(\vec{x})$ has the sign of $P_\lambda(\vec{x})$ for every $\vec{x}$. Hence the depth-two sigmoidal circuit $C(\vec{x}) := \mathsf{Sigmoid}(D_{\lambda,\delta}(\vec{x}))$ satisfies $C(\vec{x}) > 1/2$ if $f(\vec{x}) = 1$ and $C(\vec{x}) < 1/2$ otherwise. The circuit C has $2pm + 1 \leq 2(n+s)p + 1 = s \cdot \mathsf{poly}(n)$ gates. Its output gate has weights ± 1 and bias $-\delta\theta$. □

Combined with the simulation of threshold circuits with sigmoidal circuits in Lemma 3.13, we can generalize this simulation to larger depth.

Theorem 7.3. *Let $d \geq 1$. Suppose that $f : \{0,1\}^n \to \{0,1\}$ is computed by a depth-$(d+1)$ circuit C of size s consisting of THR gates. Then f is computed with no separation by a depth-$(d+1)$ sigmoidal circuit of size $\mathsf{poly}(n, s)$, where the output gate has weights in $\{\pm 1\}$.*

Proof Sketch. Without loss of generality assume that the circuit is layered. The proof is almost identical to that of Theorem 1.9 with two changes. First, we write each gate G_i at depth d as a sum of ETHR gates, and then simulate each exact threshold gate using the difference of two sigmoids; see (20). Using this method, we can replace the top two layers by two layers of sigmoid gates, where the output gate has weights in $\{\pm 1\}$. Second, the first $d-1$ layers of C are simulated approximately. By perturbing the biases, we may assume that the linear form of every gate in C is nonzero on every input. For each threshold gate $H : \vec{z} \mapsto \mathbb{I}[\langle \vec{w}, \vec{z} \rangle + \beta \geq 0]$, we replace it with a sigmoidal gate

$$S_{H,\lambda_0}(\vec{z}) := \mathsf{Sigmoid}(\lambda_0 \langle \vec{w}, \vec{z} \rangle + \lambda_0 \beta).$$

Note that $\lim_{\lambda_0 \to \infty} S_{H,\lambda_0}(\vec{z}) = H(\vec{z})$ (this is similar to the standard simulation in Lemma 3.13). We replace each gate in the first $d-1$ layer by S_{H,λ_0}, where λ_0 will be determined later. For each

gate H in the $d-1$ layer, we use $\tilde{H}(\vec{x})$ to denote the subcircuit at H in the original circuit, and $\tilde{S}_{H,\lambda_0}(\vec{x})$ to denote the subcircuit at S_{H,λ_0} after the transformation. We can see by induction on the layers that for every input $\vec{x} \in \{0,1\}^n$, $\lim_{\lambda_0 \to \infty} \tilde{S}_{H,\lambda_0}(\vec{x}) = \tilde{H}(\vec{x})$.

Let $D_{\lambda,\delta}(\vec{z})$ and $P_\lambda(\vec{z})$ be the functions as in (20) and (21) to approximate the top two layers, where $\vec{z} \in \{0,1\}^\ell$ denotes the input of the top two layers, i.e., the outputs of the gates in layer $d-1$. Let $H_1, \ldots, H_\ell$ be the gates in layer $d-1$. The final circuit is of form

$$C(\vec{x}) := \mathsf{Sigmoid}\Big(D_{\lambda,\delta}(\tilde{S}_{H_1,\lambda_0}(\vec{x}), \ldots, \tilde{S}_{H_\ell,\lambda_0}(\vec{x}))\Big).$$

Let $\vec{z} = \vec{z}(\vec{x}) := (\tilde{H}_1(\vec{x}), \ldots, \tilde{H}_\ell(\vec{x}))$. Fix λ and δ as before, so that $\mathsf{sgn}(D_{\lambda,\delta}(\vec{z})) = \mathsf{sgn}(P(\vec{z}))$ for every $\vec{z} \in \{0,1\}^\ell$, where $\mathsf{sgn}(P(\vec{z}(\vec{x}))) = f(\vec{x})$ for every $\vec{x} \in \{0,1\}^n$. As $D_{\lambda,\delta}$ is continuous and $\{0,1\}^n$ is finite, we can then fix a large enough $\lambda_0 \in \mathbb{R}$ such that the approximation error between $\tilde{S}_{H_i,\lambda_0}(\vec{x})$ and $\tilde{H}_i(\vec{x})$ in the first $d-1$ layers does not change the sign of $D_{\lambda,\delta}(\vec{z})$ for every $\vec{x} \in \{0,1\}^n$. Subsequently, $C(\vec{x})$ computes $f(\vec{x})$ with no separation, and C has $\mathsf{poly}(n,s)$ gates. □

References

[AA22] Amol Aggarwal and Josh Alman. Optimal-degree polynomial approximations for exponentials and gaussian kernel density estimation. *arXiv preprint arXiv:2205.06249*, 2022.

[AB02] Martin Anthony and Peter L. Bartlett. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 2002.

[ABFR94] James Aspnes, Richard Beigel, Merrick L. Furst, and Steven Rudich. The expressive power of voting polynomials. *Comb.*, 14(2):135–148, 1994.

[ACHS24] Manasseh Ahmed, Tsun-Ming Cheung, Hamed Hatami, and Kusha Sareen. Communication complexity and discrepancy of halfplanes. In *40th International Symposium on Computational Geometry, SoCG 2024*, volume 293 of *LIPIcs*, pages 5:1–5:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024.

[ACW16] Josh Alman, Timothy M. Chan, and R. Ryan Williams. Polynomial representations of threshold functions and algorithmic applications. In Irit Dinur, editor, *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, Hyatt Regency, New Brunswick, New Jersey, USA, October 9-11, 2016*, pages 467–476. IEEE Computer Society, 2016.

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

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

[AT19] Kazuyuki Amano and Shoma Tate. On XOR lemmas for the weight of polynomial threshold functions. *Inf. Comput.*, 269, 2019.

[BBGL08] Saugata Basu, Nayantara Bhatnagar, Parikshit Gopalan, and Richard J. Lipton. Polynomials that sign represent parity and descartes' rule of signs. *Comput. Complex.*, 17(3):377–406, 2008.

[Bop97] Ravi B. Boppana. The average sensitivity of bounded-depth circuits. *Inf. Process. Lett.*, 63(5):257–261, 1997.

[BPR96] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. On the combinatorial and algebraic complexity of quantifier elimination. *J. ACM*, 43(6):1002–1045, 1996.

[BPR06] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in real algebraic geometry*. Springer, 2006.

[Bru90] Jehoshua Bruck. Harmonic analysis of polynomial threshold functions. *SIAM J. Discret. Math.*, 3(2):168–177, 1990.

[BW96] Peter L. Bartlett and Robert C. Williamson. The VC dimension and pseudodimension of two-layer neural networks with discrete inputs. *Neural Computation*, 8(3):625–628, 1996.

[CM22] Arkadev Chattopadhyay and Nikhil S Mande. A short list of equalities induces large sign-rank. *SIAM Journal on Computing*, 51(3):820–848, 2022.

[CSS18] Ruiwen Chen, Rahul Santhanam, and Srikanth Srinivasan. Average-case lower bounds and satisfiability algorithms for small threshold circuits. *Theory Comput.*, 14(1):1–55, 2018.

[CTW26] Lijie Chen, Avishay Tal, and Yichuan Wang. Superquadratic lower bounds for depth-2 linear threshold circuits. In Aditya Bhaskara and Artur Czumaj, editors, *Proceedings of the 58th Annual ACM Symposium on Theory of Computing, STOC 2026, Salt Lake City, UT, USA, June 22-26, 2026*, pages 874–885. ACM, 2026.

[DBS25] Giuseppe Alessio D'Inverno, Monica Bianchini, and Franco Scarselli. VC dimension of Graph Neural Networks with Pfaffian activation functions. *Neural Networks*, 182:106924, 2025.

[DS96] Bhaskar DasGupta and Georg Schnitger. Analog versus discrete neural networks. *Neural Comput.*, 8(4):805–818, 1996.

[DYZ+26] Gavin Dooley, Andy Yang, Yijia Jessica Zhu, David Chiang, Peter Cholak, and Anand Pillay. Vc dimension and expressivity of real-valued transformers. *arXiv preprint arXiv:2609.31999*, 2026.

[ES16] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

[FKL+01] Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt, and Hans Ulrich Simon. Relations between communication complexity, linear arrangements, and computational complexity. In Ramesh Hariharan, Madhavan Mukund, and V. Vinay, editors, *FST TCS 2001: Foundations of Software Technology*

and Theoretical Computer Science, 21st Conference, Bangalore, India, December 13-15, 2001, Proceedings, volume 2245 of *Lecture Notes in Computer Science*, pages 171–182. Springer, 2001.

[FLY22] Zhiyuan Fan, Jiatu Li, and Tianqi Yang. The exact complexity of pseudorandom functions and the black-box natural proof barrier for bootstrapping results in computational complexity. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 962–975, 2022.

[For02] Jürgen Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *J. Comput. Syst. Sci.*, 65(4):612–625, 2002.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[GHR92] Mikael Goldmann, Johan Håstad, and Alexander A. Razborov. Majority gates VS. general weighted threshold gates. *Comput. Complex.*, 2:277–300, 1992.

[GJ93] Paul Goldberg and Mark Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 361–369, 1993.

[Gro24] Martin Grohe. The descriptive complexity of graph neural networks. *TheoretiCS*, 3, 2024.

[Hås94] Johan Håstad. On the size of weights for threshold gates. *SIAM J. Discret. Math.*, 7(3):484–492, 1994.

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

[HG91] Johan Håstad and Mikael Goldmann. On the power of small-depth threshold circuits. *computational complexity*, 1(2):113–129, 1991.

[HHTT23] Pooya Hatami, William M. Hoza, Avishay Tal, and Roei Tell. Depth-d threshold circuits vs. depth-(d+1) AND-OR trees. In Barna Saha and Rocco A. Servedio, editors, *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pages 895–904. ACM, 2023.

[HJ13] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2013.

[HMP+93] András Hajnal, Wolfgang Maass, Pavel Pudlák, Mario Szegedy, and György Turán. Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46(2):129–154, 1993.

[How03] John Mackintosh Howie. *Complex analysis*. Springer, 2003.

[HP10] Kristoffer Arnsfelt Hansen and Vladimir V. Podolskii. Exact threshold circuits. In *Proceedings of the 25th Annual IEEE Conference on Computational Complexity, CCC 2010, Cambridge, Massachusetts, USA, June 9-12, 2010*, pages 270–279. IEEE Computer Society, 2010.

[HS+86] Geoffrey E Hinton, Terrence J Sejnowski, et al. Learning and relearning in boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(282-317):2, 1986.

[IMZ12] Russell Impagliazzo, Raghu Meka, and David Zuckerman. Pseudorandomness from shrinkage. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012*, pages 111–119, 2012.

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

[IPS97] Russell Impagliazzo, Ramamohan Paturi, and Michael E Saks. Size–depth tradeoffs for threshold circuits. *SIAM Journal on Computing*, 26(3):693–707, 1997.

[IZBK23] Omri Isac, Yoni Zohar, Clark Barrett, and Guy Katz. DNN verification, reachability, and the exponential function problem. In *34th International Conference on Concurrency Theory (CONCUR 2023)*, volume 279 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 26:1–26:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.

[Juk12] Stasys Jukna. *Boolean Function Complexity - Advances and Frontiers*, volume 27 of *Algorithms and combinatorics*. Springer, 2012.

[Kan14] Daniel M. Kane. The correct exponent for the gotsman-linial conjecture. *Comput. Complex.*, 23(2):151–175, 2014.

[Kho91] Askold G. Khovanskiĭ. *Fewnomials*, volume 88 of *Translations of Mathematical Monographs*. American Mathematical Society, 1991.

[KKL17] Valentine Kabanets, Daniel M Kane, and Zhenjian Lu. A polynomial restriction lemma with applications. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 615–628, 2017.

[KM97] Marek Karpinski and Angus Macintyre. Polynomial bounds for vc dimension of sigmoidal and general pfaffian neural networks. *Journal of Computer and System Sciences*, 54(1):169–176, 1997.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[Koi96] Pascal Koiran. VC dimension in circuit complexity. In Steven Homer and Jin-Yi Cai, editors, *Proceedings of the Eleventh Annual IEEE Conference on Computational Complexity, Philadelphia, Pennsylvania, USA, May 24-27, 1996*, pages 81–85. IEEE Computer Society, 1996.

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

[KS97] Pascal Koiran and Eduardo D. Sontag. Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54(1):190–198, 1997.

[KW16] Daniel M. Kane and Ryan Williams. Super-linear gate and super-quadratic wire lower bounds for depth-two and depth-three threshold circuits. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 633–643. ACM, 2016.

[LSŠ08] Troy Lee, Adi Shraibman, and Robert Špalek. A direct product theorem for discrepancy. In *Proceedings of the 23rd Annual IEEE Conference on Computational Complexity, CCC 2008*, pages 71–80. IEEE Computer Society, 2008.

[Maa97a] Wolfgang Maass. Bounds for the computational power and learning complexity of analog neural nets. *SIAM J. Comput.*, 26(3):708–732, 1997.

[Maa97b] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

[MCPZ13] James Martens, Arkadev Chattopadhya, Toni Pitassi, and Richard Zemel. On the representational efficiency of restricted boltzmann machines. *Advances in neural information processing systems*, 26, 2013.

[MP43] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[MP88] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, expanded edition, 1988.

[MSS94] Wolfgang Maass, Georg Schnitger, and Eduardo D Sontag. A comparison of the computational power of sigmoid and boolean threshold circuits. In *Theoretical Advances in Neural Computation and Learning*, pages 127–151. Springer, 1994.

[MSS22] William Merrill, Ashish Sabharwal, and Noah A Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, 2022.

[New64] Donald J. Newman. Rational approximation to $|x|$. *Michigan Mathematical Journal*, 11(1):11–14, 1964.

[Nis93] Noam Nisan. The communication complexity of threshold gates. In Dézső Miklós, Vera T. Sós, and Tamás Szőnyi, editors, *Combinatorics, Paul Erdős is Eighty*, volume 1 of *Bolyai Society Mathematical Studies*, pages 301–315. János Bolyai Mathematical Society, Budapest, 1993.

[PP87] Penco P. Petrushev and Vasil A. Popov. *Rational Approximation of Real Functions*, volume 28 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1987.

[PS94] Ramamohan Paturi and Michael E. Saks. Approximating threshold circuits by rational functions. *Inf. Comput.*, 112(2):257–272, 1994.

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

[Raz92] Alexander A. Razborov. On small depth threshold circuits. In Otto Nurmi and Esko Ukkonen, editors, *Algorithm Theory - SWAT '92, Third Scandinavian Workshop on Algorithm Theory, Helsinki, Finland, July 8-10, 1992, Proceedings*, volume 621 of *Lecture Notes in Computer Science*, pages 42–52. Springer, 1992.

[ROS94] Vwani P. Roychowdhury, Alon Orlitsky, and Kai-Yeung Siu. Lower bounds on threshold and related circuits via communication complexity. *IEEE Trans. Inf. Theory*, 40(2):467–474, 1994.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, third edition, 1976.

[Sak93] Michael E. Saks. Slicing the hypercube. In K. Walker, editor, *Surveys in Combinatorics, 1993*, volume 187 of *London Mathematical Society Lecture Note Series*, pages 211–256. Cambridge University Press, 1993.

[She13] Alexander A. Sherstov. Making polynomials robust to noise. *Theory Comput.*, 9:593–615, 2013.

[Sho90] Victor Shoup. New algorithms for finding irreducible polynomials over finite fields. *Mathematics of Computation*, 54(189):435–447, 1990.

[SÖ15] Can Eren Sezener and Erhan Öztop. Minimal sign representation of boolean functions: Algorithms and exact results for low dimensions. *Neural Comput.*, 27(8):1796–1823, 2015.

[Son97] Eduardo D. Sontag. Shattering all sets of k points in "general position" requires $(k-1)/2$ parameters. *Neural Computation*, 9(2):337–348, 1997.

[SRK92] Kai-Yeung Siu, Vwani P. Roychowdhury, and Thomas Kailath. Computing with almost optimal size neural networks. In Stephen Jose Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems 5, [NIPS Conference, Denver, Colorado, USA, November 30 - December 3, 1992]*, pages 19–26. Morgan Kaufmann, 1992.

[SRK94] Kai-Yeung Siu, Vwani P. Roychowdhury, and Thomas Kailath. Rational approximation techniques for analysis of neural networks. *IEEE Transactions on Information Theory*, 40(2):455–466, 1994.

[SRV24] Itay Safran, Daniel Reichman, and Paul Valiant. How many neurons does it take to approximate the maximum? In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3156–3183. SIAM, 2024.

[SRV25] Itay Safran, Daniel Reichman, and Paul Valiant. Depth separations in neural networks: Separating the dimension from the accuracy. In *Proceedings of the Thirty Eighth Conference on Learning Theory (COLT)*, volume 291 of *Proceedings of Machine Learning Research*, pages 5108–5142. PMLR, 2025.

[Tam16] Suguru Tamaki. A satisfiability algorithm for depth two circuits with a sub-quadratic number of symmetric and threshold gates. *Electron. Colloquium Comput. Complex.*, TR16, 2016.

[VRPS21] Gal Vardi, Daniel Reichman, Toniann Pitassi, and Ohad Shamir. Size and depth separation in approximating benign functions with neural networks. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 4195–4223. PMLR, 2021.

[VS20] Gal Vardi and Ohad Shamir. Neural networks with small weights and depth-separation barriers. *Advances in neural information processing systems*, 33:19433–19442, 2020.

[Wil14] Ryan Williams. Nonuniform ACC circuit lower bounds. *J. ACM*, 61(1):2:1–2:32, 2014.

A Further Related Work

ReLUs, RBMs, and piece-wise polynomial activations. Martens, Chattopadhyay, Pitassi, and Zemel [MCPZ13] studied the representational efficiency of Restricted Boltzmann Machines (RBMs) [HS+86], characterizing the RBM free energy as a depth-2 network of "softplus" ($\ln(1+e^y)$) units with unit output weights and relating it to "hardplus" (ReLU) networks and threshold circuits. While they showed that size-$O(n^2)$ RBMs can represent any symmetric distribution, they proved $2^{\Omega(n)}$ size lower bounds for computing $\mathrm{IP2}_n$ assuming weights are polynomially bounded. Similarly, Maass [Maa97a] showed that for networks with piece-wise polynomial activations, polynomial rational weights suffice up to a polynomial size overhead.

In the unbounded-weight regime, Vardi, Reichman, Pitassi, and Shamir [VRPS21] (see also [VS20]) proved an $\Omega(n/\log n)$ gate lower bound for constant-depth ReLU networks computing $\mathsf{IP2}_n$, contrasting with our $O(\sqrt{n \log n})$ sigmoidal upper bound (Theorem 1.7).

For functions with *real-valued* inputs, Maass [Maa97b] established a stronger $\Omega(n)$ gate lower bound against unrestricted sigmoidal networks of arbitrary depth (with zero separation, and even allowing a mix of sigmoidal and threshold gates) computing the Element Distinctness function $\mathrm{ED}_n : (\mathbb{R}^+)^n \to \{0,1\}$, which outputs 1 if $x_i = x_j$ for some $i \neq j$ and 0 if $|x_i - x_j| \geq 1$ for all $i \neq j$. Crucially, Maass's proof exploits the continuous input domain $(\mathbb{R}^+)^n$ to scale the first input coordinate by a single programmable multiplier $\lambda \in \mathbb{R}^+$, allowing a network with k gates to shatter *every* subset $S \subset \mathbb{R}^+$ of size $n-1$ using $k+1$ programmable parameters; invoking Sontag's $(2w+1)$ upper bound [Son97] on the "Sontag dimension" (the maximum d such that *all* sets of d points in general position are shattered by w parameters) immediately yields $k \geq (n-4)/2$.

Because this scaling and general-position argument requires *continuous* inputs, the best known gate lower bound for *Boolean-input* functions against unrestricted sigmoidal circuits of general depth remains $\Omega(n^{1/4})$ [Koi96].

Size lower bounds for neural networks with continuous inputs and outputs are known. These apply for a variety of activation functions including signoids. Eldan and Shamir [ES16] provide a construction of a function that can be approximated in the ℓ^2-norm by a depth 3 network but requires exponential size to be approximated by a depth-2 network. Safran, Reichman and Valiant [SRV24] prove a linear size lower bounds for neural networks approximating the maximum function: their lower bound holds regardless of the depth of the network and includes sigmoidal networks. These lower bounds rely on the continuity of the domain in essential ways [SRV25] and do not imply size lower bounds for sigmooidal networks computing Boolean functions.

Communication complexity. Communication complexity [KN97] is a standard tool for depth-2 threshold circuit lower bounds, yielding $\Omega(n)$ gate lower bounds for $\mathrm{IP2}_n$ against $\mathsf{THR} \circ \mathsf{THR}$ [Nis93, ROS94] and $2^{\Omega(n)}$ lower bounds against $\mathsf{MAJ} \circ \mathsf{THR}$ [Raz92, HMP+93] and $\mathsf{THR} \circ \mathsf{MAJ}$ [FKL+01, For02, LSŠ08, She13, ACHS24]. While real communication complexity has been studied in connection with analog networks [Koi96], it does not directly yield gate lower bounds for Boolean-input sigmoidal circuits with unrestricted weights, since a single real-valued gate output can encode $\Omega(n)$ bits of information (as each boolean input could have a different output, communicating one output out of 2^n would require n bits).

B Missing Proofs

B.1 Proof of the Parameter Compression Lemma

Lemma 5.13 (Parameter compression)**.** *Let $m \geq 1$, $\varepsilon \in (0,1)$, $\vec{w} \in \mathbb{R}^m$, and $b \in \mathbb{R}$. Then there are $\vec{w}' \in \mathbb{R}^m$ and $b' \in \mathbb{R}$, whose entries have absolute values at most $(m+1)! \cdot \ln(3/\varepsilon)$, such that for every $\vec{z} \in \{0,1\}^m$, $\left|\mathsf{Sigmoid}(\langle \vec{w}, \vec{z} \rangle + b) - \mathsf{Sigmoid}(\langle \vec{w}', \vec{z} \rangle + b')\right| \leq \varepsilon$.*

Proof. Let $g(\vec{z}) := \mathsf{Sigmoid}(\langle \vec{w}, \vec{z} \rangle + b)$ and $B := \lceil 1/\varepsilon \rceil + 1$, so that $1/B < \varepsilon$ and $B \leq 3/\varepsilon$. For every $\vec{z} \in \{0,1\}^m$, pick $j_{\vec{z}} \in \{0, 1, \ldots, B-1\}$ such that $g(\vec{z}) \in [j_{\vec{z}}/B, (j_{\vec{z}}+1)/B]$. As $\mathsf{Sigmoid}$ is increasing, $\mathsf{Sigmoid}(\langle \vec{w}', \vec{z} \rangle + b')$ lies in the same interval if and only if

$$\mathsf{Sigmoid}^{-1}(j_{\vec{z}}/B) \leq \langle \vec{w}', \vec{z} \rangle + b' \leq \mathsf{Sigmoid}^{-1}((j_{\vec{z}}+1)/B),$$

where for the boundary cases, the left (resp. right) inequality is dropped if $j_{\vec{z}} = 0$ (resp. $j_{\vec{z}} = B-1$). As $\mathsf{Sigmoid}^{-1}(y) = \ln(y/(1-y))$, these bounds have absolute values at most $\ln(B-1)$.

Let $P \subseteq \mathbb{R}^{m+1}$ be the set of $(\vec{w}', b')$ satisfying all these inequalities, and in addition having every coordinate of the same sign as the corresponding coordinate of $(\vec{w}, b)$ (e.g., $w'_i \geq 0$ if $w_i \geq 0$, and $w'_i \leq 0$ otherwise). Then P is a polyhedron that contains $(\vec{w}, b)$. Because of the sign constraints, P contains no line, so it has a vertex. Each defining inequality of P has the form $\langle \vec{a}, (\vec{w}', b') \rangle \leq c$ with $\vec{a} \in \{0, \pm 1\}^{m+1}$ and $|c| \leq \ln(B-1)$, and each vertex is the unique solution of $m+1$ linearly independent such inequalities taken with equality. Fix $(\vec{w}', b')$ to be any vertex, and write the corresponding $m+1$ tight inequalities as $A\vec{u} = \vec{c}$, where $\vec{u} := (\vec{w}', b')$, $A \in \{0, \pm 1\}^{(m+1)\times(m+1)}$ is invertible, and every entry of $\vec{c}$ has absolute value at most $\ln(B-1)$. By Cramer's rule, $u_i = \det(A_i)/\det(A)$, where A_i is obtained from A by replacing its i-th column with $\vec{c}$. Here $\det(A)$ is a nonzero integer. Moreover, by expanding $\det(A_i)$ along its i-th column, we know that

$$|\det(A_i)| \;\leq\; \sum_{r=1}^{m+1} |c_r| \cdot m! \;\leq\; (m+1)! \cdot \ln(B-1) \;\leq\; (m+1)! \cdot \ln(3/\varepsilon),$$

where the second inequality follows from the fact that every $m \times m$ minor of A has entries in $\{0, \pm 1\}$ and hence absolute value at most $m!$. Therefore, for every $\vec{z}$, both $g(\vec{z})$ and $\mathsf{Sigmoid}(\langle \vec{w}', \vec{z} \rangle + b')$ lie in an interval of length $1/B < \varepsilon$. □

B.2 Proof of the Restriction Lemma

Lemma B.1 (implied by [KKL17, Lemma 1.6])**.** *Let $f(\vec{x}) = \mathsf{sgn}(p(\vec{x}))$, where $p(\vec{x})$ is a degree-d multilinear polynomial. Then for every $0 < \delta \leq 1/16$ and $m \leq n/16$,*

$$\Pr_{\rho \leftarrow \mathcal{R}_m} [f|_\rho \text{ is not } \delta\text{-almost a constant}] \leq (\sqrt{m/n} + \delta) \cdot (\log(n/m) \cdot \log \delta^{-1})^{O(d^2)}.$$

Recall that a Boolean function is said to be δ-almost a constant if for some $b \in \{0,1\}$, all but a δ fraction of input evaluates to b. We say that a polynomial $p(\vec{x})$ is δ-almost, ε-concentrated within $[a,b]$ if there exists $z \in [a,b]$ such that all but a δ fraction of input $\vec{x}$, $|p(\vec{x}) - z| \leq \varepsilon$. The following lemma is a simple corollary of the restriction lemma.

Lemma B.2. *For any degree-d multilinear polynomial $p(\vec{x})$ with n variables, $m \le n/16$, $\delta \in (0,1)$ with $\delta \le \sqrt{m/n}$, $\varepsilon > 0$, $a, b \in \mathbb{R}$ such that $b - a \ge \varepsilon$, over a random restriction $\rho \leftarrow \mathcal{R}_m$, with probability*

$$1 - O\left(\sqrt{m/n} \cdot (b-a)\varepsilon^{-1}\big(\log(n/m) \cdot \log((b-a)\varepsilon^{-1}\delta^{-1})\big)^{O(d^2)}\right),$$

at least one of the following three conditions holds:

- *$p|_\rho(\vec{z}) \le a$ for all but δ-fraction of $\vec{z} \in \{0,1\}^m$,*
- *$p|_\rho(\vec{z}) \ge b$ for all but δ-fraction of $\vec{z} \in \{0,1\}^m$,*
- *$p|_\rho$ is δ-almost ε-concentrated within $[a, b]$.*

Proof. Set $N := \lceil (b-a)/\varepsilon \rceil$, $\varepsilon' := (b-a)/N \in [\varepsilon/2, \varepsilon]$, and $\tau_i := a + i\varepsilon'$ for $i = 0, 1, \ldots, N$, so $\tau_0 = a$ and $\tau_N = b$. We adopt the convention $\mathsf{sgn}(x) = 1$ if $x > 0$ and $\mathsf{sgn}(x) = 0$ otherwise. For each i, let $g_i(\vec{x}) := \mathsf{sgn}(p(\vec{x}) - \tau_i)$, the sign of a degree-$d$ multilinear polynomial. Let δ' be a parameter to be determined later.

Applying Lemma B.1 to each g_i and by union bound, with probability at least

$$1 - (N+1) \cdot (\sqrt{m/n} + \delta') \cdot \big(\log(n/m) \cdot \log(\delta'^{-1})\big)^{O(d^2)}, \tag{22}$$

every $g_i|_\rho$ is δ'-almost a constant; that is, for each i there exists $c_i \in \{0,1\}$ such that the set $B_i := \{\vec{z} \in \{0,1\}^m : g_i|_\rho(\vec{z}) \ne c_i\}$ has $|B_i|/2^m \le \delta'$.

Condition on this event. Let $B := \bigcup_{i=0}^{N} B_i$, where $|B|/2^m \le (N+1)\delta'$. We set $\delta' := \min\{\delta/(N+1), 1/16\}$ so $|B|/2^m \le \delta$, and Lemma B.1 applies as $\delta' \le 1/16$. In particular, there exists a $\vec{z}_0 \notin B$. Subsequently, the tuple $(c_0, \ldots, c_N)$ is non-increasing: assume not, i.e., $c_i = 0$ but $c_{i+1} = 1$, we will have

$$g_i|_\rho(\vec{z}_0) = c_i = 0, \quad g_{i+1}|_\rho(\vec{z}_0) = c_{i+1} = 1, \qquad (\text{by } \vec{z}_0 \notin B)$$

which is impossible by the definition of g_i's. Let $i^* \in \{-1, 0, \ldots, N\}$ be the index of the last 1 in $(c_0, \ldots, c_N)$ (with $i^* = -1$ if all entries are 0). If $i^* = -1$, then $p|_\rho(\vec{z}) \le \tau_0 = a$ for every $\vec{z} \notin B$, satisfying the first bullet. If $i^* = N$, then $p|_\rho(\vec{z}) > \tau_N = b$ for every $\vec{z} \notin B$, satisfying the second bullet. Otherwise $i^* \in \{0, \ldots, N-1\}$, $\tau_{i^*} < p|_\rho(\vec{z}) \le \tau_{i^*+1}$ for any $\vec{z} \notin B$, satisfying the third bullet. It then follows by invoking the parameters to Equation (22), using $\delta' \le \delta \le \sqrt{m/n}$ so that $\sqrt{m/n} + \delta' \le 2\sqrt{m/n}$. □

Lemma 5.22. *For any degree-d multilinear polynomial $p(\vec{x})$ with n variables, $t \ge 2$, $k \ge 1$, $m \le n/16$, $\delta \in (0, 1/16)$ with $\delta \le \sqrt{m/n}$, over a random restriction $\rho \leftarrow \mathcal{R}_m$, with probability at least*

$$1 - O\left(\sqrt{m/n} \cdot kt \cdot \big(\log(n/m) \cdot \log(tk/\delta)\big)^{O(d^2)}\right)$$

there is a degree-(dk) polynomial $q : \mathbb{R}^m \to \mathbb{R}$ such that $\mathsf{PSigmoid}_p|_\rho, q : \{0,1\}^m \to \mathbb{R}$ are δ-almost $t^{-(k+1)}$-close to each other.

Proof. Apply Lemma B.2 to p with interval $[-(k+1)\ln t, (k+1)\ln t]$, $\varepsilon := 1/t$, and $\delta := \delta$. Over $\rho \leftarrow \mathcal{R}_m$, with probability at least

$$\begin{aligned} &1 - O\left(\sqrt{m/n} \cdot tk \ln t \cdot \big(\log(n/m) \cdot \log(tk \ln t \cdot \delta^{-1})\big)^{O(d^2)}\right) \\ &\ge 1 - O\left(\sqrt{m/n} \cdot kt \cdot \big(\log(n/m) \cdot \log(tk/\delta)\big)^{O(d^2)}\right), \end{aligned}$$

at least one of the three conditions of Lemma B.2 holds. We show that in each case, $\mathsf{PSigmoid}_p|_\rho$ admits a degree-(dk) polynomial $q : \mathbb{R}^m \to \mathbb{R}$ that is δ-almost $t^{-(k+1)}$-close.

- (*Case 1.*) Suppose $p|_\rho(\vec{z}) \le -(k+1)\ln t$ for all but δ-fraction of $\vec{z} \in \{0,1\}^m$. Then $\mathsf{Sigmoid}(p|_\rho(\vec{z})) \le e^{-(k+1)\ln t} = t^{-(k+1)}$ for these $\vec{z}$. It suffices to choose the constant polynomial $q(\vec{z}) \equiv 0$.
- (*Case 2.*) Similarly, it suffices to choose the constant polynomial $q(\vec{z}) := 1$.
- (*Case 3.*) Suppose $p|_\rho$ is δ-almost $(1/t)$-concentrated within $[-(k+1)\ln t, (k+1)\ln t]$, i.e., there exists $\mu \in [-(k+1)\ln t, (k+1)\ln t]$ such that $|p|_\rho(\vec{z}) - \mu| \le 1/t$ for all but δ-fraction of $\vec{z} \in \{0,1\}^m$. Let q_T be the degree-k Taylor approximation of $\mathsf{Sigmoid}$ at μ; by Taylor's theorem and Corollary 5.3, $|\mathsf{Sigmoid}(y) - q_T(y)| \le (2/\pi)^{k+1}|y-\mu|^{k+1} \le t^{-(k+1)}$ whenever $|y-\mu| \le 1/t$. Therefore, the degree-dk polynomial $q(\vec{z}) := q_T(p|_\rho(\vec{z}))$ satisfies the condition.

This completes the proof. □

B.3 Rational Approximation of Sigmoid

In this subsection, we prove Lemma 4.5. We will need Newman's theorem on the rational approximation of the absolute value function $|y|$.

Theorem B.3 (Newman's theorem [New64]; see also [PP87, Section 4.1])**.** *There is an absolute constant C_N such that for every integer $n \ge 1$, there is a rational function N_n of degree at most n with $\big|N_n(y) - |y|\big| \le C_N e^{-\sqrt{n}}$ for all $y \in [-1,1]$.*

Similar to [PS94, Corollary 3], we need an approximation of the sign function via rational functions when the input is bounded away from 0. Formally:

Proposition B.4 (Approximate sign function)**.** *For every $B \ge 2$ and $\eta \in (0,1)$, there is a rational function S of degree $O(\ln^2(B/\eta))$ with no pole in $[-B,B]$ such that for every $y \in [-B,B]$, $S(y) \in [-\eta, 1+\eta]$; moreover, $S(y) \in [1-\eta, 1+\eta]$ if $y \ge 1$, and $S(y) \in [-\eta,\eta]$ if $y \le -1$.*

Proof. Let $n := \lceil \ln^2(C_N(B+1)/(2\eta))\rceil = O(\ln^2(B/\eta))$ and $N(y) := (B+1)\cdot N_n(y/(B+1))$, where N_n is the rational function of Theorem B.3. Then $\big|N(y) - |y|\big| \le C_N(B+1)e^{-\sqrt{n}} \le 2\eta$ for $|y| \le B+1$; in particular, N has no pole on $[-B-1, B+1]$. Define

$$S(y) := \frac{1}{2} + \frac{N(y+1) - N(y-1)}{4},$$

a rational function of degree at most $2n$. For $y \in [-B,B]$, we have $|y \pm 1| \le B+1$, and hence $S(y)$ approximates within an error of η the function $\frac{1}{2} + \frac{|y+1|-|y-1|}{4}$, which equals 1 for $y \ge 1$, equals 0 for $y \le -1$, and lies in $[0,1]$ for all y. □

We are now ready to prove Lemma 4.5, restated below. It follows essentially the proof of [SRK94, Theorem 10]; for completeness, we provide a self-contained proof.

Lemma 4.5 (Rational approximation of $\mathsf{Sigmoid}$)**.** *There is a constant C_0 such that for every $a \ge 1$ and $\varepsilon \in (0, 1/2]$, there is a rational function r of degree at most $C_0 \ln^2(a/\varepsilon)$, with no pole in $[-a,a]$, such that $|r(x) - \mathsf{Sigmoid}(x)| \le \varepsilon$ for every $x \in [-a,a]$.*

Proof. Let $\eta = \Omega(\varepsilon)$ be a parameter to be determined, and $L := O(\ln(1/\varepsilon))$ so that $e^{-L+1} \le \eta$. At a high level, we approximate $\mathsf{Sigmoid}(x)$ on $[-a, L]$ using the Taylor expansion of e^{-x}, on $(L,a]$ by the constant 1; the two approximations are combined using the rational approximation of the sign function in Proposition B.4. Note that when $L \ge a$, it suffices to use only Taylor approximation.

Let $T_k(y) = \sum_{j=0}^{k-1} (-y)^j/j!$ be the degree-$(k-1)$ Taylor approximation of e^{-y}. It follows that for any $y \in [-(L+1), L+1]$, there exists ξ between 0 and y such that

$$|T_k(y) - e^{-y}| \le e^{-\xi} \cdot \frac{|y|^k}{k!} \le e^{L+1} \cdot \frac{(L+1)^k}{(k/e)^k}.$$

Let k be the least odd integer with $k \ge 10(L+1)$, so that $k = O(L)$. Then the error term is at most $e^{L+1} \cdot (e/10)^{10(L+1)} \le e^{-L+1} \le \eta$. Note that $\mathsf{Sigmoid}(x) = \frac{1}{1+e^{-x}}$, and the function $x \mapsto \frac{1}{1+x}$ is 1-Lipschitz over $\mathbb{R}_+$. Also, since $k-1$ is even, $T_k(y) > 0$ for all $y \in \mathbb{R}$. Therefore, the function $r_0(x) := 1/(1+T_k(x))$ is a degree-$(k-1)$ rational function without poles in $\mathbb{R}$ that approximates $\mathsf{Sigmoid}(x)$ with error η for every $x \in [-(L+1), L+1]$.

Moreover, for $x \le -(L+1)$, all terms of $T_k(x)$ are non-negative, so $T_k(x) \ge T_k(-(L+1)) \ge e^{L+1} - \eta$. In that case, $r_0(x) \le e^{-(L+1)}$ and $\mathsf{Sigmoid}(x) \le e^{-(L+1)}$, and thus $|r_0(x) - \mathsf{Sigmoid}(x)| \le \eta$.

Now we combine the approximations above. Let S be the rational function of Proposition B.4 with $B := a + L$ and error η, and let

$$r(x) := r_0(x) + S(x - L) \cdot \big(1 - r_0(x)\big).$$

For $x \in [-a, a]$ we have $|x - L| \le B$, so r is a rational function of degree at most $k + O(\ln^2(B/\eta)) = O(\ln^2(a/\varepsilon))$ with no pole in $[-a, a]$. Write $S_0 = S(x - L)$.

- $S_0 \in [-\eta, 1+\eta]$;
- $S_0 \in [1-\eta, 1+\eta]$ if $x \ge L+1$ and $S_0 \in [-\eta, \eta]$ if $x \le L-1$.

For every $x \in [-a, a]$, we will show that $|r(x) - \mathsf{Sigmoid}(x)| \le \varepsilon$ by a case analysis. Note that

$$r(x) - \mathsf{Sigmoid}(x) = (r_0(x) - \mathsf{Sigmoid}(x))(1 - S_0) + S_0(1 - \mathsf{Sigmoid}(x)).$$

For the first term: if $x \le L+1$, then $|r_0(x) - \mathsf{Sigmoid}(x)| \le \eta$ and $|1 - S_0| \le 1 + \eta$; if $x > L+1$, then $|r_0(x) - \mathsf{Sigmoid}(x)| \le 1$ (as $r_0(x), \mathsf{Sigmoid}(x) \in (0,1)$) and $|1 - S_0| \le \eta$. In both cases the first term is at most $\eta(1+\eta) \le 2\eta$. For the second term: if $x \le L-1$, then $|S_0| \le \eta$ and $0 \le 1 - \mathsf{Sigmoid}(x) \le 1$; if $x > L-1$, then $|S_0| \le 1+\eta$ and $0 \le 1 - \mathsf{Sigmoid}(x) \le e^{-x} \le e^{-(L-1)} \le \eta$. Therefore, for any $x \in [-a, a]$,

$$|r(x) - \mathsf{Sigmoid}(x)| \le 4\eta.$$

It then suffices to set $\eta := \varepsilon/4$. □

AI Methodology

Following the recommendation of "AI and TCS: The Next Six Months" white paper, we describe the usage of AI tools in this research project in detail.

Activity	**AI used**
Asking the research question	No
Coming up with the approach	No
Proof development	Yes
Writing and exposition	No
Checking for bugs	Yes
Other supporting tasks	Yes

Details. We acknowledge the use of various Generative AI tools (including Anthropic Fable 5, Fable 5.1, Opus 5.5, GPT 5.6, and GPT 6) for brainstorming, literature research, solving subproblems, and composing parts of proofs; all generated contents are reviewed and edited by the authors for clarity. Intellectual contributions of Generative AI tools are summarized below.

- The authors had a (wrong) proof sketch of a super-linear wire lower bound for parity; the idea leads us to the results in Section 5. After that, the upper bound Theorem 4.1 (disproving our super-linear wire lower bound) was observed by Fable 5 when authors asked it to formalize the (wrong) proof sketch. The proof was edited extensively by the authors for clarity.
- The lower bound in Section 4.2 is first observed by GPT 5.6 Sol in interaction with authors. The current proof (with an explicit comparison to [PS94, SRK94]) is written by the authors.
- Theorem 1.8 was observed in interaction with GPT 6 Astra. The authors constructed a function separating $\mathsf{Sigmoid} \circ \mathsf{Sigmoid}$ with small weights and $\mathsf{MAJ} \circ \mathsf{MAJ}$, and asked GPT 6 Astra to find more functions separating these classes. The authors observed that all functions are from $\mathsf{THR} \circ \mathsf{MAJ}$, and after asked by the authors, GPT 6 Astra proved the equivalence result. Theorem 1.9 was further proved by GPT 6 Astra as requested to prove similar results for $\mathsf{THR} \circ \mathsf{THR}$ using the same approach. The current proof is simplified and rewritten by the authors with the help of Claude Opus 5.5 and Fable 5.1.

Most of the queries to LLMs are given in the context of an earlier draft of the paper, which we do not attempt to describe in detail.

Other parts of the paper (e.g. Section 5 and Section 6.1) are not significantly affected by Generative AI tools. Generative AI tools (ranging from Claude Opus 4.7 to Fable 5.1, as well as GPT 5.5 to GPT 6) are also used in routine tasks (calculating the parameters with a given proof skeleton, checking bugs, typesetting, etc) that do not constitute substantial conceptual contributions.

All AI-generated content was independently verified and edited by the authors, who take full responsibility for the correctness of all claims and proofs in this paper.

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il