4

A COMPARISON OF THE COMPUTATIONAL POWER OF SIGMOID AND BOOLEAN THRESHOLD CIRCUITS

W. Maass, G. Schnitger*, and E. D. Sontag**

Institute for Theoretical Computer Science,
Technische Universitaet Graz, Austria. Email: maass@igi.tu-graz.ac.at

** Fachbereich Mathematik/Informatik,*
Universität Paderborn, Germany. Email: georgs@uni-paderborn.de

*** Dept. of Mathematics, Rutgers University, New Brunswick, NJ 08903.*
Email: sontag@control.rutgers.edu

1 INTRODUCTION

Research on neural networks has led to the investigation of massively parallel computational models that consist of analog computational elements. Usually these analog computational elements are assumed to be smooth threshold gates, i.e.γ-gates for some nondecreasing differentiable function $\gamma : \mathbb{R} \rightarrow \mathbb{R}$. A γ-gate with weights $w_1, \ldots, w_m \in \mathbb{R}$ and threshold $t \in \mathbb{R}$ is defined to be a gate that computes the function $(x_1, \ldots, x_m) \mapsto \gamma(\sum_{i=1}^{m} w_i x_i - t)$ from $\mathbb{R}^m$ into $\mathbb{R}$. A γ-circuit is defined as a directed acyclic circuit that consists of γ-gates. The most frequently considered special case of a smooth threshold circuit is the sigmoid threshold circuit, which is a σ-circuit for $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\sigma(x) = \dfrac{1}{1 + \exp(-x)}$.

Smooth threshold circuits (γ-circuits for "smooth" functions γ) have become the standard model for the investigation of learning on multi-layer artificial neural nets ([K], [HKP], [RM], [SS1], [SS2], [WK]). In fact, the most common learning algorithm for multi-layer neural nets, the Backwards-Propagation algorithm, can only be implemented on γ-circuits for differentiable functions γ.

V. Roychowdhury et al. (eds.), *Theoretical Advances in Neural Computation and Learning*

Another motivation for the investigation of smooth threshold circuits is the desire to explore simple models for the (very complicated) information processing mechanisms in neural systems of living organisms. In a first approximation one may view the current firing rate of a neuron as its current output ([S], [RM], [K]). The firing rates of neurons are known to change between a few and several hundred firings per second. Hence a smooth threshold gate provides a somewhat better computational model for a neuron than a digital element that has just two different output signals.

In this paper we examine the power of smooth threshold circuits for computing Boolean functions. In particular, we compare their power with that of boolean threshold circuits (i.e. s-circuits for the "heaviside" function s, with $s(x) = 1$ if $x \geq 0$ and $s(x) = 0$ if $x < 0$). In the literature one often refers to such "boolean threshold circuits" as "linear threshold circuits", or simply as "threshold circuits".

The most surprising result of this paper is the existence of a boolean function F_n, that can be computed by a large class of γ-circuits (containing σ-circuits) with small weights in depth 2 and size 5 (Theorem 4.1), but which cannot be computed with any weight size by constant size boolean threshold circuits of depth 2 (Theorem 4.2). A witness for this difference in computational power is the boolean function F_n with

$$F_n(\vec{x}, \vec{y}) := Majority(\vec{x}) \oplus Majority(\vec{y}),$$

where $\vec{x}$ and $\vec{y}$ are n-bit vectors.

The proof of this lower bound result for boolean threshold circuits (Theorem 4.2) is of independent interest. First, this proof demonstrates that the restriction method is not only useful in order to prove lower bounds for AC^0-circuits, but also for threshold circuits. Secondly, this proof exploits some previously unused potential in a standard tool for the analysis of threshold circuits: the ϵ-Discriminator Lemma of $[HMPST]$. It is essential for our proof that the ϵ-Discriminator Lemma holds not just for the uniform distribution over the input space (as it is stated in [HMPST]), but for any distribution. Hence we have the freedom to construct such a distribution in a malicious manner, where we exploit specific "weak points" of the considered threshold circuit. This extra power of the (generalized) ϵ-Discriminator Lemma is crucial: in Remark 4.2 we show that its conventional version is insufficient for the proof of Theorem 4.2.

Subsequent to the preliminary version [MSS] of this paper, DasGupta and Schnitger [DS] have shown that the boolean function SQ_n with

$$SQ_n(x_1, \ldots x_n, y_1, \ldots, y_{n^2}) = ((\sum_{i=1}^{n} x_i)^2 \geq \sum_{i=1}^{n^2} y_i)$$

leads to a depth-independent separation of boolean threshold circuits and γ-circuits. In particular, SQ_n is computable by a large class of γ-circuits in constant size, whereas **any** boolean threshold circuit requires size $\Omega(\log n)$. However, the gate function γ has to satisfy more stringent differentiability requirements than are necessary for the constant-size computation of F_n. For the case of analog input, [DS] also provides a comparison of the approximation power of various γ-circuits.

In order to compute a boolean function on an analog computational device one has to adopt a suitable output convention (similar to the conventions that are used to carry out digital computations on real-world computers, which consists of non-digital computational elements such as transistors).

Definition 4.1 *A γ-circuit C computes a boolean function $F : \{0,1\}^n \to \{0,1\}$ with separation ε if there is some $t_C \in \mathbb{R}$ such that for any input $(x_1, \ldots, x_n) \in \{0,1\}^n$ the output gate of C outputs a value which is at least $t_C + \varepsilon$ if $F(x_1, \ldots, x_n) = 1$, and at most $t_C - \varepsilon$ otherwise.*

A computation without separation at the output gate appears to be less interesting, since then an infinitesimal change in the output of any γ-gate in the circuit may invert the output bit. Hence we consider in this paper computations on γ-circuits C_n with separation at least $\dfrac{1}{p(n)}$ for some polynomial p (where n is the number of input bits of C_n). One nice feature of this convention is that, for Lipschitz bounded gate functions γ and polynomial size γ-circuits C_n of constant depth and with polynomially bounded weights, it allows a tolerance of $\frac{1}{poly(n)}$ for **all** γ-gates in C_n.

We will give in Theorem 4.3 a "separation boosting" result, which says that for any constant depth d one may demand for polynomial size γ-circuits with polynomially bounded weights just as well a separation of size $\Omega(1)$ without changing the class of boolean functions that can be computed.

An extended abstract of this paper has previously appeared in [MSS]. In this full version of the paper we have strengthened the claim of Theorem 4.3 by imposing a less stringent condition on γ.

2 SIGMOID THRESHOLD CIRCUITS FOR THE XOR OF MAJORITIES

We write (NL) for the following property of a function $\gamma : \mathbb{R} \to \mathbb{R}$,

(NL) *There is some rational number s so that:*

1. *γ is differentiable on some open interval containing s, and*
2. *$\gamma''(s)$ exists and is nonzero.*

Obviously the function σ satisfies (NL).

Observe that property (NL) is basically the requirement that the function be nonlinear; for instance, if γ'' happens to be everywhere defined, then (NL) is precisely equivalent to γ not being a linear function. The nonlinearity of γ is obviously a necessary assumption for Theorem 4.1, since otherwise a γ-circuit can only compute linear functions.

Without loss of generality, we will assume that

$$c := \frac{\gamma''(s)}{4} > 0$$

for some point s as in the definition. If this value where to be negative, we simply replace γ by $-\gamma$ in what follows.

Lemma 4.1 *Assume that γ satisfies (NL). Define the function*

$$\theta(x) := \gamma(x+s) + \gamma(-x+s) \; .$$

Then, the function θ is even, and there exists some $\epsilon > 0$ so that the following property holds:

$$\theta(a+h) - \theta(a) \;\geq\; ch^2 \;\text{ for all }\; a, h \in [0, \epsilon] \; .$$

Proof: Note that $\theta(-x) = \theta(x)$ directly from the definition, so θ is even. Moreover, θ is differentiable on some open interval containing $x = 0$, because γ is differentiable in a neighborhood of s, and evenness implies that $\theta'(0) = 0$. Observe also that $\theta''(0)$ exists, and in fact

$$\theta''(0) = 2\gamma''(s) = 8c > 0 .$$

By definition of $\theta''(0)$ (just write $\theta'(l) = \theta'(0)+\theta''(0)l+r(l)$ with $\lim_{l\to 0} \frac{r(l)}{l} = 0$), there is some $\epsilon > 0$ so that

$$\theta'(l) \geq \frac{\theta''(0)l}{2} = 4cl \tag{4.1}$$

for all each $l \in [0, 2\epsilon]$. Because $\theta'(l) \geq 4cl > 0$ for $l > 0$, it follows that θ is strictly increasing on $[0, 2\epsilon]$. We are only left to prove that this ϵ is so that the last property holds.

Pick any $a, h \in [0, \epsilon]$. Assume that $h \neq 0$, as otherwise there is nothing to prove. As a and $a+h$ are both in the interval $[0, 2\epsilon]$, and θ is strictly increasing there, it follows that

$$\theta(a+h) - \theta(a) > \theta(a+h) - \theta(a+\frac{h}{2})$$

and by the Mean Value Theorem this last expression equals $\theta'(l)\frac{h}{2}$ for some $l \in (a+\frac{h}{2}, a+h)$.

Since $l < a+h \leq 2\epsilon$, we may apply inequality (4.1) to obtain

$$\theta(a+h) - \theta(a) > 2clh .$$

The result now follows from the fact that $l > a + \frac{h}{2} \geq \frac{h}{2}$. □

Theorem 4.1 *Assume that $\gamma : \mathbb{R} \to \mathbb{R}$ satisfies (NL). Then there exists for every $n \in \mathbb{N}$, a γ-circuit C_n of depth 2 with 5 gates (and rational weights and thresholds of size O(1)) that computes F_n with separation $\Omega(1/n^2)$.*

Proof: With θ and ϵ as in Lemma 4.1 one has $\theta(a) > \theta(b) \Leftrightarrow |a| > |b|$ for any $a, b \in [-\epsilon, +\epsilon]$. Hence any two nonzero reals $u, v \in [-\epsilon/2, +\epsilon/2]$ have different sign if and only if $\theta(u-v) - \theta(u+v) > 0$. Let $x_1, \ldots, x_n, y_1, \ldots, y_n \in \{0, 1\}$ be arbitrary and set

$$u := \frac{\epsilon}{2}(\frac{4(x_1 + \ldots + x_n) - 2n + 1}{4n}),$$

$$v := \frac{\epsilon}{2}\left(\frac{4(y_1 + \ldots + y_n) - 2n + 1}{4n}\right).$$

Then we obtain

$$\theta(u - v) - \theta(u + v) > 0 \Leftrightarrow F_n(\vec{x}, \vec{y}) = 1.$$

Furthermore Lemma 4.1 implies that

$$|\theta(u - v) - \theta(u + v)| \geq 4c \cdot min\{u^2, v^2\} = \Omega(1/n^2).$$

Hence, we can achieve separation $\Omega(1/n^2)$ by using a γ-gate on level two of circuit C_n that checks whether $\theta(u - v) - \theta(u + v) > 0$. Such a γ-gate exists: Since $\gamma''(s) \neq 0$, there is some t with $\gamma'(t) \neq 0$. Now transform $\theta(u-v)-\theta(u+v)$ into a suitable neighborhood of t and choose a suitable rational approximation of $\theta(t)$ as threshold. □

Corollary 4.1 *Assume that $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ satisfies (NL) and γ is monotone. Then there exists for every $n \in \mathbb{N}$ a γ-circuit C_n of depth 2 and size 5 (with rational weights and thresholds of size polynomial in n) that computes F_n with separation $\Omega(1)$.*

Proof: Multiply the weights of the γ-gate on level two of the circuit C_n with n^2 and transform the threshold accordingly. In this way we can ensure that the weighted sum computed at the top gate has distance $\Omega(1)$ from its threshold. □

Remark 4.1 For computations with real (rather than Boolean) inputs, there has been some work dealing with the differences in capabilities between sigmoidal and threshold devices; in particular [So] studies questions of interpolation and classification related to learnability (VC dimension). □

3 BOOLEAN THRESHOLD GATES ARE LESS POWERFUL

Theorem 4.2 *No family $(C_n \mid n \in \mathbb{N})$ of constant size boolean threshold circuits of depth 2 (with unrestricted weights and thresholds) can compute the function F_n.*

Assume, by way of contradiction, that there exist such circuits C_n, each with at most k' gates on level one. We can demand that all weights are integers and that the level 2 gate has weights of absolute value at most $2^{O(k' \log k')}$ ([Mu],[MT]). Thus we can assume, after appropriate duplication of level one gates, that the gate on level 2 has only weights from $\{-1, 1\}$. Let k be an upper bound on the resulting number of gates.

In the next section we use the restriction method to eliminate those gates on level one of C_n whose weights for the x_i (y_i) have drastically different sizes. It turns out that we cannot achieve this goal for all gates. For example, if all the weights w_i (for the x_i) are much larger than the weights u_i (for the y_i), then we can only limit the variance of the weights w_i (see condition b. in Definition 4.2). Nevertheless, the restriction method allows us to "regularize" *all* bottom gates of C_n (see Lemma 4.3). In section 3.2 we show that the resulting regularized gates behave predictably for certain distributions (see Lemma 4.4). The argument for the proof of Theorem 4.2 concludes in section 3.3 with a non-standard application of the ε-Discriminator Lemma .

3.1 The Restriction Method

Our goal will be to fix certain inputs such that all bottom gates of C_n will have a normal form as described in the following definition.

Definition 4.2 *Let G be a boolean threshold gate (with $2m$ inputs: $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$) that outputs 1 if and only if $\sum_{i=1}^{m} w_i x_i + \sum_{i=1}^{m} u_i y_i \geq t$. Assume that the numbering is such that*

$$|w_1| \leq \ldots \leq |w_m| \text{ and } |u_1| \leq \ldots \leq |u_m|.$$

We say that G is l-regular if and only if all w_i have the same sign (negative, zero, or positive) and all u_i have the same sign. Additionally, one of the following conditions has to hold,

a. *G is constant.*

b. $\forall i\ (|w_i| \geq m^{1/8}|u_i|)$ *and* $|w_m| \leq 60|w_1|$.

c. $\forall i\ (|u_i| \geq m^{1/8}|w_i|)$ *and* $|u_m| \leq 60|u_1|$.

d. $|w_m| \leq 30(1+l)|w_1|$ *and* $|u_m| \leq 30(1+l)|u_1|$.

First we will transform a single threshold gate to a regular gate.

Lemma 4.2 *Let G be an arbitrary threshold gate that outputs 1 if and only if*

$$\sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} u_i y_i \geq t.$$

Then there are sets $M_x \subseteq \{1, \ldots, n\}$ and $M_y \subseteq \{1, \ldots, n\}$ of size $\frac{n}{60}$ each and an assignment $\mathcal{A} : \{x_i : i \notin M_x\} \cup \{y_i : i \notin M_y\} \to \{0, 1\}$ such that

a. when values are assigned according to $\mathcal{A}$, $F_{n/60}$ will be obtained as the corresponding subfunction of F_n, and

b. G, when restricted to the remaining free variables, is $n^{1/8}$-regular.

Proof: First we determine a set $M'_x \subseteq \{1, \ldots, n\}$ of size $n/3$ such that all w_i (with $i \in M'_x$) are either all positive, all negative or all zero. A set $M'_y \subseteq \{1, \ldots, n\}$ of size $n/3$ is chosen analogously to enforce the same property for the coefficients u_i (with $i \in M'_y$).

Set $m = n/3$. After possibly renumbering the indices, we can assume that $M'_x = M'_y = \{1, \ldots, m\}$. We can also assume that $|w_1| \leq \ldots \leq |w_m|$ as well as $|u_1| \leq \ldots \leq |u_m|$. We define

$$\mathcal{R} := \{1, \ldots, \frac{m}{4}\}, \ \ \mathcal{S} := \{\frac{m}{4} + 1, \ldots, \frac{3m}{4}\} \text{ and } \mathcal{T} := \{\frac{3m}{4} + 1, \ldots, m\}.$$

By assigning 1's to the x_i's with $i \in \mathcal{R}$ and 0's to the x_i's with $i \in \mathcal{T}$ or vice versa, and by assigning 1's to the y_i's with $i \in \mathcal{R}$ and 0's to the y_i's with $i \in \mathcal{T}$ or vice versa, we obtain four partial assignments.

Let us now interpret G as a threshold gate of the remaining variables x_i ($i \in \mathcal{S}$) and y_i ($i \in \mathcal{S}$). By choosing one of the four assignments, we can "move" the threshold of the resulting gate over a distance d with

$$d = \sum_{i \in \mathcal{T}} |w_i| - \sum_{i \in \mathcal{R}} |w_i| + \sum_{i \in \mathcal{T}} |u_i| - \sum_{i \in \mathcal{R}} |u_i|.$$

If for none of these four partial assignments the threshold gate G gives constant output, we have

$$d \leq \sum_{i \in \mathcal{S}} |w_i| + \sum_{i \in \mathcal{S}} |u_i|.$$

This implies that

$$\sum_{i\in\mathcal{T}}(|w_i|+|u_i|) \le \sum_{i\in\mathcal{R}\cup\mathcal{S}}(|w_i|+|u_i|). \tag{4.2}$$

Set $a = \sum_{i\in\mathcal{R}\cup\mathcal{S}}(|w_i|+|u_i|)/(3m/4)$ and $b = \sum_{i\in\mathcal{T}}(|w_i|+|u_i|)/(m/4)$. Then (4.2) implies for these "averages" of $|w_i|+|u_i|$ over $\mathcal{R}\cup\mathcal{S}$ respectively $\mathcal{T}$ that $b \le 3a$.

We subdivide the set S by introducing the sets

$$\mathcal{P} = \{\frac{3m}{4} - \frac{2m}{10} + 1, \ldots, \frac{3m}{4} - \frac{m}{10}\} \text{ and } \mathcal{Q} = \{\frac{3m}{4} - \frac{m}{10} + 1, \ldots, \frac{3m}{4}\}.$$

Since $|w_i|+|u_i|$ is a non-decreasing function of i we have for all $i \in \mathcal{R}\cup\mathcal{S}$ (and in particular for all $i \in \mathcal{P}\cup\mathcal{Q}$)

$$|w_i|+|u_i| \le b \le 3a. \tag{4.3}$$

Furthermore, we have for all $i \in \mathcal{P}$

$$|w_i|+|u_i| \ge a/10, \tag{4.4}$$

since otherwise $|w_i|+|u_i| < a/10$ for all $i \in (\mathcal{R}\cup\mathcal{S}) - (\mathcal{P}\cup\mathcal{Q})$, and we would get

$$\begin{aligned}
\sum_{i\in\mathcal{R}\cup\mathcal{S}}(|w_i|+|u_i|) &= \sum_{i\in(\mathcal{R}\cup\mathcal{S})-(\mathcal{P}\cup\mathcal{Q})}(|w_i|+|u_i|) + \sum_{i\in\mathcal{P}\cup\mathcal{Q}}(|w_i|+|u_i|)\\
&\le (\frac{3}{4} - \frac{2}{10})m \cdot \frac{a}{10} + 3a \cdot \frac{2m}{10}\\
&= m \cdot a(\frac{3}{4}\cdot\frac{1}{10} + \frac{2}{10}(3 - \frac{1}{10}))\\
&< \frac{3m \cdot a}{4},
\end{aligned}$$

which is a contradiction to the definition of a.

(4.3) and (4.4) jointly imply that

$$\max_{i\in\mathcal{P}\cup\mathcal{Q}}(|w_i|+|u_i|) \le 30 \min_{i\in\mathcal{P}\cup\mathcal{Q}}(|w_i|+|u_i|). \tag{4.5}$$

Case 1: $\forall\, i \in \mathcal{P}(|w_i| \ge m^{1/8}|u_i| \;\vee\; |u_i| \ge m^{1/8}|w_i|)$.

We can find a subset $\mathcal{P}' \subseteq \mathcal{P}$ of size $m/20$ such that

$$\forall i \in \mathcal{P}' \; (|w_i| \geq m^{1/8}|u_i|) \; \textit{or} \; \forall i \in \mathcal{P}' \; (|u_i| \geq m^{1/8}|w_i|).$$

In the former case, (4.5) implies that

$$\begin{aligned}\max_{i \in \mathcal{P}'} |w_i| &\leq 30 \min_{i \in \mathcal{P}'}(|w_i| + |u_i|) \\ &\leq 30(1 + m^{-1/8}) \min_{i \in \mathcal{P}'} |w_i| \\ &\leq 60 \min_{i \in \mathcal{P}'} |w_i|.\end{aligned}$$

Set $M_x = M_y = \mathcal{P}'$ and fix the remaining variables such that exactly half of the x_i's and half of the y_i's are 0.

Analogously, in the latter case we obtain $\max_{i \in \mathcal{P}'} |u_i| \leq 60 \min_{i \in \mathcal{P}'} |u_i|$. M_x and M_y are obtained as above.

Case 2: Otherwise.

Then $\exists i_0 \in \mathcal{P} \; (|w_{i_0}| < m^{1/8}|u_{i_0}| \wedge |u_{i_0}| < m^{1/8}|w_{i_0}|)$. We have for all $i \in \mathcal{Q}$:

$$|w_i| + |u_i| \leq 30(|w_{i_0}| + |u_{i_0}|) \leq \min\{30|w_{i_0}|(1 + m^{1/8}), 30|u_{i_0}|(1 + m^{1/8})\}.$$

Thus we have $\max_{i \in \mathcal{Q}} |w_i| \leq 30(1+m^{1/8}) \min_{i \in \mathcal{Q}} |w_i|$ and $\max_{i \in \mathcal{Q}} |u_i| \leq 30(1+m^{1/8}) \min_{i \in \mathcal{Q}} |u_i|$.

Choose M_x to be an arbitrary subsets of $\mathcal{Q}$ of size $\frac{n}{60}$, set $M_y = M_x$ and fix the remaining variables in the same fashion as before. □

If we perform the "regularization process" for all bottom gates of C_n, then we obtain the following result.

Lemma 4.3 *There are sets $M_x, M_y \subseteq \{1, \ldots, n\}$ of size $m = \frac{n}{60^k}$ and there is an assignment $\mathcal{A} : \{x_i : i \notin M_x\} \cup \{y_i : i \notin M_y\} \rightarrow \{0, 1\}$ such that*

a. when values are assigned according to $\mathcal{A}$, F_m will be obtained as the corresponding subfunction of F_n, and

b. all level one gates of C_n, when restricted to the free variables, are $n^{1/8}$-regular.

Proof: Apply Lemma 4.2 successively to each of the k level one gates of C_n. Let M_x be the set of indices of those variables x_i which did not receive a value during the processing of all gates by Lemma 4.2. M_y is defined analogously.

$\mathcal{A}$ is the union of all partial assignments that have been made in this process. □

We write D_n for the circuit that results from C_n by the restriction of Lemma 4.3. Observe that D_n computes the function F_m (for $m = \frac{n}{60^k}$).

3.2 The Likely Behavior of a Threshold Gate

In this section we will exploit the result of our regularization process. In particular, in Lemma 4.4, we will show that, for the input distribution defined below, a weighted sum with small variance in weight sizes "almost" behaves as if all the weights were identical.

For the integer s, $1 \leq s \leq m$, set $U(s) = \{\vec{x} \in \{0,1\}^m : \sum_{i=1}^{m} x_i = s\}$. $X(s)$ is the random variable which assigns to each $\vec{x} \in U(s)$ the value $\sum_{i=1}^{m} w_i x_i$; all elements of $U(s)$ are equally likely. Obviously $E(X(s)) = \frac{s}{m} \sum_{i=1}^{m} w_i$.

In the following, we will assume that the w_i's are either all positive or all negative.

Proposition 4.1 *Set* $W = \sum_{i=1}^{m} w_i^2$ *and* $g = max\{\frac{|w_i|}{|w_j|} : 1 \leq i, j \leq m\}$. *Then*

$$W \leq \frac{m}{s^2} g^2 \cdot E(X(s))^2.$$

Proof: Set $MIN = min\{|w_i| : 1 \leq i \leq m\}$. We get

$$W \leq \frac{1}{m}(m^2 \cdot g^2 \cdot MIN^2). \tag{4.6}$$

Also, $E(X(s))^2 = (\frac{s}{m} \cdot \sum_{i=1}^{m} w_i)^2 \geq (\frac{s}{m} m \cdot MIN)^2$. Thus

$$m^2 \cdot MIN^2 \leq (\frac{m}{s})^2 \cdot E(X(s))^2. \tag{4.7}$$

If we replace $m^2 \cdot MIN^2$ in (4.6) according to (4.7), we get

$$W \leq \frac{m}{s^2} \cdot g^2 \cdot E(X(s))^2. \qquad \Box$$

Proposition 4.2 $Var(X(s)) \leq \frac{s}{m} W$.

Proof: We have $Var(X(s)) = E(X(s)^2) - E(X(s))^2$. Also,

$$\begin{aligned}
E(X(s)^2) &= \frac{1}{\binom{m}{s}} \sum_{\vec{x} \in U(s)} (\sum_{i=1}^{m} w_i x_i)^2 \\
&= \frac{1}{\binom{m}{s}} (\sum_{\vec{x} \in U(s)} \sum_{i=1}^{m} {w_i}^2 x_i + 2 \sum_{\vec{x} \in U(s)} \sum_{1 \leq i < j \leq m} w_i w_j x_i x_j) \\
&= \frac{1}{\binom{m}{s}} (\sum_{i=1}^{m} {w_i}^2 \sum_{\vec{x} \in U(s)} x_i + 2 \sum_{1 \leq i < j \leq m} w_i w_j \sum_{\vec{x} \in U(s)} x_i x_j) \\
&= \frac{s}{m} \cdot \sum_{i=1}^{m} {w_i}^2 + \frac{2s(s-1)}{m(m-1)} \sum_{1 \leq i < j \leq m} w_i w_j.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
E(X(s))^2 &= (\frac{1}{\binom{m}{s}} \sum_{\vec{x} \in U(s)} \sum_{i=1}^{m} w_i x_i)^2 \\
&= (\frac{1}{\binom{m}{s}} \sum_{i=1}^{m} w_i \sum_{\vec{x} \in U(s)} x_i)^2 \\
&= (\frac{s}{m} \sum_{i=1}^{m} w_i)^2 \\
&= \frac{s^2}{m^2} \sum_{i=1}^{m} {w_i}^2 + \frac{2s^2}{m^2} \sum_{1 \leq i < j \leq m} w_i w_j.
\end{aligned}$$

In summary, we obtain

$$Var(X(s)) \leq (\frac{s}{m} - \frac{s^2}{m^2}) \sum_{i=1}^{m} {w_i}^2 \leq \frac{s}{m} W. \qquad \square$$

Lemma 4.4 *If* $\max\{\frac{|w_i|}{|w_j|} : 1 \leq i,j \leq m\} = O(m^{1/8})$, *then*

$$Pr(|X(s) - E(X(s))| \geq \frac{|E(X(s))|}{m^{1/4}}) = O(\frac{m^{3/4}}{s}).$$

Proof: By Chebyshev's inequality , we get for any t $(t > 0)$

$$Pr(|X(s) - E(X(s))| \geq t) \leq \frac{Var(X(s)}{t^2}.$$

Thus, for $t = \frac{|E(X(s))|}{m^{1/4}})$, we obtain

$$Pr(|X(s) - E(X(s))| \geq \frac{|E(X(s))|}{m^{1/4}}) \leq \frac{Var(X(s)) \cdot m^{1/2}}{E(X(s))^2}.$$

Proposition 4.2 implies

$$\frac{Var(X(s)) \cdot m^{1/2}}{E(X(s))^2} \leq \frac{s \cdot W \cdot m^{1/2}}{m \cdot E(X(s))^2},$$

and with Proposition 4.1

$$\frac{s \cdot W \cdot m^{1/2}}{m \cdot E(X(s))^2} \leq \frac{m}{s} g^2 m^{-1/2} = O(\frac{m}{s} m^{-1/4}). \qquad \square$$

3.3 A Non-standard Application of the Discriminator Lemma

Let G be some boolean threshold gate with weights $w_1, \ldots, w_m, u_1, \ldots, u_m$ and threshold t. Set

$$a := \frac{\sum_{i=1}^{m} w_i}{m} \text{ and } b := \frac{\sum_{i=1}^{m} u_i}{m}.$$

With G we can thus associate the two-dimensional threshold function $ax + by \geq t$. Similarly, with F_m we associate the two-dimensional function $F : \{0, \ldots, m\}^2 \to \{0, 1\}$, where $F(x, y) = 1$ if and only if

$$(x \geq \frac{m}{2} \wedge y < \frac{m}{2}) \vee (x < \frac{m}{2} \wedge y \geq \frac{m}{2}).$$

Let L be the line $ax + by = t$ in $\mathbb{R}^2$ (where t is the threshold of G). Let x' (y') be the x-coordinate (y-coordinate) of the intersection of L and $y = \frac{m}{2}$ ($x = \frac{m}{2}$). Set $x' = \infty$ ($y' = \infty$) if the line L is horizontal (vertical). We define

$$D(G) := min\{|x' - \frac{m}{2}|, |y' - \frac{m}{2}|\}.$$

Proposition 4.3 *Let r be an integer with $0 \leq r \leq \frac{m}{2}$ and let U_r be the uniform distribution over $V_r = \{\frac{m}{2} - r, \ldots, \frac{m}{2} + r\}$. Then*

$$Pr_{U_r \times U_r}[(x, y) \in V_r^2 \mid ax + by \geq t \ \wedge \ F(x, y) = 1\} \leq \frac{1}{2} + \frac{D(G) + 1}{2r + 1}.$$

Proof: Let X be the area enclosed by the two lines $ax + by = t$ and $ax + by = (a + b) \cdot \frac{m}{2}$. (The latter is the line through $(\frac{m}{2}, \frac{m}{2})$.) Intersect X with the set V_r^2 and call the intersection X_r.

Let us assume that $D(G) = |x' - \frac{m}{2}|$. Then X_r will contain at most $D(G) + 1$ points per row of V_r^2. Thus $|X_r| \leq (2r + 1) \cdot (D(G) + 1)$.

On the other hand, the halfspace $ax + by \geq (a + b) \cdot \frac{m}{2}$ contains exactly one half of all the elements of the set $\{(x, y) \in V_r^2 \mid F(x, y) = 1\}$. □

Let us consider the case that all weights w_i are identical and all weights u_i are identical. If $D(G)$ is "small", then G will not show any significant advantage in predicting F for a subcollection of the $\frac{m}{2} + 1$ distributions U_r mentioned in Proposition 4.3. If on the other hand $D(G)$ is large (say proportional to m), then we can trivialize G by choosing a distribution with a small value for r.

Our goal is to carry out a similar argument for arbitrary gates G. Consequently we introduce a collection Q_r of distributions over $\{0, 1\}^m$ with

$$Q_r(\vec{x}) = \begin{cases} 0 & , \text{ if } |\sum_{i=1}^m x_i - \frac{m}{2}| > r \\ \dfrac{1}{(2r + 1) \cdot \binom{m}{\sum_{i=1}^m x_i}} & , \text{ otherwise.} \end{cases}$$

Note that the probability of a string only depends on its number of ones. The appropriate value for the parameter r will be determined later.

Finally we define for the considered threshold gate G with input variables $x_1, \ldots, x_m, y_1, \ldots, y_m$,

$$\begin{aligned} ADV_r(G) \quad := \quad & Pr_{Q_r \times Q_r}[G(\vec{x}, \vec{y}) = 1 | F_m(\vec{x}, \vec{y}) = 1] \\ & - Pr_{Q_r \times Q_r}[G(\vec{x}, \vec{y}) = 1 | F_m(\vec{x}, \vec{y}) = 0]. \end{aligned}$$

Lemma 4.5 *Set* $m = \frac{n}{60^k}$. *Assume that the boolean threshold gate* G *with input variables* $x_1, \ldots, x_m, y_1, \ldots y_m$ *is* $n^{1/8}$*-regular, and that* n *is sufficiently large. Furthermore assume that the natural number* $r \in [m^{31/32}, \frac{m}{4}]$ *satisfies*

$$D(G) \leq r/(64k) \quad or \quad D(G) \geq 4r\ .$$

Then

$$|ADV_r(G)| \leq \frac{1}{2k}.$$

Proof:

Case 1: $D(G) \leq \frac{r}{64 \cdot k}$.

We know that G is $n^{1/8}$-regular. We proceed by examining the three different cases (see Definition 4.2.).

Case 1.1: $\forall i\ (|w_i| \geq m^{1/8}|u_i|)$ and $|w_m| \leq 60|w_1|$.

This implies that $|a| \geq m^{1/8} \cdot |b|$. Hence the line L is very "steep". We have in this case,

$$\begin{aligned} & (\max\{x \in [0, m] : \exists y \in [0, m]\ ((x, y) \in L)\} \\ & - \min\{x \in [0, m] : \exists y \in [0, m]\ ((x, y) \in L)\}) \quad \leq \quad m^{7/8}. \end{aligned}$$

Thus, the set $\{x \in [0, m] : \exists x', y' \in [0, m]\ (|x - x'| \leq m^{3/4} \wedge (x', y') \in L\}$ is contained in an interval of length $m^{7/8} + 2m^{3/4} + 1 \leq 3 \cdot m^{7/8}$. This implies that

$$|\{(x, y) \in \{\frac{m}{2} - r, \ldots, \frac{m}{2} + r\}^2 : P(x, y)\}| \leq 3m^{7/8} \cdot m = 3m^{15/8}, \tag{4.8}$$

where $P(x, y)$ is equivalent to

$$(ax + by < t) \ \wedge \ \exists(x', y') \in [0, m]^2(\ (ax' + by' \geq t) \wedge (|x - x'| \leq m^{3/4}) \ ).$$

As a first step towards estimating $ADV_r(G)$ we consider the set

$$S = \{(\vec{x}, \vec{y}) \in U : G(\vec{x}, \vec{y}) = 1 \ \wedge \ F_m(\vec{x}, \vec{y}) = 1\},$$

where

$$U := \{(\vec{x}, \vec{y}) \in \{0,1\}^{2m} : \frac{m}{2} - r \leq \sum_{i=1}^{m} x_i, \sum_{i=1}^{m} y_i \leq \frac{m}{2} + r\}.$$

One shows that S is contained in the following two sets,

$$S_1 = \{(\vec{x}, \vec{y}) \in U : |\sum_{i=1}^{m} w_i x_i - (\sum_{i=1}^{m} x_i \cdot \sum_{i=1}^{m} w_i)/m| \geq (\sum_{i=1}^{m} x_i \cdot \sum_{i=1}^{m} w_i)/m^{5/4}\}$$

and $S_2 = \{(\vec{x}, \vec{y}) \in U : F_m(\vec{x}, \vec{y}) = 1 \ \wedge \ Q(\vec{x}, \vec{y})\}$, where $Q(\vec{x}, \vec{y})$ is equivalent to

$$\exists \vec{x'}, \vec{y'} \in \{0,1\}^m \ (|\sum_{i=1}^{m} x'_i - -\sum_{i=1}^{m} x_i| \leq m^{3/4} \ \wedge \ a \cdot \sum_{i=1}^{m} x'_i + b \cdot \sum_{i=1}^{m} y'_i \ \geq \ t).$$

Intuitively, the set S_1 consists of all those inputs (that are relevant for Q_r) on which our approximation of G by $ax + by \geq t$ fails. We will show later that this set has small probability. S_2 on the other hand is the collection of all relevant inputs on which the approximation (in a quite liberal sense) succeeds.

Let us verify the inclusion $S \subseteq S_1 \cup S_2$. Fix any $(\vec{x}, \vec{y}) \in S - S_1$. We then have $\sum_{i=1}^{m} w_i x_i + u_i y_i \geq t$ and

$$\begin{aligned} |\sum_{i=1}^{m} w_i x_i - a \cdot \sum_{i=1}^{m} x_i| \ &\leq \ (\sum_{i=1}^{m} x_i \cdot |a|)/m^{1/4} \\ &\leq \ m^{3/4} \cdot |a|. \end{aligned} \tag{4.9}$$

We need to find vectors $\vec{x'}, \vec{y'}$ according to the definition of set S_2.

If $a \geq 0$, we pick some $\vec{x'}$ such that $\sum_{i=1}^{m} x'_i = \sum_{i=1}^{m} x_i + m^{3/4}$. This is possible, since $U \subseteq \bigcup_{i=m/4}^{3m/4} \{0,1\}^i$. We then have with (4.9) $a \cdot \sum_{i=1}^{m} x'_i \geq \sum_{i=1}^{m} w_i x_i$.

If $a < 0$, we pick some $\vec{x'}$ such that $\sum_{i=1}^{m} x'_i = \sum_{i=1}^{m} x_i - m^{3/4}$. We then have $a \cdot \sum_{i=1}^{m} x'_i = a \cdot \sum_{i=1}^{m} x_i - a \cdot m^{3/4} = a \cdot \sum_{i=1}^{m} x_i + |a| \cdot m^{3/4} \geq \sum_{i=1}^{m} w_i x_i$.

Furthermore, we pick some vector $\vec{y'}$ with $b \cdot \sum_{i=1}^{m} y'_i \geq \sum_{i=1}^{m} u_i y_i$ according to the following procedure: if all components of u_i are positive or all components are zero, then set $\vec{y'} = (1, \ldots, 1)$. Otherwise all components are negative and we set $\vec{y'} = (0, \ldots, 0)$.

This concludes our proof of inclusion, since property $Q(\vec{x}, \vec{y})$ holds for the pair $(\vec{x'}, \vec{y'})$. It is obvious that $Pr_{Q_r \times Q_r}[S_1 \mid F_m(\vec{x}, \vec{y}) = 1] \leq 2 \cdot Pr_{Q_r \times Q_r}[S_1]$. If we apply Lemma 4.4 for $s \in [\frac{m}{2} - r, \frac{m}{2} + r] \subseteq [\frac{m}{4}, \frac{3m}{4}]$, we obtain $Pr_{Q_r \times Q_r}[S_1] = O(m^{-1/4})$.

In order to give an upper bound on $Pr_{Q_r \times Q_r}[S_2]$ we observe that $S_2 \subseteq S_3 \cup S_4$, where

$$S_3 = \{(\vec{x}, \vec{y}) : (F_m(\vec{x}, \vec{y}) = 1) \ \wedge \ (a \cdot \sum_{i=1}^{m} x_i + b \cdot \sum_{i=1}^{m} y_i \geq t)\},$$

and

$$S_4 = \{(\vec{x}, \vec{y}) : (F_m(\vec{x}, \vec{y}) = 1) \ \wedge \ (a \cdot \sum_{i=1}^{m} x_i + b \cdot \sum_{i=1}^{m} y_i < t) \ \wedge \ R(\vec{x}, \vec{y})\}.$$

$R(\vec{x}, \vec{y})$ is equivalent to

$$\exists \vec{x'}, \vec{y'} \ (|\sum_{i=1}^{m} x_i - \sum_{i=1}^{m} x'_i| \leq m^{3/4} \ \wedge \ a \cdot \sum_{i=1}^{m} x'_i + b \cdot \sum_{i=1}^{m} y'_i \geq t).$$

It follows from Proposition 4.3 that

$$Pr_{Q_r \times Q_r}[S_3 \mid F_m(\vec{x}, \vec{y}) = 1] \leq \frac{1}{2} + \frac{D(G) + 1}{2r + 1}.$$

Also, it is obvious that

$$Pr_{Q_r \times Q_r}[S_4 \mid F_m(\vec{x}, \vec{y}) = 1] \leq 2 \cdot Pr_{Q_r \times Q_r}[S_4].$$

Furthermore, by (4.8), $Pr_{Q_r \times Q_r}[S_4] \leq \frac{3 \cdot m^{15/8}}{(2r+1)^2}$. Thus we have

$$\begin{aligned}
Pr_{Q_r \times Q_r}[S \mid F_m(\vec{x},\vec{y}) = 1] &\leq Pr_{Q_r \times Q_r}[S_1 \cup S_3 \cup S_4 \mid F_m(\vec{x},\vec{y}) = 1] \\
&\leq O(m^{-1/4}) + \frac{1}{2} + \frac{D(G)+1}{2r+1} + \frac{6m^{15/8}}{(2r+1)^2} \\
&\leq \frac{1}{2} + \frac{1}{64k} + O(m^{-1/16}).
\end{aligned}$$

We will obtain the same upper bound for the probability of

$$S' = \{(\vec{x},\vec{y}) \in U : G(\vec{x},\vec{y}) = 0 \ \wedge \ F_m(\vec{x},\vec{y}) = 1\}.$$

Thus, since

$$Pr_{Q_r \times Q_r}[S \mid F_m = 1] + Pr_{Q_r \times Q_r}[S' \mid F_m = 1] = 1,$$

we get

$$|Pr_{Q_r \times Q_r}[S \mid F_m(\vec{x},\vec{y}) = 1] - \frac{1}{2}| \leq \frac{1}{64k} + O(m^{-1/16}).$$

One shows analogously for $T = \{(\vec{x},\vec{y}) \in U : G(\vec{x},\vec{y}) = 1 \ \wedge \ F_m(\vec{x},\vec{y}) = 0\}$ that

$$|Pr_{Q_r \times Q_r}[T \mid F_m(\vec{x},\vec{y}) = 0] - \frac{1}{2}| \leq \frac{1}{64k} + O(m^{-1/16}).$$

Thus, $|ADV_r(G)| \leq \frac{1}{32k} + O(m^{-1/16})$.

Case 1.2: $\forall i \ (|u_i| \geq m^{1/8}|w_i|)$ and $|u_m| \leq 60|u_1|$.

The argument is analogous to Case 1.1.

Case 1.3: $|w_m| \leq 30(1+n^{1/8})|w_1|$ and $|u_m| \leq 30(1+n^{1/8})|u_1|$.

We first observe that the set S is contained in the union of the sets S_1' and S_2', where

$$S_1' = \{(\vec{x},\vec{y}) \in U : P'(\vec{x},\vec{y})\},$$

and $P'(\vec{x},\vec{y})$ is equivalent to

$$(|\sum_{i=1}^m w_i x_i - a \cdot \sum_{i=1}^m x_i| \geq \frac{|a| \cdot \sum_{i=1}^m x_i}{m^{1/4}}) \ \vee \ (|\sum_{i=1}^m u_i y_i - b \cdot \sum_{i=1}^m y_i| \geq \frac{|b| \cdot \sum_{i=1}^m y_i}{m^{1/4}});$$

$S_2' = \{(\vec{x},\vec{y}) \in U : F_m(\vec{x},\vec{y}) = 1 \ \wedge \ Q'(\vec{x},\vec{y})\}$, and $Q'(\vec{x},\vec{y})$ is equivalent to

$$\exists \vec{x'}, \vec{y'} \in \{0,1\}^m \ (|\sum_{i=1}^m x_i' - -\sum_{i=1}^m x_i| \leq m^{3/4} \ \wedge$$

$$|\sum_{i=1}^{m} y_i' - \sum_{i=1}^{m} y_i| \leq m^{3/4} \ \wedge \ a \cdot \sum_{i=1}^{m} x_i' + b \cdot \sum_{i=1}^{m} y_i' \ \geq \ t).$$

Lemma 4.4 implies that $Pr_{Q_r \times Q_r}[S_1' \ |F_m(\vec{x}, \vec{y}) = 1] \leq 2 \cdot Pr_{Q_r \times Q_r}[S_1'] = O(m^{-1/4})$. With an argument analogous to Case 1.1 we get $S_2' \subseteq S_3 \cup S_4'$ where

$$S_4' = \{(\vec{x}, \vec{y}) : (F_m(\vec{x}, \vec{y}) = 1) \ \wedge \ (a \cdot \sum_{i=1}^{m} x_i + b \cdot \sum_{i=1}^{m} y_i < t) \ \wedge \ R'(\vec{x}, \vec{y})\}.$$

$R'(\vec{x}, \vec{y})$ is equivalent to

$$\exists \vec{x'}, \vec{y'} \ (|\sum_{i=1}^{m} x_i - \sum_{i=1}^{m} x_i'| \leq m^{3/4} \ \wedge$$

$$|\sum_{i=1}^{m} y_i - \sum_{i=1}^{m} y_i'| \leq m^{3/4} \ \wedge \ a \cdot \sum_{i=1}^{m} x_i' + b \cdot \sum_{i=1}^{m} y_i' \geq t).$$

We have already shown that

$$Pr_{Q_r \times Q_r}[S_3 \mid F_m(\vec{x}, \vec{y}) = 1] \leq \frac{1}{2} + \frac{D(G) + 1}{2r + 1}.$$

Furthermore, it is obvious that

$$Pr_{Q_r \times Q_r}[S_4' \mid F_m(\vec{x}, \vec{y}) = 1] \leq \frac{4 \cdot m \cdot m^{3/4}}{(2r + 1)^2}.$$

The remaining argument is now analogous to Case 1.1.

Case 2: $D \geq 4r$.

The analysis is now far simpler. The probability of the set S_1 (resp. S_1') is computed as before. As for S_3 we now get

$$Pr_{Q_r \times Q_r}[S_3 \mid F_m(\vec{x}, \vec{y}) = 1] \in \{0, 1\}.$$

For S_4 we obtain

$$Pr_{Q_r \times Q_r}[S_4 \mid F_m(\vec{x}, \vec{y}) = 1] = 0.$$

The same applies to S_4'. This follows, since the set U will be entirely contained in one of the halfspaces of $\{(\vec{x}, \vec{y}) : a \cdot \sum_{i=1}^{m} x_i + b \cdot \sum_{i=1}^{m} y_i = t\}$. □

In order to prove Theorem 4.2 we observe that for sufficiently large n we can find r such that for *each* of the at most k gates G on level one of D_n:

$$D(G) \leq \frac{r}{64k} \text{ or } D(G) \geq 4r.$$

(A value for r can be found whenever k is bounded from above by the number of possible "r-intervals". This is the case, provided $k \leq c \cdot \log_k(m)$ for a suitably small constant c. This in turn is satisfied for $k \leq d \cdot \frac{\log n}{\log \log n}$ for a suitably small constant d.)

The ϵ-Discriminator Lemma of [HMPST] can be generalized to hold for any distribution over the input space. We apply it here to the distribution $Q_r \times Q_r$ over the input space $\{0,1\}^{2m}$ of the circuit D_n (which computes the function F_m).

Since the weights of the gate on level two of D_n are from $\{-1,1\}$, we get $|ADV_r(G)| \geq \frac{1}{k}$ for some gate G on level one of D_n. But this contradicts Lemma 4.5. □

Thus we get a lower bound of $\Omega(\frac{\log n}{\log \log n})$ for the size of depth 2 threshold circuits (with weights from $\{-1,1\}$ for the top gate) computing F_n. For unrestricted threshold circuits our lower bound will be $\Omega(\frac{\log \log n}{\log \log \log n})$ ([M],[MT]).

Remark 4.2 It is not possible to prove Theorem 4.2 with the customary version of the ϵ-Discriminator Lemma, where one considers the uniform distribution over the input space. Consider for example the threshold gate G defined by

$$G(x_1, \ldots x_n, y_1, \ldots y_n) = 1 \Leftrightarrow \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i \geq c \cdot \sqrt{n}.$$

For appropriate c one has $ADV(G) = \Omega(1)$ (where $ADV(G)$ is defined like $ADV_r(G)$, but with regard to the uniform distribution over $\{0,1\}^{2n}$). This happens, because a "large discrepancy" in x-sum and y-sum is more likely if we assume $\sum_{i=1}^{n} x_i \geq \frac{n}{2}$ and $\sum_{i=1}^{n} y_i \leq \frac{n}{2}$ than if we assume (say) $\sum_{i=1}^{n} x_i \geq \frac{n}{2}$ and $\sum_{i=1}^{n} y_i \geq \frac{n}{2}$. This phenomenon has been independently observed by Bultman [B]. □

Corollary 4.2 *The class of boolean functions computable by constant size boolean threshold circuits of depth 2 with integer weights of polynomial size is properly contained in the class of boolean functions computable by constant size*

σ-circuits of depth 2 with polynomial size rational weights (even with common polynomial size denominator) and separation $\frac{1}{poly}$.

The same statement holds if one considers arbitrary real weights for both types of circuits (still with separation $\frac{1}{poly}$).

Proof: It is quite easy to simulate boolean threshold circuits of size s and constant depth d by sigmoid threshold circuits of the *same* size and depth. The containment is proper as a consequence of Theorems 4.2 and 4.1. □

4 SIMULATION RESULTS AND SEPARATION BOOSTING

$TC_d^0(\gamma)$ is the class of those families $(g_n \mid n \in \mathbb{N})$ of boolean functions that are computable, with separation $\Omega(\frac{1}{poly(n)})$, by polynomial size, depth d γ-circuits whose weights are reals of absolute value at most $poly(n)$. TC_d^0 ([HMPST]) is the corresponding class of families of boolean functions computable by polynomial size, depth d boolean threshold circuits whose weights are polynomial size integers.

Theorem 4.3 *Let $\gamma : \mathbb{R} \to [0,1]$ be a nondecreasing function that is Lipschitz-bounded and converges fast to 0 (resp. 1) in the following sense:*

$$\exists\, \varepsilon > 0\ \exists\, x_0 > 0\ \forall\, x \geq x_0 \left(\gamma(-x) \leq \frac{1}{x^{\varepsilon}} \wedge 1 - \gamma(x) \leq \frac{1}{x^{\varepsilon}} \right).$$

Then the following holds.

(a) For every $d \in \mathbb{N}$, $TC_d^0 = TC_d^0(\gamma)$.

(b) The class $TC_d^0(\gamma)$ does not change if we demand separation $\Omega(1)$.

Observe, that the above class of functions also includes the standard sigmoid σ.

Proof: Assume that $(g_n|n \in \mathbb{N})$ is a family of boolean functions in $TC_d^0(\gamma)$. Thus $(g_n|n \in \mathbb{N})$ can be computed with separation $\frac{1}{p(n)}$ by some family $(C_n|n \in$

$\mathbb{N}$) of γ-circuits of depth d with the number of gates and the size of weights bounded by $q(n)$ (for some polynomials p and q). Since γ is Lipschitz-bounded, and since the depth d of C_n is a constant, there exists a polynomial $r(n)$ with the following property:

> If the gate function of each gate G in C_n is replaced by some arbitrary function $\gamma_G : \mathbb{R} \to \mathbb{R}$ (where the functions γ_G may be different for different gates G) such that
>
> $$\forall x \in \mathbb{R}(|\gamma(x) - \gamma_G(x)| \leq \frac{1}{r(n)}),$$
>
> then for each input $x_1, \ldots, x_n$ of C_n the value of the output gate of the new circuit differs from the value of the output gate of C_n by at most $\frac{1}{2p(n)}$.

In order to construct a boolean threshold circuit C_n^b that computes g_n, one replaces in C_n each internal γ-gate that outputs $\gamma(\sum_{j=1}^{m} \alpha_j y_j - \theta)$ for inputs $y_1, \ldots, y_m \in [0,1]$ (with reals $\alpha_1, \ldots, \alpha_m, \theta$ of polynomial size in n) by a weighted sum

$$S(\vec{y}) := \sum_{k=1}^{l} \frac{1}{2r(n)} H_k(y_1, \ldots, y_m)$$

of $l := 2r(n)$ boolean threshold gates $H_1, \ldots, H_l$ (which use the same weights $\alpha_1, \ldots, \alpha_m$ as G). The function S is chosen to be a step function which approximates γ such that for all $y_1, \ldots, y_m \in [0,1]$,

$$|\gamma(\sum_{j=1}^{m} \alpha_j y_j - \theta) - S(\vec{y})| \leq \frac{1}{2r(n)}.$$

In a second step, one replaces each of the boolean threshold gates H_k by a boolean threshold gate H'_k whose weights and thresholds are integers of polynomial size. We set

$$S'(\vec{y}) := \sum_{k=1}^{l} \frac{1}{2r(n)} H'_k(y_1, \ldots, y_m).$$

The threshold gates H'_k are chosen such that

$$\forall y_1, \ldots, y_m \in [0,1](|S(\vec{y}) - S'(\vec{y})| \leq \frac{1}{2r(n)}).$$

Let C'_n be the circuit that results from C_n by replacing in the described manner each *internal* γ-gate in C_n by an array of boolean threshold gates H'_k. For every input, the value of the output gates of C_n and C'_n differ by at most $\frac{1}{2p(n)}$. Hence we can replace the output gate of C'_n by a boolean threshold gate with integer weights and threshold of polynomial size such that the resulting boolean threshold circuit C^b_n computes g_n. This shows that $(g_n | n \in \mathbb{N}) \in TC^0_d$.

In order to prove the other inclusion assume that $(g_n | n \in \mathbb{N}) \in TC^0_d$ is computed by a family $(B_n | n \in \mathbb{N})$ of boolean threshold circuits of depth d, where B_n has at most $p(n)$ gates and its weights and thresholds are integers of absolute value at most $q(n)$ (for some polynomials p and q with $p(n) \cdot q(n) \geq 2$ for all $n \in \mathbb{N}$). Without loss of generality, we assume that for each circuit input the weighted sum at each gate in B_n has distance at least 1 from its threshold (if this is not the case, first multiply all weights and thresholds of gates in B_n by 2, and then lower each threshold by 1). In addition we assume for simplicity that $x_0 = 1$ in the assumption about γ.
By the assumption about γ there exists some $l \in \mathbb{N}$ such that

$$\forall\, x \geq 1 (\gamma(-x^l) \leq \frac{1}{x} \text{ and } 1 - \gamma(x^l) \leq \frac{1}{x}).$$

Let B'_n be a boolean threshold circuit that results from B_n by multiplying first all weights and thresholds of gates in B_n by $2[2p(n)q(n)]^l$. It is obvious that B'_n also computes the boolean function g_n. In addition, for each circuit input the weighted sum at each gate in B'_n has distance at least $2[2p(n)q(n)]^l$ from its threshold.
Let C_n be the γ-circuit that results if we replace each boolean threshold gate in B'_n by a γ-gate with the same weights and threshold. Then one shows by induction on the depth of a gate G in C_n that for every boolean circuit input, the output of G differs by at most δ_n from the output of the corresponding gate in B'_n, where $\delta_n := \dfrac{1}{2p(n)q(n)}$.

In the induction step one exploits that

$$p(n) \cdot q(n) \cdot 2 \cdot [2p(n)q(n)]^l \cdot \delta_n = [2p(n)q(n)]^l.$$

This implies that a change of at most δ_n in each of the at most $p(n)$ inputs of G causes a change of at most $[2p(n)q(n)]^l$ in the value of the weighted sum that reaches G, and so this weighted sum has distance at least

$$2[2p(n)q(n)]^l - [2p(n)q(n)]^l = [2p(n)q(n)]^l$$

from the threshold. Therefore the output value of the γ-gate G differs by at most

$$\max\left\{1-\gamma([2p(n)q(n)]^l), \gamma(-[2p(n)q(n)]^l)\right\} \leq \frac{1}{2p(n)q(n)} = \delta_n$$

from the output of the corresponding boolean threshold gate in B'_n.

The preceding argument implies that for any $n \geq 2$ the γ-circuit C_n with outer threshold $\frac{1}{2}$ computes the boolean function g_n with separation $\frac{1}{4}$. □

Remark 4.3 One can also simulate polynomial size σ-circuits with weights of absolute value at most $2^{poly(n)}$ by polynomial size boolean threshold circuits with 0-1 weights; however in this case the circuit depth increases by a constant factor. This simulation can be extended to the case of real-valued inputs, where we assume that polynomially many bits of each real input are given as inputs to the simulating boolean threshold circuit. □

Remark 4.4 More recently it has been shown (see the paper by Maass in this volume, or the extended abstract [M]) that for neural nets with arbitrary piecewise polynomial activation functions γ (with polynomially many polynomial pieces of bounded degree) and *arbitrary real weights*, the class of boolean functions that can be computed in constant depth and polynomial size (with *arbitrarily small separation*) is contained in TC^0. □

REFERENCES

[B] W.J. Bultman, "Topics in the Theory of Machine Learning and Neural Computing", *Ph.D Dissertation*, University of Illinois at Chicago, 1991.

[DS] B. DasGupta and G. Schnitger, "The power of Approximating: a Comparison of Activation Functions", in *Advances in Neural Information Processing Systems 5*, S.E. Hanson, J.D. Cowan and C.L. Giles eds, pp. 615-622, 1993.

[HKP] J. Hertz, A. Krogh and R.G. Palmer, "Introduction to the Theory of Neural Computation", *Addison-Wesley*, Redwood City, 1991.

[HMPST] A.Hajnal, W. Maass, P. Pudlak, M. Szegedy and G. Turan, "Threshold Circuits of Bounded Depth", *Proc. of the 28th Annual Symp. on Foundations of Computer Science*, pp. 99-110, 1987. To appear in *J. Comp. Syst. Sci.*.

[K] C.C. Klimasauskas, "The 1989 Neuro-Computing Bibliography", *MIT Press*, Cambridge, 1989.

[M] W. Maass, "Bounds for the computational power and learning complexity of analog neural nets", *Proc. of the 25th ACM Symposium on the Theory of Computing*, 1993, 335-344.

[MSS] W. Maass, G. Schnitger, E. D. Sontag, "On the computational power of sigmoid versus boolean threshold circuits", *Proc. of the 32nd Annual IEEE Symp. on Foundations of Computer Science*, 1991, 767 - 776

[MT] W. Maass and G. Turan, "How Fast Can a Threshold Gate Learn?", in: *Computational Learning Theory and Natural Learning Systems: Constraints and Prospects*, G. Drastal, S.J. Hanson and R. Rivest eds., *MIT Press*, to appear.

[Mu] S. Muroga, "Threshold Logic and its Applications", *Wiley*, New York, 1971.

[S] E.L. Schwartz, "Computational Neuroscience", *MIT Press*, Cambridge, 1990.

[So] E.D. Sontag, "Feedforward Nets for Interpolation and Classification", *J. Comp. Syst. Sci.*, **45**, 1992, 20-48.

[SS1] E.D. Sontag and H.J. Sussmann, "Backpropagation can give rise to spurious local minima even for networks without hidden layers", *Complex Systems* **3**, pp. 91-106, 1989.

[SS2] E.D. Sontag and H.J. Sussmann, "Backpropagation separates where Perceptrons do", *Neural Networks*, **4**, pp. 243-249, 1991.

[WK] S.M. Weiss, C.A. Kulikowsky, "Computer Systems that Learn", *Morgan Kaufmann*, San Mateo, 1991.