A Lower Bound for Monotone Perceptrons

(Revised Version)

Frederic Green
Department of Mathematics and Computer Science
Clark University
Worcester, Massachusetts 01610

Abstract

It is proved that there is a monotone function in $AC_4^{(0)}$ which requires exponential size monotone perceptrons of depth 3. This solves the monotone version of a problem which, in the general case, would imply an oracle separation of PP^{PH}.

1 Introduction

Recently perceptrons (see definition below) have received some renewed attention because of their relevance to some important issues in structural complexity theory. In [6] it was shown that perceptrons cannot compute parity unless they are of exponential size [1]. This has the consequence that $(\exists A)(\oplus\mathrm{P} \not\subseteq \mathrm{PP}^{\mathrm{PH}^A})$. A result for perceptrons of depth 2 was used by Beigel [5] to provide an oracle A such that $\mathrm{P}^{\mathrm{NP}^A} \not\subseteq \mathrm{PP}^A$. Some interesting results for perceptrons were also obtained by Beigel, Reingold and Spielman [3] as a consequence of the closure under intersection of PP. Further related results and extensions have been obtained by Beigel, Reingold and Spielman [4], Aspnes, Beigel, Furst and Rudich [2], and Tarui [11].

An interesting open question left by the work of [6] is whether there is an oracle separating the hierarchy $\mathrm{PP}^{\mathrm{PH}}$. More precisely, is there an oracle A such that the following is true: $(\forall d)(\mathrm{PP}^{\Sigma_{d+1}^{p,A}} \not\subseteq \mathrm{PP}^{\Sigma_d^{p,A}})$. Indeed, in light of surprising results such as Toda's theorem [12] one might wonder if it is possible that $\mathrm{PP}^{\mathrm{PH}}$ is a proper hierarchy in any relativized world. One step in this direction was provided by [5], since the result of that paper trivially implies an oracle A such that $\mathrm{PP}^{\mathrm{NP}^A} \not\subseteq \mathrm{PP}^A$. [5] also leaves open the possibility that there is an oracle such that the separations between the levels in $\mathrm{PP}^{\mathrm{PH}}$ are via levels of the polynomial hierarchy. That is, it is possible that there is an oracle A such that $(\forall d)(\Sigma_{d+2}^{p,A} \not\subseteq \mathrm{PP}^{\Sigma_d^{p,A}})$ or even $(\forall d)(\Delta_{d+2}^{p,A} \not\subseteq \mathrm{PP}^{\Sigma_d^{p,A}})$.

In this paper we investigate a circuit problem which is motivated by the above questions. The associated circuit problems are also of intrinsic interest. We suspect that certain $\mathrm{AC}^{(0)}$ boolean functions of depth $d+1$ cannot be computed by even exponentially large perceptrons of depth d. In fact this problem can be reduced to a problem about circuits of a specific depth via the Håstad switching lemma [7]. If we could show the result for $d = 3$, it then follows for all $d > 3$ by applying that lemma. Lacking the techniques to prove this in full generality, we concentrate here on monotone perceptrons. Our main result is that certain monotone depth 4 $\mathrm{AC}^{(0)}$ functions cannot be computed by exponentially large depth 3 *monotone* perceptrons. This result represents a refinement of a result due to Yao [13] (whose techniques are also adopted). However to explain the relationship with Yao's work and to describe our results more precisely, we must introduce the basic concepts and notation. A threshold gate over the N Boolean inputs $x_1, ..., x_N$ with weights $w_i, i = 1, ..., N$ and threshold t is a Boolean gate which outputs 1 iff $\sum_{i=1}^m w_i x_i \geq t$. Throughout this paper we consider threshold circuits of bounded weight, that is with w_i bounded from above by a constant independent of N. A threshold gate is *monotone* if all its weights are non-negative.

Definition 1 A *perceptron* is a circuit consisting of a single threshold gate whose inputs are the outputs of constant depth boolean circuits over the basis { $\wedge, \vee$, not }. We refer to these subcircuits as the *AND/OR subcircuits.* A perceptron is *monotone* if its threshold gate is monotone and its AND/OR subcircuits contain no negations.

[1]In [6] perceptrons were referred to as $\mathrm{PP}^{\mathrm{PH}}$-circuits. In [5] the connection with perceptrons was first noted, and we adopt this older terminology here.

Definition 2 The class of sets recognizable by perceptrons of polynomial size and depth k is denoted $\mathrm{TAC}_k^{(0)}$, and the union over k of these classes is denoted $\mathrm{TAC}^{(0)}$. The monotone versions of these classes are denoted (respectively) $\widehat{\mathrm{TAC}}_k^{(0)}$ and $\widehat{\mathrm{TAC}}^{(0)}$.

In general, we denote the monotone version of any circuit class C by $\widehat{C}$. When it is unambiguous, we use the same notation for a complexity class and the associated set of circuits.

Notation 3 *Let A and B be classes of circuits. If there is a function computable by circuits in class A which requires circuits in class B of exponential size, we write $A \not\subseteq_{exp} B$.*

Yao [13] has shown the following,

Theorem 4 (Yao) *For all k, $\widehat{\mathrm{TC}}_{k+1}^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TC}}_k^{(0)}$.*

In fact this theorem was proved in [13] via the following stronger result,

Theorem 5 (Yao) *For all k, $\widehat{\mathrm{AC}}_{2k}^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TC}}_k^{(0)}$.*

Since $\widehat{\mathrm{AC}}_k^{(0)} \subseteq \widehat{\mathrm{TAC}}_k^{(0)}$ and $\widehat{\mathrm{TAC}}_k^{(0)} \subseteq \widehat{\mathrm{TC}}_k^{(0)}$ for any k, the following is immediate,

Corollary 6 *For all k, $\widehat{\mathrm{AC}}_{2k}^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TAC}}_k^{(0)}$ and therefore $\widehat{\mathrm{TAC}}_{2k}^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TAC}}_k^{(0)}$.*

In this paper we sharpen Theorem 5 and Corollary 6 for $k = 3$, that is, we show

Theorem 7 *$\widehat{\mathrm{AC}}_4^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TAC}}_3^{(0)}$ and therefore $\widehat{\mathrm{TAC}}_4^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TAC}}_3^{(0)}$.*

While this paper was being written, a stronger result due to Håstad and Goldmann [8] appeared, which states that for all k, $\widehat{\mathrm{AC}}_{k+1}^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TC}}_k^{(0)}$. The results reported here were obtained independently and differ in some important technical details. A discussion of Håstad and Goldmann's results and their relationship with this paper is given at the end of this section.

To repeat our main motivation more precisely, if Theorem 7 could be shown to hold in the *non*-monotone case, i.e., if $\mathrm{AC}_4^{(0)} \not\subseteq_{\mathrm{exp}} \mathrm{TAC}_3^{(0)}$, then an application of the Håstad switching lemma [7] would establish a separation of the hierarchy $\mathrm{PP}^{\mathrm{PH}}$ relative to some oracle. Depth 4 versus depth 3 is critical, since it would be the base case in such a proof. The switching lemma says that, applying a random restriction to the inputs, the lowest levels of AND and OR can be switched with high probability. Suppose for example we originally have AND's at the second lowest level and OR's at the lowest level. Applying a random restriction, with high probability we can then make any AND of OR's into an OR of AND's and reduce the depth by absorbing one layer of OR's with the OR's above it. It is clear that the minimum number of levels of AND's and OR's to which we can reduce via this method is 2. Since neither AND's nor OR's can be "absorbed" into a threshold gate, the smallest depth perceptron to which we can reduce consists of a threshold gate over depth 2 AND/OR subcircuits.

Observe that Corollary 6 does not immediately imply the second clause of Theorem 7. It is not clear that $\widehat{\mathrm{TAC}}_{k+1}^{(0)} \subseteq \widehat{\mathrm{TAC}}_k^{(0)}$ implies $\widehat{\mathrm{TAC}}_{2k}^{(0)} \subseteq \widehat{\mathrm{TAC}}_k^{(0)}$. This is because $\widehat{\mathrm{TAC}}^{(0)}$ circuits are not uniform on every level (the root is a threshold while lower levels consists of AND's and OR's). For the same reason, $\widehat{\mathrm{TAC}}_{2k}^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TAC}}_{k}^{(0)}$ is not known to imply $\widehat{\mathrm{TAC}}_{k+1}^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TAC}}_{k}^{(0)}$, nor is this known in the non-monotone case. Similarly it is not known if the collapse of $\mathrm{PP}^{\mathrm{PH}}$ translates upwards, i.e., if $\mathrm{PP}^{\Sigma_{d+1}^p} \subseteq \mathrm{PP}^{\Sigma_d^p}$ implies $\mathrm{PP}^{\mathrm{PH}} \subseteq \mathrm{PP}^{\Sigma_d^p}$.

The canonical $\mathrm{AC}_k^{(0)}$ functions $f_{k,N}$ (see section 2 for definitions), which we use for achieving the separations, are monotone. Upper bounds for monotone perceptrons computing $f_{k,N}$ imply identical upper bounds for non-monotone perceptrons computing arbitrary $\mathrm{AC}_k^{(0)}$ languages. Hence it makes sense to look for monotone simulations of $f_{k,N}$ before looking for more general simulations of $\mathrm{AC}_k^{(0)}$ functions. Our results show that small monotone perceptrons computing these functions cannot exist.

It is necessary to address a result due to Allender [1], which issues a warning as to the significance of results about monotone threshold circuits. Essentially due to the fact that Toda's theorem [12] relativizes, Yao's result (Theorem 5 above) does not carry over to the non-monotone case, at least not via exponential lower bounds. Allender showed explicitly that any $\mathrm{AC}^{(0)}$ circuit can be simulated by a depth-3 *non*-monotone $\mathrm{TC}^{(0)}$-circuit of sub-exponential but superpolynomial size. However, the situation with perceptrons may very well be different. It seems quite possible that $\mathrm{AC}^{(0)}$ is *not* contained in some specific level of $\mathrm{TAC}^{(0)}$. Thus far there is evidence for this, but only in the known lower bounds for $\mathrm{TAC}_2^{(0)}$. Minsky and Papert prove there is an $\mathrm{AC}_2^{(0)}$ function which cannot be computed by any $\mathrm{TAC}_2^{(0)}$-circuit with $\sqrt{N}/2$ bottom fan-in, regardless of circuit size or weights (see [10], section 3.2). Beigel [5] sharpened this result for perceptrons with small weights.

We now describe our results in more detail and clarify the relationship between this paper and the one of Håstad and Goldmann [8]. Håstad and Goldmann prove that for all k, $\widehat{\mathrm{AC}}_{k+1}^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TC}}_{k}^{(0)}$. They also adopt similar techniques to Yao's, using a more detailed version of his inductive argument. Their theorem clearly implies the main Theorem 7 as stated above. However the more precise statement given in the main result, Theorem 12 (section 3), is not implied by their results, and the techniques of the current paper take advantage of the special circuits we have to deal with. Theorem 12 gives a trade-off between the size of the AND/OR subcircuits and the number of these subcircuits, and this leads to a lower bound which is closer to optimal. The upper bound is $2^{(N+1)lgN}$. The lower bound of Håstad and Goldmann is 2^{cN} while the one obtained here is $2^{cNlglgN}$ (see Corollary 14). If the AND/OR subcircuits are sufficiently "small" (namely, 2^{N^ϵ} for $\epsilon < 1$), our lower bound is $N^{\alpha N}$ (see Corollary 13), which is much closer to optimal. For simpler circuits we obtain results that are optimal. In Theorem 10 we present the optimal improvement of Minsky and Papert's result on $\mathrm{AC}_2^{(0)}$ vs. $\mathrm{TAC}_2^{(0)}$ in the monotone setting, and Theorem 11 gives the optimal lower bound for $\widehat{\mathrm{AC}}_3^{(0)}$ vs. $\widehat{\mathrm{TAC}}_2^{(0)}$.

Thus in the context of perceptrons, the current result strengthens quantitatively the main message of the Håstad and Goldmann paper. Both the Håstad and Goldmann results and the current results show that in the monotone setting, in some circumstances threshold gates are no more powerful than AND and OR gates. Håstad and Goldmann show that to compute the canonical depth k $\mathrm{AC}^{(0)}$ function we need a depth $k-1$ monotone threshold circuit of exponential size. We show here (at least for $k = 4$) that to compute the canonical depth k $\mathrm{AC}^{(0)}$ function we need a depth $k-1$

monotone perceptron whose size is almost the same as what we would need if the single threshold gate were replaced by an OR gate. In this sense, the results obtained here explore the limits of Yao's techniques as applied to perceptrons.

2 Preliminaries

We define our notation and conventions and introduce additional necessary concepts. Many of these are from Yao [13], sometimes in a slightly different form.

Let $f_{k,N}(x_1, x_2, ..., x_{N^k})$ denote the canonical $\mathrm{AC}_k^{(0)}$-function. The defining circuit for $f_{k,N}$ consists of an $\vee$ ($\wedge$) as the top gate if k is even (odd), and $k-1$ alternating levels of $\wedge$'s and $\vee$'s below that. The fan-in of all gates in the defining circuit is N, so that the number of inputs is N^k. More precisely, we define $f_{1,N}(x_1, ..., x_N) = \bigwedge_{i=1}^{N} x_i$, and

$$f_{k,N}(x_1, ..., x_{N^k}) = \begin{cases} \bigvee_{i=1}^{N} f_{k-1,N}(x_{(i-1)N^{k-1}+1}, x_{(i-1)N^{k-1}+2}, ..., x_{(i-1)N^{k-1}+N^{k-1}}) & \text{if } k \text{ is even} \\ \bigwedge_{i=1}^{N} f_{k-1,N}(x_{(i-1)N^{k-1}+1}, x_{(i-1)N^{k-1}+2}, ..., x_{(i-1)N^{k-1}+N^{k-1}}) & \text{if } k \text{ is odd.} \end{cases}$$

Clearly for any k, $f_{k,N} \in \widehat{\mathrm{AC}}_k^{(0)}$.

The notation $f_{k,N}$ will be used both for the function and the defining circuit. Note that the defining circuits for $f_{k,N}$ are "stratified", that is, the output of any gate can only go to gates on the next highest level. It will be useful to refer to individual gates in $f_{k,N}$, as well as the set of inputs that can affect the output of that gate. Since it will be of central importance, consider $f_{4,N}$. The "level 1" gate is the top ($\vee$) gate, the "level 2" gates are the $\wedge$-gates with wires leading to the level 1 gate, and so forth. We refer to the set of inputs which can affect a given gate on a given level as a "block". For example, the first level 2 $\wedge$-block for $f_{4,N}$ consists of the inputs $\{x_1, x_2, ..., x_{N^3}\}$, and the l^{th} level 2 $\wedge$-block consists of inputs $\{x_{(l-1)N^3+1}, ..., x_{lN^3}\}$. Note that we only apply the terminology of $\wedge$ and $\vee$ blocks to the defining circuits of $f_{k,N}$, and not to arbitrary circuits. See Figure 1 for an illustration of these concepts for $f_{4,3}$.

The following definition is essentially due to Yao, although he phrases it in terms of probability distributions.

Definition 8 For any function $h : \{0,1\}^N \to \{0,1\}$, a pair of sets (p, q) is called a *separator* if p is a subset of $h^{-1}(1)$ and q is a subset of $h^{-1}(0)$.

We use the same separators as in [13]. $(p_{k,N}, q_{k,N})$ denotes the separator for $f_{k,N}$. Let $(x_1, ..., x_{N^k})$ denote the inputs to $f_{k,N}$. $p_{1,N}$ consists of the single element in which all of the x_i's, $i \in \{1, ..., N\}$ are 1 (more formally, $p_{1,N} = \{(1, ..., 1)\}$ where there are N 1's). $q_{1,N}$ consists of all settings of $(x_1, ..., x_N)$ in which exactly one of the x_i's, $i \in \{1, ..., N\}$ is 1 (i.e., $q_{1,N}$ consists of all input settings of the form $(0, ..., 0, 0, 1, 0, 0, ..., 0)$). The remaining separators are defined inductively.

- Even k: Define $p_{k,N}$ to include all input settings of the following form. Choose one of the level 2 $\wedge$-blocks of $f_{k,N}$. Set the inputs in that block according to an element of $p_{k-1,N}$, and set all others to 0.

  Define $q_{k,N}$ to include all input settings in which the inputs in each level 2 $\wedge$-block are set independently according to some element of $q_{k-1,N}$.

- Odd k: Define $p_{k,N}$ to include all settings in which the inputs in each level 2 $\vee$-block are set independently according to some element of $p_{k-1,N}$.

  Define $q_{k,N}$ to include all input settings of the following form. Choose one of the level 2 $\vee$-blocks of $f_{k,N}$. Set the inputs in that block according to an element of $q_{k-1,N}$, and set all other inputs to 1.

For any set p of input settings and any boolean function $h : \{0,1\}^N \to \{0,1\}$, $Pr_p(h(x) = 1)$ denotes the probability that $h(x) = 1$ when the input setting x is chosen randomly and uniformally from the set p. We will always use the sets $p_{k,N}$ and $q_{k,N}$ in this context, and we use the phrase "choosing from p" to mean "choosing an element randomly and uniformally from the set p".

The ϵ-discriminator method for proving lower bounds on threshold circuits was introduced by Hajnal, Maass, Pudlák, Szegedy and Turán in [9]. For the separator (p,q) and a subcircuit C, define the *advantage* Δ of C as $\Delta = Pr_p(C(x) = 1) - Pr_q(C(x) = 1)$. If we can find a separator such that the advantage of any circuit of the same type as C is small, then it follows that a circuit consisting of a threshold over subcircuits of this type must be large. This is stated precisely in the following lemma, which is Lemma 3.3 in reference [9]. It is valid for non-monotone as well as monotone circuits.

Lemma 9 *[9] Let T be a threshold circuit with subcircuits $C_1, C_2, ..., C_m$, so that T outputs 1 if and only if $\sum_{i=1}^{m} C_i(x) \geq t$ for some threshold t. Let T compute the boolean function $f : \{0,1\}^n \to \{0,1\}$, and let (p,q) be a separator for f. Let $\Delta_i = Pr_p(C_i(x) = 1) - Pr_q(C_i(x) = 1)$. Then if $\Delta_i \leq s^{-1}$ for all $i \in \{1, ..., m\}$, then $m \geq s$.*

The intuitive reason for choosing the given separator is as follows. $p_{k,N}$ is chosen so that the fewest possible inputs are 1 when $f_{k,N}$ is 1. $q_{k,N}$ is chosen such that the fewest number of inputs are 0 when $f_{k,N}$ is 0. By choosing $p_{k,N}$ in this way we tend to make $Pr_p(C_i(x) = 1)$ small, and by choosing $q_{k,N}$ as we have, we tend to make $Pr_q(C_i(x) = 1)$ large. Putting the two together conspires to make Δ_i as small as possible. This strategy makes essential use of the monotonicity of the perceptron.

3 Main Results

In this section we prove Theorem 7, which asserted that $\widehat{\mathrm{AC}}_4^{(0)} \not\subseteq_{exp} \widehat{\mathrm{TAC}}_3^{(0)}$. We first present two simpler results which illustrate the proof technique and are also interesting in themselves. The first one says that in any $\widehat{\mathrm{TAC}}_2^{(0)}$ circuit which computes $f_{2,N}$, the subcircuits (which are single AND or OR gates) must have fan-in N and the circuit size must be N, which are both optimal. In the non-monotone case, as stated in the introduction, the best known lower bound on the fan-in is $\sqrt{N}/2$, as proved by Minsky and Papert. It would be interesting to see if the optimal bound in the monotone case extends to the non-monotone case.

Theorem 10 *Let T be a $\widehat{\mathrm{TAC}}_2^{(0)}$ circuit which computes $f_{2,N}$. Then the fan-in of at least one of the $\wedge$ or $\vee$ gates of T must be at least N and T must consist of at least N gates.*

Proof: Each subcircuit (in this case, a single gate) of T can be either an $\vee$, in which case we will refer to it as D, or an $\wedge$, in which case we will refer to it as C. For notational convenience, we drop the N subscript on the seperators, that is, we denote $p_{k,N}$ and $q_{k,N}$ by p_k and q_k respectively. For any $\wedge$ or $\vee$ gate G of T, define $\Delta = Pr_{p_2}(G=1) - Pr_{q_2}(G=1)$. By Lemma 9, if T has finitely many gates, for some gate G we must have $\Delta > 0$. This gate may be an $\vee$ or an $\wedge$. We consider these two possibilities separately.

Case 1 (OR): Let $\Delta = Pr_{p_2}(D=1) - Pr_{q_2}(D=1)$. Suppose that $\Delta > 0$. Observe that $Pr_{q_2}(D=1) = 1$ (and therefore $\Delta \leq 0$) unless D is of a special form: It has at most one input from each level 2 $\wedge$-block. Thus we may suppose D is of this form, and that it gets its inputs from s $\wedge$-blocks. Note that since there are N $\wedge$-blocks, it follows that $s \leq N$. Choosing from q_2, the probability that an input from a given level 2 $\wedge$-block is zero is $\frac{1}{N}$. Since the assignments to level 2 $\wedge$-blocks are independent in q_2, it follows that $Pr_{q_2}(D=0) = \frac{1}{N^s}$ and therefore,

$$Pr_{q_2}(D=1) = 1 - \frac{1}{N^s}.$$

Now choosing from p_2, the probability that an input from a given level 2 $\wedge$-block is 1 is $\frac{1}{N}$. Since s such blocks contribute to D, this implies that $Pr_{p_2}(D=1) \leq \frac{s}{N}$. Combining these observations, we have,

$$\Delta \leq \frac{s}{N} + \frac{1}{N^s} - 1.$$

For any fixed N, the quantity $\frac{s}{N} + \frac{1}{N^s}$ is increasing for $1 \leq s \leq N$. If $s = N-1$, note that $\frac{s}{N} + \frac{1}{N^s} = 1 - \frac{1}{N} + \frac{1}{N^{N-1}}$, which implies that $\Delta < 0$. Thus if $\Delta > 0$, it follows that $s \geq N$ and therefore $s = N$, as claimed. Note that this also implies that $\Delta \leq N^{-N} \leq \frac{1}{N}$.

Case 2 (AND): Let C denote the $\wedge$ gate. Define $\Delta = Pr_{p_2}(C=1) - Pr_{q_2}(C=1)$, and again assume $\Delta > 0$. Now $Pr_{p_2}(C=1) = 0$ (and hence $\Delta \leq 0$) unless C is of a special form: All of its inputs can come from at most one $\wedge$-block. Suppose C is of this form and that the fan-in of C is s. Again, $s \leq N$. The probability that $C = 1$ choosing from p_2 is independent of s:

$$Pr_{p_2}(C=1) \leq \frac{1}{N}.$$

For this same C, it is easy to see that $Pr_{q_2}(C=1) = 1 - \frac{s}{N}$. Therefore,

$$\Delta \leq \frac{1}{N} - (1 - \frac{s}{N}).$$

Similar to case 1, we find that if $\Delta > 0$, then $s = N$, as claimed. Note that again we have established that $\Delta \leq \frac{1}{N}$.

In both cases, we have the inequality $\Delta \leq \frac{1}{N}$, which, by Lemma 9, implies that the size of the circuit must be at least N. ▮

We further illustrate the technique of the main theorem by proving optimal lower bounds for $\widehat{\mathrm{TAC}}_2^{(0)}$ versus $\widehat{\mathrm{AC}}_3^{(0)}$. The proof is simpler in detail but follows the same broad outline as the proof of Theorem 12 which follows. The N^N lower bound is optimal because in the defining circuit for $f_{3,N}$ we can use the distributive law to switch the two upper AND and OR levels, yielding N^N AND gates, and then collapse the lower two levels of AND's, to yield an OR of N^N AND's each of fan-in N^2.

Theorem 11 *If $N > 2$, any depth 2 monotone perceptron T computing $f_{3,N}(x)$ requires N^N gates.*

Proof: We use the same notations and conventions as in the previous proof. We consider each case for a subcircuit of T ($\vee$, which we denote as D, or $\wedge$, which we refer to as C) separately. Since in the defining circuit for $f_{3,N}$ the only $\vee$'s are at level 2, we lose no precision in referring to the set of inputs that affect these gates as $\vee$-blocks. Defining $\Delta = Pr_{p_3}(G = 1) - Pr_{q_3}(G = 1)$, it suffices, by Lemma 9, to show that $\Delta \leq N^{-N}$.

Case 1 (OR): Let $\Delta = Pr_{p_3}(D = 1) - Pr_{q_3}(D = 1)$. Observe that $Pr_{q_3}(D = 1) = 1$ (and therefore $\Delta \leq 0$) unless D is of a special form: All of its inputs come from at most one $\vee$-block, and furthermore there can be at most one input from each level 3 $\wedge$-block. Suppose D is of this form, and that its inputs come from s $\wedge$-blocks ($s \leq N$). For such a gate, it is not hard to see that

$$Pr_{q_3}(D = 0) = \frac{1}{N} \cdot \frac{1}{N^s} = \frac{1}{N^{s+1}}$$

and therefore,

$$Pr_{q_3}(D = 1) = 1 - \frac{1}{N^{s+1}}.$$

Now choosing from p_3, the probability that there are inputs from a given level 3 $\wedge$-block that are 1 is $\frac{1}{N}$. Since s such blocks contribute to D, this implies that $Pr_{p_3}(D = 1) \leq \frac{s}{N}$. Combining these observations, we have,

$$\Delta \leq \frac{s}{N} + \frac{1}{N^{s+1}} - 1.$$

Arguing exactly as in case 1 of Theorem 10, if $\Delta \geq 0$, it follows that $s = N$. Hence, $\Delta \leq N^{-N-1} < N^{-N}$, as claimed.

Case 2 (AND): Let C denote the $\wedge$ gate. Let $\Delta = Pr_{p_3}(C = 1) - Pr_{q_3}(C = 1)$. Now $Pr_{p_3}(C = 1) = 0$ (and hence $\Delta \leq 0$) unless C has inputs from at most one level 3 $\wedge$-block of any $\vee$-block. Suppose C is of this form and that the fan-in of C is s. Then,

$$Pr_{p_3}(C = 1) \leq \frac{1}{N^s}.$$

For this same C, it is easy to see that $Pr_{q_3}(C = 1) \geq 1 - \frac{s}{N}$. Therefore,

$$\Delta \leq \frac{1}{N^s} - (1 - \frac{s}{N}).$$

Similar to case 1, we find that if $\Delta \geq 0$, then $s \geq N$. Hence $\Delta \leq N^{-N}$, as claimed. ∎

As in the proof of Theorem 10, a crucial result in the proof of the previous theorem is the lower bound N on the fan-in s of the $\wedge$ or $\vee$-gates of the perceptron. It should be noted that the lower bound on fan-in is tight if the perceptron is a threshold of $\vee$'s, although in that case the bound on size is not tight since then the optimal bound would be N^{N+1}, as demonstrated in case 1 of the proof of Theorem 11. Similarly, the lower bound on size is tight if the perceptron is a threshold of $\wedge$'s, although in that case the bound on fan-in is not tight since then the optimal bound would be N^2. Again it would be interesting to see if the Minsky and Papert lower bound on the fan-in, $\sqrt{N}/2$, can be improved to N in the non-monotone case.

We now turn to the main result. We derive a trade-off between the size of the AND/OR subcircuits and the number of such subcircuits. We first give the lower bound on the number of gates when the maximum size of an AND/OR subcircuit is given by any function $u(N)$ bounded from above by $N^{const \cdot N}$. Later we consider various possibilities for $u(N)$.

Theorem 12 *Let $0 < \delta < 1$ be a constant and $u : \mathbb{N} \to \mathbb{N}$ be any function such that $u(N) \leq N^{\delta N}$ for all N. There exist constants $N_0 > 0$ and $0 < \beta < 1$ such that the following is true. Let T be any depth 3 perceptron which computes $f_{4,N}(x)$ where $N \geq N_0$ and the largest of the AND/OR subcircuits of T has $u(N)$ gates. Then the number of AND/OR subcircuits in T is at least*

$$min(N^{\beta N}, (\frac{Nlg(N)}{lg(Nu(N))})^{\beta N}).$$

Proof: Let N_0 be chosen such that $N_0^{(1-\delta)N_0} > N_0^3$ and let β be such that $\beta + \delta \leq 1 - \frac{2}{N_0}$. Assume throughout the following that $N > N_0$. Each subcircuit of T can be either an $\vee$-of-$\wedge$'s, in which case we refer to it as D, or an $\wedge$-of-$\vee$'s, in which case we refer to it as C. We consider each of these cases separately. For any subcircuit G of T, let $\Delta = Pr_{p_4}(G = 1) - Pr_{q_4}(G = 1)$. By Lemma 9, it suffices to show that either $\Delta \leq N^{-\beta N}$ or $\Delta \leq (\frac{lg(Nu(N))}{Nlg(N)})^{\beta N}$. Also as in the proofs of Theorems 10 and 11, drop the N subscript on the separators, denoting $p_{k,N}$ and $q_{k,N}$ by p_k and q_k respectively.

Case 1 (OR-OF-AND's): In this case $\Delta = Pr_{p_4}(D = 1) - Pr_{q_4}(D = 1)$. Let the $\wedge$-gates of D be denoted C_i. Let $c \leq N$ be the number of level 2 $\wedge$-blocks from which D takes its inputs. Let D_l ($l \in \{1, ..., c\}$) denote the function obtained from D when we restrict all inputs in f_4 to be 0 except those in level 2 $\wedge$-block l. It is clear that for input settings in p_4, $D = 1 \Leftrightarrow \bigvee_{l=1}^{c} (D_l = 1)$. We thus have

$$Pr_{p_4}(D = 1) = Pr_{p_4}(\bigvee_{l=1}^{c} (D_l = 1)).$$

Using the fact that the D_l's are mutally exclusive and the fact that $Pr_{p_4}(D_l = 1) = \frac{1}{N} Pr_{p_3}(D_l = 1)$, this implies,

$$\begin{aligned} Pr_{p_4}(D = 1) &= \sum_{l=1}^{c} Pr_{p_4}(D_l = 1) \\ &= \frac{1}{N} \sum_{l=1}^{c} Pr_{p_3}(D_l = 1) \end{aligned}$$

Now $\bigvee_{l=1}^{c}(D_l = 1) \Rightarrow D = 1$ because D is monotone. Hence $Pr_{q_4}(\bigvee_{l=1}^{c}(D_l = 1)) \leq Pr_{q_4}(D = 1)$, so that we have

$$\begin{aligned} Pr_{q_4}(D = 1) &\geq 1 - Pr_{q_4}(\bigwedge_{l=1}^{c}(D_l = 0)) \\ &= 1 - \prod_{l=1}^{c} Pr_{q_3}(D_l = 0) \end{aligned}$$

where the equality holds because the D_l's are independent in q_4 (for each l, D_l depends only on inputs from a level 2 $\wedge$-block, and in q_4 these inputs are assigned independently according to q_3). Thus we find that

$$\Delta \leq \frac{1}{N}\sum_{l=1}^{c} Pr_{p_3}(D_l = 1) + \prod_{l=1}^{c} Pr_{q_3}(D_l = 0) - 1. \tag{1}$$

Let C_i^l denote the function that C_i computes when all inputs except those in level 2 $\wedge$-block l are 0. Like C_i, C_i^l is also an $\wedge$-gate. Fix an l and consider assignments to level 2 $\wedge$-block l chosen from p_3. Observe that the only way in which we can have $Pr_{p_3}(C_i^l = 1) \neq 0$ is that in which the set of inputs to C_i^l is a subset of a level 2 $\wedge$-block, and furthermore the subset can be partitioned into groups of inputs, each group from at most one of the level 4 $\wedge$-blocks of a level 3 $\vee$-block of f_4. Suppose $Pr_{p_3}(C_i^l = 1) \neq 0$ and that C_i^l has inputs from exactly k_i^l such $\wedge$-blocks. (Refer to Figure 2 for clarification.) The probability that any one of these $\wedge$-blocks is on is $\frac{1}{N}$, and hence for any C_i^l such that $Pr_{p_3}(C_i^l = 1) \neq 0$, we have,

$$Pr_{p_3}(C_i^l = 1) = \frac{1}{N^{k_i^l}}.$$

Let s_l denote the number of C_i^l's, i.e., s_l is the fan-in of D_l. Thus

$$\begin{aligned} Pr_{p_3}(D_l = 1) &= Pr_{p_3}(\bigvee_{i=1}^{s_l}(C_i^l = 1)) \\ &\leq \sum_{i=1}^{s_l} Pr_{p_3}(C_i^l = 1) \\ &\leq \sum_{i=1}^{s_l} \frac{1}{N^{k_i^l}} \\ &\leq \frac{s_l}{N^{k_{min}^l}} \end{aligned}$$

where $k_{min}^l = \min_i(k_i^l)$. By hypothesis, $s_l \leq u(N)$. Therefore,

$$Pr_{p_3}(D_l = 1) \leq \frac{u(N)}{N^{k_{min}^l}}.$$

Then since $c \leq N$, we find

$$\frac{1}{N}\sum_{l=1}^{c} Pr_{p_3}(D_l = 1) \leq \max_l Pr_{p_3}(D_l = 1) \leq \frac{u(N)}{N^{k_m}} \tag{2}$$

where $k_m = \min_l(k^l_{min})$. Now for any l and i, $Pr_{q_3}(D_l = 0) \leq Pr_{q_3}(C^l_i = 0)$. For any assignment in q_3, C^l_i can be 0 only if one of its inputs from a level 4 $\wedge$-block is 0. Thus for any i such that $Pr_{p_3}(C^l_i = 1) \neq 0$, we have $Pr_{q_3}(C^l_i = 0) \leq \frac{k^l_i}{N}$. Therefore

$$\prod_{l=1}^{c} Pr_{q_3}(D_l = 0) \leq \frac{k_m}{N}. \tag{3}$$

Putting this together with eq. 2 and eq. 1,

$$\Delta \leq \frac{u(N)}{N^{k_m}} + \frac{k_m}{N} - 1. \tag{4}$$

Consider two possibilities for the size of k_m. First suppose $k_m \geq \frac{lg(u(N))}{lg(N)} + 1$. (Note that since k^l_{min} is defined only if $Pr_{p_3}(D_l = 1) \neq 0$, the inequality $k_m \geq \frac{lg(u(N))}{lg(N)} + 1$ really means that for all l, either $Pr_{p_3}(D_l = 1) = 0$ or $k^l_{min} \geq \frac{lg(u(N))}{lg(N)} + 1$.) Then $\frac{lg(u(N))}{lg(N)} + s \leq k_m \leq \frac{lg(u(N))}{lg(N)} + s + 1$, for some integer s with $1 \leq s \leq N - \frac{lg(u(N))}{lg(N)}$. Then by eq. 4,

$$\Delta \leq \frac{1}{N^s} + \frac{1}{N}\left(\frac{lg(u(N))}{lg(N)} + s + 1\right) - 1.$$

If $\Delta \geq 0$ it follows that, $\frac{1}{N^s} \geq 1 - \frac{1}{N}(\frac{lg(u(N))}{lg(N)} + s + 1)$, which, by the upper bound on $u(N)$, yields

$$\frac{1}{N^s} \geq 1 - \delta - \frac{s+1}{N}.$$

If $s = (1-\delta)N - 2$, the above inequality implies $N^{(1-\delta)N} \leq N^3$, which contradicts the choice of N. Again, as noted in the proof of Theorem 10, $\frac{1}{N^s} + \frac{s}{N}$ is an increasing function of s for $1 \leq s \leq N$, and therefore $s \geq (1-\delta)N - 1$, which implies $s \geq \beta N$, by the choice of β. Therefore,

$$\Delta \leq \frac{1}{N^s} \leq N^{-\beta N},$$

as desired.

Now consider the case $k_m < \frac{lg(u(N))}{lg(N)} + 1$. We show that $\Delta \leq (\frac{lg(Nu(N))}{Nlg(N)})^{\beta N}$. For suppose that for σ values of l ($1 \leq \sigma \leq N$), $Pr_{p_3}(D_l = 1) \neq 0$ and $k^l_{min} < \frac{lg(u(N))}{lg(N)} + 1 = \frac{lg(Nu(N))}{lg(N)}$. Wlog suppose this is true for $1 \leq l \leq \sigma$. The best upper bound we can put on $Pr_{p_3}(D_l = 1)$ for $1 \leq l \leq \sigma$ is 1. For the remaining $N - \sigma$ values of l,

$$Pr_{p_3}(D_l = 1) \leq \frac{u(N)}{N^{k^l_{min}}} \leq \frac{u(N)}{N^{\frac{lg(u(N))}{lg(N)}+1}} = \frac{1}{N}.$$

Therefore,

$$\frac{1}{N}\sum_{l=1}^{c} Pr_{p_3}(D_l = 1) \leq \frac{\sigma}{N} + \frac{N-\sigma}{N^2}.$$

Combining these observations with eqs. 1 and 3, we find that

$$\begin{aligned}\Delta &\leq \frac{\sigma}{N} + \frac{N-\sigma}{N^2} + \frac{lg(u(N))}{Nlg(N)} + \frac{1}{N} - 1 \\ &\leq \frac{1}{N}(1 - \frac{1}{N})\sigma + \frac{1}{N}(2 + \delta N - N).\end{aligned}$$

Thus if $\Delta \geq 0$, it follows that,

$$\sigma \geq (1-\delta)N - 2.$$

Since $c \geq \sigma$, applying eq. 1 again we then have,

$$\begin{aligned}
\Delta &\leq \prod_{l=1}^{c} Pr_{q_3}(D_l = 0) \\
&\leq \prod_{l=1}^{\sigma} \min_i Pr_{q_3}(C_i^l = 0) \\
&\leq \prod_{l=1}^{\sigma} \frac{k_{min}^l}{N} \\
&\leq \left(\frac{lg(Nu(N))}{Nlg(N)}\right)^{\sigma} \\
&\leq \left(\frac{lg(Nu(N))}{Nlg(N)}\right)^{(1-\delta)N-2}.
\end{aligned}$$

Using the choice of β, this implies that $\Delta \leq \left(\frac{lg(Nu(N))}{Nlg(N)}\right)^{\beta N}$, as desired. (End case 1).

Case 2 (AND-OF-OR'S): Similar to case 1, we let $\Delta = Pr_{p_4}(C=1) - Pr_{q_4}(C=1)$. Let the $\vee$-gates of C be denoted D_i. Let C_l denote the function obtained from C when we restrict all inputs in f_4 to be 0 except those in $\wedge$-block l ($l \in \{1, ..., c\}$, where $c \leq N$ is the number of level 2 $\wedge$-blocks from which C takes its inputs). The same considerations as in case 1 yield

$$\begin{aligned}
Pr_{p_4}(C=1) &= Pr_{p_4}(\bigvee_{l=1}^{c}(C_l = 1)) \\
&= \sum_{l=1}^{c} Pr_{p_4}(C_l = 1) \\
&= \frac{1}{N}\sum_{l=1}^{c} Pr_{p_3}(C_l = 1) \\
&\leq \max_l(Pr_{p_3}(C_l = 1)).
\end{aligned}$$

(Note that the final inequality is the first departure with the argument of case 1.) Denote the l for which $Pr_{p_3}(C_l = 1)$ is a maximum by l'. Thus $Pr_{p_4}(C=1) \leq Pr_{p_3}(C^{l'} = 1)$. For the quantity $Pr_{q_4}(C=1)$ we again follow an argument similar to that of case 1 to obtain,

$$\begin{aligned}
\Delta &\leq Pr_{p_3}(C^{l'} = 1) + \prod_{l=1}^{c} Pr_{q_3}(C_l = 0) - 1 \\
&\leq Pr_{p_3}(C^{l'} = 1) + Pr_{q_3}(C^{l'} = 0) - 1 \\
&\leq Pr_{p_3}(C^{l'} = 1) - Pr_{q_3}(C^{l'} = 1)
\end{aligned}$$

(The alert reader will wonder why a similar strategy was not followed in case 1, since it appears to be simpler here. The reason is that the product in the second term on the LHS of eq. 1 is crucial to the argument in case 1, but not here.)

A dual version of the argument for case 1, now simplified since we are using the distribution p_3, gives the desired bound on Δ. (End case 2). ▮

The restriction $N^{\delta N}$ on the size of the AND/OR subcircuits in Theorem 12 is "optimal" in the following sense. The proof would not work if $\delta = 1$. Allowing $\delta = 1$ means we allow the $\wedge$ and $\vee$ gates of T to have fan-in N^N (that is, we may take $u(N) = N^N$). This fan-in is just large enough that the circuit T could compute f_4: simply expand the lowest two levels of f_4 (this gives N^N gates for each $\vee$-gate expanded), collapse the resulting two adjacent levels of $\wedge$'s, and set the threshold to 1 (so the threshold gate is an $\vee$-gate).

If the upper bound on $u(N)$ is realized then trivially we have circuits whose size is close to optimal (that is, $N^{\delta N}$). This is also true if $u(N)$ is sufficiently small.

Corollary 13 *Let ϵ be any constant such that $0 < \epsilon < 1$. There exist constants $0 < \alpha < 1$ and N_0 such that for any $N > N_0$, if T is any depth 3 perceptron which computes $f_{4,N}$ and the size of the AND/OR subcircuits of T is at most 2^{N^ϵ}, the size of T is at least $N^{\alpha N}$.*

Proof: In Theorem 12, consider any $u(N) \leq 2^{N^\epsilon}$. It is easy to see, for some $0 < \epsilon' < 1$ and for sufficiently large N, it holds that

$$\left(\frac{Nlg(N)}{lg(Nu(N))}\right) \geq N^{\epsilon'}.$$

This implies that the number of gates of T is at least $N^{\epsilon' \beta N}$, from which the result follows by taking $\alpha = \epsilon' \beta$. ▮

For intermediate subcircuit sizes (that is, between 2^{N^ϵ} and $N^{\delta N}$), we have not been able to obtain bounds that are as close to optimal. The best one can do using Theorem 12 is given in the following.

Corollary 14 *Let $0 < \delta < 1$ be a constant. There exists a constant $N_0 > 0$ and a constant α, $0 < \alpha < 1$ such that the following is true. Let T be any depth 3 perceptron which computes $f_{4,N}(x)$ where the number of inputs is N^4 ($N \geq N_0$) and the fan-in of any of the $\wedge$ or $\vee$-gates of T is bounded from above by $N^{\delta N}$. Then the number of gates in T is at least $2^{\alpha N lg(lg(N))}$.*

Proof: Let $u(N)$ be as in Theorem 12. If $u(N) \geq 2^{N lg(lg(N))}$, we are done. If $u(N) < 2^{N lg(lg(N))}$, then it is not hard to see that if $\beta < 1$, there is a constant $\alpha < 1$ such that for sufficiently large N

$$\left(\frac{Nlg(N)}{lg(Nu(N))}\right)^{\beta N} \geq 2^{\alpha N lg(lg(N))}.$$

Applying Theorem 12, the result follows. ▮

4 Acknowledgements

I wish to thank the Departament de Llenguatges i Sistemes Informàtics, Barcelona, where this paper was originally written, for their hospitality. Conversations on this subject with Arthur Chou and Richard Beigel are gratefully acknowledged. The author is greatful to two anonymous referees for comments which greatly improved the paper.

References

[1] E. Allender, "A note on the power of threshold circuits", *Proceedings of the 30th IEEE Symposium on Foundations of Computer Science* (1989) 580-584.

[2] J. Aspnes, R. Beigel, M. Furst and S. Rudich, "The Expressive Power of Voting Polynomials", in Proceedings of the 23rd Annual ACM Symposium on Theory of Computing, ACM Press (1991) 402-409.

[3] R. Beigel, N. Reingold, and D. Spielman, "PP is closed under intersection", in *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing*, IEEE Computer Society Press (1991) 1-9. Also, J. Comp. Syst. Sci., to appear.

[4] R. Beigel, N. Reingold, and D. Spielman, "The Perceptron Strikes Back", in *Proceedings of the Sixth Annual Conference on Structure in Complexity Theory*, IEEE Computer Society Press (1991) 286-291.

[5] R. Beigel, "Perceptrons, PP, and the Polynomial Hierarchy", in *Proceedings of the Seventh Annual Conference on Structure in Complexity Theory*, IEEE Computer Society Press (1992) 14-19.

[6] F. Green, "An oracle separating $\oplus$P from $\mathrm{PP}^{\mathrm{PH}}$", *Information Processing Letters* **37** (1991) 149-153.

[7] J. Håstad, "Computational limitations for small-depth circuits", The MIT press, Cambridge 1987.

[8] J. Håstad and M. Goldmann, "On the Power of Small-Depth Threshold Circuits", *Computational Complexity* **1** (1991) 113-129.

[9] A. Hajnal, W. Maass, P. Pudlák, M. Szegedy, and G. Turán, "Threshold circuits of bounded depth", in *Proceedings 28th Annual IEEE Symposium on Foundations of Computer Science*, IEEE Computer Society Press (1987) 99-110.

[10] M. L. Minsky and S. A. Papert, "Perceptrons" (expanded edition), MIT Press, Cambridge, 1988.

[11] J. Tarui, "Randomized Polynomials, Threshold Circuits, and the Polynomial Hierarchy", in 8th Annual Symposium on Theoretical Aspects of Computer Science, Springer-Verlag LNCS 480, (1991) 238-250.

[12] S. Toda, "On the computational power of PP and $\oplus$P", in *Proceedings 30th IEEE Symposium on Foundations of Computer Science*, IEEE Computer Society Press (1989) 514-519.

[13] A. C.-C. Yao, "Circuits and local computation", in *Proceedings of the 21st ACM Symposium on Theory of Computing*, ACM Press (1989) 186-196.