PERCEPTRONS, PP, AND THE POLYNOMIAL HIERARCHY

Richard Beigel

Abstract. We construct a predicate that is computable by a perceptron with linear size, order 1, and exponential weights, but which cannot be computed by any perceptron having subexponential ($2^{n^{o(1)}}$) size, subpolynomial ($n^{o(1)}$) order and subexponential weights. A consequence is that there is an oracle relative to which P^{NP} is not contained in PP.

Key words. Perceptron; PP; polynomial hierarchy; weight.

Subject classifications. 68Q15.

1. Introduction

A *perceptron* is a depth-2 circuit with a threshold gate at the root and AND-gates at the remaining level. The *order* of a perceptron is the maximum fanin of its AND-gates. The *weight* of a perceptron is the maximum absolute value of the weights on the inputs to its threshold gate. The *size* of a perceptron is the number of AND-gates it contains. Perceptrons are an important computational model, which is used in practice, and which has been studied by Minsky & Papert (1988), Beigel *et al.* (1994), Aspnes *et al.* (1991), Tarui (1993), and Beigel *et al.* (1991b).

Minsky & Papert (1988) constructed several predicates that require exponential weight (assuming that distinct AND-gates compute distinct functions). However, all of their examples require exponential size as well.[1]

[1]Minsky and Papert did construct predicates that were computable as thresholds of a small number of basis functions, but not with small weights. However, the basis involved was ad hoc, being designed specifically to make the weights large. The usual basis for perceptrons consists of the AND-functions, as in this paper.

In this paper, we define a predicate called ODD-MAX-BIT that is computable by perceptrons having linear size, order 1, and exponential weight. However, ODD-MAX-BIT is not computable by perceptrons having quasipolynomial size, polylogarithmic order, and subexponential weight.

We think that quasipolynomial size and polylogarithmic order are theoretically important for two reasons. First, perceptrons of this size and order arise naturally in the work of Beigel *et al.* (1994), Beigel *et al.* (1991b), Tarui (1993), and Aspnes *et al.* (1991). Second, if we replace the threshold gate at the root of a perceptron with some other symmetric gate like a parity gate, then the notion of size and order still make sense; such circuits having quasipolynomial size and polylogarithmic order arise in the work of Yao (1990), Beigel & Tarui (1994), Tarui (1993), Allender (1989), and Allender & Hertrampf (1994). In all those papers, the weights are quasipolynomial as well.

Quasipolynomial size, polylogarithmic order, and quasipolynomial weight come up quite often for an additional reason. When translating between nondeterministic Turing machine complexity and circuit complexity in the manner of Furst *et al.* (1984), polynomial time translates into quasipolynomial size, polylogarithmic order, and quasipolynomial weight. Relativizable upper bounds for nondeterministic Turing machines with a particular acceptance mechanism translate into upper bounds for depth-2 circuits with a corresponding gate at the root. Lower bounds for circuits translate into separations of Turing machine complexity classes via oracles.

Toda (1991) has shown that the polynomial hierarchy is contained in P^{PP}. Toda's result has been extended by Toda & Ogiwara (1992) and independently by Tarui (1993). Tarui shows that the polynomial hierarchy is probabilistically m-reducible to PP with zero-sided error. Beigel *et al.* (1991a) have shown that $\mathrm{P}^{\mathrm{NP[log]}}$ is contained in PP. Their result has been improved by Gundermann *et al.* (1990) who showed that $\mathrm{P}^{\mathrm{C_{=}P[log]}} \subseteq \mathrm{PP}$, and by Beigel *et al.* (1994) who showed that $\mathrm{P}^{\mathrm{PP[log]}} = \mathrm{PP}$. People have asked whether some of those techniques can be extended to show that more of the polynomial hierarchy is contained in PP. Our lower bound for ODD-MAX-BIT yields an oracle relative to which P^{NP} is not contained in PP, and in fact $\mathrm{P}^{\mathrm{NP}[f(n)]}$ is contained in PP iff $f(n) = O(\log n)$. (Independently, Fu 1992 has observed that Minsky and Papert's one-in-a-box theorem yields an oracle relative to which a weaker separation holds: $\mathrm{NP}^{\mathrm{NP}} \not\subseteq \mathrm{PP}$.) Since the techniques of Toda (1991), Beigel *et al.* (1991a), Toda & Ogiwara (1992), Tarui (1993), Gundermann *et al.* (1990), and Beigel *et al.* (1994) relativize, this means that other techniques will be needed in order to determine how much of the polynomial hierarchy is contained in PP.

2. Threshold Circuits

Throughout this paper we assume that the weights on the inputs to a threshold gate are integral. If we allow a perceptron to contain identical AND-gates then there is a partial tradeoff between size and weight in perceptrons, because an AND-gate having weight w can be replaced with $|w|$ identical AND-gates each having weight $w/|w|$. Identical AND-gates can only be a nuisance when trying to prove lower bounds. We eliminate them, as well as logical negations.

We say that a perceptron is in *clean* form if it contains no logical negations and no identical AND-gates. The following lemma is essentially due to Minsky & Papert (1988).

LEMMA 2.1. *If f is computed by a perceptron with size s, weight w, and order d, then f is computed by a perceptron in clean form with size $2^d s$, weight sw, and order d.*

PROOF. For each AND-gate, replace each negated input $\overline{x}$ by $1 - x$, and replace "and" by multiplication. Expand using the distributive laws of arithmetic. Each term in the expansion is a conjunction of the inputs to f. The total number of terms obtained by expanding all AND-gates in this way is at most $2^d s$. Each term is contributed at most once per AND-gate, so when we collect terms, no weight is greater than sw in absolute value. □

DEFINITION 2.2. ODD-MAX-BIT *is the set of all strings over $\{0,1\}^*$ whose rightmost 1 is in an odd-numbered position, i.e., the set of strings of the form $x10^k$ where the length of x is even.*

By associating the weight 2^i to the input x_i if i is odd, -2^i if i is even, we obtain a family of perceptrons having size N and order 1 which compute N-bit instances of the predicate ODD-MAX-BIT. (Similar ideas were applied to $\mathrm{P}^{\mathrm{NP[log]}}$ by Hemachandra & Wechsung 1988.)

3. Polynomial Approximations

We will use an extension of Markov's theorem in approximation theory. We follow the treatment of Rivlin & Cheney (1966). If P is a real-valued function and Y is nonempty, define

$$\begin{aligned} ||P||_Y &= \max\nolimits_{x\in Y} P(x), \\ ||P|| &= \max\nolimits_{x\in[-1,1]} P(x). \end{aligned}$$

Henceforth let P denote a polynomial of degree d. Let P' denote the derivative of P. A. A. Markov's theorem, which can be found in Cheney (1966), Lorentz (1966) or Pólya & Szegö (1972, Vol. II, Part VI, Prob. 83), says that

$$||P'|| \le d^2||P||.$$

The following lemma is due to Ehlich & Zeller (1964). See Rivlin & Cheney (1966) for an improved version.

Lemma 3.1 (Ehlich and Zeller). *Let m be a positive integer, and let $Y = \{-1, -1+\frac{2}{m}, -1+\frac{4}{m}, \ldots, 1-\frac{2}{m}, 1\}$. Assume that $d^2(d^2-1)/m^2 < 6$. Then*

$$||P|| \le ||P||_Y \cdot \frac{1}{1-\frac{1}{6}d^2(d^2-1)/m^2}.$$

The following lemma is immediate.

Lemma 3.2. *Let p be a degree-d polynomial. Suppose that $p(0) = 0$ and $W \le p(x) < 3W$ for $x = 1, \ldots, m$, where m is a positive integer. Then $d > \sqrt{(\sqrt{87}-9)m}$.*

Proof. Let $P(x) = p(\frac{1}{2}m(x+1)) - \frac{3}{2}W$, and let Y be as in Lemma 3.1. Then $||P||_Y \le \frac{3}{2}W$. Since $P(-1) = -\frac{3}{2}W$ and $P(-1+\frac{2}{m}) \ge -\frac{1}{2}W$, we have $||P'|| \ge \frac{1}{2}mW$. If $d^2(d^2-1)/m^2 \ge 6$ then $d > \sqrt{\sqrt{6}m}$, so we are done. Otherwise we can apply Markov's theorem and Lemma 3.1.

$$\begin{aligned} \frac{1}{2}mW &\le ||P'|| \\ &\le d^2||P||_Y \cdot \frac{1}{1-\frac{1}{6}d^2(d^2-1)/m^2} \\ &\le d^2\frac{3}{2}W \cdot \frac{1}{1-\frac{1}{6}d^2(d^2-1)/m^2}. \end{aligned}$$

Therefore,

$$\begin{aligned}
m &\le 3d^2 \cdot \frac{1}{1-\frac{1}{6}d^2(d^2-1)/m^2}, \\
m &< 3d^2 \cdot \frac{1}{1-\frac{1}{6}d^4/m^2}, \\
1-\frac{1}{6}d^4/m^2 &< 3d^2/m, \\
0 &< \frac{1}{6}(d^2/m)^2 + 3d^2/m - 1.
\end{aligned}$$

Since d^2/m is nonnegative, $d^2/m > \sqrt{87} - 9$. $\square$

4. Hardness of ODD-MAX-BIT

Let $\max(S)$ denote the lexically maximum string belonging to the finite set S. In this section, let $\exp(k)$ denote 2^k.

LEMMA 4.1. *Let $d \ge 1$. If C is a perceptron in clean form having size s, weight w, and order d which recognizes* ODD-MAX-BIT $\cap \{0,1\}^N$ *then*

$$w \ge \frac{1}{s}\exp\left(\left\lfloor \frac{(N-1)}{2\lceil d^2/(\sqrt{87}-9)\rceil}\right\rfloor\right).$$

PROOF. Let $x = x_1, \ldots, x_N$ denote the input to C. We identify the vector x with the set $X = \{i : x_i = 1\}$. By assumption, C accepts x iff $\max(X)$ is odd. Let T denote C's threshold gate. Each input to T depends on a set $S \subseteq \{1, \ldots, N\}$ such that $0 \le |S| \le d$. For each S let $w(S)$ denote the weight given to the corresponding input to T (0 if there is no such input).

Let

$$c(X) = \sum_{S\subseteq X,\, |S|\le d} w(S)$$

denote the *total weight of* X. Then C accepts x iff $c(X) > 0$. Hence $c(\{1\}) \ge 1$. For $0 \le i \le d$, let

$$c_i(X) = \sum_{S\subseteq X,\, |S|=i} w(S)$$

denote the weight of X due to subsets of size i. Clearly

$$c(X) = \sum_{0 \le i \le d} c_i(X).$$

Suppose we have found X such that $c(X) = -W < 0$ and $\max(X) = i$, where $i \le N - 2m$. Let $m = \lceil d^2/(\sqrt{87} - 9) \rceil$, and $M = \{i+1, i+3, \ldots, i+2m-1\}$. We will find $Y \subseteq M$ such that $c(X \cup Y) \ge 2W$. Similarly if $c(X) > 0$, we will find $Y \subseteq M$ such that $c(X \cup Y) \le -2c(X)$. We start with $X = \{1\}$ and $W \ge 1$, and we iterate $\lfloor (N-1)/2m \rfloor$ times to obtain a set $X \subseteq \{1, \ldots, N\}$ with $|c(X)| \ge \exp(\lfloor (N-1)/2m \rfloor)$, so $w \ge \frac{1}{s}\exp(\lfloor (N-1)/2m \rfloor) = \frac{1}{s}\exp\left(\left\lfloor \frac{(N-1)}{2\lceil d^2/(\sqrt{87}-9)\rceil} \right\rfloor\right)$.

It remains to show that the desired set Y always exists. We will consider only the case $c(X) = -W < 0$, because the other case is entirely similar. Fix X, i, m, and M as above. For each nonempty $Y \subseteq M$, we have $c(X \cup Y) \ge 0$. For each $S \subseteq M$ let

$$b(S) = \sum_{X \cup R = X \cup S,\, |R| \le d} w(R)$$

denote the weight due to S. Let

$$u_k(Y) = \operatorname*{ave}_{S \subseteq Y,\, |S| = k} b(S)$$

denote the average weight due to a k-element subset of Y. Note for each $j \ge k$ that $u_k(M)$ is equal to $\operatorname{ave}_{Y \subseteq M, |Y| = j} u_k(Y)$. Now

$$\begin{aligned} c(X \cup Y) &= \sum_{S \subseteq Y} b(S) \\ &= c(X) + \sum_{S \subseteq Y,\, 1 \le |S| \le d} b(S) \\ &= c(X) + \sum_{1 \le k \le d} \sum_{S \subseteq Y,\, |S| = k} b(S) \\ &= c(X) + \sum_{1 \le k \le d} \binom{|Y|}{k} u_k(Y). \end{aligned}$$

Therefore, for every nonempty $Y \subseteq M$ we have

$$\sum_{1 \le k \le d} \binom{|Y|}{k} u_k(Y) = c(X \cup Y) - c(X) \ge W. \tag{4.1}$$

If, for some $Y \subseteq M$ we have

$$\sum_{1\le k\le d} \binom{|Y|}{k} u_k(Y) \ge 3W$$

then $c(X \cup Y) \ge 2W$ and we are done, so assume that for every $Y \subseteq M$ we have

$$\sum_{1\le k\le d} \binom{|Y|}{k} u_k(Y) < 3W. \tag{4.2}$$

Combining (4.1) and (4.2), we have for every nonempty $Y \subseteq M$

$$W \le \sum_{1\le k\le d} \binom{|Y|}{k} u_k(Y) < 3W.$$

Recall that for $j \ge k$, $u_k(M)$ is equal to $\mathrm{ave}_{Y\subseteq M,|Y|=j}\, u_k(Y)$. Therefore, for all j, $\binom{j}{k} u_k(M)$ is equal to $\mathrm{ave}_{Y\subseteq M,|Y|=j} \binom{|Y|}{k} u_k(Y)$. Applying the operator $\mathrm{ave}_{Y\subseteq M,|Y|=j}$ to all three terms in the inequality above, we obtain for $1 \le j \le m$

$$W \le \sum_{1\le k\le d} \binom{j}{k} u_k(M) < 3W.$$

Define a dth degree polynomial

$$p(z) = \sum_{1\le k\le d} u_k(M) \binom{z}{k}.$$

Then $p(0) = 0$, and $W \le p(z) < 3W$ for $z = 1, 2, \ldots, m$.

By Lemma 3.2, $d > \sqrt{(\sqrt{87} - 9)m}$. But $m = \lceil d^2/(\sqrt{87} - 9)\rceil$. This contradiction completes the proof that the set Y always exists. □

The following theorem is immediate.

Theorem 4.2. ODD-MAX-BIT *is not recognized by any family of perceptrons having size* $2^{n^{o(1)}}$, *order* $n^{o(1)}$, *and weight* $2^{n^{o(1)}}$.

5. An oracle for $\mathrm{P}^{\mathrm{NP}} \not\subseteq \mathrm{PP}$

The structure of the oracle construction is as follows. We define a test language

$$\text{ODD-MAX-ELEMENT}^A = \{0^n : \max(A \cap \{0,1\}^n) \text{ ends in a } 1\},$$

which clearly belongs to $\mathrm{P}^{\mathrm{NP}^A}$. The correspondence between oracle Turing machines and circuits is as in Furst *et al.* (1984). We will just sketch the basic idea. We construct an oracle A such that $\text{ODD-MAX-ELEMENT}^A \notin \mathrm{PP}^A$ by an initial segment argument. In order to defeat a polynomial-time probabilistic oracle TM M we choose m such that $A \cap \{0,1\}^m$ is as yet completely undefined. By convention we assume that a computation of M includes the oracle answers. Fix an input 0^m. Let $N = 2^m$. We will construct a perceptron C of size $2^{\text{polylog}\, N}$, weight 1, and order polylog N that simulates M. Its input consists of $N = 2^m$ bits: the characteristic sequence of $A \cap \{0,1\}^m$. For each of the $2^{\text{polylog}\, N}$ computations of M, we construct an AND-gate that verifies the oracle answers in the computation; each such AND-gate has fanin $m^{O(1)} = \text{polylog}\, N$. If a computation accepts, then we give its AND-gate weight $+1$; if it rejects, then we give its AND-gate weight -1. The perceptron C accepts the characteristic sequence of $A \cap \{0,1\}^m$ if and only if M accepts 0^m when using oracle A. We choose $A \cap \{0,1\}^m$ so that C accepts or rejects incorrectly. The construction fails only if there is a family of perceptrons having size $2^{\text{polylog}\, N}$, weight 1, and order polylog N which compute the predicate ODD-MAX-BIT.

We have shown that there is no family of perceptrons having having size $2^{\text{polylog}\, N}$, order polylog N, and weight 1 which compute N-bit instances of the predicate ODD-MAX-BIT. Thus the desired oracle must exist.

Let $\mathrm{P}^{\mathrm{NP}^A[f(n)]}$ denote the class of languages accepted by a deterministic polynomial-time bounded oracle Turing machine that is allowed at most $f(n)$ queries to an NP^A oracle. The following proposition follows from the standard connection, which we sketched above, between oracle Turing machine complexity and circuit complexity.

PROPOSITION 5.1.

- *If, for every oracle A,* ODD-MAX-ELEMENT^A *belongs to* PP^A, *then N-bit instances of* ODD-MAX-BIT *can be decided by perceptrons having size* $2^{\text{polylog}\, N}$, *weight* 1, *and order* polylog N.
- *If, for every oracle A,* $\mathrm{P}^{\mathrm{NP}^A[f(n)]}$ *belongs to* PP^A, *then* $(2^{f(n)} - 1)$*-bit instances of* ODD-MAX-BIT *can be decided by perceptrons having size* $2^{n^{O(1)}}$, *weight* 1, *and order* $n^{O(1)}$.

THEOREM 5.2. *If $f(n) \neq O(\log n)$ then there exists an oracle A such that* $\mathrm{P}^{\mathrm{NP}^A[f(n)]}$ *is not contained in* PP^A.

PROOF. Suppose that $\mathrm{P}^{\mathrm{NP}^A[f(n)]} \subseteq \mathrm{PP}^A$. Then $(2^{f(n)}-1)$-bit instances of ODD-MAX-BIT can be decided by perceptrons with size $2^{n^{O(1)}}$, weight 1, and order $n^{O(1)}$. Therefore they can be decided by perceptrons in clean form with size $s = 2^{n^{O(1)}}$, weight $w = 2^{n^{O(1)}}$, and order $d = n^{O(1)}$. By Lemma 4.1, $w \geq \frac{1}{s}2^{(2^{f(n)}-1)/8d^2}$. Therefore,

$$\begin{aligned}
\frac{1}{s}2^{(2^{f(n)}-2)/8d^2} &= 2^{n^{O(1)}}, \\
2^{(2^{f(n)}-2)/8d^2} &= s2^{n^{O(1)}}, \\
2^{(2^{f(n)}-2)/8d^2} &= 2^{n^{O(1)}}, \\
(2^{f(n)}-2)/8d^2 &= n^{O(1)}, \\
2^{f(n)}-2 &= n^{O(1)}8d^2, \\
2^{f(n)}-2 &= n^{O(1)}, \\
2^{f(n)} &= n^{O(1)}, \\
f(n) &= O(\log n). \ \Box
\end{aligned}$$

COROLLARY 5.3. *There exists an oracle A such that $\mathrm{P}^{\mathrm{NP}^A}$ is not contained in PP^A.*

The following corollary was obtained independently by Fu (1992).

COROLLARY 5.4 (INDEPENDENTLY BY FU). *There exists an oracle A such that PH^A is not contained in PP^A.*

Because $\Theta_2^p = \mathrm{P}^{\mathrm{NP}[\log]} \subseteq \mathrm{PP}$, we obtain the following old result of Buss & Hay (1991).

COROLLARY 5.5 (BUSS AND HAY). *There exists an oracle A such that $(\Theta_2^p)^A \subset (\Delta_2^p)^A$.*

Acknowledgements

I am grateful to Nick Reingold, Lane Hemachandra, Gerd Wechsung, Tomas Feder, Anna Karlin, Samuel Karlin, Bill Gasarch, John Gill, Joan Feigenbaum, Bin Fu, Jun Tarui, Lance Fortnow, Mario Szegedy, Noam Nisan, and Ted Rivlin for helpful discussions; to Allen Cohn, Anna Karlin, Jeff Westbrook, and Bob Floyd for their hospitality during my visit to Stanford; and to Martin Schultz, Michael Fischer, and Nick Reingold for helping to make that visit possible.

This work was supported in part by NSF grants CCR-8808949 and CCR-8958528.

References

E. Allender, A note on the power of threshold circuits. In *Proceedings of the 30th Ann. Symp. Found. Comput. Sci.*, 1989, 580–584.

E. Allender and U. Hertrampf, Depth reduction for circuits of unbounded fan-in. *Inform. and Comput.* **108** (1994). To appear.

J. Aspnes, R. Beigel, M. Furst, and S. Rudich, The expressive power of voting polynomials. In *Proceedings of the 23rd Ann. ACM Symp. Theor. Comput.*, 1991, 402–409. A revised version is to appear in *Combinatorica*.

R. Beigel and J. Tarui, On ACC. This Journal.

R. Beigel, L. A. Hemachandra, and G. Wechsung, Probabilistic polynomial time is closed under parity reductions. *Inform. Process. Lett.* **37**(2) (1991a), 91–94.

R. Beigel, N. Reingold, and D. Spielman, The perceptron strikes back. In *Proceedings of the 6th Ann. Conf. Structure in Complexity Theory*, 1991b, 286–291.

R. Beigel, N. Reingold, and D. Spielman, PP is closed under intersection. *J. Comput. System Sci.* **48** (1994). To appear.

S. R. Buss and L. E. Hay, On truth table reducibility to SAT. *Inform. and Comput.* **91**(1) (1991), 86–102.

E. W. Cheney, *Approximation Theory*. McGraw-Hill, 1966.

H. Ehlich and K. Zeller, Schwankung von polynomen zwischen gitterpunkten. *Math. Z.* **86** (1964), 41–44.

B. Fu, Separating PH from PP by relativization. *Acta Math. Sinica* **8**(3) (1992), 329–336.

M. Furst, J. B. Saxe, and M. Sipser, Parity, circuits, and the polynomial-time hierarchy. *Math. Systems Theory* **17**(1) (1984), 13–27.

T. Gundermann, N. Nasser, and G. Wechsung, A survey of counting classes. In *Proceedings of the 5th Ann. Conf. Structure in Complexity Theory*. IEEE Computer Society Press, 1990, 140–153.

L. Hemachandra and G. Wechsung, On the power of probabilistic polynomial time: $\mathrm{P}^{\mathrm{NP}[\log]} \subseteq \mathrm{PP}$. Technical Report CUCS-372-88, Columbia Dept. of Computer Science, New York, NY, 1988.

G. G. Lorentz, *Approximation of Functions*. Holt, Rinehart and Winston, New York, 1966.

M. L. Minsky and S. A. Papert, *Perceptrons*. MIT Press, Cambridge, MA, 1988. Expanded version of the original 1968 edition.

G. Pólya and G. Szegö, *Problems and Theorems in Analysis*. Springer-Verlag, Berlin, 1972.

T. J. Rivlin and E. W. Cheney, A comparison of uniform approximations on an interval and a finite subset thereof. *SIAM Numer. Anal.* **3**(2) (1966), 311–320.

J. Tarui, Probabilistic polynomials, AC^0 functions, and the polynomial-time hierarchy. *Theoret. Comput. Sci.* **113** (1993), 167–183.

S. Toda, PP is as hard as the polynomial-time hierarchy. *SIAM J. Comput.* **20**(5) (1991), 865–877.

S. Toda and M. Ogiwara, Counting classes are at least as hard as the polynomial-time hierarchy. *SIAM J. Comput.* **21**(2) (1992), 316–328.

A. C.-C. Yao, On ACC and threshold circuits. In *Proceedings of the 31st Ann. Symp. Found. Comput. Sci.*, 1990, 619–627.

Manuscript received August 20, 1992

Richard Beigel
Dept. of Computer Science
P.O. Box 208285
New Haven, CT 06520-8285
USA
`beigel-richard@cs.yale.edu`