Lower Bounds for Perceptrons Solving some Separation Problems and Oracle Separation of AM from PP

Nikolai K. Vereshchagin*
Department of Mathematical Logic
Moscow State University, Moscow 119899, Russia
E-mail: ver@mech.math.msu.su

Abstract

We prove that perceptrons separating Boolean matrices in which each row has a one from matrices in which many rows have no one must have either large total weight or large order. This result extends one-in-a-box theorem by Minsky and Papert [13] stating that perceptrons of small order cannot decide if each row of a given Boolean matrix has a one. As a consequence, we prove that $\mathrm{AM} \cap \text{co-AM} \not\subseteq \mathrm{PP}$ *under some oracle*[1]*. This contrasts the fact that* $\mathrm{MA} \subseteq \mathrm{PP}$ *under any oracle.*

1 Introduction

A *perceptron* is a depth-2 circuit with a threshold gate at the root and AND-gates at the remaining level. Each input of the threshold gate is labeled by an integer called its *weight*.

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. The *total weight* is the sum of absolute values of the weights on the inputs to its threshold gate. The perceptron outputs one on an input if the sum of weights on all true AND's is greater than its threshold.

Perceptrons have been studied by Minsky and Papert in [13]. Among their results we distinguish the following two theorems. The first one states that any perceptron computing parity function of n variables must have order at least n. The second theorem states that any perceptron recognizing whether all rows in a given Boolean matrix of size $n \times 4n^2$ contain 1 has order at least n (one-in-a-box theorem). Beigel in [3] constructed a boolean function of n variables that is computable by a perceptron having exponential total weight and order 1 and that is not computable by perceptrons having quasipolynomial ($2^{\mathrm{poly}(\log(n))}$) total weight and polylogarithmic ($\mathrm{poly}(\log(n))$) order. To be more precise, he proved lower bound $d^2 \log w = \Omega(n)$ for order d and total weight w of perceptrons computing that function.

We extend Minsky and Papert's one-in-a-box theorem in following direction. We consider separation problems instead of problems of predicate computation (i.e., decision problem). Let Π stand for the following separation problem: to separate Boolean matrices in which any row contains 1 from matrices in which many rows (e.g., a fraction 0.99 of all rows) contain zeros only. Obviously, any perceptron solving Π also recognizes if given matrix has a one in each row. Our theorem states that problem Π is not solvable by perceptrons having order $o(\sqrt{m})$ and total weight $2^{o(n)}$, where n is the number of rows and m is the number of columns (Theorem 3.1). This implies that perceptrons of polylogarithmic order and quasipolynomial total weight cannot solve Π. The preliminary version of this result appeared in [16]. The proof given there is more complicated compared with the present proof. It uses the Riesz' theorem in the functional analysis whereas the present proof uses the duality theorem in the linear programming instead. And the lower bound proven here is better than one proven in [16].

Quasipolynomial total weight and polylogarithmic order come up quite often by the following reason.

*This research was in part supported by a grant from the American Mathematical Society former Soviet Union Aid Fund, the grant MQT000 from the International Science Foundation, by a grant from "Cultural Initiative" foundation and by NSF grant CCR-8957604. Work done in part while visiting the University of Rochester.

[1] This result was presented at the Conference on Structure in Complexity Theory'92 [17].

When translating between nondeterministic Turing machine complexity and circuit complexity in the manner of Furst, Saxe, and Sipser [11], polynomial time translates into quasipolynomial total weight and polylogarithmic order. 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. (In other words, lower bounds for circuits translate into separations of Turing machine complexity classes via oracles.) In particular, PP-machines translate into perceptrons having polylogarithmic order and quasipolynomial total weight.

So due to Minsky and Papert's theorem on parity function we get that $\oplus\mathrm{P} \not\subseteq \mathrm{PP}$ relative to some oracle. The one-in-a-box theorem implies that $\mathrm{NP}^{\mathrm{NP}} \not\subseteq \mathrm{PP}$ under some oracle [10]. The above mentioned result by Beigel implies that $\mathrm{P}^{\mathrm{NP}} \not\subseteq \mathrm{PP}$ under some oracle [3]. Our result on the above defined separation problem Π shows that $\mathrm{AM} \not\subseteq \mathrm{PP}$ relative to some oracle. Some slight improvement of that result involves that $\mathrm{AM} \cap \text{co-AM} \not\subseteq \mathrm{PP}$ under some oracle [17].

The class PP is interesting by the following three reasons. First, this class has the following interpretation. Random input r of the probabilistic machine M that recognizes a language L can be regarded as a voter and the output $M(x,r)$ of M on the input word x and random input r can be regarded as the opinion of voter r about whether x is in L. From this point of view PP is the class of all languages L such that membership of x in L can be determined via election with $2^{\mathrm{poly}(|x|)}$ voters, every voter being uniformly polynomial time bounded.

Second, as shown in [15], the class PP proved to be surprisingly powerful: polynomial hierarchy PH is Turing reducible to PP.

Third, PP is closed under polynomial truth table reductions (see [5, 9]). Thus the class PP has a rather regular structure.

The paper is organized as follows. The next section contains two known theorems used in proving our lower bound. In Section 3, the bound itself is proved. In Section 4 we apply the method to prove that $\mathrm{AM} \cap \text{co-AM}$ is not contained in PP relative to some oracle.

2 Auxiliary theorems

We will use the following two well known results.

Theorem 2.1 *(Chernoff inequality [7]) Let $\xi_1, \ldots, \xi_n$ be independent random values in the set $\{0,1\}$ such that $\mathrm{Prob}[\xi_i = 1] = p$ for all i. Then for any $\delta \in (0; p(1-p)]$,*

$$\mathrm{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - p\right| \geq \delta\right] \leq 2e^{-\frac{\delta^2 n}{2p(1-p)}}.$$

The following lemma by Farkash is a version of the duality theorem in the theory of linear programming (see, for example [1]).

Lemma 2.1 *Let*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1t}x_t &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2t}x_t &= b_2 \\ &\ldots \\ a_{s1}x_1 + a_{s2}x_2 + \ldots + a_{st}x_t &= b_s \end{aligned}$$

be a linear system of equations, where $x_1, x_2, \ldots, x_t$ range the set of nonnegative reals. It has a solution if and only if there are no $y_1, y_2, \ldots, y_s$ such that

$$\begin{aligned} a_{11}y_1 + a_{21}y_2 + \ldots + a_{s1}y_s &\geq 0 \\ a_{12}y_1 + a_{22}y_2 + \ldots + a_{s2}y_s &\geq 0 \\ &\ldots \\ a_{1t}y_1 + a_{2t}y_2 + \ldots + a_{st}y_s &\geq 0 \\ b_1y_1 + b_2y_2 + \ldots + b_sy_s &< 0. \end{aligned}$$

3 The extension of one-in-a-box theorem

Definition 3.1 A perceptron is a depth-2 circuit having a threshold gate at the bottom and AND-gates at remaining level. Inputs of AND-gates are Boolean variables or their negations. Each AND-gate is labeled by a natural number called the weight of that AND-gate. The total weight of a perceptron is the sum of absolute values of weights on all its AND-gates. The order of a perceptron is the maximal fanin of its AND-gates.

Let P be a perceptron, and φ be an assignment of values to its variables. The *weight of* φ, written $W_P(\varphi)$, is the sum of weights on all AND's which are true on φ. The perceptron outputs 1 on input φ if $W_P(\varphi)$ is greater than the threshold of its threshold gate and 0 otherwise. The output value is denoted by $P(\varphi)$.

Let M be a Boolean matrix having n rows and m columns. Any matrix of such size can be defined in usual way by means of mn Boolean values. When we say that a perceptron P has such a matrix M as input

we mean that those Boolean values are assigned to its input variables. In this case we denote the output of P by $P(M)$.

A matrix is called *good* if its every row contains a one. A matrix is called *bad* if it is not good. Let q be a real in the segment [0;1]. A matrix is called *q-bad* if a fraction at least q of its rows contain no ones.

One-in-a-box theorem by Minsky and Papert states that perceptrons deciding whether the input Boolean matrix of size $n \times 4n^2$ is good must have order at least n.

We say that a perceptron P separates good matrices from q-bad matrices of size $n \times m$ if $P(M) = 1$ for any good matrix of size $n \times m$ and $P(M) = 0$ for any q-bad matrix of size $n \times m$. Note that for any m, n there is a perceptron of order m and total weight m separating good matrices from 1-bad matrices of size $n \times m$.

Theorem 3.1 *Let $1/2 \leq q < 1$, $\varepsilon = 1 - q$. If there exists a perceptron of order d and total weight w which separates good matrices from q-bad matrices of size $n \times m$, then*

$$d \geq \sqrt{(3/8)\varepsilon m} \tag{1}$$

or

$$w \geq 0.5e^{(2/15)\varepsilon n}. \tag{2}$$

Proof. Let m, n be integers. Denote the set of Boolean matrices having n rows and m columns by $\mathcal{M}$. By M_{ij} we denote the element of matrix M in ith row and jth column. Let μ be a probability distribution in the set $\mathcal{M}$. For a property S of matrices in $\mathcal{M}$ we denote by $\mathrm{Prob}_{\mu(M)}[S(M)]$ the probability that a matrix M taken at random with respect to μ satisfies S. Let d be an integer. Let us say that two probability distributions in the set $\mathcal{M}$, μ and ν, are *d-indistinguishable* if

$$\begin{aligned}&\mathrm{Prob}_{\mu(M)}[M_{i_1 j_1} = b_1, \ldots, M_{i_u j_u} = b_u]\\ &\quad = \quad \mathrm{Prob}_{\nu(M)}[M_{i_1 j_1} = b_1, \ldots, M_{i_u j_u} = b_u]\end{aligned}$$

for any sequence $\langle i_1, j_1\rangle, \ldots, \langle i_u, j_u\rangle$ of indices such that $u \leq d$ and for any sequence of bits $b_1, \ldots, b_u$.

The theorem easily follows from the following two lemmas.

Lemma 3.1 *If there exist d-indistinguishable probability distributions μ and ν in $\mathcal{M}$ such that, with respect to μ, a random matrix is good with probability 1 and with respect to ν a random matrix is q-bad with probability at least $1-p$, then any perceptron of order d separating good matrices from q-bad matrices has total weight at least p^{-1}.*

Lemma 3.2 *If $d < \sqrt{(3/8)\varepsilon m}$ and $0 < \varepsilon \leq 1/2$, then there exist d-indistinguishable probability distributions μ and ν in $\mathcal{M}$ such that, with respect to μ, a random matrix is good with probability 1 and with respect to ν a random matrix is $(1-\varepsilon)$-bad with probability at least $1 - 2e^{-(2/15)\varepsilon n}$.*

Let us prove first Lemma 3.1.

Proof of Lemma 3.1. Let d, μ, ν, q, p as in Lemma 3.1. Let P be perceptron of order d and total weight w separating good matrices from q-bad matrices in $\mathcal{M}$. Let E_μ and E_ν stand for the average with respect to distributions μ and ν. We claim that

$$\mathrm{E}_\mu W_P(M) = \mathrm{E}_\nu W_P(M). \tag{3}$$

Let us prove this claim. Let $C(M)$ stand for the Boolean function computed by an AND-gate C in P. Let l be the total number of AND-gates in P, ith gate being C_i and having weight w_i. Then $\mathrm{E}_\mu W_P(M) = \sum_{i=1}^{l} w_i \mathrm{E}_\mu C_i(M) = \sum_{i=1}^{l} w_i \mathrm{Prob}_{\mu(M)}[C_i(M) = 1]$. Therefore, it suffices to prove that $\mathrm{Prob}_{\mu(M)}[C(M) = 1] = \mathrm{Prob}_{\nu(M)}[C(M) = 1]$ for every AND-gate C in P. Let C be the conjunction $\bigwedge_{s=1}^{u}(M_{i_s j_s} = b_s)$, where b_s's are 0 or 1. Then

$$\begin{aligned}&\mathrm{Prob}_{\mu(M)}[C(M) = 1]\\ &\quad = \quad \mathrm{Prob}_{\mu(M)}[M_{i_1 j_1} = b_1, \ldots, M_{i_u j_u} = b_u].\end{aligned}$$

Thus d-indistinguishability of μ and ν implies (3).

Let t be the value of the threshold of the threshold-gate. Since a matrix taken at random with respect to μ is good with probability 1, we have $\mathrm{E}_\mu W_P(M) \geq t+1$. On the other hand, since a matrix taken at random with respect to ν is q-bad with probability at least $1 - p$, we have $\mathrm{E}_\nu W_P(M) \leq (1-p)t + pw$. Therefore, $t + 1 \leq (1-p)t + pw$. Thus, $t + 1 \leq t + pw$ which implies $1 \leq pw$.□

Proof of Lemma 3.2. Assume that the conditions of Lemma 3.2 are fulfilled. Let σ be a probability distribution in the segment $\{1, 2, \ldots, m\}$ and τ be a probability distribution in the segment $\{0, 1, 2, \ldots, m\}$ to be defined later. To take a matrix at random with respect to μ pick independent random $k_1, k_2, \ldots, k_n$ in $\{1, 2, \ldots, m\}$ with respect to σ. Then for each $i \leq n$ take a random string with exactly k_i ones as ith row of the matrix.

The distribution ν is defined in the same way but instead of the distribution σ we take the distribution τ. The distribution τ in turn will be obtained from some distribution ρ in the segment $\{1,2,\ldots,m\}$ by means of the following transformation:

$$\mathrm{Prob}_{\tau(k)}[k=a] = \begin{cases} 0.6\varepsilon \mathrm{Prob}_{\rho(k)}[k=a], & \text{if } a \in \{1,2,\ldots,m\}; \\ 1-0.6\varepsilon, & \text{if } a = 0. \end{cases}$$

Let us note that if ν is obtained from a distribution ρ in the way described above, then a matrix taken at random with respect to ν is $(1-\varepsilon)$-bad with probability at least $1-2e^{-(2/15)\varepsilon n}$. This is a direct consequence of the Chernoff inequality. Indeed, with respect to ν, each row of a random matrix has only zeros with probability $p = 1-0.6\varepsilon$. Take $\delta = 0.4\varepsilon$. Since we assume that $\varepsilon \leq 1/2$, δ and p satisfy the conditions of Theorem 2.1. By applying the inequality we conclude that a random matrix has less than $(1-\varepsilon)n$ zero rows with probability at most

$$\begin{aligned} & 2e^{-\frac{\delta^2 n}{2p(1-p)}} = 2e^{-\frac{(0.4\varepsilon)^2 n}{2\cdot 0.6\varepsilon(1-0.6\varepsilon)}} \\ = \; & 2e^{-\frac{0.16\varepsilon n}{1.2(1-0.6\varepsilon)}} \leq 2e^{-\frac{0.16\varepsilon n}{1.2}} = 2e^{-2/15\varepsilon n}. \end{aligned}$$

So we have to define probability distributions σ and ρ in the set $\{1,2,\ldots,m\}$. We will do that in such a way that distributions σ and τ will have the same first d moments, that is,

$$\mathrm{E}_{\sigma(k)} k^i = \mathrm{E}_{\tau(k)} k^i \tag{4}$$

for all $1 \leq i \leq d$. Let us prove that this implies the d-indistinguishability of μ and ν.

Indeed, we claim that the value

$$\mathrm{Prob}_{\mu(M)}[M_{i_1 j_1} = b_1, \ldots, M_{i_u j_u} = b_u]$$

is a polynomial in $\mathrm{E}_{\sigma(k)} k$, $\mathrm{E}_{\sigma(k)} k^2$, ..., $\mathrm{E}_{\sigma(k)} k^d$ for any sequence $\langle i_1, j_1\rangle, \ldots, \langle i_u, j_u\rangle$ of indices of length at most d and for any sequence of bits $b_1, \ldots, b_u$. Let us prove this claim.

Let $\langle i_1, j_1\rangle, \ldots, \langle i_u, j_u\rangle$ be a sequence of indices of length at most d and $b_1, \ldots, b_u$ be a sequence of bits. Denote for each $i \leq n$ by c_i the number of $l \in \{1,\ldots,u\}$ such that $i_l = i$ and $b_l = 1$ and by e_i the number of $l \in \{1,\ldots,u\}$ such that $i_l = i$ and $b_l = 0$. Then

$$\mathrm{Prob}_{\mu(M)}[M_{i_1 j_1} = b_1, \ldots, M_{i_u j_u} = b_u] = \prod_{i=1}^{n} \mathrm{E}_{\sigma(k)} \frac{k(k-1)\cdots(k-c_i+1)(m-k)(m-k-1)\cdots(m-k-e_i+1)}{m(m-1)\cdots(m-c_i-e_i+1)}$$

Evidently, for any i, the value $\mathrm{E}_{\sigma(k)} \frac{k(k-1)\cdots(k-c_i+1)(m-k)(m-k-1)\cdots(m-k-e_i+1)}{m(m-1)\cdots(m-c_i-e_i+1)}$ is a linear combination of values $\mathrm{E}_{\sigma(k)} k^r$, $r = 0, 1, \ldots, c_i + e_i$. As $c_i + e_i \leq d$, the claim is proved.

Recall that ν is obtained from τ in the same way as as μ is obtained from σ. Therefore, (4) implies the d-indistinguishability of μ and ν.

It is easy to see that

$$\mathrm{E}_{\tau(k)} k^i = 0.6\varepsilon \mathrm{E}_{\rho(k)} k^i$$

for any $i \geq 1$. Thus we have to prove that there are probability distributions σ and ρ in $\{1,2,\ldots,m\}$ satisfying (5). To do this we will apply the Farkash lemma because the dual problem is easier than the original one.

Sublemma 3.2.1 *The following conditions are equivalent:*

1) there are probability distributions σ and ρ in the segment $\{1,2,\ldots,m\}$ such that

$$\mathrm{E}_{\sigma(k)} k^i = 0.6\varepsilon \mathrm{E}_{\rho(k)} k^i, \text{ for } i = 1,2,\ldots,d, \tag{5}$$

2) there is no polynomial $p(x)$ of degree at most d such that

$$-p(0)\frac{1-0.6\varepsilon}{0.6\varepsilon} < p(j) \leq 0 \text{ for } j = 1,2,\ldots,m.$$

Proof. The condition 1) is equivalent to the existence of a nonnegative solution of the system

$$\begin{aligned} 0.6\varepsilon(1^1 x_1 + 2^1 x_2 + \ldots + m^1 x_m) - (1^1 x_{m+1} + 2^1 x_{m+2} + \ldots + m^1 x_{2m}) &= 0 \\ 0.6\varepsilon(1^2 x_1 + 2^2 x_2 + \ldots + m^2 x_m) - (1^2 x_{m+1} + 2^2 x_{m+2} + \ldots + m^2 x_{2m}) &= 0 \\ \ldots & \\ 0.6\varepsilon(1^d x_1 + 2^d x_2 + \ldots + m^d x_m) - (1^d x_{m+1} + 2^d x_{m+2} + \ldots + m^d x_{2m}) &= 0 \\ x_1 + x_2 + \ldots + x_m &= 1 \\ x_{m+1} + x_{m+2} + \ldots + x_{2m} &= 1. \end{aligned}$$

By Farkash Lemma this means that there are no $p_1, p_2, \ldots, p_d, y, z$ such that

$$0.6\varepsilon \sum_{i=1}^{d} p_i j^i + y \geq 0 \text{ for } j = 1,2,\ldots,m, \tag{6}$$

$$-\sum_{i=1}^{d} p_i j^i + z \geq 0 \text{ for } j = 1,2,\ldots,m, \tag{7}$$

$$y + z < 0. \tag{8}$$

We interpret the numbers $-z, p_1, p_2, \ldots, p_d$ as coefficients of the polynomial $p(u) = -z + p_1 u + \cdots + p_d u^d$.

Then (7) means that $p(j) \leq 0$ for all $j = 1, 2, \ldots, m$. The existence of y such that (6) and (8) hold means that $0.6\varepsilon(p(j) - p(0)) + p(0) > 0$ for all $j = 1, 2, \ldots, m$. The last inequality can be rewritten as $-p(0)\frac{1-0.6\varepsilon}{0.6\varepsilon} < p(j)$, so 1) and 2) are equivalent.□

So it remains to prove that there is no polynomial $p(x)$ of small degree such that $-p(0)\frac{1-0.6\varepsilon}{0.6\varepsilon} < p(j) \leq 0$ for all $j = 1, 2, \ldots, m$. We will use the following theorem by A. A. Markov. Let for a polynomial $P(x)$, $\|P\|$ stand for the maximum absolute value of P on $[-1; 1]$. Let $P(x)$ be a polynomial of degree d. Markov's theorem (see, for example, [6, 12, 14]) says that $\|P'\| \leq d^2\|P\|$.

The following simplified version of Ehlich and Zeller lemma [8] is an easy corollary of the Markov's theorem. Let

$$Y = \{-1, -1 + 2/m, -1 + 4/m, \ldots, 1 - 2/m, 1\}$$

and let $\|P_Y\|$ stand for the maximum absolute value of $P(x)$ on Y. Then

$$\|P\|\left(1 - \frac{2d^2}{m}\right) \leq \|P_Y\|. \tag{9}$$

Indeed, by Lagrange theorem and by Markov's theorem we have

$$\|P\| - \|P_Y\| \leq \|P'\|\frac{2}{m} \leq \|P\|\frac{2d^2}{m}.$$

This easily implies (9).

So suppose there is a polynomial $p(x)$ of degree d such that $-p(0)\frac{1-0.6\varepsilon}{0.6\varepsilon} < p(j) \leq 0$ for all $j = 1, 2, \ldots, m$. We have to prove that $d \geq \sqrt{(3/8)\varepsilon m}$.

Observe that $p(0)$ is positive. Denote $\frac{1-0.6\varepsilon}{0.6\varepsilon}$ by α. Let $P(x) = p(\frac{1}{2}m(x+1)) + \frac{\alpha-1}{2}p(0)$. Then $\|P_Y\| = \frac{\alpha+1}{2}p(0)$, therefore

$$(1 - 2d^2/m)\|P\| \leq \frac{\alpha+1}{2}p(0). \tag{10}$$

On the other hand, since $|P(-1 + 2/m) - P(-1)| = |p(1) - p(0)| \geq p(0)$, Lagrange theorem implies that $\|P'\| \geq p(0)m/2$. Therefore, by Markov's theorem

$$\|P\| \geq p(0)\frac{m}{2d^2}. \tag{11}$$

Combining (10) and (11) we obtain

$$\left(1 - \frac{2d^2}{m}\right)p(0)\frac{m}{2d^2} \leq \frac{\alpha+1}{2}p(0).$$

Therefore

$$\frac{d^2}{m} \geq \frac{1}{\alpha+3} = \frac{0.6\varepsilon}{1+1.2\varepsilon} \geq \frac{0.6\varepsilon}{1.6} = \frac{3}{8}\varepsilon.\square$$

4 Applying lower bounds for perceptrons to separation of AM ∩ co-AM from PP

We consider languages over the binary alphabet $\mathbf{B} = \{0, 1\}$. The set of all binary words of length n is denoted by $\mathbf{B}^n$. Functions with binary values are called predicates. All Turing machines output 0, 1.

Definition 4.1 A language L belongs to PP iff there is a polynomial time probabilistic Turing machine T such that $x \in L \Leftrightarrow \mathrm{Prob}[T(x) = 1] > 1/2$.

Definition 4.2 $L \in$ MA iff there are a polynomial p and polynomial time computable predicate $Q(x, r, s)$ such that

$$\begin{aligned} x \in L &\Rightarrow \exists s \in \mathbf{B}^{p(|x|)}\ \mathrm{Prob}_r[Q(x, r, s)] > 2/3, \\ x \notin L &\Rightarrow \forall s \in \mathbf{B}^{p(|x|)}\ \mathrm{Prob}_r[Q(x, r, s)] < 1/3, \end{aligned}$$

where probability is with respect uniform distribution in $\mathbf{B}^{p(|x|)}$.

Definition 4.3 $L \in$ AM iff there exist a polynomial p and polynomial computable predicate $Q(x, r, s)$ such that

$$\begin{aligned} x \in L &\Rightarrow \mathrm{Prob}_r[\exists s \in \mathbf{B}^{p(|x|)} Q(x, r, s)] > 2/3, \\ x \notin L &\Rightarrow \mathrm{Prob}_r[\exists s \in \mathbf{B}^{p(|x|)} Q(x, r, s)] < 1/3, \end{aligned}$$

where probability is with respect to uniform distribution in $\mathbf{B}^{p(|x|)}$.

In the paper [2], it is proven that MA $\subseteq$ AM. The proof is relativizable. In [17], it is proven that MA $\subseteq$ PP. The proof relativizes, too.

The lower bound in Theorem 3.1 suffices to construct an oracle under which AM $\not\subseteq$ PP. To construct an oracle under which AM ∩ co-AM $\not\subseteq$ PP we need a lower bound for perceptrons solving another separation problem. Let us define it.

Let $\mathcal{M}_n$ stand for the family of Boolean matrices of size $n \times n$ and let $\mathcal{N}_n = \mathcal{M}_n \times \mathcal{M}_n$. Let $D = \langle M_0, M_1\rangle$ be a pair of matrices in $\mathcal{N}_n$. Say that D is of type 0 [type 1] if all rows in M_0 [M_1] contain 1 and at least 2/3 of rows in M_1 [M_0] contain no 1.

Theorem 4.1 *There exists $\delta > 0$ such that the following holds for large enough n. If there exists a perceptron of order d and total weight w separating elements in $\mathcal{N}_n$ having type 0 from elements in $\mathcal{N}_n$ having type 1, then $d \geq \delta n^{1/2}$ or $w \geq 2^{\delta n}$.*

Proof. Let n be an integer. Let P be a perceptron of order d and total weight w separating elements in $\mathcal{N}_n$ having type 0 from elements in $\mathcal{N}_n$ having type

1. Let us use Lemma 3.2. Let $\varepsilon = 1/3$, $\delta = 0.01$ and $m = n$. Suppose that $d \leq \delta n^{1/2}$. Then the conditions of Lemma 3.2 are fulfilled. Therefore there exist probability distributions μ and ν in $\mathcal{M}_n$ such that the following hold:

1) a matrix M taken at random with respect to μ is good with probability 1;

2) a matrix M taken at random with respect to ν is not 2/3-bad with probability at most $2e^{-(2/15)(1/3)n}$;

3) μ and ν are d-indistinguishable.

Denote $2e^{-(2/15)(1/3)n}$ by p. Consider the following probability distributions κ and λ in $\mathcal{N}_n$. To produce a random pair $\langle M_0, M_1\rangle$ of matrices with respect to κ take M_0 at random with respect to μ and take M_1 at random with respect to ν. To produce a random pair $\langle M_0, M_1\rangle$ of matrices with respect to λ take M_0 at random with respect to ν and take M_1 at random with respect to μ. Then

$$\mathrm{Prob}_{\kappa(D)}[D \text{ has type } 0\ ] \geq 1 - 2p,$$
$$\mathrm{Prob}_{\lambda(D)}[D \text{ has type } 1\ ] \geq 1 - 2p.$$

As we have shown above, 3) implies that

$$\mathrm{E}_\kappa W_P(D) = \mathrm{E}_\lambda W_P(D).$$

Let t be the value of the threshold of the threshold-gate. Obviously we can assume that $|t| < w$. We have $\mathrm{E}_\kappa W_P(D) \geq (1-2p)(t+1) - 2pw$ and $\mathrm{E}_\lambda W_P(D) \leq (1-2p)t + 2pw$. Therefore, $(1-2p)(t+1) - 2pw \leq (1-2p)t + 2pw$, which implies the inequality $w \geq 1/(6p) = (1/12)e^{(2/15)(1/3)n} \geq e^{\delta n}$ for large enough n.□

Theorem 4.2 [17] *There is an oracle A such that* $\mathrm{AM}^A \cap \text{co-AM}^A \not\subseteq \mathrm{PP}^A$.

Proof. Let A be an oracle and let $j \in \mathbf{N}$. We will consider the value of A on words of length $2j+1$ as a pair of Boolean matrices of size $2^j \times 2^j$. Denote that pair by A_j. Let us say that A_j is *correct* if A_j has type 1 or type 0. Associate with any oracle A the language $L(A) = \{1^j \mid A_j \text{ has type } 0\}$. We will construct an oracle A such that A_j is correct for all $j \in \mathbf{N}$ and $L(A) \notin \mathrm{PP}^A$. From the former condition we can easily deduce that $L(A)$ is in $\mathrm{AM}^A \cap \text{co-AM}^A$.

To ensure $L(A) \notin \mathrm{PP}^A$ let us enumerate all polynomial probabilistic machines and denote i-th machine by PP_i. Define first A in such a way that A_j is correct for all $j \in \mathbf{N}$. We will make steps with numbers 0, 1, 2, ... On step i we will ensure that $L(A)$ differs from the language recognized by PP_i^A. To this end we will change the value of A on finite number of words in such a way that for some $j \in \mathbf{N}$

$$1^j \in L(A) \not\Leftrightarrow \mathrm{Prob}[PP_i^A(1^j) = 1] > 1/2. \qquad (12)$$

After changing we will fix the value of A on all words which the truth value of (12) depends on. This means that on later steps we will not change the value of A on those words.

Let us describe ith step. Choose j such that no value of A (the oracle constructed on $(i-1)$th step) on words of length $2j+1$ is fixed and sufficiently large (how large should be j we shall see later). Let $n = 2^j$. Then A_j is in $\mathcal{N}_n$. For any $D \in \mathcal{N}_n$ denote by $A[D]$ the oracle obtained from A by replacing A_j with D. Let us prove that there is a correct $D \in \mathcal{N}_n$ such that (12) is true for $A[D]$.

Let d be maximal number of queries that PP_i can make on input 1^j. Let q be the number of random strings used by PP_i on input 1^j. Evidently $d \leq \mathrm{poly}(j) = \mathrm{poly}(\log n)$ and $q \leq 2^{\mathrm{poly}(j)} = 2^{\mathrm{poly}(\log n)}$. Let us construct a perceptron P of order d and of total weight $w = 2^d q$ such that

$$P(D) = 1 \Leftrightarrow \mathrm{Prob}[PP_i^{A[D]}(1^j) = 1] > 1/2. \qquad (13)$$

Let r be a random string used by PP_i on input 1^j and $v = v(1)v(2)\cdots v(d)$ be a binary string of length d such that PP_i accepts the input 1^j provided oracle answers to queries are respectively $v(1), \ldots, v(d)$. Associate with every such pair $\langle v, r\rangle$ the following AND-gate C. Denote by $u_1, \ldots, u_d$ the questions to oracle made by PP_i on input 1^j with random string r provided the oracle answers are respectively $v(1), \ldots, v(d)$. On assignment $D \in \mathcal{N}_n$ gate C produces 1 if $u_k \in A[D] \Leftrightarrow v(k) = 1$ for all $k \in \{1, 2, \ldots, d\}$. Let weight on C be 1. Let the threshold of perceptron P be $q/2$. It is easy to verify that

$$W_P(D) = q\mathrm{Prob}[PP_i^{A[D]}(1^j) = 1],$$

which implies (13).

Obviously, the order of P is $d = \mathrm{poly}(j) = \mathrm{poly}(\log n)$ and the total weight of P is $2^d q = 2^{\mathrm{poly}(j)} = 2^{\mathrm{poly}(\log n)}$. By Theorem 4.1 if j is large enough then P cannot separate pairs of type 0 from pairs of type 1 in $\mathcal{N}_n$. Thus there exists a correct $D \in \mathcal{N}_n$ such that

$$D \text{ has type } 0 \not\Leftrightarrow \mathrm{Prob}[PP_i^{A[D]}(1^j) = 1] > 1/2$$

and we are done.□

5 Conclusion

Theorem 3.1 states that any perceptron of small total weight separating good matrices from q-bad ones has large order. This leaves the possibility that perceptrons of small order and large total weight can separate good matrices from q-bad ones (for some $q < 1$). Since one-in-a-box theorem involves restrictions on the order only, the better extension of one-in-a-box theorem would be a theorem stating that perceptrons of small order and arbitrary total weight can do the job. Recently Beigel obtained such a lower bound [4]. He proved that perceptrons separating good matrices of size $n \times n$ from q-bad ones must have order superpolylogarithmic in n (for any fixed $q < 1$). The problem if perceptrons of order $n^{o(1)}$ can do that job remains open. Note that Beigel's bound also suffices to separate AM from PP via oracles, and since his proof is rather simple this yields an alternative simplification of oracle separation of AM from PP.

Acknowledgments

The author is sincerely grateful to Vladimir Borisenko, Konstantin Gorbunov, Frederic Green, Lane Hemaspaandra, Andrey Muchnik, Alexander Razborov, Alexander Shen and Yuri Tyurin.

References

[1] S. A. Ashmanov. *Linear Programming.* Nauka, Moscow, 1981. (in Russian).

[2] L. Babai. Trading group theory for randomness. In *17th Annual ACM Symposium on Theory of Computing*, pages 421–429, 1985.

[3] R. Beigel. Perceptrons, PP, and the polynomial time hierarchy. In *7th Annual Conference on Structure in Complexity Theory*, pages 14–19, Boston, MA, July 1992.

[4] R. Beigel. Personal communication. 1994.

[5] R. Beigel, N. Reingold, and D. Spielman. PP is closed under intersection. In *23th Annual ACM Symposium on Theory of Computing*, pages 1–9, 1991.

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

[7] H. Chernoff. A measure of assymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[8] H. Ehlich and K. Zeller. Schwankung von Polynomen zwischen Gitterpunkten. *Mathematische Zeitschrift*, 86:41–44, 1964.

[9] L. Fortnow and N. Reingold. PP is closed under truth table reductions. In *6th Annual Conference on Structure in Complexity Theory*, pages 13–15, 1991.

[10] B. Fu. Separating PH from PP by relativisation. Preprint, 1990.

[11] N. Furst, J. Saxe, and M. Sipser. Parity, circuits and the polynomial time hierarchy. *Mathematical Systems Theory*, 17:13–27, 1984.

[12] G. G. Lorentz. *Approximation of Functions.* Hort, Rinehart and Winston, New York, 1966.

[13] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1988. (Expanded edition, first edition appeared in 1967.).

[14] G. Pólya and G. Szegö. *Problems and Theorems in Analysis.* Springer Verlag, 1972.

[15] S. Toda. On the computational power of PP and $\oplus$P. In *30th Annual IEEE Symposium on Foundation of Computer Science*, pages 514–519, 1989.

[16] N. K. Vereshchagin. Lower bounds for perceptrons solving some separation problems and oracle separation of AM from PP. Technical Report 498, Computer Science Department, University of Rochester, 1988.

[17] N. K. Vereshchagin. On the power of PP. In *7th Annual Conference on Structure in Complexity Theory*, pages 138–143, 1992.