ORACLE SEPARATION OF COMPLEXITY CLASSES AND LOWER BOUNDS FOR PERCEPTRONS SOLVING SEPARATION PROBLEMS

Nikolai K. Vereshchagin

Abstract. In the first part of the paper we prove that, relative to a random oracle, the class NP has infinite sets having no infinite Co-NP-subsets (Co-NP-immune sets). In the second part we prove that perceptrons separating Boolean matrices in which each row has a one from matrices in which many rows (say 99% of them) have no ones must have large size or large order. This result partially strengthens one-in-a-box theorem by Minsky and Papert [16] stating that perceptrons of small order cannot decide if each row of 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. Introduction

We study problems of oracle separations of complexity classes and related combinatorial problems.

Many important problems in Complexity theory remain open. The most known one is if the classes P and NP coincide. It is also unknown if NP = Co-NP and if NP $\cap$ Co-NP = P.

In the paper [3], it was shown that all these problems have no relativizable solutions. More exactly, oracles A and B were constructed such that $\mathrm{P}^A = \mathrm{NP}^A$ (and, therefore, $\mathrm{P}^A = \mathrm{NP}^A \cap \text{Co-NP}^A$) and $\mathrm{NP}^B \neq \text{Co-NP}^B$ (and, therefore, $\mathrm{P}^B \neq \mathrm{NP}^B$). By using the same technique one can construct an oracle C for which $\mathrm{NP}^C \cap \text{Co-NP}^C \neq \mathrm{P}^C$.

As the relationships between the classes Co-NP^A, NP^A and P^A depend on the oracle A, it is natural to ask what happens for "typical" oracle A. This question is studied in the first part of the present paper.

A possible refinement of the notion of typicalness is randomness with respect to the uniform measure. In the present paper we just study the relationships between the classes NP^A, Co-NP^A and P^A for oracle A being random with respect to the uniform measure. More exactly, we say that assertion $S(A)$ holds for random A,

1991 *Mathematics Subject Classification.* Primary 68Q15.

Translated from Izvestija Rossijskoj Akademii Nauk, Vol. ?? (1995), No. 6, pp. ??–??.

Received November 28, 1994

Preliminary versions of parts of the present papers were published in *Third Israel Symposium on Theory of Computing and Systems (ISTCS'95)*[23,24].

This research was in part supported by a grant from the American Mathematical Society, 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.

or for almost all A, if the uniform measure of the set $\{A \mid S(A)\}$ is equal to 1. In the Introduction, we consider only properties $S(A)$ satisfying the two following conditions: the set $\{A \mid S(A)\}$ is measurable and $S(A)$ is stable with respect to any changing the value of A on a finite set of arguments. Due to 0-1-law of A. N. Kolmogorov, for such properties either $S(A)$ holds for random A, or $\neg S(A)$ holds for random A.

The study of Complexity theory relative to a random oracle began in [6], where it was proved that $\mathrm{P}^A \neq \mathrm{NP}^A \neq \text{Co-NP}^A$ for random A. It was also proved in [6] that for random A there exists an infinite NP^A-set which is P^A-immune, that is, has no infinite P^A-subsets. In [22], it was proved that both NP-sets and Co-NP-sets are P-inseparable relative to a random oracle. It was proved also in that paper that the class Co-NP has an infinite NP-immune set relative to a random oracle.

Let us look at these results from the point of view of analogy between Recursion theory and Complexity theory. According to this analogy P-sets correspond to decidable sets, NP-sets correspond to recursively enumerable sets and Co-NP-sets correspond to complements of recursively enumerable sets. More exactly, we will consider the complexity theory relativized with a random oracle. Thus, in our case, decidable sets correspond to P^A-sets, recursively enumerable sets correspond to NP^A-sets and complements of recursively enumerable sets correspond to Co-NP^A-sets, where A is a random oracle. As we have said, in Complexity theory the analogs of the following theorems are true: the theorem on the existence of recursively enumerable undecidable set, the theorem on the existence of recursively enumerable set whose complement is not recursively enumerable, the theorem on the existence of recursively enumerable inseparable sets, the theorem on the existence of a simple set (recursively enumerable set whose complement is infinite but has no infinite recursively enumerable subsets). The analogs of the following theorems are false: the theorem saying that any infinite recursively enumerable set has an infinite decidable subset and the theorem on separability of sets having recursively enumerable complements.

It remains an open problem at present if the analog of Post's theorem is true (if both a set and its complement are recursively enumerable, then the set is decidable), i.e., it is unknown if $\mathrm{NP}^A \cap \text{Co-NP}^A = \mathrm{P}^A$ is true for random A. As mentioned in [[BI87]], if $\mathrm{NP}^A \cap \text{Co-NP}^A = \mathrm{P}^A$ for random A, then $\mathrm{AM} \cap \text{Co-AM} = \mathrm{BPP}$, in particular, the problem of graph isomorphism is in BPP. Thus, it cannot be hoped to prove that $\mathrm{NP}^A \cap \text{Co-NP}^A = \mathrm{P}^A$ for random A. No absolute consequences of the assertion "$\mathrm{NP}^A \cap \text{Co-NP}^A \neq \mathrm{P}^A$ for random A" are known; so one may hope to prove that $\mathrm{NP}^A \cap \text{Co-NP}^A \neq \mathrm{P}^A$ for random A.

It remained an open problem if the class NP has a Co-NP-immune set relative to a random oracle. Some advance in solving this problem was made in [15], where it was proved that relative to a random oracle there exists an infinite NP-set all of whose Co-NP-subsets are "thin" in a sense.

In the present paper we solve this problem: we prove that the class NP has Co-NP-immune sets relative to a random oracle. Moreover, we prove that, relative to a random oracle, there exist a set $L \in \mathrm{P}$ and an NP-set $L_1 \subseteq L$ such that L_1 is Co-NP-immune and $L \setminus L_1$ is NP-immune. The second theorem implies the theorem in [22] stating that the class Co-NP has NP-immune sets relative to a random oracle.

All the result cited above are presented in the Table 1. The analogous results on generic oracle are also shown there.

TABLE 1

	Relative to random oracle	Relative to generic oracle	Recursion theory
$P \neq NP$	+	+	+
$NP \neq$ Co-NP	+	+	+
NP-sets are P-inseparable	+	?	+
Co-NP-sets are P-inseparable	+	+	−
$P = NP \cap$ Co-NP	?	?	+
NP has P-immune sets	+	−	−
NP has Co-NP-immune sets	+	−	−
Co-NP has NP-immune sets	+	+	+

In the second part of the present paper we prove that relative to an oracle the class AM is not included in the class PP and we study the complexity of the related separation problem for perceptrons (the relevant definitions will be given later).

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 [16]. 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 [4] 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 (poly(log(n))) order. To be more precise, he proved the 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 recognizing whether given matrix has a one in each row also solves Π. 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 5). This implies that perceptrons of polylogarithmic order and quasipolynomial total weight cannot solve Π. The preliminary version of this result appeared in [20]. 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 [20].

Quasipolynomial total weight and polylogarithmic order come up quite often by the following reason. When translating between nondeterministic Turing machine complexity and circuit complexity in the manner of Furst, Saxe, and Sipser [13], 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 [12]. The above mentioned result by Beigel implies that $\mathrm{P}^{\mathrm{NP}} \not\subseteq \mathrm{PP}$ under some oracle [4]. 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 [21].

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 [19], the class PP proved to be surprisingly powerful: the polynomial hierarchy PH is Turing reducible to PP.

Third, PP is closed under polynomial truth table reductions (see [5,11]). Thus the class PP has a rather regular structure.

2. Definitions

A *language* is a subset of the set $\mathbf{B}^*$ of all the words over the binary alphabet $\mathbf{B} = \{0,1\}$. An *oracle* is function from the set $\mathbf{B}^*$ into the set $\mathbf{B}$.

An *oracle machine* is a Turing machine having an extra tape called *oracle tape*; this tape has a read/write head. That head can write only zeros and ones. To run an oracle machine on an input we must supply it with an oracle. Let A be an oracle. Then machine works as usual two tape Turing machine with one exception. If oracle machine gets into a certain state, then the word u written on oracle tape (starting from the first cell up to the cell where the head is now) is considered as a question to the oracle. In this case oracle provides its answer $A(u)$ in the cell viewed by the head. The time needed for oracle to provide its answer is assumed to be 1.

Let A be an oracle. The class P^A consists of all languages L such that there exists a deterministic polynomial time oracle Turing machine M such that $x \in L \Leftrightarrow M^A(x) = 1$, where $M^A(x)$ stands for the output of M on input x with oracle A. The class NP^A consists of all languages L such that there exist a deterministic

polynomial time oracle Turing machine M and a polynomial p such that

$$x \in L \Leftrightarrow \exists y \in \mathbf{B}^* \ \big(|y| = p(|x|) \ \& \ M^A(x,y) = 1\big),$$

where $|u|$ denotes the length of word u. The pair $N = \langle M, p\rangle$ is called *nondeterministic* machine and we set

$$N^A(x) = \begin{cases} 1, & \text{if } \exists y \in \mathbf{B}^* \\ & (|y| = p(|x|) \ \& \ M^A(x,y) = 1); \\ 0, & \text{otherwise.} \end{cases}$$

A possible computation of machine N on input x is a computation of M on any input of the form $\langle x, y\rangle$, where $|y| = p(|x|)$. Denote by L_{N^A} the language $\{x \in \mathbf{B}^* \mid N^A(x) = 1\}$.

Definition 1. $L \in$ AM if there exist a polynomial p and polynomial time computable predicate $Q(x,r,s)$ such that

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

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

Definition 2. $L \in$ MA if there exist a polynomial p and polynomial time computable predicate $Q(x,r,s)$ such that

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

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

Definition 3. A language L belongs to PP if there is a polynomial time probabilistic Turing machine T such that $x \in L \iff \mathrm{Prob}\,[T(x) = 1] > 1/2$.

In the paper [2] it is proven that MA $\subseteq$ AM. The proof relativizes. In [21] it is proven that MA $\subseteq$ PP. The proof relativizes, too.

Let L_1, L_2 and L be languages. We say that L *separates* L_1 from L_2 if $L_1 \subseteq L$ and $L_2 \subseteq \mathbf{B}^* \setminus L$. Let $\mathcal{C}$ and $\mathcal{C}'$ be families of languages. Say that $\mathcal{C}$-languages are $\mathcal{C}'$-separable if for any two disjoint languages L_1 and L_2 in $\mathcal{C}$ there exists a language L in $\mathcal{C}'$ which separates L_1 from L_2. A language is called $\mathcal{C}$-immune if it is infinite and has no infinite subsets in $\mathcal{C}$.

Let $S(A)$ be a property of an oracle A. We say that $S(A)$ holds for random A, or for almost all A, if the uniform measure of the set $\{A \mid S(A)\}$ is equal to 1.

3. The results on a random oracle

Theorem 1. *Relative to a random oracle, there are* $L \in$ P *and* NP*-set* $L_1 \subseteq L$ *such that both* L_1 *and* $L \setminus L_1$ *are infinite,* L_1 *has no infinite* Co-NP*-subsets and* $L \setminus L_1$ *has no infinite* NP*-subsets.*

Proof. Define the sequence $\{t_i\}$ of integers by induction: $t_0 = 1$, $t_{i+1} = 2^{t_i}$. Let i be a natural number. An *i-block* is a set of the form

$$B_w = \{wv \mid v \in \mathbf{B}^*, \ |v| = \log_2 t_i\},$$

where w is a binary word of length t_i. Thus, each i-block consists of t_i words (of length $t_i + \log_2 t_i$).

Let

$$L = \{1^{t_i} \mid \ i \in \mathbf{N}\}.$$

We say that an oracle A is *identically zero in block* B if $\forall u \in B \ A(u) = 0$.

For any oracle A let

$$L_1^A = \{1^{t_i} \mid \ i \in \mathbf{N} \text{ and } A \text{ is identically zero in some } i\text{-block}\},$$
$$L_0^A = L \setminus L_1^A$$

Obviously, L is in P^A and L_1^A is in NP^A for any A.

Both languages L_0^A and L_1^A are infinite for almost all A, because

$$\mathrm{Prob}\,[1^{t_i} \in L_0^A] = 1 - \mathrm{Prob}\,[1^{t_i} \in L_1^A] = (1 - 2^{-t_i})^{2^{t_i}} \to e^{-1},$$

as $i \to \infty$.

Let us prove that

1) L_0^A has no infinite NP^A-subsets for almost all A and
2) L_1^A has no infinite $\mathrm{Co\text{-}NP}^A$-subsets for almost all A.

We claim that instead of 1) it suffices to prove the following assertion

1′) there exists $c < 1$ such that for any nondeterministic machine N the probability of event "L_{N^A} is infinite and $L_{N^A} \subseteq L_0^A$" is less than c,

and instead of 2) it suffices to prove the following assertion

2′) there exists $c < 1$ such that for any nondeterministic machine N the probability of event "$\mathbf{B}^* \setminus L_{N^A}$ is infinite and $\mathbf{B}^* \setminus L_{N^A} \subseteq L_1^A$" is less than c.

Let us prove first the implication 1′) $\Rightarrow$ 1). Assume that 1′) is true but the probability of event "L_0^A has an infinite NP^A-subset" is positive. Then there exists a nondeterministic machine N such that the probability of event "L_{N^A} is infinite and $L_{N^A} \subseteq L_0^A$ " is positive. Call any set of oracles of the form

$$\{A \mid A(w_1) = b_1, \ldots, A(w_j) = b_j\}$$

where $w_1, \ldots, w_j \in \mathbf{B}^*, b_1, \ldots, b_j \in \mathbf{B}$, an *interval*. A simple theorem in the measure theory states that if a set of oracles S has positive measure and $c < 1$, then there exists an interval Γ such that

$$\mathrm{Prob}\,[A \in S \mid A \in \Gamma] > c. \tag{2}$$

Let us apply that theorem to the set

$$S = \{A \mid L_{N^A} \text{ is infinite and } L_{N^A} \subseteq L_0^A\}.$$

Thus, there exists an interval Γ such that (2) is true. Let $w_1, \ldots, w_j \in \mathbf{B}^*, b_1, \ldots, b_j \in \mathbf{B}$ be such that $\Gamma = \{A \mid A(w_1) = b_1, \ldots, A(w_j) = b_j\}$. For any oracle A let us define the new oracle A' as follows:

$$A'(u) = \begin{cases} b_l, & \text{if } u = w_l, \text{ where } l \in \{1, \ldots, j\}; \\ A(u), & \text{if } u \notin \{w_1, \ldots, w_j\}. \end{cases}$$

It is easy to see that

$$\text{Prob}\,[A' \in S] = \text{Prob}\,[A \in S \mid A \in \Gamma] > c.$$

Let $k = \max_{m \le j} |w_m|$. We can easily construct a nondeterministic machine N_1 such that

$$N_1^A(x) = \begin{cases} 0, & \text{if } |x| + \log_2 |x| \le k; \\ N^{A'}(x), & \text{otherwise.} \end{cases}$$

We claim that if $A' \in S$, then $L_{N_1^A}$ is infinite and $L_{N_1^A} \subseteq L_0^A$. Indeed, let A' be in S, that is, $L_{N^{A'}}$ is infinite and $L_{N^{A'}} \subseteq L_0^{A'}$. Then, obviously, $L_{N_1^A}$ is infinite. And $L_{N_1^A} \subseteq L_0^A$, as if $N_1^A(x) = 1$, then $|x| + \log_2 |x| > k$ and $N^{A'}(x) = 1$, consequently, $x \in L_0^{A'}$, which implies that $x \in L_0^A$ (because $|x| + \log_2 |x| > k$). Therefore,

$$\text{Prob}\left[L_{N_1^A} \text{ is infinite and } L_{N_1^A} \subseteq L_0^A\right] \ge \text{Prob}\,[A' \in S] > c.$$

The obtained contradiction proves the implication $1') \Rightarrow 1)$. We can prove the implication $2') \Rightarrow 2)$ in a similar way.

So, we have to prove $1')$ and $2')$. Let N be an arbitrary nondeterministic machine. Let us prove that the probabilities of both events

(3) $$\text{“}L_{N^A} \text{ is infinite and } L_{N^A} \subseteq L_0^A\text{”}$$

(4) $$\text{“}\mathbf{B}^* \setminus L_{N^A} \text{ is infinite and } \mathbf{B}^* \setminus L_{N^A} \subseteq L_1^A\text{”}$$

do not exceed 0.9. Obviously, if L_{N^A} is infinite and $L_{N^A} \subseteq L_0^A$ then there exist infinitely many i such that $N^A(1^{t_i}) = 1$. Thus, to prove that Prob $[(3)] \le 0.9$ it suffices to prove the implication

$$\text{Prob}\left[\exists^\infty i\ N^A(1^{t_i}) = 1\right] > 0.9 \Rightarrow \text{Prob}\left[\exists i\ (N^A(1^{t_i}) = 1\ \&\ 1^{t_i} \in L_1^A)\right] \ge 0.1$$

where $\exists^\infty i$ means “there exist infinitely many i such that”. Likewise, to prove that Prob $[(4)] \le 0.9$ it suffices to prove the implication

$$\text{Prob}\left[\exists^\infty i\ N^A(1^{t_i}) = 0\right] > 0.9 \Rightarrow \text{Prob}\left[\exists i\ (N^A(1^{t_i}) = 0\ \&\ 1^{t_i} \in L_0^A)\right] \ge 0.1.$$

We will prove these two implications in parallel. The combined implication is as follows: for any $\delta \in \{0, 1\}$

(9) $$\text{Prob}\left[\exists^\infty i\ N^A(1^{t_i}) = \delta\right] > 0.9 \Rightarrow$$

(10) $$\text{Prob}\left[\exists i\ (N^A(1^{t_i}) = \delta\ \&\ 1^{t_i} \in L_\delta^A)\right] \ge 0.1.$$

Let us fix $\delta \in \{0, 1\}$. Assume that (9) is true. Let us prove (10). Assertion (9) implies that

(11)
$$\sum_{i=k}^{\infty} \text{Prob}\left[N^A(1^{t_i}) = \delta\ \&\ N^A(1^{t_{i-1}}) \ne \delta, N^A(1^{t_{i-2}}) \ne \delta, \ldots, N^A(1^{t_k}) \ne \delta\right] > 0.9$$

for all $k \in \mathbf{N}$.

Denote the event

$$N^A(1^{t_{i-1}}) \ne \delta, N^A(1^{t_{i-2}}) \ne \delta, \ldots, N^A(1^{t_k}) \ne \delta$$

by $P_{ik\delta}(A)$.

Lemma 1. *If k is large enough, then for any $i \geq k$,*

$$\begin{aligned} &\text{Prob}\left[1^{t_i} \in L^A_\delta \ \&\ N^A(1^{t_i}) = \delta \ \&\ P_{ik\delta}(A)\right] > \\ &(1/3)\text{Prob}\left[N^A(1^{t_i}) = \delta \ \&\ P_{ik\delta}(A)\right] - \epsilon_i, \end{aligned} \tag{12}$$

where ϵ_i is a sequence having the convergent series $\sum_{i=1}^\infty \epsilon_i$.

We will prove the lemma later. Let us continue now the proof of the theorem. By summing (12) over $i \geq k$ we get

$$\begin{aligned} &\sum_{i=k}^\infty \text{Prob}\left[1^{t_i} \in L^A_\delta \ \&\ N^A(1^{t_i}) = \delta \ \&\ P_{ik\delta}(A))\right] > \\ &(1/3)\sum_{i=k}^\infty \text{Prob}\left[N^A(1^{t_i}) = \delta \ \&\ P_{ik\delta}(A)\right] - \sum_{i=k}^\infty \epsilon_i. \end{aligned} \tag{13}$$

Combining (11) and (13) we get

$$\sum_{i=k}^\infty \text{Prob}\left[1^{t_i} \in L^A_\delta \ \&\ \mathbf{N}^A(1^{t_i}) = \delta \ \&\ P_{ik\delta}(A)\right] > 0.3 - \sum_{i=k}^\infty \epsilon_i. \tag{14}$$

As the events whose probabilities are summed in the left hand side of (14) are pairwise disjoint and included in the event

$$\exists i \ (N^A(1^{t_i}) = \delta \ \&\ 1^{t_i} \in L^A_\delta), \tag{15}$$

we get that the probability of (15) is greater then $0.3 - \sum_{i=k}^\infty \epsilon_i$. By taking k so large that $\sum_{i=k}^\infty \epsilon_i < 0.2$ we get (10).

Thus the proof of the theorem is completed. However, it remains to prove Lemma 1.

Proof of the Lemma 1. Let us fix a sufficiently large k (how large should be k we will see later). Let us fix an arbitrary $i \geq k$. Denote by D_i the set of all binary words of length less than $t_i + \log_2 t_i$ and by F_i the set of all functions from D_i into $\mathbf{B}$. Denote by $f|S$ the restriction of function f to the set S. For any j, the event $N^A(1^{t_j}) \neq \delta$ depends only on the value of A on the words of length bounded by a polynomial of t_j (because N on input 1^{t_j} can put to the oracle questions of the length bounded by a polynomial of t_j). As $t_i = 2^{t_{i-1}}$, for sufficiently large k, for any $i \geq k$, the event $P_{ik\delta}(A)$ depends only on $A|D_i$. Therefore, we may assume that this is the case.

We will prove that for all $f \in F_i$,

$$\begin{aligned} &\text{Prob}\left[1^{t_i} \in L^A_\delta \ \&\ N^A(1^{t_i}) = \delta \mid A|D_i = f\right] > \\ &(1/3)\text{Prob}\left[N^A(1^{t_i}) = \delta \mid A|D_i = f\right] - \epsilon_i, \end{aligned} \tag{16}$$

where $\{\epsilon_i\}$ is a sequence such the series $\sum_{i=1}^\infty \epsilon_i$ converges. Let us prove that this implies (12). Assume that (16) is true for any $f \in F_i$. Then by multiplying (16) by $\text{Prob}\left[A|D_i = f\right]$ we get

$$\begin{aligned} &\text{Prob}\left[1^{t_i} \in L^A_\delta \ \&\ N^A(1^{t_i}) = \delta \ \&\ A|D_i = f\right] > \\ &(1/3)\text{Prob}\left[N^A(1^{t_i}) = \delta \ \&\ A|D_i = f\right] - \epsilon_i \text{Prob}\left[A|D_i = f\right]. \end{aligned} \tag{17}$$

By summing (17) over all $f \in F_i$ such that $P_{ik\delta}(A)$ is true provided $A|D_i = f$, we get

$$\begin{aligned}&\mathrm{Prob}\left[1^{t_i} \in L^A_\delta \ \&\ N^A(1^{t_i}) = \delta \ \&\ P_{ik\delta}(A)\right] > \\ &(1/3)\mathrm{Prob}\left[N^A(1^{t_i}) = \delta \ \&\ P_{ik\delta}(A)\right] - \epsilon_i \mathrm{Prob}\left[P_{ik\delta}(A)\right] \geq \\ &(1/3)\mathrm{Prob}\left[N^A(1^{t_i}) = \delta \ \&\ P_{ik\delta}(A)\right] - \epsilon_i,\end{aligned}$$

that is, we get (12)

So we have to prove that (16) is true for all $f \in F_i$. Let us fix an arbitrary $f \in F_i$. Obviously, the number of queries made by M during possible computations on input 1^{t_i}, is bounded by a polynomial of t_i. Denote that polynomial by $q(t_i)$.

To proceed further we have to handle the cases $\delta = 0$ and $\delta = 1$ separately.

Lemma 2.

$$\begin{aligned}&\mathrm{Prob}\left[N^A(1^{t_i}) = 1 \ \&\ 1^{t_i} \in L^A_1 \mid A|D_i = f\right] > \\ &\left(1 - (1 - 2^{-t_i})^{2^{t_i} - q(t_i)}\right)\mathrm{Prob}\left[N^A(1^{t_i}) = 1 \mid A|D_i = f\right].\end{aligned} \tag{18}$$

Proof. Let $r = q(t_i)$. An *r-circumstance* is a set of oracles of the form

$$\{A \mid A|D_i = f,\ A|B_1 = f_1, A|B_2 = f_2, \ldots, A|B_r = f_r,\ A|E = g\}, \tag{19}$$

where $B_1, \ldots, B_r$ are i-blocks, $f_1, \ldots, f_r$ are functions from $B_1, \ldots, B_r$, respectively, into $\mathbf{B}$, E is a finite set being disjoint with D_i and disjoint with all the i-blocks, and g is a function from E into $\mathbf{B}$. Let

$$W = \{A \mid N^A(1^{t_i}) = 1, A|D_i = f\}.$$

Lemma 3. *The set W can be represented as a finite union of r-circumstances.*

Proof. Let us prove first that the set W can be represented as finite union of intervals of the form $\{A \mid A(w_1) = b_1, \ldots, A(w_r) = b_r,\ A|D_i = f\}$. Let A be an oracle such that $A|D_i = f$. Assume that $N^A(1^{t_i}) = 1$. Fix an accepting computation of N^A on input 1^{t_i}. Let the oracle be queried about the value on the words $w_1, \ldots, w_r$. Then the set W contains the entire interval $\{A \mid A(w_1) = b_1, \ldots, A(w_r) = b_r,\ A|D_i = f\}$. Since the number of possible computations of N on input 1^{t_j} with different oracles is finite, we are done.

Thus, the set W can be represented as a finite union of intervals of the form $\{A \mid A(w_1) = b_1, \ldots, A(w_r) = b_r,\ A|D_i = f\}$. Obviously any interval having this form is a finite union of r-circumstances.

Denote by $\Gamma_1, \ldots, \Gamma_n$ the r-circumstances such that $W = \Gamma_1 \cup \cdots \cup \Gamma_n$. Thus, we have to prove that

$$\mathrm{Prob}\left[1^{t_i} \in L^A_1 \mid A \in \Gamma_1 \cup \cdots \cup \Gamma_n\right] > 1 - (1 - 2^{-t_i})^{2^{t_i} - r}. \tag{20}$$

Call r-circumstance (19) *bad* if f_l is identically zero for some $l \leq r$. Any oracle A in a bad r-circumstance satisfies the assertion $1^{t_i} \in L^A_1$. Therefore, if we remove from the union $\Gamma_1 \cup \cdots \cup \Gamma_n$ all bad r-circumstances, the probability

$$\mathrm{Prob}\left[1^{t_i} \in L^A_1 \mid A \in \Gamma_1 \cup \cdots \cup \Gamma_n\right]$$

will not increase. Thus it suffices to prove (20) in the case when all $\Gamma_1, \dots, \Gamma_n$ are not bad. Clearly, it suffices to prove that for all $m \le n$,

(21)
$$\text{Prob}\,[A \text{ is identically zero in some } i\text{-block} \mid A \in \Gamma_m \setminus (\Gamma_1 \cup \dots \cup \Gamma_{m-1})] > 1 - (1 - 2^{-t_i})^{2^{t_i} - r}$$

provided $\Gamma_m \setminus (\Gamma_1 \cup \dots \cup \Gamma_{m-1})$ is not empty.

Let us fix $m \le n$. Let Γ_m be equal to (19). Denote by $C_1, \dots, C_{2^{t_i}-r}$ all the i-blocks that do not belong to the set $\{B_1, \dots, B_r\}$. For any $l \le 2^{t_i} - r$ let

(22) $$p_l = \text{Prob}\,[A \text{ is identically zero in } C_l \mid A \in \Gamma_m \setminus (\Gamma_1 \cup \dots \cup \Gamma_{m-1}) \text{ and } A \text{ is not identically zero in all the blocks } C_1, \dots, C_{l-1}].$$

Of course, it may happen that some of p_l's is undefined because the set

$$\{A \in \Gamma_m \setminus (\Gamma_1 \cup \dots \cup \Gamma_{m-1}) \mid A \text{ is not identically zero in all the blocks } C_1, \dots, C_{l-1}\}$$

is empty. This means that any $A \in \Gamma_m \setminus (\Gamma_1 \cup \dots \cup \Gamma_{m-1})$ is identically zero in some of the blocks $C_1, \dots, C_{l-1}$. Thus in this case the probability (21) is equal to 1, so we are done.

Obviously, the probability (21) is equal to $1 - (1 - p_1)(1 - p_2) \cdots (1 - p_{2^{t_i}-r})$. Thus to prove Lemma 2 it suffices to prove the following

Lemma 4. *$p_l \ge 2^{-t_i}$ for any $l \le 2^{t_i} - r$ such that p_l is defined.*

Proof. Let us fix arbitrary $l \le 2^{t_i} - r$ such that p_l is defined. Let us prove that $p_l \ge 2^{-t_i}$. Define for any oracle A the new oracle A' as follows:

$$A'(u) = \begin{cases} 0, & \text{if } u \in C_l; \\ A(u), & \textit{otherwise}. \end{cases}$$

Let us call a set of oracles U *monotone* if $A \in U$ implies $A' \in U$. We claim that the set

$$U = \{A \mid A \in \Gamma_m \setminus (\Gamma_1 \cup \dots \cup \Gamma_{m-1}) \text{ and } A \text{ is not identically zero in all the blocks } C_1, \dots, C_{l-1}\}$$

is monotone. Recall that

$$\Gamma_m = \{A \mid A|D_i = f,\ A|B_1 = f_1, \dots, A|B_r = f_r,\ A|E = g\}.$$

Recall that C_l does not belong to the set $\{B_1, \dots, B_r\}$. This implies that Γ_m is monotone.

Since all the sets $\Gamma_1, \dots, \Gamma_{m-1}$ are not bad, the complement of the set $\Gamma_1 \cup \dots \cup \Gamma_{m-1}$ is monotone, too. Obviously, the set

$$\{A \mid A \text{ is not identically zero in all the blocks } C_1, \dots, C_{l-1}\}$$

is monotone. Therefore, U is monotone as an intersection of monotone sets.

Thus Lemma 4 easily follows from the following

Lemma 5. *Any monotone set V of positive measure satisfies the inequality*

$$\text{Prob}\,[A \text{ is identically zero in } C_l \mid A \in V] \geq 2^{-t_i}.$$

Proof of Lemma 5. Let V satisfy the assumptions of the Lemma 5. Denote by H the set of all the functions from C_l into $\mathbf{B}$. Denote by $\mathbf{0}$ the identically zero function in H. Let h be an arbitrary function in H. We claim that there exists a measure preserving 1–1–function from the set

$$\mathcal{B} = \{A \in V \mid A|C_l = h\}$$

into the set

$$\mathcal{D} = \{A \in V \mid A|C_l = \mathbf{0}\}.$$

Indeed, let π be a permutation of H such that $\pi(h) = \mathbf{0}$. Then the mapping $A \mapsto A''$, where

$$A''(u) = \begin{cases} \pi(A|C_l)(u), & \text{if } u \in C_l; \\ A(u), & \text{otherwise.} \end{cases}$$

maps $\mathcal{B}$ into $\mathcal{D}$ because $A'' = A'$ if $A \in \mathcal{B}$ and V is monotone.

Thus for any h,

$$\text{Prob}\,[A \in V \ \&\ A|C_l = h] \leq \text{Prob}\,[A \in V \ \&\ A|C_l = \mathbf{0}].$$

Therefore,

$$\text{Prob}\,[A \in V] = \sum_h \text{Prob}\,[A \in V \ \&\ A|C_l = h] \leq 2^{|C_l|}\text{Prob}\,[A \in V \ \&\ A|C_l = \mathbf{0}].$$

Thus Lemma 4 and Lemma 2 are proved.

It is easy to verify that $\lim_{i\to\infty}\left(1-(1-2^{-t_i})^{2^{t_i}-q(t_i)}\right) = 1-e^{-1} > 1/3$, therefore Lemma 2 implies Lemma 1 in the case $\delta = 1$ (in this case we set $\epsilon_i = 0$).

It remains to handle with the case $\delta = 0$. Let $a_i = (1-2^{-t_i})^{2^{t_i}} = \text{Prob}\,[1^{t_i} \in L_0^A]$ and $\epsilon_i = (1-2^{-t_i})^{2^{t_i}-q(t_i)} - a_i$. Obviously,

$$\epsilon_i = O(q(t_i)2^{-t_i}).$$

Thus the series $\sum_{i=1}^{\infty}\epsilon_i$ converges.

Lemma 1 for $\delta = 0$ immediately follows from the following

Lemma 6.

(23)

$$\text{Prob}\left[1^{t_i} \in L_0^A \ \&\ N^A(1^{t_i}) = 0 \mid A|D_i = f\right] > a_i\text{Prob}\left[N^A(1^{t_i}) = 0 \mid A|D_i = f\right] - \epsilon_i.$$

Proof. By Lemma 2,

$$\begin{aligned}(24)\qquad &\text{Prob}\left[1^{t_i} \in L_1^A,\ N^A(1^{t_i}) = 1 \mid A|D_i = f\right] \\ &> (1-a_i-\epsilon_i)\text{Prob}\left[N^A(1^{t_i}) = 1 \mid A|D_i = f\right].\end{aligned}$$

In the next equations we shall omit the condition $A|D_i = f$ to make them readable. We have

$$\begin{aligned}
&\text{Prob}\,[1^{t_i} \in L_1^A,\ N^A(1^{t_i}) = 0] \\
&= \text{Prob}\,[1^{t_i} \in L_1^A] - \text{Prob}\,[1^{t_i} \in L_1^A,\ N^A(1^{t_i}) = 1] \\
&\le 1 - a_i - (1 - a_i - \epsilon_i)\text{Prob}\,[N^A(1^{t_i}) = 1] \\
&= (1 - a_i)\text{Prob}\,[N^A(1^{t_i}) = 0] + \epsilon_i \text{Prob}\,[N^A(1^{t_i}) = 1] \\
&\le (1 - a_i)\text{Prob}\,[N^A(1^{t_i}) = 0] + \epsilon_i. \\
&\text{Prob}\,[1^{t_i} \in L_0^A,\ N^A(1^{t_i}) = 0] \\
&= \text{Prob}\,[N^A(1^{t_i}) = 0] - \text{Prob}\,[1^{t_i} \in L_1^A,\ N^A(1^{t_i}) = 0] \\
&\ge \text{Prob}\,[N^A(1^{t_i}) = 0] - (1 - a_i)\text{Prob}\,[N^A(1^{t_i}) = 0] - \epsilon_i \\
&= a_i \text{Prob}\,[N^A(1^{t_i}) = 0] - \epsilon_i.
\end{aligned}$$

The Lemma 6 is proved.

Corollary 2. [22] *For random A there exists an infinite* NP^A*-set having no infinite* Co-NP^A*-subsets.*

Corollary 3. *For random A there exists an infinite* Co-NP^A*-set having no infinite* NP^A*-subsets.*

The rest of the paper is organized as follows. The next section contains two well known theorems which are used to prove our lower bound for perceptrons. In Section 5, the bound itself is proved. In Section 6 we apply the method to prove that $\text{AM} \cap \text{Co-AM}$ is not contained in PP relative to an oracle.

4. Auxiliary theorems

We will use the following two well known results.

Theorem 4. *(Chernoff inequality* [9]*) Let $\xi_1, \ldots, \xi_n$ be independent random values in the set $\{0, 1\}$ such that* $\text{Prob}\,[\xi_i = 1] = p$ *for all i. Then for any $\delta \in (0; p(1-p)]$,*

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

The following lemma by Farkash is a version of the duality theorem in the theory of linear programming (see, for example [1,17]).

Lemma 7. *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*

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

5. The extension of one-in-a-box theorem

Definition 4. 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 $W_P(\phi)$, is the sum of weights on all AND's which are true on ϕ. The perceptron outputs 1 on input ϕ if $W_P(\phi)$ is greater than the threshold of its threshold gate and 0 otherwise. The output value is denoted by $P(\phi)$.

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 every its 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 5. *Let* $0 < \varepsilon \leq 1/2$. *If there exists a perceptron of order* d *and total weight* w *which separates good matrices from* $(1-\varepsilon)$*-bad matrices of size* $n \times m$, *then*

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

or

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

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

$$\begin{aligned}&\mathrm{Prob}_{\mu(M)}[M_{i_1j_1}=b_1,\dots,M_{i_uj_u}=b_u]\\ =&\mathrm{Prob}_{\nu(M)}[M_{i_1j_1}=b_1,\dots,M_{i_uj_u}=b_u]\end{aligned}$$

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

The theorem easily follows from the following two lemmas.

Lemma 8. *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 9. *If $d<\sqrt{(3/8)\varepsilon m}$ and $0<\varepsilon\le 1/2$, 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 8.

Proof of Lemma 8. Let d,μ,ν,q,p be as in Lemma 8. 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

$$\mathrm{E}_\mu W_P(M)=\mathrm{E}_\nu W_P(M). \tag{27}$$

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 any AND-gate C in P. Let C be the conjunction $\bigwedge_{s=1}^u(M_{i_sj_s}=b_s)$, where b_s's are 0 or 1. Then

$$\mathrm{Prob}_{\mu(M)}[C(M)=1]=\mathrm{Prob}_{\mu(M)}[M_{i_1j_1}=b_1,\dots,M_{i_uj_u}=b_u].$$

Thus d-indistinguishability of μ and ν implies (27).

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 $\mathrm{E}_\mu W_P(M)\ge t+1$. 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 $\mathrm{E}_\nu W_P(M)\le(1-p)t+pw$. Therefore, $t+1\le(1-p)t+pw$. Thus, $t+1\le t+pw$ which implies $1\le pw$.

Proof of Lemma 9. Assume that the conditions of Lemma 9 are fulfilled. Let σ be a probability distribution in the segment $\{1,2,\dots,m\}$ and τ be a probability distribution in the segment $\{0,1,2,\dots,m\}$. Distributions σ and τ will 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 is 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 some 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 4. By applying the inequality we conclude that a random matrix has less than $(1 - \varepsilon)n$ zero rows with probability at most

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

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,

$$\mathrm{E}_{\sigma(k)} k^i = \mathrm{E}_{\tau(k)} k^i \tag{28}$$

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 $\mathrm{E}_{\sigma(k)} k$, $\mathrm{E}_{\sigma(k)} k^2$, $\ldots$, $\mathrm{E}_{\sigma(k)} k^d$ 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. For any $i \leq n$ denote 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

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

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, (28) 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 exist probability distributions σ and ρ in $\{1, 2, \ldots, m\}$ satisfying (29). To do this we will apply the Farkash lemma because the dual problem is easier than the original one.

Lemma 10. *The following conditions are equivalent:*
1) there are probability distributions σ and ρ in the segment $\{1, 2, \ldots, m\}$ such that

$$\mathrm{E}_{\sigma(k)} k^i = 0.6\varepsilon \mathrm{E}_{\rho(k)} k^i, \text{ for } i = 1, 2, \ldots, d, \tag{29}$$

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

$$0.6\varepsilon \sum_{i=1}^{d} p_i j^i + y \geq 0 \text{ for } j = 1, 2, \ldots, m, \tag{30}$$

$$-\sum_{i=1}^{d} p_i j^i + z \geq 0 \text{ for } j = 1, 2, \ldots, m, \tag{31}$$

$$y + z < 0. \tag{32}$$

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 (31) means that $p(j) \leq 0$ for all $j = 1, 2, \ldots, m$. The existence of y such that (30) and (32) 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 degree d such that $-p(0)\frac{1-0.6\varepsilon}{0.6\varepsilon} < p(j) \le 0$ 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, [8,14,18]) says that $\|P'\| \le d^2\|P\|$.

The following simplified version of Ehlich and Zeller lemma [10] 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

$$\|P\|\left(1 - \frac{2d^2}{m}\right) \le \|P_Y\|. \tag{33}$$

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

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

This easily implies (33).

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

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

$$(1 - 2d^2/m)\|P\| \le \frac{\alpha+1}{2}p(0). \tag{34}$$

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

$$\|P\| \ge p(0)\frac{m}{2d^2}. \tag{35}$$

Combining (34) and (35) we obtain

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

Therefore

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

6. Oracle separation of AM ∩ Co-AM from PP

The lower bound in Theorem 5 suffices to construct an oracle under which AM $\not\subseteq$ PP. To construct an oracle under which AM$\cap$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 a one and at least 2/3 of rows in M_1 [M_0] contain no ones.

Theorem 6. *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 make use of Lemma 9. Let $\varepsilon = 1/3$, $\delta = 0.01$ and $m = n$. Suppose that $d \leq \delta n^{1/2}$. Then the conditions of Lemma 9 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

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

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 7. [21] *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 time 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 $j \in \mathbf{N}$

$$1^j \in L(A) \not\Leftrightarrow \text{Prob}\,[\mathrm{PP}_i^A(1^j) = 1] > 1/2. \tag{36}$$

After changing we will fix the value of A on all words which the truth value of (36) 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 (36) 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 $d \le \mathrm{poly}(j) = \mathrm{poly}(\log n)$ and $q \le 2^{\mathrm{poly}(j)} = 2^{\mathrm{poly}(\log n)}$. Let us construct a perceptron P of order d and of total weight $w = 2^d q$ such that

$$P(D) = 1 \Leftrightarrow \mathrm{Prob}\,[\mathrm{PP}_i^{A[D]}(1^j) = 1] > 1/2. \tag{37}$$

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

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

which implies (37).

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 6 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

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

and we are done.

7. Conclusion

Theorem 5 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 (personal communication). 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. (Russian)
2. L. Babai, *Trading Group Theory for Randomness*, 17th Annual ACM Symposium on Theory of Computing, 1985, pp. 421–429.
3. T. Baker and J. Gill and R. Solovay, *Relativization of P=?NP Question*, SIAM Journal on Computing **4** (1975), no. 4, 431–442.
4. R. Beigel, *Perceptrons, PP, and the Polynomial Time Hierarchy*, Seventh Annual Conference on Structure in Complexity Theory, Boston, MA, 1992, pp. 14–19.
5. R. Beigel and N. Reingold and D. Spielman, *PP is closed under intersection*, 23th Annual ACM Symposium on Theory of Computing, 1991, pp. 1–9.
6. C. H. Bennett and J. Gill, *Relative to a random oracle $P \neq NP \neq coNP$ with probability 1*, SIAM Journal on Computing **10** (1981), 96–113.
7. M. Blum and R. Impagliazzo, *General Oracle and Oracle Classes*, 28th Annual IEEE Symposium on Foundation of Computer Science, New York, 1987, pp. 118–126.
8. E. W. Cheney, *Approximation Theory*, McGraw-Hill, 1966.
9. H. Chernoff, *A Measure of Assymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations*, Annals of Mathematical Statistics **23** (1952), 493–509.
10. H. Ehlich and K. Zeller, *Schwankung von Polynomen zwischen Gitterpunkten*, Mathematische Zeitschrift **86** (1964), 41–44.
11. L. Fortnow and N. Reingold, *PP is closed under truth table reductions*, Sixth Annual Conference on Structure in Complexity Theory, 1991, pp. 13–15.
12. Bin Fu, *Separating PH from PP by Relativization* (1990), Preprint.
13. N. Furst and J. Saxe and M. Sipser, *Parity, Circuits and the Polynomial Time Hierarchy*, Mathematical Systems Theory **17** (1984), 13-27.
14. G. G. Lorentz, *Approximation of Functions*, Hort, Rinehart and Winston, New York, 1966.
15. S. Kurtz and S. Mahaney and J. Royer, *Average Dependence and Random Oracle*, Seventh Annual Conference on Structure in Complexity Theory, Boston, MA, 1992, pp. 306–317.
16. M. Minsky and S. Papert, *Perceptrons*, MIT Press, Cambridge, MA, 1988, Expanded edition, first edition appeared in 1967.
17. C. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, New Jersey, 1982.
18. G. Pólya and G. Szegö, *Problems and Theorems in Analysis*, Springer Verlag, 1972.
19. S. Toda, *On the computational power of PP and $\oplus P$*, 30th Annual IEEE Symposium on Foundation of Computer Science, 1989, pp. 514-519.
20. N. K. Vereshchagin, *Lower Bounds for Perceptrons Solving some Separation Problems and Oracle Separation of AM from PP* (1994), Computer Science Department, University of Rochester, Technical report 498.
21. ———, *On the Power of PP*, Seventh Conference on Structure in Complexity Theory, 1992, pp. 138–143.
22. ———, *Relationships between NP-sets, Co-NP-sets and P-sets relative to random oracles*, Izvestiya Vysshyh Uchebnyh Zavedenij. Seria Matematika (1993), no. 3, 31–39 (Russian); Proc. of Eighth Annual IEEE Conference Conference on Structure in Complexity Theory,, San-Diego CA, 1993, pp. 132–138.
23. ———, *Lower Bounds for Perceptrons Solving some Separation Problems and Oracle Separation of AM from PP*, Third Israel Symposium on Theory of Computing and Systems, 1995, pp. 46–51.
24. ———, *NP-sets are Co-NP-immune Relative to a Random Oracle*, Third Israel Symposium on Theory of Computing and Systems, 1995, pp. 40–45.

Translated by THE AUTHOR

Department of Mathematical Logic, Faculty of Mechanics and Mathematics, Moscow State University, Moscow 119899, Russia

E-mail address: ver@mech.math.msu.su