Hardness Results for General Two-Layer Neural Networks

Christian Kuhlmann[*]
Lehrstuhl Mathematik & Informatik
Fakultät für Mathematik
Ruhr-Universität Bochum
D-44780 Bochum
email: kuhlmann@lmi.ruhr-uni-bochum.de

Abstract

We deal with the problem of learning a general class of 2-layer neural networks in polynomial time. The considered neural networks consist of k linear threshold units on the hidden layer and an arbitrary binary output unit.

We show NP-completeness of the consistency problem for classes that use an arbitrary set of binary output units containing a function which depends on all input dimensions. Thereby k is allowed to be polynomial in the input size. Those classes enclose a variety of multilayer neural networks like the class of multilayer feedforward threshold units. We obtain an analogous result for classes of 2-layer neural networks with any fixed nontrivial output unit.

Further we present a hardness result for approximation. We prove that it is NP-hard to find a 2-layer neural network of constant size with output unit PARITY that approximately (up to a constant factor) maximizes the fraction of correctly classified examples in the given training set. We further develop a general tool to prove this type of hardness results for neural networks.

1 Introduction

Two-layer neural network classifiers are an important class of neural networks in various fields of computer science. A natural question in this regard is if we can efficiently compute an appropriate network which performs well in the setting of consideration, i.e. which separates a representative set of data *correctly*, *optimally* or *approximately optimally*, respectively.

All three cases, the problem of learning neural networks that exactly, optimally and approximately optimally classify a training set, have been studied in different ways in the last years. Several, mostly negative, results in this framework arose in the recent time.

Blum and Rivest [4] consider the class of two-layer neural networks with two linear threshold units on the hidden layer and functions like AND, OR, XOR as the output unit. They prove that the decision problem of whether there is a network that exactly classifies a training set is NP-complete. They also show a similar hardness result for a conjunction of k linear threshold units. DasGupta, Siegelmann and Sontag [5] extend the result of Blum and Rivest to two-layer neural networks with piecewise linear hidden units. Schmitt [13] examines the question whether the restriction of the samples, such that they have a limited overlap, and a restriction of the weights of the neurons simplify the problems. Hammer [7] shows hardness for the decision problem of the class of multilayer feedforward threshold units.

Amaldi and Kann [1] prove hardness of identifying finite conjunctions of a number of halfspaces, hyperplanes and their complements which optimally classify a training set.

For the learning problem of finding a neural network that approximately optimally classifies a training set, two main branches were investigated. The first is represented by 'robust learning' where, for *each* $\epsilon > 0$, an efficient learner has to identify a hypothesis with error rate within ϵ from the error rate of an optimal classifying hypothesis, in time polynomial in the sample size and $1/\epsilon$. Höffgen, Simon and Van Horn [9] show that robust learning of halfspaces is NP-hard. The second branch is the problem of identifying a hypothesis which classifies within a *fixed* error rate from the error rate of the optimal classifying hypothesis. Finding such hypotheses is generally much easier than identifying hypotheses which perform optimal, and for practical use often sufficient. Arora, Babai, Stern and Sweedyk [2] show NP-hardness for the class of linear threshold functions if the fixed error is a constant multiple of the optimal error rate. Höffgen, Simon and Van Horn [9] obtain similar results. Bartlett and Ben-David [3] extend this type of results to larger hypothesis classes. They consider a 2-layer neural network consisting of k linear threshold units and a conjunction (linear threshold function, respectively) as the output unit and show NP-hardness to find a network with proportion of correctly classified data within c/k (c/k^3, respectively) from optimal. They substantiate the same result for classes with an arbitrary set of output units containing the conjunction.

We extend the above results in the following way. First we show NP-completeness for the consistency problem of the class of two-layer neural networks with k linear threshold

[*] The author likes to thank Hans Ulrich Simon, Shai Ben-David, Jürgen Forster, Michael Schmitt and Peter Kohlmann for helpful hints and discussions. The author gratefully acknowledges the support of the German-Israeli Foundation for Scientific Research and Development (grant I-403-001.06/95).

units and an arbitrary output unit that depends on at least two inputs. This class contains the one considered by Blum and Rivest. We also show NP-completeness for the consistency problem of the same class admitting an arbitrary set of output units that contains a function depending on all inputs. It turns out that the class of multilayer feedforward threshold units considered by Hammer [7] is a special case of this class. For both results, k can even be polynomial in n.

Finally, we turn towards approximation and extend the result of Bartlett and Ben-David [3]. We show that it is NP-hard to find a two-layer neural network with k hidden linear threshold units that maximizes the fraction of correctly classified examples in the given training set, as long as the set of output units contains PARITY.

Our paper is structured as follows. After formal definitions of the discussed problems and preliminary results in section 2, we describe our main results in section 3. Section 4 introduces a general technique to prove hardness with respect to the consistency and approximation problem for two-layer neural networks with k hidden linear threshold units. Finally, Sections 5 and 6 deal with the proofs of our main results, applying the general technique developed before. In the former section we show NP-completeness for the described classes. In the latter section we prove the NP-hardness result.

2 The models, definitions, and preliminary results

In this section we will formalize the problems we are interested in.

Let $\mathcal{X}$ be the *instance space* and, for each instance $x \in \mathcal{X}$, let $\mathcal{Y}_x$ be the *solution space of* x. The *profit function* $\mathcal{Z}$ assigns to every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}_x$ a value $\mathcal{Z}(x, y) \in [0, 1]$. Let $\mathcal{P} = (\mathcal{X}, (\mathcal{Y}_x)_{x\in\mathcal{X}}, \mathcal{Z})$. Then the *decision problem of* $\mathcal{P}$ is the problem of deciding whether for a given instance $x \in \mathcal{X}$ there exists a solution $y \in \mathcal{Y}_x$ such that $\mathcal{Z}(x, y) = 1$. Further, for the *approximation problem of* $\mathcal{P}$ *with error-rate* ϵ, the algorithm has to identify a solution $y \in \mathcal{Y}_x$ for a given instance $x \in \mathcal{X}$ which satisfies

$$\mathcal{Z}(x, y) \geq (1 - \epsilon) opt_{\mathcal{P}}(x)$$

where $opt_{\mathcal{P}}(x) = \max_{y\in\mathcal{Y}_x}(\mathcal{Z}(x, y))$. For $\epsilon = 0$, we call this problem *maximization problem*.

Regarding decision problems, a well-known example is SET-SPLITTING where the instance space is the set $\mathcal{G}$ of hypergraphs $G = (V, E)$ with $V \subseteq \mathsf{N}$, $E \subseteq 2^V$. The solution space of G contains all *2-colorings* $\tau : V \to \{1, 2\}$ that *2-colors* G, and the profit function $\mathcal{Z}$ outputs the fraction of the number of edges which are *2-colored* by τ, i.e. $\mathcal{Z}(G, \tau) = |\{I \in E : \tau(I) = \{1, 2\}\}|/|E|$.

If we restrict the instance space to the set $\mathcal{G}^2$ of graphs ($|I| = 2$ for all $I \in E$), the corresponding approximation problem are called MAX 2-CUT.

We will obtain our hardness results by reduction relying on the following two basic theorems. The first concerns our decision results and was proven by Lovàsz [6].

Theorem 2.1 SET-SPLITTING *is NP-complete.*

The following theorem is a corollary of a theorem proven by Kann, Khanna, Lagergren and Panconesi [10] is the basis for our approximation results.

Theorem 2.2 MAX 2-CUT *is NP-hard with error-rate* $\epsilon < 1/34$.

A special class of decision problems are *consistency problems*. In this class, the instance space is $\mathcal{S} = (\mathcal{S}_n)_{n\geq 1}$, where $\mathcal{S}_n$ is the set consisting of all finite sequences $S = (v^i, l_i)_{i\in I_S}$ of labeled vectors $(v^i, l_i) \in V_n \times \{-1, 1\}$ called *samples (in* V_n*)*. V_n is a K-Vektorspace. The solution space of a sample in $\mathcal{S}_n$ is a set $\mathcal{F}_n$ of decision functions $f : V_n \to \{-1, 1\}$. Let $\mathcal{F} = (\mathcal{F}_n)_{n\geq 1}$. Finally, the profit function outputs the fraction of points of a sample S *consistent* with a function f, i.e. $\mathcal{Z}(S, f) = |\{i : f(v^i) = l_i, i \in I_S\}|/|I_S|$.

We denote the consistency problem of $\mathcal{F}$ by CONS$(\mathcal{F})$ and the corresponding maximization problem and approximation problem by MAX$(\mathcal{F})$ and APPROX$(\mathcal{F}, \epsilon)$, respectively.

This paper deals with a special class of decision functions, the *class* $\mathcal{F}^{k,\phi_k} = (\mathcal{F}_n^{k(n),\phi_{k(n)}})_{n\geq 1}$ *of two-layer neural networks with k linear classifiers and output unit* ϕ_k, where k is a function in n with positive integer values, and

$$\mathcal{F}_n^{j,\phi} = \{F : F(v) = \phi(f_1(v), ..., f_j(v)), f_i \in LT_n\},$$

where LT_n is the class of *linear threshold functions* sgn$(w \cdot v + \theta)$ with $w \in V_n$, and *threshold* $\theta \in K$. $w \cdot v$ denotes the inner product of two vectors, and sgn$(x) = 1$, if $x \geq 0$ and sgn$(x) = -1$ otherwise. Moreover, $\phi \in \mathcal{B}_j$ where $\mathcal{B}_j$ is the class of all boolean functions $\{-1, 1\}^j \to \{-1, 1\}$. For a subset $\Phi \subseteq \mathcal{B}_j$, we define $\mathcal{F}_n^{j,\Phi} = \bigcup_{\phi\in\Phi} \mathcal{F}_n^{j,\phi}$ and consequently $\mathcal{F}^{k,\Phi_k} = (\mathcal{F}_n^{k(n),\Phi_{k(n)}})_{n\geq 1}$.

With respect to these classes, we ask for the complexity of algorithms solving the problems CONS$(\mathcal{F}^{k,\Phi_k})$ and APPROX$(\mathcal{F}^{k,\Phi_k}, \epsilon)$.

We call a sequence $(\Phi_{k(n)})_{n\geq 1}$ of sets $\Phi_i \subseteq \mathcal{B}_i$ *well-behaving*, if there is a polynomial p and a representation-scheme, i.e. a surjective mapping, $\mathcal{R} : \Sigma^* \to \cup_{n\geq 1}\mathcal{B}_{k(n)}$ with the following properties:

1. For all $n \geq 1$ and all $\phi \in \Phi_{k(n)}$: $size(\phi) \leq p(n)$.

2. There is an algorithm that solves the decision problem

   Instance: $s \in \Sigma^{\leq p(n)}$
   Question: $\mathcal{R}(s) \in \Phi_{k(n)}$?

   in time polynomial in n.

Hence, if $(\Phi_{k(n)})_{n\geq 1}$ is well-behaving, we can guess a word in $s \in \Sigma^{\leq p(n)}$ and check in polynomial time, if $\mathcal{R}(s) \in \Phi_{k(n)}$.

Further, since it is possible to replace weights and threshold of a threshold unit by values of representation length polynomial in the input size (see e.g. Raghavan [12] and Håstad [8]), we can write down all the weights and thresholds of any neural network in $\mathcal{F}^{k,\Phi_k}$ in polynomial time, in order to test the agreement with respect to the given sample. This implies

Lemma 2.3 *Let* $k(n)$ *be polynomial in* n *and let* $(\Phi_{k(n)})_{n\geq 1}$ *be a well-behaving sequence of sets* $\Phi_i \subseteq \mathcal{B}_i$*. Then*

$$\text{CONS}(\mathcal{F}^{k,\Phi_k}) \in NP.$$

Hence, in order to show NP-completeness for the consistency problem of $\mathcal{F}^{k,\Phi_k}$, it suffices to find a reduction from a problem that is already known to be NP-complete.

We continue with further definitions. The *size* $|S|$ of a sample $S = (v^i, l_i)_{i \in I_S}$ is the size of the index set I_S. Further, $S_1 \sqcup S_2$ is the concatenation of two samples S_1 and S_2.

For a threshold function f, P_f denotes the *separating hyperplane of* f, i.e. $P_f = \{v : w \cdot v + \theta = 0\}$. We say that f *separates* a pair of points (v^1, v^2), if $f(v^1) = -f(v^2)$. It is easy to see that in this case P_f intersects the line between v^1 and v^2 at exactly one point, i.e. there is exactly one $\lambda \in [0, 1]$ such that $w \cdot (\lambda v^1 + (1 - \lambda)v^2) + \theta = 0$. We call this point $\lambda v^1 + (1 - \lambda)v^2$ the *separating point of* (v^1, v^2) *with respect to* f.

Consider a boolean function $\phi \in \mathcal{B}_k$ and a vector $b \in \{-1, 1\}^k$. For an index set $I \subseteq \{1, \ldots, k\}$, $b_{(I)}$ denotes the boolean vector b with a negated coordinate at each position $i \in I$. Further, $b_{(i|1)}$ ($b_{(i|-1)}$, respectively) denotes vector b where the i-th coordinate is set to 1 (-1, respectively). We say that b is *critical (for* ϕ*) with respect to* i, if

$$\phi(b) = -\phi(b_{(i)}).$$

We say that ϕ *depends on all dimensions*, if for each i there is a vector b^i which is critical for i. We call the set of such vectors $(b^1, ..., b^k)$ *a witness set of* ϕ. The following lemma is an implication of a result[1] shown by Simon [14].

Lemma 2.4 *Let* ϕ *depend on all dimensions. Then there exists a vector* b^{eff}*, which is critical for at least two different coordinates.*

A vector b^{eff} with the above property is called *effective for* ϕ. In order to simplify our reduction, we introduce the following relation on subsets of $\mathcal{B}_k$. For $\Phi, \Psi \subseteq \mathcal{B}_k$ we write $\Phi \sim \Psi$, iff there exist boolean values $\sigma_0, \ldots, \sigma_k \in \{-1, +1\}$ such that

$$\Psi = \{\sigma_0(\phi \circ \sigma) : \phi \in \Phi\},$$

where $\sigma(x_1, \ldots, x_k) = (\sigma_1 x_1, \ldots, \sigma_k x_k)$. Obviously $\sim$ is an equivalence relation. For $\phi, \psi \in \mathcal{B}_k$ with $\{\phi\} \sim \{\psi\}$ we simply write $\phi \sim \psi$. Finally, $[\phi]$ denotes the equivalence class of $\{\phi\}$ with respect to $\sim$.

For instance, if a function $\phi \in \mathcal{B}_2$ depends on all dimensions, and if b^{eff} is an effective vector of ϕ, then, with $\psi(b) = \phi(b^{\text{eff}})\phi(b_1^{\text{eff}} b_1, b_2^{\text{eff}} b_2)$, obviously $\phi \sim \psi$ and $1 = \psi(1, 1) = -\psi(-1, 1) = -\psi(1, -1)$, i.e. $\psi \in \{\textsc{And}, -\textsc{Xor}\}$. Since $\textsc{Xor} \sim -\textsc{Xor}$, it follows

Lemma 2.5 *If* $\phi \in \mathcal{B}_2$ *depends on all dimensions, then* $\phi \in [\textsc{And}] \cup [\textsc{Xor}]$. •

In the following lemma we see, that NP-completeness of $\textsc{Cons}(\mathcal{F}^{k,\Phi_k})$ is preserved for the whole equivalence class of Φ.

Lemma 2.6 *If* $\textsc{Cons}(\mathcal{F}^{k,\Phi_k})$ *is NP-complete and* $\Phi_k \sim \Psi_k$*, then* $\textsc{Cons}(\mathcal{F}^{k,\Psi_k})$ *is NP-complete.*

Proof We reduce $\textsc{Cons}(\mathcal{F}^{k,\Phi_k})$ to $\textsc{Cons}(\mathcal{F}^{k,\Psi_k})$. Since $\Phi_k \sim \Psi_k$, there exist $\sigma_0, \ldots, \sigma_k \in \{+1, -1\}$ such that $\Psi_k = \{\sigma_0(\phi \circ \sigma), \phi \in \Phi_k\}$. Define function ρ that maps $S = (v^i, l_i)_{i \in I_S}$ to the sample $\tilde{S} = (v^i, \sigma_0 l_i)_{i \in I_S}$. Let $\phi(f_1, \ldots, f_k)$ be a solution for S. Since the size of S is finite, we can assume that $|w^l \cdot v^i + \theta| > 0$ for all $l = 1, \ldots, k$ and $i \in I_S$. This implies that with the functions $\tilde{f}_l(v^i) = \text{sgn}(\sigma_l w^l \cdot v^i + \sigma_l \theta_l) \in LT_n$, $l = 1, \ldots, k$, we have $\sigma_l f_l(v^i) = \tilde{f}_l(v^i)$, which yields $\phi(f_1(v^i), \ldots, f_k(v^i)) = \sigma_0 \psi(\sigma_1 f_1(v^i), \ldots, \sigma_k f_k(v^i)) = \sigma_0 \psi(\tilde{f}_1(v^i), \ldots, \tilde{f}_k(v^i))$. Hence, $\psi \in \Psi_k$ and $\psi(\tilde{f}_1, \ldots, \tilde{f}_k)$ is a solution for $\tilde{S}$. The converse direction is analogous. •

3 Main results

We present our main results in this paper. The first two theorems are NP-completeness results for the class of two-layer neural networks with linear threshold units and arbitrary nontrivial output units.

Theorem 3.1 *Assume that for sufficiently large* n, $2 \leq k(n) \leq \frac{n-3}{2}$ *and* $\phi_{k(n)} \in \mathcal{B}_{k(n)}$ *depends on at least two dimensions. Then* $\textsc{Cons}(\mathcal{F}^{k,\phi_k})$ *is NP-complete.*

The following result similar except that it admits a class of output units.

Theorem 3.2 *Assume that for sufficiently large* n, $2 \leq k(n) \leq \frac{n-3}{2}$ *and* $\Phi_{k(n)} \subseteq \mathcal{B}_{k(n)}$*, where* $\Phi_{k(n)}$ *contains a function* ϕ *depending on all dimensions. Then* $\textsc{Cons}(\mathcal{F}^{k,\Phi_k})$ *is NP-hard.*

If, in addition, Φ_k *is well-behaving, then* $\textsc{Cons}(\mathcal{F}^{k,\Phi_k})$ *is NP-complete.*

The last theorem is an NP-hardness result of approximation concerning two-layer neural networks with linear threshold units and a class of output units containing $\textsc{Parity}^2$.

Theorem 3.3 *Let* $k \geq 2$ *be constant and* $\Phi \subseteq \mathcal{B}_k$ *with* $\textsc{Parity} \in \Phi$*. Then* $\textsc{Approx}(\mathcal{F}^{k,\Phi}, \epsilon)$ *is NP-hard for*

$$\epsilon = \tfrac{1}{384 + 128k}.$$

4 General technique

In this section we develope a general tool to show hardness results of the consistency and approximation problem of two layer neural networks with linear classifiers. We first introduce a reduction ρ from $\textsc{Set-Splitting}$ to $\textsc{Cons}(\mathcal{F}^{2,\textsc{Xor}})$ proposed by Blum and Rivest [4] and show that this reduction is also a valid reduction for the broader class $\mathcal{F}^{2,\Phi}$, if Φ contains a function depending on all dimensions.

In theorem 4.3 we present conditions to show NP-completeness of $\textsc{Cons}(\mathcal{F}^{k,\Phi_k})$ from the properties of $\mathcal{F}^{2,\Phi}$. Theorem 4.4 is the counterpart for NP-hardness in approximation.

Let $G = (V, E)$ be a hypergraph and $I \in E$. We define the sample H_G^I consisting of $(\pm e^i, -1)$, $(e^I, 1)$ and $(\bar{0}, 1)$ for all $i \in I$ where e^i denotes the i-th unit vector and $e^I =$

[1] Simon [14] even shows the significantly better lower bound $\log k$ for the number of coordinates for which there exists a critical vector

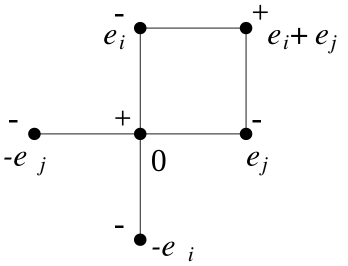

Figure 1: The sample $H_G^{\{i,j\}}$

$\sum_{i\in I} e^i$. Figure 1 illustrates this sample for an edge $\{i,j\}$. Further let H_G consist of all points that appear in H_G^I for all $I \in E$. Then ρ is the function that maps G to H_G.

Lemma 4.1 *Let $\psi(f_1, f_2) \in \mathcal{F}_n^{2,\psi}$ for an arbitrary boolean function ψ. If f_1 is constant on $\{e^i : i \in I\}$, then $\psi(f_1, f_2)$ is not consistent with H_G^I.*

Proof Let $h_j(v) = w^j \cdot v + \theta_j$ and $f_j(v) = \text{sgn}(h_j(v))$, $j = 1, 2$. Without loss of generality[3] we can assume that $\theta_1, \theta_2 \geq 0$. This implies in particular $\psi(1,1) = 1$.

Suppose now that $\psi(f_1, f_2)$ is consistent with H_G^I. Since $\bar{0}$ and e^i have different labels for all $i \in I$ and f_1 is constant on $\{e^i : i \in I\}$, either $f_1 \equiv -1$ or $f_2 \equiv -1$ on $\{e^i : i \in I\}$. Let $f_1 \equiv -1$ on $\{e^i : i \in I\}$ (the other case is treated similarly). Since $\theta_1 \geq 0$, $0 > h_1(e^i) = w_i^1 + \theta_1 > w_i^1$ for all i. This implies $h_1(e^I) = \sum_{j\in I} w_j^1 + \theta_1 < 0$ and $h_1(-e^i) = -w_i^1 + \theta_1 \geq 0$. Hence,

$$f_1(e^I) = f_1(e^i) = -f_1(\bar{0}) = -f_1(-e^i)$$

for all i. Since $\bar{0}$ and $-e^i$ have different labels and $f_1(\bar{0}) = f_1(-e^i)$, $f_2(-e^i) = -f_2(\bar{0}) = -1$ for all i. Together with $\theta_2 \geq 0$ we obtain $0 > h_2(-e^i) = -w_i^2 + \theta_2 > -w_i^2$. This gives us $h_2(e^i) = w_i^2 + \theta_2 \geq 0$ and $h_2(e^I) = \sum_{j\in I} w_j^2 + \theta_2 \geq 0$. This implies $f_2(e^i) = f_2(e^I)$ and, finally,

$$\psi(f_1(e^i), f_2(e^i)) = \psi(f_1(e^I), f_2(e^I))$$

which is a contradiction to the alternately labeled e^i and e^I. ●

Theorem 4.2 *Let $\Phi \subseteq \mathcal{B}_2$ be a set which contains a function depending on all dimensions. Then* CONS$(\mathcal{F}^{2,\Phi})$ *is NP-complete.*

Proof

We use the reduction ρ to show hardness for class Φ of output units. Due to Lemma 2.5 and 2.6, without loss of generality we can assume that Φ contains AND or $-$XOR. Assume first that τ 2-colors G, $V = \{1, \ldots, n\}$. Define $f_i(v) = \text{sgn}(w^i \cdot v + 1/2)$ with

$$w_j^1 = \begin{cases} -1, & \text{if } \tau(j) = 1 \\ n, & \text{if } \tau(j) = 2 \end{cases} \quad \text{and} \quad w_j^2 = \begin{cases} -1, & \text{if } \tau(j) = 2 \\ n, & \text{if } \tau(j) = 1 \end{cases}$$

Obviously $f_1(\bar{0}) = f_2(\bar{0}) = 1$ and $f_1(\pm e^i) = -f_2(\pm e^i)$. Since G is 2-colored, also $f_1(\pm e^I) = -f_2(\pm e^I) = 1$ for all $I \in E$. Therefore, AND(f_1, f_2) and $-$XOR(f_1, f_2) are consistent with $\rho(G)$.

Let conversely $\psi(f_1, f_2)$ be a solution $\mathcal{F}^{2,\Phi}$ on $\rho(G)$. Define the mapping τ as follows: $\tau(i) = 1$, if $f_1(e^i) = 1$, and $\tau(i) = 2$, otherwise. Suppose that τ is not a solution, i.e. that an edge $I \in E$ has a monochromatic coloring, say $\tau(I) = \{1\}$. According to the definition of τ, $f_1(e^i) = 1$ for all $i \in I$. Then according to Lemma 4.1, $\psi(f_1, f_2)$ is not consistent with H_G^I. Hence, $\psi(f_1, f_2)$ is not a solution for CONS$(\mathcal{F}^{2,\psi})$. ●

Consider two samples H and $\tilde{H}$ in V_n. We say that $\tilde{H}$ *employs $\mathcal{F}^{k,\Phi_k}$ for H*, iff for all $\phi(f_1, \ldots, f_k) \in \mathcal{F}^{k,\Phi_k}$ consistent with $H \sqcup \tilde{H}$ at least $k - 2$ functions of $f_1, \ldots, f_k$ are constant on H.

Theorem 4.3 *Let m be constant. If there are polynomial-time computable functions ρ, η mapping a hypergraph G with $|V| = n$ to a sample in V_{n+m} such that the following holds*

- *For each 2-colorable $G \in \mathcal{G}$, there exists $F \in \mathcal{F}^{k,\Phi_k}$ consistent with $\rho(G) \sqcup \eta(G)$*
- *$\eta(G)$ employs $\mathcal{F}^{k,\Phi_k}$ for $\rho(G)$,*

then the consistency problem CONS$(\mathcal{F}^{k,\Phi_k})$ *is NP-complete.*

Proof We show that $\varrho : G \mapsto \rho(G) \sqcup \eta(G)$ is a reduction from SET-SPLITTING to CONS$(\mathcal{F}^{k,\Phi_k})$. First, assume that G is 2-colorable. Because of the first condition, there exists $F \in \mathcal{F}^{k,\Phi_k}$ consistent with $\rho(G) \sqcup \eta(G)$.

The converse direction relies on the second condition: Assume that there exists a function $F = \phi(f_1, ..., f_k)$ consistent with $\rho(G) \sqcup \eta(G)$. Since $k - 2$ functions f_i are constant on $\rho(G)$, there is a function $\psi \in \mathcal{B}_2$ and two functions of $f_1, ..., f_k$, say f_1, f_2, with $\phi(f_1, ..., f_k) = \psi(f_1, f_2)$ on $\rho(G)$. Together with Lemma 4.1 and Theorem 4.2 we obtain a solution for SET-SPLITTING. ●

Theorem 4.4 *Let m, k be constant, and let η be a polynomial-time computable function that maps a graph G with $|V| = n$ to a sample in V_{n+m}. Assume that for each $G \in \mathcal{G}^2$ the following holds:*

- *For all 2-colorings $\tau : V \to \{1, 2\}$ there exists $F \in \mathcal{F}^{k,\Phi}$ such that for all 2-colored edges I and all vertices $i \in V$, F is consistent with $H_G^{\{i\}}$, H_G^I and $\eta(G)$.*
- *There exist samples $\eta_G(I)$ with $\eta(G) = \sqcup_{I\in E}\eta_G(I)$ such that $\eta_G(I)$ employs $\mathcal{F}^{k,\Phi}$ for H_G^I, for all $I \in E$.*

Then APPROX$(\mathcal{F}^{k,\Phi}, \epsilon)$ *is NP-hard for $\epsilon = 1/(384 + 128z)$, where $z = \max_{G,I\in E} |\eta_G(I)|$.*

For the detailed proof we refer to the appendix (A).

[2] PARITY denotes the function that outputs 1, if the number of positive arguments is odd, and -1 otherwise

[3] Otherwise let $\sigma_j = \text{sgn}(\theta_j)$ for $j = 1, 2$. Since H_G^I is finite, we can assume that $|\hat{f}_j(v)| > 0$. With the transformation $\phi(b_1, b_2) \to \phi(\sigma_1 b_1, \sigma_2 b_2)$ and $\hat{f}_j \to \sigma_j \hat{f}_j$ for $j = 1, 2$ we obtain the function $\tilde{\phi}(\tilde{f}_1, \tilde{f}_2) \in \mathcal{C}_n^{2,\tilde{\phi}}$ with $\tilde{\theta}_1, \tilde{\theta}_2 \geq 0$, that has the same behaviour on H_G^I as $\phi(f_1, f_2)$.

5 NP-completeness

In this section we will prove the first two main theorems 3.1 and 3.2.

Proof of Theorem 3.2 We will use Theorem 4.3 to show this theorem, i.e. we will construct the functions ρ, η and verify the required conditions.

Consider an arbitrary fixed n and let $k = k(n)$. For $k = 2$, we can apply Theorem 4.2. Assume now that $k \geq 3$. Let Φ be an arbitrary subset of $\mathcal{B}_k$ containing the function ϕ which depends on each dimension. Due to Lemma 2.6, we can w.l.o.g. assume that $b^{\text{eff}} = (1, \ldots, 1)$, $\phi(b^{\text{eff}}) = 1$, and that coordinates 1 and 2 are critical for b^{eff}. Let $(b^1, \ldots, b^k)$ be a witness set for ϕ.

In order to construct ρ and η, we require a polynomial-time algorithm V-SYSTEM which with input n outputs a system $U_n = (u^{ij})_{\substack{i=3,\ldots,k \\ j=1,\ldots,kn}}$ of vectors $u^{ij} \in V_n$ with the following properties:

(U1) Every n vectors of U_n are in general position.

(U2) $u_i^{ij} = 0$ for all i, j

(U3) $\text{sgn}(u_l^{ij}) = b_l^i$ for all $i, l = 1, \ldots, k, j = 1, \ldots, kn$, $i \neq l$.

(U4) $size(u_i^{ij}) \leq \alpha(n) = poly(n)$ for $i \neq l$ ($size(x)$ corresponds to the binary representation of x).

Such a polynomial-time algorithm exists. It obtains the vector system by generating vectors from a system of appropriate linear independent polynomials. For the explicit construction of V-SYSTEM we refer to the appendix (B). Then we define ρ that maps a hypergraph $G = (V, E)$, $V = \{3, \ldots, n\}$ to the sample in $\hat{H}_G$ in V_n consisting of $(t \pm e^i, -1), (t + e^I, +1)$ and $(t, 1)$ for all $i \in I$, $I \in E$, where

$$t = (0, 0, 2, \ldots, 2).$$

$\hat{H}_G$ is the same construction as H_G (see section 4), up to a translation of the examples by vector t. Further, η maps G to the sample $G_n = \bigsqcup_{i=3,\ldots,k} G_n^i$ with

$$G_n^i = \bigsqcup_{j=1,\ldots,kn} \Big((u^{ij} + \epsilon e^i, \phi(b^i_{(i|1)})) \sqcup (u^{ij} - \epsilon e^i, \phi(b^i_{(i|-1)})) \Big)$$

and $\epsilon = 2^{-50n^4\alpha}$. Given an algorithm V-SYSTEM fulfilling the above conditions, ρ and η are obviously polynomial-time computable. We call the pairs of points $(t, t + e^i)$, $(t, t - e^i)$ and $(t + e^i, t + e^I)$ for $i = 3, \ldots, n$, $I \in E$ *pairs in* $\hat{H}_G$, and $(u^{ij} + \epsilon e^i, u^{ij} - \epsilon e^i)$ for $i = 3, \ldots, n$ *pairs in* G_n^i. Figure 2 illustrates the samples G_n and $\hat{H}_G$.

Now we prove the first condition in Theorem 4.3. Assume that τ 2-colors G. We define $f = (f_1, \ldots, f_k)$, where $f_i(v) = \text{sgn}(w^i v + \theta_i)$ with

$$\begin{array}{ll} w^1 = (r, 0, \alpha_1, \ldots, \alpha_{n-2}) & \theta_1 = -w^1 t + 1/2 \\ w^2 = (0, r, \beta_1, \ldots, \beta_{n-2}) & \theta_2 = -w^2 t + 1/2 \\ w^i = e^i & \theta_i = 0, \quad i = 3, \ldots, k, \end{array}$$

$$\alpha_i = \begin{cases} -1, & \text{if } \tau(i) = 1 \\ n, & \text{if } \tau(i) = 2 \end{cases} \quad \text{and} \quad \beta_i = \begin{cases} -1, & \text{if } \tau(i) = 2 \\ n, & \text{if } \tau(i) = 1 \end{cases}$$

$$\text{and} \quad r = n^2 2^{2\alpha}.$$

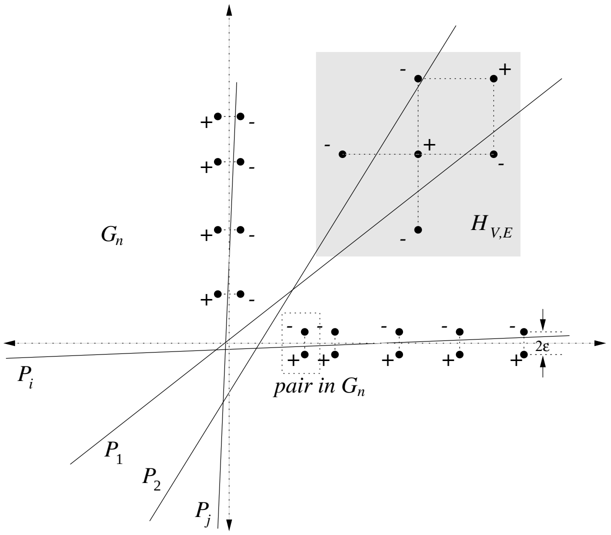

Figure 2: The examples of G_n^i, G_n^j and $H_G^{\{i,j\}}$ projected to V_2. The pairs in G_n 'employ' $k - 2$ hyperplanes such that they are not able to separate the pairs in H_G.

The following calculation will show that $\phi(f)$ is consistent on $\hat{H}_G \sqcup G_n$. Consider first an arbitrary pair $(u^{ij} \pm \epsilon e^i)$ in G_n. Then with (U3)

$$\begin{aligned} f_1((u^{i,j} \pm \epsilon e^i) &= \text{sgn}(w^1 u^{i,j} \pm \epsilon w^1 e^i + \theta_1) \\ &= \text{sgn}\Big(\underbrace{r u_1^{i,j}}_{A} + \underbrace{\textstyle\sum_{l \geq 3} \alpha_{l-2}(u_l^{i,j} - 2) + \epsilon \alpha_{i-2} + 1/2}_{B} \Big). \end{aligned}$$

For A we obtain

$$|r u_1^{i,j}| = 4(n\alpha)^2 |u_1^{i,j}| \geq 4(n\alpha)^2 \alpha^{-1} = 4n^2\alpha.$$

On the other hand, the amount of B is bounded from above by $2n^2\alpha$, since $|\alpha_l| \leq n$ and $|u_l^{i,j}| < \alpha$. Hence, $f_1(u^{i,j} \pm \epsilon e^i) = \text{sgn}(u_1^{i,j}) = b_1^i$ and, similarly, $f_2(u^{ij} \pm \epsilon e^i) = b_2^i$. Further, $f_l(u^{ij} \pm \epsilon e^i) = \text{sgn}(w^l u^{ij} \pm \epsilon w^l e^i) = \text{sgn}(u_l^{ij}) = b_l^i$ for all $l \geq 3$ with $l \neq i$, and $f_i(u^{ij} \pm \epsilon e^i) = \text{sgn}(w^i u^{ij} \pm \epsilon w^i e^i) = \text{sgn}(\pm w^i e^i) = \pm 1$. This implies

$$\phi(f(u^{ij} \pm \epsilon e^i)) = \phi(b^i_{(i|\pm 1)}),$$

i.e. $\phi(f)$ is consistent on G_n.

Let us turn towards the points in $\hat{H}_G$. First, $f_1(t) = \text{sgn}(w^1 t + \theta_1) = \text{sgn}(w^1 t - w^1 t + 1/2) = 1$ and, likewise, $f_2(t) = 1$. Also, $f_l(t) = \text{sgn}(t_l) = \text{sgn}(2) = 1$ for $l = 3, \ldots, n$. Further, the definition of α_i, β_i yields $f_1(t + e^I) = \text{sgn}(w^1 e^I + 1/2) = \text{sgn}(\sum_{i \in I} \alpha_{i-2} + 1/2) = 1$, since I is 2-colored. Similarly, $f_2(t + e^I) = 1$. Also, $f_l(t + e^I) = \text{sgn}(2 + \delta_I^l) = 1$ for all $i = 3, \ldots, n$. Finally, from the definition of α_i, β_i we obtain $f_1(t \pm e^i) = \text{sgn}(\pm w^1 e^i + 1/2) = \text{sgn}(\pm\alpha_{i-2} + 1/2) = -\text{sgn}(\pm\beta_{i-2} + 1/2) = -f_2(t \pm e^i)$ for $i = 3, \ldots, n$. Further, $f_l(t \pm e^i) = \text{sgn}(2 + \delta_i^l) = 1$ for $l = 3, \ldots, n$. Hence, $\phi(f(t \pm e^i) = \phi(-1, 1, \ldots, 1)$ or $= \phi(1, -1, 1, \ldots, 1)$, which is -1. Consequently, all points in $\hat{H}_G$ are correctly classified by $\phi(f)$.

Now we show that the second condition in Theorem 4.3 holds. Because of condition U4, $|u_l^{ij}| < 2^{\alpha}$, which yields that $||u^{ij}|| < n2^{\alpha}$. Since for $(v, l) \in \hat{H}_G$, $|v_l| < 3n$ for all l, we conclude that

$$\hat{H}_G \sqcup G_n \subseteq B_0(n2^{\alpha}) = \{v \in V_n : ||v|| < n2^{\alpha}\}.$$

The following lemma formally states that for the ϵ chosen in our construction, a hyperplane separating at least n pairs in G_n^i is 'close' to the hyperplane spanned by $U^i = (u^{ij})_j$.

Lemma 5.1 *Let f separate n pairs in G_n^i for $i \geq 3$. Then for each $v \in P_f \cap B_0(n2^{\alpha})$, $|v_i| < 2^{-\alpha}$.*

For the proof of this lemma we refer to appendix (B). Since for each pair in $\hat{H}_G$ and $G_n^{i'}$ for $i' \neq i$, and each of their separating points v, $|v_i| > 2^{-\alpha}$, we immediately obtain the following

Corollary 5.2 *Let f separate at least n pairs in G_n^i. Then f does not separate any pair neither in G_n^j for all $i \neq j$ nor in $\hat{H}_G$.*

Let $\phi(f)$ be consistent on $\hat{H}_G \sqcup G_n$. Since each pair in $\hat{H}_G$ and G_n has alternate labels, $P_{f_1}, \ldots, P_{f_k}$ separate each pair in $\hat{H}_G \sqcup G_n$. Let P_{f_1}, P_{f_2} separate points in $\hat{H}_G$. Then, due to Corollary 5.2, P_{f_1}, P_{f_2} separate at most $n-1$ pairs in each G_n^i. Since G_n^i has kn pairs, there is a hyperplane P_{f_l} separating at least n pairs in G_n^i. But, again, owing to Lemma 5.2, P_{f_l} separates no pair of $G_n^{i'}$ for each $i' \neq i$. Hence, each hyperplane of $P_{f_3}, \ldots, P_{f_k}$ separates at least n pairs of one of $G_n^3, \ldots, G_n^k$. I.e., $P_{f_3}, \ldots, P_{f_k}$ do not separate pairs in $\hat{H}_G$. •

Proof of Theorem 3.1 Let $\phi \in \mathcal{B}_k$ be a function depending on k' dimensions $i_1, \ldots, i_{k'}$, $2 \leq k' \leq k$. Then there exists $\phi' \in \mathcal{B}_{k'}$ depending on all dimensions, such that $\phi(b) = \phi'(b_{i_1}, \ldots, b_{i_{k'}})$ for all b. This implies that $\mathcal{F}^{k,\phi} = \mathcal{F}^{k',\phi'}$. Since $\textsc{Cons}(\mathcal{F}^{k',\phi'})$ is NP-complete (due to Theorem 3.2), also $\textsc{Cons}(\mathcal{F}^{k,\phi})$ is NP-complete. •

We consider a *multilayer feedforward perceptron (MLP)* $(n, h, n_1, \ldots, n_h)$ which is a neural network of linear threshold units, with h hidden layers with n_i units on the i-th layer, n_1 input units receiving data from V_n and one output unit. Each unit receives input from all units on the ancestral layer (see also [7]). One can consider an MLP as a two-layer neural network with n_1 linear classifiers realizing a collection Φ of boolean functions in $\mathcal{B}_{n_1}$. Clearly, each MLP can compute the AND function (if every unit on $n_2, \ldots, n_h$ and the output unit compute the AND) which depends on all dimensions. Furthermore, it is easy to see, that Φ is a well-behaving set. This gives us

Corollary 5.3 *Let $h \geq 1, 2 \leq n_1 \leq \frac{n-3}{2}, n_2, \ldots, n_h$ be polynomial in n. Then the consistency problem of the MLP $(n, h, n_1, \ldots, n_h)$ is NP-complete.*

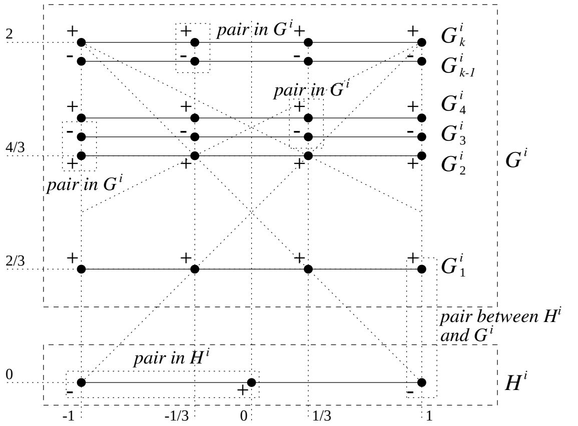

Figure 3: The examples and pairs on $H^i \cup G^i$

6 The approximation problem

Proof of Theorem 3.3 We apply Theorem 4.4 by constructing mappings (ρ_G) and (η_G) and verifying the two required properties. First, we address the construction of the mappings. For this purpose we consider a 2-graph $G = (V, E)$, $V = \{1, \ldots, n-1\}$. ρ_G maps an edge $I = \{i, j\} \in E$ to $H^{i,j} = H_G^I$ (see section 4). Moreover, let H^i consist of the labeled examples $(e^i, -1)$, $(-e^i, -1)$ and $(\bar{0}, 1)$ of $H^{i,j}$ that lie on the i-axis.

For η_G, consider the points $g_{i,r,\mu} = (\lambda_r e^n + \mu e^i)$ for $i = 1, \ldots, n$ and $r = 1, \ldots, k$, where $\lambda_1 = 2/3$ and $\lambda_r = \frac{2}{3}\left(2 + \frac{r-2}{k-2}\right)$ for $r \geq 2$. We define the samples $G^i = \bigsqcup_{r=1,\ldots,k} G_r^i$ on V_n where

$$G_r^i = (g_{i,r,\mu}, l_r)_{\mu \in \{-1, -\frac{1}{3}, \frac{1}{3}, 1\}},$$

and define the labels by $l_r = 1$ for all even r and $r = 1$, and $l_r = -1$ otherwise. Finally, η_G maps the edge $\{i, j\} \in E$ to $G^{i,j} = G^i \sqcup G^j$. Figure 3 (p. 6) illustrates the position of the examples in $(\bar{0}, \pm e^i) \sqcup G^i$. For all $i = 1, \ldots, n$, $r = 2, \ldots, k-1$, and $\mu \in \{\pm 1, \pm\frac{1}{3}\}$, we call $(g_{i,r,\mu}, g_{i,r+1,\mu})$ a *pair in G^i*. We call $(\pm e^i, \bar{0})$ *pairs in H^i* and $(\pm e^i, g_{i,1,\pm 1})$ *pairs between H^i and G^i*. Note that all pairs are labeled alternately and, therefore, have to be separated by at least one function of $f_1, \ldots, f_k$ for $\phi(f_1, \ldots, f_k)$ consistent with $H^i \sqcup G^i$.

We prove the first condition of Theorem 4.4: Let an arbitrary function $\tau : \{1, \ldots, n-1\} \to \{1, 2\}$ be given. Theorem 4.2 implies that there exist functions $f_1(v) = \text{sgn}(w^1 \cdot v + \theta_1), f_2(v) = \text{sgn}(w^2 \cdot v + \theta_2)$ such that $-\textsc{Xor}(f_1, f_2)$ is consistent with $H^{i,j}$ for all edges $\{i, j\} \in E$ that are 2-colored by τ (by simply removing the edges which are colored monochromatically). In this setting, w_n^1, w_n^2 can be arbitrary, since for all examples in $H^{i,j}$ the n-th component is 0. Set

$$w_n^1 = -w_n^2 = 4 \max_{\hat{\iota}=1,2,r=1,\ldots,n-1} (|w_r^{\hat{\iota}}| + |\theta_{\hat{\iota}}|).$$

Then for $|\lambda| \leq 1$, $w^1(\frac{2}{3}e^n + \lambda e^i) + \theta_1 \geq (\frac{2}{3} - \frac{1}{2})w_n^1 > 0$ and $w^2(\frac{2}{3}e^n + \lambda e^i) + \theta_2 \leq (-\frac{2}{3} + \frac{1}{2})w_n^1 < 0$. In other words,

$f_1(v) = -f_2(v) = 1$ for all examples v in $G^{i,j}$. For the remaining $\hat{\imath} = 3, \ldots, k$ we define $w^{\hat{\imath}} = e^n$ and $\theta_{\hat{\imath}} = -\frac{1}{3}(2+\frac{\hat{\imath}-3/2}{k-2})$. Obviously, if $\hat{\imath} \geq 3$, $f_{\hat{\imath}}(v) = -1$ for all $(v,l) \in H^{i,j}$ and $f_{\hat{\imath}}(g_{i,r,\mu}) = 1$ iff $\hat{\imath} \leq i$. Hence, $\text{PARITY}(f_1, \ldots, f_k)$ is consistent with $H^{i,j} \sqcup G^{i,j}$.

Now we will show that for each $i \neq j$, η_G employs $\mathcal{F}_n^{k,\Phi}$ for ρ_G. For this purpose, we need the following geometric observations, which can be easily verified (we refer to appendix (C)): For each threshold function f,

(a) f separates at most one pair in H^i and at most four pairs in G^i.

(b) if f separates a pair in H^i, it separates at most one pair in G^i or between H^i and G^i.

(c) If f separates 4 pairs in G^i, then f does not separate pairs in H^i and between H^i and G^i. If f additionally separates pairs in H^j, then f does not separate pairs between H^j and G^j.

(d) If f separates 8 pairs in $G^{i,j}$, then it does not separate pairs in $H^{i,j}$.

Let $F = \phi(f_1, \ldots, f_k) \in \mathcal{F}_n^{k,\Phi}$ be consistent with $H^{i,j} \sqcup G^{i,j}$. Assume that $\tilde{k}$ functions of $f_1, \ldots, f_k$ separate $H^i \sqcup H^j$.

If $\tilde{k} = 2$, due to (a), both functions separate pairs in H^i *and* H^j, since four pairs have to be separated. Due to (b), the functions separate at most 4 pairs in $G^{i,j}$ or between $H^{i,j}$ and $G^{i,j}$. Since the number of pairs in $G^{i,j}$ or between $H^{i,j}$ and $G^{i,j}$ is $8(k-2)+4$, together with (a), this implies that the remaining $k-2$ functions separate exactly 8 pairs each. (d) yields that non of the $k-2$ functions separates pairs in $H^{i,j}$.

If $\tilde{k} = 3$, at least one function f_1 separates pairs in H^i *and* H^j. Due to (b), this function separates at most 2 pairs in $G^{i,j}$ or between $H^{i,j}$ and $G^{i,j}$. The other two functions f_2, f_3 separate at most 5 pairs. Hence, at most 12 pairs are separated by the three functions altogether. Therefore, at least $8(k-2)-8 = 8(k-3)$ pairs in $G^{i,j}$ or between $H^{i,j}$ and $G^{i,j}$ must be separated by $k-3$ functions, i.e., with (a), each of the $k-3$ functions separates exactly 8 pairs. This implies that f_1 and f_2, f_3 separate exactly 2 and 5 pairs on G^i, respectively. (b), (c) imply that at most two pairs between $H^{i,j}$ and $G^{i,j}$ are separated, which contradicts the consistency.

If $\tilde{k} > 3$, then each function which separates pairs in $H^i \sqcup H^j$, separates at most 5 pairs in $G^{i,j}$ or between $H^{i,j}$ and $G^{i,j}$. Consequently, at least $8(k-2)+4-5\tilde{k}$ pairs have to be separated by $k-\tilde{k}$ functions. For $\tilde{k} = 4$, $8(k-2)+4-5\tilde{k} = 8(k-4)$, hence, exactly 8 pairs are separated by each of the $k-\tilde{k}$ functions while the four functions which separate pairs in $H^i \sqcup H^j$, separates exactly 5 pairs in $G^{i,j}$ or between $H^{i,j}$ and $G^{i,j}$. But due to (c), this implies that no pair between $H^{i,j}$ and $G^{i,j}$ is separated, in contradiction to our assumption. If $\tilde{k} > 4$, then $8(k-2)+4-5\tilde{k} > 8(k-\tilde{k})$, which implies that the $k-\tilde{k}$ functions cannot separate all remaining pairs, which is again a contradiction to the consistency.

Hence, at least $k-2$ functions are constant on $H^{i,j}$. Therefore, the assumptions of Theorem 4.4 are satisfied and, consequently, the proof completed. •

References

[1] Edoardo Amaldi and Viggo Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209:237–260, 1998.

[2] Sanjeev Arora, Laszlo Babai, Jaques Stern, and Z. Sweedyk. Hardness of approximate optima in lattices, codes, and linear systems. *Journal of Computer and System Sciences*, 54(2):317–331, 1997.

[3] Peter Bartlett and Shai Ben-David. Hardness results for neural network approximation problems. *Proceedings of the 4th European Conference on Computational Learning Theory (Lecture Notes in Artificial Intelligence)*, 4(1572):50–62, 1999.

[4] A.L. Blum and R.L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5(1):117–127, 1992.

[5] Bhaskar DasGupta, Hava T. Siegelmann, and Eduardo D. Sontag. On the complexity of training neural networks with continuous activation functions. *IEEE Transactions on Neural Networks*, 6(6):1490–1504, 1995.

[6] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, San Francisco, 1992.

[7] Barbara Hammer. Some complexity results for perceptron networks. *Proceedings of the 8th International Conference on Artificial Neural Networks*, 2:639–644, 1998.

[8] Johan Håstad. The size of weights for threshold gates. *SIAM J. Discrete Math.*, 7(3):484–492, 1994.

[9] Klaus-U. Höffgen, Hans U. Simon, and Kevin S. Van Horn. Robust trainability of single neurons. *Journal of Computer and System Sciences*, 50(1):114–125, 1995.

[10] Viggo Kann, Sanjeev Khanna, Jens Lagergren, and Alessandro Panconesi. On the hardness of approximating max-k-cut and its dual. *Technical Report CJTCS-1997-2, Chicago Journal of Theoretical Computer Science*, pages 317–331, 1997.

[11] C. H. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences*, 43:425–440, 1991.

[12] Prabhakar Raghavan. Learning in threshold networks. *Proceedings fo the 1988 Workshop on Computational Learning Theory*, pages 19–27, 1988.

[13] Michael Schmitt. *Komplexität neuronaler Lernprobleme*. Peter Lang, 1996.

[14] Hans-Ulrich Simon. A tight $\omega(\log \log n)$-bound on the time for parallel RAM's to compute nondegenerated boolean functions. *Information and Control*, 55(1-3):102–107, 1982.

A Appendix to Section 4

Proof of Theorem 4.4

We define two mappings ϱ, ξ and show that ϱ, ξ is an L-reduction from APPROX(2-CUT) to APPROX($\mathcal{F}^{k,\Phi_k}$). For the L-reduction (see e.g. [11]) we have to show the following two required conditions for constant α and β

L1 $opt_{\mathcal{F}^{k,\Phi}}(\varrho(G)) \leq \alpha opt_{2\text{-CUT}}(G)$

L2 $opt_{2\text{-CUT}}(G) - \mathcal{Z}((G), \xi_{S_G}(F))$
$\leq \beta(opt_{\mathcal{F}^{k,\Phi}}(\varrho(G)) - \mathcal{Z}(\varrho(G), F))$

With Theorem 2.2 and functions satisfying L1 and L2, we obtain that APPROX($\mathcal{F}^{k,\Phi}, \epsilon$) is NP-hard for $\epsilon < 1/(64\alpha\beta)$.

Let n be arbitrary and $V = \{1, \ldots, n\}$. W.l.o.g. let $|\eta_G(I)| = z$ for all (G) and all $I \in E$ (by adding copies of labeled instances, if necessary). With respect to a 2-graph (G) we define the sample

$$S_G = \sqcup_{I \in E} \left(H_G^I \sqcup \eta_G(I)\right)$$

Obviously, the length of the sample is $|S_G| = (6 + 2z)|E|$. So we define

$$\varrho : (G) \mapsto S_G.$$

Let $\xi_{S_G} = \xi(S_G, \cdot)$ be the function that maps $F \in \mathcal{F}^{k,\Phi_k}$ to the solution $\xi_{S_G}(F)$ of (G). For $F = \phi(f_1, \ldots, f_k)$ we define

$$\xi_{S_G}(F)(i) = \left\{ \begin{array}{ll} 2, & \text{if } F(e^i) = F(-e^i) = f_{l_i}(e^i) = -1 \\ 1, & \text{otherwise} \end{array} \right.$$

where l_i is the smallest index, for which $f_{l_i}(e^i) \neq f_{l_i}(-e^i)$.

Now we will show that the inequalities L1 and L2 are satisfied and, therefore, ϱ, ξ is an L-reduction.

Properties of ϱ and ξ_{S_G}

- *Let $I = \{i_1, i_2\} \in E$. If $\xi_{S_G}(F)(i_1) = \xi_{S_G}(F)(i_2)$, then F is not consistent with $H_G^I \sqcup \eta_G(I)$. Inparticular,*

  $$\mathcal{Z}(S_G, F) \leq \frac{(5 + 2z)|E| + |E|\mathcal{Z}((G), \xi_{S_G}(F))}{|S_G|}. \quad (1)$$

  Certainly, every pair of examples with alternating labels must be separated by at least one of the functions $f_1, \ldots, f_k$. Since every line that contains such a pair contains exactly one separating point, $e^i, \bar{0}$ and $-e^i, \bar{0}$ cannot be separated by the same function, for all i.

  Assume that $l_{i_1} \neq l_{i_2}$. W.l.o.g. let $l_{i_1} < l_{i_2}$. Then $f_{l_{i_2}}(e^{i_1}) = f_{l_{i_2}}(-e^{i_1})$, since otherwise $l_{i_2} \leq l_{i_1}$. In other owrds, there exists i_3 different from i_1, i_2, such that $f_{l_{i_3}}(e^{i_1}) = f_{l_{i_3}}(-e^{i_1})$. Hence, there are at least three functions that are constant on H_G^I. This leads to a contradiction of the assumption that $\eta_G(I)$ employs $\mathcal{F}^{k,\Phi}$ for H_G^I.

  We consider the case $l_{i_1} = l_{i_2} =: l$ and assume that F is consistent with H_G^I. I.e. $F(e^{i_j}) = F(-e^{i_j}) = -1$ for $j = 1, 2$. Since $\xi_{S_G}(F)(i_1) = \xi_{S_G}(F)(i_2)$, the definition of ξ_{S_G} yields $f_l(e^{i_1}) = f_l(e^{i_2})$. The assumption that $\eta_G(I)$ employs $\mathcal{F}^{k,\Phi}$ for H_G^I gives us that at least $k - 2$ functions of $f_1, \ldots, f_k$ are constant on H_G^I. Consequently, there exists $l' \neq l$ and a function $\psi \in \mathcal{B}_2$ with $F \equiv \psi(f_l, f_{l'})$. Since $f(e^{i_1}) = f(e^{i_2})$, Lemma 4.1 gives us that F is not consistent on H_G^I, in condradiction to the assumption.

  We have therefore shown that for every edge $I \in E$ colored monochromatically by ξ_{S_G}, at least on example of F is not classified correctly. Hence, also inequality 1 is proven.

- *For all colorings τ there exists a function F satisfying*

  $$\mathcal{Z}(S_G, F) \geq \frac{(5 + 2z)|E| + |E|\mathcal{Z}((G), \tau)}{|S_G|} \quad (2)$$

  This inequality is a consequence of the first condition, i.e. there exists a function F such that F is consistent with $\eta_G(E)$ ($2z|E|$ examples), with $H_G^{\{i\}}$ for all $i \in V$ ($5|E|$ examples) and with H_G^I for all edges I that are 2-colored by τ (additionally at least $|E|\mathcal{Z}((G), \tau)$ examples).

With the above inequalities 1 and 2 we obtain

$$opt_{\mathcal{F}^{k,\Phi}}(S_G) = \frac{(4 + 2z)|E| + |E|opt_{2\text{-CUT}}(G)}{|S_G|} \quad (3)$$

Since $opt_{2\text{-CUT}}(G) \geq 1/2$, this yields

$$\begin{array}{rcl} opt_{\mathcal{F}^{k,\Phi}}(S_G) & \leq & \frac{2(4+2z)|E|opt_{2\text{-CUT}}(G)+|E|opt_{2\text{-CUT}}(G)}{(6+2z)|E|} \\ & = & 2\frac{(6+2z)|E|-|E|/2}{(6+2z)|E|}opt_{2\text{-CUT}}(G) \\ & = & 2\left(1 - \frac{|E|}{2|S_G|}\right) opt_{2\text{-CUT}}(G) \\ & \leq & 2\ opt_{2\text{-CUT}}(G) \end{array}$$

Hence, inequality L1 is satisfied for $\alpha = 2$. By simple transformation of 1 and 3 we inparticularly obtain

$$\begin{array}{rcl} opt_{2\text{-CUT}}(G) & = & \dfrac{|S_G|opt_{\mathcal{F}^{k,\Phi}}(S_G) - a - |E|(4 + 2z)}{|E|} \\ \mathcal{Z}(G, \xi_{S_G}(F)) & \geq & \dfrac{|S_G|\mathcal{Z}(S_G, F) - a - |E|(4 + 2z)}{|E|} \end{array}$$

Therefore,

$$\begin{array}{l} opt_{2\text{-CUT}}(G) - \mathcal{Z}(G, \xi_{S_G}(F)) \\ \leq \dfrac{|S_G|}{|E|}\Big(opt_{\mathcal{F}^{k,\Phi}}(S_G) - \mathcal{Z}(S_G, F)\Big), \end{array}$$

which satisfies the equality L2 for $\beta = \frac{|S_G|}{|E|} = 6 + 2z$. Since MAX 2-CUT is NP-hard with error-rate $\epsilon \leq 1/64$, MAX($\mathcal{F}^{k,\Phi}$) is NP-hard with error-rate

$$\tilde{\epsilon} \leq \epsilon/(\alpha\beta) = \frac{1}{64(6 + 2z)} = \frac{1}{384 + 128z}.$$

•

B Appendix to Section 5

The algorithm V-SYSTEM Basically the algorithm has to compute k polynomials of degree $\leq 2k+2$ and evaluate them at a number of points.

Let $v(x) = (p_1(x), \ldots, p_n(x))^T$, where $p_1, \ldots, p_n$ are linear independent polynomials (i.e. $\sum_i \lambda_i p_i \equiv 0 \Rightarrow \lambda_i = 0, \forall i$) of degree $\leq n-1$.

Lemma B.1 *For arbitrary $x_1, \ldots, x_n \in K$ pairwise different, $v(x_1), \ldots, v(x_n)$ are linearly independent.*

Proof Consider the matrix $(v(x_1), \ldots, v(x_n))$ and its row-vectors $\tilde{v}_i = (p_i(x_1), \ldots, p_i(x_n))$, $i = 1, \ldots, n$. Assume that $\sum_i \lambda_i \tilde{v}_i = 0$, i.e. $\sum_i \lambda_i p_i(x_j) = 0 \forall j = 1, \ldots, n$. Since $\sum_i \lambda_i p_i$ is a polynomial of degree $\leq n-1$, $\sum_i \lambda_i p_i \equiv 0$ and since $p_1, \ldots, p_n$ are linearly independent, $\lambda_i = 0 \forall i$. Hence, the matrix $(v(x_1), \ldots, v(x_n))$ has rank n and, therefore, the vectors are linearly independent. •

We will construct U with a system using polynomials which we will represent as products of the linear factors $\pi_r(x) = x - r$. Obviously, $\pi_r(x) < 0$, iff $x < r$, $\pi_r(x) > 0$, iff $x > r$ and $\pi_r(x) = 0 \Leftrightarrow x = r$.

Consider the following n polynomials $p_1, \ldots, p_n$ on the interval $[0, \ldots, 2(k+1)]$

$$\begin{array}{ll} p_l = \pi_{2l} \left(\prod_{r \in J_l} \pi_{2r+1}\right) \pi_0^{k+l+1-|J_l|} & \text{for } l = 1, \ldots, k \\ p_l = \pi_0^{l-k-1} & \text{for } l = k+1, \ldots, 2k+3 \\ p_l = \pi_0^{l-1} & \text{for } l = 2k+4, \ldots, n \end{array}$$

with $J_l \subseteq \{1, \ldots, k\}$, where

$$r \in J_l \Leftrightarrow \left\{ \begin{array}{llll} & (r \notin \{l-1, l\} & \text{and} & b_l^r = -b_l^{r+1}) \\ \text{or} & (r = l-1 & \text{and} & b_l^r = b_l^{r+2}). \end{array} \right.$$

The above $b_l^{k+1} = 1$ for all l. Obviously, the polynomials x^{l-k-1} for $l = k+1, \ldots, 2k+3$, x^{l-1} for $l = 2k+4, \ldots, n$ and $p_1, \ldots, p_k$ are n polynomials with pairwise different degree $\leq n-1$ and are, therefore linearly independent.

Lemma B.2 $p_l(x_0) = 0$ *and sgn*$(p_l(x) - p_i(x)) = b_l^i$ *for all* $|x - x_i| \leq \frac{1}{2}(2k+1)^{-(2k+2)}$, $l \neq i$.

Proof It is easy to see that our construction yields $p_l(x_0) = 1$, and sgn$(p_l(x_i) - p_i(x_i)) = b_l^i$ for $l \neq i$. Observe that

$$\begin{array}{rcl} |p_i(x) - 1| & \leq & (2k+1)^{2k+2}|x - x_i| \\ |p_l(x) - 1| & \geq & \min(|x - s_{i-1}|, |x - s_i|) \end{array}$$

for $|x - x_i| \leq 1$ and $l \neq i$. Hence,

$$\begin{array}{rcl} |p_i(x) - 1| & \leq & 1/2 \\ |p_l(x) - 1| & > & 1/2 \end{array}$$

for $|x - x_i| = \frac{1}{2}(2k+1)^{-(2k+2)}$. Therefore, sgn$(p_l(x) - p_i(x)) = b_l^i$ for $|x - x_i| \leq \frac{1}{2}(2k+1)^{-(2k+2)}$. •

Let $x_{ij} = 2i - \frac{1}{2j}(2k+1)^{-(2k+2)}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, kn$ which obviously satisfies $|x_{ij} - x_i| < \frac{1}{2}(2k+1)^{-(2k+2)}$ for all i, j. Due to Lemma B.2, property (U3) is satisfied. Finally, we obtain the vectors

$$u^{ij} = p_i(0)v(x_{ij}) - p_i(x_{ij})v(0) = v(x_{ij}) - p_i(x_{ij})\bar{1}.$$

Consequently, $u_i^{ij} = p_i(x_{ij}) - p_i(x_{ij}) = 0$ and sgn$(u_l^{ij}) =$ sgn$(p_l(x_{ij}) - p_i(x_{ij})) = b_l^i$ which gives us properties (U1) and (U2). To show property (U4), consider the representation $x_{ij} = \frac{m_1}{m_2}$ with integers $m_1 = 4ij(2k+1)^{2k+1} - 1$ and $m_2 = 2j(2k+1)^{2k+1}$. Obviously, $m_1 \leq m$, and $2km_2 \leq m$ for $4n(2k+1)^{2k+4} < n^{3n} =: m$. Then

$$p_l(x_{i,j}) = \left\{ \begin{array}{l} \frac{1}{(n_{i,j})^{k+l+2}} \Big[(m_{i,j} - 2ln_{i,j}) \\ \left(\prod_{r \in J_r} (m_{i,j} - (2r+1)n_{i,j})\right) \\ (m_{i,j})^{k+l+1-|J_l|} \Big] \text{ for } l = 1, \ldots, k \\ \frac{(m_{i,j})^{k-l-1}}{(n_{i,j})^{k-l-1}} \text{ for } l = k+1, \ldots, 2k+3 \\ \frac{(m_{i,j})^{l-1}}{(n_{i,j})^{l-1}} \text{ for } l = 2k+4, \ldots, n. \end{array} \right.$$

Obviously, numerator and denominator of the above representation are integers bounded by n^{3n^2}, i.e. $size(p_l(x_{ij})) \leq 1 + 2\lceil 3n^2 \log n \rceil$. Hence, $size(p_l(x_{ij}) - p_i(x_{ij})) \leq 6 + 12n^2\lceil \log n \rceil =: \alpha(n) = O(n \log n) = poly(n)$.

First, V-SYSTEM computes $p_1, \ldots, p_k$ where each polynomial needs $O(k)$ comparisons. Then the algorithm evaluates p_l at x_{ij} for all l, i, j. Since x_{ij} has a polynomial binary representation length, the algorithm is polynomial in k and n.

Preliminaries for Lemma 5.1 Note that for x with binary representation length $size(x)$, we obtain the bounds $|x| \leq 2^{size(x)}$ and, if $x \neq 0$, $|x| \geq 2^{-size(x)}$ According to property (U4), $size(u_l^{ij}) \leq \alpha$. We will use the standard determinant det and $V(v^1, \ldots, v^i) = \det(v^i \cdot v^j)_{i,j=1\ldots,i}$ which is the square of the standard volume form. Consider an arbitrary set of vectors $u^0, \ldots, u^{n-1}$ of U. Since the vectors are in general position, $u^1 - u^0, \ldots, u^i - u^0$ are linearly independent. A straight forward calculation shows that $size(u_l^i - u_l^0) \leq 3\alpha$ and $size((u^i - u^0) \cdot (u^j - u^0)) \leq 6n\alpha$ for all i, j. Further, $size(V(u^1 - u^0, \ldots, u^i - u^0)) \leq 27n^4\alpha = poly(n)$. Since $u^0, \ldots, u^{n-1}$ are in general position, $2^{-27n^4\alpha} \leq V(u^1 - u^0, \ldots, u^i - u^0) \leq 2^{27n^4\alpha}$.

Lemma B.3 *Let $v^1, \ldots, v^n, w^1, \ldots, w^n \in V_n$ with $|v_l^i| \leq \delta$ and $|w_l^i| \leq \lambda$. Then* $|\det(v^1 + w^1, \ldots, v^n + w^n)| \geq |\det(v^1, \ldots, v^n)| - \delta 2^{2n^2\alpha}$

Proof Straightforward analysis. •

A system $\tilde{U} = (\tilde{u}^{ij})_{ij}$ is called *ϵ-close*, iff for each i, j, $\tilde{u}_{ij} = u_{ij} + \delta_{ij}e^i$, $|\delta_{ij}| \leq \epsilon$.

Lemma B.4 *Let $\tilde{U}$ ϵ-close to U with $\epsilon = 2^{-50n^4\alpha}$, and let $\tilde{u}^0, \ldots, \tilde{u}^{n-1}$ vectors of $\tilde{U}$. Then $V(\tilde{u}^1 - \tilde{u}^0, \ldots, \tilde{u}^{n-1} - \tilde{u}^0) > 2^{-28n^4\alpha}$, in particular, $\tilde{u}^0, \ldots, \tilde{u}^{n-1}$ are in general position.*

Proof

$$(\tilde{u}^i - \tilde{u}^0) \cdot (\tilde{u}^j - \tilde{u}^0) = (u^i - u^0) \cdot (u^j - u^0) + \lambda^i_j$$

with

$$\begin{aligned} \lambda^i_j &= \delta_i(e^i - u^0) \cdot (u^j - u^0) + \delta_0(\tilde{u}^i - e^0) \cdot (u^j - u^0) \\ &+ \delta_i(\tilde{u}^i - \tilde{u}^0) \cdot (e^j - u^0) + \delta_0(\tilde{u}^i - \tilde{u}^0) \cdot (\tilde{u}^j - e^0) \end{aligned}$$

This implies $|\lambda^i_j| \leq 4\epsilon 2^{9n\alpha}$. With Lemma B.3 we obtain

$$\begin{aligned} & V(\tilde{u}^1 - \tilde{u}^0, \ldots, \tilde{u}^{n-1} - \tilde{u}^0) \\ & \geq V(u^1 - u^0, \ldots, u^i - u^0) - \frac{1}{2} 2^{-27n^4\alpha} > 0 \end{aligned}$$

if $4\epsilon 2^{9n\alpha} \leq \frac{1}{2} 2^{-5n^3 9n\alpha} = \frac{1}{2} 2^{-45n^4\alpha}$, i.e. inparticular if $\epsilon \leq 2^{-50n^4\alpha} \leq 2^{-45n^4\alpha - 9n\alpha - 2}$. •

Proof of Lemma 5.1 Let f intersect pairs $(u^0 \pm \epsilon e^i)$, $\ldots$, $(u^{n-1} \pm \epsilon e^i)$, where $u^0, \ldots, u^{n-1} \in U^i$. Then

$$\begin{aligned} P_f &= \Big\{ v = \tilde{u}^0 + \lambda_1 v^1 + \cdots + \lambda_{n-1} v^{n-1} : \\ & \qquad v^j = \tilde{u}^j - \tilde{u}^0, \lambda_j \in K, j = 1, \ldots n-1 \Big\} \end{aligned}$$

where $\tilde{u}^j$ are separating points of the pairs $(u^j \pm \epsilon e^i)$, $j = 0, \ldots, n-1$. Assuming $v \in P_f \cap B_0(n2^\alpha)$, we have

$$2n^\alpha > ||\lambda_1 v^1 + \cdots + \lambda_{n-1} v^{n-1}|| \geq \lambda_j ||(v^j)'||, \qquad (4)$$

where $(v^j)'$ is the orthogonal projection of vector v^j on the hyperplane $[v^1, \ldots, v^{j-1}, v^{j+1}, \ldots, v^{n-1}]^\perp$. From Lemma B.4 an the preliminaries we obtain the bounds $2^{-28n^4\alpha} \leq V(v^1, \ldots, v^{j-1}, (v^j,) v^{j+1}, \ldots, v^{n-1}) \leq 2^{27n^4\alpha}$ for $i = 1, 2$. Since

$$V(v^1, \ldots, v^{n-1}) = V(v^1, \ldots, v^{j-1}, v^{j+1}, \ldots, v^{n-1}) ||(v^j)'||^2,$$

we have $||(v^j)'|| \geq 2^{-28n^4\alpha} > 0$. Applying this bound to 4, we obtain

$$|\lambda_j| \leq 2n 2^{\alpha + 28n^4\alpha} \leq 2^{30n^4\alpha}.$$

Finally,

$$\begin{aligned} |v_i| &= |\tilde{u}^0_i + \lambda_1(\tilde{u}^1_i - \tilde{u}^0_i) + \cdots + \lambda_{n-1}(\tilde{u}^{n-1}_i - \tilde{u}^0_i)| \\ &\leq |\tilde{u}^0_i| + \lambda_1 |\tilde{u}^1_i - \tilde{u}^0_i| + \cdots + \lambda_{n-1} |\tilde{u}^{n-1}_i - \tilde{u}^0_i| \\ &\leq \epsilon + 2\epsilon(|\lambda_1| + \cdots + |\lambda_{n-1}|) \\ &< 6n^2 2^{30n^4\alpha} \epsilon < 2^{30n^4\alpha - 50n^4\alpha} \\ &< 2^{-\alpha}. \end{aligned}$$

•

C Appendix to Section 6

We proof the claims used in section 6. For a vektor v let $[v] = \{\lambda v : \lambda \in K\}$ be the line that contains v. Observe that the pairs in H^i lie on the line $[e^i]$, and the pairs in G^i lie on the four lines $\mu e^i + [e^n]$, $\mu = \pm 1, \pm \frac{1}{3}$. For a line $v + [u]$:

Lemma C.1 *Let $f(v) = sgn(w \cdot v + \theta)$. Either $w \cdot (v + [u])$ is constant, or there is a unique $\lambda \in K$ with $w \cdot (v + \lambda u) + \theta = 0$.*

Proof Assume there exist $\lambda_1 \neq \lambda_2$ with $w \cdot (v + \lambda_1 u) + \theta = w \cdot (v + \lambda_2 u) + \theta = 0$. Then $w \cdot ((\lambda_2 - \lambda_1) u) = 0$, which implies $w \cdot ([u]) = 0$ and, hence, $w \cdot (v + [u])$ is constant. •

We call two points v_1, v_2 of a set S *neighboured*, if there is no point $v_0 \in S$ on the line between them, i.e. $S \cap \{v = \lambda v_1 + (1-\lambda) v_2 : 0 < \lambda < 1\} = \emptyset$. Note that all pairs defined on H and G are neighboured. Then we have the following

Corollary C.2 *At most one pair of a collection of neighboured pairs lying on one line can be separated by a linear threshold function f.* •

Now we are ready to show the following claims: Let f be a threshold function on V_n. Then

(a) *f_l separates at most one pair in H^i and at most four pairs in G^i.*

Since all pairs lie on one and four lines, respectively, Corollary C.2 proves this claim.

(b) *f_l which separates a pair in H^i, separates at most one pair in G^i or between H^i and G^i.*

Assume that f separates a pair in H^i, i.e. $f(e^i) = -f(-e^i)$. Without loss of generality let $f(e^i) = 1$, i.e.

$$\begin{aligned} w \cdot e^i + \theta &> 0 & (5) \\ w \cdot (-e^i) + \theta &\leq 0. & (6) \end{aligned}$$

1. Assume further that f separates two pairs in G^i, i.e. for $\mu_1 \neq \mu_2$ with $|\mu_1|, |\mu_2| \leq 1$, $f(\mu_l e^i + \frac{4}{3} e^n) = -f(\mu_l e^i + 2e^n), l = 1, 2$. We will obtain a contradiction.

Case: $f(\mu_l e^i + 2e^n) = 1, l = 1, 2$, i.e.

$$\begin{aligned} w \cdot (\mu_l e^i + 2e^n) + \theta = \mu_l w_i + 2w_n + \theta &> 0 (7) \\ w \cdot (\mu_l e^i + \tfrac{4}{3} e^n) + \theta = \mu_l w_i + \tfrac{4}{3} w_n + \theta &\leq 0 (8) \end{aligned}$$

Then 5 and 8 yield

$$\tfrac{4}{3} w_n < (1 - \mu_l) w_i. \qquad (9)$$

The difference of 7 and 8 give us

$$|\mu_2 - \mu_1| w_i < \tfrac{2}{3} w_n. \qquad (10)$$

Let wlog $\mu_2 > \mu_1$. 9 and 10 imply $(\mu_2 - \mu_1) < \frac{1}{2}(1 - \mu_2)$. Since for different μ_l, $\mu_2 - \mu_1 \geq \frac{2}{3}$, we have $\frac{2}{3} < \frac{1}{2}(1 - \mu_2)$ which is equivalent to $\mu_2 < -\frac{1}{3}$. Since $-1 \leq \mu_1 < \mu_2 < -\frac{1}{3}$, $\mu_2 - \mu_1 < \frac{2}{3}$, which is a contradiction.

Case: $f(\mu_l e^i + 2e^n) = -1, l = 1, 2$, is analogous to the previous case.

Case: $f(\mu_1 e^i + 2e^n) = 1$ and $f(\mu_2 e^i + 2e^n) = -1$, i.e. $\mu_1 w_i + 2w_n + \theta > 0$, $\mu_1 w_i + \frac{4}{3} w_n + \theta \leq 0$, and $\mu_2 w_i + 2w_n + \theta \leq 0$, $\mu_2 w_i + \frac{4}{3} w_n + \theta > 0$. The difference of the first and third inequality yields $(\mu_1 - \mu_2) w_i > 0$, and the difference of the second and fourth inequality implies $(\mu_1 - \mu_2) w_i < 0$ which is obviously a contradiction.

Case: $f(\mu_1 e^i + 2e^n) = -1$ and $f(\mu_1 e^i + 2e^n) = 1$ is, again, analogous to the previous case.

2. Assume now that f separates a pair between H^i and G^i. Note that 5 and 6 implies $w_i > 0$ and $w_i \geq |\theta|$.

Case: f separates a pair $(e^i, e^i + \frac{2}{3}e^n)$, i.e. $f(e^i + \frac{2}{3}e^n) = -1$ which gives $w_i + \theta > 0$ and $w_i + \frac{2}{3}w_n + \theta \leq 0$. Then $w_n < 0$ and $\mu w_i + \lambda w_n \leq 0$ for all $-1 \leq \mu \leq 1$ and $\lambda > 2/3$. But this implies that f does not separate a pair of the form $(\mu w_i + \frac{4}{3}w_n, \mu w_i + 2w_n)$ and, in particular, pairs in G^i.

Case: $f(-e^i + \frac{2}{3}e^n) = 1$ is analogous to the previous case.

(c) *If f_l separates 4 pairs in G^i, then f_l does not separate pairs in H^i or between H^i and G^i.If f_l additionally separates pairs in H^j, then f_l does not separate pairs between H^j and G^j.*

Assume that f separates four pairs in G^i. Then

$$f(e^i + \tfrac{4}{3}e^n) \quad = \quad -f(e^i + 2e^n) \tag{11}$$

$$f(-e^i + \tfrac{4}{3}e^n) \quad = \quad -f(-e^i + 2e^n). \tag{12}$$

Case: If $f(e^i + \frac{4}{3}e^n) = -f(-e^i + \frac{4}{3}e^n)$, then $f(e^i + 2e^n) = -f(-e^i + 2e^n)$, and we obtain $w_i > 0$ and $w_i < 0$ from the first and second equation, respectively, which is a contradiction.

Case: $f(e^i + \frac{4}{3}e^n) = f(-e^i + \frac{4}{3}e^n) = -1$. Then

$$w_i + \tfrac{4}{3}w_n + \theta \quad \leq \quad 0 \tag{13}$$

$$-w_i + \tfrac{4}{3}w_n + \theta \quad \leq \quad 0 \tag{14}$$

$$w_i + 2w_n + \theta \quad > \quad 0 \tag{15}$$

$$-w_i + 2w_n + \theta \quad > \quad 0. \tag{16}$$

$(15) - (13)$ implies $w_n > 0$. $(16) - (13)$ and $(15) - (14)$ imply $|w_i| \leq (1/3)w_n$. $(14) - (13)$ implies $\theta \leq -\frac{4}{3}w_n$. Hence, for $-1 \leq \mu \leq 1$, $\lambda \leq 2/3$,

$$\mu w_i + \lambda w_n + \theta \leq w_n - \tfrac{4}{3}w_n < 0.$$

Consequently, f does not separate pairs in H^i or between H^i and G^i.

Case: $f(e^i + \frac{4}{3}e^n) = f(-e^i + \frac{4}{3}e^n) = 1$ is analogous to the previous case.

Assume now that f additionally separates a pair in H^j.

Case: If $f(e^i + \frac{4}{3}e^n) = -f(-e^i + \frac{4}{3}e^n)$, we obtain a contradiction (see the first case in (c)).

Case: $f(e^i + \frac{4}{3}e^n) = f(-e^i + \frac{4}{3}e^n) = -1$. Then for all $-1 \leq \mu \leq 1$, $f(\mu e^i + \frac{4}{3}e^n) = -f(\mu e^i + 2e^n) = -1$, in particular, for $\mu = 0$. Then the proof of (b) yields the claim.

(d) *If f_l separates 8 pairs in $G^{i,j}$, then it does not separate pairs in $H^{i,j}$.*

Due to (c) f does not separate pairs in H^i and H^j. We still have to check the pairs $(e^i, e^{\{i,j\}})$ and $(e^j, e^{\{i,j\}})$, which are in $H^{i,j}$ and not in $H^i \cup H^j$. Assume that f separates 8 pairs in $G^{i,j}$.

Case: $f(e^l + \frac{4}{3}e^n) = -f(-e^l + \frac{4}{3}e^n) = -1$ for $l \in \{i, j\}$ leads to a contradiction analogous to the proof of (c).

Case: $f(e^l + \frac{4}{3}e^n) = f(-e^l + \frac{4}{3}e^n) = -1$, $l = i, j$. Then inequalities 13-16 are valid for i and j. Again, $(15) - (13)$ implies $w_n > 0$. $(16) - (13)$ and $(15) - (14)$ imply $|w_l| \leq (1/3)w_n$, $l = i, j$. $(14) - (13)$ implies $\theta \leq -\frac{4}{3}w_n$.

Hence, for arbitrary $-1 \leq \mu_1, \mu_2 \leq 1$,

$$\mu_1 w_i + \mu_2 w_j + \theta \leq \tfrac{2}{3}w_n - \tfrac{4}{3}w_n = -\tfrac{2}{3}w_n < 0,$$

i.e. $f(v) = -1$ for all points in $H^{i,j}$. Therefore, no pair in $H^{i,j}$ is separated.

The other cases are similar.