J. Yaghob (Ed.): ITAT 2015 pp. 167–171
Charles University in Prague, Prague, 2015

Limitations of One-Hidden-Layer Perceptron Networks

Věra Kůrková

Institute of Computer Science, Czech Academy of Sciences,
vera@cs.cas.cz,
WWW home page: http://www.cs.cas.cz/∼vera

Abstract: Limitations of one-hidden-layer perceptron networks to represent efficiently finite mappings is investigated. It is shown that almost any uniformly randomly chosen mapping on a sufficiently large finite domain cannot be tractably represented by a one-hidden-layer perceptron network. This existential probabilistic result is complemented by a concrete example of a class of functions constructed using quasi-random sequences. Analogies with central paradox of coding theory and no free lunch theorem are discussed.

1 Introduction

A widely-used type of a neural-network architecture is a network with one-hidden-layer of computational units (such as perceptrons, radial or kernel units) and one linear output unit. Recently, new hybrid learning algorithms for feedforward networks with two or more hidden layers, called deep networks [9, 3], were successfully applied to various pattern recognition tasks. Thus a theoretical analysis identifying tasks for which shallow networks require considerably larger model complexities than deep ones is needed. In [4, 5], Bengio et al. suggested that a cause of large model complexities of shallow networks with one hidden layer might be in the "amount of variations" of functions to be computed and they illustrated their suggestion by an example of representation of d-dimensional parities by Gaussian SVM.

In practical applications, feedforward networks compute functions on finite domains in $\mathbb{R}^d$ representing, e.g., scattered empirical data or pixels of images. It is well-known that shallow networks with many types of computational units have the "universal representation property", i.e., they can exactly represent any real-valued function on a finite domain. This property holds, e.g., for networks with perceptrons with any sigmoidal activation function [10] and for networks with Gaussian radial units [15]. However, proofs of universal representation capabilities assume that networks have numbers of hidden units equal to sizes of domains of functions to be computed. For large domains, this can be a factor limiting practical implementations. Upper bounds on rates of approximation of multivariable functions by shallow networks with increasing numbers of units were studied in terms of variational norms tailored to types of network units (see, e.g., [11] and references therein).

In this paper, we employ these norms to derive lower bounds on model complexities of shallow networks representing finite mappings. Using geometrical properties of high-dimensional spaces we show that a representation of almost any uniformly randomly chosen function on a "large" finite domain by a shallow perceptron networks requires "large" number of units or "large" sizes of output weights. We illustrate this existential probabilistic result by a concrete construction of a class of functions based on Hadamard and quasi-noise matrices. We discuss analogies with central paradox of coding theory and no free lunch theorem.

The paper is organized as follows. Section 2 contains basic concepts and notations on shallow networks and dictionaries of computational units. Section 3 reviews variational norms as tools for investigation of network complexity. In Section 3, estimates of probabilistic distributions of sizes of variational norms are proven. In section 4, concrete examples of functions which cannot be tractably represented by perceptron networks are constructed using Hadamard and pseudo-noise matrices. Section 5 is a brief discussion.

2 Preliminaries

One-hidden-layer networks with single linear outputs (*shallow networks*) compute input-output functions from sets of the form

$$\operatorname{span}_n G := \left\{ \sum_{i=1}^{n} w_i g_i \,|\, w_i \in \mathbb{R},\, g_i \in G \right\},$$

where G, called a *dictionary*, is a set of functions computable by a given type of units, the coefficients w_i are called output weights, and n is the number of hidden units. This number is sometimes used as a measure of *model complexity*.

In this paper, we focus on representations of functions on finite domains $X \subset \mathbb{R}^d$. We denote by

$$\mathscr{F}(X) := \{f \,|\, f : X \to \mathbb{R}\}$$

the *set of all real-valued functions on* X. On $\mathscr{F}(X)$ we have the Euclidean inner product defined as

$$\langle f, g \rangle := \sum_{u \in X} f(u) g(u)$$

and the Euclidean norm

$$\|f\| := \sqrt{\langle f, f \rangle}.$$

To distinguish the inner product $\langle .,. \rangle$ on $\mathscr{F}(X)$ from the inner product on $X \subset \mathbb{R}^d$, we denote it $\cdot$, i.e., for $u, v \in X$,

$$u \cdot v := \sum_{i=1}^{d} u_i v_i.$$

We investigate networks with units from the dictionary of *signum perceptrons*

$$P_d(X) := \{\mathrm{sgn}(v \cdot . + b) : X \to \{-1,1\} \,|\, v \in \mathbb{R}^d, b \in \mathbb{R}\}$$

where $\mathrm{sgn}(t) := -1$ for $t < 0$ and $\mathrm{sign}(t) := 1$ for $t \geq 0$. Note that from the point of view of model complexity, there is only a minor difference between networks with signum perceptrons and those with Heaviside perceptrons as

$$\mathrm{sgn}(t) = 2\vartheta(t) - 1$$

and

$$\vartheta(t) := \frac{\mathrm{sgn}(t)+1}{2},$$

where $\vartheta(t) := 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$.

3 Model Complexity and Variational Norms

A useful tool for derivation of estimates of numbers of units and sizes of output weights in shallow networks is the concept of a variational norm tailored to network units introduced in [12] as an extension of a concept of variation with respect to half-spaces from [2]. For a subset G of a normed linear space $(\mathscr{X}, \|.\|_{\mathscr{X}})$, *G-variation (variation with respect to the set G)*, denoted by $\|.\|_G$, is defined as

$$\|f\|_G := \inf\{c \in \mathbb{R}_+ \mid f/c \in \mathrm{cl}_{\mathscr{X}} \mathrm{conv}\,(G \cup -G)\},$$

where $\mathrm{cl}_{\mathscr{X}}$ denotes the closure with respect to the norm $\|\cdot\|_{\mathscr{X}}$ on $\mathscr{X}$, $-G := \{-g \mid g \in G\}$, and

$$\mathrm{conv}\, G := \left\{ \sum_{i=1}^{k} a_i g_i \mid a_i \in [0,1], \sum_{i=1}^{k} a_i = 1, g_i \in G, k \in \mathbb{N} \right\}$$

is the convex hull of G. The following straightforward consequence of the definition of G-variation shows that in all representations of a function with "large" G-variation by shallow networks with units from the dictionary G, the number of units must be "large" or absolute values of some output weights must be "large".

Proposition 1. *Let G be a finite subset of a normed linear space $(\mathscr{X}, \|.\|_{\mathscr{X}})$, then for every $f \in \mathscr{X}$,*

$$\|f\|_G = \min\left\{ \sum_{i=1}^{k} |w_i| \,\middle|\, f = \sum_{i=1}^{k} w_i g_i, w_i \in \mathbb{R}, g_i \in G \right\}.$$

Note that classes of functions defined by constraints on their variational norms represent a similar type of a concept as classes of functions defined by constraints on both numbers of gates and sizes of output weights studied in theory of circuit complexity [16].

To derive lower bounds on variational norms, we use the following theorem from [13] showing that functions which are "not correlated" to any element of the dictionary G have large variations.

Theorem 2. *Let $(\mathscr{X}, \|.\|_{\mathscr{X}})$ be a Hilbert space with inner product $\langle .,. \rangle_{\mathscr{X}}$ and G its bounded subset. Then for every $f \in \mathscr{X} - G^{\perp}$,*

$$\|f\|_G \geq \frac{\|f\|^2}{\sup_{g \in G} |\langle f, g \rangle_{\mathscr{X}}|}.$$

The following theorem shows that when a dictionary $G(X)$ is not "too large", then for a "large" domain X, almost any randomly chosen function has large $G(X)$-variation. We denote by

$$S_r(X) := \{f \in \mathscr{F}(X) \,|\, \|f\| = r\}$$

the sphere of radius r in $\mathscr{F}(X)$ and for $f \in \mathscr{F}(X)$, $f^o := \frac{f}{\|f\|}$. The proof of the theorem is based on geometry of spheres in high-dimensional Euclidean spaces. In large dimensions, most of areas of spheres lie very close to their "equators" [1].

Theorem 3. *Let d be a positive integer, $X \subset \mathbb{R}^d$ with $\mathrm{card} X = m$, $G(X)$ a subset of $\mathscr{F}(X)$ with card $G(X) = n$ such that for all $g \in G(X)$, $\|f\| \leq r$, μ be a uniform probability measure on $S_r(X)$, and $b > 0$. Then*

$$\mu(\{f \in S_r(X) \,|\, \|f\|_{G(X)} \geq b\}) \geq 1 - 2n\, e^{-\frac{m}{2b^2}}.$$

Proof. Denote for $g \in S_r(X)$ and $\varepsilon \in (0,1)$,

$$C(g,\varepsilon) := \{h \in S_r^{m-1} \,|\, |\langle h^o, g^o \rangle| \geq \varepsilon\}.$$

As $C(g,\varepsilon)$ is equivalent to a polar cap in $\mathbb{R}^{\mathrm{card} X}$, whose measure is exponentially decreasing with the dimension m, we have

$$\mu(C(g,\varepsilon)) \leq e^{-\frac{m\varepsilon^2}{2}}$$

(see, e.g., [1]). By Theorem 2,

$$\{f \in S_r(X) \,|\, \|f\|_{G(X)} \geq b\} = S_r(X) - \bigcup_{g \in G} C(g, 1/b).$$

Hence the statement follows. □

Theorem 3 can be applied to dictionaries $G(X)$ on domains $X \subset \mathbb{R}^d$ with $\mathrm{card} X = m$, which are "relatively small". In particular, dictionaries of signum and Heaviside perceptrons are relatively "small". Estimates of their sizes can be obtained from bounds on numbers of linearly separable dichotomies to which finite subsets of $\mathbb{R}^d$ can be partitioned. Various estimates of numbers of dichotomies have been derived by several authors starting from results by Schläfli [17]. The next bound is obtained by combining a theorem from [7, p.330] with an upper bound on partial sum of binomials.

Theorem 4. *For every d and every $X \subset \mathbb{R}^d$ such that* $\mathrm{card}\, X = m$,

$$\mathrm{card}\, P_d(X) \leq 2\sum_{i=0}^{d}\binom{m-1}{i} \leq 2\frac{m^d}{d!}.$$

Combining Theorems 3 and 4, we obtain a lower bound on measures of sets of functions having variations with respect to signum perceptrons bounded from below by a given bound b.

Corollary 1. *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* $\mathrm{card}\, X = m$, *μ a uniform probability measure on $S_{\sqrt{m}}(X)$, and $b > 0$. Then*

$$\mu(\{f \in S_{\sqrt{m}}(X) \,|\, \|f\|_{P_d(X)} \geq b\}) \geq 1 - 4\frac{m^d}{d!}e^{-\frac{m}{2b^2}}.$$

For example, for the domain $X = \{0,1\}^d$ and $b = 2^{\frac{d}{4}}$, we obtain from Corollary 1 a lower bound

$$1 - \frac{2^{d^2+2}e^{-(2^{\frac{d}{2}}-1)}}{d!}$$

on the probability that a function on $\{0,1\}^d$ with the norm $2^{d/2}$ has variation with respect to signum perceptrons greater or equal to $2^{\frac{d}{4}}$. Thus for large d almost any uniformly randomly chosen function on the d-dimensional Boolean cube $\{0,1\}^d$ of the same norm $2^{d/2}$ as signum perceptrons, has variation with respect to signum perceptrons depending on d exponentially.

4 Construction of Functions with Large Variations

The results derived in the previous section are existential. In this section, we construct a class of functions, which cannot be represented by shallow perceptron networks of low model complexities. We construct such functions using Hadamard matrices. We show that the class of Hadamard matrices contains circulant matrices with rows being segments of pseudo-noise sequences which mimic some properties of random sequences.

Recall that a *Hadamard matrix* of order m is an $m \times m$ square matrix M with entries in $\{-1,1\}$ such that any two distinct rows (or equivalently columns) of M are orthogonal. Note that this property is invariant under permutating rows or columns and under sign flipping all entries in a column or a row. Two distinct rows of a Hadamard matrix differ in exactly $m/2$ positions.

The next theorem gives a lower bound on variation with respect to signum perceptrons of a $\{-1,1\}$-valued function constructed using a Hadamard matrix.

Theorem 5. *Let M be an $m \times m$ Hadamard matrix, $\{x_i \,|\, i = 1,\ldots,m\} \subset \mathbb{R}^d$, $\{y_j \,|\, j = 1,\ldots,m\} \subset \mathbb{R}^d$, $X = \{x_i \,|\, i = 1,\ldots,m\} \times \{y_j \,|\, j = 1,\ldots,m\} \subset \mathbb{R}^{2d}$, and $f_M : X \to \{-1,1\}$ be defined as $f_M(x_i,y_j) =: M_{i,j}$. Then*

$$\|f_M\|_{P_d(X)} \geq \frac{\sqrt{m}}{log_2 m}.$$

Proof. By Theorem 2,

$$\|f_M\|_{P_d(X)} \geq \frac{\|f_M\|^2}{\sup_{g \in P_d(X)} \langle f_M, g\rangle} = \frac{m^2}{\sup_{g \in P_d(X)} \langle f_M, g\rangle}.$$

For each $g \in P_d(X)$, let $M(g)$ be an $m \times m$ matrix defined as $M(g)_{i,j} = g(x_i,y_j)$. It is easy to see that

$$\langle f_M, g\rangle = \sum_{i,j} M_{i,j}M(g)_{i,j}.$$

Using suitable permutations, we reorder rows and columns of both matrices $M(g)$ and M in such a way that each row and each column of the reordered matrix $\bar{M}(g)$ starts with a (possibly empty) initial segment of -1's followed by a (possibly empty) segment of 1's. Denoting $\bar{M}$ the reordered matrix M we have

$$\langle f_M, g\rangle = \sum_{i,j} M_{i,j}M(g)_{i,j} = \sum_{i,j} \bar{M}_{i,j}\bar{M}(g)_{i,j}.$$

As the property of being a Hadamard matrix is invariant under permutations of rows and columns, we can apply Lindsay lemma [8, p.88] to submatrices of the Hadamard matrix $\bar{M}$ on which all entries of the matrix $\bar{M}(g)$ are either -1 or 1. Thus we obtain an upper bound $m\sqrt{m}$ on the differences of $+1$s and -1s in suitable submatrices of $\bar{M}$. Iterating the procedure at most $\log_2 m$-times, we obtain an upper bound $m\sqrt{m}\log_2 m$ on $\sum_{i,j} \bar{M}_{i,j}\bar{M}(g)_{i,j} = \langle f_M, g\rangle$. Thus

$$\|f_M\|_{P_d(X)} \geq \frac{m^2}{m\sqrt{m}\log_2 m} = \frac{\sqrt{m}}{\log_2 m}.$$

□

Theorem 5 shows that functions whose representations by shallow perceptron networks require numbers of units or sizes of output weights bounded from below by $\frac{\sqrt{m}}{\log_2 m}$ can be constructed using Hadamard matrices. In particular, when the domain is d-dimensional Boolean cube $\{0,1\}^d$, where d is even, the lower bound is $\frac{2^{d/4}}{d/2}$. So the lower bounds grows with d exponentially.

Recall that if a Hadamard matrix of order m exists, then $m = 1$ or $m = 2$ or m is divisible by 4 [14, p.44]. It is conjectured that there exists a Hadamard matrix of every order divisible by 4. Listings of Hadamard matrices of various orders can be found at Neil Sloane's library of Hadamard matrices.

We show that suitable Hadamard matrices can be obtained from pseudo-noise sequences. An infinite sequence

$a_0, a_1, \ldots, a_i, \ldots$ of elements of $\{0,1\}$ is called *k-th order linear recurring sequence* if for some $h_0, \ldots, h_k \in \{0,1\}$

$$a_i = \sum_{j=1}^{k} a_{i-j} h_{k-j} \mod 2$$

for all $i \geq k$. It is called *k-th order pseudo-noise (PN) sequence* (or *pseudo-random sequence*) if it is k-th order linear recurring sequence with minimal period $2^k - 1$.

A $2^k \times 2^k$ matrix L is called *pseudo-noise* if for all $i = 1, \ldots, 2^k$, $L_{1,i} = 0$ and $L_{i,1} = 0$ and for all $i = 2, \ldots, 2^k$ and $j = 2, \ldots, 2^k$

$$L_{i,j} = \bar{L}_{i-1,j-1}$$

where the $(2^k - 1) \times (2^k - 1)$ matrix $\bar{L}$ is a circulant matrix with rows formed by shifted segments of length $2^k - 1$ of a k-th order pseudo-noise sequence.

PN sequences have many useful applications because some of their properties mimic those of random sequences. A run is a string of consecutive 1's or a string of consecutive 0's. In any segment of length $2^k - 1$ of k-th order PN-sequence, one-half of the runs have length 1, one quarter have length 2, one-eighth have length 3, and so on. In particular, there is one run of length k of 1's, one run of length $k-1$ of 0's. Thus every segment of length $2^k - 1$ contains $2^{k/2}$ ones and $2^{k/2} - 1$ zeros [14, p.410].

Let $\tau : \{0,1\} \to \{-1,1\}$ be defined as $\tau(x) = -1^x$ (i.e., $\tau(0) = 1$ and $\tau(1) = -1$). The following theorem states that a matrix obtained by applying τ to entries of a pseudo-noise matrix is a Hadamard matrix.

Theorem 6. *Let L be a $2^k \times 2^k$ pseudo-noise matrix and L_τ be the $2^k \times 2^k$ matrix with entries in $\{-1,1\}$ obtained from L by applying τ to all its entries. Then L_τ is a Hadamard matrix.*

Proof. We show that inner product of any two rows of L_τ is equal to zero. The autocorrelation of a sequence $a_0, a_1, \ldots, a_i, \ldots$ of elements of $\{0,1\}$ with period $2^k - 1$ is defined as

$$\rho(t) = \frac{1}{2^k - 1} \sum_{j=0}^{2^k - 1} -1^{a_j + a_{j+t}}.$$

For every pseudo-noise sequence,

$$\rho(t) = -\frac{1}{2^k - 1}$$

for every $t = 1, \ldots, 2^k - 2$ [14, p. 411]. Thus the inner product of every two rows of the matrix $\bar{L}_\tau$ is equal to -1. As all elements of the first column of L_τ are equal to 1, inner product of every pair of its rows is equal to zero. □

Theorem 5 implies that for every pseudo-noise matrix L of order 2^k and $X \subset \mathbb{R}^d$ such that $\mathrm{card} X = 2^k \times 2^k$, there exists a function $f_{L_\tau} : X \to \{-1,1\}$ induced by the matrix L_τ obtained from L by replacing 0's with 1's and 1's with -1's such that

$$\|f_{L_\tau}\|_{P_d(X)} \geq \frac{2^{k/2}}{k}.$$

So the variation of f_{L_τ} with respect to signum perceptrons depends on k exponentially. In particular, setting $X = \{0,1\}^d$, where $d = 2k$ is even, we obtain a function of d variables with variation with respect to signum perceptrons growing with d exponentially as

$$\|f_{L_\tau}\|_{P_d(X)} \geq \frac{2^{d/4}}{d/2}.$$

Representation of this function by a shallow perceptron network requires number of units or sizes of some output weights depending on d exponentially.

It is easy to show that for each even integer d, the function induced by Sylvester-Hadamard matrix

$$M_{u,v} = -1^{u \cdot v},$$

where $u, v \in \{0,1\}^{d/2}$, can be represented by a two-hidden-layer network with $d/2$ units in each hidden layer.

5 Discussion

We proved that almost any uniformly randomly chosen function on a sufficiently large finite set in $\mathbb{R}^d$ has large variation with respect to signum perceptrons and thus it cannot be tractably represented by a shallow perceptron network.

It seems to be a paradox that although representations of almost all functions by shallow perceptron networks are "untractable", it is difficult to construct such functions. The situation can be rephrased in analogy with the title of an article from coding theory "Any code of which we cannot think is good" [6] as "representation of almost any function of which we cannot think by shallow perceptron networks is untractable". A central paradox of coding theory concerns the existence and construction of the best codes. Virtually every linear code is good (in the sense that it meets the Gilbert-Varshamov bound on distance versus redundancy), however despite the sophisticated constructions for codes derived over the years, no one has succeeded in demonstrating a constructive procedure that yields such good codes.

The only class of functions having "large" variations which we succeeded to construct is the class described in section 4 based on Hadamard matrices. Among these matrices belong quasi-noise (quasi-random) matrices with rows obtained as shifts of segments of quasi-noise sequences. These sequences have been used in construction of codes, interplanetary satellite picture transmission, precision measurements, acoustics, radar camouflage, and light diffusers. Pseudo-noise sequences permit design of surfaces that scatter incoming signals very broadly making reflected energy "invisible" or "inaudible".

It should be emphasized that similarly as "no free lunch theorem" [18], our results assume uniform distributions of functions to be represented. However, probability distributions of functions modeling some practical tasks of interest (such as colors of pixels in a photograph) might be highly non uniform.

Acknowledgments. This work was partially supported by the grant COST LD13002 of the Ministry of Education of the Czech Republic and institutional support of the Institute of Computer Science RVO 67985807.

References

[1] Ball, K.: An elementary introduction to modern convex geometry. In: S. Levy, (ed.), Falvors of Geometry, 1–58, Cambridge University Press, 1997

[2] Barron, A. R.: Neural net approximation. In: K. S. Narendra, (ed.), Proc. 7th Yale Workshop on Adaptive and Learning Systems, 69–72, Yale University Press, 1992

[3] Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning **2** (2009), 1–127

[4] Bengio, Y., Delalleau, O., Le Roux, N.: The curse of dimensionality for local kernel machines. Technical Report 1258, Département d'Informatique et Recherche Opérationnelle, Université de Montréal, 2005

[5] Bengio, Y., Delalleau, O., Le Roux, N.: The curse of highly variable functions for local kernel machines. In: Advances in Neural Information Processing Systems 18, 107–114, MIT Press, 2006

[6] Coffey, J. T., Goodman, R. M.: Any code of which we cannot think is good. IEEE Transactions on Information Theory **36** (1990), 1453–1461

[7] Cover, T.: Geometrical and statistical properties of systems of linear inequalities with applictions in pattern recognition. IEEE Trans. on Electronic Computers **14** (1965), 326–334

[8] Erdös, P., Spencer, J. H.: Probabilistic Methods in Combinatorics. Academic Press, 1974

[9] Hinton, G. E., Osindero, S., Teh, Y. W.: A fast learning algorithm for deep belief nets. Neural Computation **18** (2006), 1527–1554

[10] Ito, Y.: Finite mapping by neural networks and truth functions. Mathematical Scientist **17** (1992), 69–77

[11] Kainen, P. C., Kůrková, V., Sanguineti, M.: Dependence of computational models on input dimension: Tractability of approximation and optimization tasks. IEEE Trans. on Information Theory **58** (2012), 1203–1214

[12] Kůrková, V.: Dimension-independent rates of approximation by neural networks. In: K. Warwick and M. Kárný, (eds.), Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, 261–270, Birkhäuser, Boston, MA, 1997

[13] Kůrková, V., Savický, P., Hlaváčková, K.: Representations and rates of approximation of real-valued Boolean functions by neural networks. Neural Networks **11** (1998), 651–659

[14] MacWilliams, F. J., Sloane, N. J. A.: The theory of error-correcting codes. North Holland, New York, 1977

[15] Micchelli, C. A.: Interpolation of scattered data: Distance matrices and conditionally positive definite functions. Constructive Approximation **2** (1986), 11–22

[16] Roychowdhury, V., Siu, K. -Y., Orlitsky, A.: Neural models and spectral methods. In: V. Roychowdhury, K. Siu, and A. Orlitsky, (eds.), Theoretical Advances in Neural Computation and Learning, 3–36, Springer, New York, 1994

[17] Schläfli, L.: Theorie der vielfachen Kontinuität. Zürcher & Furrer, Zürich, 1901

[18] Wolpert, D. H., Macready, W. G.: No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation **1(1)** (1997), 67–82