Representations of Highly-Varying Functions by One-Hidden-Layer Networks

Věra Kůrková

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 18207 Prague, Czech Republic
vera@cs.cas.cz

Abstract. Limitations of capabilities of one-hidden-layer networks are investigated. It is shown that for networks with Heaviside perceptrons as well as for networks with kernel units used in SVM, there exist large sets of d-variable functions which cannot be tractably represented by these networks, i.e., their representations require numbers of units or sizes of weighs depending on d exponentially. Our results are derived using the concept of variational norm from nonlinear approximation theory and the concentration of measure property of high dimensional Euclidean spaces.

Keywords: model complexity of neural networks, one-hidden-layer networks, highly-varying functions, tractability of representations of multivariable functions by neural networks.

1 Introduction

Originally, biologically inspired neural networks were modeled as as multilayer distributed computational systems. Later, one-hidden-layer architectures became dominant in applications due to relatively simple optimization procedures needed for adjustment of their parameters (see, e.g., [1, 2] and the references therein). In some literature, one-hidden-layer networks are called shallow networks to distinguish them from deep ones containing more hidden layers.

In addition to a variety of successful applications of one-hidden-layer networks, also theoretical confirmation of their capabilities has been obtained. Shallow networks with many types of computational units are known to be universal approximators, i.e., they can approximate up to any desired accuracy all continuous functions on compact subsets of $\mathbb{R}^d$. In particular, the universal approximation property holds for shallow networks with perceptrons having any non-polynomial activation function [3, 4] and with radial and kernel units satisfying mild conditions [5–7], [8, p.153]). Moreover, all functions defined on finite subsets of $\mathbb{R}^d$ can be represented exactly by one-hidden-layer networks with either sigmoidal perceptrons [9] or with Gaussian radial units [10].

Proofs of the universal approximation capability of shallow networks require potentially unlimited numbers of hidden units. These numbers representing model complexities are critical factors for practical implementations. Dependence of

L. Rutkowski et al. (Eds.): ICAISC 2014, Part I, LNAI 8467, pp. 67–76, 2014.

model complexities of shallow networks on their input dimensions, types of units, functions to be approximated, and accuracies of approximation have been studied using tools from nonlinear approximation theory (see, e.g., [11] and references therein). Inspection of upper bounds on rates of approximation by shallow networks led to descriptions of various families of functions that can be well approximated by shallow networks with reasonably small numbers of computational units of various types. On the other hand, cases when numbers of networks units are untractably large are less understood. Only few lower bounds on rates of approximations by shallow networks are known and the estimates are mostly non constructive and hold for types of computational units that are not commonly used [12, 13]. Moreover, in some cases, sizes of weights can be more critical factors for successful learning than numbers of network units [14].

Recently, new hybrid learning algorithms were developed for deep networks [15, 16]. Training networks with more than one hidden layer involves complicated nonlinear optimization procedures and thus generally it is more difficult than training shallow ones. Hence, it is desirable to develop some theoretical background for characterization of tasks whose computations by networks with shallow architectures would require networks with considerably higher complexities than computations by deep networks. Bengio et al. [17] suggested that a cause of difficulties in representing functions by shallow networks tractably can be their "amount of variations". As a class of function with high-variations they considered the parities on d-dimensional Boolean cubes $\{0,1\}^d$. They proved that a classification of points in $\{0,1\}^d$ according to their parities by support vector machine (SVM) with Gaussian kernel units cannot accomplish this task with less than $2^{d/2}$ units.

On the other hand, it is well-known and easy to verify that for any d, the d-dimensional parity can be represented by a one-hidden-layer Heaviside perceptron network with d units. Indeed, parity can be visualized as a plane wave orthogonal to the diagonal of the cube in the direction of the vector $(1,\ldots,1)$ (see, e.g.,[18, 19]). So some functions are highly-varying with respect to one type of computational units, while they are "varying" much less with respect to another type of units. Thus it is reasonable to consider the notion of a highly-varying function with respect to a type of computational units.

In this paper, we propose to formalize this concept in terms of a norm called variation with respect to a set of functions. This norm has been studied in nonlinear approximation theory and plays an important role in estimates of rates of approximation by neural networks (see, e.g., [11] and the references therein). We show that the size of the variational norm of a function with respect to a dictionary of computational units reflects both the number of hidden units and sizes of output weights in a shallow network with units from the dictionary representing such function. Using the concept of variational norm, we describe classes of d-variable functions whose representations by networks with a given type of units with increasing numbers of inputs d are not tractable in the sense that representations of such functions by these network require numbers of units or some of sizes of output weights to grow exponentially with d. Using concentration of

measure property in high-dimensional Euclidean spaces we estimate probability distributions of sizes of variations. We show that for popular dictionaries (such as dictionaries formed by SVM and by Heaviside perceptrons) with increasing dimension d almost any randomly chosen Boolean function has large variational norm (depending on d exponentially). Our results imply that for large d, in sets of functions with constant Euclidean norms most Boolean real valued functions cannot be tractably represented by Heaviside perceptron networks or by SVMs. We illustrate general existential results by an example of a concrete class of non tractable functions. Some preliminary results from this paper appeared as work in progress in local conference proceedings [20].

The paper is organized as follows. Section 2 contains basic concepts on shallow networks, dictionaries of computational units and Boolean functions. Section 3 presents a mathematical formalization of the concept of a "highly-varying function", shows that it is related to large sizes of networks representing such functions or large output weights of these networks. In Section 4 estimates of probabilistic measures of sets of functions with variations depending on d exponentially are derived and illustrated by an example of a class of functions which cannot be tractably represented by one-hidden-layer Heaviside perceptron networks. Section 5 is a brief disussion.

2 Preliminaries

One-hidden-layer networks with single linear outputs, 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 output weights, and n is the number of hidden units. This number can be interpreted as a measure of *model complexity*. In this paper we use the term *shallow network* meaning one-hidden-layer network with a single linear output. By

$$\operatorname{span} G := \bigcup_{n \in \mathbb{N}} \operatorname{span}_n G$$

is denoted the set of functions computable by one-hidden-layer networks with units from the dictionary G with any number of hidden units.

We investigate growth of complexities of networks representing functions of increasing numbers of variables d. Let D be an infinite subset of the set of positive integers, $\mathcal{F} = \{f_d \,|\, d \in D\}$ a class of functions and $\{G_d \,|\, d \in D\}$ a class of dictionaries, such that for every $d \in D$, f_d is a function of d variables and G_d is formed by functions of d variables. We call the problem of representing the set $\mathcal{F}$ by networks from $\{\operatorname{span} G_d \,|\, d \in D\}$ *tractable* if for every $d \in D$, there exists a network in $\operatorname{span}_{n_d} G$ representing f_d as its input-output function

such that n_d and absolute values of all output weights in the network grow with d polynomially. Note that different concepts of tractability were used in other contexts (see, e.g., [11]).

In this paper, we focus on representations of real-valued functions on finite subsets of $\mathbb{R}^d$ by shallow networks with units from several dictionaries. We denote by $H_d(X)$ the dictionary of functions on $X \subset \mathbb{R}^d$ computable by *Heaviside perceptrons*, i.e.,

$$H_d(X) := \{\vartheta(v \cdot . + b) : X \to \{0,1\} \,|\, v \in \mathbb{R}^d, b \in \mathbb{R}\}\,,$$

where ϑ denotes the *Heaviside activation function* defined as

$$\vartheta(t) := 0 \text{ for } t < 0 \text{ and } \vartheta(t) := 1 \text{ for } t \geq 0.$$

Note that H_d is the *set of characteristic functions of half-spaces*. The dictionary $S_d(X)$ is formed by functions on X computable by perceptrons with *signum activation function* $\mathrm{sgn} : \mathbb{R} \to \{-1, 1\}$ defined as

$$\mathrm{sgn}(t) := -1 \text{ for } t < 0 \text{ and } \mathrm{sign}(t) := 1 \text{ for } t \geq 0.$$

We denote

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

For a kernel $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we denote by $F_{K_d}(X)$ the dictionary of kernel units, i.e.,

$$F_{K_d}(X) := \{K_d(., x) : X \to \mathbb{R} \,|\, x \in X\}.$$

The set of real-valued functions on the *d-dimensional Boolean cube* $\{0,1\}^d$ is denoted

$$\mathcal{B}(\{0,1\}^d) := \{f \,|\, f : \{0,1\}^d \to \mathbb{R}\}.$$

It is a linear space isomorphic to the Euclidean space $\mathbb{R}^{2^d}$. Thus on $\mathcal{B}(\{0,1\}^d)$ we have the *Euclidean inner product* defined as

$$\langle f, g\rangle := \sum_{u \in \{0,1\}^d} f(u)g(u)$$

and the *Euclidean norm* $\|f\|_2 := \sqrt{\langle f, f\rangle}$. By $\cdot$ is denoted the inner product on $\{0,1\}^d$, defined as $u \cdot v := \sum_{i=1}^d u_i v_i$.

3 Highly-Varying Functions

In this section, we investigate a mathematical formalization of the observation of Bengio et al. [17] that representations of highly-varying functions might require large networks. We show that the concept of a variational norm from approximation theory can play a role of a measure of tractability of representations of classes of functions by shallow networks.

For a subset G of a normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$, *$G$-variation (variation with respect to the set G)*, denoted by $\|.\|_G$, is defined as

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

where $-G := \{-g \mid g \in G\}$, $\mathrm{cl}_{\mathcal{X}}$ denotes the closure with respect to the norm $\|\cdot\|_{\mathcal{X}}$ on $\mathcal{X}$, and $\operatorname{conv} G := \left\{\sum_{i=1}^k a_i g_i \mid a_i \in [0,1], \sum_{i=1}^k a_i = 1, g_i \in G, k \in \mathbb{N}\right\}$ is the convex hull of G.

Variation with respect to a set of functions was introduced by Kůrková [21] as an extension of Barron's [22] concept of variation with respect to sets of characteristic functions. Barron investigated the set of characteristic functions of half-spaces, which corresponds to the dictionary of functions computable by Heaviside perceptrons. For $d = 1$, variation with respect to half-spaces coincides up to a constant with the concept of total variation from integration theory. Variational norms play an important role in estimates of approximation rates by one-hidden-layer networks (see, e.g., [11, 23, 24] and the references therein).

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

Proposition 1. *Let G be a bounded subset of a normed linear space $(\mathcal{X}, \|.\|)$, then for every $f \in \mathcal{X}$,*
(i) $\|f\|_G \leq \left\{\sum_{i=1}^k |w_i| \,\middle|\, f = \sum_{i=1}^k w_i\, g_i\,,\ w_i \in \mathbb{R},\ g_i \in G, k \in \mathbb{N}\right\}$;
(ii) for G finite with $\operatorname{card} G = k$,
$\|f\|_G = \min \left\{\sum_{i=1}^k |w_i| \,\middle|\, f = \sum_{i=1}^k w_i\, g_i\,,\ w_i \in \mathbb{R},\ g_i \in G\right\}$.

Proposition 1 implies that families of sets of d-variable functions $\{F_d \,|d \in D\}$ with G_d-variations growing with d exponentially cannot be tractably represented by networks with units from G_d.

Note that G-variation is a norm and thus by multiplying f by suitable constants we can obtain functions with arbitrarily large or small variations. However, in neurocomputing we are interested in computation of functions with similar sizes as computational units. For example, in dictionaries $H_d(X)$ and $P_d(X)$ $F_{K_d}(X)$ formed by functions on a finite subset X of $\mathbb{R}^d$, the supremum of l_2-norms of their elements is $2^{\operatorname{card} X/2}$. Thus we explore variational norms of functions in the spheres of radii $2^{\mathrm{card}(X)/2}$ in the Euclidean spaces $\mathcal{B}(X)$.

To describe classes of functions with large variations, we use the following lower bound on variational norm from [19] (see also [25, 26]). By $G^\perp$ is denoted the *orthogonal complement of G*.

Theorem 1. *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Hilbert space and G its bounded subset. Then for every $f \in \mathcal{X} \setminus G^\perp$, $\|f\|_G \geq \frac{\|f\|^2}{\sup_{g \in G} |g \cdot f|}$.*

Theorem 1 implies that functions which are "almost orthogonal" to G have large variations. To take advantage of this theorem, we use the *angular pseudometrics* δ on the unit sphere S^{m-1} in $\mathbb{R}^m$ defined as

$$\delta(f,g) = \arccos |f \cdot g|.$$

Note that this pseudometrics defines the distance as the minimum of the two angles between f and g and between f and $-g$ (it is a pseudometrics as the distance of antipodal vectors is zero).

The next corollary of Theorem 1 states that functions which have large distances measured by an angular pseudometrics δ from the set G have large G-variations.

Corollary 1. *Let m be a positive integer, $G \subset S^{m-1}$, and $f \in S^{m-1}$ such that has for some $\alpha \in (0, \pi/2)$ and all $g \in G$, the angular distance $\delta(f,g) \geq \alpha$. Then $\|f\|_G \geq \frac{1}{\cos \alpha}$.*

4 Sets of Functions with Large Variations

In this section we show that for reasonably "small" dictionaries G formed by functions on finite subsets X of $\mathbb{R}^d$ with $\operatorname{card} X = m$ there exist "large subsets" of spheres in $\mathbb{R}^m$ consisting of functions with "large" G-variations. The following theorem estimates probability that a randomly chosen vector $f \in S^{m-1}$ has G-variation larger than $\frac{1}{\cos \alpha}$. Its proof is based on a geometrical property of high-dimensional Euclidean spaces called "concentration of measure". This property implies that for large dimensions m, most of the areas of spheres S^{m-1} in m-dimensional spaces $\mathbb{R}^m$ lie "close" to the equators of these spheres (see, e.g., [27]).

Theorem 2. *Let m be a positive integer, μ a uniform measure on S^{m-1} such that $\mu(S^{m-1}) = 1$, G a finite subset of S^{m-1} with $\operatorname{card} G = k$, $\alpha \in (0, \pi/2)$, and $V_\alpha = \{f \in S^m \mid \|f\|_G \geq \frac{1}{\cos \alpha}\}$. Then $\mu(V_\alpha) \geq 1 - k\, e^{-\frac{m(\cos \alpha)^2}{2}}$.*

Proof. By Corollary 1, V_α contains all $f \in S^{m-1}$ satisfying for all $g \in G$, $|f \cdot g| \leq \cos \alpha$, i.e., all f with $\delta(f,g) = \arccos |f \cdot g| \geq \alpha$. Let $C(g,\varepsilon)$ denotes the spherical cap with a center $g \in G$ and the angle $\alpha = \arccos \varepsilon$ defined as $C(g,\varepsilon) = \{h \in S^{m-1} \mid h \cdot g \geq \varepsilon\}$. So f is not contained in any of the spherical caps $C(g,\varepsilon)$ with a center $g \in G$. With d increasing, the normalized measures of the spherical caps are decreasing exponentially fast: $\mu(C(g,\varepsilon)) \leq e^{-\frac{m\varepsilon^2}{2}}$ (see, e.g., [28, p.11]). Thus $\mu(V_\alpha) \geq 1 - k\, e^{-\frac{m(\cos \alpha)^2}{2}}$.

Combining Theorem 2 with "relatively small"sizes of the dictionaries $H_d(\{0,1\}^d)$, $P_d(\{0,1\}^d)$, and $F_{K_d}(\{0,1\}^d)$, induced by a bounded kernel $K_d : \{0,1\}^d \times \{0,1\}^d \to \mathbb{R}$ (such as the Gaussian), we obtain an estimate of the fraction of the area of the sphere of radius $2^{d/2}$ in the space $\mathcal{B}(\{0,1\}^d) \simeq \mathbb{R}^{2^d}$ which contains functions with variations depending on d exponentially. By S_r^{m-1} we denote the sphere of radius r in $\mathbb{R}^m$.

Theorem 3. *Let d be a positive integer, μ a uniform measure on $S^{2^d-1}_{2^{d/2}}$ such that $\mu(S^{2^d-1}_{2^{d/2}}) = 1$, G a dictionary formed by functions on $\{0,1\}^d$ such that for all $g \in G$, $\|g\|_2 \leq 2^{d/2}$, $\alpha \in (0, \pi/2)$, and $V_\alpha(G) = \{f \in S^{2^d-1} \,|\, \|f\|_G \geq \frac{1}{\cos\alpha}\}$.*
(i) If $G = H_d(\{0,1\}^d)$, then $\mu(V_\alpha(H_d(\{0,1\}^d))) \geq 1 - 2^{d^2}\, e^{-\frac{2^d(\cos\alpha)^2}{2}}$;
(ii) if $G = P_d(\{0,1\}^d)$, then $\mu(V_\alpha(H_d(\{0,1\}^d))) \geq 1 - 2^{d^2}\, e^{-\frac{2^d(\cos\alpha)^2}{2}}$;
(iii) if $G_{K_d}(\{0,1\}^d)$, where $K : \{0,1\}^d \times \{0,1\}^d$ is a kernel such that $\sup_{x\in\{0,1\}^d} |K(x,x) \leq 1$, then $\mu(V_\alpha(G_{K_d}(\{0,1\}^d))) \geq 1 - 2^d\, e^{-\frac{2^d(\cos\alpha)^2}{2}}$.

Proof. (i) and (ii) follow from Theorem 2 and an upper bound $2^{d^2 - d\log_2 d + \mathcal{O}(d)}$ on the dictionary card $H_d(\{0,1\}^d)$ [29, 30]. Thus cardinalities of both dictionaries $H_d(\{0,1\}^d)$ and $P_d(\{0,1\}^d)$ are smaller than 2^{d^2}, which is much smaller than the cardinality 2^{2^d} of the whole space $\mathcal{B}(\{0,1\}^d)$. The Euclidean norm of all elements of $P_d(\{0,1\}^d)$ is $2^{d/2}$, which is the maximal value of the Euclidean norms of elements of $H_d(\{0,1\}^d)$.
(iii) follows from Theorem 2 and the cardinality 2^d of the dictionary $G_{K_d}(\{0,1\}^d)$ formed by kernel units centered at the vertices of the Boolean cube $\{0,1\}^d$.

Theorem 3 holds for any kernel with $\sup_{x\in\{0,1\}^d} |K(x,x)| = 1$ and implies that representations of most functions from $\mathcal{B}(\{0,1\}^d)$ having their Euclidean norms equal to $2^{d/2}$ by SVM induced by the kernel K are not tractable, i.e., their representations require exponentially large numbers of units or exponentially large sizes of output weights.

Setting $\cos\alpha = 2^{-d/4}$, we obtain from Theorem 3 the lower bound

$$1 - e^{-\frac{2^{d/2} - 2d^2}{2}}$$

on the relative size of the subset of the ball of radius $2^{d/2}$ in $\mathcal{B}(\{0,1\}^d)$ containing functions with variations with respect to half-spaces larger or equal to $2^{d/4}$. So by Proposition 1, for large d almost any randomly chosen real-valued Boolean function with the norm $2^{d/2}$ cannot be tractably represented by a shallow Heaviside perceptron network.

Theorem 3 showing that for large d, almost any function on the sphere of radius $2^{\mathrm{card}X}$ has variation depending on d exponentially is existential. However, to construct concrete examples of such functions is not easy. The only example of which we are aware is the function "inner product mod 2" which serves in theory of circuit complexity as ana example of a function which does not belong to the class $\widehat{LT_2}$ of depth-2 polynomial-size threshold gate circuits with weights being polynomially bounded integers (see, e.g., [18]). For every even positive integer d, let $\beta_d : \{0,1\}^d \to \{-1,1\}$ be defined for all $x \in \{0,1\}^d$ as

$$\beta_d := (-1)^{l(x)\cdot r(x)}$$

where $l(x), r(x) \in \{0,1\}^{d/2}$ are defined for every $i = 1, \ldots \frac{d}{2}$ as $l(x)_i := x_i$ and $r(x)_i := x_{\frac{d}{2}+i}$. When the range $\{-1,1\}$ is replaced with $\{1,0\}$, functions computing inner products of $l(x)$ with $r(x)$ mod 2 are obtained.

The following theorem is a corollary of a lower bound on the variational norm from [19, Theorem 3.7]. Recall the $h = \Omega(g(d))$ for two functions $g, h : \mathbb{N} \to \mathbb{R}$ meaning that there exist a positive constant c and $n_0 \in \mathbb{N}$ such that for all $n \geq n_o$ one has $h(n) \geq c\, g(n)$ [31].

Theorem 4. *Let d be an even integer, then $\|\beta_d\|_{H_d(\{0,1\}^d)} \geq \|f\|_{P_d(\{0,1\}^d)} = \Omega(2^{d/6})$.*

By Theorem 4 and Proposition 1 we get the following corollary.

Corollary 2. *Let d be an even integer and $\beta_d(x) = \sum_{i=1}^m w_i \vartheta(v_i \cdot x + b_i)$ be a representations of the function $\beta_d : \{0,1\}^d \to \{-1,1\}$ by a one-hidden-layer Heaviside perceptron network. Then $\sum_{i=1}^m |w_i| = \Omega(2^{2d/6})$.*

Corollary 2 implies that a representation of a class of d-variable Boolean functions $\{\beta_d \,|\, d \text{ even}\}$ by one-hidden-layer Heaviside perceptron networks is not tractable. These functions cannot be represented by Heaviside perceptron networks with both numbers of units and sums of absolute values of output weights polynomially bounded.

5 Discussion

We investigated model complexities of one-hidden-layer networks representing high-dimensional functions. We showed that the concept of variational norm with respect to a dictionary studied on approximation theory reflects both numbers of units and sizes of output weights in representing networks with units from the dictionary. Using properties of high-dimensional spaces, we proved that for networks with common units (such as perceptrons and SVM kernel units) with increasing input dimension d most of the functions require networks with number of units or sizes of output weights depending on d exponentially. An essential condition in our arguments is a relatively small size of these dictionaries. The upper bound $2^{d^2 - d \log_2 d + \mathcal{O}(d)}$ on the dictionary of Heaviside perceptrons on the Boolean cube was derived already in 19th century by one of the founders of high-dimensional geometry [29].

Our results hold for functions of comparable norms as network units. Note that also in theory of circuit complexity (see, e.g., [18]), there are studied representations of functions of fixed Euclidean norms by networks with gates computing functions with the same norms by networks with constrains on both numbers of units and their output weights. In particular, in this theory there are studied representations of Boolean functions with values in $\{-1, 1\}$ by networks composed from signum perceptrons. All these functions have Euclidean norms equal to $2^{d/2}$.

Acknowledgments. This work was partially supported by 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. Fine, T.L.: Feedforward Neural Network Methodology. Springer, Heidelberg (1999)
2. Chow, T.W.S., Cho, S.Y.: Neural Networks and Computing: Learning Algorithms and Applications. World Scientific (2007)
3. Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6, 861–867 (1993)
4. Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numerica 8, 143–195 (1999)
5. Park, J., Sandberg, I.: Approximation and radial-basis-function networks. Neural Computation 5, 305–316 (1993)
6. Mhaskar, H.N.: Versatile Gaussian networks. In: Proc. of IEEE Workshop of Nonlinear Image Processing, pp. 70–73 (1995)
7. Kůrková, V.: Some comparisons of networks with radial and kernel units. In: Villa, A.E.P., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds.) ICANN 2012, Part II. LNCS, vol. 7553, pp. 17–24. Springer, Heidelberg (2012)
8. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
9. Ito, Y.: Finite mapping by neural networks and truth functions. Mathematical Scientist 17, 69–77 (1992)
10. Micchelli, C.A.: Interpolation of scattered data: Distance matrices and conditionally positive definite functions. Constructive Approximation 2, 11–22 (1986)
11. Kainen, P.C., Kůrková, V., Sanguineti, M.: Dependence of computational models on input dimension: Tractability of approximation and optimization tasks. IEEE Transactions on Information Theory 58, 1203–1214 (2012)
12. Maiorov, V.: On best approximation by ridge functions. J. of Approximation Theory 99, 68–94 (1999)
13. Maiorov, V., Pinkus, A.: Lower bounds for approximation by MLP neural networks. Neurocomputing 25, 81–91 (1999)
14. Bartlett, P.L.: The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. IEEE Trans. on Information Theory 44, 525–536 (1998)
15. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18, 1527–1554 (2006)
16. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning 2, 1–127 (2009)
17. Bengio, Y., Delalleau, O., Roux, N.L.: The curse of highly variable functions for local kernel machines. In: Advances in Neural Information Processing Systems 18, pp. 107–114. MIT Press (2006)
18. Roychowdhury, V., Siu, K., Orlitsky, A.: Neural models and spectral methods. In: Roychowdhury, V., Siu, K., Orlitsky, A. (eds.) Theorertical Advances in Neural Computation and Learning, pp. 3–36. Kluwer Academic Press (1997)
19. Kůrková, V., Savický, P., Hlaváčková, K.: Representations and rates of approximation of real-valued Boolean functions by neural networks. Neural Networks 11, 651–659 (1998)
20. Kůrková, V.: Representations of highly-varying functions by perceptron networks. In: Informačné Technológie - Aplikácie a Teória - ITAT 2013, Košice, UPJŠ (2013)

21. Kůrková, V.: Dimension-independent rates of approximation by neural networks. In: Warwick, K., Kárný, M. (eds.) Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, pp. 261–270. Birkhäuser, Boston (1997)
22. Barron, A.R.: Neural net approximation. In: Narendra, K. (ed.) Proc. 7th Yale Workshop on Adaptive and Learning Systems, pp. 69–72. Yale University Press (1992)
23. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. on Information Theory 39, 930–945 (1993)
24. Kůrková, V., Sanguineti, M.: Comparison of worst-case errors in linear and neural network approximation. IEEE Transactions on Information Theory 48, 264–275 (2002)
25. Kůrková, V.: Minimization of error functionals over perceptron networks. Neural Computation 20, 250–270 (2008)
26. Kůrková, V.: Complexity estimates based on integral transforms induced by computational units. Neural Networks 33, 160–167 (2012)
27. Matoušek, J.: Lectures on Discrete Geometry. Springer, New York (2002)
28. Ball, K.: An elementary introduction to modern convex geometry. In: Levy, S. (ed.) Falvors of Geometry, pp. 1–58. Cambridge University Press (1997)
29. Schläfli, L.: Theorie der vielfachen Kontinuität. Zürcher & Furrer, Zürich (1901)
30. Schläfli, L.: Gesamelte Mathematische Abhandlungen, Band 1. Birkhäuser, Basel (1950)
31. Knuth, D.E.: Big omicron and big omega and big theta. SIGACT News 8, 18–24 (1976)