J. Hlaváčová (Ed.): ITAT 2017 Proceedings, pp. 100–105
CEUR Workshop Proceedings Vol. 1885, ISSN 1613-0073, © 2017 V. Kůrková

Bounds on Sparsity 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 (shallow) perceptron networks to sparsely represent multivariable functions is investigated. A concrete class of functions is described whose computation by shallow perceptron networks requires either large number of units or is unstable due to large output weights. The class is constructed using pseudo-noise sequences which have many features of random sequences but can be generated using special polynomials. Connections with the central paradox of coding theory are discussed.

1 Introduction

To identify and explain efficient network designs, it is necessary to develop a theoretical understanding to the influence of a proper choice of network architecture and type of units on reducing network complexity. Bengio and LeCun [5], who recently revived the interest in deep networks, conjectured that "most functions that can be represented compactly by deep architectures cannot be represented by a compact shallow architecture". On the other hand, a recent empirical study demonstrated that shallow networks can learn some functions previously learned by deep ones using the same numbers of parameters as the original deep networks [1].

It is well-known that shallow networks with merely one hidden layer of computational units of many common types can approximate within any accuracy any reasonable function on a compact domain and also can exactly compute any function on a finite domain [9, 22]. All these universality type results are proven assuming that numbers of network units are potentially infinite or, in the case of finite domains, are at least as large as sizes of the domains. However, in practical applications, various constraints on numbers and sizes of network parameters limit feasibility of implementations.

Whereas many upper bounds on numbers of units in shallow networks which are sufficient for a given accuracy of function approximation are known (see, e.g, the survey article [10] and references therein), fewer lower bounds are available. Some bounds hold merely for types of computational units that are not commonly used (see, e.g., [20, 19]). Proofs of lower bounds are generally much more difficult than derivation of upper ones.

Characterization of tasks which can be computed by considerably sparser deep networks than shallow ones requires proving lower bounds on complexity of shallow networks which are larger than upper bounds on complexity of deep ones. An important step towards this goal is exploration of functions which cannot be computed or approximated by shallow networks satisfying various sparsity constraints.

Investigation of sparsity of artificial neural networks has a biological motivation. A number of studies confirmed that only a small fraction of neurons have a high rate of firing at any time (sparse activity) and that each neuron is connected to only a limited number of other neurons (sparse connectivity) [17]. The most simple measure of sparse connectivity between hidden units and network outputs is the number of non zero output weights. This number is in some literature called "l_0-pseudo-norm". However, it is neither a norm nor a pseudo-norm. Its minimization is a not convex problem and its solution is NP-hard [23]. Thus instead of "l_0-pseudo-norm", l_1 and l_2-norms have been used as measures of network sparsity as they can be implemented as stabilizers in weight-decay regularization techniques (see, e.g., [8] and references therein). Also in online dictionary learning, l_1-norms were used as stabilizers [21, 7].

Bengio et al. [4] suggested that a cause of large model complexities of shallow networks might be in the "amount of variations" of functions to be computed. In [14], we presented some examples showing that sparsity of shallow networks computing the same input-output functions strongly depends on types of their units. We proposed to use as a measure of sparsity variational norms tailored to dictionaries of computational units. These norms were used as tools in nonlinear approximation theory. We showed that variational norms can be employed to obtain lower bounds on sparsity measured by l_1-norms. For many dictionaries of computational units, we derived lower bounds on these norms using a probabilistic argument based on the Chernoff Bound [14, 15]. The bounds hold for almost all functions representing binary classifiers on sufficiently large finite domains. In [13] we complemented these probabilistic results by a concrete construction of binary classifiers with large variational norms with respect to signum perceptrons.

In this paper, we investigate sparsity of shallow networks computing real-valued functions on finite rectangular domains. Such domains can be 2-dimensional (e.g., pixels of photographs) or high-dimensional (e.g., digitized high-dimensional cubes), but typically they are quite large. We describe a construction of a class of functions on such

domains based on matrices with orthogonal rows. Estimating variational norms of these functions from bellow, we obtain lower bounds on l_1-norms of shallow networks with signum perceptrons. We show that these networks must have either large numbers of hidden units or some of their output weights must be large. Both are not desirable as large output weights ma lead to non stability of computation. We illustrate our general construction by a concrete class of circulant matrices generated by pseudo-noise sequences. We discuss the effect of pseudo-randomness on network complexity.

The paper is organized as follows. Section 2 contains basic concepts on shallow networks and dictionaries of computational units. In Section 3, sparsity is investigated in terms of l_1-norm and norms tailored to computational units. In Section 4, a concrete construction of classes of functions with large variational norms based on orthogonal matrices is described. In Section 5, the general results are illustrated by a concrete example of matrices obtained from pseudo-noise sequences. Section 6 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 the simplest 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

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

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

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

where $\operatorname{sgn}(t) := -1$ for $t < 0$ and $\operatorname{sign}(t) := 1$ 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

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

and

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

where $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. An advantage of signum perceptrons is that all units from the dictionary $P_d(X)$ have the same size of norms equal to $\sqrt{\operatorname{card} X}$.

3 Measures of Sparsity

The most simple measure of sparse connectivity between the hidden layer and the network output is the number of non-zero output weights. In some literature, the number of non-zero coefficients among w_i's in an input-output function

$$f = \sum_{i=1}^{n} w_i g_i \tag{1}$$

from $\operatorname{span} G$ is called an *"l_0-pseudo-norm"* in quotation marks and denoted $\|w\|_0$. However, it is neither a norm nor a pseudo-norm. The quantity $\|w\|_0$ is always an integer and thus $\|\cdot\|_0$ does not satisfy the homogeneity property of a norm ($\|\lambda x\| = |\lambda| \|x\|$ for all λ). Moreover, the "unit ball" $\{w \in \mathbb{R}^n \,|\, \|w\|_0 \leq 1\}$ is non convex and unbounded.

Minimization of the "l_0-pseudo-norm" of the vector of output weights is a difficult non convex optimization task. Instead of "l_0", *l_1-norm* defined as

$$\|w\|_1 = \sum_{i=1}^{n} |w_i|$$

and *l_2-norm* defined as

$$\|w\|_2 = \sqrt{\sum_{i=1}^{n} w_i^2}$$

of output weight vectors $w = (w_1, \ldots, w_n)$ have been used in weight-decay regularization [8]. These norms can be implemented as stabilizers modifying error functionals which are minimized during learning. A network with a large l_1 or l_2-norm of its output-weight vector must have either a large number of units or some output weights must be large. Both of these properties are not desirable as they imply a large model complexity or non stability of computation caused by large output weights.

Many dictionaries of computational units are overcomplete and thus the representation (1) as a linear combination of units from the dictionary need not be unique. For a finite dictionary G, the minimum of the l_1-norms of output-weight vectors of shallow networks with units from G computing f is equal to a norm tailored to the dictionary G. This norm, called *G-variation*, has been used as a tool for estimation of rates of approximation of functions by networks with increasing "l_0-pseudo-norms". G-variation

is defined for a bounded subset G of a normed linear space $(\mathscr{X}, \|.\|)$ as

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

where $-G := \{-g \mid g \in G\}$, $\mathrm{cl}_{\mathscr{X}}$ denotes the closure with respect to the topology induced by the norm $\|\cdot\|_{\mathscr{X}}$, and conv is the convex hull. Variation with respect to the dictionary of Heaviside perceptrons (called *variation with respect to half-spaces*) was introduced by Barron [2] and we extended it to general sets in [11].

As G-variation is a norm, it can be made arbitrarily large by multiplying a function by a scalar. Also in theoretical analysis of approximation capabilities of shallow networks, it has to be taken into account that the approximation error $\|f - \mathrm{span}_n G\|$ in any norm $\|.\|$ can be made arbitrarily large by multiplying f by a scalar. Indeed, for every $c > 0$,

$$\|cf - \mathrm{span}_n G\| = c\|f - \mathrm{span}_n G\|.$$

Thus, both G-variation and errors in approximation by $\mathrm{span}_n G$ have to be studied either for sets of normalized functions or for sets of functions of a given fixed norm.

G-variation is related to l_1-sparsity, it can be used for estimating its lower bounds. The proof of the next proposition follows easily from the definition.

Proposition 1. *Let G be a finite subset of $(\mathscr{X}, \|.\|)$ with $\mathrm{card}\, G = k$. Then, for every $f \in \mathscr{X}$*

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

Another important property of G-variation is its role in estimates of rates of approximation by networks with small "l_0-pseudo-norms". This follows from the Maurey-Jones-Barron Theorem [3]. Here we state its reformulation from [11, 16, 12] in terms of G-variation merely for finite dimensional Hilbert space $(\mathscr{F}(X), \|.\|)$ with the Euclidean norm. By G^o is denoted the set of normalized elements of G, i.e., $G^o = \{\frac{g}{\|g\|} \mid g \in G\}$.

Theorem 2. *Let $X \subset \mathbb{R}^d$ be finite, G be a finite subset of $\mathscr{F}(X)$, $s_G = \max_{g\in G} \|g\|$, and $f \in \mathscr{F}(X)$. Then for every n,*

$$\|f - \mathrm{span}_n G\| \le \frac{\|f\|_{G^o}}{\sqrt{n}} \le \frac{s_G \|f\|_G}{\sqrt{n}}.$$

Theorem 2 together with Proposition 1 imply that for any function that can be l_1-sparsely represented by a shallow network with units from a dictionary G and for any n, there exists an input-output function f_n of a network with n units such that $f_n = \sum_{i=1}^{n} w_i g_i$ (and so $\|w\|_0 \le n$) such that

$$\|f - f_n\| \le \frac{s_G \|f\|_G}{\sqrt{n}}.$$

Lower bounds on variational norms can be obtained by geometrical arguments. The following theorem from [16] shows that functions which are "nearly orthogonal" to all elements of a dictionary G have large G-variations.

Theorem 3. *Let $(\mathscr{X}, \|.\|_{\mathscr{X}})$ be a Hilbert space and G its bounded subset. Then for every $f \in \mathscr{X} \setminus G^{\perp}$,*

$$\|f\|_G \ge \frac{\|f\|^2}{\sup_{g\in G} |g \cdot f|}.$$

4 Constructive Lower Bounds on Variational Norms

In this section, we derive lower bounds on l_1-sparsity of shallow signum perceptron networks from lower bounds of variational norms with respect to signum perceptrons.

Theorem 3 implies that functions which are nearly orthogonal to all elements of a dictionary have large variations. The inner product

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

of any two functions f, g on a finite domain X is invariant under reordering of the points in X.

To estimate inner products of functions with signum perceptrons on sets of points in general positions is quite difficult, so we focus on functions on square domains

$$X = \{x_1, \ldots, x_n\} \times \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$$

formed by points in grid-like positions. For example, pixels of pictures in $\mathbb{R}^d$ as well as digitized high-dimensional cubes can form such square domains. Functions on square domains can be represented by square matrices. For a function f on $X = \{x_1, \ldots, x_n\} \times \{y_1, \ldots, y_n\}$ we denote by $M(f)$ the $n \times n$ matrix defined as

$$M(f)_{i,j} = f(x_i, y_j).$$

On the other hand, an $n \times n$ matrix M induces a function f_M on X such that

$$f_M(x_i, y_j) = M_{i,j}.$$

We prove a lower bound on variation with respect to signum perceptrons for functions on square domains represented by matrices with orthogonal rows. To obtain these bounds from Theorem 3, we have to estimate inner products of these functions with signum perceptrons. We derive estimates of these inner products using the following two lemmas. The first one is from [13] and the second one follows from the Cauchy-Schwartz Inequality.

Lemma 1. *Let $d = d_1 + d_2$, $\{x_i \mid i = 1, \ldots, n\} \subset \mathbb{R}^{d_1}$, $\{y_j \mid j = 1, \ldots, n\} \subset \mathbb{R}^{d_2}$, and $X = \{x_1, \ldots, x_n\} \times \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$. Then for every $g \in P_d(X)$ there exists a*

reordering of rows and columns of the $n \times n$ matrix $M(g)$ such that in the reordered matrix each row and each column starts with a (possibly empty) initial segment of -1's followed by a (possibly empty) segment of $+1$'s.

Lemma 2. *Let M be an $n \times n$ matrix with orthogonal rows, $v_1, \ldots, v_n$ be its row vectors, $a = max_{i=1,\ldots,n} \|v_i\|$. Then for any subset I of the set of indices of rows and any subset J of the set of indices of columns of the matrix M,*

$$\left| \sum_{i \in I} \sum_{j \in J} M_{i,j} \right| \leq a \sqrt{\operatorname{card} I \operatorname{card} J}.$$

The following theorem gives a lower bound on variation with respect to signum perceptrons for functions induced by matrices with orthogonal rows.

Theorem 4. *Let M be an $n \times n$ matrix with orthogonal rows, $v_1, \ldots, v_n$ be its row vectors, $a = max_{i=1,\ldots,n} \|v_i\|$, $d = d_1 + d_2$, $\{x_i \,|\, i = 1, \ldots, n\} \subset \mathbb{R}^{d_1}$, $\{y_j \,|\, j = 1, \ldots, m\} \subset \mathbb{R}^{d_2}$, $X = \{x_i \,|\, i = 1, \ldots, n\} \times \{y_j \,|\, j = 1, \ldots, n\} \subset \mathbb{R}^d$, and $f_M : X \to \mathbb{R}$ be defined as $f_M(x_i, y_j) = M_{i,j}$. Then*

$$\|f_M\|_{P_d(X)} \geq \frac{a}{\lceil \log_2 n \rceil}.$$

Sketch of the proof.
For each signum perceptron $g \in P_d(X)$, we permute both matrices: the one induced by g and the one with orthogonal rows. To estimate the inner product of the permuted matrices, we apply Lemmas 1 and 2 to a partition of both matrices into submatrices. The submatrices of permuted $M(g)$ have all entries either equal to $+1$ or all entries equal to -1. The permuted matrix M has orthogonal rows and thus we can estimate from above the sums of entries of its submatrices with the same rows and columns as submatrices of $M(g)$. The partition is constructed by iterating at most $\lceil \log_2 n \rceil$-times the same decomposition. By Theorem 3,

$$\|f_M\|_{P_d(X)} \geq \frac{\|f_M\|^2}{\sup_{g \in P_d(X)} \langle f_M, g \rangle} \geq \frac{a}{\lceil \log_2 n \rceil}.$$

□

Theorem 4 shows that shallow perceptron networks computing functions generated by orthogonal $n \times n$ matrices must have l_1-norms bounded from bellow by $\frac{a}{\lceil \log_2 n \rceil}$.

All signum perceptrons on a domain X with $\operatorname{card} X = n \times n$ have norms equal to n. The largest lower bound implied by Theorem 4 for functions induced by $n \times n$ matrices with orthogonal rows, which have norms equal to n, is achieved for matrices where all rows have the same norms equal to $\sqrt{n}$. Such functions have variations with respect to signum perceptrons at least

$$\frac{\sqrt{n}}{\lceil \log_2 n \rceil}.$$

In particular, when the domain is the $2d$-dimensional Boolean cube $\{0,1\}^{2d} = \in^d \times \{0,1\}^d$, then the lower bound is

$$\frac{2^{d/2}}{d}.$$

So the lower bounds grows with d exponentially.

5 Computation of Functions Induced by Pseudo-Noise Sequences by Shallow Perceptron Networks

In this section, we apply our general results to a class of circulant matrices with rows formed by shifted segments of pseudo-noise sequences. These sequences are deterministic but exhibit some properties of random 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 $2^k - 1$. PN-sequences are generated by *primitive polynomials*. A polynomial

$$h(x) = \sum_{j=0}^{m} h_j x^j$$

is called *primitive polynomial of degree m* when the smallest integer n for which $h(x)$ divides $x^n + 1$ is $n = 2^m - 1$.

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 a 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 [18, p.410].

An important property of PN-sequences is their low autocorrelation. 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

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

For every PN-sequence and for every $t = 1, \ldots, 2^k - 2$,

$$\rho(t) = -\frac{1}{2^k - 1} \tag{3}$$

[18, p. 411].

Let $\tau : \{0,1\} \to \{-1,1\}$ be defined as

$$\tau(x) = -1^x$$

(i.e., $\tau(0) = 1$ and $\tau(1) = -1$). We say that a $2^k \times 2^k$ *matrix $L_k(\alpha)$ is induced by a k-th order PN-sequence* $\alpha = (a_0, a_1, \ldots, a_i, \ldots)$ when for all $i = 1, \ldots, 2^k$, $L_{i,1} = 1$, for all $j = 1, \ldots, 2^k$, $L_{1,j} = 1$, and for all $i = 2, \ldots, 2^k$ and $j = 2, \ldots, 2^k$

$$L_k(\alpha)_{i,j} = \tau(A_{i-1,j-1})$$

where A is the $(2^k - 1) \times (2^k - 1)$ circulant matrix with rows formed by shifted segments of length $2^k - 1$ of the sequence α. The next proposition following from the equations (2) and (3) shows that for any PN-sequence α the matrix $L_k(\alpha)$ has orthogonal rows.

Proposition 5. *Let k be a positive integer, $\alpha = (a_0, a_1, \ldots, a_i, \ldots)$ be a k-th order PN-sequence, and $L_k(\alpha)$ be the $2^k \times 2^k$ matrix induced by α. Then all pairs of rows of $L_k(\alpha)$ are orthogonal.*

Applying Theorem 4 to the $2^k \times 2^k$ matrice $L_k(\alpha)$ induced by a k-th order PN-sequence α we obtain a lower bound of the form $\frac{2^{k/2}}{k}$ on variation with respect to signum perceptrons of the function induced by the matrix $L_k(\alpha)$. So in any shallow perceptron network computing this function, the number of units or sizes of some output weights depend on k exponentially.

6 Discussion

We investigated limitations of shallow perceptron networks to sparsely represent real-valued functions. We considered sparsity measured by the l_1-norm which has been used in weight-decay regularization techniques [8] and in online dictionary learning [7]. We proved lower bounds on l_1-norms of output weight vectors of shallow signum perceptron networks computing functions on square domains induced by matrices with orthogonal rows. We illustrated our general results by an example of a class of matrices constructed using pseudo-noise sequences. These deterministic sequences mimic some properties of random sequences. We showed that shallow perceptron networks, which compute functions constructed using these sequences, must have either large numbers of hidden units or some of their output weights must be large.

There is an interesting analogy with the central paradox of coding theory. This paradox is expressed in the title of the article "Any code of which we cannot think is good" [6]. It was proven there that any code which is truly random (in the sense that there is no concise way to generate the code) is good (it meets the Gilbert-Varshamov bound on distance versus redundancy). However despite sophisticated constructions for codes derived over the years, no one has succeeded in finding a constructive procedure that yields such good codes. Similarly, computation of "any function of which we cannot think" (truly random) by shallow perceptron networks might be untractable. Our results show that computation of functions exhibiting some randomness properties by shallow perceptron networks is difficult in the sense that it requires networks of large complexities. Such functions can be constructed using deterministic algorithms and have many applications. Properties of pseudo-noise sequences were exploited for constructions of codes, interplanetary satellite picture transmission, precision measurements, acoustics, radar camouflage, and light diffusers. These sequences permit designs of surfaces that scatter incoming signals very broadly making reflected energy "invisible" or "inaudible".

Acknowledgments. This work was partially supported by the Czech Grant Agency grant GA 15-18108S and institutional support of the Institute of Computer Science RVO 67985807.

References

[1] L. J. Ba and R. Caruana. Do deep networks really need to be deep? In Z. Ghahrani and et al., editors, *Advances in Neural Information Processing Systems*, volume 27, pages 1–9, 2014.

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

[3] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory*, 39:930–945, 1993.

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

[5] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large-Scale Kernel Machines*. MIT Press, 2007.

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

[7] D. L. Donoho. For most large undetermined systems of linear equation the minimal l_1-norm is also the sparsest. *Communications in Pure and Applied Mathematics*, 59:797–829, 2006.

[8] T. L. Fine. *Feedforward Neural Network Methodology*. Springer, Berlin Heidelberg, 1999.

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

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

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

[12] V. Kůrková. Complexity estimates based on integral transforms induced by computational units. *Neural Networks*, 33:160–167, 2012.

[13] V. Kůrková. Constructive lower bounds on model complexity of shallow perceptron networks. *Neural Computing and Applications*, DOI 10.1007/s00521-017-2965-0, 2017.

[14] V. Kůrková and M. Sanguineti. Model complexities of shallow networks representing highly varying functions. *Neurocomputing*, 171:598–604, 2016.

[15] V. Kůrková and M. Sanguineti. Probabilistic lower bounds for approximation by shallow perceptron networks. *Neural Networks*, 91:34–41, 2017.

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

[17] S. B. Laughlin and T. J. Sejnowski. Communication in neural networks. *Science*, 301:1870–1874, 2003.

[18] F. MacWilliams and N. A. Sloane. *The Theory of Error-Correcting Codes*. North Holland Publishing Co., 1977.

[19] V. E. Maiorov and R. Meir. On the near optimality of the stochastic approximation of smooth functions by neural networks. *Advances in Computational Mathematics*, 13:79–103, 2000.

[20] V. E. Maiorov and A. Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25:81–91, 1999.

[21] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[22] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[23] A. M. Tillmann. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letters*, 22:45–49, 2015.