Limitations of Shallow Networks

Věra Kůrková

Abstract Although originally biologically inspired neural networks were introduced as multilayer computational models, shallow networks have been dominant in applications till the recent renewal of interest in deep architectures. Experimental evidence and successful applications of deep networks pose theoretical questions asking: When and why are deep networks better than shallow ones? This chapter presents some probabilistic and constructive results on limitations of shallow networks. It shows implications of geometrical properties of high-dimensional spaces for probabilistic lower bounds on network complexity. The bounds depend on covering numbers of dictionaries of computational units and sizes of domains of functions to be computed. Probabilistic results are complemented by constructive ones built using Hadamard matrices and pseudo-noise sequences.

1 Introduction

Originally, biologically inspired neural networks were introduced as multilayer computational models, but later one-hidden-layer (shallow) architectures became dominant in applications (see, e.g., [18, 31] and the references therein). Although multilayer networks with sigmoidal and convolutional units used as filters were proposed for pattern recognition tasks by LeCun [41, 42] already in 1990s, their training by back-propagation was inefficient till the advent of fast graphic processing units (GPU). While development of GPU was motivated commercially as a tool for computer games, they enabled the revival of interest in multilayer architectures. Around 2006, a group of researchers from the Canadian Institute for Advanced Research (Bengio, Hinton, LeCun) exploited them in training networks with several convolutional and pooling layers (see, e.g., the survey article [43]). These networks were

V. Kůrková (✉)
Institute of Computer Science of the Czech Academy of Sciences,
Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic
e-mail: vera@cs.cas.cz

L. Oneto et al. (eds.), *Recent Trends in Learning From Data*, Studies
in Computational Intelligence 896, https://doi.org/10.1007/978-3-030-43883-8_6

called *deep* [10, 20] to distinguish them from *shallow* ones with merely one hidden layer. Currently, deep networks are the state of the art in areas such as text classification, musical genre recognition, speech recognition, time-series prediction, object detection, localization, video and tomography images recognition, biomedical image analysis, hyperspectral image analysis, and in combination with tree search in automatic game playing (AlphaGO).

While experimental research of deep networks is rapidly evolving, theoretical analysis complementing the empirical evidence is still in its early stages. There are fundamental wide open questions related to the role of depth of network architectures: Why should deep networks be better than shallow ones and under which conditions?

Bengio and LeCun, who 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" [9]. However, reservations about overall lower complexity of deep networks over shallow ones have appeared. An 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 [4]. Mhaskar et al. [50] suggested that due to their hierarchical structure, deep networks could outperform shallow networks in visual recognition of pictures with objects of different scales. Characterization of functions, which can be computed by deep networks of smaller model complexities than shallow ones, can be derived by comparing lower bounds on numbers of units in shallow networks with upper bounds on numbers of units in deep ones.

It has long been known that under mild conditions on types of computational units, shallow networks have the *universal representation property*, i.e., they can exactly compute any real-valued function on a finite domain [22]. However, the arguments proving this property assume that the number of units in the last hidden layer is potentially as large as the size of the domain. Obviously, not all functions require networks with such high numbers of units. For shallow networks, various upper bounds on numbers of hidden units needed for a given approximation accuracy in dependence on their types, input dimensions, and types of functions to be computed are known (see, e.g., [23] and the references therein).

Derivation of lower bounds is much more difficult than derivation of upper ones. Poggio et al. [53] proposed as a potential tool for comparison of deep and shallow networks an application of the topological approach for obtaining lower bounds on complexity of shallow networks exhibiting the "curse of dimensionality" (i.e., an exponential dependence on the number of parameters [8]) from [14]. However, applicability of topological methods is limited only to classes of networks where best or near best approximation of functions can be obtained by a continuous selection of network parameters. We proved in [28–30] that in many common classes of networks such continuous selection is not possible due their nonlinear and non-convex nature. Other lower bounds hold merely for types of computational units that are not commonly used such as perceptrons with specially designed activation functions [47] or the lower bounds merely prove existence of worst-case errors in Sobolev spaces asymptotically [46].

In this chapter, we survey recent results on complexity and sparsity of shallow networks. Minimization of "l_0-pseudonorm", which formalizes the concept of network sparsity measured by the number of hidden units in a shallow network, is a difficult non convex optimization problem. Thus we focus on investigation of minima of l_1-norms of output-weight vectors. We present several arguments showing that l_1-norm is a good approximation of "l_0-pseudonorm" (it approximates its convexification, can be used as a stabilizer in weight-decay regularization [18], and is related to variational norm tailored to a dictionary of computational units).

In practical applications, feedforward networks compute functions on finite domains (formed, e.g., by pixels of pictures, discretized cubes, or scattered vectors of data), which are often quite large. Functions on finite domains form linear spaces which are isomorphic to Euclidean spaces of dimensions equal to sizes of domains. Geometry of high-dimensional spaces has many counter-intuitive features, which have consequences for correlations between functions on large domains. We show that combination of concentration of measure property of high-dimensional spaces with characterization of dictionaries of computational units in terms of their capacity and coherence described by their covering numbers leads to lower bounds on variational norms and l_1-norms of output-weight vectors of shallow networks. Applying these estimates to dictionaries with power-type covering numbers, we conclude that computation of almost any uniformly randomly chosen function on a large domain requires either large number of units or is unstable as some output weights are large. Finally, we illustrate the probabilistic results by a concrete construction of a class of functions induced by matrices, which have large variational norms with respect to the dictionary of signum perceptrons [36].

The chapter is organized as follows. Section 2 contains basic concepts and notations on feedforward networks and dictionaries of computational units. In Sect. 3, various measures of network sparsity and their relationships are studied. In Sect. 4, properties of high-dimensional spaces are applied to obtain estimates of correlations between functions to be computed and computational units. In Sect. 5, lower bounds on variational and l_1-norms formulated in terms of covering numbers of dictionaries and sizes of the domains are derived. In Sect. 6, some estimates of sizes of dictionaries of computational units popular in neurocomputing are presented. In Sect. 7 probabilistic results are complemented by constructive ones. Section 8 contains some examples and Sect. 9 is a brief discussion.

2 Preliminaries

For $X \subset \mathbb{R}^d$, we denote by

$$\mathcal{F}(X) := \{f \mid f : X \to \mathbb{R}\}$$

the *set of all real-valued functions on* X. In practical applications, domains $X \subset \mathbb{R}^d$ are finite, but their sizes card X and/or input dimensions d can be quite large.

Fixing a linear ordering $\{x_1, \dots, x_m\}$ of elements of X we define an isomorphism $\iota : \mathcal{F}(X) \to \mathbb{R}^m$ as $\iota(f) := (f(x_1), \dots, f(x_m))$ and thus we identify $\mathcal{F}(X)$ with the finite dimensional Euclidean space $\mathbb{R}^m$. On $\mathcal{F}(X)$ we denote the induced inner product by

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

the Euclidean norm $\|f\|_2 := \sqrt{\langle f, f \rangle}$, and by $S_1(X)$ the unit sphere in $\mathcal{F}(X)$

$$S_1(X) = \{f \in \mathcal{F}(X) \mid \|f\| \le 1\}.$$

By

$$\mathcal{B}(X) := \{f \mid f : X \to \{-1, 1\}\}$$

we denote the *subset of* $\mathcal{F}(X)$ *formed by functions with values in* $\{-1, 1\}$.

For any norm or "pseudonorm" $\|.\|$ on $\mathbb{R}^d$ or $\mathcal{F}(X)$, we denote by

$$B_r(\|.\|) = \{w \in \mathbb{R}^n \mid \|w\| \le r\}$$

the ball of radius r in $\|.\|$.

The set of input-output functions of a *feedforward network with a single linear output* has the form

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

where $w = (w_1, \dots, w_n)$ is the vector of output weights and G is a parameterized family of functions called a *dictionary*. The dictionary depends on the network architecture and types of computational units. The simplest architecture is a *shallow (one-hidden-layer) network*, where G is a parameterized family of functions computable by a given type of computational units. In the case of a *deep network* with several hidden layers, G is formed by combinations and compositions of functions representing units from lower layers. Formally, a dictionary can be described as

$$G(X) = G_\phi(X, Y) := \{\phi(\cdot, y) : X \to \mathbb{R} \mid y \in Y\},$$

where $\phi : X \times Y \to \mathbb{R}$ is a function of two variables: an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter vector $y \in Y \subseteq \mathbb{R}^s$.

Popular computational units are *perceptrons* which compute functions of the form $\psi(v \cdot x + b)$, where $v \in \mathbb{R}^d$ is a *weight vector*, $b \in \mathbb{R}$ a *bias*, and $\psi : \mathbb{R} \to \mathbb{R}$ is an *activation function* (such as Heaviside, sigmoidal, rectified linear).

By

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

we denote the set of functions computable by networks with at most n units in the last hidden layer. Sets of the form $\mathrm{span}_n\, G$ are invariant under multiplication by scalars, i.e., $c\mathrm{span}_n\, G = \mathrm{span}_n\, G$ for all $c \in \mathbb{R}$. As for all $c > 0$

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

with proper choices of scalars, examples of functions with arbitrarily large or small errors in approximation by sets $\mathrm{span}_n\, G$ in any norm $\|.\|$ can be obtained. So approximation and representation of functions by networks with a linear output have to be studied for cases when functions to be approximated and function from G have the same norms, e.g., when all functions are normalized or in the case of binary classification, they have values in $\{-1, 1\}$ rather than in $\{-0, 1\}$.

3 Approximate Measures of Sparsity

It has long been known that many feedforward networks have the *universal representation property*, i.e., they can exactly compute any function on a finite domain. Ito [22] proved the following sufficient condition on a dictionary of computational units that guarantees that shallow networks with units from the dictionary have the universal representation property.

Theorem 1 *Let m be a positive integer, $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, and $G_\phi(X, Y) = \{\phi(\cdot, y) : X \to \mathbb{R} \mid y \in Y\}$ be such that there exist $y_1, \ldots, y_m \in Y$ for which the $m \times m$ square matrix Φ defined as $\Phi_{i,j} = \phi(x_i, y_j)$ is regular, then $\mathcal{F}(X) =$* $\mathrm{span}_m\, G_\phi(X, Y)$.

Regularity of the matrix Φ implies that for any $f : \{x_1, \ldots, x_m\} \to \mathbb{R}$, the family of m linear equation

$$f(x_i) = \sum_{j=1}^{m} w_j \phi(x_i, y_j),\ \ i = 1, \ldots, m \tag{1}$$

with m unknown has a solution. Any solution $(w_1, \ldots, w_m)$ can be used as an output-weight vector of a representation of f as an input-output function of a network with units from G_ϕ of the form

$$f(x) = \sum_{i=1}^{m} w_i \phi(x, y_i). \tag{2}$$

Ito [22] verified that shallow networks with sigmoidal perceptrons satisfy the condition of Theorem 1 and thus have the universal representation property. It is easy to check that Theorem 1 also implies that this property is possessed by shallow networks with any positive definite kernel (e.g., Gaussian, Laplace). Positive definiteness of a kernel guarantees that the matrix induced by the kernel with $x_i = y_i$, $i = 1, \dots, m$ is regular.

The parameters $y_1, \dots, y_m$, for which the matrix Φ is regular, as well as the solution $w_1, \dots, w_m$ of the family of m linear equations (1) need not to be unique. Thus there might exist many representations of a function f as an input-output function of a shallow network with units from G_ϕ. However, potentially all $w_1, \dots, w_m$ might be nonzero. Thus for large domains X, networks whose existence is guaranteed by universality results such as Theorem 1 might be too large for efficient implementations.

Many dictionaries popular in neurocomputing are linearly independent on infinite domains (see, e.g., [1, 26, 27, 39, 57]). Representations of functions as input-output functions of shallow networks with units from such dictionaries are unique up to permutations of hidden units and, in some cases, also sign-flips. In contrast, such dictionaries restricted to finite domains typically are linearly dependent. The condition of being equal on the whole $\mathbb{R}^d$ or its sufficiently large compact subset is much stronger than the condition requiring equality merely on its finite discrete subset. In some literature, dictionaries which are not linearly independent are called *overcomplete*. Such dictionaries allow multiple representations of functions.

For a function $f \in \mathcal{F}(X)$ and a dictionary G, we denote by

$$W_f(G) := \{w = (w_1, \dots, w_n) \in \mathbb{R}^n \mid f = \sum_{i=1}^{n} w_i g_i,\, g_i \in G, n \in \mathbb{N}\} \tag{3}$$

the set of output-weight vectors of shallow networks with units from G representing f. When G induces a class of shallow networks having the universal representation capability, then sets $W_f(G)$ are nonempty for all $f \in \mathcal{F}(X)$. It follows from the definition that sets $W_f(G)$ are convex.

Proposition 1 *Let $X \subset \mathbb{R}^d$, $G \subset \mathcal{F}(X)$, and $f \in \mathcal{F}(X)$, then $W_f(G)$ is convex.*

It is desirable to find among all representations of f as an input-output function of a shallow network with units from G the most sparse ones, i.e., in the set $W_f(G)$ to find vectors with the smallest number of nonzero entries.

Formally, for a vector $w \in \mathbb{R}^n$, the *number of its non-zero entries* is denoted $\|w\|_0$. It is called "*l_0-pseudonorm*" in quotation marks as it is neither a norm nor a pseudonorm. It satisfies the triangle inequality, but it does not satisfy the homogeneity condition, which requires $|\lambda| \, \|w\| = \|\lambda w\|$ for all $\lambda \in \mathbb{R}$. The values of $\|.\|_0$ are only integers and its "balls" are not convex. "l_0-pseudonorm" satisfies the equation

$$\|w\|_0 = \sum_{i=1}^{n} w_i^0$$

and it is a limit

$$\lim_{p\to\infty} \|w\|_p = \|w\|_0$$

of l_p-functionals.

$W_f(G)$ is convex and any continuous function on a convex set achieves its minimum. But $\|.\|_0$ is not continuous. Minimization of "l_0-pseudonorm" is a difficult non convex problem which has been studied in signal processing (see, e.g, [15, 16]). It was proven that in some cases, it is NP-hard [60].

Due to its non homogeneity, "l_0-pseudonorm" is invariant under multiplication by scalars. In contrast, any norm can be made arbitrarily large or small by multiplying a function by a suitable scalar. Thus investigation of relationships of "l_0-pseudonorm" to various norms has sense only for functions from restricted ambient sets. The following proposition from [52] shows that when the ambient set is the unit ball in l_2-norm, then the ball of radius $\sqrt{r}$ in the l_1-norm (hyperoctahedron) is a good approximation of the convexification of the "ball" of radius r in "l_0-pseudonorm".

Proposition 2 *For every positive integer m and every $r > 0$, balls in $\|.\|_0$, $\|.\|_1$, and $\|.\|_2$ in $\mathbb{R}^d$ satisfy*

$$\operatorname{conv}\left(B_r(\|.\|_0) \cap B_1(\|.\|_2)\right) \subset B_{\sqrt{r}}(\|.\|_1) \cap B_1(\|.\|_2) \subset 2\operatorname{conv}\left(B_r(\|.\|_0) \cap B_1(\|.\|_2)\right).$$

In neurocomputing, instead of "l_0-pseudonorm", l_1 and l_2-norms have been used as stabilizers in weight-decay regularization techniques [18]. Acting as a stabilizer, l_2-norm penalizes even a small number of large output weights but it can tolerate many small ones, while l_1-norm stabilizers penalize many small output weights as well as few large ones. This can be illustrated by a simple example of a weight vector $w \in \mathbb{R}^m$, with $w_i = \frac{c}{m}$ for all $i = 1, \dots, m$. Then $\|w\|_1 = c$, while $\|w\|_2 = \frac{c}{\sqrt{m}}$. So their difference increases with growing dimension m. Networks with large l_1-norms of output-weight vectors have either large numbers of units or some of their output weights are large. None of these properties is desirable: implementation of networks with large numbers of units might not be feasible, while large output weights might lead to an instability of computation.

In addition to approximating the convexification of "l_0-pseudonorm" and penalizing many small output weights, l_1-norm have several other properties which make it a good approximate measure of sparsity. We denote

$$W_f(G)_1^* := \{w^* \in W_f(G) \mid \|w^*\|_1 = \min_{w\in W_f(G)} \|w\|_1\}$$

and

$$W_f(G)_2^* := \{w^* \in W_f(G) \mid \|w^*\|_2 = \min_{w\in W_f(G)} \|w\|_2\}$$

the subsets of $W_f(G)$ formed by output-weight vectors of minimal l_1 and l_2-norms, resp.

Proposition 3 *Let $X \subset \mathbb{R}^d$, $f \in \mathcal{F}(X)$, and $G = \{g_1, \dots, g_k\} \subset \mathcal{F}(X)$. If $W_f(G)$ is non empty, then $W_f(G)^*_1$ is non empty and convex.*

Proof l_1-norm is continuous and every continuous function on a convex set achieves its minimum, so $W_f(G)^*$ is non empty. Its convexity follows from the definition. □

Note that l_2-norm does not satisfy an analogy to Proposition 3, namely the set $W_f(G)^*_2$ of vectors with minimal l_2-norms contains only one point. Indeed, the strict convexity of l_2-norm implies that

$$\|aw_1 + (1-a)w_2\|_2 < a\|w_1\|_2 + (1-a)\|w_2\|_2$$

for all $a \in (0, 1)$. Thus Proposition 3 shows another advantage of l_1-norm over l_2-norm.

Moreover, the minimal value of the l_1-norm of an output-weight vector of a network computing a function f can be expressed in terms of a norm generated by a dictionary G called *G-variation*. It is defined for a bounded subset G of a normed linear space $(\mathcal{X}, \|.\|)$ as

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

where $-G := \{-g \mid g \in G\}$, $\mathrm{cl}_{\mathcal{X}}$ denotes the closure with respect to the topology induced by the norm $\|\cdot\|_{\mathcal{X}}$, and conv is the convex hull. If the set over which the infimum is taken is empty, then $\|f\|_G := \infty$. So G-variation is the Minkowski functional of its unit ball $\mathrm{cl}_{\mathcal{X}} \operatorname{conv}(G \cup -G)$.

Variation with respect to Heaviside perceptrons (called *variation with respect to half-spaces*) was introduced in [6] and extended to general dictionaries in [34]. It was shown in [37] that infimum in the definition of G-variation can be replaced with minimum.

The next proposition shows that $\|f\|_G$ bounds the minimum of values of l_1-norms of output-weight vectors of networks with units from G computing f. Its proof follows directly from the definition.

Proposition 4 *Let $X \subset \mathbb{R}^d$, G be a bounded subset of $\mathcal{F}(X)$, and $f \in \mathcal{F}(X)$ such that $\|f\|_G$ is finite. Then $\|f\|_G \le \|w\|_1$ for all $w \in W_f(G)$. When G is finite, then $\|f\|_G = \|w^*\|_1$ for all $w^* \in W_f(G)^*_1$.*

Besides of being a lower bound on approximate measure of sparsity expressed in terms of l_1-norm, G-variation is also a critical factor in upper bounds on rates of approximation by networks with increasing "l_0-pseudonorms" of output-weight vectors. The following theorem is a special case holding for the Hilbert space $\mathcal{F}(X)$ of the Maurey–Jones–Barron Theorem [7] as reformulated in terms of a variational norm in [35, 37].

Theorem 2 *Let $X \subset \mathbb{R}^d$ be finite, G be a subset of $\mathcal{F}(X)$, $s_G = \max_{g \in G} \|g\|_2$, and $f \in \mathcal{F}(X)$. Then for every n,*

$$\|f - \operatorname{span}_n G\|_2 \leq \frac{s_G \, \|f\|_G}{\sqrt{n}}.$$

By Theorem 2 there exist functions computable by shallow networks with at most n hidden units from the dictionary G (networks with output-weight vectors with "l_0-pseudonorms" at most n) approximating f within $\frac{s_G \|f\|_G}{\sqrt{n}}$.

4 Correlation and Concentration of Measure

As mentioned above, l_2-errors in approximation by families of the form $\operatorname{span}_n G$ can only be compared when functions to be approximated and elements of dictionaries G have the same l_2-norms (e.g., when all functions are normalized). The Euclidean distance of normalized functions on the unit sphere $S_1(X)$ in $\mathcal{F}(X)$ is related to the *angular pseudometrics* ρ on $S_1(X)$ defined as

$$\rho(f, g) = \arccos |\langle f, g \rangle|.$$

Note that ρ is not metrics, it is merely a pseudometrics, because the distance between f and $-f$ is zero. It is related to the l_2-metrics by the formula

$$\|f - g\|_2 = 2 \sin(\alpha/2) \quad \text{where} \quad \rho(f, g) = \alpha.$$

It can be described in terms of *correlation* defined as the inner product $\langle f, g \rangle$. The more correlated functions are, the better they can approximate each other.

As $\mathcal{F}(X)$ is isometric to the Euclidean space $\mathbb{R}^{\operatorname{card} X}$, with increasing size of the domain X, effects of high-dimensional geometry become apparent. In particular on high-dimensional spheres, inner products with any fixed function tend to concentrate around their median. Let

$$C(g, \varepsilon) = \{f \in S^{m-1} \mid \langle f, g \rangle \geq \varepsilon\} \tag{4}$$

denotes the *spherical cap* centered at a fixed vector g, which contains all vectors f which have the angular distance from g at most $\alpha = \arccos \varepsilon$ or equivalently the inner product $\langle f, g \rangle$ is at least ε (see Fig. 1).

Using classical calculus (integration in spherical polar coordinates), one can compute the relative area of the unit sphere S^{m-1} in the m-dimensional Euclidean space $\mathbb{R}^m$, which is occupied by the spherical cap

$$\mu(C(g, \varepsilon)) \leq e^{-\frac{m\varepsilon^2}{2}} \tag{5}$$

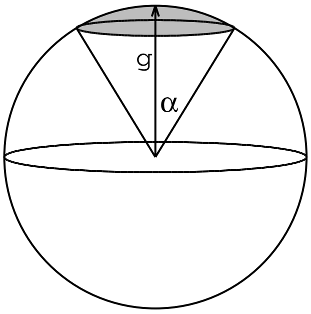

Fig. 1 Spherical cap

(see, e.g., [5]). For a fixed angle α, with increasing dimension m the normalized surface area μ of such cap decreases exponentially fast to zero. When ε is small, the complement of $C(g, \varepsilon) \cup C(-g, \varepsilon)$ contains vectors which are nearly orthogonal to g. The upper bound (5) implies that most of the area of a high-dimensional sphere is concentrated around its "equator".

The exponential decrease of sizes of "polar caps" (5) is the very essence of two properties of high-dimensional spaces called the *curse of dimensionality* and the *blessing of dimensionality*. While there are only m exactly orthogonal unit vectors in $\mathbb{R}^m$, for a fixed $\varepsilon > 0$, the number of ε-quasiorthogonal vectors (with absolute values of inner products at most ε) grows with m exponentially. So the number of highly uncorrelated functions grows with the size of their domain exponentially. On the other hand, (5) implies the phenomenon of *concentration of measure*. The upper bound (5) on the size of a spherical cap can be rephrased as follows: inner products of a fixed vector in the sphere S^{m-1} with uniformly randomly chosen vectors concentrate around zero. A generalization obtained by replacing the inner product with a sufficiently smooth (Lipschitz) function on the sphere leads to the Lèvy Lemma [44]. It states that almost all values of a Lipschitz function on a high-dimensional sphere are close to their median.

Recall that on a metric space (S, ρ) a function $h : S \to \mathbb{R}$ is called *Lipschitz with a constant c* if for all $x, y \in S$, $|h(x) - h(y)| \leq c\rho(x, y)$. For a probability measure P on S^{m-1} and a function $F : S^{m-1} \to \mathbb{R}$, the *median* of F is defined as

$$\text{med}(F) := \sup\{t \in \mathbb{R} \,|\, \text{P}[F(x) \leq t] = 1/2\}$$

and it satisfies $\text{P}[F(x) < \text{med}(F)] = 1/2$ and $\text{P}[F(x) > \text{med}(F)\} = 1/2$ [49, p. 337].

Theorem 3 (Lévy Lemma) *Let m be a positive integer,* P *be the uniform probability measure (normalized surface measure) on* S^{m-1}, $F : S^{m-1} \to \mathbb{R}$ *be a function, and* $\varepsilon \in [0, 1]$. *Then*
(i) for F continuous with modulus of continuity ω

$$\text{P}[|\, F(x) - \text{med}(F)\,| > \omega(\varepsilon)] \leq 2e^{-\frac{m\varepsilon^2}{2}};$$

(ii) for F 1-Lipschitz.

$$\mathrm{P}[|\, F(x) - \mathrm{med}(F)\,| > \varepsilon] \leq 2e^{-\frac{m\varepsilon^2}{2}}.$$

Note that the median of a Lipschitz function is related to its mean value $\mathrm{E}[F] = \int_{S^{m-1}} F(x)\, d\mathrm{P}(x)$ as follows [49, p. 338].

Proposition 5 *Let m be a positive integer,* P *be the uniform probability measure (normalized surface measure) on S^{m-1} and $F : S^{m-1} \to \mathbb{R}$ be 1-Lipschitz. Then*

$$|\mathrm{med}(F) - \mathrm{E}[F]| \leq \frac{12}{\sqrt{m}}.$$

Thus on high-dimensional spheres, Lipschitz functions are almost constant. This property of high-dimensional spheres is a special case of properties called the *concentration of measure phenomenon*. Similar property was also discovered in probability theory, where it has been studied in terms of bounds on large deviations of sums of random variables by Hoeffding [21], Chernoff [11], and Azuma [3]. Concentration of measure is also the essence of the proof of the Johnson–Lindenstrauss Flattening Lemma [49, p. 358]. It guarantees a possibility of dimension reduction of d-dimensional data by a random projection to a lower dimension bounded from below by $\frac{8}{\varepsilon} \log d$ such that the projection is a near-isometry (preserves distances within a multiplicative factor $1 \pm \varepsilon$).

5 Probabilistic Lower Bounds on Approximate Measures of Sparsity

Concentration of measure phenomenon has implications for correlations of functions on large domains with functions from dictionaries of computational units. They play a crucial role in estimating variational norms and the minima of l_1-norms of output-weight vectors of networks. Here, we state a special case of a geometric characterization of G-variation proven in [37]. Its proof is based on Hahn-Banach Theorem on separation of a point from convex set by a hyperplane. By $G^{\perp}$ is denoted the *orthogonal complement of* G in the Hilbert space $\mathcal{F}(X)$.

Theorem 4 *Let X be a finite subset of $\mathbb{R}^d$ and G be a bounded subset of $\mathcal{F}(X)$. Then for every $f \in \mathcal{F}(X) \setminus G^{\perp}$, $\|f\|_G \geq \frac{\|f\|^2}{\sup_{g \in G} |\langle g, f \rangle|}$.*

Theorem 4 gives a geometric insight into the concept of variational norm. It implies that functions which are "nearly orthogonal" to all elements of a dictionary have large variations and thus cannot be computed by networks having "small" l_1-norms of output-weigh vectors (see Fig. 2).

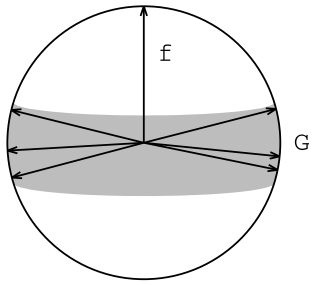

Fig. 2 Function nearly orthogonal to G

The next Corollary gives a lower bound on probability that a uniformly randomly chosen normalized function on X is in the "polar cap"

$$C(g,\varepsilon) = \{f \in S_1(X) \mid \langle f, g\rangle \geq \varepsilon\}$$

of angle $\arccos \varepsilon$ around a given g in the space $\mathcal{F}(X)$ (we use the same notation $C(g,\varepsilon)$ as for "polar cap" in S^{m-1}).

Corollary 1 *Let $X \subset \mathbb{R}^d$ be finite with* $\operatorname{card} X = m$, *$g \in S_1(X)$, and $\varepsilon \in [0,1]$. Then for f uniformly randomly chosen in $S_1(X)$*

$$\mathrm{P}[|\langle f, g\rangle| > \varepsilon] \leq 2e^{\frac{-m\varepsilon^2}{2}}.$$

Proof Let $F_g : S_1(X) \to \mathbb{R}$ be defined as $F_g(f) = \langle f, g\rangle$. By the Cauchy–Schwartz Inequality $|F_g(f_1) - F_g(f_2)| \leq \|g\|_2 \, \|f_1 - f_2\|_2$. Thus F_g is 1-Lipschitz. By symmetry, its median is zero. So the statement follows from the Lévy Lemma (Therom 3). □

Corollary 1 shows that for a fixed normalized function g on a large domain X, most of the area of $S_1(X)$ (the complement of the union of the two "polar caps" formed by functions close to g and $-g$, formally $\{f \in S_1(X) \mid |\langle f, g\rangle| > \varepsilon\} \cup \{f \in S_1(X) \mid |\langle f, g\rangle| > \varepsilon\}$), contains functions which are nearly orthogonal to g.

It implies that if a dictionary G is not large enough to outweigh the factor $2e^{\frac{-m\varepsilon^2}{2}}$, then most functions in $S_1(X)$ have inner products with all elements of G at most ε. It means that for a large domain X, most functions in $S_1(X)$ are nearly orthogonal to all elements of G and thus by Theorem 4, they have G-variation larger than $\frac{1}{\varepsilon}$ (see Fig. 3). Combining Corollary 1 and Theorem 4, we obtain for a finite dictionary the following probabilistic lower bound on G-variation.

Theorem 5 *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* $\operatorname{card} X = m$, P *be a uniform probability measure on $S_1(X)$, $b > 0$, and $G \subset S_1(X)$ be finite with* $\operatorname{card} G = k$. *Then for f uniformly randomly chosen in $S_1(X)$*

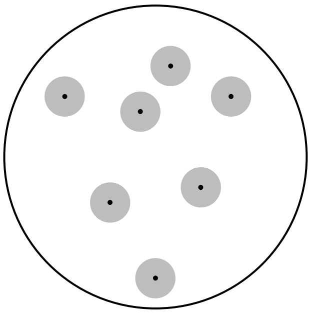

Fig. 3 Spherical caps around elements of G

$$\mathrm{P}[\|f\|_G \geq b] \geq 1 - 2k\, e^{-\frac{2m}{b^2}}.$$

Theorem 5 estimates probability that a uniformly randomly chosen function on $S_1(X)$ has G-variation at least b. Hence by Proposition 4, any representation of such function as an input-output function of a shallow network with units from G has l_1-norm of output-weight vector larger than b, too.

Similar estimate holds for dictionaries of binary-valued functions. Its proof in [38] uses a discrete version of concentration of measure in the form of Chernoff-Hoeffding Bound on sums of independent variables.

Theorem 6 *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* $\operatorname{card} X = m$, P *be a uniform probability measure on $c\mathcal{B}(X)$, $b > 0$, and $G \subset \mathcal{B}(X)$ be finite with* $\operatorname{card} G = k$. *Then for f uniformly randomly chosen in $\mathcal{B}(X)$*

$$\mathrm{P}\,(\|f\|_G \geq b) \geq 1 - k\, e^{-\frac{m}{2b^2}}.$$

Estimates from Theorem 5 can be extended also to networks with units from infinite dictionaries. Their "sizes" can be described in terms of covering and packing numbers. They were introduced in [32] as a way to measure sizes of subsets of metric spaces using as measuring units small balls. For $\varepsilon > 0$, an *ε-net in G* is a set $\{g_1, \ldots, g_n\} \subseteq G$ such that the family of the closed balls $B_\varepsilon(g_i)$ of radii ε centered at g_i covers G. The *ε-covering number* denoted $\mathcal{N}_\varepsilon(G)$ of a subset G of a metric space $\mathcal{S}$ is the cardinality of a minimal ε-net in G, i.e.,

$$\mathcal{N}_\varepsilon(G) := \min\Big\{n \in \mathbb{N}_+ \mid (\exists f_1, \ldots, f_m \in G)\,(G \subseteq \bigcup_{i=1}^{n} B_\varepsilon(f_i))\Big\}.$$

When the set over which the minimum is taken is empty, then $\mathcal{N}_\varepsilon(G) := +\infty$. Note that all covering numbers of a compact set are finite. *Packing number* $\mathcal{M}_\varepsilon(G)$ is defined as the maximal number of disjoint balls that fit in a set, i.e.,

$$\mathcal{M}_\varepsilon(G) := \min\bigl\{n \in \mathbb{N}_+ \mid (\exists f_1, \dots, f_m \in G)\,(\bigcup_{i=1}^{n} B_\varepsilon(f_i) \subseteq G)\bigr\}.$$

Packing numbers are closely related to covering numbers as

$$\mathcal{M}_{2\varepsilon}(G) \le \mathcal{N}_\varepsilon(G) \le \mathcal{M}_\varepsilon(G).$$

In the following theorem, we assume that the covering numbers of subsets G of $S_1(X)$ are considered with respect to the angular pseudometrics $\rho(f, g) = \arccos |\langle f, g\rangle|$.

Theorem 7 *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* $\mathrm{card}\, X = m$*,* P *be a uniform probability measure on $S_1(X)$, $b > 0$, and $G \subset S_1(X)$ has finite covering numbers. Then for f uniformly randomly chosen in $S_1(X)$*

$$\mathrm{P}[\|f\|_G \ge b] \ge 1 - 2\mathcal{N}_{\arccos(2/b)}(G)\, e^{-\frac{2m}{b^2}}.$$

Proof For $g \in S_1(X)$ and $\varepsilon > 0$, let $C(g, \varepsilon) = \{f \in S_1(X) \mid \langle f, g\rangle \ge \varepsilon\}$. Let $\alpha = \arccos(2/b)$ and $\{g_1, \dots, g_n\}$ be a minimal α-net in G in the angular pseudometrics ρ. Then by the triangle inequality $\bigcup_{i=1}^n (C(g_i, 2/b) \cup C(-g_i, 2/b)) \supseteq \bigcup_{g\in G} C(g, 1/b)$. By Theorem 4, $\|f\|_{G(X)} \ge \frac{1}{\sup_{g\in G} |\langle f,g\rangle|}$. Thus $\{f \in S_1(X) \mid \|f\|_G \ge b\} \supseteq S_1(X) \setminus \bigcup_{g\in G} C(g, 1/b) \supseteq S_1(X) \setminus \bigcup_i^n (C(g_i, 2/b) \cup C(-g_i, 2/b))$. By Corollary 1, $\mathrm{P}[f \in C(g, 1/b)] \le 2e^{-\frac{2m}{b^2}}$. Thus $\mathrm{P}[f \in S_1(X) \setminus \bigcup_{i=1}^n (C(g_i, 2/b) \cup C(-g_i, 2/b))] \ge 1 - 2\mathcal{N}_\alpha(G)e^{-\frac{2m}{b^2}}$ and so the statement holds. □

Combining Theorem 7 with Proposition 4 we obtain the following lower bound on the l_1-norm of the output-weight vector of any network with units from G computing a uniformly randomly chosen real-valued function on X.

Corollary 2 *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* $\mathrm{card}\, X = m$*,* P *be a uniform probability measure on $S_1(X)$, $b > 0$, $G \subset S_1(X)$ has finite covering numbers, and f be a function uniformly randomly chosen from $S_1(X)$. Then for f uniformly randomly chosen in $S_1(X)$*

$$\mathrm{P}\bigl[(\forall w \in W_f(G))\,(\|w\|_1 \ge b)\bigr] \ge 1 - 2\mathcal{N}_{\arccos(2/b)}(G)\, e^{-\frac{2m}{b^2}}.$$

For example for $b = m^{1/4}$, Theorem 7 and Corollary 2 give the lower bound

$$1 - 2\mathcal{N}_{\arccos(2m^{-1/4})}(G)\, e^{-2\sqrt{m}}$$

on probability that a uniformly randomly chosen normalized function f on X has G-variation and l_1-norm of output-weight vector of any network with units from G

computing f are at least $m^{1/4}$. If a dictionary G is "relatively small" (in the sense that its covering number $\mathcal{N}_{\arccos(2m^{-1/4})}(G)$ do not outweigh the factor $e^{-2\sqrt{m}}$, then almost any uniformly randomly chosen normalized function on X has G-variation larger than $m^{1/4}$. Covering numbers are called *power-type* if there exists $c > 0$ and a positive integer s such that $\mathcal{N}_\varepsilon(G) \leq \left(\frac{c}{\varepsilon}\right)^s$. So our results apply to dictionaries with power-type covering numbers. In particular, for $X = \{0, 1\}^d$ the d-dimensional Boolean cube and a power-type dictionary, our estimates imply for almost any uniformly randomly chosen function in $S_1(\{0, 1\}^d)$ the lower bound $2^{d/4}$ on the l_1-norms of all output-weight vectors of networks computing such function.

6 Sizes of Dictionaries of Computational Units

Our analysis of approximate measures of sparsity of networks with units from a dictionary G shows that when covering numbers of G grow only polynomially with the size of the domain X, then for almost any uniformly randomly chosen function on a sufficiently large X, l_1-norms of output-weight vectors of all networks with units from G computing f must be large. Such networks have large numbers of units or some of their output weights must be large. Both are not desirable.

Some estimates of covering numbers of dictionaries are known for shallow networks, where dictionaries are formed by functions computable by basic computational units. Much more complicated dictionaries formed by compositions of functions which are used in deep networks are less understood.

For finite dictionaries G, all covering numbers are bounded from above by card G. For some values of ε, covering numbers of finite dictionaries can even be smaller than their sizes. This can happen when a dictionary is highly coherent. Finite dictionaries are either formed by functions with finite ranges or functions with finite sets of parameters. Examples of dictionaries formed by binary-valued functions are dictionaries of Heaviside or signum perceptrons. Estimates of their sizes follow from the upper bound $2\frac{m^d}{d!}$ on the number of linearly separable dichotomies of m points in $\mathbb{R}^d$ proven by Schläfli already in the 19th century (see [54]).

The dictionary of signum perceptrons

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

occupies a relatively small subset of the set $\mathcal{B}(X)$ of all functions on X with values in $\{-1, 1\}$. The following upper bound is a direct consequence of an upper bound on the number of linearly separable dichotomies of m points in $\mathbb{R}^d$ from [13] combined with a well-known estimate of partial sum of binomials (see [40]).

Theorem 8 *For every d and every $X \subset \mathbb{R}^d$ with* card $X = m$,

$$\mathrm{card}\, P_d(X) \leq 2\frac{m^d}{d!}.$$

Thus the size of the dictionary $P_d(X)$ of signum perceptrons grows with the size of the domain $X \subset \mathbb{R}^d$ with card$X = m$ only polynomially with the polynomial degree equal to d, while the size 2^m of the set $\mathcal{B}(X)$ of all functions from X to $\{-1, 1\}$ grows with m exponentially. Estimate of the size of the dictionary of signum perceptrons combined with Theorem 6 gives the following bounds.

Theorem 9 *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* card $X = m$*,* P *be a uniform probability measure on $\mathcal{B}(X)$, f uniformly randomly chosen in $\mathcal{B}(X)$, and $b > 0$. Then*

$$\mathrm{P}\left(\|f\|_{P_d(X)} \geq b\right) \geq 1 - 4\frac{m^d}{d!}e^{-\frac{m}{2b^2}}.$$

Thus for large domains X, almost any uniformly randomly chosen function from X to $\{-1, 1\}$ has large variation with respect to signum perceptrons and so it cannot be l_1-sparsely represented by a shallow network with signum perceptrons. In particular, for card $X = 2^d$ and $b = 2^{\frac{d}{4}}$, Theorem 9 implies the following corollary.

Corollary 3 *Let d be a positive integer, $X \subset \mathbb{R}^d$ with* card $X = m$*,* P *be a uniform probability measure on $\mathcal{B}(X)$, and f uniformly randomly chosen in $\mathcal{B}(X)$. Then*

$$\mathrm{P}\left(\|f\|_{P_d(X)} \geq 2^{\frac{d}{4}}\right) \geq 1 - 4\frac{2^{d^2}}{d!}e^{-(2^{\frac{d}{2}}-1)}.$$

Covering numbers of the whole set $S_1(X)$ of all normalized functions on a finite set X are growing exponentially with card X. This follows from estimates of the *quasiorthogonal dimension* $\dim_\varepsilon m$ of $\mathbb{R}^m$. It is defined for $\varepsilon \in [0, 1]$ as the maximal number of unit vectors such that each pair of distinct ones has inner product at most ε, i.e.,

$$\dim_\varepsilon m = \max\{\mathrm{card}\, U \subset S^{m-1} \mid (\forall u, v \in U, u \neq v)(|u \cdot v| \leq \varepsilon)\}.$$

Quasiorthogonal dimension can be expressed as the packing number of S^{m-1}. It was proven in [25] that for a fixed $\varepsilon > 0$, the quasiorthogonal dimension $\dim_\varepsilon m$ grows exponentially with the dimension m as

$$\lceil e^{\frac{m\varepsilon^2}{2}} \rceil \leq \dim_\varepsilon m$$

(for arguments based on graph theory see [24]).

Let $\lambda(t) = 0$ for $t \leq 0$ and $\lambda(t) = t$ for $t \geq 0$ and let $L(X)$ denote the dictionary

$$L(X) := \{\lambda(e \cdot . + b) : X \to \mathbb{R} \mid e \in S^{d-1}, b \in \mathbb{R}\}.$$

The dictionary $L(X)$ has infinite range, but has the same size equal to the number of characteristic functions of half-spaces of $X \subset \mathbb{R}^d$.

Some dictionaries with infinite ranges have finite sets of parameters and thus they are finite. For X finite, let

$$K(X) := \{K_y \mid y \in X\},$$

where $K : X \times X \to \mathbb{R}$ is a symmetric positive definite kernel and $K_y(x)=K(x, y)$. Such dictionaries are used in SVM and their sizes are equal to card Y [2, 51].

Covering numbers of dictionaries $G_\phi(X, Y)$, for which the function $T_\phi : Y \to \mathcal{F}(X)$ defined as $T_\phi(y)(x) = \phi(x, y)$ is Lipschitz can be derived from covering numbers of the set of parameters Y.

Covering numbers in the angular pseudometrics are related to covering numbers in l_2. Indeed, for $f, g \in S_1(X)$, with $\rho(f, g) = \alpha$, we have $\|f - g\|_2 = 2\sin(\alpha/2)$. Various estimates of covering numbers in l_2-norm are known. For example, any subset G of the set of functions on a finite domain X with range $\{0, 1\}$ which has a finite VC-dimension has power-type covering numbers in l_2 [19]. It was shown in [48] that for any Lipschitz continuous sigmoidal function, $\mathcal{L}^2$-covering numbers of the dictionary of sigmoidal perceptrons on any bounded domain $\Omega \subset \mathbb{R}^d$ grow as $\left(\frac{1}{\varepsilon}\right)^\beta$, where $\beta > 0$.

7 Constructions of Functions with Large Variations

By Theorem 4, functions which are nearly orthogonal to all elements of a dictionary G have large G-variations. To construct an example of a class of functions with large variation with respect to signum signum perceptrons, we consider functions on square domains of the form

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

Such functions can be represented by square matrices. For a function f on $X = \{x_1, \dots, x_n\} \times \{y_1, \dots, y_n\}$ we denote by $M(f)$ the $n \times n$ matrix defined as

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

An $n \times n$ matrix M induces a function f_M on X such that

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

The inner product of two functions f and g on a square domain $X = \{x_1, \dots, x_n\} \times \{y_1, \dots, y_n\}$ is equal to the sum of entries of the matrices $M(f)$ and $M(g)$, i.e.,

$$\langle f, g \rangle = \sum_{i,j}^{n} M(f)_{i,j} M(g)_{i,j}.$$

Thus it is invariant under permutations of rows and columns performed jointly on both matrices $M(f)$ and $M(g)$. So to estimate inner products of functions represented by matrices we can reorder rows and columns whenever it is convenient.

An advantage of square domains is that on such domains matrices $M(g)$ representing signum perceptrons $g \in P_d(X)$ can be reordered in such a way that each row and each column of the reordered matrix starts with a segment of -1's followed by a segment of $+1$'s as stated in the next lemma from [38].

Lemma 1 *Let* $d = d_1 + d_2$, $\{x_i \mid i = 1, \dots, n\} \subset \mathbb{R}^{d_1}$, $\{y_j \mid j = 1, \dots, n\} \subset \mathbb{R}^{d_2}$, *and* $X = \{x_1, \dots, x_n\} \times \{y_1, \dots, y_n\} \subset \mathbb{R}^d$. *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.*

Proof Choose an expression of $g \in P_d(X)$ as $g(z) = \mathrm{sign}(a \cdot z + b)$, where $z = (x, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $a \in \mathbb{R}^d = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, and $b \in \mathbb{R}$. Let a_l and a_r denote the left and the right part, resp, of a, i.e., $a_{li} = a_i$ for $i = 1, \dots, d_1$ and $a_{ri} = a_{d_1+i}$ for $i = 1, \dots, d_2$. Then $\mathrm{sign}(a \cdot z + b) = \mathrm{sign}(a_l \cdot x + a_r \cdot y + b)$. Let ρ and κ be permutations of the set $\{1, \dots, n\}$ such that $a_l \cdot x_{\rho(1)} \le a_l \cdot x_{\rho(2)} \le \dots \le a_l \cdot x_{\rho(n)}$ and $a_r \cdot y_{\kappa(1)} \le a_r \cdot y_{\kappa(2)} \le \dots \le a_r \cdot y_{\kappa(n)}$.

Denote by $M(g)^*$ the matrix obtained from $M(g)$ by permuting its rows and columns by ρ and κ, resp. It follows from the definition of the permutations ρ and κ that each row and each column of $M(g)^*$ starts with a (possibly empty) initial segment of -1's followed by a (possibly empty) segment of $+1$'s. □

The reordering assembling -1's and $+1$'s in the matrix representing a signum perceptron (guaranteed by Lemma 1) reduces estimation of their inner products with functions $f : X \to \{-1, 1\}$ to estimation of differences of -1's and $+1$'s in submatrices of $M(f)$.

A class of matrices whose submatrices have relatively small differences of -1's and $+1$'s is the class of Hadamard matrices. A *Hadamard matrix* of order n is an $n \times n$ square matrix M with entries in $\{-1, 1\}$ such that any two distinct rows (or equivalently columns) of M are orthogonal. It follows directly from the definition that this property is invariant under permutations of rows and columns and sign flips of all elements in a row or a column. Note that Hadamard matrices were introduced as extremal ones among all $n \times n$ matrices with entries in $\{-1, 1\}$ as they have the largest determinants equal to $\sqrt{n}$. The well-known Lindsay Lemma bounds from above differences of $+1$'s and -1's in submatrices of Hadamard matrices (see, e.g., [17, p. 88]).

Lemma 2 (Lindsay) *Let* n *be a positive integer and* M *be an* $n \times n$ *Hadamard matrix. Then for any subset* I *of the set of indices of rows and any subset* J *of the set of indices of columns of* M,

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

Constructing a partition of a matrix induced by a signum perceptron into submatrices, which have all entries either equal to $+1$ or all entries equal to -1, and applying the Lindsay Lemma to a corresponding partition of a Hadamard matrix, one obtains the following lower bound on variation with respect to signum perceptrons for functions induced by Hadamard matrices (for details of the proof see [38]).

Theorem 10 *Let $d = d_1 + d_2$, $\{x_i \mid i = 1, \dots, n\} \subset \mathbb{R}^{d_1}$, $\{y_j \mid j = 1, \dots, n\} \subset \mathbb{R}^{d_2}$, $X = \{x_i \mid i = 1, \dots, m\} \times \{y_j \mid j = 1, \dots, m\} \subset \mathbb{R}^d$, and $f_M : X \to \{-1, 1\}$ be defined as $f_M(x_i, y_j) = M_{i,j}$, where M is an $n \times n$ Hadamard matrix. Then*

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

Theorem 10 combined with Proposition 4 implies the following corollary.

Corollary 4 *Let $d = d_1 + d_2$, $\{x_i \mid i = 1, \dots, n\} \subset \mathbb{R}^{d_1}$, $\{y_j \mid j = 1, \dots, n\} \subset \mathbb{R}^{d_2}$, $X = \{x_i \mid i = 1, \dots, n\} \times \{y_j \mid j = 1, \dots, n\} \subset \mathbb{R}^d$, and $f_M : X \to \{-1, 1\}$ be defined as $f_M(x_i, y_j) = M_{i,j}$, where M is an $n \times n$ Hadamard matrix. Then f_M cannot be computed by a shallow signum perceptron network having both the number of units and absolute values of all output weights depending on $\log_2 n$ polynomially.*

Corollary 4 shows that functions induced by Hadamard matrices cannot be computed by shallow signum or Heaviside perceptrons with numbers of units and sizes of output weights considerably smaller than sizes of their domains. Numbers of units and sizes of output weights in these networks cannot be bounded by polynomials of $\log_2$ of the sizes of their domains. Theorem 10 can be applied to domains containing sufficiently large squares, for example domains representing pictures formed by two-dimensional squares with $2^k \times 2^k$ pixels or digitized d-dimensional cubes.

Corollary 5 *Let k be a positive integer and $f_M : \{0, 1\}^k \times \{0, 1\}^k \to \{-1, 1\}$ be defined as $f_M(x_i, y_j) = M_{i,j}$, where M is a $2^k \times 2^k$ Hadamard matrix. Then*

$$\|f_M\|_{P_d(\{0,1\}^{2k})} \geq \frac{2^{k/2}}{k}.$$

Functions generated by $2^k \times 2^k$ Hadamard matrices cannot be computed by shallow signum perceptron networks with the sum of the absolute values of output weights bounded by a polynomial of k. This implies that the numbers of units and absolute values of all output weights in these networks cannot be bounded by any polynomial of k. Similarly, functions defined on $2k$-dimensional discretized cubes of sizes $s^{2k} = s^k \times s^k$ cannot be computed by networks with numbers of signum perceptrons and output weights smaller than

$$\frac{s^{k/2}}{\lceil k \log_2 s \rceil}. \tag{6}$$

8 Examples

An example of a class of functions with variation with respect to Gaussian kernel units with centers in the Boolean cube $\{0, 1\}^d$ increasing with d exponentially is the class of d-dimensional parities. Let

$$G_{K,a} = G_{K,a}(\{0, 1\}^d) := \{e^{-a\|\cdot - y\|^2} \mid y \in \{0, 1\|^d\}$$

denotes the *dictionary of Gaussian kernel units with centers in* $\{0, 1\}^d$ and $p_d : \{0, 1\}^d \to \{-1, 1\}$, where

$$p_d(v) := -1^{v \cdot u},$$

for all $u = (1, \dots, 1) \in \{0, 1\}^d$ is the *parity function*. The following lower bound from [38] shows that $G_{K,a}$-variation of p_d grows with d exponentially.

Theorem 11 *For every positive integer d and every $a > 0$,*

$$\|p_d\|_{G_{K,a}(\{0,1\}^d)} > 2^{d/2}.$$

Proof By Theorem 4,

$$\|p_d\|_{G_{K,a}} \geq \frac{\|p^d\|}{\sup_{g \in G_{K,a}(\{0,1\}^d)} |\langle p^d, g\rangle|}.$$

Let g_0 be the Gaussian centered at $(0, \dots, 0)$, then $\langle p^d, g_0\rangle = \sum_{k=0}^{d} (-1)^k \binom{d}{k} e^{-ak}$. By the binomial formula,

$$\langle p^d, g_0\rangle = \sum_{k=0}^{d} (-1)^k \binom{d}{k} e^{-ak} = (1 - e^{-a})^d.$$

Using a suitable transformation of the coordinate system, we obtain the same value of the inner product with p^d for the Gaussian g_x centered at any $x \in \{0, 1\}^d$ such that $p_d(x) = 1$. When the Gaussian g_x is centered at x with $p_d(x) = -1$, we get the same absolute value of the inner product by replacing p_d with $-p_d$ and by a transformation of the coordinate system. As $\|p_d\| = 2^{d/2}$, we get $\|p_d\|_{G_{K,a}} \geq \frac{2^{d/2}}{(1-e^{-a})^d} > 2^{d/2}$. □

Applying Corollary 4 to a variety of types of Hadamard matrices one obtains many examples of functions which cannot be computed by shallow perceptron networks with numbers of units and sizes of output weights bounded by

$$p(\log_2 \operatorname{card} X),$$

where p is a polynomial and X is the domain of the function.

Recall that if a Hadamard matrix of order $n > 2$ exists, then n is divisible by 4 (see, e.g., [45, p. 44]). It is conjectured that there exists a Hadamard matrix of every order divisible by 4. Various constructions of Hadamard matrices are known, such as Sylvester's recursive construction of $2^k \times 2^k$ matrices, Paley's construction based on quadratic residues, as well as constructions based on Latin Squares, and on Steiner triples.

Two Hadamard matrices are called equivalent when one can be obtained from the second one by permutations of rows and columns and sign flips of all entries in a row or a column. Listings of known constructions of Hadamard matrices and enumeration of non-equivalent Hadamard matrices of some orders can be found in [56].

The oldest construction of a class of $2^k \times 2^k$ matrices with orthogonal rows and columns was discovered by Sylvester [58]. A $2^k \times 2^k$ matrix is called *Sylvester-Hadamard* and denoted $S(k)$ if it is constructed recursively starting from the matrix

$$S(2) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and iterating the Kronecker product

$$S(l+1) = S(2) \otimes S(l) = \begin{vmatrix} S(l) & S(l) \\ S(l) & -S(l) \end{vmatrix}$$

for $l = 1, \ldots, k-1$. Corollary 5 implies that functions generated by $2^k \times 2^k$ Sylvester-Hadamard matrices cannot be represented by shallow signum perceptron networks with numbers of units and sizes of output weights smaller than $\frac{2^{k/2}}{k}$.

The following theorem from [38] shows that model complexities of signum or Heaviside perceptron networks computing functions generated by Sylvester-Hadamard matrices can be considerably decreased when two hidden layers are used instead of merely one hidden layer.

Theorem 12 *Let $S(k)$ be a $2^k \times 2^k$ Sylvester-Hadamard matrix, $h_k : \{0,1\}^k \times \{0,1\}^k \to \{-1,1\}$ be defined as $h_k(u,v) = S(k)_{u,v}$. Then h_k can be represented by a network with one linear output and two hidden layers with k Heaviside perceptrons in each hidden layer.*

An interesting class of functions with large variations with respect to perceptrons can be obtained by applying Theorem 10 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. They have been used in acoustics [55].

An infinite sequence $a_0, a_1, \dots, a_i, \dots$ of elements of $\{0, 1\}$ is called kth *order linear recurring sequence* if for some $h_0, \dots, h_k \in \{0, 1\}$

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

for all $i \geq k$. It is called *k-th order pseudo-noise (PN) sequence* (or *pseudo-random sequence*) if it is kth 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 kth 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 [45, p. 410].

An important property of PN-sequences is their low autocorrelation. The *autocorrelation* of a sequence $a_0, a_1, \dots, a_i, \dots$ of elements of $\{0, 1\}$ with period $2^k - 1$ is defined as

$$\kappa(t) = \frac{1}{2^k - 1} \sum_{j=0}^{2^k-1} -1^{a_j + a_{j+t}}. \tag{7}$$

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

$$\kappa(t) = -\frac{1}{2^k - 1} \tag{8}$$

[45, 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 Eqs. (7) and (8) shows that for any PN-sequence α the matrix $L_k(\alpha)$ has orthogonal rows.

Proposition 6 *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 10 to the $2^k \times 2^k$ matrices $L_k(\alpha)$ induced by a kth 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.

9 Discussion

Although current hardware allows to implement networks with large numbers of parameters, reducing network complexity is highly desirable as it can considerably improve efficiency of computation. Various studies show that also brain has sparse connectivity (each neuron is connected to only a limited number of other neurons). To obtains some theoretical understanding to limitations of shallow architectures, we investigated lower bounds on complexity of shallow networks.

As minimization of "l_0-pseudonorm" (which measures the number of hidden units in a shallow network) is a difficult non convex problem, we focused on approximate measures of network sparsity. We presented several arguments for using l_1-norm of output weight vectors as an approximate measure of network sparsity: Balls in l_1-norm are good approximations of convexifications of intersections of "balls" in "l_0" with unit balls in the ambient Euclidean metric, in contrast to l_2-norm, acting as a stabilizer l_1 penalizes even large number of output weights, l_1 has been used in weight-decay regularization [18], in statistical learning in the Lasso method [59], and is related to variational norm tailored to dictionary of computational units.

Applying geometrical properties of high-dimensional Euclidean spaces (the concentration of measure) we derived probabilistic lower bounds on minima of variational and l_1-norms of output-weight vectors in terms of covering numbers of dictionaries. As for many types of dictionaries used in shallow networks, covering numbers are power-type, the bounds imply that almost any uniformly randomly chosen normalized function on a large domain is highly uncorrelated with all elements of such dictionaries. Covering numbers of dictionaries formed by compositions of computational units characterizing deep networks are much less understood, but it is likely that they have much larger covering numbers than simple dictionaries used in shallow networks.

Although probabilistic results prove that there are many functions with large variations, it is not easy to find concrete constructions of such functions. 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" [12]. 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. The results presented in this chapter indicate that computation of functions exhibiting some randomness properties by shallow perceptron networks is difficult in the sense that it requires networks of large complexities. Some of such functions can be constructed using deterministic algorithms and have many useful applications. For example, 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" [55].

It should be emphasized that Theorem 7 and Corollary 2 assume uniform probability distribution of functions to be computed. The assumption of uniform distribution of computational tasks (sometimes implicit) is quite common. For example, in the No Free Lunch Theorem [61], it is assumed that all functions are equally likely. However in real tasks, relevance of functions for a give application area are far from being uniform. Recently, we derived some estimates of complexity of networks computing randomly chosen functions from nonuniform probability distributions [33].

Acknowledgements This work was partially supported by the Czech Grant Foundation grant GA18-23827S and institutional support of the Institute of Computer Science RVO 67985807.

References

1. Albertini, F., Sontag, E.: For neural networks, function determines form. Neural Netw. **6**(7), 975–990 (1993)
2. Anguita, D., Ghio, A., Oneto, L., Ridella, S.: Selecting the hypothesis space for improving the generalization ability of support vector machines. In: IEEE International Joint Conference on Neural Networks (2011)
3. Azuma, K.: Weighted sums of certain dependent random variables. Tohoku Math. J. **19**, 357–367 (1967)

4. Ba, L.J., Caruana, R.: Do deep networks really need to be deep? In: Ghahramani, Z. et al. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 1–9 (2014)
5. Ball, K.: An elementary introduction to modern convex geometry. In: Levy, S. (ed.) Flavors of Geometry, pp. 1–58. Cambridge University Press, Cambridge (1997)
6. Barron, A.R.: Neural net approximation. In: Narendra, K.S. (ed.) Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems, pp. 69–72. Yale University Press (1992)
7. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory **39**, 930–945 (1993)
8. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
9. Bengio, Y., LeCun, Y.: Scaling learning algorithms towards AI. In: Bottou, L., Chapelle, O., DeCoste, D., Weston, J. (eds.) Large-Scale Kernel Machines. MIT Press, Cambridge (2007)
10. Bengio, Y.: Learning deep architectures for AI. Found. Trends Mach. Learn. **2**, 1–127 (2009)
11. Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Stat. **23**, 493–507 (1952)
12. Coffrey, J.T., Goodman, R.Y.: Any code of which we cannot think is good. IEEE Trans. Inf. Theory **25**(6), 1453–1461 (1990)
13. Cover, T.: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electron. Comput. **14**, 326–334 (1965)
14. DeVore, R.A., Howard, R., Micchelli, C.: Optimal nonlinear approximation. Manuscr. Math. **63**, 469–478 (1989)
15. Donoho, D.: For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Commun. Pure Appl. Math. **59**, 797–829 (2006)
16. Donoho, D.L., Tsaig, Y.: Fast solution of 1-norm minimization problems when the solution may be sparse. IEEE Trans. Inf. Theory **54**, 4789–4812 (2008)
17. Erdös, P., Spencer, H.: Probabilistic Methods in Combinatorics. Academic, Cambridge (1974)
18. Fine, T.L.: Feedforward Neural Network Methodology. Springer, Berlin (1999)
19. Haussler, D.: Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. J. Comb. Theory A **69**(2), 217–232 (1995)
20. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**, 1527–1554 (2006)
21. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963)
22. Ito, Y.: Finite mapping by neural networks and truth functions. Math. Sci. **17**, 69–77 (1992)
23. Kainen, P.C., Kůrková, V., Sanguineti, M.: Dependence of computational models on input dimension: tractability of approximation and optimization tasks. IEEE Trans. Inf. Theory **58**, (2012)
24. Kainen, P.C., Kůrková, V.: Quasiorthogonal dimension. In: Kosheleva, O., Shary, S., Xiang, G., Zapatrin, R. (eds.) Beyond Traditional Probabilistic Data Processing Techniques: Interval, Fuzzy, etc. Methods and Their Applications. Springer, Berlin (2020, to appear)
25. Kainen, P.C., Kůrková, V.: Quasiorthogonal dimension of Euclidean spaces. Appl. Math. Lett. **6**(3), 7–10 (1993)
26. Kainen, P., Kůrková, V.: Functionally equivalent feedforward neural network. Neural Comput. **6**(3), 543–558 (1994)
27. Kainen, P., Kůrková, V.: Singularities of finite scaling functions. Appl. Math. Lett. **9**(2), 33–37 (1996)
28. Kainen, P.C., Kůrková, V., Vogt, A.: Approximation by neural networks is not continuous. Neurocomputing **29**, 47–56 (1999)
29. Kainen, P.C., Kůrková, V., Vogt, A.: Geometry and topology of continuous best and near best approximations. J. Approx. Theory **105**, 252–262 (2000)
30. Kainen, P.C., Kůrková, V., Vogt, A.: Continuity of approximation by neural networks in L_p-spaces. Ann. Oper. Res. **101**, 143–147 (2001)
31. Kecman, V.: Learning and Soft Computing. MIT Press, Cambridge (2001)
32. Kolmogorov, A.: Asymptotic characteristics of some completely bounded metric spaces. Dokl. Akad. Nauk. SSSR **108**, 585–589 (1956)

33. Kůrková, V., Sanguineti, M.: Classification by sparse neural networks. IEEE Trans. Neural Netw. Learn. Syst. **30**(9), 2746–2754 (2019)
34. 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)
35. Kůrková, V.: High-dimensional approximation and optimization by neural networks. In: Suykens, J. et al. (eds.) Advances in Learning Theory: Methods, Models, and Applications (NATO Science Series III: Computer & Systems Sciences, vol. 190), pp. 69–88. IOS Press, Amsterdam (2003)
36. Kůrková, V.: Sparsity and complexity of networks computing highly-varying functions. In: International Conference on Artificial Neural Networks, pp. 534–543 (2018)
37. Kůrková, V.: Complexity estimates based on integral transforms induced by computational units. Neural Netw. **33**, 160–167 (2012)
38. Kůrková, V.: Constructive lower bounds on model complexity of shallow perceptron networks. Neural Comput. Appl. **29**, 305–315 (2018)
39. Kůrková, V., Kainen, P.C.: Comparing fixed and variable-width Gaussian networks. Neural Netw. **57**(10), 23–28 (2014)
40. Kůrková, V., Sanguineti, M.: Model complexities of shallow networks representing highly varying functions. Neurocomputing **171**, 598–604 (2016)
41. LeCun, Y., et al.: Handwritten digit recognition with a back-propagation network. In: Proceedings of Advances in Neural Information Processing Systems, pp. 396–404 (1990)
42. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**, 2278–2324 (1998)
43. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
44. Lévy, P.: Problèmes concrets d'analyse fonctionelle. Gauthier Villards, Paris (1951)
45. MacWilliams, F., Sloane, N.A.: The Theory of Error-Correcting Codes. North Holland Publishing Co., Amsterdam (1977)
46. Maiorov, V.E., Meir, R.: On the near optimality of the stochastic approximation of smooth functions by neural networks. Adv. Comput. Math. **13**, 79–103 (2000)
47. Maiorov, V.E., Pinkus, A.: Lower bounds for approximation by MLP neural networks. Neurocomputing **25**, 81–91 (1999)
48. Makovoz, Y.: Random approximants and neural networks. J. Approx. Theory **85**, 98–109 (1996)
49. Matoušek, J.: Lectures on Discrete Geometry. Springer, New York (2002)
50. Mhaskar, H.N., Liao, Q., Poggio, T.: Learning functions: when is deep better than shallow. Center for Brains, Minds & Machines, pp. 1–12 (2016)
51. Oneto, L., Ridella, S., Anguita, D.: Tikhonov, Ivanov and Morozov regularization for support vector machine learning. Mach. Learn. **103**(1), 103–136 (2015)
52. Plan, Y., Vershynin, R.: One-bit compressed sensing by linear programming. Commun. Pure Appl. Math. **66**, 1275–1297 (2013)
53. Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., Liao, Q.: Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. Int. J. Autom. Comput. **14** (5), 503–519 (2017). https://doi.org/10.1007/s11633-017-1054-2
54. Schläfli, L.: Gesamelte Mathematische Abhandlungen, vol. 1. Birkhäuser, Basel (1950)
55. Schröder, M.: Number Theory in Science and Communication. Springer, New York (2009)
56. Sloane, N.A.: A library of Hadamard matrices. http://www.research.att.com/~njas/hadamard/
57. Sussman, H.J.: Uniqueness of the weights for minimal feedforward nets with a given input-output map. Neural Netw. **5**(4), 589–593 (1992)
58. Sylvester, J.J.: Thoughts on inverse orthogonal matrices, simultaneous sign successions, and tessellated pavements in two or more colours, with applications to Newton's rule, ornamental tile-work, and the theory of numbers. Philos. Mag. **34**, 461–475 (1867)
59. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)
60. Tillmann, A.: On the computational intractability of exact and approximate dictionary learning. IEEE Signal Process. Lett. **22**, 45–49 (2015)
61. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(67), (1997)