Soft Computing (2021) 25:12641–12648
https://doi.org/10.1007/s00500-021-05938-4

FOCUS

Correlations of random classifiers on large data sets

Věra Kůrková[1] · Marcello Sanguineti[2]

Accepted: 3 June 2021 / Published online: 15 June 2021

Abstract
Classification of large data sets by feedforward neural networks is investigated. To deal with unmanageably large sets of classification tasks, a probabilistic model of their relevance is considered. Optimization of networks computing randomly chosen classifiers is studied in terms of correlations of classifiers with network input–output functions. Effects of increasing sizes of sets of data to be classified are analyzed using geometrical properties of high-dimensional spaces. Their consequences on concentrations of values of sufficiently smooth functions of random variables around their mean values are applied. It is shown that the critical factor for suitability of a class of networks for computing randomly chosen classifiers is the maximum of sizes of the mean values of their correlations with network input–output functions. To include cases in which function values are not independent, the method of bounded differences is exploited.

Keywords Random classifiers · Optimization of feedforward networks · Binary classification · Concentration of measure · Method of bounded differences

1 Introduction

In practical applications, feedforward networks compute functions on finite domains (formed, e.g., by pixels of pictures or scattered vectors of features), which are typically large and high-dimensional. Computational difficulties of multidimensional tasks, called the *curse of dimensionality* (Bellman 1957), have long been well known. In particular, the numbers of parameters needed for computation of certain types of tasks grow exponentially with increasing dimension. However, it was observed that proper choices of a type of network computational units and/or network architecture can considerably reduce this number, sometimes even from an exponential dependence to merely a linear one (see, e.g., Barron 1993; Gnecco and Sanguineti 2008, 2011; Kainen et al. 2003, 2012; Mhaskar 2004 and the references therein). Experimental evidence was in some cases confirmed by mathematical arguments, e.g., Gaussian SVM vs. shallow perceptron networks (see, e.g., Bengio et al. 2006; Kůrková and Sanguineti 2016).

Even for domains of moderate sizes, sets of all classification and uncorrelated regression tasks are unmanageably large—they grow exponentially with increasing sizes of domains. However in real applications, most of these functions are not likely to represent any task of interest. In Kůrková and Sanguineti (2017), we introduced a proba-

V.K. was partially supported by the Czech Science Foundation grant GA 19-05704S and by institutional support of the Institute of Computer Science RVO 67985807. M.S. was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824160 and by the Project PDGP 2018/20 DIT.AD016.001 "Technologies for Smart Communities" of INM (Institute for Marine Engineering) of CNR (National Research Council of Italy), where he is Research Associate. M.S. is a member of GNAMPA-INdAM (Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni—Istituto Nazionale di Alta Matematica), Affiliated Researcher at IIT—Italian Institute of Technology (Advanced Robotics Research Line), Genova, and Visiting Professor at IMT—School for Advances Studies (AXES Research Unit), Lucca.

✉ Věra Kůrková
vera@cs.cas.cz

Marcello Sanguineti
marcello.sanguineti@unige.it

[1] Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic

[2] Department of Computer Science, Bioengineering, Robotics, and Systems Engineering (DIBRIS), University of Genoa, Genoa, Italy

bilistic approach to model prior knowledge that a function represents a tasks which can occur in a given application area. In Kůrková and Sanguineti (2019), we derived probabilistic estimates of model complexities of shallow networks computing functions randomly chosen from uniform and product probabilities.

In solving high-dimensional problems, the curse of dimensionality is complemented by the *blessing of dimensionality*, which includes concentration of values of random variables around their mean values and possibilities of reduction of data dimensionality by random projections (see, e.g., Dubhashi and Panconesi 2009; Gallicchio and Scardapane 2020; Gorban et al. 2016, 2019; Matoušek 2002). Most concentration inequalities hold for sums of independent random variables. Such estimates are applicable to cases satisfying the *naive Bayes assumption*, e.g., when the presence of a particular feature or symptom in a class is unrelated to the presence of any other feature. Results based on this assumption work quite well for data with many equally important features. The naive Bayes assumption has been successfully used in text categorization since 1960s. However in many other real problems, the independence of labels assigned to feature vectors cannot be guaranteed.

We investigate capabilities of efficient approximation of randomly chosen functions by neural networks. We explore approximation errors measured in the l_2-norm in terms of correlations of random classifiers with input–output mappings implementable by classes of feedforward networks. We propose a probabilistic model of relevance of computational tasks including distributions that may not satisfy the naive Bayes assumption (i.e., when classes are not assigned to network inputs independently). Allowing violation of the independence hypothesis strongly limits possibilities of theoretical analysis of such tasks. Indeed, most mathematical tools based on the concentration of measure phenomenon assume independence of random variables or even uniform probability. Concentration-type results holding without the assumption of independence of variables are rare and hold only for functions which do not "vary too much." We explore possibilities of applying concentration inequalities based on Azuma–Hoeffding theorems to correlations of randomly chosen classifiers with network input–output functions. We apply the *method of bounded differences* to cases when correlations satisfy a coordinate-wise Lipschitz condition with sufficiently small parameters. To deal with more general probabilities, we apply an average Lipschitz condition. We exploit these conditions to estimate correlations of input–output functions with random classifiers chosen according probability distributions, where conditional mean values do not "vary too much."

We show that on large domains, correlations of randomly chosen classifiers with any fixed binary-valued function are concentrated around their mean value. Thus, approximation errors of randomly chosen functions according to suitable distributions behave almost deterministically. Suitability of a given class of networks for approximation of relevant functions described by a probability distribution depends on the maximum of the mean values of correlations of random functions with input–output functions from the given class.

The paper is organized as follows. In Sect. 2, we introduce notations and basic concepts on approximation of functions on finite domains by feedforward networks with linear outputs. In Sect. 3, a probabilistic model of relevance of functions for a given application domain is presented. In Sect. 4, correlations of randomly chosen classifiers with input–output functions are studied using the method of bounded differences. Section 5 extends the estimates to cases which do not satisfy the independence condition. Section 6 addresses consequences of the probabilistic estimates for feedforward networks. Section 7 is a brief discussion.

2 Approximation of functions on finite domains by input–output functions

Let

$$U := \{x_1, \dots, x_m\} \subset \mathbb{R}^d$$

be a finite set. We denote by

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

the *set of all real-valued functions on U*. $\mathcal{F}(U)$ is isometric with the m-dimensional Euclidean space $\mathbb{R}^m$ and thus functions on U can be seen as m-dimensional vectors. $\mathcal{F}(U)$ inherits from $\mathbb{R}^m$ the inner product

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

and the associated Euclidean norm

$$\|f\|_2 := \sqrt{\langle f, f \rangle}.$$

We denote by

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

the *set of all functions on U with values in* $\{-1, 1\}$. This set, which can be seen as the Hamming cube $\{-1, 1\}^m$, models the set of all binary classifiers on U. Instead of using the Hamming distance, we measure errors in the l_2-norm inherited from $\mathbb{R}^m$. For $r > 0$, we denote by

$$S_r(U) := \{f : U \to \mathbb{R} \mid \|f\|_2 = r\}$$

the sphere of radius r in $\mathcal{F}(U)$.

We focus on approximation of functions from subsets of $\mathcal{F}(U)$ by functions computable by *feedforward networks with linear outputs*. Such networks compute functions of the form

$$\mathrm{span}_n \{\mathcal{G}(., v) \mid v \in V\} := \left\{ \sum_{j=1}^{n} w_i \mathcal{G}(., v_j) \mid v_j \in V \right\},$$

where n is the number of units in the last hidden layer, $w_1, \dots, w_n$ are output weights, V is a set of inner network parameters, and $\mathcal{G} : U \times V$ is a function of two vector variables $u \in U$ and $v \in V$. The parameterized family of functions

$$\mathcal{G} = \mathcal{G}_V(U) := \{\mathcal{G}(., v) : U \to \mathbb{R} \mid v \in V\}$$

depends on the network architecture and the types of its computational units. For a shallow network with one hidden layer, G is a dictionary of computational units. For a network with more hidden layers, it is formed by compositions of hidden units from subsequent hidden layers.

Sets of the form $\mathrm{span}_n \mathcal{G}$ are invariant under multiplication by scalars, i.e., $c\, \mathrm{span}_n \mathcal{G} = \mathrm{span}_n \mathcal{G}$ for all $c \in \mathbb{R}$. As for any norm $\|.\|$ and $c > 0$ one has

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

examples of functions with arbitrarily large or small errors measured by metrics induced by any norm $\|.\|$ in approximation by sets of the form $\mathrm{span}_n \mathcal{G}$ can be obtained by multiplication by suitable constants c. So estimates of errors in approximation of functions by networks with a linear output only have sense when functions to be approximated and approximating functions have the same norms. Sometimes, it is convenient to consider normalized functions. In the case of binary classification, we consider functions with values in $\{-1, 1\}$ instead of $\{0, 1\}$. Such functions have the same norms equal to $\sqrt{m}$, where m is the size of the domain.

Errors measured by the l_2-norm of functions approximated by functions of the same l_2-norm can be studied in terms of their *correlations*, expressed as inner products. For all $f, g \in \mathcal{F}(U)$,

$$\|f - g\|_2^2 = \|f\|_2^2 + \|g\|_2^2 - \langle f, g\rangle$$

and in particular for normalized functions $f = f^\circ$ and $g = g^\circ$,

$$\|f - g\|_2^2 = 2 - \langle f, g\rangle.$$

In the sequel, we take into account that correlations of functions on large domains are influenced by geometrical properties of high-dimensional spaces.

3 Probabilistic model of relevance of computational tasks

We consider a finite set $U := \{x_1, \dots, x_m\} \subset \mathbb{R}^d$, whose elements represent data to be classified (such as feature vectors, colors of linearly ordered pixels of photographs or scattered vectors of medical symptoms). Often, data sets to be processed are large.

The numbers of binary and multiclass classifiers on U grow with card $U = m$ exponentially. Also, the numbers of uncorrelated real-valued functions (nearly orthogonal) on the domain U grow with its size exponentially. These numbers were studied in Kainen and Kůrková (1993), Kainen and Kůrková (2020) in terms of the *quasiorthogonal dimension* $\dim_\varepsilon m$, defined as the maximal number of unit vectors in $\mathbb{R}^m$ such that the absolute values of their inner products are at most ε. It was proven there that $\dim_\varepsilon m$ is bounded from below by $e^{m\varepsilon^2}$.

Nevertheless, in practical applications most functions from these unmanageably large sets are not likely to represent any task of interest. Thus, it is not necessary to search for networks capable of an efficient computation of all functions on a given finite domain. Although many classes of networks enjoy the *universal representation property* (i.e., they can exactly compute all of them, see, e.g., Ito 1992), for many tasks their use is limited by an increasing model complexity. Arguments proving the universal representation property assume that networks potentially have numbers of units equal to the sizes of finite domains of functions to be represented. A prior knowledge about probability that certain functions are irrelevant or that their relevance is small can considerably reduces the requirements on expressive power of classes of neural networks. In particular, it can help to select types of networks capable of computing or approximating relevant functions with much smaller numbers of units (see, e.g., Gnecco et al. 2011a, b; Kainen et al. 2012; Kůrková 2012).

We model relevance of computational tasks by a probability distribution on a subset $\mathcal{T}(U)$ of the set $\mathcal{F}(U)$ of all functions on U. In particular, we focus on the case where $\mathcal{T}(U)$ is formed by the set of all binary classifiers $\mathcal{B}(U)$. Let P be a probability measure on $\mathcal{T}(U)$. A function $f \in \mathcal{T}(U)$ randomly chosen with respect to P induces random variables

$$X_1 := f(x_1), \dots, X_m := f(x_m).$$

4 Concentration of values of inner products

To investigate correlations of a fixed function h with classifiers randomly chosen according to a given probability on U, we define a function Φ_h of m random variables as

$$\Phi_h(X_1,\dots,X_m) := \sum_{i=1}^{m} h(x_i)X_i\,. \tag{1}$$

When the random variables are generated by a function f on $\{x_1,\dots,x_m\}$, i.e., $X_1 = f(x_1),\dots,X_m = f(x_m)$, one has

$$\Phi_h(X_1,\dots,X_m) = \sum_{i=1}^{m} h(x_i)f(x_i) = \langle h, f\rangle.$$

Geometry of high-dimensional spaces implies that almost all values of sufficiently smooth functions of large numbers of random variables concentrate around their mean values. One of such smoothness conditions is a version of Lipschitz property considered with respect to coordinates. We call a function

$$\Phi : A_1 \times \dots \times A_m \to \mathbb{R}$$

Coordinate-Wise Lipschitz (CWL) with parameters $c_1,\dots,c_m$ if for all $i = 1,\dots,m$ and all vectors $a, a' \in A_1 \times \dots \times A_m$, which differ just in the i-th coordinate,

$$|\Phi(a) - \Phi(a')| \le c_i\,.$$

Recall that the *Hamming distance* of two vectors $a, a' \in \{-1, 1\}$ is defined as the number of entries in which they differ. So for $A_i = \{-1, 1\}$, $i = 1,\dots,m$, the CWL condition implies Lipschitz continuity on $\{-1, 1\}^m$ with parameter $\bar{c} = \max_{i=1,\dots,m} c_i$ with respect to the Hamming distance.

The following theorem provides a concentration of measure inequality for functions of independent random variables satisfying the CWL condition with parameters $c_1,\dots,c_m$, such that the l_2-norm $\|c\|_2$ of the parameter vector $c = (c_1,\dots,c_m)$ is sufficiently small.

Theorem 1 (Dubhashi and Panconesi 2009, p.70) *Let $X_1,\dots,X_m$ be independent random variables with values in ranges $A_1,\dots,A_m$, resp., and $\Phi : A_1 \times \dots \times A_m \to \mathbb{R}$ a function satisfying the CWL condition with parameters $c_1,\dots,c_m$. Then, for every $t > 0$ one has*

$$\mathrm{P}\Big[\,|\Phi - E(\Phi)| > t\Big] \le e^{-2t^2/\gamma}\,, \tag{2}$$

where $\gamma := \sum_{i=1}^{m} c_i^2$.

The bound (2) does not explicitly reflect the role of the number m of random variables. To get some insight into its role, we set $t := \lambda\, m$. Then, we obtain the bound

$$\mathrm{P}\left[\frac{|\Phi - E(\Phi)|}{m} > \lambda\right] \le e^{-2\lambda^2 m^2/\gamma}\,. \tag{3}$$

The bound 3 implies

$$\mathrm{P}\left[\frac{|\Phi - E(\Phi)|}{m} \le \lambda\right] > 1 - e^{-2\lambda^2 m^2/\gamma} \tag{4}$$

which shows that when m is large and γ does not outweigh m^2, then almost all values of $\frac{\Phi}{m}$ are concentrated around $\frac{E(\Phi)}{m}$. Thus, Theorem 1 can be applied to functions of random variables satisfying the CWL condition with the parameter vector $c = (c_1,\dots,c_m)$ such that $\|c\|_2^2$ grows with m at a subquadratic rate.

Applying Theorem 1 to inner products of a fixed real-valued function with randomly chosen binary-valued functions, we obtain the following upper bound.

Theorem 2 *Let P be a product probability distribution on the set $\mathcal{B}(U)$ of all binary classifiers on $U \subset \mathbb{R}^d$, and h be a real-valued function on U. Then, for every $\lambda > 0$ and f randomly chosen according to P, the following inequalities hold*

(i)

$$\mathrm{P}\Big[\Big|\langle f^\circ, h^\circ\rangle - E\langle f^\circ, \frac{h}{\|h\|_2}\rangle\Big| > \lambda\Big] \le e^{-\frac{m\lambda^2}{2}}\,;$$

(ii) when $\|h\|_2 = \sqrt{m}$,

$$\mathrm{P}\Big[\Big|\langle f^\circ, h^\circ\rangle - E\langle f^\circ, h^\circ\rangle\Big| > \lambda\Big] \le e^{-\frac{m\lambda^2}{2}}\,.$$

Proof For the function Φ_h defined in (1), we have $|\Phi_h(X_1,\dots,X_{i-1}, 1, X_{i+1},\dots,X_m) - \Phi_h(X_1,\dots,X_{i-1}, -1, X_{i+1},\dots,X_m)| = 2|h(x_i)|$ for all $i = 1,\dots,m$. Thus, Φ_h satisfies the CWL condition with coefficients $c_i = 2|h(x_i)|$. As P is a product probability, $X_1,\dots,X_m$ are independent and so the statement follows by Theorem 1 with $t = \|h\|_2\sqrt{m}$ and $\gamma = \|c\|_2^2 = 4\|h\|_2^2$. □

Theorem 2 shows that on large domains, correlations of a fixed function (in particular an input–output function of a feedforward network) with binary classifiers randomly chosen according to a product distribution behave almost deterministically, in the sense that most correlations concentrate around their mean value. Thus, if there exists a network input–output function h having relatively large mean value of inner products with randomly chosen binary classifiers, then with a high probability a random classifier can be well approximated by h.

5 Concentration of correlations in some dependent cases

The theorems in the previous section assume that the random variables $X_1 = f(x_1), \ldots, X_m = f(x_m)$ are independent, which means that the probability distribution P can be expressed as a product of probability distributions of values of X_i, $i = 1, \ldots, m$. In real applications, often this assumption cannot be guaranteed. But independence of random variables is an essential part of proofs of most concentration inequalities (see, e.g. Dubhashi and Panconesi 2009). Without this assumption, only few concentration inequalities hold, where a weakening of the independence condition is compensated by strengthening smoothness properties of the function of random variables. Such concentration inequalities, which follow from the *Azuma–Hoeffding Inequality* and theory of martingales (see, e.g., Azuma 1967; Chung and Lui 2005; Dubhashi and Panconesi 2009, Theorem 5.1, p. 62; Dubhashi and Panconesi 2009, p. 58) are called *Bounded Difference Conditions* (Dubhashi and Panconesi 2009).

Here, we use an extension of the coordinate-wise Lipschitz (CWL) condition employed in the previous section. To define it, we first introduce some notation. For random variables X, Y, by $E(X|Y = y)$ is denoted the *conditional expectation of X on occurrence of value y of the random variable Y*. We use boldface to abbreviate sequences of random variables, sequences of real numbers, and parts of these sequences formed by the first i elements, i.e., $\mathbf{X} := (X_1, \ldots, X_m)$, $\mathbf{X_i} := (X_1, \ldots, X_i)$, $\mathbf{a} := (a_1, \ldots, a_m)$, and $\mathbf{a_i} := (a_1, \ldots, a_i)$.

Let $X_1, \ldots, X_m$ be random variables with values in the sets $A_1, \ldots, A_m$, respectively. A function $\Phi : A_1 \times \cdots \times A_m \to \mathbb{R}$ satisfies the *Averaged Lipschitz (AL) condition with parameters $c_1, \ldots, c_m$ with respect to a sequence $\mathbf{X} := (X_1, \ldots, X_m)$ of random variables* if the conditional expectation satisfies for $a_1, \bar{a}_1 \in A_1$

$$\left| E(\Phi \mid X_1 = a_1) - E(\Phi \mid X_1 = \bar{a}_1) \right| \leq c_1 \tag{5}$$

and for all $i = 2, \ldots, m$ and $a_i, \bar{a}_i \in A_i$

$$\Big| E(\Phi \mid \mathbf{X_{i-1}} = \mathbf{a_{i-1}}, X_i = a_i) - E(\Phi \mid \mathbf{X_{i-1}} = \mathbf{a_{i-1}}, X_i = \bar{a}_i) \Big| \leq c_i \tag{6}$$

(Dubhashi and Panconesi 2009, p. 68, equation (5.7)). Intuitively, the AL condition requires that the difference between two partial averages of Φ obtained by assigning some fixed values to the first $i-1$ random variables, two different values to the i-th variable, and setting randomly (according to the given distribution) the values of the remaining variables, is bounded by c_i.

For functions satisfying the AL condition, the following extension of Theorem 1 holds without the assumption of independence of random variables.

Theorem 3 (Dubhashi and Panconesi 2009, p. 68) *Let $X_1, \ldots, X_m$ be a sequence of random variables with values in sets $A_1, \ldots, A_m$, resp., and $\Phi : A_1 \times \ldots \times A_m \to \mathbb{R}$ be a function satisfying the AL condition with parameters $c_1, \ldots, c_m$ with respect to the random variables $X_1, \ldots, X_m$. Then, for every $t > 0$ one has*

$$\mathrm{P}\left[\left| \Phi - E(\Phi) \right| > t \right] \leq e^{-2t^2/\gamma}, \tag{7}$$

where $\gamma := \sum_{i=1}^m c_i^2$.

To show the effect of increasing the number m of random variables, we set $t := m\lambda$. Then, we get from (7) the bound

$$\mathrm{P}\left[\left| \frac{\Phi - E(\Phi)}{m} \right| > \lambda \right] \leq e^{-2\lambda^2 m^2/\gamma}. \tag{8}$$

Similarly as in Theorem 1, also in Theorem 3 the size of the parameter γ is critical for usefulness of the bound (8).

Theorem 3 does not assume that the random variables $X_1, \ldots, X_m$ are independent and so it can be used in more realistic scenarios than Theorem 1. To apply it to the function Φ_h representing the inner product with a fixed binary-valued function h (in particular, with an input–output function of a neural network), we first define a condition on random variables. A sequence $X_1, \ldots, X_m$, $m > 2$, of random variables with values in $\{-1, 1\}$ is called *conditionally dependent (CD) with parameters $b_1, \ldots, b_{m-2}$* if for every sequence $a_1, \ldots, a_m \in \{-1, 1\}$ and every $j = 2, \ldots, m$

$$\left| E(X_j | X_1 = 1) - E(X_j | X_1 = -1) \right| \leq \frac{b_1}{m-2} \tag{9}$$

and every $i = 2, \ldots, m-2$ and every $j = i+1, \ldots, m$,

$$\begin{aligned} \Big| E(X_j | \mathbf{X_{i-1}} = \mathbf{a_{i-1}}, X_i = 1) \\ - E(X_j | \mathbf{X_{i-1}} = \mathbf{a_{i-1}}, X_i = -1) \Big| \\ \leq \frac{b_i}{m-i-1}. \end{aligned} \tag{10}$$

Note that when $X_1, \ldots, X_m$ are independent, the mean values $E(X_j | \mathbf{X_{i-1}} = \mathbf{a_{i-1}}, X_i = 1)$ and $E(X_j | \mathbf{X_{i-1}} = \mathbf{a_{i-1}}, X_i = -1)$ are equal and thus the inequalities (9) and (10) hold with $b_i = 0$ for all $i = 1, \ldots, m$. The following proposition shows that any set of random variables with values in $\{-1, 1\}$ is conditionally dependent, but the parameters might be rather large.

Proposition 1 *Every sequence $X_1, \ldots, X_m$ of random variables with values in $\{-1, 1\}$ is conditionally dependent with parameters $b_i = 2(m - i - 1)$, $i = 1, \ldots, m-2$.*

Proof By the definition of the CD condition, to prove the statement it is sufficient to verify that for $j = i+1, \ldots, m$,

$$\begin{aligned} & E(X_j \,|\, \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, a_i = 1) \\ & \quad - E(X_j \,|\, \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, a_i = -1\,) \leq 2. \end{aligned}$$

By the definition of conditional probability, we have

$$\begin{aligned} & E(X_j \,|\, \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, a_i = 1) \\ & = \frac{\mathrm{P}(X_j = 1 \& \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = 1) \; - \; \mathrm{P}(X_j = -1 \& \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = 1}{P(\mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = 1)} \end{aligned}$$

and

$$\begin{aligned} & E(X_j \,|\, \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, a_i = -1) \\ & = \frac{\mathrm{P}(X_j{=}1 \& \mathbf{X}_{i-1}{=}\mathbf{a}_{i-1}, X_i{=}1) \; - \; \mathrm{P}(X_j{=}-1 \& \mathbf{X}_{i-1}{=}\mathbf{a}_{i-1}, X_i = -1)}{P(\mathbf{X}_{i-1}{=}\mathbf{a}_{i-1}, X_i = -1)}. \end{aligned}$$

Set $t_{1,1} = \mathrm{P}(X_j = 1 \& \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = 1)$,

$$\begin{aligned} t_{-1,1} &= \mathrm{P}(X_j = -1 \& \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = 1), \\ t_{1,-1} &= \mathrm{P}(X_j = 1 \& \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = -1) \\ t_{-1,-1} &= \mathrm{P}(X_j = -1 \& \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = -1). \end{aligned}$$

As $t_{1,1}, t_{-1,1}, t_{1,-1}, t_{-1,-1} \in [0, 1]$ and $t_{1,1} + t_{-1,1} + t_{1,-1} + t_{-1,-1} \leq 1$, we obtain

$$\frac{t_{1,1} - t_{-1,1}}{t_{1,1} + t_{-1,1}} - \frac{t_{1,-1} - t_{-1,-1}}{t_{1,-1} + t_{-1,-1}} \leq 2.$$

□

The next theorem shows that the property of conditional dependence is useful in cases when it holds with sufficiently small parameters. It gives a probabilistic estimate of the deviation of inner products from their mean value which depends on $b := \max_{i=1,\ldots,m-2} b_i$.

Theorem 4 *Let $U := \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ be finite, $h \in \mathcal{B}(U)$, and P a probability distribution on $\mathcal{B}(U)$ such that for f randomly chosen according to P, the random variables $X_1 := f(x_1), \ldots, X_m := f(x_m)$ are conditionally dependent with parameters $b_1, \ldots, b_{m-2}$. Then, for every $\lambda > 0$ and $b := \max_{i=1,\ldots,m-2} b_i$, the following inequalities hold*

$$(i)\ \mathrm{P}\Big[|\langle f, h\rangle - E\langle f, h\rangle| > m\lambda\Big] \leq e^{-\frac{2m\lambda^2}{(2+b)^2}};$$

$$(ii)\ \mathrm{P}\Big[|\langle f^\circ, h^\circ\rangle - E\langle f^\circ, h^\circ\rangle| > \lambda\Big] \leq e^{-\frac{2m\lambda^2}{(2+b)^2}}.$$

Proof To apply Theorem 3 to the function $\Phi_h(X_1, \ldots, X_m) := \sum_{i=1}^m h(x_i) X_i = \langle f, h\rangle$, we estimate parameters of the conditional dependence of random variables $X_1, \ldots, X_m$. For every $i \in \{1, \ldots, m\}$ and every fixed sequence $a_1, \ldots, a_m$ with $a_i \in \{-1, 1\}$ and for all $j = 1, \ldots, i-1$, we have

$$\begin{aligned} & |E(\Phi_h \,|\mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = 1) - E(\Phi_h \,|\, \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, X_i = -1)| \\ & \; = |\sum_{j=1}^{i-1} h(x_j) a_j + h(x_i) \\ & \quad + \sum_{j=i+1}^{m} h(x_j) E(X_j \,|\, \mathbf{X}_i = \mathbf{a}_i, a_i = 1\,) \\ & \quad - \sum_{j=1}^{i-1} h(x_j) a_j {+} h(x_i) {-} \sum_{j=i+1}^{m} h(x_j) E(X_j \,|\, \mathbf{X}_{i-1}{=}\mathbf{a}_{i-1}, a_i = -1\,)\,| \\ & \; = \Big| 2h(x_i) + \sum_{j=i+1}^{m} h(x_j) \Big(E(X_j \,|\, \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, a_i = 1) \\ & \quad - E(X_j \,|\, \mathbf{X}_{i-1} = \mathbf{a}_{i-1}, a_i = -1\,) \Big) \Big|. \end{aligned} \tag{11}$$

If $i \leq m-2$, then (11) is bounded from above by $2 + (m - i - 1)\frac{b_i}{m-i-1} = 2 + b_i$. As $|h(x_i)| = 1$, for $i = m-1$, (11) is bounded by 2, and for $i = m$, it is bounded by 1. Thus Φ_h satisfies the AL condition with $\gamma = \sum_{i=1}^{m-2} (2{+}b_i)^2 {+} 2 {+} 1 \leq m(2+b)^2$ and so the statement follows from the bound (8), which is implied by Theorem 3. □

Theorem 4 shows that inner products of a fixed binary-valued function h with randomly chosen classifiers f for which the random variables $X_1 = f(x_1), \ldots, X_m = f(x_m)$ are conditionally dependent with sufficiently small parameters are concentrated around their mean value. The larger the domain U and the smaller the parameters, the sharper the concentration. For example, for $b = 2$, Theorem 4 gives the bound

$$\mathrm{P}\Big[|\langle f^\circ, h^\circ\rangle - E(\langle f^\circ, h^\circ\rangle)| > \lambda\Big] \leq e^{-\frac{m\lambda^2}{8}}. \tag{12}$$

Setting $\lambda = m^{-1/4}$, we get from the bound (12) an exponentially decreasing upper bound $e^{-\frac{m^{1/2}}{8}}$ on the probability that the inner product of a normalized randomly chosen function does not deviate from the mean value by more than $m^{-1/4}$.

Note that random variables satisfy the CD condition with all parameters $b_i = 2$ provided

$$\begin{aligned} & \Big| E(X_j | \mathbf{X_{i-1}} = \mathbf{a_{i-1}}, X_i = 1) \\ & \qquad - E(X_j | \mathbf{X_{i-1}} = \mathbf{a_{i-1}}, X_i = -1) \Big| \\ & \quad \leq \frac{2}{m - i - 1}. \end{aligned} \tag{13}$$

This holds for sequences of random variables, for which with increasing i fixing either $X_i = 1$ or $X_i = -1$ has increasing influence on the expectation. For such sequences, the smallest difference of expectations is for fixing either $X_1 = 1$ or $X_1 = -1$. This can happen when random variables are ordered from the least significant one to the most significant one.

By Proposition 1, for every distribution P on $\mathcal{B}(X)$, the function Φ_h is conditionally dependent with parameters $b_i = 2(m-i-1)$. Then, $b = \max_{i=1,\dots,m-2} b_i = b_1 = 2(m-2)$ and the upper bound in Theorem 4 is

$$e^{-\frac{2m\lambda^2}{(2+b)^2}} = e^{-\frac{m\lambda^2}{2(m-1)^2}}.$$

Thus, for a general distribution, the upper bound following from Theorem 4 does not guarantee concentration of values of inner products.

6 Optimal approximation of random classifiers by feedforward networks

As for normalized functions, approximation errors in the l_2-norm decrease with increasing correlations, our results show that a critical factor for suitability of a class of networks for computing tasks modeled by a probability P is the maximum of the mean values of inner products with network input–output functions. For a class of feedforward networks with linear outputs computing input–output functions from the set $\mathrm{span}_n\, \mathcal{G}_V(X)$, where

$$\mathcal{G}_V(U) := \{\mathcal{G}(., v) : U \to \mathbb{R} \mid v \in V\},$$

and for a probability P on $\mathcal{B}(X)$, we define

$$M_P(n, V) := \max_{h \in \mathrm{span}_n \mathcal{G}_V(X)} \{E\langle f, h\rangle\}.$$

By Theorems 1 and 3, concentration of correlations holds for probability distributions with respect to which the function Φ_h is sufficiently "smooth." In such cases, the correlations behave, in a qualitative way, "almost deterministically."

The critical factor for suitability of a class of networks for computing tasks whose relevance is characterized by the probability P is the size of $M_P(n, V)$. The worst case is when P is the uniform distribution, i.e., when there is no prior knowledge available. Then, due to the symmetry, the mean values of inner products of any fixed function h with uniformly randomly chosen functions from $\mathcal{B}(X)$ are equal to zero. Thus, almost all randomly chosen functions are orthogonal to h. A class of networks can be suitable for computation of uniformly randomly chosen classifiers only when its set of input–output functions is "large enough" so that small neighborhoods of its elements of radii $e^{-\frac{m\lambda^2}{2}}$ in angular distance cover the set of all binary classifiers $\mathcal{B}(X)$.

Also for some nonuniform distributions, $M_P(n, V)$ can be small. In such cases, almost any randomly chosen function is nearly orthogonal to all input–output functions computable by a given class of networks and so the situation is similar to the case of uniform probability. Such networks are not suitable for tasks characterized by P unless they generate exponentially many input–output functions covering with their small neighborhoods the set $\mathcal{B}(X)$. The size and covering capability of the set $\mathrm{span}_n \mathcal{G}_V(U)$ can be increased by increasing n and/or V.

On the other hand, if for some input–output function h, the mean value of inner products is large (and hence also $M_P(n, V)$ is large), then almost all randomly chosen functions can be quite well approximated by that input–output function. Networks computing functions with large mean values of inner products with random classifiers are optimal for tasks characterized by the probability distribution P.

If the maximum $M_P(n, V)$ of the mean values of inner products has medium size, then some improvement can be achieved if one extends the set of input–output functions by increasing the number n of units in the last hidden layer, or by extending the set V of network parameters, or by choosing different types of units.

7 Discussion

To deal with unmanageably large numbers of classification tasks on data sets of even moderate sizes, we investigated a probabilistic model describing relevance of tasks. We explored optimization of feedforward networks with linear outputs for classification tasks in terms of correlations between randomly chosen functions according to a given probability and network input–output functions. As correlations are related to approximation errors measured in the l_2-norm, their distributions provide some insights into suitability of various classes of networks for computing random classifiers selected according to a given probability distribution. To include also cases where assignments of class labels to feature vectors are not independent, we exploited extensions of the Azuma–Hoeffding inequalities which hold without the naive Bayes assumption. By applying the method of averaged bounded differences to distributions with respect to which inner products with network input–output functions are sufficiently "smooth," we showed that in some cases correlations of randomly chosen function concentrate around their mean values. Thus, on large domains, errors in approximation of normalized random functions behave almost deterministically. This is rather surprising and provides qualitative insights into capabilities of feedforward networks to approximate large classes of functions. We showed that the critical factor for optimization of various classes of feedforward networks for a given type of classification tasks is the maximum of the mean values of the inner products of input–output functions with randomly chosen classifiers.

When a prior knowledge is limited and a class of tasks is described by an almost-uniform distribution, optimal sets of input–output functions must be large enough to cover the set of all tasks by small neighborhoods of the input–output functions. Some understanding of network capabilities can be obtained by investigating covering numbers of dictionaries of computational units and sets of input–output functions. Another important characterization of these sets is their coherence (Tropp 2004), defined as the maximum of absolute values of inner products of pairs of distinct input–output functions and thus it is related to correlation. Coherence measures how much two input–output functions are similar. Although coherence only reflects extreme correlations, it is easy to calculate and captures in a significant way the behavior of sets of input–output functions and computational units.

Funding This study was funded by the Czech Science Foundation grant GA 19-05704S and by institutional support of the Institute of Computer Science RVO 67985807 (V. Kůrková) and by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824160 (M. Sanguineti). M. Sanguineti is Research Associate at the Institute for Marine Engineering of National Research Council of Italy under project PDGP 2018/20 DIT.AD016.001, Affiliated Researcher at IIT—Italian Institute of Technology (Advanced Robotics Research Line), Genova, and Visiting Professor at IMT—School for Advances Studies (AXES Research Unit), Lucca.

Declarations

Conflict of Interest Author Věra Kůrková declares that she has no conflict of interest. Author Marcello Sanguineti declares that he has no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Azuma K (1967) Weighted sums of certain dependent random variables. Tohoku Math J 19:357–367

Barron AR (1993) Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans Inf Theory 39:930–945

Bellman R (1957) Dynamic Programming. Princeton University Press, Princeton

Bengio Y, Delalleau O, Roux NL (2006) The curse of highly variable functions for local kernel machines. Adv Neural Inf Process Syst 18:107–114

Chung F, Lui L (2005) Concentration inequalities and martingale inequalities: a survey. Internet Math 3:79–127

Dubhashi D, Panconesi A (2009) Concentration of measure for the analysis of randomized algorithms. Cambridge University Press, Cambridge

Gallicchio C, Scardapane S (2020) Deep randomized neural networks. In: Oneto L, Navarin N, Sperduti A, Anguita D (eds) Recent trends in learning from data. Studies in computational intelligence, vol 896. Springer, Switzeland, pp 44–68

Gnecco G, Kůrková V, Sanguineti M (2011) Can dictionary-based computational models outperform the best linear ones? Neural Netw 24:881–887

Gnecco G, Kůrková V, Sanguineti M (2011) Some comparisons of complexity in dictionary-based and linear computational models. Neural Netw 24:171–182

Gnecco G, Sanguineti M (2008) Approximation error bounds via Rademacher's complexity. Appl Math Sci 2(4):153–176

Gnecco G, Sanguineti M (2011) On a variational norm tailored to variable-basis approximation schemes. IEEE Trans Inf Theory 57:549–558

Gorban A, Tyukin I, Prokhorov D, Sofeikov K (2016) Approximation with random bases: Pro et contra. Inf Sci 364–365:129–145

Gorban AN, Makarov VA, Tyukin IY (2019) The unreasonable effectiveness of small neural ensembles in high-dimensional brain. Phys Life Rev 29:55–88

Ito Y (1992) Finite mapping by neural networks and truth functions. Math Sci 17:69–77

Kainen P, Kůrková V, Sanguineti M (2003) Minimization of error functionals over variable-basis functions. SIAM J Optim 14:732–742

Kainen PC, Kůrková V (1993) Quasiorthogonal dimension of Euclidean spaces. Appl Math Lett 6(3):7–10

Kainen PC, Kůrková V (2020) Quasiorthogonal dimension. In: Beyond traditional probabilistic data processing techniques: interval, fuzzy, etc. Methods and their applications. Studies in computational intelligence, vol 835, pp 615–629. Springer

Kainen PC, Kůrková V, Sanguineti M (2012) Dependence of computational models on input dimension: tractability of approximation and optimization tasks. IEEE Trans Inf Theory 58:1203–1214

Kůrková V (2012) Complexity estimates based on integral transforms induced by computational units. Neural Netw 33:160–167

Kůrková V, Sanguineti M (2016) Model complexities of shallow networks representing highly varying functions. Neurocomputing 171:598–604

Kůrková V, Sanguineti M (2017) Probabilistic lower bounds for approximation by shallow perceptron networks. Neural Netw 91:34–41

Kůrková V, Sanguineti M (2019) Classification by sparse neural networks. IEEE Trans Neural Netw Learning Syst 30(9):2746–2754

Matoušek J (2002) Lectures on discrete geometry. Springer, New York

Mhaskar HN (2004) On the tractability of multivariate integration and approximation by neural networks. J Complex 20:561–590

Tropp A (2004) Greed is good: algorithmic results for sparse approximation. IEEE Trans Inf Theory 50:2231–2242