Complexity of Neural Network Approximation with Limited Information: a Worst Case Approach

Mark Kon, *Boston University*
and
Leszek Plaskota, *Warsaw University*

June 8, 2000

Abstract

In neural network theory the complexity of constructing networks to approximate input-output (i-o) functions is of interest. We study this in the more general context of approximating elements f of a normed space F using partial information about f. We assume information about f and the size of the network are limited, as is typical in radial basis function networks. We show complexity can be essentially split into two independent parts, information ε-complexity and neural ε-complexity. We use a worst case setting, and integrate elements of information-based complexity and nonlinear approximation. We consider deterministic and/or randomized approximations using information possibly corrupted by noise. The results are illustrated by examples including approximation by piecewise polynomial neural networks.

1 Introduction

In Kon and Plaskota (2000), an information complexity theory for radial basis function (RBF) neural networks is studied. It is shown that two types of complexity, information complexity and neural complexity, interact in simple ways to determine the complexity of function approximation. Information complexity involves the amount of information about the unknown input-output (i-o) function f needed to approximate it to tolerance ϵ (by any approximation engine). Neural complexity involves the number of hardware processors (neurons) needed in a network for this approximation. We assume that (as for general feedforward networks) each processor P_j computes a single function $d_j(x)$, and that the network computes linear combinations of the $d_j(x)$. In standard RBF networks, the $d_i(x)$ are generally simple transformations of a single function, the reproducing kernel for a Hilbert space.

We connect the above theory to a formulation for more general classes of neural nets and function approximation paradigms. This connects information-based complexity theory and nonlinear approximation theory, yielding an approach to what might be called information-based nonlinear approximation.

Our problem at its most basic is to approximate an element f of a normed linear space F. The following two approaches to this problem seem most typical.

The first approach assumes available information about f is partial and/or noisy. If F is a space of multivariate functions, information about f might consist of sample values, which in addition can be corrupted by noise. Lack of complete information causes approximation error, since many functions generally share the same information. This is typical of problems in *information-based complexity* (IBC) and problems related to scientific computation, and numerical computation in particular. Computation of high dimensional integrals (e.g., in financial mathematics) is a primary example. The reader is referred to the monographs of Novak (1988), Plaskota (1996), Traub, Wasilkowski and Woźniakowski (1988), and Traub and Werschulz (1999).

In the second approach one assumes complete (unlimited) information about f, but is limited computationally to approximations in a special class or set (e.g., of available "computers"). If this set is a finite-dimensional linear subspace of F, we are in the domain of classical approximation theory. Recently, however, attention has been devoted to *nonlinear approximation* (NA), where the set of possible approximations is nonlinear. There, on the premise of limited computational resources, one seeks a k-term approximation of the form $\tilde{f} = \sum_{j=1}^{k} a_j d_j$, where d_j are selected from a *dictionary* $\mathcal{D} \subset F$ of "basis functions" optimized for functions in F. Here approximation error comes from restrictions on approximations; see, e.g., DeVore and Temlyakov (1995, 1997), DeVore (1998), and DeVore and Lorentz (1993).

Nonlinear approximation has been applied to signal and data compression; see, e.g., Bergeaud and Mallat (1996) for applications to so-called matching pursuit methodologies. Choices of $\mathcal{D}$ depend on whether the signal is a speech signal (with $\mathcal{D}$ a class of phoneme signals), a cardiac signal (a class of "standard" heartbeats plus variations), a stock trend (a class of wavelets appropriate to Brownian motion), or from a space of smooth functions (an appropriate class of RBF's). This approach is present in solutions of approximation-theoretic problems in feedforward perceptron models of neural network theory. See, e.g., Chui, Li, and Mhaskar (1996), Mhaskar and Micchelli (1995), or Pinkus (1999) for a survey of this topic.

The above two assumptions seem to have little in common, but there are situations where both of them are present. Consider a neural network for approximating an i-o function f (encoding a real-world phenomenon). The network depends on parameters chosen in a *learning process* based on collection of *examples* $Nf = (f(x_1), ..., f(x_n))$ of f. This involves collecting information and using it, and limited information is generic. On the other hand, with the assumption of limited neural resources, nonlinear approximation becomes central. Thus IBC and NA find common ground in neural network theory.

In this paper we study a combined model for the approximation problem, where information as well as allowed approximations are limited. The notion of information is adopted from IBC, and that of approximation from NA. The term *network* denotes a k-term approximation, which can be viewed as an (artificial) neural network with a single hidden layer containing k neurons. Examples are RBF and feedforward perceptron networks. Our goal is to construct a network approximating f with error at most ε. We seek the number $n = n(\varepsilon)$ of observations of f and the number $k = k(\varepsilon)$ of hidden neurons necessary and sufficient to perform this task. The analysis is done in the worst case setting.

When the space F is a Sobolev class, an optimal choice of dictionary $\mathcal{D}$ consists of translates of the reproducing kernel for F (Kon and Plaskota, 2000). Our goal here is to show that these RBF results extend to more general function classes.

We now introduce the two notions of complexity. *Information ε-complexity*, $\mathrm{IC}^{\mathrm{wor}}(\varepsilon)$, is the number of observations necessary and sufficient to construct an ε-approximation in the IBC model (limited information and unlimited approximations). *Neural ε-complexity*, $\mathrm{NC}^{\mathrm{wor}}(\varepsilon)$, is the number of neurons necessary and sufficient to obtain an ε-approximation in NA model (unlimited information and limited approximations). Both quantities have been studied, though the term 'neural complexity' is new as used here. Obviously, $\mathrm{IC}^{\mathrm{wor}}(\varepsilon)$ and $\mathrm{NC}^{\mathrm{wor}}(\varepsilon)$ provide lower bounds for the number of observations and neurons, respectively, in the combined model. It turns out, however, that they essentially provide upper bounds as well, in both deterministic and randomized approximation settings, and for information possibly contaminated by noise. Almost optimal approximations (networks) are essentially compositions of the best approximations from IBC and from NA. Thus (as shown for RBF's in Kon and Plaskota (2000)), complexity in the combined model can be essentially split into information complexity and neural complexity. Interestingly, the word 'essentially' above can sometimes be dropped (i.e., the lower bounds are sharp) as shown in an example of Section 3, where approximation in Hilbert spaces is analyzed. Thus the combined model is not only where IBC and NA meet, but also where they split.

Generally, randomized or Monte Carlo approximations are usually not much better for information complexity in the worst case setting than worst case approximations. We show a corresponding result for neural complexity and the above combined setting.

When only information is limited, optimal approximations often depend linearly on information, and these are actually n-term approximations. Then we have $\mathrm{NC}^{\mathrm{wor}}(\varepsilon) \leq \mathrm{IC}^{\mathrm{wor}}(\varepsilon)$ and the question is whether it is possible to use fewer neurons than observations for an ε-appoximation. This is illustrated in the problem of L_∞-approximation of functions $f : [0, 1] \to \mathbb{R}$ from the Hölder class $C_{r,\alpha}$ or Sobolev class $W_{r,p}$, where approximations are restricted to piecewise polynomials of degree s and are based on observations of f. Let, for instance, $s = r$ and information be exact. Then, in the Hölder class, we need $\Theta(\varepsilon^{-1/(r+\alpha)})$ observations and neurons, and (almost) optimal approximations use equidistant knots; hence the knots are independent of f. In the Sobolev class with $1 < p < \infty$ we need $\Theta(\varepsilon^{-1/(r+1-1/p)})$ observations, but we can reduce the number of neurons to $\Theta(\varepsilon^{1/(r+1)})$. The final approximation uses different knots for different f's.

The paper is organized as follows. In Section 2, we introduce essential notions of information ε-complexity and neural ε-complexity. In Section 3, we define the combined model and show basic facts about best approximations and complexity. We also give an important example of approximation in a Hilbert space. In Section 4, we analyze whether randomized approximations can be better than nonrandomized approximations. In Section 5, we briefly discuss noisy information. Our results are applied in Section 6 to piecewise polynomial approximation of Hölder and Sobolev classes of functions.

We use worst case machinery and deterministic or randomized approximations to answer our complexity questions. In a forthcoming paper we will study the corresponding questions in an average case setting.

2 The two notions of complexity

We first formally define the two crucial notions of ε-complexity. We use a rather general framework. We assume that F is an arbitrary normed space with norm $\|\cdot\|$, and we want to approximate elements $f \in F$. Typically, F is a space of multivariate functions $f : D \to \mathbb{R}$ where D is a subset of $\mathbb{R}^d$.

Denoting by $\tilde{f} = A(f)$ an approximation to $f \in F$, we define the error of $A : F \to F$ on a given class $\mathcal{F} \subset F$ as

$$e^{\mathrm{wor}}(A) \;=\; \sup_{f \in \mathcal{F}} \|f - A(f)\|.$$

2.1 Information ε-complexity

Suppose first that we do not have full knowledge of f. We can, however, collect some information about f by evaluating (or observing) values of some functionals at f. More specifically, the information is given as

$$Nf \;=\; (L_1 f, L_2 f, \ldots, L_n f), \tag{1}$$

where L_j are from a given class $\mathcal{L}$ of functionals. For instance, if F is a space of multivariate functions, then $\mathcal{L}$ may consist of function evaluations. If the L_j are selected in advance, the information is non-adaptive. We also formally allow adaptive information in which case the choice of L_j depends on previously obtained values (not on f itself!) $y_i = L_i f$, $1 \le i \le j-1$, so that $L_j = L_j(\cdot; y_1, \ldots, y_{j-1})$. We call $y = Nf$ *information about* f. The mapping $N : F \to Y$, where Y is the set of all possible values of information, will be called *information*. See, e.g., Traub, Wasilkowski, and Woźniakowski, 1988, for more detailed definitions and discussion.

Any approximation $A(f)$ is in this case a function of $y = Nf$ rather than f. Hence, the mapping $A : F \to F$ can be decomposed as $A(\cdot) = \varphi(N(\cdot))$ where $\varphi : Y \to F$. We write $A = (N, \varphi)$. *Radius of information* measures uncertainty in information and is defined as

$$\mathrm{rad}(N) \;=\; \inf_{\varphi : Y \to F} e^{\mathrm{wor}}(N, \varphi).$$

Equivalently,

$$\mathrm{rad}(N) \;=\; \sup_{y \in Y} \mathrm{rad}(\mathcal{F}_y),$$

where $\mathcal{F}_y = \{h \in \mathcal{F} : \; Nh = y\}$ and $\mathrm{rad}(\mathcal{F}_y)$ is the usual (Chebyshev) radius of $\mathcal{F}_y$.

The radius of information has been extensively studied in *information-based complexity*, see, e.g., Novak (1988), Traub, Wasilkowski, and Woźniakowski (1988). We recall that the error $\mathrm{rad}(N)$ is achieved if (but not only if) $\varphi(y)$ is the center of $\mathcal{F}_y$, provided the center exists. Furthermore, for any *interpolatory* approximation A, i.e., one for which $\varphi(y)$ is an arbitrary element from $\mathcal{F}_y$, we have

$$e^{\mathrm{wor}}(N, \varphi) \;\le\; 2 \cdot \mathrm{rad}(N). \tag{2}$$

Let $\mathcal{N}_n$ be the class of all information with at most n function evaluations for any f. Then

$$r_n^{\mathrm{wor}} \;=\; \inf_{N \in \mathcal{N}_n} \mathrm{rad}(N)$$

measures how much the uncertainty can be reduced using at most n observations.

The *information ε-complexity* is defined as the minimal number of observations from which it is possible to construct approximation with error ε for any $f \in \mathcal{F}$. That is,

$$\mathrm{IC}^{\mathrm{wor}}(\varepsilon) \;=\; \min\{\, n : \;\; \text{there exists } A = (N, \varphi) \text{ such that } N \in \mathcal{N}_n \text{ and } e^{\mathrm{wor}}(A) \le \varepsilon \,\}.$$

To stress the dependence on $\mathcal{F}$ and $\mathcal{L}$, we will sometimes write $\mathrm{IC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{L}; \varepsilon)$ instead of $\mathrm{IC}^{\mathrm{wor}}(\varepsilon)$.

Example 1 We recall one particular and well known result that will be used later. Let F be an infinite dimensional separable Hilbert space over $\mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\| \cdot \|$. Let $\{\xi_j\}_{j \ge 1}$ be a complete orthonormal system in F. We define the class $\mathcal{F}$ to be an ellipsoid,

$$\mathcal{F} \;=\; \Big\{ f \in F : \; \sum_{j=1}^{\infty} \langle \xi_j, f \rangle^2 / \gamma_j \le 1 \Big\}, \tag{3}$$

where $\gamma_1 \ge \gamma_2 \ge \gamma_3 \ge \cdots \ge 0$ is a fixed sequence. Suppose that the available information about f consists of observations of arbitrary functionals $L \in F^*$ at f. It is well known that then the best n observations are nonadaptive and given as

$$y_j \;=\; \langle \xi_j, f \rangle, \qquad 1 \le j \le n, \tag{4}$$

and the best approximation (assuming no restrictions) is

$$\tilde{f} \;=\; \varphi(y) \;=\; \sum_{j=1}^{n} y_j \xi_j, \tag{5}$$

which is also the center of the corresponding set $\mathcal{F}_y$. Moreover,

$$r_n^{\mathrm{wor}} \;=\; \sqrt{\gamma_{n+1}}.$$

For instance, if $\gamma_j \approx j^{-2p}$ with $p > 0$, then $r_n^{\mathrm{wor}} \approx n^{-p}$ and $\mathrm{IC}^{\mathrm{wor}}(\varepsilon) \approx \varepsilon^{-1/p}$.

2.2 Neural ε-complexity

Suppose now that we have full knowledge of f, but we restrict approximations $A(f)$ to be k-term approximations. That is, let $\mathcal{D} \subset F$ be a given *dictionary*. We are interested in approximations of the form

$$\tilde{f} \;=\; A(f) \;=\; \sum_{j=1}^{k} a_j d_j, \tag{6}$$

where $a_j \in \mathbb{R}$ and $d_j \in \mathcal{D}$.

Suppose for a moment that F is a space of multivariate functions. Then $\tilde{f}$ can be viewed as a neural network with a single hidden layer consisting of k neurons $d_j \in \mathcal{D}$. For instance, if the dictionary consists of radial basis functions then we deal with RBF networks. If $\mathcal{D}$ consists of functions $d = \sigma(\langle \vec{w}, \cdot \rangle_2 - z)$, where $\vec{w} \in \mathbb{R}^d$ and $z \in \mathbb{R}$ are arbitrary, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed *activation function*, then we deal with feedforward perceptron networks. We will use the term *network* for the approximation (6).

Let F_k be the set of all networks (6) consisting of at most k neurons (or the set of k-term approximations). The minimal error of such approximations on a class $\mathcal{F} \subset F$ is defined as

$$s_k^{\rm wor} \;=\; \sup_{f \in \mathcal{F}} \; \inf_{\tilde{f} \in F_k} \; \|f - \tilde{f}\|,$$

or, equivalently,

$$s_k^{\rm wor} \;=\; \inf_{A : F \to F_k} \; e^{\rm wor}(A).$$

The quantity $s_k^{\rm wor}$ has been studied in approximation theory for different spaces F and dictionaries $\mathcal{D}$, see, e.g., DeVore and Temlyakov (1995, 1997), Mhaskar (1996).

We now define

$${\rm NC}^{\rm wor}(\varepsilon) \;=\; \min\{\, k : \;\; \text{there exists } A : F \to F_k \;\; \text{such that} \;\; e^{\rm wor}(A) \le \varepsilon \,\}$$

to be the minimal k for which it is possible to construct a k-term approximation for any $f \in \mathcal{F}$. We will call ${\rm NC}^{\rm wor}(\varepsilon)$ the *neural ε-complexity.* To stress the dependence on $\mathcal{F}$ and $\mathcal{D}$ we will sometimes write ${\rm NC}^{\rm wor}(\mathcal{F}, \mathcal{D}; \varepsilon)$ instead of ${\rm NC}^{\rm wor}(\varepsilon)$.

Example 2 Consider the problem of approximating an f in the ellipsoid (3) of Example 1. We now assume unlimited information, but we want to produce a k-term approximation to f using the dictionary

$$\mathcal{D} \;=\; \{\, \xi_j \,:\, j \ge 1 \,\}, \tag{7}$$

where ξ_j's are as in Example 1. It can be easily seen that then the optimal k-term approximation is

$$\tilde{f} \;=\; \psi(f) \;=\; \sum_{j \in S} \langle \xi_j, f \rangle f,$$

where $S \subset \mathbb{N}$ is the set of k indices for which the coefficients $|\langle \xi_j, f \rangle|$ are largest possible.

Note that the worst case error of ψ is attained at an f for which $|\langle \xi_j, f\rangle|$ are all the same for $1 \le j \le k+1 \le m$, and $\langle \xi_j, f\rangle = 0$ for $j \ge m+1$, where m is an integer. Further considerations give the following formula. Let m^* be the largest integer in the set of all $m \ge k+1$ for which

$$m - 1 - k \;\le\; \gamma_m \cdot \sum_{j=1}^{m-1} \gamma_j^{-1}, \tag{8}$$

or $m^* = \infty$ if (8) holds for all such m. (Note that $m^* < \infty$ if $\lim_{j\to\infty} \gamma_j = 0$.) Then

$$s_k^{\mathrm{wor}} \;=\; \begin{cases} \sqrt{(m^*-k)/\left(\sum_{j=1}^{m^*} \gamma_j^{-1}\right)} & \text{if } \; m^* < \infty, \\ \lim_{m\to\infty} \sqrt{m/\left(\sum_{j=1}^{m} \gamma_j^{-1}\right)} & \text{if } \; m^* = \infty. \end{cases}$$

For instance, if $\gamma_j \approx j^{-2p}$, $p > 0$, then $m^* \approx (1 + (2p)^{-1})k$, $s_k^{\mathrm{wor}} \approx (1 + (2p)^{-1})^{-p}\, k^{-p}$, and

$$\mathrm{NC}^{\mathrm{wor}}(\varepsilon) \;\approx\; \left(\frac{2p}{2p+1}\right) \left(\frac{1}{\varepsilon}\right)^p.$$

3 Complexity in the combined model

Our purpose is to study the combined model. That is, we assume that both information and approximations are limited. For $f \in F$, we want to construct an approximation (network) $\tilde{f} = A(f)$ of the form (6) based on information $y = Nf$ about f of the form (1). Note that then A can be decomposed into an information part $N : F \to Y$ and an approximation part $\varphi : Y \to \cup_{k\ge 1} F_k$, so that $A(f) = \varphi(Nf)$. Our goal is to determine the number n of observations and the number k of neurons that are necessary and sufficient to construct an approximation with error $e^{\mathrm{wor}}(A) \le \varepsilon$.

We first characterize the minimal error of approximation.

Lemma 1 *Let N be any information (1). Then*

$$\max\{\mathrm{rad}(N),\, s_k^{\mathrm{wor}}\} \;\le\; \inf_{\varphi: Y \to F_k} e^{\mathrm{wor}}(N, \varphi) \;\le\; 2\cdot \mathrm{rad}(N) \,+\, s_k^{\mathrm{wor}}.$$

Hence, for the minimal error of approximations that use n observations and k neurons, we have

$$\max\{\, r_n^{\mathrm{wor}}, s_k^{\mathrm{wor}}\} \;\le\; \inf_{N \in \mathcal{N}_n}\, \inf_{\varphi: Y \to F_k} e^{\mathrm{wor}}(N, \varphi) \;\le\; 2\cdot r_n^{\mathrm{wor}} \,+\, s_k^{\mathrm{wor}}.$$

Proof. Since the lower bound is obvious, we show only the upper bound. To this end, we present an approximation $A^*(f) = \varphi^*(y)$, $y = Nf$, whose error is not bigger than the upper bound plus some $\eta > 0$. The approximation is given as

$$\varphi^*(y) \;=\; \psi(\varphi^I(y)),$$

where $\varphi^I : Y \to F$ is an interpolatory approximation using N, and $\psi : F \to F_k$ is a mapping with error $e^{\mathrm{wor}}(\psi) \le s_k^{\mathrm{wor}} + \eta$. Hence the network is constructed in two steps. We first choose the interpolatory approximation $g = \varphi^I(y)$, and then construct an almost optimal network for g consisting of at most k neurons.

Using the triangle inequality and (2), we now obtain for any $f \in \mathcal{F}$

$$\|f - \varphi^*(Nf)\| \le \|f - \varphi^I(Nf)\| + \|g - \psi(g)\| \le 2 \cdot \mathrm{rad}(N) + s_k^{\mathrm{wor}} + \eta.$$

Since η can be arbitrarily small, this gives the upper bound.

The second part of the lemma follows by taking infima with respect to information $N \in \mathcal{N}_n$. □

Note that the upper bound of Lemma 1 can be improved in many cases. Let $\tilde{\mathcal{F}}$ be the convex hull of $\mathcal{F}$. Suppose that for any information y we can find an element $a(y) \in \tilde{\mathcal{F}}$ such that $\sup_{f \in \mathcal{F}_y} \|f - a(y)\| \le \mathrm{rad}(\mathcal{F}_y) + \eta$. Then the minimal error of k-term approximation can be bounded by

$$\inf_{\varphi : Y \to F_k} e^{\mathrm{wor}}(N, \varphi) \le \mathrm{rad}(N) + s_k^{\mathrm{wor}},$$

i.e., we can get rid of the factor 2. Indeed, in this case we can apply the approximation as in the proof of Lemma 1 with the interpolatory part $\varphi^I(y)$ replaced by $a(y)$, and use the fact that the minimal errors s_k^{wor} of k-term approximations over $\mathcal{F}$ and over $\tilde{\mathcal{F}}$ are equal.

The following simple example shows that, in general, we cannot improve the bounds any further.

Example 3 Let $F = \mathbb{R}^2$ with the l_1-norm, and $\mathcal{F} = \{ f \in F : \|f\|_\infty \le 1 \}$. Let information $Nf = f_2$ (the second coordinate of f) and the dictionary $\mathcal{D} = \{(1, 0)\}$. Then the center of $\mathcal{F}_y$ is $(0, y)$ and $\mathrm{rad}(N) = 1$. We also have $s_0^{\mathrm{wor}} = 2$, but for $k \ge 1$ we have $s_k^{\mathrm{wor}} = 1$ and the best k-term approximation to $f = (f_1, f_2)$ is $\psi(f) = (f_1, 0)$. However, if only limited information $y = f_2$ is available about f, then the best k-term approximation is just zero, and its error is 2. Hence this error equals $\max\{\mathrm{rad}(N), s_k^{\mathrm{wor}}\}$ for $k = 0$, and $\mathrm{rad}(N) + s_k^{\mathrm{wor}}$ for $k \ge 1$.

As a consequence of Lemma 1, we obtain the following theorem.

Theorem 1 *In order to construct a network (6) in the class $\mathcal{F}$ with error at most ε, it is necessary to use at least $\mathrm{IC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{L}; \varepsilon)$ observations and $\mathrm{NC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{D}; \varepsilon)$ neurons, and it is sufficient to use $\mathrm{IC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{L}; \alpha\varepsilon)$ observations and $\mathrm{NC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{D}; \beta\varepsilon)$ neurons, where $\alpha, \beta > 0$ are arbitrary numbers satisfying $2\alpha + \beta \le 1$.*

Proof. Necessity follows immediately from the lower bound of Lemma 1. For sufficiency, let $n = \mathrm{IC}^{\mathrm{wor}}(\alpha\varepsilon)$ and $k = \mathrm{NC}^{\mathrm{wor}}(\beta\varepsilon)$. We can choose $N \in \mathcal{N}_n$ such that for some φ we have $e^{\mathrm{wor}}(N,\varphi) \le \alpha\varepsilon$, and $\psi : F \to F_k$ such that $e^{\mathrm{wor}}(\psi) \le \beta\varepsilon$. Proceeding as in the proof of the upper bound of Lemma 1 we obtain that the error of the approximation $A(f) = \psi(\varphi^I(Nf))$ is at most $2\alpha\varepsilon + \beta\varepsilon \le \varepsilon$. □

Thus the problem of constructing ε-networks can be essentially split into two separate and independent parts corresponding to information complexity and neural complexity. It is enough to know both complexities to determine complexity of constructing ε-networks.

For some concrete problems we are able to provide more precise analysis. Here is an important example.

Approximation in a Hilbert space

Consider the problem defined in Examples 1 and 2. Suppose first that we want to construct an approximation using n observations and k neurons. Then we can use the following strategy. We first observe the coefficients (4), i.e., $y_j = \langle \xi_j, f\rangle$ for $1 \le j \le n$. If $n \le k$ then the final approximation is given by (5), otherwise it is given by $\tilde{f} = A^*(f) = \sum_{j\in S} y_j\,\xi_j$, where S is the set of k indices $j \in \{1,2,\dots,n\}$ for which $|y_j|$ are largest possible.

Bounds for the error of A^* follow from Lemma 1. It turns out, however, that this error actually equals the lower bound, which means that A^* is optimal.

Theorem 2 *We have*

$$e^{\mathrm{wor}}(A^*) \;=\; \max\{\, r_n^{\mathrm{wor}}, s_k^{\mathrm{wor}}\},$$

i.e., A^* is an optimal approximation that uses n observations and k neurons.

Proof. Let f_n be the orthogonal projection onto $\mathrm{span}\{\,\xi_j : 1 \le j \le n\,\}$, i.e., $f_n = \sum_{j=1}^n \langle \xi_j, f\rangle \xi_j$. Then $A^*(f) = A^*(f_n)$ and

$$\|f - A^*(f)\|^2 \;=\; \|f - f_n\|^2 \;+\; \|f_n - A^*(f_n)\|^2.$$

Setting $a^2 = \sum_{j=1}^n \langle \xi_j, f\rangle^2/\gamma_j \le 1$ we have $\|f - f_n\|^2 \le (1-a^2)\,(r_n^{\mathrm{wor}})^2$ and $\|f_n - A^*(f_n)\|^2 \le a^2 (s_k^{\mathrm{wor}})^2$, where we used the fact that both $(r_n^{\mathrm{wor}})^2$ and $(s_k^{\mathrm{wor}})^2$ are homogeneous with respect to the squared 'radius' b of the ellipsoid $\mathcal{F}_b = \{\, f \in F : \sum_{j=1}^\infty \langle \xi_j, f\rangle^2/\gamma_j \le b\,\}$. Hence

$$e^{\mathrm{wor}}(A^*) \;\le\; \max_{0\le a\le 1} \sqrt{(1-a^2)(r_n^{\mathrm{wor}})^2 \,+\, a^2 (s_k^{\mathrm{wor}})^2} \;=\; \max\{\, r_n^{\mathrm{wor}}, s_k^{\mathrm{wor}}\}.$$

By Lemma 1, $\max\{\, r_n^{\mathrm{wor}}, s_k^{\mathrm{wor}}\}$ is also the lower bound for any approximation that uses n observations and k neurons. Hence A^* is optimal and the formula for its error follows. □

Let us briefly discuss relations between $r_n^{\rm wor}$ and $s_k^{\rm wor}$. Observe first that if $n \le k$ then $s_k^{\rm wor} \le r_n^{\rm wor}$, which is due to the fact that, in this case, (5) is an n-term approximation. If $k < n$ then any of the two minimal errors can dominate. For $k = n$ we have

$$\sqrt{\frac{1}{n+1}} \cdot r_n^{\rm wor} \;\le\; s_n^{\rm wor} \;\le\; r_n^{\rm wor}. \tag{9}$$

The lower bound follows from the fact that the squared minimal error of an n-term approximation for

$$f \;=\; \beta \cdot \sum_{i=1}^{n+1} \xi_j,$$

where $\beta = \left(\sum_{j=1}^{n+1} \gamma_j^{-1}\right)^{-1}$, equals just β, and $\beta \ge \gamma_{n+1}/(n+1)$. Furthermore, both bounds in (9) are sharp. Indeed, if $\gamma_j = 1$ for $1 \le j \le n+1$ and $\gamma_j = 0$ otherwise, then $(n+1)^{-1/2} \cdot r_n^{\rm wor} = s_n^{\rm wor}$. On the other hand, if $\gamma = 1$ for all $j \ge 1$, then $r_n^{\rm wor} = s_n^{\rm wor}$.

We can draw the following conclusion from Theorem 2. In order to construct an approximation with error at most ε, it is necessary and sufficient to use $n = {\rm IC}^{\rm wor}(\varepsilon)$ observations and $k = {\rm NC}^{\rm wor}(\varepsilon)$ neurons. By (9) we have ${\rm NC}^{\rm wor}(\varepsilon) \le {\rm IC}^{\rm wor}(\varepsilon)$, and the ratio of the two complexities can be arbitrarily large. For instance, if $\gamma_j \approx j^{-2p}$ with $p > 0$, then

$$\frac{{\rm NC}^{\rm wor}(\varepsilon)}{{\rm IC}^{\rm wor}(\varepsilon)} \;\approx\; \frac{2p}{2p+1},$$

which follows from Examples 1 and 2.

4 Randomization

We now consider *non-deterministic* approximations where the information is obtained and/or the network is built depending on a random parameter t. Thus an approximation procedure is now formally defined as a family $A = \{A_t\}_{t \in T}$, where T is an arbitrary measurable set in $\mathbb{R}^s$, $s \ge 1$, with some probability measure ω. For a given $t \in T$ we have $A_t = (N_t, \varphi_t)$, i.e., the approximation to f is obtained as $\tilde{f} = \varphi_t(y)$, where $y = N_t(f)$ is information (1) about f. This means that we randomize with respect to information and/or networks.

The main question is whether randomization can reduce complexity in the combined model. To give an answer, we first formally define the notions of error and complexity in the non-deterministic case.

The error of a random approximation $A = \{A_t\}$ is defined as [1]

$$e^{\rm ran}(A, \omega) \;=\; \sup_{f \in \mathcal{F}} \int_T \|f - A_t(f)\| \, \omega(dt).$$

[1]To avoid technical difficulties with measurability of integrands, by an integral we mean here and subsequently the upper integral.

We also define the complexity of information $N = \{N_t\}$ as

$$\mathrm{ic}^{\mathrm{ran}}(N, \omega) \;=\; \sup_{f \in \mathcal{F}} \int_T n_t(N, f)\, \omega(dt),$$

where $n_t(N, f)$ is the number of observations of $f \in F$ with the random parameter $t \in T$ (or, equavalently, the cardinality of $N_t f$), and we define the complexity of the network approximation $\varphi = \{\varphi_t\}$ as

$$\mathrm{nc}^{\mathrm{ran}}(\varphi, \omega) \;=\; \sup_{f \in \mathcal{F}} \int_T k_t(\varphi, f)\, \omega(dt),$$

where $k_t(\varphi, f)$ is the number of neurons used in the network approximation $\varphi_t(f)$ of f with parameter t.

Finally, we define randomized information complexity and randomized neural complexity as

$$\begin{aligned} \mathrm{IC}^{\mathrm{ran}}(\varepsilon) \;=\; \inf\{ \;\; & \mathrm{ic}^{\mathrm{ran}}(N, \omega) : \quad N = \{N_t\} \quad \text{and} \quad \omega \quad \text{such that} \\ & \text{for some} \quad \varphi = \{\varphi_t\} \quad \text{is} \quad e^{\mathrm{ran}}(N, \varphi, \omega) \le \varepsilon \,\}, \end{aligned} \tag{10}$$

and

$$\begin{aligned} \mathrm{NC}^{\mathrm{ran}}(\varepsilon) \;=\; \inf \Big\{ \;\; & \mathrm{nc}^{\mathrm{ran}}(\varphi, \omega) : \quad \varphi = \{\varphi_t\} \quad \text{and} \quad \omega \quad \text{such that} \\ & \sup_{f \in \mathcal{F}} \int_T \|f - \varphi_t(f)\|\, \omega(dt) \le \varepsilon \Big\}. \end{aligned} \tag{11}$$

We now show a result corresponding to Theorem 1 from the deterministic setting.

Theorem 3 *In order to obtain a randomized approximation (6) in the class $\mathcal{F}$ with error at most ε, it is necessary to use information with at least $\mathrm{IC}^{\mathrm{ran}}(\mathcal{F}, \mathcal{L}; \varepsilon)$ observations and a network with at least $\mathrm{NC}^{\mathrm{ran}}(\mathcal{F}, \mathcal{D}; \varepsilon)$ neurons, and it is sufficient to use information with $\mathrm{IC}^{\mathrm{ran}}(\mathcal{F}, \mathcal{L}; \alpha\varepsilon)$ observations and a network with $\mathrm{NC}^{\mathrm{ran}}(\mathcal{F}, \mathcal{D}; \beta\varepsilon)$ neurons. Here $\alpha, \beta > 0$ are arbitrary numbers satisfying $2\alpha + \beta < 1$.*

Proof. Since the lower bound is again obvious, we concentrate on the upper bound. Let $\eta > 0$ be arbitrary such that $(2\alpha + \beta)\varepsilon + \eta \le \varepsilon$. We take $N = \{N_{t_1}\}_{t_1 \in T_1}$ and $\varphi = \{\varphi_{t_1}\}_{t_1 \in T_1}$ with $t_1 \sim \omega_1$, such that $e^{\mathrm{ran}}(N, \varphi, \omega_1) \le \alpha\varepsilon$ and $\mathrm{ic}^{\mathrm{ran}}(N, \omega_1) \le \mathrm{IC}^{\mathrm{ran}}(\alpha\varepsilon) + \eta$. Then we define $\tilde{\varphi} = \{\tilde{\varphi}_{t_1}\}_{t_1 \in T_1}$ such that for all t_1 and y we have $\tilde{\varphi}_{t_1}(y) \in \mathcal{F}$ and

$$\|\tilde{\varphi}_{t_1}(y) - \varphi_{t_1}(y)\| \;\le\; \inf_{h \in \mathcal{F}} \|h - \varphi_{t_1}(y)\| + \eta.$$

Note that then

$$\|f - \tilde{\varphi}_{t_1}(y)\| \;\le\; \|f - \varphi_{t_1}(y)\| + \|\tilde{\varphi}_{t_1}(y) - \varphi_{t_1}(y)\| \;\le\; 2 \cdot \|f - \varphi_{t_1}(y)\| + \eta.$$

We also take $\psi = \{\psi_{t_2}\}_{t_2 \in T_2}$ with $t_2 \sim \omega_2$ such that for all $f \in \mathcal{F}$ is $\int_{T_2} \|f - \psi_{t_2}(f)\|\, \omega_2(dt_{t_2}) \le \beta\varepsilon$, and $\mathrm{nc}^{\mathrm{ran}}(\psi, \omega_2) \le \mathrm{NC}^{\mathrm{ran}}(\beta\varepsilon) + \eta$.

Now we let $T = T_1 \times T_2$, $\omega = \omega_1 \times \omega_2$, and define information $N^* = \{N_t^*\}_{t \in T}$ and approximation $\varphi^* = \{\varphi_t^*\}_{t \in T}$ as $N_t^* = N_{t_1}$ and $\varphi_t^*(\cdot) = \psi_{t_2}(\tilde{\varphi}_{t_1}(\cdot))$, where $t = (t_1, t_2)$. Then

$$\begin{aligned} \|f - \varphi_t^*(N_t f)\| &= \|f - \psi_{t_2}(\tilde{\varphi}_{t_1}(N_{t_1} f))\| \\ &\leq \|f - \tilde{\varphi}_{t_1}(N_{t_1} f)\| + \|\tilde{\varphi}_{t_1}(N_{t_1} f) - \psi_{t_2}(\tilde{\varphi}_{t_1}(N_{t_1} f))\|, \end{aligned}$$

so that for any $f \in \mathcal{F}$

$$\begin{aligned} &\int_T \|f - \varphi_t^*(N_t f)\| \, \omega(dt) \\ &\leq \int_{T_2} \int_{T_1} \|f - \tilde{\varphi}_{t_1}(N_{t_1} f)\| \, \omega_1(dt_1) \omega_2(dt_2) \\ &\quad + \int_{T_1} \int_{T_2} \|\tilde{\varphi}_{t_1}(N_{t_1} f) - \psi_{t_2}(\tilde{\varphi}_{t_1}(N_{t_1} f))\| \, \omega_2(dt_2) \omega_1(dt_1) \\ &\leq 2 \int_{T_1} \|f - \varphi_{t_1}(N_{t_1} f)\| \, \omega_1(dt_1) + \eta + \beta\varepsilon \leq (2\alpha + \beta)\varepsilon + \eta. \end{aligned}$$

We also have

$$\mathrm{ic}^{\mathrm{ran}}(N^*, \omega) = \mathrm{ic}^{\mathrm{ran}}(N, \omega_1) \leq \mathrm{IC}^{\mathrm{ran}}(\alpha\varepsilon) + \eta$$

and

$$\mathrm{nc}^{\mathrm{ran}}(\varphi^*, \omega) = \sup_{f \in \mathcal{F}} \int_{T_1} \int_{T_2} k_{(t_1, t_2)}(f) \, \omega_2(dt_2) \omega_1(dt_1) \leq \mathrm{NC}^{\mathrm{ran}}(\beta\varepsilon) + \eta,$$

since for any fixed t_1, the expected value of $k_{(t_1,t_2)}(f)$ is the expected number of neurons in approximation of $\tilde{\varphi}_{t_1}(N_{t_1} f) \in \mathcal{F}$ using the appoximation ψ.

Since η can be arbitrarily small, the proof is complete. $\square$

Theorem 3 implies that randomization does not help much in the combined model if and only if it does not help for both information complexity and neural complexity. We obviously have

$$\mathrm{IC}^{\mathrm{ran}}(\varepsilon) \leq \mathrm{IC}^{\mathrm{wor}}(\varepsilon) \quad \text{and} \quad \mathrm{NC}^{\mathrm{ran}}(\varepsilon) \leq \mathrm{NC}^{\mathrm{wor}}(\varepsilon).$$

The question of whether randomization can significantly help for information complexity has been studied in IBC. It is known that, even though we can randomize the choice of functionals to be observed as well as the number of them, for our general approximation problem the answer is usually negative. See, e.g., Novak (1988) for sufficient conditions for randomization not to help.

The question of whether randomization helps for neural complexity seems not to have been studied yet. Obviously, the only way to reduce the neural complexity would be by randomizing the number k of neurons in approximations. It turns out, however, that this can help only a little.

Theorem 4 *For any $\varepsilon > 0$ and $m > 1$ we have*

$$\Big(1 - \frac{1}{m}\Big) \cdot \mathrm{NC}^{\mathrm{wor}}(m\varepsilon) \leq \mathrm{NC}^{\mathrm{ran}}(\varepsilon) \leq \mathrm{NC}^{\mathrm{wor}}(\varepsilon).$$

Proof. It suffices to prove the left hand inequality. To this end, we first show the following auxiliary fact. For $f \in \mathcal{F}$ and $\varepsilon_1 > 0$, let

$$k(f, \varepsilon_1) \;=\; \min \{\, k : \, \exists \phi \in F_k \quad \text{s.t.} \quad \|f - \phi\| \le \varepsilon_1 \,\}.$$

Suppose that there exists $f \in \mathcal{F}$ and a convex function $\gamma : (0, \infty) \to (0, \infty)$ such that

$$k(f, \varepsilon_1) \;\ge\; \gamma(\varepsilon_1), \qquad \forall \varepsilon_1 > 0.$$

Then

$$\mathrm{NC}^{\mathrm{ran}}(\varepsilon) \;\ge\; \gamma(\varepsilon). \tag{12}$$

Indeed, let $\varphi = \{\varphi_t\}$ be such that $\int_T \|f - \varphi(f)\| \, \omega(dt) \le \varepsilon$ and assume that ϕ uses a network from F_k with probability p_k to approximate f. [2] Letting $\varepsilon_k = \inf_{\phi \in F_k} \|f - \phi\|$ we then have $\varepsilon \ge \sum_{k=0}^{\infty} p_k \varepsilon_k$ and the average number of neurons used for f is $l = \sum_{k=0}^{\infty} p_k k$. This and convexity of γ give

$$l \;\ge\; \sum_{k=0}^{\infty} p_k k(f, \varepsilon_k) \;\ge\; \sum_{k=0}^{\infty} p_k \gamma(\varepsilon_k) \;\ge\; \gamma\Big(\sum_{k=0}^{\infty} p_k \varepsilon_k \Big) \;\ge\; \gamma(\varepsilon),$$

which yields (12).

Now, let $\varepsilon > 0$ and $f \in \mathcal{F}$ be a function for which $\mathrm{NC}^{\mathrm{wor}}(m\varepsilon)$ is attained. Define

$$\gamma(\varepsilon_1) \;=\; \begin{cases} (1 - \varepsilon_1/(m\varepsilon)) \cdot \mathrm{NC}^{\mathrm{wor}}(m\varepsilon) & 0 < \varepsilon_1 < m\varepsilon, \\ 0 & \varepsilon_1 \ge m\varepsilon. \end{cases}$$

Then γ is convex and $\gamma(\varepsilon_1) \le k(f, \varepsilon_1)$, $\forall \varepsilon_1$. By (12) we then have

$$\mathrm{NC}^{\mathrm{ran}}(\varepsilon) \;\ge\; \gamma(\varepsilon) \;=\; \Big(1 - \frac{1}{m}\Big) \cdot \mathrm{NC}^{\mathrm{wor}}(m\varepsilon),$$

as claimed. □

Remark 1 One can actually show the following. For $f \in \mathcal{F}$ and $\varepsilon > 0$, let $\tilde{k}(f, \varepsilon)$ be the lower convex envelope of $k(f, \varepsilon)$, i.e., $\tilde{k}(f, \varepsilon) \le k(f, \varepsilon)$, $\forall \varepsilon > 0$, and for any other function $k_1(f, \varepsilon)$ satisfying the last inequality we have $k_1(f, \varepsilon) \le \tilde{k}(f, \varepsilon)$, $\forall \varepsilon$. (Note that such a function exists.) Then

$$\mathrm{NC}^{\mathrm{ran}}(\varepsilon) \;=\; \sup_{f \in \mathcal{F}} \tilde{k}(f, \varepsilon).$$

Thus, since $k(f, \varepsilon)$ is not convex (exept for some trivial cases), randomization always helps; however, by Theorem 4, we can gain only a little.

[2] We can assume without loss of generality that $t \mapsto k_t(\varphi, f)$ is measurable.

5 Noisy information

In this section, we discuss a generalization of Theorem 3 to the case of noisy information. That is, we assume that each piece of information is given as

$$y_j \;=\; L_j f + z_j, \qquad 1 \le j \le n,$$

where z_j is a *random noise* in the jth observation. More precisely, we assume that z_j's are independent random variables distributed according to some known distribution p_{σ_j} on $\mathbb{R}$. The parameter σ_j represents noise level and $\sigma_j = 0$ corresponds to the situation when there is no noise with probability one. A primary example is the Gaussian noise in which case $p_\sigma \;=\; Normal(0, \sigma^2)$.

Formally, *noisy information* is a pair (N, Δ), where N represents the choice of functionals L_j to be observed and Δ represents the choice of precisions σ_j. Both, L_j and σ_j, as well as the total number n of observations, can be in general selected adaptively based on the previously obtained values $y_1, \ldots, y_{j-1}$; see Plaskota (1996) for more details. An approximation to f is, as always, given as $\varphi(y)$ where y is noisy information about f. Thus an approximation procedure is a triple $A = (N, \Delta, \varphi)$.

Similarly to the noiseless case, in the randomized setting an approximation procedure is a family $A = \{A_t\}_{t \in T}$, $t \sim \omega$, where $A_t = (N_t, \Delta_t, \varphi_t)$ is a deterministic procedure for any t.

The error of A in the 'noisy' case is defined by adding another integral which is due to noise,

$$e^{\mathrm{ran-noi}}(N, \Delta, \varphi) \;=\; \sup_{f \in \mathcal{F}} \int_T \int_{Y_{f,t}} \|f - \varphi_t(y)\| \, \pi_{f,t}(dy) \omega(dt).$$

Here $Y_{f,t} \subset \cup_{i=1}^{\infty} \mathbb{R}^i$ is the set of all possible values of information for given f and parameter t, and $\pi_{f,t}$ is the probability distribution of information about f in $Y_{f,t}$. Note that, since any deterministic approximation can be treated as non-deterministic one by letting T be a singleton, we do not write separate definitions for the two kinds of approximation.

We also have to introduce the cost of noisy information in order to be able to define information complexity. Our model of cost is again taken from Plaskota (1996); namely, the cost of obtaining a single value $y_j = L_j f + z_j$ with noise $z_j \sim p_{\sigma_j}$ equals $c(\sigma_j)$, where $c : [0, \infty) \to [0, \infty]$ is a given nonincreasing cost function. For instance, $c(\sigma) = 1$ if $\sigma \ge \sigma_0 \ge 0$, and $c(\sigma) = \infty$ if $\sigma < \sigma_0$, corresponds to the situation when all the observations are performed with fixed precision σ_0. These assumptions imply the complexity of information,

$$\mathrm{ic}^{\mathrm{ran-noi}}(N, \Delta, \omega) \;=\; \sup_{f \in \mathcal{F}} \int_T \int_{Y_{f,t}} \mathrm{cost}_{f,t}(y) \, \pi_{f,t}(dy) \omega(dt),$$

where $\mathrm{cost}_{f,t}(y)$ is the cost of obtaining information y about f with parameter t.

Finally, information ε-complexity, $\mathrm{IC}^{\mathrm{ran-noi}}(\varepsilon)$, is defined as in (10) with N, $\mathrm{ic}^{\mathrm{ran}}$, and e^{ran} replaced by (N, Δ), $\mathrm{ic}^{\mathrm{ran-noi}}$, and $e^{\mathrm{ran-noi}}$, respectively. We can similarly define information ε-complexity, $\mathrm{NC}^{\mathrm{wor-noi}}(\varepsilon)$, for noisy but non-randomized information. Note that the presence of noise does not change the definition of neural ε-complexity $\mathrm{NC}^{\mathrm{ran}}(\varepsilon)$. However, in the combined model, the final approximation $\varphi_t(y)$ to f depends not only on the random parameter t, but also on the noise, because the information y about f does.

It turns out that a result analogous to Theorem 3 holds in the case of noisy information.

Theorem 5 *In order to obtain a randomized approximation (6) with error at most ε using noisy information, it is necessary to pay at least $\mathrm{IC}^{\mathrm{ran-noi}}(\mathcal{F},\mathcal{L};\varepsilon)$ for observations and use at least $\mathrm{NC}^{\mathrm{ran}}(\mathcal{F},\mathcal{D};\varepsilon)$ neurons, and it is sufficient to pay $\mathrm{IC}^{\mathrm{ran-noi}}(\mathcal{F},\mathcal{L};\alpha\varepsilon)$ for observations and use $\mathrm{NC}^{\mathrm{ran}}(\mathcal{F},\mathcal{D};\beta\varepsilon)$ neurons. Here $\alpha,\beta>0$ are arbitrary numbers satisfying $2\alpha+\beta<1$.*

Proof. The proof follows the proof of Theorem 3 with obvious modifications corresponding to the presence of noise. □

Recall that randomization does not help much for neural complexity. We also have that randomization does not help for information complexity in the presence of noise, as shown in Plaskota (1996a). Thus Theorem 5 can be interpreted as follows. Even if randomized approximations are allowed, the best approximations use essentially $\mathrm{NC}^{\mathrm{wor-noi}}(\varepsilon)$ observations and $\mathrm{NC}^{\mathrm{wor}}(\varepsilon)$ neurons.

6 Example: piecewise polynomial networks

We illustrate the obtained results using the well known example of one dimensional piecewise polynomial networks. [3] We assume $F=C(D)$ with $D=[0,1]$ and the uniform norm,

$$\|f\| = \max_{0\le x\le 1} |f(x)|.$$

The neurons evaluate functions from the dictionary

$$\mathcal{D} = \mathcal{D}_s = \{\, w(\min(u,\cdot)) : \, w\in \mathrm{Poly}(s),\ u\in D\,\},$$

where $\mathrm{Poly}(s)$ are polynomials of degree at most s, $s\ge 1$. Obviously, in this case the set F_k of all k-term approximations (6) includes all continuous piecewise polynomials with at most k pieces. We want to approximate functions from the Hölder class $\mathcal{F}=C_{r,\alpha}$,

$$C_{r,\alpha} = \left\{ f\in C^r(D): \ |f^{(r)}(x_1)-f^{(r)}(x_2)| \le |x_1-x_2|^{\alpha},\ 0\le x_1,x_2\le 1 \right\}$$

with $r\ge 0$ and $0<\alpha\le 1$, or from the Sobolev class $\mathcal{F}=W_{r,p}$,

$$W_{r,p} = \left\{ f\in C^r(D): \ f^{(r)}\text{-abs. cont.}, \ \|f^{(r+1)}\|_p \le 1 \right\}$$

[3]Multivariate piecewise polynomial networks would be more interesting, but very little is known about lower bounds for neural complexity in this case.

with $r \geq 0$ and $1 \leq p \leq \infty$. Information y about f is given by

$$y \;=\; (f(x_1) + z_1, f(x_2) + z_2, \ldots, f(x_n) + z_n\,),$$

where the noise is Gaussian with fixed variance σ^2. In this case, $\mathcal{L} = \{\, L : \exists x \in D, \text{ s.t. } Lf = f(x) \text{ for all } f \in \mathcal{F}\,\}$. We restrict our considerations to deterministic approximations as randomization does not help much for these problems; see Novak (1988) and Plaskota (1996a), and Theorem 4.

We first consider the deterministic and noiseless situation ($\sigma = 0$).

Theorem 6 *(i) For information complexity we have*

$$\mathrm{IC}^{\mathrm{wor}}(C_{r,\alpha};\varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+\alpha}} \qquad \textit{and} \qquad \mathrm{IC}^{\mathrm{wor}}(W_{r,p};\varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1/q}},$$

where $1/p + 1/q = 1$.

(ii) For neural complexity we have the following. If $1 \leq s \leq r-1$ *then*

$$\mathrm{NC}^{\mathrm{wor}}(C_{r,\alpha}, \mathcal{D}_s;\varepsilon) \;=\; \mathrm{NC}^{\mathrm{wor}}(W_{r,p}, \mathcal{D}_s;\varepsilon) \;=\; +\infty.$$

If $s \geq r$ *then*

$$\mathrm{NC}^{\mathrm{wor}}(C_{r,\alpha}, \mathcal{D}_s;\varepsilon) \;\asymp\; \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+\alpha}} \qquad \textit{and} \qquad \mathrm{NC}^{\mathrm{wor}}(W_{r,p}, \mathcal{D}_s;\varepsilon) \;\asymp\; \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1}}.$$

Proof. The proof of (i) can be found, e.g., in Heinrich (1993) or Novak (1988). Part (ii) can be shown using a standard technique see, e.g., DeVore, Howard, and Micchelli (1989), or DeVore (1998). We only mention that in case $\mathcal{F} = W_{r,p}$ with $1 \leq p < \infty$, the knots $0 = x_0 < x_1 < \cdots < x_k \leq 1$ of the (almost) optimal approximation $\phi(x) = \sum_{j=1}^{k} w_j(\min(x_j, x))$ essentially depend on f, and they are selected such that

$$(x_j - x_{j-1})^{r+1/q}\Big(\int_{x_{j-1}}^{x_j} |f^{(r+1)}(u)|^p du\Big)^{1/p} \;\leq\; \Big(\frac{1}{k}\Big)^{r+1}, \qquad j = 1, 2, \ldots, k. \tag{13}$$

In all other cases, equidistant knots (and sampling) are optimal. □

We now comment on Theorem 6. For dictionary $\mathcal{D}_s$ with polynomials of degree $s < r$, it is impossible to construct networks with finite error in any of the Hölder or Sobolev classes, since the neural complexity is in this case infinite. Let $s \geq r$. Then, to construct a network with error $\varepsilon > 0$ in Hölder class $C_{r,\alpha}$, it is necessary and sufficient to use $\Theta(\varepsilon^{-1/(r+\alpha)})$ function values and the same amount of neurons. Moreover, the (almost) optimal sample points x_j are equidistant and equal to the knots in expansion (6). In particular, they are chosen independently of f. The same applies for the Sobolev class $W_{r,\infty}$, i.e., we have to use $\Theta(\varepsilon^{-1/(r+1)})$ equidistant samples and knots. The situation changes for $W_{r,p}$ with $1 \leq p < \infty$. If $r = 0$ and $p = 1$ then it is again impossible to approximate with finite error in $W_{0,1}$ since information complexity is infinite. For $r > 0$

or $1 < p < \infty$ the equidistant sampling at $\Theta(\varepsilon^{-1/(r+1/q)})$ points is still (almost) optimal, but we need only $\Theta(\varepsilon^{-1/(r+1)})$ neurons. The knots x_j of the network approximating f are in this case selected adaptively, i.e., these depend on the obtained information about the values of f. For instance, assume a minimal smoothness $r = 0$, and $p = 2$. Then we need $\Theta(\varepsilon^{-2})$ samples, but only $\Theta(\varepsilon^{-1})$ neurons.

We also comment on practical construction of (almost) optimal networks. The construction is quite easy in cases where $\mathrm{IC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{D}_s; \varepsilon) \asymp \mathrm{NC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{D}_s; \varepsilon)$. We just take piecewise polynomial interpolation ϕ of f based on equidistant sampling. In Sobolev classes with $\mathrm{IC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{D}_s; \varepsilon) \gg \mathrm{NC}^{\mathrm{wor}}(\mathcal{F}, \mathcal{D}_s; \varepsilon)$, the situation is more complicated, since the first step of the algorithm from the proof of Lemma 1 requires finding a function $f_y \in W_{r,p}$ interpolating data $y_j = f(j/n)$, $0 \le j \le n$, with $n = n(\varepsilon) = \Theta(\varepsilon^{-1/(r+1/q)})$ (which is known to exist). This is in general not an easy task. In case $p = 2$, f_y can be chosen as the natural spline of degree $2r + 1$ interpolating data $y = Nf$. In the second step, we select $k = k(\varepsilon) = \Theta(\varepsilon^{-1/(r+1)})$ knots x_j for f_y using condition (13), and then we interpolate f_y at x_j's by a network with k neurons. Note that the resulting network $\phi = \phi(y)$ does not necessarily interpolate the original function f at any of the x_j.

In the 'noisy' case $\sigma > 0$, by Plaskota (1996a) we have

$$\mathrm{NC}^{\mathrm{wor-noi}}(\mathcal{F}, \sigma; \varepsilon) \asymp \sigma^2 \Big(\frac{1}{\varepsilon}\Big)^{\gamma+2} \ln\Big(\frac{1}{\varepsilon}\Big),$$

where $\mathcal{F} = C_{r,\alpha}$ or $\mathcal{F} = W_{r,p}$, and γ is the exponent for the class $\mathcal{F}$ in the noiseless case, see Theorem 6. Hence, in any case, we need $\Omega(\varepsilon^{-2})$ samples, which is large as compared to the number of neurons needed. For instance, for minimal smoothness $r = 0$, and for $p = 2$, we have $\mathrm{IC}^{\mathrm{ran-noi}}(\sigma, \varepsilon) \asymp \sigma^2 \varepsilon^{-4} \ln(1/\varepsilon)$ while $\mathrm{NC}^{\mathrm{wor}}(\varepsilon) \asymp \varepsilon^{-1}$.

References

1. Bergeaud F., and Mallat, S. (1996), Matching pursuit: Adaptive representations of images and sounds, *Computational and Appl. Math.* **15**.
2. Chui, C.K., Li, X., and Mhaskar, H.N. (1996), Limitation capabilities of neural networks with one hidden layer, *Advances in Comput. Math.* **5**, pp. 233-243.
3. DeVore, R.A. (1998), Nonlinear Approximation, *Acta Numerica*, pp. 51–150.
4. DeVore, R.A., Howard, R., and Micchelli, C.A. (1989), Optimal nonlinear approximation, *Manuscripta Mathematica* **63**, pp. 469–478.
5. DeVore, R.A. and Lorentz, G.G. (1993), *Constructive Approximation*, Springer.
6. DeVore, R.A., and Temlyakov, V.N. (1995), Nonlinear approximation by trigonometric sums, *J. Fourier Anal. Appl.* **2**, pp. 29–48.
7. DeVore, R.A., and Temlyakov, V.N. (1997), Nonlinear approximation in finite dimensional spaces, *J. of Complexity* **13**, pp. 489–508.
8. Heinrich, S. (1993), Random approximation in numerical analysis, in *Proc. of the Functional Analysis Conf., Essen, 1991*, (Bierstedt *et al.*, Eds.), pp. 123–171, Dekker, New York.

9. Kon, M.A., and Plaskota, L. (2000), Information complexity of neural networks, to appear in *Neural Networks*.
10. Mhaskar, H.N. (1996), Neural networks for optimal approximation of smooth and analytic functions, *Neural Computation* **8**, pp. 164–177.
11. Mhaskar H.N., and Micchelli, C.A. (1995), Degree of approximation by neural and translation networks with a single hidden layer, *Advances in Appl. Math.* **161**, pp. 151-183.
12. Novak, E. (1988), *Deterministic and Stochastic Error Bounds in Numerical Analysis*, Lecture Notes in Math., Vol. 1349, Springer, Berlin.
13. Pinkus, A. (1999), Approximation theory of the MLP model in neural networks, *Acta Numerica*, pp. 143–195.
14. Plaskota, L. (1996), *Noisy information and computational complexity*, Cambridge Univ. Press, Cambridge.
15. Plaskota, L. (1996a), Worst case complexity of problems with random information noise, *J. of Complexity* **12**, pp. 416–439.
16. Traub, J.F., Wasilkowski, G.W., and Woźniakowski, H. (1988), *Information-based Complexity*, Academic Press, New York.
17. Traub, J.F., and Werschulz (1999), *Complexity and Information*, Cambridge University Press, Cambridge.

Authors addresses:

Mark Kon
Boston University
Department of Mathematics & Statistics
111 Cummington Street, Boston, MA 02215
email: mkon@math.bu.edu

Leszek Plaskota
Warsaw University
Department of Mathematics, Informatics, & Mechanics
ul. Banacha 2, 02–097 Warsaw, Poland
email: leszekp@mimuw.edu.pl