

INFORMATION-BASED NONLINEAR APPROXIMATION: AN AVERAGE CASE SETTING

MARK KON AND LESZEK PLASKOTA

ABSTRACT. *Nonlinear approximation* has usually been studied under deterministic assumptions and complete information about the underlying functions. In the present paper we assume only *partial information*, e.g., function values at some points, and we are interested in the *average case* error and complexity of nonlinear approximation. We show that the problem can be essentially decomposed in two independent problems related to average case nonlinear (restricted) approximation from complete information, and to average case unrestricted approximation from partial information. The results are then applied to average case piecewise polynomial approximation on $C([0, 1])$ based on function values with respect to r -fold Wiener measure. In this case, to approximate with error ε it is necessary and sufficient to know the function values at $\Theta \left(\left(\varepsilon^{-1} \ln^{1/2}(1/\varepsilon) \right)^{1/(r+1/2)} \right)$ equidistant points and use $\Theta \left(\varepsilon^{-1/(r+1/2)} \right)$ adaptively chosen break points in piecewise polynomial approximation.

1. INTRODUCTION

Nonlinear approximation (NA) relies on approximating a function f using a nonlinear manifold that consists of k -term linear combinations of functions from a dictionary. Such approximation is constructed based on full knowledge about f , and one is interested in the error as k goes to infinity, see, e.g., [1] for a review. Since in practice the underlying function is usually given via its (exact or noisy) values at finitely many points, the assumption that f is completely known can be sometimes questioned. On the other hand, there is *information-based complexity (IBC)* theory where partial information is essential; see, e.g., [7]. In this situation the study of nonlinear approximation based on partial and/or noisy information becomes quite natural.

In [2, 3], *information-based nonlinear approximation (IBNA)* was studied in the context of neural networks. The authors considered the *worst case setting* and established, among other things, the following remarkable property. Suppose one wants to know how many function evaluations and how many terms in approximation he needs to approximate the function with error $\varepsilon > 0$. It turns out that this problem can be essentially solved by decomposing it in two problems corresponding to NA and IBC. Two crucial notions used are information complexity and neural complexity, which mean, respectively, the minimal number of function evaluations and the minimal number of

Date: November 7, 2004.

1991 Mathematics Subject Classification. 41A46, 41A10.

Key words and phrases. nonlinear approximation, information-based complexity, average case error, Wiener measure.

terms in approximation sufficient to approximate with error ε . Since the two quantities have been studied (mostly independently) in NA and IBC, the results from both theories can be integrated to obtain results in IBNA.

In the present paper we study information-based nonlinear approximation in the *average case setting*. That is, assuming the functions are distributed according to a probability measure, we ask for the average number of function evaluations and average number of terms in approximation necessary and sufficient to approximate within the average error ε .

Information complexity in the average case setting has been the topic of extensive study in IBC, but, unlike in the worst case, average case NA seems not to have been regularly studied yet. Therefore the purpose of this paper is twofold. First, we want to establish general results corresponding to those from the worst case. This is done in Sections 3 and 4 where we show that, again, to obtain the complexity results and construct best approximations in the average case IBNA, it suffices to combine the corresponding results and approximations in the average case IBC and NA. Second, we want to provide a first thorough analysis of complexity of average case nonlinear approximation for a nontrivial problem. This is done in Section 5, where piecewise polynomial uniform approximation with respect to r -fold Wiener measure is considered.

We now use the problem of Section 5 to illustrate results obtained in this paper. Suppose we want to uniformly approximate a random function $f : [0, 1] \rightarrow \mathbb{R}$ distributed according to the r -fold Wiener measure. The approximation is based on n evaluations, $f(x_1), \dots, f(x_n)$, and given as a piecewise polynomial of degree at most s ($s \geq r$) with k break points. We stress that adaption is allowed, i.e., the choice of successive points at which the function is evaluated and the number n of them may depend on previously obtained values (adaptive information), and the break points of the approximation and the number k of them may depend on the gathered information about f (adaptive approximation). We ask how the points x_j in information and the break points in approximation should be chosen to approximate f with average error ε at minimal cost. The cost is measured by the average values of n and k . Are the optimal information and approximation adaptive or not?

The information complexity for this problem has been studied in IBC and it is known that f should be evaluated at

$$n \asymp \left(\frac{\ln^{1/2}(1/\varepsilon)}{\varepsilon} \right)^{\frac{1}{r+1/2}}$$

equidistant points x_j . The best information is then nonadaptive. If the same x_j s are used as the break points then we will already obtain nonadaptive approximation with error ε and $k = n$. Actually, this is best we can get from nonadaptive approximation. However, we can do better by selecting the break points adaptively depending on variability of the underlying function. As a result, we can get rid of the log factor, obtaining

$$k \asymp \left(\frac{1}{\varepsilon} \right)^{\frac{1}{r+1/2}}.$$

Further reduction of the cost is impossible; see Section 5 for details.

These results can also be interpreted as follows. If complete information about f were available then the best convergence of the average error would be $k^{-(r+1/2)}$ where, as before, k is the number of break points in piecewise polynomial approximation. Since complete information is usually unrealistic, it makes sense to ask what the minimal knowledge about f is, measured by the number n of function values, that allows to approximate f with average error of the same order. The answer is that it is necessary and sufficient to know a number proportional to $n = k(\ln k)^{1/(2r+1)}$ values at equally spaced points.

2. BASIC CONCEPTS

Let F be a real separable Banach space equipped with a probability measure μ on the Borel sets of F . Let G be another normed space such that F is continuously embedded in G . By $\|\cdot\|$ we denote the norm in G . Any $A : F \rightarrow G$ such that $f \mapsto \|f - A(f)\|$ is a measurable mapping is called an approximation operator (or just *approximation*). The (q th average) error of A is defined as

$$(1) \quad e(A) = \left(\int_F \|f - A(f)\|^q \mu(df) \right)^{1/q},$$

where $1 \leq q < \infty$.

We put two restrictions on possible approximations. First, we assume that $A(f)$ is for any $f \in F$ an element of a specified set $\Phi \subset G$, i.e., the range $A(F) \subset \Phi$. We also assume that

$$\Phi = \bigcup_{k=1}^{\infty} \Phi_k,$$

where $\Phi_1 \subset \Phi_2 \subset \Phi_3 \subset \dots$. Although the general results of this paper hold for arbitrary Φ_k s, we will have in mind the nonlinear approximation. That is, we choose a set $\mathcal{D} \subset G$, called a *dictionary*, and define

$$(2) \quad \Phi_k = \left\{ \sum_{j=1}^k a_j f_j : a_j \in \mathbb{R}, f_j \in \mathcal{D} \right\}.$$

If $A(F) \subset \Phi_k$ then A is called a k -term approximation. A k -term approximation is *nonadaptive* if $A(F)$ is in a linear space spanned by some k elements of \mathcal{D} . Otherwise A is *adaptive*, in which case the f_j s in (2) and/or the number of them are chosen depending on f .

The second restriction is that $A(f)$ is based on some *information* about f . This is defined as the value $y = Nf$ of a measurable mapping $N : F \rightarrow Y$ with range $Y \subset \bigcup_{n=0}^{\infty} \mathbb{R}^n$. More specifically, let $\Lambda \subset F^*$ be a set of permissible information functionals. Then *nonadaptive* information is a mapping $N : F \rightarrow \mathbb{R}^n$,

$$Nf = (L_1 f, L_2 f, \dots, L_n f)$$

where $L_j \in \Lambda$, $1 \leq j \leq n$. The number n is usually called *cardinality* of N . In *adaptive* information the choice of the functionals L_j depends on previously obtained values $y_i = L_i f$, $1 \leq i \leq j-1$. In this case, we formally have

$$Nf = (L_1 f, L_2(f; y_1), \dots, L_n(f; y_1, \dots, y_{n-1}))$$

where $L_j(\cdot; y_1, \dots, y_{j-1}) \in \Lambda$, $\forall y_i$, $1 \leq j \leq n$. We can also vary the number n of functionals obtaining information of varying cardinality. In this case, we gather information by evaluating the successive $y_j = L_j(f; y_1, \dots, y_{j-1})$ until the condition $(y_1, \dots, y_j) \in Y$ is reached. For this to be well defined, we assume that for any infinite sequence $y = (y_1, y_2, y_3, \dots)$ there exists an index n such that $(y_1, \dots, y_n) \in Y$. For details and further discussion see, e.g., [7].

Any approximation $A : F \rightarrow \Phi$ that is based on some information $N : F \rightarrow Y$ can be decomposed as

$$A = \varphi \circ N$$

where the mapping $\varphi : Y \rightarrow \Phi$. Furthermore,

$$\begin{aligned} e(\varphi \circ N) &= \left(\int_F \|f - \varphi(Nf)\|^q \mu(df) \right)^{1/q} \\ &= \left(\int_Y \int_F \|f - \varphi(y)\|^q \mu(df|y) \mu_N(dy) \right)^{1/q}, \end{aligned}$$

where $\mu_N = \mu N^{-1}$ is the a priori distribution of information y on Y , and $\mu(\cdot|y)$ is the conditional distribution on F given information y .

Remark 1. In a more general model one assumes that information is corrupted by some random *noise*. In this case, the information operator is defined as a mapping $\tilde{N} : \tilde{F} \rightarrow Y$ with $\tilde{F} = F \times \mathbb{R}^\infty$. Noisy information about f is given as $y = \tilde{N}(f, x)$,

$$y = (L_1 f + x_1, L_2(f; y_1) + x_2, \dots, L_n(f; y_1, \dots, y_{n-1}) + x_n),$$

where x_j , $j \geq 1$, are independent random variables with known distribution. Hence an approximation is a mapping $A = \varphi \circ \tilde{N} : \tilde{F} \rightarrow G$. Denoting by $\tilde{\mu}$ the joint distribution on $F \times \mathbb{R}^\infty$, we have

$$\begin{aligned} e(\varphi \circ \tilde{N}) &= \left(\int_{\tilde{F}} \|f - \varphi(\tilde{N}(f, x))\|^q \tilde{\mu}(d(f, x)) \right)^{1/q} \\ &= \left(\int_Y \int_F \|f - \varphi(y)\|^q \mu(df|y) \mu_{\tilde{N}}(dy) \right)^{1/q}. \end{aligned}$$

See Plaskota (1996) for more details.

3. THE OPTIMAL k -TERM APPROXIMATION

In this section we assume that information $N : F \rightarrow Y$ (adaptive or nonadaptive, with fixed or varying cardinality) is given, and we seek the best possible choice of $\varphi : Y \rightarrow \Phi_k$, so that the error of the k -term approximation $A = \varphi \circ N$ is minimized. As it will turn out, the minimal error depends on two independent quantities which are related to the set Φ_k and information N , respectively. We will define them in turn.

The first quantity, denoted s_k , is the average distance of elements $f \in F$ from Φ_k ,

$$s_k = \left(\int_F \inf_{\tilde{f} \in \Phi_k} \|f - \tilde{f}\|^q \mu(df) \right)^{1/q}.$$

Equivalently, s_k is the minimal error that can be achieved by k -term approximations from *complete* information about f , i.e.,

$$s_k = \inf_{\psi: F \rightarrow \Phi_k} e(\psi).$$

The second quantity, called the (*q*th average) *radius of information* and denoted $r(N)$ is defined as

$$r(N) = \left(\int_Y (\text{rad}(\mu(\cdot|y)))^q \mu_N(dy) \right)^{1/q},$$

where $\text{rad}(\nu)$ is the radius of a measure,

$$\text{rad}(\nu) = \left(\inf_{g \in F} \int_F \|f - g\|^q \nu(df) \right)^{1/q}.$$

That is, $r(N)$ is the average radius of the conditional measures $\mu(\cdot|y)$ with respect to information $y = Nf$.

It is well known that $r(N)$ is also the minimal error of approximations of the form $A = \varphi \circ N$ with arbitrary $\varphi : Y \rightarrow G$ (the approximation is based on information N , but the restriction $\varphi(y) \in \Phi$ is relaxed). That is,

$$r(N) = \inf_{\varphi: Y \rightarrow G} e(\varphi \circ N).$$

Recall that

$$(3) \quad \text{rad}(\nu) \leq \left(\int_F \int_F \|f_1 - f_2\|^q \nu(df_1) \nu(df_2) \right)^{1/q} \leq 2 \cdot \text{rad}(\nu).$$

Theorem 1. *The minimal error of the k -term approximation based on given information $N : F \rightarrow Y$ satisfies*

$$\max(r(N), s_k) \leq \inf_{\varphi: Y \rightarrow \Phi_k} e(\varphi \circ N) \leq 2 \cdot \max(2r(N), s_k).$$

Proof. The lower bound is obvious. To show the upper bound, we use a nondeterministic argument. That is, suppose for a moment that we apply the approximation $y \mapsto \psi(g)$ with a deterministic component $\psi : F \rightarrow \Phi_k$, and with the nondeterministic component g which is chosen randomly according to the conditional distribution $\mu(\cdot|y)$ on F . Using the decomposition of μ with respect to information y , inequality $(a+b)^q \leq 2^{q-1}(a^q + b^q)$

(for $a, b > 0$), and (3), the error of such approximation satisfies

$$\begin{aligned}
& \left(\int_Y \int_F \left(\int_F \|f - \psi(g)\|^q \mu(df|y) \right) \mu(dg|y) \mu_N(dy) \right)^{1/q} \\
& \leq 2^{1-1/q} \left(\int_Y \left(\int_F \int_F \|f - g\|^q \mu(dg|y) \mu(df|y) \right) \mu_N(dy) \right. \\
& \quad \left. + \int_Y \int_F \|g - \psi(g)\|^q \mu(dg|y) \mu_N(dy) \right)^{1/q} \\
& \leq 2^{1-1/q} \left(2^q \cdot \int_Y \text{rad}(\mu(\cdot|y))^q \mu_N(dy) + \int_F \|f - \psi(f)\|^q \mu(df) \right)^{1/q} \\
& \leq 2^{1-1/q} (2^q \cdot r(N)^q + e(\psi)^q)^{1/q} \\
& \leq 2 \cdot \max(2r(N), e(\psi)).
\end{aligned}$$

Now, by the mean value theorem, there exists a $\varphi : Y \rightarrow G$ such that

$$\int_F \|f - \psi(\varphi(y))\|^q \mu(df|y) \leq \int_F \int_F \|f - \psi(g)\|^q \mu(df|y) \mu(dg|y).$$

The approximation $\varphi^*(Nf) = \psi(\varphi(Nf))$ is then deterministic and its error

$$e(\varphi^* \circ N) \leq 2 \cdot \max(2r(N), e(\psi)).$$

To complete the proof, it suffices to minimize this with respect to ψ . \square

The essence of Theorem 1 is that, for given information N , the minimal error of k -term approximations is proportional to $\max(r(N), s_k)$ where both quantities, $r(N)$ and s_k , can be studied independently of each other. On the other hand, Theorem 1 is not constructive, i.e., it does not give a construction of the deterministic approximation φ^* whose error attains the upper bound. We now present another estimate of the minimal error from which the construction of φ^* will follow.

Recall first that c is a center of a measure ν iff

$$\left(\int_F \|f - c\|^q \nu(df) \right)^{1/q} = \text{rad}(\nu).$$

We assume, for simplicity, that for all y a.e. there exists a center, denoted $c(y)$, of the conditional measure $\mu(\cdot|y)$. Let

$$\begin{aligned}
\bar{s}_k(N) &= \left(\int_Y \inf_{\tilde{f} \in \Phi_k} \|c(y) - \tilde{f}\|^q \mu_N(dy) \right)^{1/q} \\
&= \inf_{\varphi: Y \rightarrow \Phi_k} \left(\int_Y \|c(y) - \varphi(y)\|^q \mu_N(dy) \right)^{1/q}.
\end{aligned}$$

Theorem 2. *The minimal error of k -term approximation based on given information $N : F \rightarrow Y$ satisfies*

$$\max(r(N), \bar{s}_k(N)/2) \leq \inf_{\varphi: Y \rightarrow \Phi_k} e(\varphi \circ N) \leq 2 \cdot \max(r(N), \bar{s}_k(N)).$$

Proof. For arbitrary $\varphi : Y \rightarrow \Phi_k$ we have

$$\begin{aligned} e(\varphi \circ N)^q &\geq \int_Y \int_F \|c(y) - \varphi(y)\| - \|f - c(y)\|^q \mu(df|y) \mu_N(dy) \\ &\geq \int_Y \int_F 2^{1-q} \|c(y) - \varphi(y)\|^q - \|f - c(y)\|^q \mu(df|y) \mu_N(dy) \\ &\geq 2^{1-q} \bar{s}_k(N)^q - r(N)^q. \end{aligned}$$

Hence, if $\bar{s}_k(N) \leq 2r(N)$ then $e(\varphi \circ N) \geq r(N) = \max(r(N), \bar{s}_k(N))$, and if $\bar{s}_k(N) \geq 2r(N)$ then

$$\begin{aligned} e(\varphi \circ N) &\geq (2^{1-q} \bar{s}_k(N)^q - (\bar{s}_k(N)/2)^q)^{1/q} \\ &= \bar{s}_k(N)/2 \\ &= \max(r(N), \bar{s}_k(N)). \end{aligned}$$

This yields

$$\inf_{\varphi: Y \rightarrow \Phi_k} e(\varphi \circ N) \geq \max(r(N), \bar{s}_k(N)/2).$$

Consider now the approximation $\varphi_\eta : Y \rightarrow \Phi_k$ such that

$$(4) \quad e(\varphi_\eta \circ N) \leq \bar{s}_k(N) + \eta,$$

where $\eta > 0$. Then

$$\begin{aligned} e(\varphi_\eta \circ N)^q &\leq \int_Y \int_F 2^{q-1} (\|f - c(y)\|^q + \|c(y) - \varphi_\eta(y)\|^q) \mu(df|y) \mu_N(dy) \\ &= 2^{q-1} (\text{rad}(N)^q + (\bar{s}_k(N) + \eta)^q)^{1/q} \\ &\leq 2^q \cdot \max(\text{rad}(N)^q, (\bar{s}_k(N) + \eta)^q). \end{aligned}$$

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

Suppose now that for all y a.e. there exists a best approximation φ^c of $c(y)$ in Φ_k . That is,

$$(5) \quad \|c(y) - \varphi^c(y)\| = \inf_{\tilde{f} \in \Phi_k} \|c(y) - \tilde{f}\|.$$

Then Theorem 2 immediately yields

Corollary 1. *We have*

$$e(\varphi^c \circ N) \leq 4 \cdot \inf_{\varphi: Y \rightarrow \Phi_k} e(\varphi \circ N).$$

Remark 2. Under closer inspection of the proof of Theorem 2 one can see that the estimate of Corollary 1 can be slightly improved; namely

$$e(\varphi^c \circ N) \leq 2^{1-1/q} (1 + 2^q)^{1/q} \cdot \inf_{\varphi: Y \rightarrow \Phi_k} e(\varphi \circ N).$$

The following example shows that this estimate is sharp, at least for $q = 1$.

Let $F = \mathbb{R}^2$ with the measure μ concentrated on two points, $(0, a)$ and $(1, a)$, $0 < a < 1$, with weights p_1 and p_2 , respectively, $p_1 < p_2$, $p_1 + p_2 = 1$. Let $\mathcal{D} = \{(1, 0), (0, 1)\}$. Take $N \equiv 0$ (zero information) and $k = 1$, i.e., we are interested in 1-term approximations. The error is measured in ℓ_1 -norm. Then the center of μ is $c = (1, a)$ and $\text{rad}(\mu) = p_1$.

The closest approximation to c is $\varphi^c = (1, 0)$, and its error $e(\varphi^c) = a + p_1$. On the other hand, for $\varphi = (0, 1)$ we have $e(\varphi) = p_2$. The ratio $e(\varphi^c)/e(\varphi)$ is arbitrarily close to 3 if $p_1 \approx p_2 \approx 1/2$ and $a \approx 1$.

Remark 3. In some cases, the approximation φ^c is optimal. Suppose that μ is a zero-mean Gaussian measure on F . Then, for any information $N : F \rightarrow Y$, the conditional distribution $\mu(\cdot|y)$ (with $y = Nf$) is also Gaussian. Furthermore, $\mu(\cdot|y)$ is symmetric about $m(y)$, and the center $c(y) = m(y)$, $\forall y$ a.e. Suppose also that the error is measured in a Hilbert norm and $q = 2$. Then for any $\varphi : Y \rightarrow \Phi_k$ we have

$$\begin{aligned} e(\varphi \circ N)^2 &= \int_Y \int_F \|f - \varphi(y)\|^2 \mu(df|y) \mu_N(dy) \\ &= \int_Y \int_F \|f - m(y)\|^2 + \|m(y) - \varphi(y)\|^2 \\ &\quad + \langle f - m(y), m(y) - \varphi(y) \rangle \mu(df|y) \mu_N(dy) \\ &\geq \int_Y \int_F \|f - m(y)\|^2 + \|m(y) - \varphi^c(y)\|^2 \mu(df|y) \mu_N(dy) \\ &= e(\varphi^c \circ N)^2, \end{aligned}$$

as claimed. One can also show that for arbitrary q and error measured in arbitrary norm the constant 4 in Corollary 1 can be replaced by 2.

Remark 4. Using the same proofs one can show that Theorems 1 and 2, and Corollary 1 hold also for noisy information \tilde{N} as defined in Remark 1. Obviously, for noisy information \tilde{N} its radius

$$r(\tilde{N}) = \left(\int_Y (\text{rad}(\mu(\cdot|y)))^q \mu_{\mathbb{R}}(dy) \right)^{1/q}.$$

If, in addition, μ and noise are Gaussian, i.e., $x_j \sim \mathcal{N}(0, \sigma^2)$, then Remark 3 is also valid.

Example 1. Let μ be a zero mean Gaussian measure on F with symmetric and positive definite covariance operator $C_\mu : F^* \rightarrow F$. Let G be a separable Hilbert space with the inner product $\langle \cdot, \cdot \rangle$. Let the dictionary $\mathcal{D} = \{\xi_j : j \geq 1\}$ where the ξ_j s form a complete orthonormal system in G .

It is easy to see that for given $f \in F$ the best k -term approximation $\psi^*(f)$ is adaptive and given as follows. Let $B_{k,f}$ be the set of k indices for which $|\langle f, \xi_j \rangle|$, $j \geq 1$, are largest possible, i.e., $\#B_{k,f} = k$, and if $i \notin B_{k,f}$, $j \in B_{k,f}$ then $|\langle f, \xi_i \rangle| \leq |\langle f, \xi_j \rangle|$. Then $\psi^*(f) = \sum_{j \in B_{k,f}} \langle f, \xi_j \rangle \xi_j$.

Suppose now that approximation is based on information $y = (L_1 f + x_1, L_2 f + x_2, \dots, L_n f + x_n)$ with $L_j \in F^*$, where the noise $x_j \sim \mathcal{N}(0, \sigma^2)$. Note that information is in general noisy, but $\sigma = 0$ corresponds to exact (noiseless) information. In this case, the center of the conditional measure $\mu(\cdot|y)$ is $c(y) = \sum_{j=1}^n z_j (C_\mu L_j)$, where z is the solution of the linear system $(\sigma^2 I_n + H)z = y$, I_n is the identity in \mathbb{R}^n , and $H = \{L_i(C_\mu L_j)\}_{i,j=1}^n$ is the Gram matrix, see, e.g., Plaskota (1996).

Hence the almost optimal (and optimal for $q = 2$) k -term approximation based on information y is $\varphi^c(y) = \psi^*(c(y))$. Observe that this approximation is in general

adaptive. It is however nonadaptive when $k \geq n$ and $\xi_j = C_\mu L_j$ for $1 \leq j \leq n$, since then $\varphi^c(y) = c(y) \in \text{span}\{\xi_1, \dots, \xi_n\}$.

4. COST AND ε -COMPLEXITY

We now consider the problem of complexity. That is, we ask for the minimal cost of obtaining an approximation with error at most ε .

The (average) cost of information $N : F \rightarrow Y$ is defined as

$$\begin{aligned} \text{cost}(N) &= \int_F n(Nf) \mu(df) \\ &= \int_Y n(y) \mu_N(dy), \end{aligned}$$

where $n(y)$ is such that $y \in \mathbb{R}^{n(y)}$. Similarly, the (average) cost of an approximation $\psi : F \rightarrow \Phi$ is defined as

$$\text{cost}(\psi) = \int_F k(\psi(f)) \mu(df),$$

where $k(\tilde{f}) = \min \{k : \tilde{f} \in \Phi_k\}$ is the number of terms in the approximation $\tilde{f} \in \Phi$.

We need the following result.

Lemma 1. *Let $N : F \rightarrow Y$ be given information and $\psi : F \rightarrow \Phi$ a given approximation. Let $\alpha, \beta > 0$ with $1/\alpha + 1/\beta < 1$. Then there exists $\varphi : Y \rightarrow \Phi$ such that*

$$e(\varphi \circ N) \leq 2\alpha^{1/q} \cdot \max(2r(N), e(\psi))$$

and

$$\text{cost}(\varphi \circ N) \leq \beta \cdot \text{cost}(\psi).$$

Proof. For any $y \in Y$ there is $g_y \in F$ such that

$$\int_F \|f - \psi(g_y)\|^q \mu(df|y) \leq \alpha \cdot \int_F \int_F \|f - \psi(g)\|^q \mu(df|y) \mu(dg|y)$$

and

$$k(\psi(g_y)) \leq \beta \cdot \int_F k(\psi(g)) \mu(dg|y).$$

Indeed, by the Chebyshev inequality the set of elements g_y for which the first inequality holds is of measure at least $(1 - 1/\alpha)$, and the set of elements g_y for which the second inequality holds is of measure at least $(1 - 1/\beta)$. Since $(1 - 1/\alpha) + (1 - 1/\beta) > 1$, both sets have nonempty intersection.

Define

$$\varphi(y) = \psi(g_y), \quad y \in Y.$$

Then

$$\begin{aligned}
e(\varphi \circ N)^q &= \int_F \|f - \psi(g_{Nf})\|^q \mu(df) \\
&= \int_Y \int_F \|f - \psi(g_y)\|^q \mu(df|y) \mu_N(dy) \\
&\leq \alpha \cdot \int_Y \int_F \int_F \|f - \psi(g)\|^q \mu(dg|y) \mu(df|y) \mu_N(dy) \\
&\leq 2^{q-1} \cdot \alpha \cdot \int_Y \int_F \int_F \|f - g\|^q + \|g - \psi(g)\|^q \mu(dg|y) \mu(df|y) \mu_n(dy) \\
&\leq 2^{q-1} \cdot \alpha \cdot (2^q r(N)^q + e(\psi)^q) \\
&\leq 2^q \alpha \cdot \max(2^q r(N)^q, e(\psi)^q),
\end{aligned}$$

and

$$\begin{aligned}
\text{cost}(\varphi \circ N) &= \int_F k(\psi(g_{Nf})) \mu(df) \\
&= \int_Y k(\psi(g_y)) \mu_N(dy) \\
&\leq \beta \cdot \int_Y \int_F k(\psi(g)) \mu(dg|y) \mu_n(dy) \\
&= \beta \cdot \text{cost}(\psi),
\end{aligned}$$

as claimed. \square

We now define the *information ε -complexity* as

$$\text{comp}^I(\varepsilon) = \inf \{ \text{cost}(N) : N : F \rightarrow Y \text{ s.t. there is } \varphi : Y \rightarrow G \text{ with } e(\varphi \circ N) \leq \varepsilon \},$$

and the *approximation ε -complexity* as

$$\text{comp}^A(\varepsilon) = \inf \{ \text{cost}(\psi) : \psi : F \rightarrow \Phi \text{ s.t. } e(\psi) \leq \varepsilon \}.$$

By Lemma 1 we have the following.

Theorem 3. *Suppose we want an approximation $A(f) = \varphi(Nf)$ with average error $e(\varphi \circ N) \leq \varepsilon$. Then*

- *it is necessary to use on average $\text{comp}^I(\varepsilon)$ functional evaluations and $\text{comp}^A(\varepsilon)$ terms in approximation, and*
- *it is sufficient to use on average $\text{comp}^I(\varepsilon/(4\alpha^{1/q}))$ functional evaluations and $\beta \cdot \text{comp}^A(\varepsilon/(2\alpha^{1/q}))$ terms in approximation.*

Here $\alpha, \beta > 0$ with $1/\alpha + 1/\beta < 1$.

Thus the minimal cost of ε -approximation depends on $\text{comp}^I(\varepsilon)$ and $\text{comp}^A(\varepsilon)$. We stress that both quantities can be studied independently of each other.

One can ask if the best approximation, say $A^* = \varphi^* \circ N^*$, uses a constant number of information functionals and terms in approximation, and whether adaption helps or not. Due to Lemma 1, these questions can also be answered by studying $\text{comp}^I(\varepsilon)$ and $\text{comp}^A(\varepsilon)$. That is, if $\text{comp}^I(\varepsilon)$ is attained by nonadaptive information with a fixed number of functionals then N^* is also nonadaptive and uses a fixed number of

functionals. Similarly, if $\text{comp}^A(\varepsilon)$ is attained by approximation with a fixed number of terms then A^* also uses a fixed number of terms.

Remark 5. The problem whether adaption helps for information complexity has already been studied, see, e.g., [8] and [7] (and [5] for noisy information) where general conditions for adaptive information not to help in case of Gaussian μ are given. The corresponding question for approximation complexity seems not to have been studied yet.

Theorem 3 is not constructive. In the following, we give a possible construction of (almost) optimal information and approximation in the case when varying the number of terms in approximation does not help.

For given $\varepsilon > 0$, let information $N^* : F \rightarrow Y$ and approximation $\psi^* : F \rightarrow \Phi$ be such that $r(N^*) \leq \varepsilon/16$, $\text{cost}(N^*) \leq C_1 \cdot \text{comp}^I(\varepsilon/16)$, and $e(\psi^*) \leq \varepsilon/8$, $\text{cost}(\psi^*) \leq C_2 \cdot \text{comp}^A(\varepsilon/8)$. Assume also that for all $f \in F$ a.e.

$$(6) \quad k(\psi^*(f)) \leq k^* \leq C_3 \cdot \text{cost}(\psi^*),$$

i.e., ψ^* uses at most k^* terms. Finally, let $\varphi^c : Y \rightarrow \Phi_{k^*}$ be defined as in (5) (provided the corresponding center exists). Applying Corollary 1 and Theorem 1 we immediately obtain that for the approximation $A^*(f) = \varphi^c(N^*f)$

$$e(\varphi^c \circ N^*) \leq \varepsilon \quad \text{and} \quad \text{cost}(\varphi^c \circ N^*) \leq C_2 C_3 \cdot \text{comp}^A(\varepsilon/8),$$

i.e., A^* is close to the cheapest approximation with error at most ε .

Remark 6. Results of this section hold also for noisy information \tilde{N} defined in Remark 1, with obvious modifications of $\text{cost}(\tilde{N})$ and $\text{comp}^I(\varepsilon)$.

5. PIECEWISE POLYNOMIAL APPROXIMATION

We consider the following approximation problem. Let

$$F = F_r = \{ f \in C^r([0, 1]) : f(0) = f'(0) = \dots = f^{(r)}(0) = 0 \}$$

with the norm $\|f\|_r = \sup_{0 \leq t \leq 1} |f^{(r)}(t)|$. Let $\mu = w_r$ be the r -fold integrated Wiener measure on F , i.e., $w_r(B) = w_0(\{f^{(r)} : f \in B\})$ where w_0 is the classical Wiener measure (Brownian motion) on F_0 . We approximate $f \in F_r$ by piecewise polynomials of degree at most s , $s \geq r$, with finitely many break points that can be chosen adaptively. Specifically, we have $G = C([0, 1])$, i.e., the error is measured in the Chebyshev norm, $q = 1$ in (1), and

$$\mathcal{D} = \mathcal{D}_s = \{ w(\min(u, \cdot)) : w \in \Pi_s, u \in [0, 1] \}.$$

Information about f is given by its values at finitely many knots,

$$y = (f(x_1), f(x_2), \dots, f(x_n)),$$

$0 \leq x_1 \leq x_2 \leq \dots \leq x_n \leq 1$. Information can be in general adaptive.

The formula for $\text{comp}^I(\varepsilon)$ is known; namely

$$\text{comp}^I(\varepsilon) \asymp \left(\frac{\ln^{1/2}(1/\varepsilon)}{\varepsilon} \right)^{\frac{1}{r+1/2}}$$

see, e.g., [4]. Furthermore, the information attaining this estimate is nonadaptive, and it evaluates f at knots that are equally spaced in $[0, 1]$. For this information, the center $c(y)$ of the conditional measure $w_r(\cdot|y)$ is the natural spline of degree $2r+1$ interpolating the data y . Note that the given bound is obtained not only by $\varphi(y) = c(y)$, but also by the piecewise polynomial interpolation of degree r .

We now concentrate on $\text{comp}^A(\varepsilon)$ which has not been studied yet. Since the piecewise polynomial interpolation with k fixed nodes is also a nonadaptive k -term approximation, we have

$$\text{comp}^A(\varepsilon) = O\left(\left(\frac{\ln^{1/2}(1/\varepsilon)}{\varepsilon}\right)^{\frac{1}{r+1/2}}\right).$$

However, it turns out that the log factor above can be removed.

Theorem 4. *We have*

$$\text{comp}^A(\varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1/2}}.$$

Theorem 4 will be proven in several steps. First we show the lower bound on $\text{comp}^A(\varepsilon)$.

Lemma 2. *Let $\varepsilon_0 > 0$ be such that $\text{comp}^A(\varepsilon_0) > 2$. If an approximation $\psi : F \rightarrow \Phi$ has average error at most ε , where $0 < \varepsilon < \varepsilon_0$, then it uses on average at least*

$$\text{comp}^A(\varepsilon) \geq C_1 \cdot \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1/2}}$$

terms, where $C_1 = C_1(\varepsilon_0) = \varepsilon_0^{1/(r+1/2)}(\text{comp}^A(\varepsilon_0) - 2)/2$.

Proof. We first show that $\lim_{\varepsilon \rightarrow 0} \text{comp}^A(\varepsilon) = \infty$, so that ε_0 exists. Indeed, otherwise an l would exist such that for arbitrarily small ε there is an approximation ψ_ε with $e(\psi_\varepsilon) \leq \varepsilon$ and $\text{cost}(\psi_\varepsilon) \leq l$. This means that the set A_ε of functions that can be approximated with error at most 3ε using $3l$ terms is of measure at least $1/3$. Since $A_{1/2} \supset A_{1/3} \supset A_{1/4} \supset \dots$, we have $w_r(\cup_{j=2}^{\infty} A_{1/j}) \geq 1/3$. This implies

$$w_r(\{f : \text{dist}(f, \Phi_{3l})\}) \geq 1/3,$$

and contradicts the fact that the last measure is zero.

Consider now two separate problems of approximating f with the average error ε on the subintervals $[0, 1/2]$ and $[1/2, 1]$, instead of on the whole interval $[0, 1]$. Denote the corresponding average complexities of these problems as $\text{comp}^A_0(\varepsilon)$ and $\text{comp}^A_1(\varepsilon)$. Then

$$(7) \quad \text{comp}^A_1(\varepsilon) + 1 \geq \text{comp}^A_0(\varepsilon) = \text{comp}^A(2^{r+1/2}\varepsilon).$$

Indeed, the first relation follows from the fact that the processes $f|_{[0,1/2]}$ and $f|_{[1/2,1]}$ differ only by a random polynomial of degree r . (Here we also use the assumption $s \geq r$.) For the second relation it is enough to observe that $g(x)$ is the r -fold Wiener process on $[0, 1/2]$ iff $f(x) = 2^{r+1/2}g(x/2)$ is the r -fold Wiener process on $[0, 1]$.

Suppose now that ψ is an approximation with the average error ε . Since the error is measured in the uniform norm, the average error of ψ on each of the two subintervals

$[0, 1/2]$ and $[1/2, 1]$ is also at most ε . Hence ψ divides each of the sub-intervals into at least $\text{comp}^A_0(\varepsilon)$ pieces (on average). Noting that one of the pieces may overlap and using (7) we obtain

$$\begin{aligned} \text{comp}^A(\varepsilon) &\geq \text{comp}^A_0(\varepsilon) + \text{comp}^A_1(\varepsilon) - 1 \\ &\geq 2 \cdot (\text{comp}^A(2^{r+1/2}\varepsilon) - 1). \end{aligned}$$

Proceeding inductively, we finally arrive at the estimate

$$\text{comp}^A(\varepsilon) \geq 2^s (\text{comp}^A(\varepsilon_0) - 2),$$

where s is the largest integer such that $2^{s(r+1/2)}\varepsilon \leq \varepsilon_0$. Hence $2^s > (\varepsilon_0/\varepsilon)^{1/(r+1/2)}/2$ and the lemma follows. \square

We now construct two concrete approximations for which the lower bound $\varepsilon^{-1/(r+1/2)}$ of Lemma 2 is attained.

Consider the following approximation ψ_ε , $\varepsilon > 0$. For $f \in F_r$, we select the knots $0 = t_0 < t_1 < \dots < t_k = 1$ in such a way that

$$(8) \quad (t_j - t_{j-1})^r \max_{t_{j-1} \leq \xi_1, \xi_2 \leq t_j} |f^{(r)}(\xi_1) - f^{(r)}(\xi_2)| = \varepsilon \cdot r!$$

for $1 \leq j \leq k-1$, and we have inequality " \leq " above for $j = k$. Then we approximate f by a continuous piecewise polynomial $\tilde{f} = \psi_\varepsilon(f)$ of degree r with the break points t_j , i.e., on each interval $[t_{j-1}, t_j]$, \tilde{f} interpolates f at t_{j-1} , t_j , and arbitrary $r-1$ points in this interval. Note that ψ_ε is an adaptive approximation with the varying number k of terms.

Lemma 3. *For all $f \in F_r$ we have $\|f - \psi_\varepsilon(f)\| \leq \varepsilon$. Furthermore, the average number of terms used by ψ_ε is finite, and for any $\varepsilon_0 > 0$ and $0 < \varepsilon < \varepsilon_0$*

$$\text{cost}(\psi_\varepsilon) \leq C_2 \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1/2}}$$

where $C_2 = C_2(\varepsilon_0) = 2 \text{cost}(\psi_{\varepsilon_0})\varepsilon_0^{1/(r+1/2)}$.

Proof. Using the error formula for Lagrange interpolation we obtain for $x \in [t_{j-1}, t_j]$ that

$$|f(x) - \tilde{f}(x)| \leq \frac{(t_j - t_{j-1})^r}{r!} \max_{t_{j-1} \leq \xi_1, \xi_2 \leq t_j} |f^{(r)}(\xi_1) - f^{(r)}(\xi_2)| \leq \varepsilon.$$

Hence the error is always at most ε .

We now show that $\text{cost}(\psi_\varepsilon) < \infty$. Let $a = \text{Prob}\{k(\psi_\varepsilon(f)) \geq 2\}$. Then $0 < a < 1$ and, due to independence of $f^{(r)}(\xi_1^1) - f^{(r)}(\xi_2^1)$ and $f^{(r)}(\xi_1^2) - f^{(r)}(\xi_2^2)$ for $\xi_1^1 < \xi_2^1 < \xi_1^2 < \xi_2^2$, we also have $\text{Prob}\{k(\psi_\varepsilon(f)) \geq 3\} \leq a^2$ and, generally, $\text{Prob}\{k(\psi_\varepsilon(f)) \geq s\} \leq a^{s-1}$. Hence

$$\text{cost}(\psi_\varepsilon) \leq (1 - a) + \sum_{j=2}^{\infty} j a^{j-1} < \infty,$$

as claimed.

To get the upper bound for $\text{cost}(\psi_\varepsilon)$, we proceed as follows. Let $k_0(f, \varepsilon)$ and $k_1(f, \varepsilon)$ denote the number of intervals $[t_{j-1}, t_j]$ selected for f when the interval $[0, 1]$ is replaced by $[0, 1/2]$ and $[1/2, 1]$, respectively. We obviously have

$$k(\psi_\varepsilon(f)) \leq k_0(f, \varepsilon) + k_1(f, \varepsilon).$$

Letting

$$\begin{aligned} f_0(x) &= 2^{r+1/2} f(x/2), \\ f_1(x) &= 2^{r+1/2} \left(f((x+1)/2) - \sum_{j=0}^r f^{(j)}(1/2) x^j / (2^j j!) \right), \end{aligned}$$

we have that f_0 and f_1 are independent r -fold Wiener processes on $[0, 1]$. Moreover, since for $i = 0, 1$

$$|f_i^{(r)}(\xi_1) - f_i^{(r)}(\xi_2)| = \sqrt{2} \cdot |f^{(r)}((\xi_1 + i)/2) - f^{(r)}((\xi_2 + i)/2)|,$$

we also have $k_i(f, \varepsilon) = k(\psi_{2^{r+1/2}\varepsilon}(f_i))$. This in turn implies

$$\text{cost}(\psi_\varepsilon) \leq 2 \cdot \text{cost}(\psi_{2^{r+1/2}\varepsilon}),$$

or more generally

$$\text{cost}(\psi_\varepsilon) \leq 2^s \cdot \text{cost}(\psi_{\varepsilon_0}),$$

where s is the smallest integer for which $2^{-s(r+1/2)}\varepsilon_0 \leq \varepsilon$. The last inequality gives the desired result. \square

A similar estimate is obtained by an approximation that uses the same number of terms for all f . More specifically, let $k \geq 1$. For $f \in F_r$, we select the knots $0 = t_0 < \dots < t_k = 1$ in such a way that the quantities

$$(t_j - t_{j-1})^r \max_{t_{j-1} \leq \xi_1, \xi_2 \leq t_j} |f^{(r)}(\xi_1) - f^{(r)}(\xi_2)|$$

are equal for all $1 \leq j \leq k$. The approximation ψ_k relies on the piecewise polynomial interpolation of degree r with pieces determined by the knots t_j . Note that ψ_k is an adaptive k -term approximation.

Lemma 4. *The average error of ψ_k satisfies*

$$e(\psi_k) = O(k^{-(r+1/2)}).$$

Proof. As in the proof of Lemma 3, the error for given f is bounded from above by

$$\varepsilon_k = \varepsilon_k(f) = \max_{1 \leq j \leq k} \frac{(t_j - t_{j-1})^r}{r!} \max_{t_{j-1} \leq \xi_1, \xi_2 \leq t_j} |f^{(r)}(\xi_1) - f^{(r)}(\xi_2)|.$$

Hence it suffices to show that the average value of ε_k is $O(k^{-(r+1/2)})$.

For $\varepsilon > 0$ and $f \in F_r$, let the points $t_j, j \geq 0$, be defined as in (8). Then $T_{\varepsilon, j} = t_j - t_{j-1}$ are independent random variables and

$$\text{Prob}(\varepsilon_k < \varepsilon) = \text{Prob} \left(\sum_{j=1}^k T_{\varepsilon, j} > 1 \right).$$

Using the fact that if $f(x)$ is an r -fold Wiener process then so is $g(x) = c^{-(r+1/2)}f(cx)$ ($c > 0$), we find that $T_{\varepsilon,j} = \varepsilon^{1/(r+1/2)}T_j$ with $T_j = T_{1,j}$. Hence we can equivalently write

$$\text{Prob}(\varepsilon_k < \varepsilon) = \text{Prob}(S_k > \varepsilon^{-1/(r+1/2)}),$$

where $S_k = \sum_{j=1}^k T_j$. Denote by F_k and f_k the distribution and density of S_k , respectively. The error of ψ_k can be estimated as (E stands for expectation)

$$\begin{aligned} E(\varepsilon_k) &= \int_0^\infty (1 - \text{Prob}(\varepsilon_k < \varepsilon)) d\varepsilon \\ &= \int_0^\infty F_k(\varepsilon^{-1/(r+1/2)}) d\varepsilon \\ &= (r+1/2) \int_0^\infty F_k(\eta) \eta^{-(r+3/2)} d\eta \\ &= -F_k(\eta) \eta^{-(r+1/2)} \Big|_0^\infty + \int_0^\infty f_k(\eta) \eta^{-(r+1/2)} d\eta. \end{aligned}$$

The first component in the last sum vanishes since

$$(9) \quad \lim_{\eta \rightarrow 0^+} F_k(\eta) \eta^{-(r+1/2)} = 0.$$

Indeed,

$$\begin{aligned} F_k(\eta) &= \text{Prob}\left(\sum_{j=1}^k T_j \leq \eta\right) \leq \text{Prob}(T_1 \leq \eta) \\ &= \text{Prob}\left((\eta^r/r!) \max_{0 \leq \xi_1, \xi_2 \leq \eta} |f^{(r)}(\xi_1) - f^{(r)}(\xi_2)| \geq 1\right) \\ &= \text{Prob}\left((\eta^{r+1/2}/r!) \max_{0 \leq \xi_1, \xi_2 \leq 1} \eta^{-1/2} |f^{(r)}(\eta\xi_1) - f^{(r)}(\eta\xi_2)| \geq 1\right) \\ &= \text{Prob}\left(\max_{0 \leq \xi_1, \xi_2 \leq 1} |f^{(r)}(\xi_1) - f^{(r)}(\xi_2)| \geq r! \eta^{-(r+1/2)}\right) \\ &\leq \text{Prob}\left(2 \max_{0 \leq \xi \leq 1} |f^{(r)}(\xi)| \geq r! \eta^{-(r+1/2)}\right). \end{aligned}$$

Since $f^{(r)}$ is just Brownian motion, the last probability tends to zero exponentially fast as $\eta \rightarrow 0$, and (9) follows.

Thus

$$E(\varepsilon_k) = \int_0^\infty f_k(\eta) \eta^{-(r+1/2)} d\eta = E(S_k^{-(r+1/2)}),$$

where we have used the standard property that $\int f_k(\eta) g(\eta) d\eta = E(g(S_k))$.

Now, letting $\gamma = E(T_j)$ and $W_k = S_k/(k\gamma) - 1$, we find that $E(W_k) = 0$, $E(W_k^2) = O(1/k)$, and

$$\begin{aligned} E\left(\left(\frac{S_k}{k\gamma}\right)^{-(r+1/2)}\right) &= E((1+W_k)^{-(r+1/2)}) \\ &= 1 - (r+1/2)E(W_k) + O(E(W_k^2)) \\ &= 1 + O(1/k). \end{aligned}$$

Hence $E(S_k^{-(r+1/2)}) \approx (k\gamma)^{-(r+1/2)}$, which completes the proof. \square

Theorem 4 is thus proven. We now show, in addition, that the use of adaption for the upper bound is crucial.

Lemma 5. *For any nonadaptive k -term approximation ψ_k^{non} we have*

$$e(\psi_k^{non}) = \Omega\left(k^{-(r+1/2)} \ln^{1/2} k\right).$$

Proof. We first show the following. Let f be a continuous zero mean Gaussian stochastic process on $[0, 1]$ with positive definite covariance kernel. Let w_f be the polynomial of degree at most s best approximating f with respect to the Chebyshev norm $\|\cdot\|_C$ on $[0, 1]$. Define the random variable

$$(10) \quad X = \|f - w_f\|_C.$$

Then there is $\sigma > 0$ such that

$$(11) \quad \text{Prob}(X \geq \alpha) \geq \sqrt{\frac{2}{\pi\sigma^2}} \int_\alpha^\infty e^{-z^2/(2\sigma^2)} dz.$$

Indeed, let $0 \leq t_0 < \dots < t_{s+1} \leq 1$ be arbitrary $s+2$ distinct points. Let v_f be the polynomial best approximating f with respect to the maximum norm on the set $\{t_j\}_{j=0}^{s+1}$. Then the t_j s are the alternation points and

$$v_f(x) = \tilde{v}_f(x) - c \cdot p_{s+1}(x),$$

where \tilde{v}_f is the polynomial of degree $s+1$ interpolating f at the t_j s, p_{s+1} is the polynomial of degree $s+1$ such that $p_{s+1}(t_j) = (-1)^j$, $0 \leq j \leq s+1$, and c is the ratio of the leading coefficients (divided differences) of v_f and p_{s+1} , i.e.,

$$c = \frac{f(t_0, \dots, t_{s+1})}{p_{s+1}(t_0, \dots, t_{s+1})}.$$

Hence

$$\|f - w_f\|_C \geq \max_{0 \leq j \leq s+1} |f(t_j) - v_f(t_j)| = |c|.$$

Since the divided difference $f(t_0, \dots, t_{s+1})$ is a nonzero linear functional of the $f(t_j)$ s, it is a zero mean Gaussian random variable with a positive variance σ , and (11) follows.

We now use (11) to prove the lemma. Divide the unit interval into $2k$ equal subintervals. Then one can select k distinct subintervals in which there are no break points of $\psi_k^{non}(f)$. It is clear that the error on $[0, 1]$ is not larger than the maximum of errors from independent approximations of f by polynomials of degree s on the chosen k subintervals. These minimal errors are independent random variables and $X_i = (2k)^{-(r+1/2)} X$, $1 \leq i \leq k$, where X is as in (10). Hence, by (11) we obtain

$$\begin{aligned} E(\|f - \psi_k^{non}(f)\|) &\geq (2k)^{-(r+1/2)} E\left(\max_{1 \leq i \leq k} X_i\right) \\ &\asymp k^{-(r+1/2)} \ln^{1/2} k, \end{aligned}$$

as claimed. \square

Summarizing, the cheapest information-based nonlinear approximation $A_\varepsilon : F \rightarrow \Phi$ with error ε is obtained as follows.

- (1) Choose $n = n(\varepsilon) \asymp \varepsilon^{-(r+1/2)} \ln(1/\varepsilon)$ and $k = k(\varepsilon) \asymp \varepsilon^{-(r+1/2)}$.
- (2) Observe $y = (f(1/n), f(2/n), \dots, f(1))$.
- (3) Find the natural spline s_y of degree $2r + 1$ interpolating the data, $s_y(j/n) = y_j$, $0 \leq j \leq n$ ($y_0 = 0$).
- (4) Apply the adaptive k -term approximation ψ_k of Lemma 4 on s_y , i.e., $A_\varepsilon(f) = \psi_k(s_y)$.

Remark 7. We have seen that for the problem of this section the difference between information complexity and approximation complexity is only by a log factor. The situation changes for noisy information with zero mean Gaussian noise of variance $\sigma^2 > 0$. Indeed, then

$$\text{comp}^I(\varepsilon) \asymp \left(\frac{\ln^{1/2}(1/\varepsilon)}{\varepsilon} \right)^{\frac{4r+4}{2r+1}}.$$

Furthermore, the cheapest approximation A_ε is in this case obtained as follows.

- (1) Choose $n = n(\varepsilon) \asymp (\varepsilon^{-2} \ln(1/\varepsilon))^{(2r+2)/(2r+1)}$ and $k = k(\varepsilon) \asymp \varepsilon^{-(r+1/2)}$.
- (2) Observe $y = (f(1/n) + x_1, f(2/n) + x_2, \dots, f(1) + x_n)$.
- (3) Find the natural spline $s_{\sigma,y}$ of degree $2r + 1$ minimizing the penalty functional

$$P(f) = \sigma^2 \cdot \int_0^1 |f^{(r+1)}(u)|^2 du + \sum_{j=1}^n (y_j - f(j/n))^2.$$

- (4) Apply ψ_k of Lemma 4 on $s_{\sigma,y}$, i.e., $A_\varepsilon(f, x) = \psi_k(s_{\sigma,y})$.

For details, see Plaskota (1998).

Acknowledgments: This research was partially supported by the U.S. National Science Foundation (M. Kon) and by the State Committee for Scientific Research of Poland (L. Plaskota).

REFERENCES

- [1] R.A. DeVore, Nonlinear approximation, *Acta Numerica* 8 (1998) 51–150.
- [2] M.A. Kon, L. Plaskota, Information complexity of neural networks, *Neural Networks* 13 (2000) 365–376.
- [3] M.A. Kon, L. Plaskota, Complexity of neural network approximation with limited information: a worst case approach, *J. Complexity* 17 (2000) 345–365.
- [4] V.E. Maiorov, G.W. Wasilkowski, Probabilistic and average linear widths in L_∞ norm with respect to r -fold Wiener measure, *J. Approx. Th.* 84 (1996) 31–40.
- [5] L. Plaskota, *Noisy Information and Computational Complexity*, Cambridge University Press, Cambridge, 1996.
- [6] L. Plaskota, Average case L_∞ approximation in the presence of Gaussian noise, *J. Approx. Th.* 93 (1998) 501–515.
- [7] J.F. Traub, G.W. Wasilkowski, H. Woźniakowski, *Information-based Complexity*, Academic Press, New York, 1988.
- [8] G.W. Wasilkowski, Information of varying cardinality, *J. Complexity* 2 (1986) 204–228.

DEPARTMENT OF MATHEMATICS AND STATISTICS, BOSTON UNIVERSITY, 111 CUMMINGTON STREET, BOSTON, MA 12215

E-mail address: `mkon@math.bu.edu`

DEPARTMENT OF MATHEMATICS, INFORMATICS AND MECHANICS, WARSAW UNIVERSITY, BANACHA 2, 02-097 WARSAW, POLAND

E-mail address: `leszekp@mimuw.edu.pl`