# Information-based nonlinear approximation: an average case setting

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

December 2004

**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 approximation. We show that the problem can be essentially split into two independent problems related to average case nonlinear (restricted) approximation from complete information, and average case unrestricted approximation from partial information. The results are then applied to average case piecewise polynomial approximation in $C([0,1])$ based on function values with respect to $r$-fold Wiener measure. In this case, to approximate with average error $\varepsilon$ 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.

*Subject classification:* 41A46, 41A10

*Key words and phrases:* nonlinear approximation, partial information, average case error, Wiener measure

*Research partially supported by the U.S. National Science Foundation

†Research partially supported by the Sate Committee for Scientific Research of Poland (KBN) under Grant 5 P03A 007 21

# 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 given 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., DeVore (1998) 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 sometimes be questioned. On the other hand, there is *information-based complexity* (*IBC*) theory, where partial information is essential, see, e.g., Traub *et al.* (1988). In this situation the study of nonlinear approximation based on partial and/or noisy information becomes quite natural.

In Kon and Plaskota (2000a,b), *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 splitting it into two corresponding problems in NA and IBC. The two crucial notions used are information complexity and neural complexity, which mean, respectively, the minimal number of function evaluations and the minimal number of terms in approximation sufficient to approximate with error $\varepsilon$. 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 $\varepsilon$.

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 **??** and **??** 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 **??** where piecewise polynomial uniform approximation with respect to the $r$-fold Wiener measure is considered.

We now use the problem of Section **??** to illustrate results obtained in this paper.

Suppose we want to uniformly approximate a random function $f : [0,1] \to \mathbb{R}$ distributed according to the $r$-fold Wiener measure. The approximation is based on $n$ evaluations, $f(t_1), \ldots, f(t_n)$, 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 $t_j$ in information and the break points in approximation should be chosen to approximate $f$ with average error $\varepsilon$ 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 was 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 $t_j$. The best information is then nonadaptive. If the same $t_j$'s are used as the break points then we will already obtain nonadaptive approximation with error $\varepsilon$ 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

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

Further reduction of the cost is impossible; see Section **??** 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 evaluations, that allows to approximate $f$ with average error of the same order. The answer is that it is necessary and sufficient to know proportionally to $n = k^{-(r+1/2)} \ln^{1/2} k$ values at equally spaced points.

# 2 Basic concepts

Let $F$ be a real separable Banach space equipped with a probability measure $\mu$ 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 $A : F \to G$ such that $f \mapsto \|f - A(f)\|$ is a measurable mapping is called an approximation operator (or

just *approximation*). The *(qth average) error* of $A$ is defined as

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

where $1 \le q < \infty$.

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 image $Im(A) \subset \Phi$. We also assume that

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

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

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

If $Im(A) \subset \Phi_k$ then $A$ is called a *k-term approximation*. A $k$-term approximation is *nonadaptive* if $Im(A)$ 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 (**??**) and/or the number $k$ of them are chosen adaptively 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 $N : F \to Y$ with range $Y \subset \cup_{n=0}^{\infty} \mathbb{R}^n$. More specifically, let $\Lambda \subset F^*$ be a set of permissible information functionals. Then *nonadaptive* information is a mapping $N : F \to \mathbb{R}^n$,

$$Nf \;=\; (\, L_1 f,\, L_2 f,\, \ldots,\, L_n f \,)$$

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

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

where $L_j(\cdot; t_1, \ldots, t_{j-1}) \in \Lambda$, $\forall t_i$, $1 \le j \le n$. We can also vary the number $n$ of functionals obtaining information of varying cardinality. In this case, we gather information by evaluating the successive $y_j = L_j(f; y_1, \ldots, y_{j-1})$ until the condition $(y_1, \ldots, y_j) \in Y$ is reached. For this to be well defined, we assume that for any infinite sequence $y = (y_1, y_2, y_3, \ldots)$ there exists an index $n$ such that $(y_1, \ldots, y_n) \in Y$. For details and further discussion see, e.g., Traub *et al.* (1988).

Any approximation $A : F \to \Phi$ that is based on some information $N : F \to Y$ can be written as the composition

$$A \;=\; \varphi \circ N$$

where the mapping $\varphi : Y \to \Phi$. 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 $\widetilde{N} : \widetilde{F} \to Y$ with $\widetilde{F} = F \times \mathbb{R}^\infty$. Noisy information about $f$ is given as $y = \widetilde{N}(f, x)$,

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

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

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

See Plaskota (1996) for more details.

# 3 The optimal $k$-term approximation

In this section we assume that information $N : F \to Y$ (adaptive or nonadaptive, with fixed or varying cardinality) is given, and we seek for the best possible choice of $\varphi : Y \to \Phi_k$, 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 information $N$ and the set $\Phi_k$, respectively. We will define them in turn.

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

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

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

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

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

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

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

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

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 $\varphi : Y \to G$ (the approximation is based on information $N$, but the restriction $\varphi(y) \in \Phi$ is relaxed.) That is,

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

Recall that

$$\mathrm{rad}(\nu) \;\le\; \left(\int_F \int_F \|f_1 - f_2\|^q \, \nu(df_1)\nu(df_2)\right)^{1/q} \;\le\; 2\cdot \mathrm{rad}(\nu). \tag{2}$$

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

$$\max\left(r(N), s_k\right) \;\le\; \inf_{\varphi: Y\to \Phi_k} e(\varphi\circ N) \;\le\; 2\cdot \max\left(2r(N), s_k\right).$$

*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 $\psi : F \to \Phi_k$, and with the nondeterministic component $g$ which is chosen randomly according to the conditional distribution $\mu(\cdot|y)$ on $F$. Using the decomposition of $\mu$ with respect to information $y$, the inequality $(a+b)^q \le 2^{q-1}(a^q+b^q)$ (for $a, b > 0$), and (**??**), the error of such an approximation satisfies

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

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

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

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

$$e(\varphi^* \circ N) \;\le\; 2 \cdot \max\left(2r(N), e(\psi)\right).$$

To complete the proof, it suffices to minimize this with respect to $\psi$. □

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

Recall first that $c$ is a center of a measure $\nu$ iff

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

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

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

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

$$\max\left(r(N), \overline{s}_k(N)/2\right) \;\le\; \inf_{\varphi : Y \to \Phi_k} e(\varphi \circ N) \;\le\; 2 \cdot \max\left(r(N), \overline{s}_k(N)\right).$$

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

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

Hence, if $\overline{s}_k(N) \le 2r(N)$ then $e(\varphi \circ N) \ge r(N) = \max(r(N), \overline{s}_k(N))$, and if $\overline{s}_k(N) \ge 2r(N)$ then

$$\begin{aligned} e(\varphi \circ N) &\ge \left(2^{1-q} \, \overline{s}_k(N)^q \;-\; (\overline{s}_k(N)/2)^q\right)^{1/q} \\ &= \overline{s}_k(N)/2 \\ &= \max\left(r(N), \overline{s}_k(N)\right). \end{aligned}$$

This yields

$$\inf_{\varphi: Y \to \Phi_k} e(\varphi \circ N) \;\geq\; \max\left(r(N), \overline{s}_k(N)/2\right).$$

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

$$e(\varphi_\eta \circ N) \;\leq\; \overline{s}_k(N) \,+\, \eta, \tag{3}$$

where $\eta > 0$. Then

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

Since $\eta$ can be arbitrarily small, this gives the upper bound. □

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

$$\|c(y) - \varphi^c(y)\| \;=\; \inf_{\tilde{f} \in \Phi_k} \|c(y) - \tilde{f}\|. \tag{4}$$

Then Theorem **??** immediately yields

**Corollary 1.** *We have*

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

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

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

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

Let $F = \mathbb{R}^2$ with the measure $\mu$ concentrated in 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 $\ell_1$-norm. Then the center of $\mu$ is $c = (1, a)$ and $\mathrm{rad}(\mu) = p_1$. 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 $\varphi^c$ is optimal. Suppose that $\mu$ is a zero-mean Gaussian measure on $F$. Then, for any information $N : F \to Y$, the conditional distribution $\mu(\cdot|y)$ (with $y = Nf$) is also Gaussian. Furthermore, $\mu(\cdot|y)$

is symmetric about its mean $m(y)$, 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 $\varphi : Y \to \Phi_k$ we have

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

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

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

$$r(\widetilde{N}) \;=\; \left( \int_Y (\mathrm{rad}(\mu(\cdot|y))^q \mu_{\widetilde{N}}(dy) \right)^{1/q}.$$

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

**Example 1.** Let $\mu$ be a zero mean Gaussian measure on $F$ with symmetric and positive definite covariance operator $C_\mu : F^* \to F$. 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 $\xi_j$'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 $y = (L_1 f + x_1, L_2 f + x_2, \ldots, L_n f + x_n)$ 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 $\varphi^c(y) = c(y) \in \mathrm{span}\{\xi_1, \ldots, \xi_n\}$.

# 4 Cost and $\varepsilon$-complexity

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

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

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

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

$$\mathrm{cst}(\psi) = \int_F k(\psi(f))\,\mu(df),$$

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

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

*and*

$$\mathrm{cst}(\varphi \circ N) \ \le \ \beta \cdot \mathrm{cst}(\psi).$$

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

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

and

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

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

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

Then

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

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

and

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

as claimed. □

We now define the *information $\varepsilon$-complexity* as

$$
\mathrm{cmp}^{\mathrm{I}}(\varepsilon) \;=\; \inf\{\, \mathrm{cst}(N) :\; N : F \to Y \text{ s.t. there is } \varphi : Y \to G \text{ with } e(\varphi \circ N) \leq \varepsilon \,\},
$$

and the *approximation $\varepsilon$-complexity* as

$$
\mathrm{cmp}^{\mathrm{A}}(\varepsilon) \;=\; \inf\{\, \mathrm{cst}(\psi) :\; \psi : F \to \Phi \text{ s.t. } e(\psi) \leq \varepsilon \,\}.
$$

By Lemma **??** we have the following.

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

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

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

Thus the minimal cost of $\varepsilon$-approximation depends on $\mathrm{cmp}^{\mathrm{I}}(\varepsilon)$ and $\mathrm{cmp}^{\mathrm{A}}(\varepsilon)$. 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 **??**, these questions can also be answered by studying $\mathrm{cmp}^{\mathrm{I}}(\varepsilon)$

and $\mathrm{cmp}^{\mathrm{A}}(\varepsilon)$. That is, if $\mathrm{cmp}^{\mathrm{I}}(\varepsilon)$ is attained by nonadaptive information with the fixed number of functionals then $N^*$ is also nonadaptive and uses the fixed number of functionals. Similarly, if $\mathrm{cmp}^{\mathrm{A}}(\varepsilon)$ is attained by approximation with the fixed number of terms then $A^*$ also uses the fixed number of terms.

**Remark 5.** The problem of whether adaption helps for information complexity has already been studied, see, e.g., Wasilkowski (1986) and Traub et al. (1988) (and Plaskota (1996) for noisy information) where general conditions for adaptive information not to help in case of Gaussian $\mu$ are given. The corresponding question for approximation complexity seems not to have been studied yet.

Theorem **??** 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 the approximation does not help.

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

$$k(\psi^*(f)) \;\le\; k^* \;\le C_3 \cdot \mathrm{cst}(\psi^*), \tag{5}$$

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

$$e(\varphi^c \circ N^*) \;\le\; \varepsilon \quad \text{and} \quad \mathrm{cst}(\varphi^c \circ N^*) \;\le\; C_2 C_3 \cdot \mathrm{cmp}^{\mathrm{A}}(\varepsilon/8),$$

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

**Remark 6.** Results of this section hold also for noisy information $\widetilde{N}$ defined in Remark **??**, with obvious modifications of $\mathrm{cst}(\widetilde{N})$ and $\mathrm{cmp}^{\mathrm{I}}(\varepsilon)$.

# 5 Piecewise polynomial approximation

We consider the following approximation problem. Let

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

with the norm $\|f\|_r = \sup_{0\le t\le 1} |f^{(r)}(t)|$. 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 $s$, $s \ge r$, with finitely many break points that can possibly be chosen adaptively. Specifically, we have $G = C([0,1])$, i.e., the error is measured in Chebyshev norm, 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,

$$y \;=\; (\, f(t_1), f(t_2), \ldots, f(t_n)\,)\,,$$

$0 \le x_1 \le x_2 \le \cdots \le x_n \le 1$. Information can in general be adaptive.

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

$$\mathrm{cmp}^{\mathrm{I}}(\varepsilon) \;\asymp\; \left(\frac{\ln^{1/2}(1/\varepsilon)}{\varepsilon}\right)^{\frac{1}{r+1/2}},$$

see, e.g., Maiorov and Wasilkowski (1996). 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 can also be obtained (up to a constant) by a piecewise polynomial interpolation of degree $r$.

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

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

It turns out, however, that the log factor above can be removed.

**Theorem 4.** *We have*

$$\mathrm{cmp}^{\mathrm{A}}(\varepsilon) \;\asymp\; \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1/2}}.$$

Theorem **??** will be proven in several steps. First we show the lower bound on $\mathrm{cmp}^{\mathrm{A}}(\varepsilon)$.

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

$$\mathrm{cmp}^{\mathrm{A}}(\varepsilon) \;\ge\; C_1 \cdot \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1/2}}$$

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

*Proof.* We first show that $\lim_{\varepsilon\to 0} \mathrm{cmp}^{\mathrm{A}}(\varepsilon) = \infty$, so that $\varepsilon_0$ exists. Indeed, otherwise an $l$ would exist such that for arbitrarily small $\varepsilon$ there is an approximation $\psi_\varepsilon$ with

$e(\psi_\varepsilon) \le \varepsilon$ and $\mathrm{cst}(\psi) \le l$. Denote $A = \{\, f :\ k(\psi_\varepsilon(f)) \le sl \,\}$ and define $\psi'_\varepsilon(f) = \psi_\varepsilon(f)$ for $f \in A$, and $\psi'_\varepsilon(f) = 0$ for $f \notin A$. Then $\psi'_\varepsilon$ is an $l$-term approximation with error

$$\begin{aligned} e(\psi'_\varepsilon) &= \int_A \|f - \psi_\varepsilon(f)\|\, w_r(df) \,+\, \int_{F_r \setminus A} \|f\|\, w_r(df) \\ &\le \varepsilon \,+\, \int_{F_r \setminus B_\varepsilon} \|f\|\, w_r(df), \end{aligned}$$

where $B_\varepsilon$ is the ball centered at the origin with $w_r(B_\varepsilon) = w_r(A)$. Since, in addition, $w_r(A) \ge 1 - 1/s$, we have $\lim_{\varepsilon\to 0} e(\psi'_\varepsilon) = 0$ and $s_l = 0$. This is however impossible, because the set of $l$-term approximations is not dense in $F_r$.

Consider now two separate problems of approximating $f$ with the average error $\varepsilon$ on the sub-intervals $[0, 1/2]$ and $[1/2, 1]$, instead of on the whole interval $[0, 1]$. Denote the corresponding average complexities of these problems as $\mathrm{cmp}^{\mathrm{A}}{}_0(\varepsilon)$ and $\mathrm{cmp}^{\mathrm{A}}{}_1(\varepsilon)$. Then

$$\mathrm{cmp}^{\mathrm{A}}{}_1(\varepsilon) \,+\, 1 \;\ge\; \mathrm{cmp}^{\mathrm{A}}{}_0(\varepsilon) \;=\; \mathrm{cmp}^{\mathrm{A}}(2^{r+1/2}\varepsilon). \tag{6}$$

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 $s \ge r$.) 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 $\psi$ is an approximation with the average error $\varepsilon$. Since the error is measured in the uniform norm, the average error of $\psi$ on each of the two sub-intervals $[0, 1/2]$ and $[1/2, 1]$ is also at most $\varepsilon$. Hence $\psi$ divides each of the sub-intervals into at least $\mathrm{cmp}^{\mathrm{A}}{}_0(\varepsilon)$ pieces (on average). Noting that two of the pieces may overlap and using (**??**) we obtain

$$\begin{aligned} \mathrm{cmp}^{\mathrm{A}}(\varepsilon) &\ge \mathrm{cmp}^{\mathrm{A}}{}_0(\varepsilon) \,+\, \mathrm{cmp}^{\mathrm{A}}{}_1(\varepsilon) \,-\, 1 \\ &\ge 2 \cdot \Big( \mathrm{cmp}^{\mathrm{A}}(2^{r+1/2}\varepsilon) \,-\, 1 \Big). \end{aligned}$$

Proceeding inductively, we finally arrive at the estimate

$$\mathrm{cmp}^{\mathrm{A}}(\varepsilon) \;\ge\; 2^s \,(\mathrm{cmp}^{\mathrm{A}}(\varepsilon_0) - 2),$$

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

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

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

$$(t_j - t_{j-1})^r \max_{t_{j-1} \le \xi_1, \xi_2 \le t_j} |f^{(r)}(\xi_1) - f^{(r)}(\xi_2)| \;=\; \varepsilon \cdot r! \tag{7}$$

for $1 \le j \le k-1$, and we have inequality "$\le$" above for $j = k$. Then we approximate $f$ by a continuous piecewise polynomial $\tilde{f} = \psi_\varepsilon(f)$ of degree $r$ with knots $t_j$, 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 $\psi_\varepsilon$ is an adaptive approximation with a varying number of terms.

**Lemma 3.** *For all $f \in F_r$ we have $\|f - \psi_\varepsilon\| \le \varepsilon$. Furthermore, $\mathrm{cst}(\psi_\varepsilon) < \infty$, and for any $0 < \varepsilon < \varepsilon_0$ we have*

$$\mathrm{cst}(\psi_\varepsilon) \;\le\; C_2 \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r+1/2}}$$

*where $C_2 = C_2(\varepsilon_0) = 2\,\mathrm{cst}(\psi_{\varepsilon_0})\varepsilon_0^{1/(r+1/2)}$.*

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

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

Hence the error is always at most $\varepsilon$.

We now show that $\mathrm{cst}(\psi_\varepsilon) < \infty$. Let $a = \mathrm{Prob}\{k(\psi_\varepsilon(f)) \ge 2\}$. 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 $\mathrm{Prob}\{k(\psi_\varepsilon(f)) \ge 3\} \le a^2$ and, generally, $\mathrm{Prob}\{k(\psi_\varepsilon(f)) \ge s\} \le a^{s-1}$. Hence

$$\mathrm{cst}(\psi_\varepsilon) \;\le\; (1-a) + \sum_{j=2}^{\infty} j a^{j-1} \;<\; \infty,$$

as claimed.

To get the upper bound for $\mathrm{cst}(\psi_\varepsilon)$, 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

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

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

$$\mathrm{cst}(\psi_\varepsilon) \;\le\; 2 \cdot \mathrm{cst}(\psi_{2^{r+1/2}\varepsilon}),$$

or more generally

$$\mathrm{cst}(\psi_\varepsilon) \;\le\; 2^s \cdot \mathrm{cst}(\psi_{\varepsilon_0}),$$

where $0 < \varepsilon < \varepsilon_0$ and $s$ is the smallest integer for which $2^{-s(r+1/2)}\varepsilon_0 \le \varepsilon$. The last inequality gives the desired result. □

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

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

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

**Lemma 4.** *The average error of $\psi_k$ satisfies*

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

*Proof.* As in the proof of Lemma **??** we show that the error for given $f$ can be estimated from above by

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

with any $1 \le j \le k$. Hence it suffices to show that the average value of $\varepsilon_k(f)$ is $O(k^{-(r+1/2)})$.

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

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

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

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

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 $\psi_k$ can be estimated as ($E$ stands for expectation)

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

The first component in the last expression vanishes, since

$$\lim_{\eta \to 0^+} F_k(\eta) \eta^{-(r+1/2)} \;=\; 0. \tag{8}$$

Indeed,

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

Since last probability tends to zero exponentially fast as $\eta \to 0$, see, e.g., Billingsley (1968), (**??**) 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. □

Theorem **??** 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 $\psi_k^{non}$ 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 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

$$X \;=\; \|f - w_f\|_C. \tag{9}$$

Then there is $\sigma > 0$ such that

$$\text{Prob}\,(X \geq \alpha) \;\geq\; \sqrt{\frac{2}{\pi\sigma^2}} \int_\alpha^\infty e^{-z^2/(2\sigma^2)} dz. \tag{10}$$

Indeed, let $0 \leq t_0 < \cdots < t_{s+1} \leq 1$ be $s+2$ arbitrary 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.,

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

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 $f(t_0, \ldots, t_{s+1})$ is a nonzero linear functional of the $f(t_j)$s, it is a zero mean Gaussian random variable with a positive variance $\sigma$, and (**??**) follows.

We now use (**??**) 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 $\psi_k^{non}$. 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 (**??**). Hence, by (**??**) we obtain

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

as claimed. □

Summarizing, the cheapest information-based nonlinear approximation $A_\varepsilon : F \to \Phi$ with error $\varepsilon$ 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 $y \;=\; (\, f(1/n),\, f(2/n), \ldots, f(1)\,)$.

3. Find the natural spline $s_y$ of degree $2r+1$ interpolating the data, $s_y(j/n) = y_j$, $0 \le j \le n$ ($y_0 = 0$).

4. Apply the adaptive $k$-term approximation $\psi_k$ of Lemma **??** on $s_y$, i.e., $A_\varepsilon(f) = \psi_k(s_y)$.

**Remark 7.** We saw 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

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

Furthermore, the cheapest approximation $A_\varepsilon$ is in this case obtained as follows.

1. Choose $n = n(\varepsilon) \asymp (\varepsilon^{-1} \ln^{1/2}(1/\varepsilon))^{(4r+2)/(4r+1)}$ and $k = k(\varepsilon) \asymp \varepsilon^{-(r+1/2)}$.

2. Observe $y = (\, f(1/n) + x_1,\, f(2/n) + x_2, \ldots, f(1) + x_n \,)$.

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 $\psi_k$ of Lemma **??** on $s_{\sigma,y}$, i.e., $A_\varepsilon(f, x) = \psi_k(s_{\sigma,y})$.

For details, see Plaskota (1998).

# References


1. Billingsley, P. (1968): *Convergence of Probability Measures.* Wiley, New York.
2. Kon, M.A., & Plaskota, L. (2000a): Information complexity of neural networks. *Neural Networks* **13**, pp.365–376.
3. Kon, M.A., & Plaskota, L. (2000b): Complexity of neural network approximation with limited information: a worst case approach. *J. of Complexity* **17**, pp.345–365.
4. Maiorov, V.E. & Wasilkowski, G.W. (1996): Probabilistic and average linear widths in $L_\infty$ norm with respect to $r$-fold Wiener measure. *J. of Approx. Theory* **84**, pp 31–40.
5. Plaskota, L. (1996): *Noisy information and computational complexity.* Cambridge University Press, Cambridge.
6. Plaskota, L. (1998): Average case $L_\infty$ approximation in the presence of Gaussian noise. *J. of Approx. Theory* **93**, pp.501–515.

7. Traub, J.F., Wasilkowski, G.W., & Woźniakowski, H. (1988): *Information-based Complexity*. Academic Press, New York.
8. De Vore, R.A. (1998): Nonlinear approximation, *Acta Numerica* **8**, pp.51–150.
9. Wasilkowski, G.W. (1986): Information of varying cardinality. *J. of Complexity* **2**, pp.204–228.

**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