CUCS-31-82

Average Case Optimality for Linear Problems

J.F. Traub
Departments of Computer Science and Mathematics
Columbia University
New York, New York

G. Wasilkowski
Institute of Informatics
University of Warsaw
Warsaw, Poland

H. Woźniakowski
Institute of Informatics
University of Warsaw
Warsaw, Poland

and

Department of Computer Science
Columbia University
New York, New York

September 1981

Revised August 1982

This research was supported in part by the National Science Foundation under Grant MCS-7823676. J.F. Traub wishes to acknowledge the support of the Advanced Research Projects Agency under Contract N00039-82-C-0427.

Abstract

We introduce an average case model and define general notions of optimal algorithm and optimal information. We prove that the same algorithm and information are optimal in the worst and average cases and that adaptive information is not more powerful than non-adaptive information.

CONTENTS

1. Introduction

In two recent monographs (Traub and Woźniakowski [80], Traub, Wasilkowski, and Woźniakowski [83]) we studied optimal reduction of uncertainty for a worst case model. With this paper we initiate a corresponding study for an average case model. This is the first of a number of papers reporting average case results. These results will eventually appear as part of a third volume devoted to the study of various probabilistic settings.

We indicate earlier work on this subject. Suldin ([59],[60]) studied average case error for the integration problem. Larkin, in a series of pioneering papers commencing with [72], studied optimal algorithms, mostly for linear problems, utilizing a Gaussian measure. Both Suldin and Larkin confine themselves to linear algorithms.

In this initial paper we confine ourselves to linear problems in a finite dimensional space. (Average case analysis for an infinite dimensional setting is studied in Wasilkowski and Woźniakowski [82a].) By a linear problem we mean a problem specified by a linear operator. Examples of linear operators are integration, interpolation, and approximation. Note that the solution of a linear system is not a linear problem since the solution does not depend linearly on the matrix element.

We restrict ourselves to the finite dimensional setting for two reasons.

1. This setting is of intrinsic interest.
2. The analysis of the infinite dimensional setting requires rather heavy mathematical machinery. In order to permit the reader to focus on the model assumptions and the results we avoid these mathematical complications in this first paper.

In this paper we specify an average case model and introduce general notions of optimal algorithm and optimal information. The following results are obtained.

1. The same algorithm is optimal in the worst and average cases.
2. The same information is optimal in the worst and average cases.
3. Adaptive information is not more powerful than nonadaptive information.

We discuss these results. Conclusions 1 and 2 are favorable to the user since the same algorithm with the same information minimizes both the worst and average error. It was established (see Traub and Woźniakowski [80, p.49] for a history) that adaptive information does not help for the worst case. Many researchers believe that this is only true in the worst case setting. We prove the counterintuitive result that adaption doesn't help even on the average.

We illustrate some of the basic concepts of this paper by

Example 1.1

Assume we wish to approximate the function f knowing some information $N(f)$ and knowing that f belongs to some given class of functions F. To be specific let $N(f) = [f(t_1),\ldots,f(t_n)]$ consist of n function samples and let F be the class of trigonometric polynomials of degree m whose r th derivative is bounded by unity.

An algorithm ϕ is any mapping acting on the information $N(f)$. An example of an algorithm is the linear algorithm

$\phi(N(f)) = \sum_{i=1}^{n} f(t_i)\,\alpha_i$ where α_i are some functions. An *algorithm is optimal* if it minimizes the error according to some error criterion. In the worst case setting the error is defined as the largest error for all f in F. In the average case setting the error is defined in terms of the L_2 norm with respect to some measure on F.

Next, assume the t_i may be varied. We say that the *information* is *optimal* if the t_i are chosen so as to minimize the worst or average case error of the optimal algorithm.

If the t_i are given independently of f, then the information is called *nonadaptive*. On a parallel computer nonadaptive information can be computed simultaneously. If the t_i depend on previously computed values of f, the information is called *adaptive*. One might hope that choosing points adaptively decreases the error. However, adaption does not help for either the worst or average case. This example will be continued in Section 8. □

We briefly summarize the contents of this paper. In Section 2 we outline the setting and results of the worst case model which we shall constrast with the results of this paper. In Section 3 we introduce an average case model and prove that the same algorithm is optimal for both the worst and average case. Very simple and elegant formulas for the worst and average radii of information are given by Theorem 3.2. In the following section the problem of optimal average information is posed and solved. The same information is optimal for both the worst and average cases. In Section 5 we show that adaptive information is no more powerful than nonadaptive information in either model.

In Section 6 we compare the intrinsic uncertainty if only the problem setting is known, with the uncertainty when n optimal evaluations are used. In Section 7 we obtain very tight complexity bounds and prove that the same algorithm enjoys nearly optimal complexity in both models. In the concluding section an example illustrates the models and some of the results.

2. Worst Case Model: Optimal Algorithms

To help the reader we begin with the relatively simple worst case model and pass next to an average case model. We summarize the setting and main results of the worst case model for a (simplified) linear problem studied in general by Traub, Wasilkowski, and Woźniakowski [83], see especially Appendix E, and Traub and Woźniakowski [80]. Although we use the terminology and notation presented there, the following account is self-contained.

Let F_1 be a finite dimensional real space and let

$$m = \dim (F_1). \tag{2.1}$$

Let F_2 be a real Hilbert space. Consider the linear operator

$$S: F_1 \rightarrow F_2. \tag{2.2}$$

The operator S is called the solution operator.

Our aim is to find an element $x = x(f)$ which approximates Sf according to some error criterion. There are many error criteria of practical importance some of which we cite here. The absolute error criterion is such that $\| S(f) - x(f) \| \le \varepsilon$ for a given nonnegative ε. The relative error criterion is such that $\| S(f) - x(f) \| / \| S(f) \| \le \varepsilon$. The absolute-relative error criterion is such that $\| S(f) - x(f) \| / (\| S(f) \| + \eta) \le \varepsilon$ with a given positive η.

Sometimes we will want to satisfy the error criterion for f from the whole space F_1 , and sometimes for only a subset of F_1. This subset can be characterized, for instance, by the condition

$\| Tf \| \leq 1$ for some operator T.

We now present a general error criterion which will include the above examples as special cases. We have chosen a formulation which will also be used for the average case. Let

(2.3) $$T : F_1 \to F_4$$

be a one-to-one linear operator where $F_4 = T(F_1)$ is a Hilbert space. We call this space F_4 (rather than F_3) to conform to the usage in Traub and Woźniakowski [80]. Let

(2.4) $$\rho : \mathbb{R}_+ \to \mathbb{R}_+$$

be a given function.

We say that an element x of F_2 is an ε-approximation to Sf iff

(2.5) $$\| Sf - x \| \, \rho(\| Tf \|) \leq \varepsilon$$

where ε is a nonnegative number.

Observe that for $\rho(x) \equiv 1$, (2.5) becomes the absolute error criterion. For $\rho(x) = 1/x$ and $T = S$, (2.5) becomes the relative error criterion. If $\rho(x) = 1/(x+\eta)$, $\eta > 0$ and $T = S$ then (2.5) becomes the absolute-relative error criterion. If $\rho(x) = 1$ for $x \leq 1$ and $\rho(x) = 0$ for $x > 1$, (2.5) becomes the absolute error criterion for elements f for which $\| Tf \| \leq 1$.

Our aim is to find an ε-approximation to Sf for all f from F_1. To find an ε-approximation, information on f is required. We assume that we know $N(f)$ where N is a linear operator. Without loss of generality we can assume that N has the form

$$N(f) = [L_1(f), L_2(f), \ldots, L_n(f)] \tag{2.6}$$

where $L_1, L_2, \ldots, L_n$ are linearly independent linear functionals and $n < m$. We say N is a (partial) information operator and n is the cardinality of N.

Since $n < m$ then there exist infinitely many elements $\tilde{f}$ from F_1 which are indistinguishable with respect to $N(f)$. (Hence N is called partial.) It is therefore impossible to recognize which element $S(\tilde{f})$ is to be approximated. Let

$$V(N,y) = \{\tilde{f} \in F_1 : N(\tilde{f}) = y\}, \quad y = N(f), \tag{2.7}$$

be the set of indistinguishable elements.

We seek an ε-approximation x of the form $x = \phi(N(f))$ where ϕ is a mapping,

$$\phi : N(F_1) \to F_2. \tag{2.8}$$

Note that $\phi(N(f))$ has to satisfy (2.5) for all $\tilde{f}$ from $V(N,y)$.

We call ϕ an (idealized) algorithm. Let $\Phi(N)$ be the class of all (idealized) algorithms, i.e., $\Phi(N)$ consists of all mappings ϕ, defined by (2.8), which use the information operator N.

We stress that our definition of algorithm is extremely general. In spite of this we can prove some negative results. This makes the negative results even stronger. If one wishes to carry out a computation, then in general the class of algorithms must be restricted. We shall see that for the problem studied in this paper, algorithms which are "optimal" in the class of idealized algorithms are relatively easy to implement in actual computation.

Let ϕ be an algorithm, $\phi \epsilon \Phi(N)$. Then

$$(2.9) \qquad e(\phi,N) = \sup_{f \epsilon F_1} \| Sf - \phi(N(f)) \| \rho(\| Tf \|)$$

is called the error of ϕ.

Note that the error of ϕ is defined as its error for the "hardest" f. That is why this model is called the worst case model. For the average case model studied in the following sections we replace the sup in (2.9) by an integral which measures the average performance of ϕ.

From (2.9) it follows that $\phi(N(f))$ is an ε-approximation to Sf for all f iff $e(\phi,N) \le \varepsilon$.

Definition 2.1

We shall say $r(N)$ is the radius of information iff

$$(2.10) \qquad r(N) = \inf_{\phi \epsilon \Phi(N)} e(\phi, N) .$$

We shall say an algorithm ϕ, $\phi \epsilon \Phi(N)$, is an optimal error algorithm iff

$$(2.11) \qquad e(\phi,N) = r(N).$$ □

Remark 2.1

The radius of information can be defined independently of the concept of algorithm and (2.10) can then be established; see the books quoted at the beginning of this section. For simplicity we here present (2.10) as the definition of radius. □

Equation (2.10) implies that we can find an ε-approximation iff $r(N) \le \varepsilon$. If $r(N) \le \varepsilon$ then an optimal error algorithm supplies an ε-approximation.

We now present a spline algorithm ϕ^s (see Traub and Woźniakowski [80, Chapter 4]) and prove that it is an optimal error algorithm.

Let $\sigma = \sigma(y)$ be an element of F_1 such that

$$N(\sigma) = y \tag{2.12}$$
$$\|T\sigma\| = \min\{\|T\tilde{f}\|: \tilde{f} \in V(N,y)\}.$$

It is obvious that such an element exists and is unique. The element $\sigma(y)$ is called a spline interpolating y. The spline algorithm ϕ^s is defined as

$$\phi^s(y) = S\,\sigma(y)\ , \quad \forall y \in N(F_1) = \mathbb{R}^n . \tag{2.13}$$

Since S is linear and σ depends linearly on y, the spline algorithm ϕ^s is a linear algorithm. Thus

$$\phi^s(y) = \sum_{i=1}^{n} L_i(f)\ S\sigma_i \tag{2.14}$$

where $y = N(f) = [L_1(f), \ldots, L_n(f)]$ and $\sigma_i = \sigma([0, \ldots, \underset{i}{1}, \ldots, 0])$. The evaluation of $\phi^s(y)$ requires the knowledge of $S\sigma_1, \ldots, S\sigma_n$. Computing the $S\sigma_i$ can be difficult, but since they are independent of y, this need be done only once and the cost of computing them may be viewed as a precomputation cost. Then to compute $\phi^s(y)$ it is enough to perform n multiplications of a real number by a m dimensional vector and $n-1$ additions of m-dimensional vectors. Hence if the $S\sigma_i$ are precomputed, then the evaluation of $\phi^s(y)$ requires at most nm scalar multiplications and $(n-1)m$ scalar additions.

The spline algorithm ϕ^s enjoys very strong optimal error properties one of which is stated in

Theorem 2.1

The spline algorithm ϕ^s is an optimal error algorithm and

$$e(\phi^s,N) = r(N) = \sup_{x\geq 0} x\rho(x) \sup_{h\in \ker N} \|Sh\| / \|Th\| \tag{2.15}$$

with the convention $0\cdot\infty = 0$. □

Proof

This result is established for a more general problem in Traub, Wasilkowski, and Woźniakowski [83], see Theorem E.1. For the simplified linear problem of this section we supply a short proof.

Let $\tilde{f} = \sigma(y) \pm h$ where $h \in \ker N$. Then $\tilde{f} \in V(N,y)$ and $(T\sigma(y), Th) = 0$. We have

$$e(\phi,N) = \sup_{y} \sup_{\tilde{f}\in V(N,y)} \| S(\tilde{f}) - \phi(y)\| \rho(\|T\tilde{f}\|) =$$

$$= \sup_{y} \sup_{h\in \ker N} \| S\sigma(y) \pm Sh - \phi(y)\| \rho(\sqrt{\|Th\|^2 + \|T\sigma(y)\|^2}).$$

Since

$$\max(\| S\sigma(y) + Sh - \phi(y) \| , \| S\sigma(y) - Sh - \phi(y) \|) \geq \| Sh \|$$

for any $\phi(y)$, we have

$$e(\phi,N) \geq \sup_{y} \sup_{h \in \ker N} \| Sh \| \, \rho(\sqrt{\| Th \|^2 + \| T\sigma(y) \|^2}) = e(\phi^S,N).$$

This proves optimality of ϕ^S. Observe that .

$$e(\phi^S,N) = \sup_{y} \sup_{x \geq 0} \rho(\sqrt{x^2 + \| T\sigma(y) \|^2}) \cdot$$

$$\sup\{ \| Sh \| : h \in \ker N, \| Th \| = x \} =$$

$$\sup_{y} \sup_{x \geq 0} x\rho(\sqrt{x^2 + \| T\sigma(y) \|^2}) \sup_{h \in \ker N} \| Sh \| / \| Th \| =$$

$$\sup_{x \geq 0} x \, \rho(x) \sup_{h \in \ker N} \| Sh \| / \| Th \|$$

which proves (2.15) and completes the proof. □

Remark 2.2

The space F_2 need not be a Hilbert space and the spaces F_1, F_2 and F_4 need not be finite dimensional in Theorem 2.1. In fact this theorem holds for any normed linear space F_2 and any Hilbert space F_4 , assuming that T(kerN) is closed. The assumption that F_2 is a Hilbert space and both F_2 and F_4 are finite dimensional will be used in the next sections. For simplicity of presentation we assume, even in this section, that F_2 is a Hilbert space and F_2 and F_4 are finite dimensional. □

3. Average Case Model: Optimal Algorithms

We introduce an average case model, and pose and solve the problem of optimal algorithms in this model. We prove that the spline algorithm defined by (2.13) is also optimal for the average case model. We find its error and compare with the worst case model.

We begin by defining a probability measure on F_1. Without loss of generality assume that $F_1 = \mathbb{R}^m$. Let $\mathbb{B}$ be a σ-field of Borel sets in $\mathbb{R}^m$. By

(3.1) $$\int_{\mathbb{R}^m} \cdot \, df = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \cdot \, df_m \; df_{m-1} \cdots df_1$$

we mean the Lebesgue integral, $f = [f_1, f_2, \ldots, f_m]$.

Let w, $w: \mathbb{R}_+ \to \mathbb{R}_+$, be a function such that

(3.2) $$\int_{\mathbb{R}^m} w(\| Tf \|) df = 1$$

The function w is a scalar weight function. Note that $\| \cdot \|$ in (3.2) denotes the norm in the Hilbert space F_4. Let A be a Borel set in F_1, $A \in \mathbb{B}$. We define a measure μ on F_1 as

(3.3) $$\mu(A) = \int_A w(\| Tf \|) df \; .$$

Note that μ is a probability measure, i.e.,

(3.4) $$\mu(F_1) = 1.$$

The measure μ generates the Lebesgue integral in F_1. This integral is denoted by $\int_A \cdot \, \mu(df)$. Thus if $g: F_1 \to \mathbb{R}$ then

$$(3.5) \qquad \int_A g(f)\ \mu(df) \overset{df}{=} \int_A g(f)\ w(\| Tf \|)df.$$

Remark 3.1

It may seem somewhat arbitrary to restrict ourselves to measures defined as in (3.3). However it is shown by Woźniakowski [82] that any measure which enjoys a certain orthogonality invariance property must be of form (3.3).

The use of orthogonal invariance is also discussed by Micchelli [82].

Remark 3.2

The operator T plays two roles in our setting. It is used with the function ρ in (2.5) to define an ε-approximation and it is used with the function w in (3.3) to define a probability measure on F_1.

Although we could analyze a more general setting with different operators in (2.5) and (3.3), we shall use only one operator to simplify our analysis and, more importantly, to show that the same (spline) algorithm is optimal for both models.

We are ready to define the average error of an algorithm ϕ.

Definition 3.1

Let $\phi \in \Phi(N)$. We shall say $e^{avg}(\phi,N)$ is the average error of ϕ iff

$$(3.6) \qquad e^{avg}(\phi,N) = \{\int_{F_1} \| S(f) - \phi(N(f))\|^2 \rho^2(\| Tf \|)\ \mu(df)\}^{\frac{1}{2}}. \qquad \square$$

Thus the squared average error of ϕ is defined as the average value of $\| S(f) - \phi(N(f)) \|^2 \rho^2(\| Tf \|)$. Recall that the worst error of ϕ is defined as $\| S(f) - \phi(N(f))\| \ \rho(\| Tf \|)$ for a worst f. Since $\| S(f) - \phi(N(f)) \|^2 \rho^2(\| Tf \|) \le \sup_{f \in F_1} \| S(f) - \phi(N(f))\|^2 \rho^2(\| Tf \|)$

and $\int_{F_1} \mu(df) = 1$ then

$$e^{avg}(\phi,N) \le e(\phi,N). \tag{3.7}$$

This verifies the expected condition that the average error of ϕ does not exceed the (worst case) error of ϕ.

We comment on Definition 3.1.

Remark 3.3

The average error is defined only for algorithms ϕ such that $\| S(f) - \phi(N(f)) \|^2 \rho^2(\| Tf \|)$ is a measurable function of f, i.e., the integral in (3.6) exists. It is possible to define the average error for an arbitrary algorithm by using the concept of local average errors, see Wasilkowski and Woźniakowski [82b]. For simplicity we restrict the class $\Phi(N)$ to algorithms with well-defined average errors. □

Remark 3.4

One may also study the p-th average error defined as

$$e_p^{avg}(\phi,N) = \{\int_{F_1} \| S(f) - \phi(N(f)) \|^p \rho^p(\| Tf \|) \mu(df)\}^{1/p}$$

for some $p \in [1,\infty]$. Note that for $p = 2$, $e_2^{avg}(\phi,N)$ coincides with $e^{avg}(\phi,N)$. We have chosen $p = 2$ to avoid technical difficulties and not to distract the reader from the main model assumptions of this paper. For $p = 1$ we have the expected value of $\| S(f) - \phi(N(f)) \| \rho(\| Tf \|)$ whereas if p tends to infinity then

$$e_{\infty}^{avg}(\phi,N) = \operatorname*{ess\,sup}_{f\in F} \| S(f) - \phi(N(f)) \| \, c(\| Tf \|).$$

This coincides with the worst case model modulo sets of measure zero. □

As in Definition 2.1 we now introduce the average radius of information and an optimal average error algorithm.

<u>Definition 3.2</u>

We shall say $r^{avg}(N)$ is the <u>average radius of information</u> iff

$$(3.8) \qquad r^{avg}(N) = \inf_{\phi\in\Phi(N)} e^{avg}(\phi,N) \;.$$

We shall say an algorithm ϕ, $\phi \in \Phi(N)$, is an <u>optimal average error algorithm</u> iff

$$(3.9) \qquad e^{avg}(\phi,N) = r^{avg}(N).$$

□

Thus, we can find an ε-approximation with average error not exceeding ε iff $r^{avg}(N) \le \varepsilon$. If $r^{avg}(N) \le \varepsilon$ then an optimal average error algorithm supplies such an ε-approximation.

We are now ready to prove that the spline algorithm, see (2.13), has minimal average error. Let $\{a_1,a_2,\ldots,a_m\}$ be an orthonormal basis of F_4 such that

$$(3.10) \qquad \begin{aligned} T(\ker N)^{\perp} &= \operatorname{lin}\{a_1,a_2,\ldots,a_n\}, \\ T(\ker N) &= \operatorname{lin}\{a_{n+1},a_{n+2},\ldots,a_m\}. \end{aligned}$$

We say two algorithms, ϕ_1 and ϕ_2 , are equal iff

$$\| Tf \|) = 0 \}) = 1 \;.$$

Theorem 3.1

The spline algorithm ϕ^s is a unique optimal average error algorithm and

$$(3.11)\quad e^{avg}(\phi^s,N) = r^{avg}(N) = \{\int_{F_1} \| Tf \|^2 \rho^2 (\| Tf \|)\, \mu(df)\}^{\frac{1}{2}} \{\frac{1}{m} \sum_{j=n+1}^{m} \| ST^{-1}a_j \|^2\}^{\frac{1}{2}}$$

□

Proof

Let $f \in F_1$. Then $f = \sum_{j=1}^{m} z_j\, T^{-1}a_j$. Note that

$$(3.12)\qquad L_i(f) = \sum_{j=1}^{n} z_j\, L_i(T^{-1}a_j).$$

Define the $n \times n$ matrix M as

$$M = (L_j(T^{-1}a_i))_{i,j=1}^{n}$$

Note that M is nonsingular and

$$(3.13)\qquad y = N(f) = [z_1,z_2,\ldots,z_n]M.$$

Let $\sigma = \sum_{j=1}^{n} z_j\, T^{-1}a_j$. Then (3.12) yields $L_i(\sigma) = L_i(f)$ and $N(\sigma) = N(f)$. Let $h \in \ker N$. Then $Th \in \mathrm{lin}\{a_{n+1},\ldots,a_m\}$ and therefore $(T\sigma,Th) = 0$. Thus σ is a spline interpolating y and

$$(3.14)\qquad \phi^s(N(f)) = S\sigma = \sum_{j=1}^{n} z_j\, ST^{-1}a_j.$$

Take an arbitrary algorithm ϕ from $\Phi(N)$. We change variables in (3.6) by setting

$$(3.15) \qquad f = \sum_{j=1}^{m} z_j \, T^{-1} a_j \, , \quad z = [z_1, z_2, \ldots, z_m].$$

Since $\{a_1, \ldots, a_m\}$ are orthonormal, $\| Tf \| = \| z \| = (\sum_{j=1}^{m} z_j^{\,2})^{½}$ and

$$|\det(T^{-1}a_1, \ldots, T^{-1}a_n)| = |\det(T^{-1})|.$$

Thus $df = |\det(T^{-1})| dz$ and (3.6) can be rewritten due to (3.15), (3.13) and (3.14), as

$$(3.16) \quad e^{avg}(\phi, N)^2 = |\det(T^{-1})| \int_{\mathbb{R}^m} \| \sum_{j=1}^{m} z_j \, ST^{-1} a_j - \phi([z_1, \ldots, z_n]M) \|^2 \rho^2(\| z \|) \; w(\| z \|) dz =$$

$$|\det(T^{-1})| \int_{\mathbb{R}^n} \{ \int_{\mathbb{R}^{m-n}} \| \phi^s([z_1, \ldots, z_n]M) + \sum_{j=n+1}^{m} z_j ST^{-1} a_j - \phi([z_1, \ldots, z_n]M) \|^2 \rho^2((\sum_{j=1}^{m} z_j^{\,2})^{½}) \; w((\sum_{j=1}^{m} z_j^{\,2})^{½}) dz_{n+1} \ldots dz_m \} \; dz_1 \ldots dz_n .$$

Note that in the expression in braces we integrate over all elements indistinguishable from f under N.

We again change variables, setting $z_i^* = z_i$ for $i = 1, 2, \ldots, n$ and $z_i^* = -z_i$ for $i = n+1, \ldots, m$. Then $dz^* = dz$ and

$$(3.17) \quad e^{avg}(\phi, N)^2 = |\det(T^{-1})| \int_{\mathbb{R}^n} \{ \int_{\mathbb{R}^{m-n}} \| \phi^s([z_1^*, \ldots, z_n^*]M) - \sum_{j=n+1}^{m} z_j^* ST^{-1} a_j - \phi([z_1^*, \ldots, z_n^*]M \|^2 \rho^2((\sum_{j=1}^{m} z_j^{*2})^{½}) w((\sum_{i=1}^{m} z_j^{*2})^{½}) dz_{n+1}^* \ldots dz_m^* \} \; dz_1^* \ldots dz_n^* .$$

Dropping the asterisk in (3.17), we add (3.16) and (3.17) getting

$$e^{avg}(\phi,N)^2 = \frac{1}{2}|\det(T^{-1})| \int_{\mathbb{R}^n}\{\int_{\mathbb{R}^{m-n}}\{\| \sum_{j=n+1}^{m} z_j ST^{-1}a_j + \phi^s([z_1,\dots,z_n]M) - \phi([z_1,\dots,z_n]M)\|^2 + \| \sum_{j=n+1}^{m} z_j ST^{-1}a_j - \phi^s([z_1,\dots,z_n]M) + \phi([z_1,\dots,z_n]M)\|^2$$

$$\rho^2(\|z\|)dz_{n+1}\cdots dz_m\}\ dz_1\cdots dz_n.$$

Note that

$$\|g_1+g_2\|^2 + \|g_1-g_2\|^2 = 2(\|g_1\|^2 + \|g_2\|^2),\quad \forall g_1,g_2 \in F_2.$$

Setting $g_1 = \sum_{j=n+1}^{m} z_j ST^{-1}a_j$ and $g_2 = \phi^s(y) - \phi(y)$ we get

$$e^{avg}(\phi,N)^2 = e^{avg}(\phi^s,N)^2 + \int_{F_1} \|\phi^s(N(f)) - \phi(N(f))\|^2 \rho^2(\|Tf\|)w(\|Tf\|)df.$$

This shows that $e^{avg}(\phi,N) \geq e^{avg}(\phi^s,N)$ and
$e^{avg}(\phi,N) = e^{avg}(\phi^s,N)$ iff
$\mu(\{f : \|\phi^s(N(f)) - \phi(N(f))\|\rho(\|Tf\|) = 0\}) = 1$
which means that $\phi^s(N(f))$ and $\phi(N(f))$ are equal.
Hence, ϕ^s is a unique optimal average error
algorithm and $e^{avg}(\phi^s,N) = r^{avg}(N)$.

To prove (3.11) observe that

$$\|\sum_{j=n+1}^{m} z_j ST^{-1}a_j\|^2 = \sum_{j=n+1}^{m} z_j^2 \|ST^{-1}a_j\|^2 + 2\sum_{i<j} z_i z_j (ST^{-1}a_i, ST^{-1}a_j).$$

Since $z_i z_j\ \rho^2(\|z\|)w(\|z\|)$ is odd then

$$\int_{\mathbb{R}^{m-n}} z_i z_j\ \rho^2(\|z\|)w(\|z\|)dz_{n+1}\cdots dz_m = 0,\ \forall i < j,\ i,j \in [n+1,m].$$

Thus we have

$$(3.18)\quad e^{avg}(\phi^s,N)^2 = |\det(T^{-1})| \sum_{j=n+1}^{m} \| ST^{-1}a_j \|^2 \int_{\mathbb{R}^m} z_j^2\, \rho^2(\| z \|)w(\| z \|)dz.$$

Note that $\int_{\mathbb{R}^m} z_j^2\, \rho^2(\| z \|)w(\| z \|)dz$ does not depend on j. Thus

$$\int_{\mathbb{R}^m} z_j^2\, \rho^2(\| z \|)w(\| z \|)dz = \frac{1}{m}\sum_{i=1}^{m} \int_{\mathbb{R}^m} z_i^2\, \rho^2(\| z \|)w(\| z \|)dz$$

$$= \frac{1}{m}\int_{\mathbb{R}^m} \| z \|^2\, \rho^2(\| z \|)w(\| z \|)dz.$$

From this we finally get

$$e^{avg}(\phi^s,N)^2 = \frac{1}{m}\int_{F_1} \| Tf \|^2 \rho^2(\| Tf \|)\mu(df) \sum_{j=n+1}^{m} \| ST^{-1}a_j \|^2$$

from which (3.11) follows. This completes the proof. □

Theorem 3.1 states that the spline algorithm is uniquely optimal for the average case. It is also optimal for the worst case due to Theorem 2.1. It is very desirable that the same algorithm is optimal for both error criteria.

Remark 3.5

For the average case we prove that the spline algorithm is the unique algorithm which minimizes the average error. For

the worst case, the optimal error algorithm is, in general, not unique. However, the spline algorithm is the unique algorithm which minimizes the local errors, see Traub and Woźniakowski [80]. For simplicity, we do not define or discuss local average errors in this paper. As we shall show in Wasilkowski and Woźniakowski [82b], an algorithm which minimizes the average error also minimizes the local average errors. Thus, the spline algorithm is the unique algorithm which minimizes the local errors for both the average and worst case models. □

We now compare the radii of information for the worst and average cases. The radius $r(N)$ of information (for the worst case) is given by (2.15). Note that $h \in \ker N$ is of the form $h = \sum_{j=n+1}^{m} x_j T^{-1} a_j$ for some numbers x_j and

$$\frac{\|Sh\|^2}{\|Th\|^2} = \sum_{i,j=n+1}^{m} x_i x_j (ST^{-1}a_i,\ ST^{-1}a_j) / \sum_{j=n+1}^{m} x_j^2 . \tag{3.19}$$

Define the $(m-n) \times (m-n)$ matrix A such that

$$A = ((ST^{-1}a_i,\ ST^{-1}a_j))_{i,j=n+1}^{m} . \tag{3.20}$$

Note that A is symmetric and positive definite and

$$\frac{\|Sh\|^2}{\|Th\|^2} = \frac{(Ax,x)}{(x,x)} = \frac{\|A^{\frac{1}{2}}x\|^2}{\|x\|^2}$$

Thus

$$\sup_{h \in \ker N} \|Sh\| / \|Th\| = \|A^{\frac{1}{2}}\|_2 = \sqrt{\lambda_{n+1}(A)}$$

where $\|A^{\frac{1}{2}}\|_2$ denotes the spectral norm of the matrix $A^{\frac{1}{2}}$ and $\lambda_{n+1}(A)$ is the largest eigenvalue of A.

Let

$$(3.21) \qquad \bar{\rho}(x) = x\ \rho(x), \quad x \geq 0,$$

and let

$$\| \bar{\rho} \|_\infty = \sup_{x \geq 0} |\bar{\rho}(x)|.$$

Then (2.15) can be rewritten as

$$(3.22) \qquad r(N) = \| \bar{\rho} \|_\infty \| A^{\frac{1}{2}} \|_2 .$$

We now express the average radius $r^{avg}(N)$ in a form similar to (3.22). The radius $r^{avg}(N)$ is given by (3.11). From (3.20) we have

$$(3.23) \quad \sum_{j=n+1}^{m} \| ST^{-1} a_j \|^2 = \sum_{j=n+1}^{m} (ST^{-1} a_j, ST^{-1} a_j) = \text{trace}(A) = \| A^{\frac{1}{2}} \|_E =$$

$$= \sqrt{\lambda_{n+1}(A) + \lambda_{n+2}(A) + \ldots + \lambda_m(A)}$$

where $\| A^{\frac{1}{2}} \|_E = \sqrt{\sum_{i,j=n+1}^{m} a_{ij}^2}$, $A^{\frac{1}{2}} = (a_{ij})$, denotes the Euclidean (or Frobenius) norm of the matrix A and $\lambda_{n+1}(A) \geq \lambda_{n+2}(A) \geq \ldots \geq \lambda_m(A) \geq 0$ are eigenvalues of A .

Let

$$(3.24) \qquad \| \bar{\rho} \|_2 = \{ \int_{F_1} \bar{\rho}^{\,2}(\| Tf \|) \mu(df) \}^{\frac{1}{2}}$$

Of course, $\| \bar{\rho} \|_2 \leq \| \bar{\rho} \|_\infty$. We can rewrite (3.11) using (3.23) and (3.24) getting

(3.25) $$r^{avg}(N) = \| \bar{\rho} \|_2 \; \| A^{\frac{1}{2}} \|_E \; / \sqrt{m}$$

Thus we have proven

Theorem 3.2

Let A and $\bar{\rho}$ be defined by (3.20) and (3.21). Then

$$r(N) = \| \bar{\rho} \|_\infty \; \| A^{\frac{1}{2}} \|_2 ,$$

(3.26)

$$r^{avg}(N) = \| \bar{\rho} \|_2 \; \| A^{\frac{1}{2}} \|_E / \sqrt{m} \; .$$ □

From the definition of the matrix norms, Theorem 3.2 can be rewritten as

Corollary 3.1

$$r(N) = \| \bar{\rho} \|_\infty \sqrt{\lambda_{n+1}(A)} ,$$

$$r^{avg}(N) = \| \bar{\rho} \|_2 \sqrt{\frac{\lambda_{n+1}(A) + \dots + \lambda_m(A)}{m}} ,$$

$$r^{avg}(N) = c \; r(N)$$

where

$$c = \frac{\| \bar{\rho} \|_2}{\sqrt{m} \; \| \bar{\rho} \|_\infty} \sqrt{1 + \frac{\lambda_{n+2}(A)}{\lambda_{n+1}(A)} + \dots + \frac{\lambda_m(A)}{\lambda_{n+1}(A)}} \in$$

$$\left[\frac{\| \bar{\rho} \|_2}{\sqrt{m} \; \| \bar{\rho} \|_\infty} , \sqrt{\frac{m-n}{m}} \right]$$ □

From Theorem 3.2 and Corollary 3.1 it follows that if all eigenvalues of A are of comparable magnitude, $\|\bar{\rho}\|_2$ and $\|\bar{\rho}\|_\infty$ are of comparable magnitude and n is much less than m, then

$$r^{avg}(N) \cong r(N).$$

On the other hand, if $\|\bar{\rho}\|_2$ is significantly smaller than $\|\bar{\rho}\|_\infty$ or the eigenvalues $\lambda_i(A)$ for $i > n+2$ are significantly smaller than $\lambda_{n+1}(A)$ (i.e., A is close to a matrix of rank one) or if n is close to m, then

$$r^{avg}(N) << r(N).$$

4. Optimal Information Operators

In the previous sections we studied optimal algorithms (for the worst and average cases) which use a given information operator N of cardinality n of the form

$$(4.1) \qquad N(f) = [L_1(f), L_2(f), \ldots, L_n(f)]$$

where the L_i are linearly independent linear functionals.

In this section we determine the best choice of linear functionals in (4.1). Since the radius $r(N)$ of information and the average radius $r^{avg}(N)$ of information are the errors of optimal algorithms, we want to select linear functionals in (4.1) in such a way that the corresponding radii of information are minimized.

Let Ψ_n be the class of all linear information operators of cardinality n of the form (4.1).

Definition 4.1

We shall say $r(n)$ $(r^{avg}(n))$ is the nth minimal radius of information (the nth minimal average radius of information) iff

$$(4.2) \qquad r(n) = \inf_{N \in \Psi_n} r(N) \qquad (r^{avg}(n) = \inf_{N \in \Psi_n} r^{avg}(N)).$$

We shall say N_n, $N_n \in \Psi_n$, is an nth optimal information operator (an nth optimal average information operator) iff

$$(4.3) \qquad r(N_n) = r(n) \qquad (r^{avg}(N_n) = r^{avg}(n)).$$

□

We exhibit nth optimal and nth optimal average information operators in terms of eigenvectors of the linear operator K_1 which is defined as follows. Let

$$K \overset{df}{=} ST^{-1} : F_4 \to K(F_4) \subset F_2.$$

By K^* we mean the adjoint operator to K, $K^*: K(F_4) \to F_4$ and

(4.4) $$(Kf,g) = (f,K^*g), \quad \forall f \in F_4, \forall g \in K(F_4).$$

Note that the inner product of the left-hand side of (4.4) is in F_2 and the inner product of the right-hand side of (4.4) is in F_4. Let

(4.5) $$K_1 \overset{df}{=} K^*K: F_4 \to F_4.$$

Of course, K_1 is symmetric and nonnegative definite. Then there exist $\lambda_i = \lambda_i(K_1)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$ and an orthonormal basis $z_1, z_2, \ldots, z_m$ of F_4 such that

(4.6) $$K_1 z_i = \lambda_i z_i , \; i = 1,2,\ldots,m.$$

Thus $\lambda_i(K_1)$ is the ith largest eigenvalue of K_1 and corresponds to the eigenvector z_i. Define the information operator

(4.7) $$N_n(f) = [(Tf,z_1), (Tf,z_2),\ldots, (Tf,z_n)]$$

Then $\text{card}(N_n) = n$ and $N_n \in \Psi_n$. We now establish the optimality of N_n.

Theorem 4.1

The information operator N_n defined by (4.7) is an nth optimal and n th optimal average information operator and

$$(4.8) \qquad r(N_n) = r(n) \quad = \|\bar{\rho}\|_\infty \cdot \sqrt{\lambda_{n+1}(K_1)}\ ,$$

$$(4.9) \quad r^{avg}(N_n) = r^{avg}(n) = \|\bar{\rho}\|_2 \sqrt{\frac{\lambda_{n+1}(K_1) + \dots + \lambda_m(K_1)}{m}}\ . \qquad \square$$

<u>Proof</u>

The optimality of N_n for the worst case and (4.8) follows from (2.15) and Theorem 5.3, Chapter 2, of Traub and Woźniakowski [80]. So we need to prove only (4.9).

We first compute the average radius of N_n . Let $h \in \ker N_n$. Then $(Th, z_i) = 0$, $i = 1,2,\dots,n$, and $Th = \sum_{j=n+1}^{m} x_j z_j$ for some x_j. Thus $z_{n+1},\dots,z_m$ form an orthonormal basis of $T(\ker N_n)$. Hence we can set $a_j = z_j$ in (3.11) for $j = n+1,\dots, m$. We can rewrite (3.11) as

$$(4.10) \quad r^{avg}(N_n)^2 = \|\bar{\rho}\|_2^2 \ \frac{1}{m} \sum_{j=n+1}^{m} (K_1 z_i, z_j) = \|\bar{\rho}\|_2^2 \sum_{j=n+1}^{m} \lambda_j(K_1)/m.$$

We now show that $r^{avg}(N) \ge r^{avg}(N_n)$ for any $N \in \Psi_n$. From (3.11) we have

$$(4.11) \qquad m(r^{avg}(N)/\|\bar{\rho}\|_2)^2 = \sum_{j=n+1}^{m} (K_1 a_j, a_j)$$

where $a_{n+1},\dots,a_m$ form an orthonormal basis of $T(\ker N)$. Then

$$\sum_{j=n+1}^{m} (K_1 a_j, a_j) \ge c \overset{df}{=} \min\{ \sum_{j=n+1}^{m} (K_1 b_j, b_j) : (b_i, b_j) = \delta_{ij} \}$$

From Theorem 4.1.4 of Chapter 2 of Marcus and Minc [64] it follows that

$$c = \sum_{j=n+1}^{m} \lambda_j(K_1).$$

Combining this with (4.11) and (4.10) we have $r^{avg}(N) \geq r^{avg}(N_n)$ which completes the proof. □

Remark 4.1

Theorem 4.1 gives us a very useful property; the same information operator is optimal for the worst and average cases. In Section 3 we proved that the same algorithm is optimal in both the worst and average case models. Thus the information (4.7) and the spline algorithm minimize the error for both models. □

Remark 4.2

Theorem 3.2 states that the radii of information can be expressed in terms of eigenvalues of the matrix A defined by (3.20). Note that for the information operator N_n, $A = ((K_1 z_i, z_j))$ is diagonal since $(K_1 z_i, z_j) = \lambda_i \delta_{ij}$. Thus $\lambda_j(A) = \lambda_j(K_1)$ for $j = n+1, \ldots, m$ and (3.26) agrees with (4.8) and (4.9) for $N = N_n$.
□

As in Section 3, we note that $r^{avg}(n) \cong r(n)$ if $\| \bar{\rho} \|_\infty$ and $\| \bar{\rho} \|_2$ are of comparable magnitude, all eigenvalues $\lambda_j(K_1)$ are of comparable magnitude and n is much less than m .

5. Adaptive Information

In the previous sections we studied linear information operators of the form

$$N(f) = [L_1(f),\ L_2(f),\ \ldots,\ L_n(f)]$$

where linearly independent linear functionals L_i are simultaneously given. Such information operators are called nonadaptive and denoted by $N = N^{non}$. A natural generalization is an adaptive linear information operator N^a defined as

(5.1) $$N^a(f) = [L_1(f),\ L_2(f;y_1),\ \ldots,\ L_n(f;y_1,\ \ldots,\ y_{n-1})]$$

where

(5.2) $$y_i = y_i(f) = L_i(f;\ y_1,\ \ldots,\ y_{i-1})$$

and L_i is a linear functional with respect to the first argument f. See Traub and Woźniakowski [80 p.47]. This means that the choice of the ith functional may now depend on the previously computed values $L_1(f),\ L_2(f;\ y_1),\ \ldots,\ L_i(f;\ y_1,\ \ldots,\ y_{i-1})$.

From (2.15) and Theorem 7.1, Chapter 2, of Traub and Woźniakowski [80] it follows that adaptive information operators are not more powerful than nonadaptive information operators for linear problems in the worst case setting.

Does adaptive information help for linear problems in the average case setting? We prove the surprising result that the answer is negative. In fact, we prove an even stronger result. We construct a nonadaptive linear information operator which has the same cardinality and which consists of the same functionals as a given adaptive information operator and whose average radius does not exceed the average radius of the given adaptive information. In order to prove this we proceed as follows.

Let N^a be an adaptive information operator of the form (5.1). Without loss of generality we can assume that the functionals L_1, $L_2(\cdot;\, y_1), \cdots, L_n(\cdot;\, y_1, \ldots, y_{n-1})$ are linearly independent for every $y_i = y_i(f)$, $i = 1, 2, \ldots, n-1$. Let ϕ be an algorithm using N^a. Then the average error of ϕ is defined by (3.6). Similarily to (2.10) and (3.8) we define the average radius $r^{avg}(N^a)$ as

$$r^{avg}(N^a) = \inf_{\phi \epsilon \Phi(N^a)} e^{avg}(\phi, N^a). \tag{5.3}$$

We now construct a nonadaptive linear information operator N^{non} which consists of the same functionals as N^a and such that $r^{avg}(N^a) \geq r^{avg}(N^{non})$. For a given vector $v = [y_1, y_2, \ldots, y_{n-1}] \epsilon \mathbb{R}^{n-1}$ define the linear functionals

$$L_{i,v}(f) = L_i(f;\, y_1, \ldots, y_{i-1}),\ i = 1,2,\ldots,n. \tag{5.4}$$

We assume that for every f, $L_{i,v}(f)$, as a function of v, has a continuous first derivative for almost all v.

Define the information operator

$$N_v^{non}(f) = [L_{1,v}(f), L_{2,v}(f), \ldots, L_{n,v}(f)]. \tag{5.5}$$

Note that N_v^{non} is a nonadaptive linear information operator of cardinality n which consists of the same functionals as N^a. Let $a_1(v), a_2(v), \ldots, a_m(v)$ be a basis of F_4 such that

$$L_{i,v}\,(T^{-1}a_j(v)) = \delta_{i,j}\,, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m,$$

(5.6)

$$(a_k(v),\; a_j(v)) = \delta_{k,j} \quad k = n+1, \ldots, m, \quad j = 1, 2, \ldots, m.$$

Since $L_{i,v}$ depends only on $y_1,\ldots y_{i-1}$, we choose $a_i(v)$ depending also on $y_1,\ldots,y_{i-1}$, i.e., $a_i(v) = a_i(y_1,\ldots,y_{i-1})$.

.Due to regularity of $L_{i,v}(f)$ we can choose $a_i(v)$ such that they are continuously differentiable for almost all v.

Let

$$(5.7) \qquad q = \inf_{v\in\mathbb{R}^{n-1}} \sum_{j=n+1}^{m} \| ST^{-1}a_j(v) \|^2.$$

Let the infimum in (5.7) be attained for $v = v^*$, i.e.,

$$\sum_{i=n+1}^{m} \| ST^{-1}a_i(v^*) \|^2 = q.$$

We are ready to prove

Theorem 5.1

$$r^{avg}(N^a) \geq r^{avg}(N_{v^*}^{non}).$$

□

Proof

We proceed similarily as in the proof in Theorem 3.1. Let

$\Phi \in \Phi(N^a)$. The average eror of Φ is defined by (3.6). We change variables in (3.6) by setting

$$(5.8) \qquad f = G(y) \overset{df}{=} \sum_{j=1}^{n} y_j \, T^{-1} a_j(v) + \sum_{j=n+1}^{m} y_j \, T^{-1} a_j(v)$$

where $v = [y_1, y_2, \ldots, y_{n-1}]$ and $y = [y_1, y_2, \ldots, y_m]$. Note that the mapping G is one-to-one. Indeed, knowing f we have, due to (5.6), $y_j = L_j(f)$, $j = 1, 2, \ldots, n$. Thus v and $a_j(v)$, $j = 1,2,\ldots,n$, are also known and $y_{n+1}, \ldots, y_m$ are a part of the unique components of f in the basis $a_1(v), \ldots, a_m(v)$. The mapping G is continuously differentiable almost everywhere. From (5.8) we have

$$(5.9) \qquad G(y) = T^{-1} Q(v) y^t$$

where $Q(v) = [a_1(v), a_2(v), \ldots, a_m(v)]$ is an orthogonal matrix and t denotes the transpose. From (5.6) we get

$$\left(\frac{\partial a_k}{\partial y_p}(v), a_j(v) \right) + \left(a_k(v), \frac{\partial a_j}{\partial y_p}(v) \right) = 0 \; .$$

Since a_k depends only on $y_1, y_2, \ldots, y_{k-1}$, we have

$\frac{\partial a_k}{\partial y_p}(v) = 0$ for $p \geq k$. Thus

$$(5.10) \qquad \left(a_k(v), \frac{\partial a_j}{\partial y_p}(v) \right) = 0 \; , \quad \forall j \, , \; \forall p \geq k \; .$$

Let $Q_p(v) = [\frac{\partial a_1}{\partial y_p}(v), \ldots, \frac{\partial a_{m(v)}}{\partial y_p}]$ and let $W_p(v) = Q^t(v) Q_p(v)$.

Due to (5.10) the (k,j) element of $W_p(v)$ is equal to

$\left(a_k(v), \frac{\partial a_j}{\partial y_p}(v) \right) = 0$ for any j and $p \geq k$. Thus the first p rows of $W_p(v)$ are equal to zero. From (5.9) we have

$$G'(y) = T^{-1} Q(y) \, \{ I + [W_1(v) y^t, \ldots, W_m(v) y^t] \} \; .$$

Since the first p components of $W_p(v)y^t$ are equal to zero, the matrix $[W_1(v)y^t, \ldots, W_m(v)y^t]$ is a lower triangular matrix with zero diagonal. This yields

$$|\det G'(y)| = |\det T^{-1}| \ .$$

Let $g(v) = \| \sum_{j=1}^{n} y_j \ T^{-1}a_j(v) \|^2$. Then $\| Tf \|^2 = g(v) + \sum_{j=n+1}^{m} y_j^2$. Using the properties of G we transform (3.6) by techniques similar to those used in (3.16) and (3.17). Thus

$$e^{avg}(\phi, N^a)^2 = \tfrac{1}{2} |\det T^{-1}| \int_{\mathbb{R}^n} \{ \int_{\mathbb{R}^{m-n}} [\| \sum_{j=1}^{n} y_j \ ST^{-1}a_j(v) + \sum_{j=n+1}^{m} y_j ST^{-1}a_j(v)$$

$$- \phi(y_1, \ldots, y_n) \|^2 + \| \sum_{j=1}^{n} y_j ST^{-1}a_j(v) - \sum_{j=n+1}^{m} y_j ST^{-1}a_j(v) - \phi(y_1, \ldots, y_n) \|^2]$$

$$\rho^2((g(v) + \sum_{j=n+1}^{m} y_j^2)^{\frac{1}{2}}) \ w \ ((g(v) + \sum_{j=n+1}^{m} y_j^2)^{\frac{1}{2}}) dy_{n+1} \cdots dy_m \} \ dy_1 \cdots dy_n \geq$$

$$|\det T^{-1}| \int_{\mathbb{R}^n} | \{ \int_{\mathbb{R}^{m-n}} \| \sum_{j=n+1}^{m} y_j ST^{-1}a_j(v) \|^2 \rho^2((g(v) + \sum_{j=n+1}^{m} y_j^2)^{\frac{1}{2}})$$

$$w((g(v) + \sum_{j=n+1}^{m} y_j{}^2)^{\frac{1}{2}})\, dy_{n+1} \cdots dy_m\}\, dy_1 \cdots dy_n =$$

$$|\det T^{-1}| \int_{\mathbb{R}^n} \{ \sum_{j=n+1}^{m} \| ST^{-1} a_j(v) \|^2 \int_{\mathbb{R}^{m-n}} y_j{}^2\, \rho^2((g(v) + \sum_{j=n+1}^{m} y_j{}^2)^{\frac{1}{2}})$$

$$w((g(v) + \sum_{j=n+1}^{m} y_j{}^2)^{\frac{1}{2}}) dy_{n+1} \cdots dy_m\}\ dy_1 \cdots dy_n\ .$$

Let

$$c(v) = \int_{\mathbb{R}^{m-n}} y_j{}^2\ \ \rho^2((g(v) + \sum_{j=n+1}^{m} y_j{}^2)^{\frac{1}{2}})\ w\ ((g(v) + \sum_{j=n+1}^{m} y_j{}^2)^{\frac{1}{2}})$$

$$dy_{n+1} \cdots dy_m .$$

Since $c(v)$ does not depend on j , (5.7) yields

$$(5.11)\ e^{avg}(\phi,N^a)^2 \geq \{ |\det T^{-1}| \int_{\mathbb{R}^n} c(v)\, dy_1 \cdots dy_n\} \sum_{j=n+1}^{m} \| ST^{-1} a_j(v^*) \|^2 .$$

Take now the nonadaptive linear information operator $N^{non}_{v^*}$ and repeat the above transformation with $v = v^*$ and with the spline algorithm $\phi^s(N^{non}_{v^*}(f)) = \sum_{j=1}^{n} L_i(f)\ ST^{-1} a_j(v^*)$. Then we find that the right-hand side of (5.11) is equal to $e^{avg}(\phi^s, N^{non}_{v^*}) = r^{avg}(N^{non}_{v^*})$. Thus $e^{avg}(\phi,N^a) \geq r^{avg}(N^{non}_{v^*})$ and this holds for every ϕ from $\Phi(N^a)$ Hence $r^{avg}(N^a) \geq r^{avg}(N^{non}_{v^*})$ which completes the proof. □

Theorem 5.1 states that for every adaptive information operator one can find a nonadaptive information operator of the same structure and cardinality as the given adaptive information and with no greater average radius. This means that adaptive information operators do not supply more information than nonadaptive ones. This result and the corresponding result for the worst case model may be summarized in

Corollary 5.1

Adaption does not help for linear problems in either the average or worst case models. □

6. How Much Can Information Reduce Uncertainty?

We considered the information operator $N = [L_1, L_2, \ldots, L_n]$ for $n \ge 1$ and proved that $r(N)$ and $r^{avg}(N)$ are sharp lower bounds on uncertainty. Observe that the radii also depend on the setting of the problem, i.e., $r(N)$ depends on S, T, N and ρ, and $r^{avg}(N)$ depends additionally on w. Thus the total information is specified by the linear operators S, T, N and the functions ρ and w. Since S, T, ρ and w are fixed we call N the information.

We pose and answer the following question. What is the uncertainty if only the setting of the problem is known? Or equivalently, what is the minimal ε for which we can find an ε-approximation knowing only S, T, ρ and w?

This corresponds formally to the zero information operator $N = 0$. By convention zero information has cardinality zero. Then an algorithm using zero information takes only one value since $\phi(N(f)) \equiv \phi(0)$. The value $\phi(0)$ should be thus an ε-approximation for all f from F_1. It is easy to observe that the proof technique of Sections 1 through 5 work for $N = 0$ with $n = 0$.

Thus, the radii of zero information are given by

$$r(0) = \|\bar{\rho}\|_\infty \sqrt{\beta_1}$$

(6.1)

$$r^{avg}(0) = \|\bar{\rho}\|_2 \sqrt{(\beta_1 + \ldots + \beta_m)/m}$$

where $\beta_i = \lambda_i(K_1)$ is the ith eigenvalue of the operator K_1 defined by (4.5). Note that (6.1) formally agrees with (4.8) and (4.9) for $n = 0$. Thus, if $r(0) \le \varepsilon$ or $r^{avg}(0) \le \varepsilon$ then we can find an

ε-approximation for the worst or average model without the evaluation of any linear functionals. Note that the optimal error and the optimal average error algorithm is equal to zero, $\phi(0) = 0$. This also formally agrees with the definition (2.14) of the spline algorithm for $n = 0$.

Let

$$(6.2) \qquad \bar{r}(n) = \frac{r(n)}{r(0)} \;, \quad \bar{r}^{avg}(n) = \frac{r^{avg}(n)}{r^{avg}(0)} \;.$$

Then $\bar{r}(n)$ and $\bar{r}^{avg}(n)$ measure how much the uncertainty is reduced after n optimal evalutions of linear functionals. From Theorem 4.1 and (6.1) we have

$$(6.3) \qquad \begin{aligned} \bar{r}(n) &= \sqrt{\frac{\beta_{n+1}}{\beta_1}} \;, \\ \bar{r}^{avg}(n) &= \sqrt{\frac{\beta_{n+1} + \ldots + \beta_m}{\beta_1 + \ldots + \beta_m}} \;. \end{aligned}$$

Note that $\bar{r}(n)$ and $\bar{r}^{avg}(n)$ are independent of the measure μ (i.e., the function w) and the function ρ. They depend only on the eigenvalues of K_1. We consider three, rather typical, distributions of eigenvalues of K_1.

Case 1.

Let $\beta_i \equiv \beta$ for some positive constant β. This corresponds, for instance, to the case when $S = \sqrt{\beta}\, T$ and the operator $K_1 = \beta I$ where I is the identity operator. Then for $n < m$,

$$r(n) = \|\bar{\rho}\|_{\infty} \sqrt{\beta} \ , \quad r^{avg}(n) = \|\rho\|_{2} \sqrt{\beta}\sqrt{(m-n)/m}$$

$$\bar{r}(n) = 1 \ , \quad \bar{r}^{avg}(n) = \sqrt{(m-n)/m} \ .$$

In the worst case it is impossible to solve the problem with any amount of information. In the average case for $n << m$, there is almost no reduction in uncertainty since all the radii are close to unity. This means that such a problem cannot be solved either in the worst or average case for small ε.

Case 2.

Let $\beta_i = cq^{2i}$ for some positive constants c and q with $q < 1$. This corresponds, for instance, to the approximation problem $S = I$ with $Tf = \sqrt{c}\ [q^{-1} f_1, q^{-2} f_2, \ldots, q^{-n} f_n]$. Then

$$r(n) = \sqrt{c}\|\bar{\rho}\|_{\infty} q^{n+1} \ , \quad r^{avg}(n) = \sqrt{c}\|\bar{\rho}\|_{2} q^{n+1} \sqrt{\frac{1-q^{2(m-n)}}{m(1-q^2)}}$$

$$\bar{r}(n) = q^n \ , \quad \bar{r}^{avg}(n) = q^n \sqrt{\frac{1-q^{2(m-n)}}{1-q^{2m}}} \ .$$

For $n << m$, $\bar{r}^{avg}(n) \cong q^n = \bar{r}(n)$. This means that the reduction of uncertainty after n evaluations is approximately the same for the worst and average case.

Case 3.

Let $\beta_i = i^{-2r}$ for $r > \frac{1}{2}$. This corresponds to the approximation problem $S = I$ with $Tf = [f_1, 2^{-r}f_2, \ldots, n^{-r}f_r]$. This choice of

S and T is a discrete analogue of the continuous approximation problem $Sf = f$, $Tf = f^{(r)}$ where f is a scalar $(r-1)$ absolutely continuous function whose nth derivative belongs to L_2. Observe that

$$\sum_{i=n+1}^{m} i^{-2r} \cong \int_{n+1}^{m} x^{-2r}dx = \frac{1}{2r-1}\,(n+1)^{-(2r-1)}\left(1-\left(\frac{n+1}{m}\right)^{2r-1}\right).$$

From this and for $n << m$ we have

$$r(n) = \|\bar{\rho}\|_{\infty}(n+1)^{-r} \quad , \quad r^{avg}(n) \cong \|\bar{\rho}\|_2 \frac{1}{\sqrt{2r-1}}(n+1)^{-r}\sqrt{\frac{n+1}{m}}$$

$$\bar{r}(n) = (n+1)^{-r} \quad , \quad \bar{r}^{avg}(n) \cong (n+1)^{-(r-\frac{1}{2})}.$$

Thus, the reduction of uncertainty is larger, in this case, for the worst case model.

7. Complexity

In this section we briefly discuss the complexity, i.e., the minimal cost, of finding an ε-approximation for the average case model. We obtain extremely tight upper and lower bounds on the complexity. We show that the spline algorithm is essentially an optimal complexity algorithm.

The complexity for the worst case model is studied in Traub and Woźniakowski [80] where very tight complexity bounds are obtained. The spline algorithm is shown to achieve nearly optimal complexity.

We first outline the model of computation. Assume that the cost of adding two vectors from F_2 and the multiplication of a vector from F_2 by a scalar is taken as unity. (Recall that F_2 is the image space of the solution operator S .) Suppose that the evaluation of an arbitrary linear functional is allowed and costs c .

To find an ε-approximation using linear information $N = [L_1, L_2, \ldots, L_n]$ we have to guarantee that $r^{avg}(N) \le \varepsilon$. Let

(7.1) $$m^{avg}(\varepsilon) = \min\{n: r^{avg}(n) \le \varepsilon\}$$

be the ε-average cardinality number. Thus $m^{avg}(\varepsilon)$ denotes the smallest cardinality of information whose average radius does not exceed ε .

Let ϕ be an algorithm using N with $e^{avg}(\phi,N) \le \varepsilon$. Since $e^{avg}(\phi,N) \ge r^{avg}(N)$, the cardinality of N has to be at least $m^{avg}(\varepsilon)$. Thus the evaluation of $N(f)$ requires the computation of at least $m^{avg}(\varepsilon)$ linear functionals. Hence the complexity of $N(f)$, i.e., the cost of computing $N(f)$, is at least $m^{avg}(\varepsilon)c$.

To produce an ε-approximation, the algorithm ϕ has to use at least $m^{avg}(\varepsilon)$ linear functionals. It is natural to postulate that the computation of $\phi(N(f))$ given $N(f)$ has complexity at least $m^{avg}(\varepsilon) - 1$. Let the algorithm complexity (total cost) of producing an ε-approximation by the algorithm ϕ be $\text{comp}^{avg}(\phi)$. A lower bound is given by

$$\text{(7.2)} \qquad \text{comp}^{avg}(\phi) \geq m^{avg}(\varepsilon)(c+1) - 1.$$

Note that (7.2) holds for any algorithm ϕ using an arbitrary linear information operator N. Let

$$\text{(7.3)} \qquad \text{comp}^{avg}(\varepsilon) = \inf\{\text{comp}^{avg}(\phi): e^{avg}(\phi,N) \leq \varepsilon\}$$

be the ε-average complexity. An algorithm ϕ is called an optimal average complexity algorithm iff

$$\text{(7.4)} \qquad \text{comp}^{avg}(\phi) = \text{comp}^{avg}(\varepsilon).$$

From (7.2) we have a lower bound on the ε-average complexity,

$$\text{(7.5)} \qquad \text{comp}^{avg}(\varepsilon) \geq m^{avg}(\varepsilon)(c+1) - 1.$$

We now show that the spline algorithm is a nearly optimal average algorithm ϕ^s using the information N_n defined by (4.7). Recall that N_n is an nth average optimal information operator, $r^{avg}(N_n) = r^{avg}(n)$. The spline algorithm ϕ^s is linear, $\phi^s(N_n(f)) = \sum_{i=1}^{n} L_i(f)g_i$ for some g_i from F_2. Since the elements g_i can be precomputed, the evaluation of $\phi^s(N_n(f))$ given $N(f)$ requires only n multiplications and $n-1$ additions each of unit cost. Thus if $n = m^{avg}(\varepsilon)$ then

$$\text{(7.6)} \qquad \text{comp}^{avg}(\phi^s) = m^{avg}(\varepsilon)(c+2) - 1.$$

Combining (7.6) with (7.5) we see that the spline algorithm is a nearly optimal average complexity algorithm.

A similar result holds for the worst case model. In fact, worst case definitions and results are obtained by deleting the superscripts "avg" in (7.1) through (7.6).

We summarize this in

Theorem 7.1

The spline algorithm is a nearly optimal complexity algorithm in both the average and worst case models. The complexity is given by

$$\text{comp}(\varepsilon) = m(\varepsilon)(c+a_1) - 1,$$

$$\text{comp}^{avg}(\varepsilon) = m^{avg}(\varepsilon)(c+a_2) - 1$$

where $a_1, a_2 \in [1,2]$. □

8. Example

We continue the example of the Introduction. Recall that example deals with the approximation of a trigonometric polynomial of degree m. We choose approximation as our example because it is of such wide interes in applications. We discussed in the Introduction why we confine ourselv in this paper to finite dimensional F_1. Throughout this section we use the approximation example while illustrating the effects of choosing various error criteria and measures.

Identifying a trigonometric polynomial with its coefficients we can set $F_1 = F_2 = F_4 = \mathbb{R}^m$ equipped with the spectral norm and

$$Sf = f \ , \ \forall f \in \mathbb{R}^m \ .$$

Without loss of generality we can assume that T is a diagonal matrix since the dependence on T is through the norm $\| Tf \|$ which is orthogonally invariant. Thus let

$$Tf = [\sqrt{\beta_1}\, f_1, \ \sqrt{\beta_2}\, f_2, \ldots, \ \sqrt{\beta_m}\, f_m]$$

where $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_m > 0$.

(i) The absolute error criterion, $\rho(x) \equiv 1$. Then $\bar{\rho}(x) = x$ and $\| \bar{\rho} \|_\infty = +\infty$. Thus implies that

$$r(n) = +\infty \quad , \ \forall n < m.$$

Thus, it is impossible to find an ε-approximation for the worst case, no matter what the value of ε.

For the average case, $\| \bar{\rho} \|_2$ may be finite or infinite depending on the function w. For instance, let

$$(8.1) \qquad w(x) = (\beta_1 \cdots \beta_m)^{\frac{1}{2}} \ \pi^{-m/2} \ e^{-x^2} \ .$$

A rather lengthy calculation shows that w satisfies (3.2), i.e., $\int_{\mathbb{R}^m} w(\| Tf \|) df = 1$, and

$$||\bar{\rho}||_2 = \sqrt{\frac{m}{2}}.$$

Hence $||\bar{\rho}||_2$ is finite although it goes to infinity with m. The nth average radius is given by

$$r^{avg}(n) = \sqrt{\frac{\beta_{n+1} + \ldots + \beta_m}{2}}.$$

We can find an ε-approximation for the average case using n evaluations whenever $r^{avg}(n) \le \varepsilon$.

On the other hand, let

(8.2) $$w(x) = (\beta_1 \ldots \beta_m)^{\frac{1}{2}} \Pi^{-m/2} \Gamma(\frac{m}{2} + 1)(1 + x^2)^{-(\frac{m}{2} + 1)}.$$

Then (3.2) holds and $||\bar{\rho}||_2 = +\infty$. Thus

$$r^{avg}(n) = +\infty \quad , \forall n < m.$$

Hence, it is impossible to find an ε-approximation for the average case (as in the worst case) no matter what the value of ε.

(ii) The relative error criterion, $\rho(x) = 1/x$. Then $\bar{\rho}(x) \equiv$ and $||\bar{\rho}||_\infty = ||\bar{\rho}||_2 = 1$ for an arbitrary function w satisfying (3.2). We have in this case

$$r(n) = \sqrt{\beta_{n+1}}$$

$$r^{avg}(n) = \sqrt{\frac{\beta_{n+1} + \ldots + \beta_m}{m}}.$$

(iii) The absolute errcr criterion for a subset of F_1. Let

$$\rho(x) = \begin{cases} 1 & 0 \le x \le 1, \\ 0 & x > 1. \end{cases}$$

Thus we approximate Sf only for elements f such that $\| Tf \| \le 1$. We have $\bar{\rho}(x) = x$ for $x \in [0,1]$ and $\bar{\rho}(x) = 0$ for $x > 1$. Hence

$$\| \bar{\rho} \|_\infty = 1.$$

Note that $\| Tf \| \le 1$ defines an ellipsoid in $\mathbb{R}^m$. We define w such that its support is on this ellipsoid, i.e.,

$$w(x) = (\beta_1 \cdots \beta_m)^{\frac{1}{2}} \pi^{-m/2} \Gamma(\tfrac{m}{2} + 1) \begin{cases} 1 & 0 \le x \le 1, \\ 0 & x > 1. \end{cases}$$

Then w satisfies (3.2) and

$$\| \bar{\rho} \|_2 = \sqrt{\frac{m}{m+2}} .$$

For large m, $\| \bar{\rho} \|_2 \cong \| \bar{\rho} \|_\infty = 1$. In this case we have

$$r(n) = \sqrt{\beta_{n+1}} ,$$

$$r^{avg}(n) = \sqrt{\frac{\beta_{n+1} + \cdots + \beta_m}{m+2}} .$$

ACKNOWLEDGEMENTS

We thank David Lee, Charles Micchelli, and Arthur Werschulz for their thorough reading of the manuscript.