CUCS-65-83

AVERAGE CASE ε-COMPLEXITY IN COMPUTER SCIENCE: A BAYESIAN VIEW

July 1983

J. B. Kadane

Carnegie-Mellon University

G. W. Wasilkowski

Columbia University

TABLE OF CONTENTS

Acknowledgements

The authors are grateful to H. T. Kung, M. I. Shamos and J. F. Traub for their roles in bringing them together. Joseph B. Kadane was supported in part by ONR Contract 014-82-K-0622 and G. W. Wasilkowski was supported in part by the National Science Foundation under Grant MCS-7823676.

Abstract

Relations between average case ϵ-complexity and Bayesian statistics are discussed. An algorithm corresponds to a decision function, and the choice of information to the choice of an experiment. Adaptive information in ϵ-complexity theory corresponds to the concept of sequential experiment. Some results are reported, giving ϵ-complexity and minimax-Bayesian interpretations for factor analysis. Results from ϵ-complexity are used to establish that the optimal sequential design is no better than optimal nonsequential design for that problem.

This paper shows that average case analysis of algorithms and information in ϵ-complexity theory is related to to optimal decisions and experiments, respectively, in Bayesian Theory. Finding such relations between problems in previously disjoint literatures is exciting both because one discovers a new set of colleagues and because results obtained in each literature can illuminate the other.

Sections 1 and 2 explain, respectively, the worst-case and average-case analysis of algorithms. Section 3 establishes the correspondence mentioned above. Finally, Section 4 discusses some results from the average case ϵ-complexity literature, and its interpretation for Bayesians. We hope that the relation reported here can lead to further fruitful results for both fields.

1 Worst Case Analysis

In this section we briefly present some major questions addressed in ϵ-complexity theory. We first discuss the worst case model which is conceptually simpler than the average case model, discussed in Section 2.

An expository account of ϵ-complexity theory (which is also known as information-centered theory) may be found in Traub and Woźniakowski (1983). The general worst case model is presented in two research monographs: Traub and Woźniakowski (1980) and Traub, Wasilkowski and Woźniakowski (1983). The first of these has an extensive annotated bibliography. Reports on the average case model are cited in Section 2.

A simple integration problem provides a suggestive illustration. We wish to approximate $\int_0^1 f(t)dt$ knowing n values of f at points t_i, $N(f) = [f(t_1), \ldots, f(t_n)]$, and knowing that f

belongs to a given class F of functions where F is a subclass of a linear space F_1. This means that for given information value $y = N(f)$ we approximate the integral of f by $\phi(y)$ where $\phi: \mathbb{R}^n \to \mathbb{R}$ is a mapping called an algorithm. In the worst case model discussed in this section, the error of ϕ is determined by its performance for the "hardest" element f, i.e.,

$$e^{\omega}(\phi,N) = \sup_{f \in F} \left| \int_0^1 f(t)dt - \phi(N(f)) \right| .$$

The radius of information $r^{\omega}(N)$, is defined as the minimal error among all algorithms that use N, and the optimal algorithm ϕ^* is defined so that its error is minimal, i.e.,

$$e^{\omega}(\phi^*,N) = r^{\omega}(N) = \inf_{\phi} e^{\omega}(\phi,N).$$

Suppose now that the points t_i may be varied. Then $N^*(f) = [f(t_1^*),\ldots,f(t_1^*)]$ is nth optimal iff the points t_i^* are chosen to minimize the radius r^{ω}. Hence, roughly speaking, an optimal algorithm ϕ^* that uses nth optimal information N^* approximates $\int_0^1 f(t)dt$ for every $f \in F$, with minimal error among all algorithms that use n function evaluations. Observe that these concepts are independent of a model of computation.

In ϵ-complexity theory we are interested in minimizing errors as well as cost. More precisely, suppose we are given $\epsilon > 0$. Then the problem is to find information N, and an algorithm ϕ that uses N, so that ϕ approximates $\int_0^1 f(t)dt$, $\forall f \in F$, with error not greater that ϵ and the cost of computing $\phi(N(f))$ is minimized. Observe that this cost (denoted by $\mathrm{comp}(\phi,f)$) is the sum of two terms: the cost of computing $y = N(f)$ (denoted by $\mathrm{comp}(N,f)$) and the cost of computing $\phi(y)$ given y (denoted by $\mathrm{comb}(\phi,y)$). Of course, $\mathrm{comp}(\phi,f) \geq \mathrm{comp}(N,f)$. A major problem of

ϵ-complexity can be stated as follows: find N^{**} and ϕ^{**} that uses N^{**} such that $e^{\omega}(\phi^{**},N^{**}) \leq \epsilon$ and

$$\sup_{f\in F} \mathrm{comp}(\phi^{**},f) = \min\{\sup_{f\in F} \mathrm{comp}(\phi,f): e^{\omega}(\phi,N) \leq \epsilon\}.$$

Of course, the choice of N^{**} and ϕ^{**} depend strongly on the model of computation, i.e., how much various operations cost. In this section we discuss a very simple model, assuming that comp (N,f) is proportional to n, say comp(N,f) = c•n, where c is so large that comb(ϕ^{*},N(f) is negligible for some optimal algorithm ϕ^{*} that uses N. Then to choose N^{**} and ϕ^{**} we must find information with the minimal number of function evaluations such that $r^{\omega}(N^{**}) \leq \epsilon$. Then for ϕ^{**} we can take the optimal error algorithm ϕ^{*} that uses N^{**}.

We now comment on the assumption that comb(ϕ,N(f)) < < cn. For many problems there exists an optimal algorithm ϕ^{*} which is linear, i.e., $\phi^{*}(y) = \sum_{i=1}^{n} y_i g_i$, $g_i \in \mathbb{R}$. Since an arithmetic operation (we take its cost to be unity) is less expensive than a function evaluation.

$$\mathrm{comb}(\phi,y) = 2n-1 < < cn.$$

Hence the above assumption is satisfied whenever ϕ^{*} is linear. This also explains one reason why we are particularly interested in linear optimal algorithms.

We now indicate another important question studied in ϵ-complexity theory. Recall that in our example the information N is of the form, $N(f) = [f(t_1), \ldots, f(t_n)]$. If the points t_i are given independently of f, then N is called nonadaptive. If the t_i depend on previously computed

information values, N is called adaptive. Nonadaptive information is desirable on a parallel computer and for distributed computation since the information can then be computed on various processors at the same time. This lowers significantly the cost comp (N,f). Adaptive information has to be computed sequentially which means that comp(N,f) remains nc. Hence, if N^{non} is nonadaptive and N^{a} is adaptive, we prefer N^{non} unless $r^{\omega}(N^{a}) << r^{\omega}(N^{non})$. This explains why we are interested in the following question: when is adaptive information more powerful than nonadaptive information?

We described some of the major questions addressed in ϵ-complexity theory by using integration as an example. The same questions can be asked in great generality where, for example, different operators are considered instead of $\int_0^1 f(t)dt$, and different information operators N are studied instead of function evaluations. For many problems optimal information and optimal algorithms are known. Sometimes this information and these algorithms are new. Furthermore, for many problems (including the integration problem) adaptive information is not more powerful than nonadaptive information. The significance of this result is that adaption is widely used by practitioners.

2 Average Case Anaysis

In the previous section we briefly discussed the worst case model where the error of an algorithm was defined by its performance for the "hardest" f. For some problems this model might be too pessimistic . Researchers in ϵ-complexity theory also analyze average case models, three of which we present in this section. For simplicity we discuss only problems defined on

Hilbert spaces. This presentation is based on Traub, Wasilkowski and Woźniakowski (1981), Wasilkowski and Woźniakowski (1982), Woźniakowski (1982) and Wasilkowski (1983a).

Let F_1 and F_2 be two real separable Hilbert spaces and let S,

$$S: F_1 \to F_2, \tag{2.1}$$

be a continuous operator. We call S a solution operator. For instance, we might take $S(f) = \int_0^1 f(t)dt$ which corresponds to the integration problem discussed above, $S(f) = f$ which corresponds to the approximation problem or $S = \Delta^{-1}$ where $\Delta u = -\sum_{i=1}^{p} \partial^2 u/\partial x_i^2$ which corresponds to a differential equation problem.

As in Section 1 we want to approximate S(f) for every $f \in F_1$ but now with the average error as small as possible. In order to define average error we assume that the space F_1 is equipped with a probability measure μ, $\mu(F_1) = 1$, defined on the Borel sets of F_1.

To find an approximation to S(f) we must know something about f. We assume $y = N(f)$ is known, where now N is defined as follows:

$$N(f) = \left[L_1(f), L_2(f,y_1), \ldots, L_n(f,y_1,\ldots,y_{n-1})\right] \in \mathbb{R}^n, \tag{2.2}$$

where $y_1 = L_1(f)$, $y_i = L_i(f,y_1, \ldots, y_{i-1})$ $(i = 2, \ldots, n)$, and for every $y \in \mathbb{R}^n$ the functionals $L_i(\bullet,y) : F_1 \to \mathbb{R}$ belong to some given class Λ of measureable functionals. Such an operator N is called an adaptive information operator and the number n of functional evaluations is called the cardinality of N, card $(N) = n$. In general, the choice of the ith evaluation depends on the

previously computed information $y_1, \ldots, y_{i-1}$. If $L_i(\bullet, y) \equiv L_i$ for every y then N is called nonadaptive. Of course, for nonadaptive information N, N(f) can be very efficiently computed in parallel.

To illustrate this concept assume, as in Section 1, that F_1 is a space of functions. Then N(f) might consist of function evaluations, $N(f) = [f(t_1), \ldots, f(t_n)]$, i.e., $L_i(f) = f(t_i)$. If the t_i's are fixed then N is nonadaptive. Otherwise, if the selection of t_2 depends on the value $f(t_1)$ and so on, then N is adaptive.

Knowing N(f) we approximate S(f) by $\phi(N(f))$ where ϕ is an algorithm that uses N. By an algorithm we mean any mapping from $N(F_1) = \mathbb{R}^n$ into F_2. Then the average error of ϕ is defined as

$$e^{avg}(\phi, N) = \left\{ \int_{F_1} \left\| S(f) - \phi(N(f)) \right\|_2^2 \mu(df) \right\}^{1/2}. \tag{2.3}$$

where the integral in (2.3) is understood as the Lebesque integral.

We pause to comment on definition (2.3)

- the average error of an algorithm ϕ is conceptually the same as the error $e^{\omega}(\phi, N)$ in the worse case model, except the supremum is replaced by an integral.

- The average error of ϕ is well defined only if ϕ is "measurable" (or more precisely only when $\left\| S(\bullet) - \phi(N(\bullet)) \right\|_2^2$ is measurable in f). Since "unmeasurable" algorithms might be as good as "measurable" ones, we would like to have the concept of average error for every algorithm. It is possible to extend the definition (2.3) so that average error is well defined for every ϕ (see Wasilkowski (1983a)) but for the

purpose of this paper we shall assume that ϕ is chosen such that (2.3) is well defined.

- The average error is defined as an average value of $\|S(f) - \phi(N(f))\|^2$. Of course, in general $\|S(f) - \phi(N(f))\|^2$ can be replaced by a different error function $E(s(f), \phi(N(f))$ (see e.g. Traub, Wasilkowski and Woźniakowski (1981), and Wasilkowski (1983a)).

Let

$$r^{avg}(N) = \inf_{\phi} e^{avg}(\phi, N) \tag{2.4}$$

be the average radius of information N. Then by an optimal average error algorithm we mean an algorithm ϕ^* that uses N and enjoys the smallest error, i.e.,

$$e^{avg}(\phi^*, N) = r^{avg}(N). \tag{2.5}$$

For a given integer n, let Ψ_n be the class of all information operators N of the form (2.2) with cardinality not greater than n. We shall say that $r^{avg}(n)$ is an nth minimal average radius if

$$r^{avg}(n) = \inf_{N \in \Psi_n} r^{avg}(N). \tag{2.6}$$

We shall say that N_n^* from Ψ_n is nth optimal if the radius of N_n^* is minimal, i.e.,

$$r^{avg}(N_n^*) = r^{avg}(n) \tag{2.7}$$

Using this notation, we now describe some of the major questions addressed in the average case model. Given the problem, i.e., solution operator S, probability measure μ, class of information operators N and error tolerance ϵ, find N^{**} and ϕ^{**} with minimal (or almost minimal) complexity such that

operators N and error tolerance ϵ, find N^{**} and ϕ^{**} with minimal (or almost minimal) complexity such that

$$e^{avg}(\phi^{**},N^{**}) \leq \epsilon.$$

The complexity of the average case model can be measured in different ways. Depending on the problem, sometimes it is defined as complexity of ϕ in the worst case, i.e.,

$$\sup_{f \in F_1} \text{comp}(\phi,f).$$

We call this Case A. Sometimes complexity is defined by the average case complexity, which is

$$\int_{F_1} \text{comp}(\phi,f)\,\mu(df),$$

which we call Case B. However, if we agree that the assumptions in Section 1 are satisfied the search for optimal ϕ^{**} and N^{**} can be simplified, namely, to find ϕ^{**} and N^{**} we need only to find an n^*th optimal information operator with the minimal n^* such that

$$r^{avg}(n^*) \leq \epsilon,$$

here called Case C. Then the optimal N^{**} is $N^{**} = N_{n^*}$ and the optimal algorithm ϕ^{**} is an optimal average error algorithm ϕ^* that uses N^{**}. The conditions which guarantee that ϕ^* is linear, i.e., $\phi_n^*((y_1, \ldots, y_n)) = \sum_{i=1}^{n} y_i g_i, g_i \in F_2$, are also studied. Another important question posed in ϵ-complexity theory is when adaptation is more powerful than nonadaptation.

We end this section by presenting the concepts of local errors and local radii.

Let N be an information operator. For simplicity assume that $N(F_1) = \mathbb{R}^n$. Define a measure $\mu_{1,N}$ on Borel sets from $\mathbb{R}^n$ as

$$\mu_{1,N}(A) = \mu(N^{-1}(A)) \quad \left(= \mu(\{f \in F_1, N(f) \in A\})\right). \tag{2.8}$$

Of course, $\mu_{1,N}$ is a probability measure. Then there exists a family of probability measures $\mu_{2,N}(\cdot \mid y)$ on F_1 such that $\mu_{2,N}$ are concentrated on $N^{-1}(y)$, i.e., $\mu_{2,N}(N^{-1}(y) \mid y) = 1$, for almost every y and

$$\mu(B) = \int_{N(B)} \mu_{2,N}(B \mid y)\mu_{1,N}(dy). \tag{2.9}$$

(For more detailed discussion see Wasilkowski (1983a)). Such measures $\mu_{2,N}(\cdot \mid y)$ are called conditional measures. Then the average error of an algorithm ϕ can be rewritten as

$$e^{avg}(\phi,N)^2 = \int_{\mathbb{R}^n} \left\{ \int_{F_1} \|S(f) - \phi(y)\|^2 \mu_{2,N}(df \mid y) \right\} \mu_{1,N}(dy)$$

$$= \int_{\mathbb{R}^n} e^{avg}(\phi,N,y)^2 \mu_{1,N}(dy), \tag{2.10}$$

where

$$e^{avg}(\phi,N,y) = \left\{ \int_{F_1} \|S(f) - \phi(y)\|^2 \mu_{2,N}(df(y)) \right\}^{1/2} \tag{2.11}$$

is called a local average error of ϕ. Let

$$r^{avg}(N,y) = \left\{ \inf_{f \in F_2} \int_{F1} \| S(f)-g \|^2 \mu_{2,N}(df \mid y) \right\}^{1/2} \tag{2.12}$$

be the local average radius of N. It is proven in Wasilkowski (1983a) that $r^{avg}(N,y)^2$, as a function of y, is μ_1 - measurable. Hence, we have

$$r^{avg}(N)^2 = \int_{\mathbb{R}^n} r^{avg}(N,y)^2 \mu_{1,N}(dy) \tag{2.13}$$

and ϕ^* is optimal iff

$$e^{avg}(\phi^*,N,y)^2 = r^{avg}(N,y)^2 = \inf_{g \in F_2} \int_{F_1} \|S(f)-g\|^2 \mu_{2,N}(df \mid y), \ \forall \ y, \text{ a.e.} \tag{2.14}$$

Finally, N_n^* is nth optimal iff

$$r^{avg}(N_n^*)^2 = \inf_{N \in \Psi_n} \int_{\mathbb{R}^n} \left\{ \inf_{g \in F_2} \int_{F_1} \|S(f) - g\|^2 \mu_{2,N}(df \mid y) \right\} \mu_{1,N}(dy). \tag{2.15}$$

3 A Bayesian Interpretation of the Average Case Model

Recall that the Bayesian scheme for the design of experiments comes in two equivalent forms, normal and extensive (Lindley, 1971). In the normal form, one chooses a decision function $\delta(x)$, depending on the data, to minimize expected loss over both the sample space X and parameter spaces Ω:

$$\min_{\delta} \int_{\Omega} \int_{X} L(\delta(x),\theta)\ p(x \mid \theta) dx d\theta. \tag{3.1}$$

In the extensive form, a Bayesian chooses an experiment e and a decision d, after observing x but not observing θ, to

$$\min_{e} \int_{x} dx \ \min_{d} \int_{\Omega} d\theta \ L(d,\theta,e,x) p(\theta \mid x,e) p(x \mid e). \tag{3.2}$$

Comparing (3.1) with (2.3) and (2.4) for the normal forms, and (3.2) with (2.15) for the extensive forms, we see identical forms, leading to the correspondence exhibited in Table 1.

Table 1: Correspondence of Notation Between Bayesian Decision Theory and ϵ-Complexity Theory (Case C)

Language of Bayesian Decision Theory		Language of ϵ-Complexity Theory	
e	experiment	N	information
x	data	y	information value
$\delta(.)$	decision function	$\phi(.)$	algorithm
d	decision	g	value of algorithm
θ	parameter	f	problem element
L	loss	$\|\|S(f)-g\|\|^2$	algorithm error
$p(\theta \mid x,e)$	posterior distribution	$\mu_{2,N}(. \mid y)$	conditional probability
$p(x \mid e)$	marginal distribution of data	$\mu_{1,N}(.)$	distribution of N
expression (3.1)	Bayes risk of experiment	$[r^{avg}(N)]^2$	squared radius of information

In both cases A and B, researchers in ϵ-complexity theory keep the error and the cost of computation separate. They want to guarantee that the error is not greater than ϵ, and then they ask about minimal cost of computation. A Bayesian might prefer a formulation in which cost and error were represented in the loss function L so that the optimal information N and algorithm ϕ would minimize

$$\int_{F_1} L(\phi,N,f)\ \mu(df)$$

unconditionally. Observe that letting the loss function

$$L(\varphi,N,f) = \mathrm{comp}(\varphi,f) + \left(\frac{1}{\psi_\epsilon(\varphi,N)} - 1\right), \tag{3.3}$$

where $\psi_\epsilon(\varphi,N) = \begin{cases} 1 & \text{if } e^{\mathrm{avg}}(\varphi,N) \le \epsilon \\ 0 & \text{otherwise.} \end{cases}$

will yield a Bayesian formulation of Case B. However, Case A appears not to have a correspondence with Bayesian statistics.

With the above as background, the role of adaptive information becomes clearer to statisticians. Adaptive information is defined to be an N dependent on past values of y, that is, an experiment dependent on past data. Thus what researchers in ϵ-complexity mean by adaptive information is related to what statisticians mean by the sequential design of experiments. To ask whether adaptation helps in the average case is to ask whether sequential experimentation yields greater expected utility.

The average case models presented in this paper is not the only one studied in ϵ-complexity theory. We now present very briefly another model (see Wasilkowski (1983b)) which has no correspondence in the Bayesian Decision Theory and whose conclusion (that probabilistic algorithms are better than nonprobabilistic algorithms) is contrary to Bayesian Decision Theory.

In the model to be presented here we have a class $\bar{\psi}$ of pairs (N,φ) where, as always N is an information operator and φ is an algorithm that uses N. In general, $\bar{\psi}$ is uncountable. Consider

a random variable R with values in $\bar{\Psi}$. This random variable defines the following probabilistic method: according to R, randomly choose $(N_R, \phi_R) \in \bar{\Psi}$ and then approximate S(f) by $x = \phi_R(N_R(f))$. Observe that if R is constant, then this probabilistic method is an ordinary algorithm discussed in this paper. For given $\epsilon > 0$ and $p \in [0,1]$ let $\mathbb{R}(\epsilon,p)$ be the set of all random variables R such that

$$\text{prob}\left(\left\|S(f) - \phi_R(N_R(f))\right\| \leq \epsilon\right) \geq p.$$

i.e., $\mathbb{R}(\epsilon,p)$ is the set of all probabilistic methods which, with probability at least p, yield an approximation with error at most ϵ. Now the problem is to find an optimal method, or equivalently an optimal R^*, with minimal complexity among all methods from $\mathbb{R}(\epsilon,p)$. Here the complexity of R is defined by the average complexity, i.e.,

$$\int_{\bar{\Psi}} \left(\int_{F_1} \text{comp}(\phi,f)\ \mu(df)\right) R(d(N,\phi)) .$$

It turns out that for many problems the optimal R^*, although discrete, is not constant. This means that in this model, probabilistic methods are better than nonprobabilistic methods.

4 An Application to Factor Analysis

In this section we report some results from the average case ϵ-complexity literature giving an interpretation of factor analysis. This and the correspondence between Bayesian statistics and ϵ-complexity yields a Bayesian interpretation of factor analysis with some minimax elements, much along the lines suggested by Manski (1981), Lambert and Duncan (1981) and Berger (1983).

Consider the problem of representing a linear space by a few vectors capturing most of the variability in a particular sense. In a factor analysis one chooses any system of vectors spanning the space spanned by the eigenvectors corresponding to the largest eigenvalues of the covariance matrix V. This problem can be displayed in the language of ϵ-complexity as follows.

Let F_1 be a separable Hilbert space (although statisticians will be more familiar with the large, finite dimensional case $F_1 = \mathbb{R}^m$, we prefer to talk about not necessarily finite spaces, since for many interesting problems in ϵ-complexity the spaces are infinite dimensional.) To represent this space means to approximate the solution operator S, which for this case is the identity operator,

$$S = I$$

possessing partial information N,

$$N(f) = [(f,z_1), \ldots, (f,z_n)]. \tag{4.1}$$

Here z_j are some vectors from F_1 and $(\bullet,\bullet)$ denotes the inner product in F_1. Assume now that the probability measure μ on F_1 is unknown and what we know is its covariance operator V. Recall that V is defined by

$$(Vg,h) = \int_{F_1} (f,g)\ (f,h)\ \mu(df), \quad \forall g,h \in F_1 \tag{4.2}$$

and for $F_1 = \mathbb{R}^m$, V is the covariance matrix of μ. Of course, V is symmetric and positive definite. Without loss of generality we can assume that V has finite trace (this is equivalent to the assumption that $\int_{F_1} \| f \|^2 \mu(df) < +\infty$) since otherwise we cannot approximate S = I

with finite error. Since μ is unknown we replace the problem (2.15) by the following one. Find N^* and ϕ^* that uses N^* such that

$$\sup_{\mu} e^{avg}(\phi^*, N^*, \mu) = \inf_{N} \inf_{\phi} \sup_{\mu} e^{avg}(\phi, N; \mu). \tag{4.3}$$

where $e^{avg}(\phi, N; \mu)$ denotes the average error of ϕ for a measure of μ.

This is, we believe, the ϵ-complexity formulation of the problem studied in factor analysis. To solve this problem we first fix N. From Wasilkowski and Woźniakowski (1982) it follows immediately that

$$\inf_{\phi} \sup_{\mu} e^{avg}(\phi, N, \mu) = e^{avg}(\phi^*; N, \mu^*) = r^{avg}(N; \mu^*) \tag{4.4}$$

where ϕ^* is the spline algorithm that uses N and μ^* is any orthogonally invariant measure. The formal definitions of spline algorithms and of orthogonal invariance can be found in the paper cited above. We only stress tht the spline algorithm is linear (i.e., is simple) and that Gaussian measures are examples of orthogonally invariant measures. Hence from (4.4) we have that the spline algorithm ϕ^* that uses N is optimal in the sense of (4.3). Furthermore, the Gaussian measure with covariance operator V is the "least favorable" measure for every information N.

We now exhibit the optimal information N^*. Let $\eta_1^*, \eta_2^*, \ldots, (\|\eta_j^*\| = 1)$ be the eigenvectors of the operator V, i.e.,

$$V\eta_j^* = \lambda_j^* \eta_j^*, \quad \lambda_1^* \geq \lambda_2^* \geq \ldots \geq 0. \tag{4.5}$$

Let

$$N^*(f) = [(f,\eta_1^*), \ldots, (f,\eta_n^*)].$$

From (4.4) and Wasilkowski and Wozniakowski (1982) it follows that

$$\inf_N \inf_\phi \sup_\mu e^{2\nu\varepsilon}(\phi,N,\mu) = e^{4\nu\varepsilon}(\phi^*,N^*,\mu^*) = r^{4\nu\varepsilon}(N^*,\mu^*)$$

$$\cdot \left\{ \sum_{i=n+1}^{\infty} \lambda_i^* \right\}^{1/2}. \tag{4.7}$$

This means that N^* defined by (4.6) is optimal and that the spline algorithm ϕ^*, which for this information has a very simple form

$$\phi^*(y) = \sum_{i=1}^{n} y_i \eta_i^*, \; y = [y_1, \ldots, y_n] \in \mathbb{R}^n,$$

is the unique optimal algorithm for the problem (4.3).

We now comment on the choice of information N^*. Suppose that instead of N^* one chooses N,

$$N(f) = \left[(f,z_1), \ldots, (f,z_n)\right],$$

where $z_1, \ldots, z_n$ spans the same space as $\eta_1^*, \ldots, \eta_n^*$, i.e.,

$$\text{lin}\{z_1,\ldots,z_n\} = \text{lin}\{\eta_1^*, \ldots, \eta_n^*\}.$$

Then information N^* and N are equivalent. More precisely, if $\phi_{N^*}^*$ and ϕ_N^* are optimal algorithms that use N^* and N, respectively, then

$$\phi_{N^*}^*(N^*(f)) = \phi_N^*(N(f)), \; \forall f \in F_1,$$

and

$$r^{avg}(N^*, \mu) = r^{avg}(N, \mu), \text{ for every } \mu.$$

Observe that N^* defined above is nonadaptive. It is natural to ask whether N^* remains optimal among all adaptive information operators. From Woźniakowski (1982) we know that adaption is not more powerful in the average if the measure μ is orthogonally invariant. Since for every N the supremum in (4.3) is attained for such measures, this implies that adaption does not help in our problem.

In the language of statistics (we refer the reader to Table 1 as necessary), $S = f$ is a random variable, N is an experiment which gives ω_j for n chosen vectors z_j ($j = 1, \ldots, n$), the value of the random variables $(S(\bullet), z_j)$, which can be written in the finite dimensional case, $z_j^T S(\bullet)$. Note that the covariance matrix V of S in the finite dimensional case is the covariance operator of μ. Knowing the matrix V, we wish to find, for fixed n, optimal vectors $\eta_1^*, \ldots, \eta_n^*$ to satisfy (4.3), that is to minimize, over experiments N and estimates ϕ^*, the loss against the least favorable distribution μ for S. (This latter aspect gives rise to the minimax character of the criterion). The nature of this optimal choice of vectors η_j^* is that they span the space spanned by the eigenvectors of V corresponding to the n largest eigenvalues. This is exactly the space of possible factor analysis of V. We therefore have a Bayes-minimax interpretation of factor analysis, and one that does not appear to be available in the statistical literature.

Another conclusion is that the experiment N^*, which is nonsequential, is optimal among all sequential experiments.

5 Conclusion

We believe that the relations between Bayesian decision theory and average case ϵ-complexity theory may have important consequences for both groups of researchers. We have found in our discussions that despite the similarities reported here, the perspectives of the two fields are rather distinct. Only as we further explore the connections between these two areas can we determine how much progress can be made in each by exploiting the relations reported here.

REFERENCES

Berger, J. O. (1982). *The Robust Bayesian Viewpoint*. Technical Report No. 82-9, Department of Statistics, Purdue University.

Lambert, D. and Duncan, G. (1981). *Bayesian learning based on partial prior information*. Technical Report No. 209, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.

Lindley, D. V. (1971). *Bayesian Statistics: A Review*. Philadelphia:SIAM.

Manski, C. F. (1981). *Learning and decision making when subjective probabilities have subjective domains*. Annals of Statistics 9, 59-65.

Traub, J. F., Wasilkowski, G. W. and Woźniakowski, H. (1981), *Average Case Optimality for Linear Problems*. To appear in Th. Comp. Sci.

_____ (1983) *Information, Uncertainty, Complexity*. Addison-Wesley, Reading, MA.

Traub, J. F. and Woźniakowski, H. (1980). *A General Theory of Optimal Algorithms*, Academic Press, New York, NY.

_____(1983) *Information and Computation*, Dept. of Computer Science Report, Columbia University, New York, NY. To appear in Advances in Computers, Vol. 23, ed. M. Yovitz, Academic Press, 1983.

Wasilkowski, G. W. (1983a). *Local Average Errors*, Dept. of Computer Science Report, Columbia University, New York, NY.

_____(1983b). *Optimal Probabilistic Methods*, in preparation.

Wasilkowski, G. W. and Woźniakowski H. (1982). *Average Case Optimal Algorithms in Hilbert Spaces*. Dept. of Computer Science Report, Columbia University, New York, NY.

Woźniakowski, H. (1982). *Can Adaption Help on the Average?* Dept. of Computer Science Report, Columbia University, New York, NY.