ROCKY MOUNTAIN
JOURNAL OF MATHEMATICS
Volume 2, Number 3, Summer 1972

GAUSSIAN MEASURE IN HILBERT SPACE AND APPLICATIONS IN NUMERICAL ANALYSIS

F. M. LARKIN

ABSTRACT. The numerical analyst is often called upon to estimate a function from a very limited knowledge of its properties (e.g. a finite number of ordinate values). This problem may be made well posed in a variety of ways, but an attractive approach is to regard the required function as a member of a linear space on which a probability measure is constructed, and then use established techniques of probability theory and statistics in order to infer properties of the function from the given information. This formulation agrees with established theory, for the problem of optimal linear approximation (using a Gaussian probability distribution), and also permits the estimation of nonlinear functionals, as well as extension to the case of "noisy" data.

1. **Introduction.** The problem which is central to the subject to be discussed occurs frequently in numerical analysis and the interpretation of experimental data. Typically, we may be given the ordinate values of a function, measured at a finite number of abscissae, and wish to interpolate, i.e. make a reasonable estimate of the function value at some other abscissa. This problem may be regarded in two parts:

(i) Construct an estimator for the required unknown value.

(ii) Determine how accurate the estimated value is likely to be.

The approach traditionally taken by numerical analysts has been to assume an algebraic form for the function in question, e.g., a polynomial of specified order, to determine the assignable parameters (coefficients) by forcing the function to satisfy the given constraints, and subsequently to refer to this constructed function for estimating any other required information.

A statistician, on the other hand, might assume a joint probability distribution, e.g., multivariate normal, for the known and unknown quantities, and thence determine a conditional distribution for the required values.

Each of these approaches has its own advantages and shortcomings. We shall be working towards a generalisation which retains some of

Received by the editors July 1, 1971.

AMS 1970 *subject classifications.* Primary 41A65, 46E20, 65J05; Secondary 28A40, 41A55, 46C20, 60G15, 60G35, 65D10.

the advantages of both points of view, by regarding the interpolating function itself as a random variable in a probability measure space constructed out of a Hilbert space.

Notice that no progress whatsoever can be made unless further assumptions are made about the dependence of the required values upon the given values. Even given the table of values

x_j	y_j
0.0	1.0
1.0	1.0
2.0	1.0
3.0	?
4.0	1.0
5.0	1.0

how confidently can we assert that $y(3.0) = 1.0$? Would we be prepared to bet evens that $|y(3.0) - 1.0| < 0.1$, or < 0.01, or $< 10^{-100}$?! Clearly, some well defined conceptual framework must be established before questions of this kind can be answered, or even *posed,* satisfactorily.

Three main fields of activity can be identified which have a bearing upon the present discussion: Optimal Approximation, Functional Integration, and the Theory of Stochastic Processes.

One part of the story begins with Sard's application (1949) of the techniques of optimal approximation in normed and seminormed spaces to the construction of practical interpolation and quadrature formulae. Since then an enormous volume of literature deriving from Sard's original ideas has been published. Much of this has dealt with spline functions and their use in the approximation of linear functionals, for example, see Ahlberg, Nilson and Walsh (1967), Holladay (1957), Schoenberg (1958, etc.), Golomb and Weinberger (1959) de Boor (1963), Birkhoff and de Boor (1964), and Schultz and Varga (1968). The two last named authors give a bibliography of publications on spline functions.

Mehlum (1964) and Schoenberg (1964) have suggested data smoothing procedures which amount to aesthetic compromises between spline interpolation and least-squares line fitting. The author (1969) ha pointed out that optimal interpolation can be interpreted as maximum likelihood estimation in a Hilbert space of normally distributed functions, and has also (1971) extended the spline smoothing work o Schoenberg as an application of the theory described later in thi paper.

Sard (1963) introduced probabilistic concepts into the theory o linear approximation, for the purpose of estimating the value of a linear functional from approximate values of other linear functionals.

A. V. Sul'din (1959, 1960) has considered minimum variance estimation of the values of linear functionals over the Wiener space of real, continuous functions $\{x(t);\ 0 \leqq t \leqq 1,\ x(0) = 0\}$. The theory of Wiener integration, i.e. integration over this special function space, has received a great deal of attention since the original work by Wiener in the 1920's, possibly because of its usefulness in the applied fields of Statistical and Quantum Mechanics as much as for its intrinsic mathematical interest. I. M. Koval'chik (1963) gives a survey of the field to that date, as well as an extensive bibliography. For physical applications see, for example, Edwards (1967) and Gel'fand and Jaglom (1960).

The theory of stochastic processes, itself intimately related to Wiener integration, provides a third viewpoint on the subject to be discussed. The work of Parzen (1959, etc.) and others, on Time Series Analysis and its formulation within the framework of Hilbert spaces possessing reproducing kernel functions, is described elsewhere in these proceedings. Kimeldorf and Wahba (1968, 1969) have pointed out the connection between spline functions associated with general, linear, differential operators and stochastic processes, and also interpret the smoothing properties of splines in terms of Bayesian estimation on stochastic processes.

Parzen (1970) has also indicated the formal similarity between certain linear estimation problems in approximation theory, stochastic processes and control theory.

With the exception of the work of Sul'din, the concepts and techniques developed in the three above mentioned fields have attracted little attention among numerical analysts. Even Sul'din's work seems to have found little application so far, perhaps because the Wiener space comprises too wide a class of functions for many numerical purposes. The subset of differentiable functions has zero measure and, since optimal approximation involves the selection of a suitable function from a complete space of possible solutions, the Wiener space would seem inappropriate to situations requiring a differentiable approximating function.

However, there are many Hilbert spaces, in particular those associated with the names of Bergman, Payley and Wiener, Sobolev and Szegö, which have already proved useful to numerical analysts, and the present work was motivated by a desire to extend the work of Sul'din to these, and other, function spaces, as well as to investigate nonlinear estimation problems. This desire is reflected by the examples given later, although these are chosen to illustrate techniques rather than to recommend specific formulae for practical use.

In numerical analysis Hilbert spaces possessing reproducing kerne functions are particularly important, for the following reason. It i often convenient to express ones knowledge about a function defined over some domain D, in terms of its ordinate values at prescribed abscissae in D. One takes the view that this information, i it is to be of any value, should increase the precision to which th values of other functionals can be localised. However, it turns ou that a knowledge of the value of an unbounded linear functional ma not contribute much to a knowledge of the value of another linea functional. Thus Hilbert spaces in which ordinate values correspon to bounded, linear functionals — which therefore possess reproducin kernel functions (Aronszajn, 1950) — offer special advantages i numerical analysis.

2. **Optimal approximation in Hilbert space.**

2.1. *The linear approximation problem.* Optimal (or "best") approximation is concerned with the problem of finding linear combinations of known values of linear functionals on a normed, linear spac (usually a Banach space and often a Hilbert space) in order to estimat values of other functionals. In numerical analysis the theory ha found application in the construction of interpolation, quadrature an derivative rules (e.g. Davis, 1963; Handscomb, 1966; Sard, 1963).

In the following we shall be concerned fundamentally with optima approximation in Hilbert spaces, especially those possessing reproducing kernels, since this will form a natural basis for the subsequen probabilistic generalisation.

Suppose we are given numerical values $\{\alpha_j; j = 1, 2, \cdots, n\}$ of th bounded, linear functionals $\{L_j; j = 1, 2, \cdots, n\}$ whose correspondin representers in some Hilbert space H are $\{g_j; j = 1, 2, \cdots, n\}$. Thu the given information relates to some element $h \in H$ satisfying th linear constraints

$$L_j h = (h, g_j) = \alpha_j; \qquad j = 1, 2, \cdots, n. \tag{2.1}$$

Using only this information, can we estimate the value of anothe bounded, linear functional, say $L_0 h$?

The answer to this question is somewhat equivocal since, although is not difficult to construct a reasonable estimator for $L_0 h$, there apparently no way of judging the accuracy of the estimate unle extra information is assumed. Knowledge of the values of a fini number of bounded, linear functionals can localise h only to a line manifold in H (Golomb and Weinberger, 1959).

Let g_0 be the representer of L_0 and consider the "error functional"

$$Rh = L_0h - \sum_{j=1}^{n} w_jL_jh = \left(h, g_0 - \sum_{j=1}^{n} \bar{w}_jg_j \right),$$

where the $\{w_j; j = 1, 2, \cdots, n\}$ are, as yet undetermined, coefficients in the estimation formula

$$L_0h \simeq \sum_{j=1}^{n} w_jL_jh, \quad \forall h \in H.$$

The bar, of course, denotes complex conjugation.

Since the functionals $\{L_j; j = 1, 2, \cdots, n\}$ are all bounded, R is bounded, and we can write

$$|Rh| \leqq \left\| g_0 - \sum_{j=1}^{n} \bar{w}_jg_j \right\| \cdot \|h\|.$$

Thus, a reasonable estimation strategy would be to choose the weights $\{w_j; j = 1, 2, \cdots, n\}$ so as to minimise

$$\|R\| = \left\| g_0 - \sum_{j=1}^{n} \bar{w}_jg_j \right\|,$$

i.e. by projecting g_0 onto the subspace of H spanned by $\{g_j; j = 1, 2, \cdots, n\}$, and then to choose

$$(2.2) \qquad (\widehat{L_0h}) = \sum_{j=1}^{n} w_jL_jh = \sum_{j=1}^{n} w_j\alpha_j = \bar{\underline{\alpha}}'\underline{w}$$

as an estimator of L_0h, given the values $\underline{\alpha} = \{\alpha_j; j = 1, 2, \cdots, n\}$.

Since $\|R\|^2$ is a positive, quadratic form in the weights, the minimisation is easily performed, the final result being

$$(2.3) \qquad \underline{w} = G^{-1}\underline{v}$$

where the nth order Gram matrix G and column vector $\underline{v}$ are defined by

$$\begin{aligned} G_{jk} &= (g_j, g_k), \\ v_j &= (g_j, g_0), \end{aligned} \qquad j, k = 1, 2, \cdots, n.$$

G, of course, is nonsingular provided the elements $\{g_j = 1, 2, \cdots, n\}$ are linearly independent.

2.2. *Solution in terms of an optimal function.* Let us now determine that element $\hat{h} \in H$ of smallest norm, subject to constraints (2.1). Introducing the real Lagrange multipliers $\{\lambda_j, \mu_j; j = 1, 2, \cdots, n\}$ we minimise the quadratic form in h

$$S = \tfrac{1}{2}\ \|h\|^2 - \sum_{j=1}^{n} \lambda_j \cdot \mathrm{Re}[(h, g_j) - \alpha_j]$$

$$+ \sum_{j=1}^{n} \mu_j \cdot \mathrm{Im}[(h, g_j) - \alpha_j]$$

with respect to h. A small change δS in S resulting from an arbitrary small change δh in h will vanish (to first order) only if

$$h = \hat{h} = \sum_{j=1}^{n} (\lambda_j + i\mu_j) g_j = \sum_{j=1}^{n} \nu_j g_j, \quad \text{say.} \tag{2.4}$$

The complex parameters $\underline{\nu} = \{\nu_j; j = 1, 2, \cdots, n\}$ are found by noting that

$$\alpha_k = (h, g_k) = \sum_{j=1}^{n} \nu_j (g_j, g_k)$$

so that $\underline{\bar{\nu}} = G^{-1}\underline{\bar{\alpha}}$. Hence

$$L_0\hat{h} = \sum_{j=1}^{n} \nu_j v_j = \underline{\bar{\nu}}'\underline{v} = \underline{\bar{\alpha}}'G^{-1}\underline{v} = (\widehat{L_0 h}), \tag{2.5}$$

where the prime denotes "complex conjugate transpose".

In other words, for *any* bounded, linear functional the "optimal estimate" $(\widehat{L_0 h})$, defined by (2.2) may be found by applying L_0 to $\hat{h}$ – the element of H having smallest norm subject to the given linear constraints. It is easily verified that the minimal norm in question is given by

$$\|h\|^2 = \underline{\bar{\alpha}}'G^{-1}\underline{\bar{\alpha}} = \underline{\alpha}'\bar{G}^{-1}\underline{\alpha} \tag{2.6}$$

where the prime denotes "complex conjugate transpose".

2.3. *Geometric interpretation.* So far we have merely proposed an intuitively reasonable estimator for $L_0 h$ and can say nothing about the magnitude of $|Rh|$. However, if we make the additional assumption that

$$\|h\|^2 \leqq r^2, \tag{2.7}$$

where r must exceed $\|\hat{h}\|$ in order to ensure compatibility with the linear constraints, it becomes possible to bound $|Rh|$. Interpreting the situation from a geometric viewpoint (see Fig. 2.1) we see that conditions (2.1) constrain h to lie in a *hyperplane* in H, while (2.7) constrains h to lie within a hypersphere, centred on the origin, in H. Provided the constraints are compatible, i.e. the hyperplane cuts the hypersphere, h is constrained to lie in a *hyperdisc* with centre at $\hat{h}$ and radius ρ given by

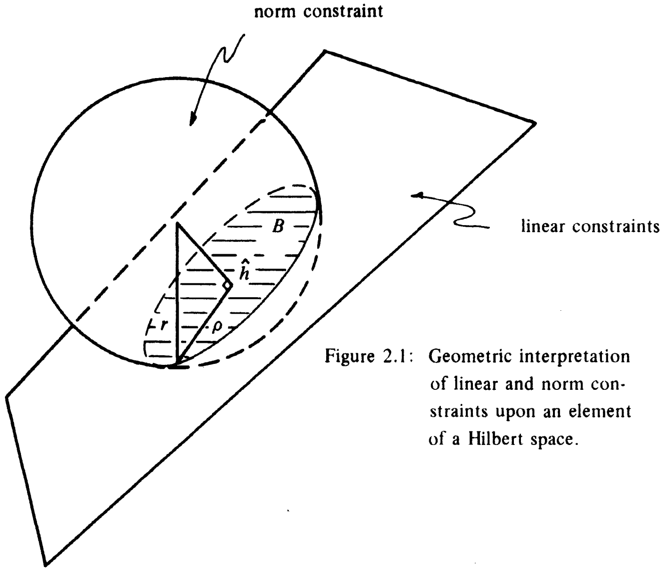

Figure 2.1: Geometric interpretation of linear and norm constraints upon an element of a Hilbert space.

$$\rho = \sqrt{r^2 - \|\hat{h}\|^2}. \tag{2.8}$$

2.4. *Optimal approximation as a minimax problem.* Let us now minimax the error in the required functional subject to the given constraints. Denote the hyperdisc by B, and consider the expression

$$\begin{aligned} e &= \inf_{h_0 \in B} \sup_{h \in B} |(h, g_0) - (h_0, g_0)| \\ &= \inf_{h_0 \in B} \sup_{h \in B} |(h - h_0, g_0^*)| \end{aligned}$$

where g_0^* is the projection of g_0 onto the linear manifold defined by (2.1). By inspection of Figure 2.2, we see that

$$e = \inf_{h_0 \in B} \max \left\{ \left| \left(\hat{h} + \rho \frac{g_0^*}{\|g_0^*\|} - h_0, g_0^* \right) \right|, \right.$$
$$\left. \left| \left(\hat{h} - \rho \frac{g_0^*}{\|g_0^*\|} - h_0, g_0^* \right) \right| \right\}.$$

Thus

$$(2.9) \qquad e = \rho\|g_0^*\| = [r^2 - \|\hat{h}\|^2]^{1/2}[\|g_0\|^2 - |(g_0, \hat{h})|^2/\|\hat{h}\|^2]^{1/2},$$

which occurs when $h_0 = \hat{h}$. Expression (2.9) agrees exactly with the well-known Hypercircle Inequality (e.g. Davis, 1963).

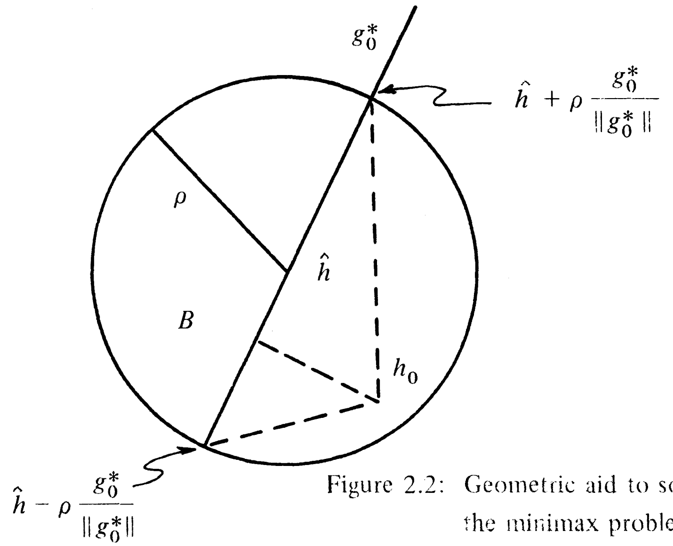
g_0^*
$\hat{h} + \rho \frac{g_0^*}{\|g_0^*\|}$
ρ
$\hat{h}$
B
h_0
$\hat{h} - \rho \frac{g_0^*}{\|g_0^*\|}$

Figure 2.2: Geometric aid to solution of the minimax problem.

Reviewing the situation, we see that $\hat{h}$ and $L_0\hat{h}$ are intuitively reasonable estimators for h and L_0h, respectively, in the case when a bound on $\|h\|$ is known. Even when such a bound is not known these estimators are still reasonable, since $\|h\|$ must be finite for h to be a member of H, but in this case the error bound (2.9) is inapplicable since r is unknown. Notice, however, that it is quite possible for other constraints, e.g.

$$(h, Ah) \leqq 1$$

where A is a positive mapping of H into itself, to preclude $\hat{h}$ from being a reasonable estimator of the solution function. This situation is illustrated in Figure 2.3.

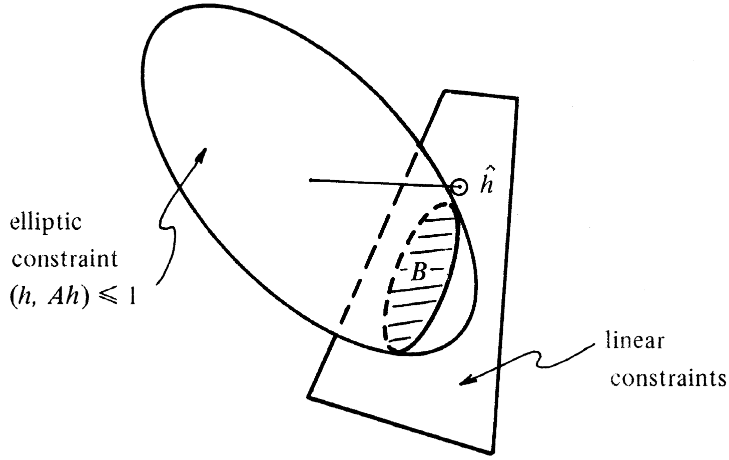

Figure 2.3: Illustration of elliptic constraint, preventing $\hat{h}$ from being a good estimator.

2.5. *Relevance of a reproducing kernel.* Let H be a Hilbert space of functions $\{h(x); x \in D\}$ where D is some real or complex domain. Suppose H possesses the reproducing kernel $K(x, \bar{y})$, so that the ordinate evaluation functionals $\{L_j h = h(x_j), \forall h \in H; j = 1, 2, \cdots, n\}$ are all bounded if D contains the abscissae $\{x_j; j = 1, 2, \cdots, n\}$.

Given an optimal linear rule

$$L_0 h \simeq \sum_{j=1}^{n} w_j L_j h = \sum_{j=1}^{n} w_j h(x_j), \quad \forall h \in H, \tag{2.10}$$

it is of practical interest to know

(i) how to adjust the abscissae so as to minimise the optimal error norm $\|R\|$,

(ii) for what subset of H the optimal estimation rule is exact,

(iii) how rapidly (if at all) $\|R\|$ tends to zero as n tends to infinity.

The first two points are discussed in another paper (Larkin, 1970), the main results of which are as follows:

THEOREM 2.1. *If the distinct abscissae $\{x_j;\ j = 1,\ 2,\ \cdots,\ n\}$ are regarded as fixed in D, the optimal approximation rule* (2.10) *is characterised by the fact that it is exact for the subspace of H spanned by $\{K(\,\cdot\,, \bar{x}_j); j = 1, 2, \cdots, n\}$.*

THEOREM 2.2. *If the weights $\{w_j;\ j = 1, 2, \cdots, n\}$ are prescribed, the optimal approximation rule* (2.10) *has the property that it is exact* *for the subspace of H spanned by* $\{(\partial/\partial\bar{x}_j)K(\cdot, \bar{x}_j); j = 1, 2, \cdots, n\}$ *provided that the derivatives, and distinct abscissae within D, exist.*

Clearly, an approximation rule which is optimal with respect to both weights and abscissae will, in general, be exact for the subspace of H spanned by the $2n$ functions $\{K(\cdot, \bar{x}_j), (\partial/\partial\bar{x}_j)K(\cdot, \bar{x}_j); j = 1, 2, \cdots, n\}$.

In general, the determination of optimal abscissae is a more difficult problem than that of finding optimal weights. However, if we extend the meaning of "optimality with respect to the abscissae" to cover the situation where $\|R\|$ attains a smallest value (not necessarily a minimum) then, as pointed out by Rabinowitz and Richter (1970), existence of optimal abscissae is trivial if D is compact.

Assuming D to be separable and $K(x, \bar{y})$ to be continuous over D^2, it is not difficult to see that $\|R\|$ can be made arbitrarily small if a sufficiently large number of suitable chosen abscissae are used. Suppose the countably infinite sequence $S = \{x_j; j = 1, 2, 3, \cdots\}$ is dense in D, then the set of functions $\{K(\cdot, \bar{x}_j); j = 1, 2, 3, \cdots\}$ is complete in H. [*Proof.* If $g_0(\cdot)$ is the Riesz representer of some bounded, linear functional L_0 and

$$L_0K(\cdot, \bar{x}_j) = (K(\cdot, \bar{x}_j), g_0) = \overline{g_0(x_j)} = 0, \quad \forall x_j \in S;$$

therefore, by the boundness of L_0 and continuity of $K(x, \bar{y})$, $g_0(x)$ vanishes for all x in D.] Hence, by the equivalence of closure and completeness (e.g. Davis, 1963), $\{K(\cdot, \bar{x}_j); j = 1, 2, \cdots\}$ must span H. However, we noted in §2.1 that for fixed abscissae $\{x_j; j = 1, 2, \cdots, n\}$ the optimal weights are found by projecting g_0 onto the subspace of H spanned by $\{K(\cdot, \bar{x}_j); j = 1, 2, \cdots, n\}$ so $\|R\|$ may be made arbitrarily small by a suitably large choice of n.

The asymptotic behaviour of $\|R\|$ for large n will, of course, depend upon the particular linear functional L_0 which is being approximated. An illustration, for the case of optimal approximation with respect to a seminorm, is given in §2.6.

Example. Let H be the Bergman-Hilbert space H of functions analytic within the unit circle in the complex plane, with inner product defined by

$$(f, g) = \iint_{|z|<1} f(z)\overline{g(z)} \cdot dx\, dy; \qquad z = x + iy; \qquad \forall f, g \in H,$$

and reproducing kernel function

$$K(z, \bar{t}) = \pi^{-1}(1 - z\bar{t})^{-2}. \tag{2.11}$$

Let

$$L_0h = \int_{-a}^{a} h(z) \cdot dz; \qquad |a| < 1, \quad \forall h \in H,$$

and let $\{z_j : |z_j| < 1; j = 1, 2, \cdots, n\}$ be given, distinct abscissae. From (2.11) the representer of L_0 is found to be $g_0(z) = 2a\pi^{-1}(1 - a^2z^2)^{-1}$. The optimal weights for the quadrature rule

$$\int_{-a}^{a} h(z) \cdot dz = L_0h \simeq \sum_{j=1}^{n} w_jh(z_j), \quad \forall h \in H, \tag{2.12}$$

satisfy the linear equations $G\underline{w} = \underline{v}$ where the (j, k)th element of the Gram matrix G is given by

$$G_{jk} = \pi^{-1}(1 - \bar{z}_jz_k)^{-2}, \qquad j, k = 1, 2, \cdots, n,$$

and the vector $\underline{v}$ is given by

$$v_j = 2a\pi^{-1}(1 - a^2\bar{z}_j^2)^{-1}, \ a \text{ real}; j = 1, 2, \cdots, n.$$

2.6. *Extension to seminorms.* In practical situations it may be required that a linear approximation rule be exact for some prescribed subspace of functions (e.g. polynomials of order less than some fixed order) and "good" for the remainder of the space. One way of achieving this is to minimise $|Rh|$ subject both to constraints (2.1) and the required exactness conditions.

The problem is conveniently formulated as follows (e.g. Handscomb, 1965):

Let F be a vector space and R a linear functional on F. Let $s(f)$ be a seminorm on F and define the norm of R with respect to s as

$$\|R\| = \sup_{f \in F; s(f) \leqq 1} |Rf|.$$

Thus R can be bounded in this norm only if Rf vanishes on the null space of s.

For the function space $L_p^{(m)}(X)$; $1 \leqq p < \infty$, an extension of the Riesz representation theorem due to Sard (1948) states that any linear functional R on F, which is bounded with respect to the seminorm

$$s(f) = \left[\int_X |f^{(m)}(x)|^p \cdot \mu\{dx\}\right]^{1/p}, \quad \forall f \in F,$$

may be represented in the form

$$Rf = \int_X r(x) \cdot f^{(m)}(x) \cdot \mu\{dx\} \tag{2.13}$$

where $r(x) \in L_{p'}(X)$; $1/p + 1/p' = 1$. The Hölder inequality applied to the integral representation (2.13) enables us to find numerical bounds on $|Rf|$ where the weights $\{w_j; j = 1, 2, \cdots, n\}$ in the optimal linear approximation rule are constrained so that $|Rf|$ vanishes over the null space of $s(\,\cdot\,)$.

In the case of a Hilbert space the theory of reproducing kernels provides another viewpoint on the problem. Suppose the reproducing kernel Hilbert space H is the direct sum of an orthogonal pair of subspaces H_1 and H_2, i.e. any $h \in H$ may be uniquely represented as

$$(2.14) \qquad h = h_1 + h_2; \qquad h_1 \perp H_2 \quad \text{and} \quad h_2 \perp H_1.$$

Suppose furthermore that the inner product over H may be represented as

$$(f, g) = (f, g)_1 + \alpha(f, g)_2; \qquad \alpha > 0, \quad \forall f, g \in H,$$

where $(\,\cdot\,,\,\cdot\,)_1$ and $(\,\cdot\,,\,\cdot\,)_2$ are norms over H_1 and H_2 respectively and seminorms over H. We then have

Lemma. *The reproducing kernel $K_\alpha(x, \bar{y})$ for H may be expressed as*

$$K_\alpha(x, \bar{y}) = K_1(x, \bar{y}) + K_2(x, \bar{y})/\alpha$$

where $K_1(x, \bar{y})$ and $K_2(x, \bar{y})$ are the reproducing kernels for H_1 and H_2 respectively.

Proof. By definition of $K_\alpha(x, \bar{y})$ we have

$$h(x) = (h(\,\cdot\,), K_\alpha(\,\cdot\,, \bar{x}))_1 + \alpha(h(\,\cdot\,), K_\alpha(\,\cdot\,, \bar{x}))_2, \quad \forall h \in H.$$

However, by (2.14) we can decompose $h(\,\cdot\,)$ and $K_\alpha(\,\cdot\,, \bar{x})$ as

$$h(\,\cdot\,) = h_1(\,\cdot\,) + h_2(\,\cdot\,), \qquad K_\alpha(\,\cdot\,, \bar{x}) = K_{\alpha 1}(\,\cdot\,, \bar{x}) + K_{\alpha 2}(\,\cdot\,, \bar{x}),$$

where h_1 and $K_{\alpha 1}$, are orthogonal to H_2, and h_2 and $K_{\alpha 2}$ are orthogonal to H_1. Thus

$$h(x) = (h_1(\,\cdot\,), K_{\alpha 1}(\,\cdot\,, \bar{x}))_1 + \alpha(h_2(\,\cdot\,), K_{\alpha 2}(\,\cdot\,, \bar{x}))_2 = h_1(x) + h_2(x).$$

The only element of H common to H_1 and H_2 is, of course, the zero element, so we can successively take $h_2 = 0$ and $h_1 = 0$ to find

$$h_1(x) = (h_1(\,\cdot\,), K_{\alpha 1}(\,\cdot\,, \bar{x}))_1, \qquad h_2(x) = \alpha(h_2(\,\cdot\,), K_{\alpha 2}(\,\cdot\,, \bar{x}))_2,$$

from which the required result is obvious.

Corollary. *If H can be represented as the direct sum of m orthogonal subspaces $H = \sum_{k=1}^{m} H_k$ and*

$$(f, g) = \sum_{k=1}^{m} \alpha_k (f, g)_k, \qquad \{\alpha_k > 0; k = 1, 2, \cdots, m\}, \quad \forall f, g \in H,$$

where $(\,\cdot\,,\,\cdot\,)_k$ *is a norm over* H_k, $k = 1, 2, \cdots, m$, *then the reproducing kernel for* H *may be expressed as*

$$K(x, \bar{y}) = \sum_{k=1}^{m} \frac{K_k(x, \bar{y})}{\alpha_k}$$

where $K_k(x, \bar{y})$ *is the reproducing kernel for* H_k, $k = 1, 2, \cdots, m$.

Now, allowing $\{\alpha_k; k = 2, 3, \cdots, m\}$ successively to approach zero, and referring to Theorem 2.1, we see that if the weights in the linear approximation rule

$$L_0 h \simeq \sum_{j=1}^{n} w_j h(x_j) \tag{2.15}$$

are chosen to minimise the seminorm $\|g_0 - \sum_{j=1}^{n} w_j K(\,\cdot\,, \bar{x}_j)\|_1$, then the rule will automatically be exact for the functions

$$\{K_k(\,\cdot\,, \bar{x}_j); j = 1, 2, \cdots, n; k = 2, 3, \cdots, m; m \leqq n\}.$$

In particular, if the orthogonal complement of H_1 in H is finite dimensional, there exists m such that $\{H_k; k = 2, 3, \cdots, m\}$ are all 1-dimensional. In this case $K_k(\,\cdot\,, \bar{x}_j)$ trivially spans H_k for all j, k, so the linear approximation rule which is optimal in this sense is necessarily exact for all h in the null space of $\|\,\cdot\,\|_1$.

Thus, by judicious choice of the seminorm $\|\,\cdot\,\|_1$ we can arrange for the optimal linear approximation rule (2.15) to be exact for some finite-dimensional space of specially favoured functions, while having a relatively small error for functions near the origin in its orthogonal complement H_1. This is the situation for the original "best approximation" rules (Sard, 1949), although these were derived by different techniques.

Example. The following example of a practically useful quadrature rule has been discussed in the literature from several different viewpoints (Krylov, 1959; Stern, 1967; Larkin, 1970). Let $H = L_2^{(2)}[0, 1]$ with inner product defined by

$$(f, g) = \int_0^1 f''(x) \cdot g''(x) \cdot dx + \alpha_2 f(0) g(0) + \alpha_3 f'(0) g'(0).$$

Let H_2 be the space of constant functions on $[0, 1]$ and H_3 the space of straight lines through the origin; H_1 is the orthogonal complement of $H_2 \oplus H_3$ in H_1, i.e. that subspace of $L_2^{(2)}[0, 1]$ whose members have vanishing ordinate and derivatives at the origin.

The inner product in H_2 is defined by

$$(f, g)_2 = f(0)g(0)$$

and the reproducing kernel is the unit function on $[0, 1]$. The inner product in H_3 is defined by

$$(f, g)_3 = f'(0)g'(0)$$

and its reproducing kernel is xy. The inner product in H_1 is defined by

$$(f, g)_1 = \int_0^1 f''(x)g''(x) \cdot dx,$$

and its reproducing kernel is

$$K_1(x, y) = \begin{cases} -y^3/6 + y^2x/2; & 0 \leqq y \leqq x, \\ -x^3/6 + x^2y/2; & x \leqq y \leqq 1. \end{cases}$$

The reproducing kernel for H is given by

$$K(x, y) = \begin{cases} -y^3/6 + y^2x/2 + yx/\alpha_3 + 1/\alpha_2; & 0 \leqq y \leqq x, \\ -x^3/6 + x^2y/2 + xy/\alpha_3 + 1/\alpha_2; & x \leqq y \leqq 1. \end{cases}$$

Defining L_0 by

$$L_0 h = \int_0^1 h(x) \cdot dx, \quad \forall h \in H,$$

we find

$$g_0(x) = x^4/24 - x^3/6 + x^2/4 + x/2\alpha_3 + 1/\alpha_2.$$

The optimal weights and abscissae, which minimise the seminorm

$$\|R\|_1 = \left\| g_0(\,\cdot\,) - \sum_{j=1}^{n} w_j K(\,\cdot\,, x_j) \right\|_1,$$

are given by

$$\begin{aligned} x_1 &= \tfrac{1}{2} - (n-1)h/2; & w_1 &= \tfrac{1}{2} - (n/2 - 1)h = w_n, \\ x_j &= x_1 + (j-1)h; & w_j &= h; j = 2, 3, \cdots, n-1; \\ x_n &= \tfrac{1}{2} + (n-1)h, \end{aligned}$$

where the "step-length" parameter h is defined by

$$h = (n - 1 + \sqrt{2/3})^{-1}.$$

The minimal value of $\|R\|_1$ can be expressed in terms of h, by substituting in the optimal abscissae and weight values, to obtain

$$\|R\|_{1,\min} = h^2/12\sqrt{5},$$

confirming the result given by Stern (1967).

2.7. *Limitations of optimal approximation.* We have seen that the theory of optimal linear approximation provides a useful framework for the construction of numerical approximation rules, of which quadrature rules form the typical example. However, there is an important class of problems which, although superficially amenable to the techniques of optimal approximation, actually serves to emphasise its limitations.

Leaving aside possible computational difficulties, which in practice seem no more severe with optimal approximation than with most other useful techniques, the three main areas of deficiency (which incidentally are shared by other approximation techniques) are as follows:

(i) The error functional R must be linear. There are important practical situations, especially occurring in the experimental sciences, in which a Hilbert space formulation necessitates that the given information, the required information and the estimation error all be treated as nonlinear quantities. Optimal approximation in its linear form is thus inapplicable to this kind of problem. An illustration is given below.

(ii) Often the given information is not only insufficient to determine the required information but is also inexact. We might elect to "smooth" the given information in some aesthetic sense, or even make a statistical estimate of the true values of the given quantities, but would this be the right approach? Under what circumstances would this preprocessing be an appropriate preliminary to optimal approximation? In its classical form the theory of optimal approximation gives no clue as to how best to extract the best possible signal from given, noisy data.

(iii) We know that, even for the problem of linear approximation, extra, nonlinear information is required in order to bound the error of approximation; typically, one requires a bound on the norm of the function sought. Often this information is either not precise enough to give realistic error bounds, or is simply not available. It would be useful to have a theory which could provide an error estimate, even of a probabilistic nature, from the "working" information alone.

By way of illustration, consider the following example, which is a simplified version of one discussed elsewhere (Larkin, 1969). Suppose an experimental scientist wishes to estimate a function $f(x)$, or one of its attributes, such as its integral over a range $[a, b]$. From fundamental physical considerations $f(x)$ is known to be nonnegative (e.g. $f(x)$ may represent the density of some physical quantity like mass, heat, or some other form of energy). Measurements are made in an attempt to determine the ordinates $\{f(x_j); a \leqq x_j \leqq b; j = 1, 2, \cdots, n\}$ but the corresponding values obtained $\{f_j^*; j = 1, 2, \cdots, n\}$ are inaccurate for two reasons:

(a) The measuring instrument has a less than perfect resolution; instead of measuring $f(x_j)$, under ideal (noise-free) circumstances it measures the values

$$Q_jf = \int_a^b q_j(x)f(x) \cdot dx; \qquad j = 1, 2, \cdots, n,$$

where the $\{q_j(\,\cdot\,); j = 1, 2, \cdots, n\}$ are known nonnegative functions.

(b) The experimental circumstances are not ideal; that is to say, the values $\{f_j^*; j = 1, 2, \cdots, n\}$ are compounded of the values $\{Q_jf; j = 1, 2, \cdots, n\}$ together with extraneous "noise" of a probabilistic nature.

Since the observations and the required integral all represent linear operations on f, we might be tempted to regard f as an element of a linear space and then use the techniques of optimal approximation. However, it is important to realise that any approximation rule we construct should not imply an $f(x)$ which becomes negative anywhere within its domain, *whatever (physically realisable) set of observations the rule may be used upon.* There could well exist a set of positive values for $\{Q_jf; j = 1, 2, \cdots, n\}$ which would imply an inadmissable estimator $\hat{f}$ constructed by the methods discussed earlier, indicating that optimal, linear approximation would not be appropriate to this problem. In addition, of course, we have to devise a justifiable technique for filtering out the signal from the noise.

A Hilbert space formulation of the above problem could be achieved in many ways, perhaps the simplest being to write

$$f(x) = h^2(x); \qquad a \leqq x \leqq b,$$

where $h(\,\cdot\,)$ is taken to be an element of some real Hilbert space H of functions with domain $[a, b]$. In this case, the quantities $\{Q_jf; j = 1, 2, \cdots, n\}$ and the required integral all become positive, quadratic functionals on H.

Thus, we see that a satisfactory treatment of this problem of the interpretation of experimental data involves both probabilistic concepts and essential nonlinearity, two features which are notably absent from the usual theory of optimal approximation.

3. **Gaussian measure on the Hilbert space.**

3.1. *General philosophy.* A natural way of introducing the required probabilistic concepts into the theory of optimal approximation would appear to be to construct a probability measure space on the Hilbert space in which the approximation is being performed. A Gaussian measure would be preferred, partly for its mathematical convenience and partly for its intuitive attractiveness in assigning relatively high likelihood to "smooth" functions (i.e. functions of small norm) and in preserving the "independence" of sections of functions in $L_2(X)$ over disjoint subsets of their domain X.

However, it turns out that an infinite-dimensional Hilbert space cannot support a fully countably additive Gaussian measure. In order to retain countable additivity, and with it the complete apparatus of Lebesgue integration theory, it is necessary to extend the Hilbert space H, by completing it with respect to a "measurable norm" (Gross, 1962), to form a Banach space B. This elegant solution to the countable additivity problem was proposed by Gross and discussed in a series of papers (1960, 1962, 1963, 1967). For convenience of presentation a companion paper (Kuelbs, Larkin and Williamson, 1971) gives an outline of that part of the theory of Gaussian measure in separable Hilbert spaces, and derived Banach spaces, necessary as a foundation for the applications described below. Here we shall merely use the relevant concepts, definitions and results, as required.

The idea of Sul'din was to evaluate the mean-square error in a linear approximation rule, evaluated over the Wiener measure space of continuous functions on the real segment $[0, 1]$ and to minimise this with respect to the weights and/or abscissae (possibly subject to exactness conditions for low order polynomials). Interpreted in the light of Gross' theory, the underlying Hilbert space H is the space of absolutely continuous functions, vanishing at the origin, with square-integrable derivatives on $[0, 1]$, and inner product given by

$$(f, g) = \int_0^1 f'(x)g'(x) \cdot dx, \quad \forall f, g \in H.$$

The norm

$$\|f\|_1 = \sup_{x \in [0, 1]} |f|$$

is measurable on H, and B, the completion of H with respect to $\| \cdot \|_1$ is the space of continuous functions on $[0, 1]$, which is known to support the countably additive Gauss-Wiener measure.

As it happens, the approximation rules resulting from this program are nothing other than rules for optimal approximation in the aforementioned Hilbert space. However, once adjusted to the idea of probability distributions over a function space we can cheerfully contemplate use of standard statistical techniques, such as determination of the joint, finite-dimensional distribution of the known and unknown quantities, and construction of maximum likelihood or minimum variance estimators. For example, an estimator for the integral of the nonnegative function f, in the problem at the end of §2, in the form

$$\int_a^b f(x) \cdot dx \simeq \sum_{j=1}^n w_j \int_a^b q_j(x) \cdot f(x) \cdot dx$$

may be constructed by minimising the functional integral

$$\int_B \left| \int_a^b h^2(x) \left[1 - \sum_{j=1}^n w_j q_j(x) \right] \cdot dx \right|^2 \cdot \mu\{dh\}$$

with respect to the weights. Here B is, of course, the completion of H with respect to some suitable (measurable) norm.

3.2. *Evaluation of functional integrals.* The companion paper gives conditions under which an integral over B results as the limit of a sequence of integrals over finite-dimensional subspaces of the underlying (real) Hilbert space H, i.e. for the validity of the relation

$$\begin{aligned} \int_H F(h) \cdot \nu\{dh\} &= \lim_{n\to\infty} \int_H F(P_n h) \cdot \nu\{dh\} \\ &= \lim_{n\to\infty} \int_B F(P_n h) \cdot \mu\{dh\} = \int_B F(h) \cdot \mu\{dh\}, \end{aligned} \tag{3.1}$$

where $\{P_n; n = 1, 2, 3, \cdots\}$ is an increasing sequence of projections converging to the identity, and $\nu\{\cdot\}$ and $\mu\{\cdot\}$ are Gaussian measures appropriately defined on H and B, respectively. Alternative expressions for the general element of the integral sequence are easily found by the usual transformation rules for finite-dimensional integrals; some useful examples follow.

We shall denote the n-fold Lebesgue integral

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} A(x_1, x_2, \cdots, x_n) \cdot dx_1 dx_2 \cdots dx_n \quad \text{by} \quad \int^{(n)} A(\underline{x}) \cdot \underline{dx}.$$

Let $F(\cdot)$ be a functional on H, λ a real, positive constant and $\underline{g} = \{g_j;\ j = 1, 2, 3, \cdots\}$ a (not necessarily orthonormal) basis for H. Let P_n be the operation of projection of elements in H onto the n-dimensional subspace spanned by $\underline{g}_n = \{g_1, g_2, \cdots, g_n\}$. Let $\underline{x} = \{x_1, x_2, \cdots, x_n\}$ be an nth order vector of real numbers and define $\hat{h}$ to be that element of H with minimal norm, subject to the constraints

$$(\hat{h}, g_j) = x_j; \qquad j = 1, 2, \cdots, n. \tag{3.2}$$

The minimising element $\hat{h}$ may be expressed as $\hat{h} = \underline{x}'G^{-1}\underline{g}_n$, where the nth order Gram matrix G has (g_j, g_k) for its (j, k)th element. We also have $\|\hat{h}\|^2 = \underline{x}'G^{-1}\underline{x}$.

Referring to the measure definitions given in the companion paper, it may be verified that

$$\begin{aligned}\int_H F(P_n h)\cdot\nu\{dh\} &= \frac{\int^{(n)} F(\hat{h})\exp[-(\lambda/2)\|\hat{h}\|^2]\cdot\underline{dx}}{\int^{(n)}\exp[-(\lambda/2)\|\hat{h}\|^2]\cdot\underline{dx}}\\ &= \int^{(n)} F(\sqrt{2\pi/\lambda}\,\underline{t}'\underline{f}_n)\exp[-\pi\underline{t}'\underline{t}]\cdot\underline{dt},\end{aligned} \tag{3.3}$$

where $\underline{f}_n = \{f_1, f_2, \cdots, f_n\}$ is a Gram orthonormalisation of $\underline{g}_n$. The positive, real number λ will be called the "dispersion parameter" of the weak distribution on H.

Let T_n be the continuous, linear transformation mapping H onto R^n defined by the relations

$$\begin{cases} T_n : h \rightarrow \underline{x} \\ \quad (h, g_j) = x_j; \qquad j = 1, 2, \cdots, n. \end{cases}$$

T_n has the generalised inverse $T_n{}^+$, i.e. that mapping from R^n into H which associates with any $\underline{x} \in R^n$ the element $\hat{h} \in H$ of smallest norm subject to constraints (3.2). In terms of this generalised inverse, we have

$$\int_H F(P_n h)\cdot\nu\{dh\} = \frac{\int^{(n)} F(T_n{}^+\underline{x})\exp[-(\lambda/2)\|T_n{}^+\underline{x}\|^2]\cdot\underline{dx}}{\int^{(n)}\exp[-(\lambda/2)\|T_n{}^+\underline{x}\|^2]\cdot\underline{dx}}. \tag{3.4}$$

For typographical convenience we shall sometimes denote the quantity

$$\int_H F(P_n h)\cdot\nu\{dh\} \quad\text{by}\quad E_n[F; \underline{g}].$$

In the case of a complex Hilbert space $E_n[F; \underline{g}]$ may be constructed in a manner formally equivalent to that in the real case. However, the vectors $\underline{x}$, $\underline{g}$, $\underline{t}$ and $\underline{f}$ are now complex and we find that the nth order real integral becomes a 2nth order integral

$$E_n[F,\underline{g}] = \int^{(2n)} F(\sqrt{2\pi/\lambda}\,[\underline{t}_R{}'\underline{f}_R - \underline{t}_I{}'\underline{f}_I + i(\underline{t}_R{}'\underline{f}_I + \underline{t}_I{}'\underline{f}_R)])$$
$$\times \exp[-\pi(\underline{t}_R{}'\underline{t}_R + \underline{t}_I{}'\underline{t}_I)] \cdot \underline{dt}_R \cdot \underline{dt}_I,$$

using the obvious notation

$$\underline{t} = \underline{t}_R + i\underline{t}_I, \qquad \underline{f} = \underline{f}_R + i\underline{f}_I,$$

where $\underline{t}_R$, $\underline{t}_I$, $\underline{f}_R$ and $\underline{f}_I$ are all real vectors. In particular, the measure of a cylinder set may be expressed in this form by choosing $F(\,\cdot\,)$ to be an indicator function. Strictly speaking, Gross's theory should be extended to cover the case of complex H; however, in this paper we restrict our complex examples to tame functions, for which no difficulty arises.

Example 1. Let us evaluate the measure of the set $S = \{h \in H : (h, g) > \alpha\}$ for some fixed $g \in H$ and real α. H is real in these examples.

The indicator function I_s of the set S is a tame function based on the one-dimensional subspace of H spanned by g, so that

$$\begin{aligned} \nu\{S\} &= \int_H I_S(h)\cdot\nu\{dh\} \\ &= \int_{-\infty}^{\infty} I_S(\sqrt{2\pi/\lambda}\cdot tg/\|g\|)\exp[-\pi t^2]\cdot dt \qquad (3.5) \\ &= \int_{(\alpha/\|g\|)\cdot\sqrt{\lambda/2\pi}}^{\infty} \exp[-\pi t^2]\cdot dt \\ &= \|g\|^{-1}\sqrt{\lambda/2\pi}\int_{\alpha}^{\infty}\exp[-\lambda t^2/2\|g\|^2]\cdot dt. \end{aligned}$$

Notice that the result depends only upon the *norm* of g, not its direction.

Example 2. For any fixed $g \in H$, (h, g) is a tame function, and so is $e^{is(h,g)}$ for any real s. We thus have

$$\begin{aligned} \int_H \exp[is(h,g)]\cdot\nu\{dh\} &= E_1[\exp[is(h,g)];g] \\ (3.6) \qquad &= \int_{-\infty}^{\infty}\exp[is\sqrt{2\pi/\lambda}\cdot\|g\|t - \pi t^2] \\ &= \exp\left[-\frac{s^2\|g\|^2}{2\lambda}\right]. \end{aligned}$$

Hence, identifying coefficients of powers of t in the absolutely convergent Taylor expansions for the exponential function, we find that

$$(3.7) \qquad \int_H (h, g)^{2n} \cdot \nu\{dh\} = \frac{(2n)!}{n!} \cdot \frac{\|g\|^{2n}}{(2\lambda)^n},$$

while expectations of odd powers of (h, g) vanish identically.

EXAMPLE 3. Let A be a finite-trace class operator mapping H onto itself, then the functional $\|h\|_1 = (h, Ah)^{1/2}$, $\forall h \in H$, is a measurable norm on H (Gross, 1960). The functional

$$F(h) = \exp\,[is(h, Ah)], \quad s \text{ real},$$

has a continuous extension to B, the Banach space formed by completing H with respect to $\|\cdot\|_1$. Hence formula (3.1) is valid for this $F(\cdot)$.

Let $\underline{b} = \{b_j; j = 1, 2, 3, \cdots\}$ be the complete orthonormal sequence of eigenfunctions of A, with corresponding, nonzero eigenvalues $\underline{\mu} = \{\mu_j; j = 1, 2, 3, \cdots\}$. We then have

$$E_n[F; \underline{b}] = \int^{(n)} \exp\left[i\frac{2\pi}{\lambda} \cdot \underline{t}'Q\underline{t} - \pi\underline{t}'\underline{t}\right] \cdot \underline{dt},$$

where Q is the nth order matrix whose (j, k)th element is given by

$$Q_{jk} = (b_k, Ab_j) = \delta_{jk}\mu_j; \qquad j, k = 1, 2, \cdots, n.$$

Thus

$$\begin{aligned} &E_n[\exp\,[is(h, Ah)]; \underline{b}] \\ (3.8) \qquad &= \prod_{j=1}^{n} \int_{-\infty}^{\infty} \exp\left[-\pi\left(1 - \frac{2is}{\lambda}\mu_j\right)t^2\right] \cdot dt \\ &= \prod_{j=1}^{n} \left(1 - \frac{2is}{\lambda}\mu_j\right)^{-1/2}. \end{aligned}$$

The infinite product converges since A has finite trace and, in particular, identification of coefficients of s leads to the conclusion

$$(3.9) \qquad \int_H (h, Ah) \cdot \nu\{dh\} = \int_B (h, Ah) \cdot \mu\{dh\} = \frac{\text{trace}\,[A]}{\lambda}.$$

3.3. *The relative likelihood of a function.* Although we have noted that a Hilbert space H cannot support a fully countably additive Gaussian measure, the work of Gross shows that, at least for the purpose of averaging a usefully large class of functionals on H, such additivity as the weak Gaussian distribution on H possesses is "good enough". Accordingly, it is tempting, and intuitively convenient, to regard $\exp[-(\lambda/2)\|h\|^2]$ as the "relative likelihood" of h in H.

Clearly the relative likelihood of h is maximised when $\|h\|$ is minimised, so an optimal approximation rule may be regarded as a maximum likelihood estimator for a bounded, linear functional. The minimising element $\hat{h}$ referred to in earlier sections may be thought of as the "most likely" element of H which could have resulted in the given values of the given bounded, linear functionals. We shall see later that, as one would expect in the case of a Gaussian distribution, the minimum variance estimator agrees with the maximum likelihood estimator of a bounded, linear functional, i.e. it also agrees with optimal approximation theory.

3.4. *Distributions of linearly transformed quantities.* Let L be a linear operator having the Hilbert space H as its domain. We define an inner product on the range $R[L]$ by means of the relation

$$(3.10) \qquad (f_1, g_1)_{R[L]} = (L^+f_1, L^+g_1)_H; \quad \forall f_1, g_1 \in R[L],$$

where, for any $h_1 \in R[L]$, L^+h is that element $\hat{h} \in H$ with smallest norm satisfying the relation $L\hat{h} = h_1$. Notice that

$$\|L\| = \sup_{h \in H} \frac{\|Lh\|_{R[L]}}{\|h\|_H} = \sup_{h \in H} \frac{\|L^+Lh\|_H}{\|h\|_H} \leqq 1,$$

since L^+L is a projection operator on H. Hence L is a bounded, linear mapping from H to the linear space $R[L]$, and it follows that $R[L]$ must be complete. Thus $R[L]$ is a Hilbert space, which we shall denote by H_1, and L^+ is the generalised inverse of L.

THEOREM 3.1. *Given a canonical normal cylinder set measure on H, the mapping $L: H \to H_1$ induces a canonical normal cylinder set measure on H_1, having the same dispersion parameter as that on H.*

PROOF. We have to show

(a) The inverse image under L of any cylinder set $C_1 \subset H_1$ is a cylinder set $C \subset H$, and

(b) $\nu_1\{C_1\} = \nu\{C\}$ where ν and ν_1 are the canonical normal cylinder set measures on H and H_1, respectively.

For simplicity of presentation we restrict H to a real Hilbert space. Since ν and ν_1 are finitely additive over their respective cylinder set algebras, it suffices to demonstrate (a) and (b) for the case

$$C_1 = \{h_1 \in H_1 : (h_1, g_1)_{H_1} \; \Delta \; \alpha\}$$

where α is a real number and Δ may denote any one of the relational symbols $\{<, \leqq, \geqq, >\}$.

Consider the set

$$\begin{aligned} S &= \{h \in H : (h, [L^+L]^*L^+g_1)_H \;\Delta\; \alpha\} \\ &= \{h \in H : (L^+Lh, L^+g_1)_H \;\Delta\; \alpha\} \\ &= \{h \in H : (Lh, g_1)_{H_1} \;\Delta\; \alpha\} = C. \end{aligned}$$

Thus C is a cylinder set in H.

Furthermore, from result (3.5) we know that

$$\nu_1\{C_1\} = \begin{cases} \displaystyle\int_{(\alpha/\|g\|)\cdot\sqrt{\lambda/2\pi}}^{\infty} \exp[-\pi t^2] \cdot dt & \text{if } \Delta = \geqq \text{ or } >, \\ \displaystyle\int_{-\infty}^{(\alpha/\|g\|)\cdot\sqrt{\lambda/2\pi}} \exp[-\pi t^2] \cdot dt & \text{if } \Delta = \leqq \text{ or } <, \end{cases}$$

and a similar result for $\nu\{C\}$. Also, writing $g = [L^+L]^*L^+g_1$ we have

$$\begin{aligned} \|g\|_H^2 &= ([L^+L]^*L^+g_1, [L^+L]^*L^+g_1)_H \\ &= (L^+g_1, [L^+L]^*L^+g_1)_H \end{aligned}$$

since $[L^+L]$ is a projection operator on H. Thus

$$\|g\|_H^2 = (L^+LL^+g_1, L^+g_1)_H = (LL^+g_1, g_1)_{H_1} = \|g_1\|_{H_1}^2$$

since LL^+ is the identity operator on H_1. Therefore $\nu_1\{C_1\} = \nu\{C\}$ and the required result is proved.

In a rather more intuitive fashion, the above result may be paraphrased as:

Theorem 3.1. *If the relative likelihood of h in H is* $\exp(-(\lambda/2)\|h\|^2)$ *and L is a linear mapping from H onto H_1 (where the inner product on H_1 is defined by relation* (3.10)) *then the relative likelihood of $Lh = h_1$ in H_1 is* $\exp(-(\lambda/2)\|L^+h_1\|^2)$.

Example. Let $\underline{g} = \{g_1, g_2, \cdots, g_n\}$ be linearly independent elements of a real Hilbert space H, and take H_1 to be the real n-dimensional Euclidean space R^n.

Define the mapping $L : H \rightarrow H_1$ by means of the relation

$$\begin{aligned} Lh &= \{(h, g_1), (h, g_2), \cdots, (h, g_n)\} \\ &= \{x_1, x_2, \cdots, x_n\} = \underline{x}, \quad \forall h \in H. \end{aligned} \tag{3.11}$$

The element $\hat{h}$ in H having smallest norm subject to (3.11), for some fixed $\underline{x}$, is given by $\hat{h} = \underline{x}'G^{-1}\underline{g}$ where the (j, k)th element of the Gram matrix G is (g_j, g_k). Hence

$$\|\hat{h}\|^2 = \|L^+\underline{x}\|^2 = \underline{x}'G^{-1}\underline{x}$$

and the relative likelihood of $\underline{x}$ in R^n is $\exp(-(\lambda/2)\underline{x}'G^{-1}\underline{x})$. However, the canonical normal distribution is countably additive on the finite-dimensional space R^n, so a genuine probability density function exists, given by

$$\rho(\underline{x}) = \left(\frac{\lambda}{2\pi}\right)^{n/2} \cdot \frac{\exp[-(\lambda/2)\underline{x}'G^{-1}\underline{x}]}{|G|^{1/2}}. \tag{3.12}$$

This result has applications to the problem of estimating the value of a given linear functional from given values of a finite number of other linear functionals, with or without the extra complication of "noise" affecting the data. Those applications are discussed in another paper (Larkin, 1971).

3.5. *Nonlinear transformations.* We consider first the problem of finding the distribution of a real, scalar, nonlinear functional on H. Suppose there exists a measurable norm $\|\cdot\|_1$ on H with respect to which the real functional $F(h)$ is continuous. The integral over B, the completion of H with respect to $\|\cdot\|_1$, of $\exp[itF(h)]$ then exists for any real t and may be found by means of relation (3.1).

THEOREM 3.2. *The characteristic function of the probability distribution of F is given by*

$$\phi_F(t) = \int_B \exp[itF(h)] \cdot \mu\{dh\}.$$

PROOF. Consider

$$J = \frac{1}{\pi}\int_{-T}^{T} \frac{\text{Sin}(\alpha t)}{t} \cdot \exp[-itx] \cdot \phi_F(t) \cdot dt$$

$$= \frac{1}{\pi}\int_{-T}^{T} \frac{\text{Sin}(\alpha t)}{t} \cdot \exp[-itx] \int_B \exp[itF(h)] \cdot \mu\{dh\} \cdot dt.$$

The conditions for reversing the order of integrations apply, so that

$$J = \frac{1}{\pi}\int_B \int_{-T}^{T} \frac{\text{Sin}(\alpha t)}{t} \cdot \exp[it(F(h) - x)] \cdot dt \cdot \mu\{dh\},$$

i.e.

$$J = \int_B g(F, T) \cdot \mu\{dh\},$$

where

$$g(F, T) = \frac{2}{\pi} \int_0^T \text{Sin}\,(\alpha t) \cdot \text{Cos}\,[t(F(h) - x)] \cdot \frac{dt}{t}\,.$$

It may be verified that

$$\lim_{T\to\infty} \; g(F, t) = \begin{cases} 0; & F < x - \alpha, \\ \frac{1}{2}; & F = x - \alpha, \\ 1; & x - \alpha < F < x + \alpha, \\ \frac{1}{2}; & F = x + \alpha, \\ 0; & F > x + \alpha, \end{cases}$$

and hence, $F(\,\cdot\,)$ being continuous and therefore measurable on B,

$$\begin{aligned} \lim_{T\to\infty} \; J &= \int_B I(h : x - \alpha < F(h) < x + \alpha) \cdot \mu\{dh\} \\ &= \mu\{h \in B : x - \alpha < F < x + \alpha\}. \end{aligned}$$

Thus, the probability density function for $x = F(h)$ is given by

$$\begin{aligned} \rho(x) &= \lim_{\alpha\to 0} \lim_{T\to\infty} \frac{1}{2\pi\alpha} \int_{-T}^{T} \frac{\text{Sin}\,(\alpha t)}{t} \cdot \exp\,[-\,itx]\,\phi_F(t) \cdot dt \\ &= \lim_{\alpha\to 0} \lim_{T\to\infty} \frac{1}{4\pi\alpha} \int_{-T}^{T} \int_{-\alpha}^{\alpha} \exp\,[it(u - x)]\,\phi_F(t) \cdot du \cdot dt, \end{aligned}$$

i.e.

$$\begin{aligned} \rho(x) &= \lim_{\alpha\to 0} \frac{1}{2\alpha} \int_{-\alpha}^{\alpha} \xi(x - u) \cdot du \\ &= \lim_{\alpha\to 0} \frac{1}{2\alpha} \int_{x-\alpha}^{x+\alpha} \xi(v) \cdot dv \end{aligned}$$

where

$$\xi(v) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \phi_F(t) \exp\,[-\,itv] \cdot dt.$$

Thus

$$\phi_F(t) = \int_B \exp\,[itF(h)] \cdot \mu\{dh\}$$

is the characteristic function for F, as required.

Using similar arguments it may be shown that, if $\underline{F} = \{F_1(h), F_2(h), \cdots, F_n(h)\}$ is a finite order vector whose elements satisfy the same conditions as $F(\cdot)$ above, then:

THEOREM 3.3. *The characteristic function of $\underline{F}$ is given by*

$$\phi_{\underline{F}}(\underline{t}) = \int_B \exp\left[i\underline{t}'\underline{F}(h)\right] \cdot \mu\{dh\}.$$

EXAMPLE. Let A be a finite-trace class operator mapping H onto itself; then, as shown in a previous example (relation (3.8)), the characteristic function of the scalar functional $F = (h, Ah)$ is given by

$$\phi_F(t) = \prod_{j=1}^{\infty} \left(1 - \frac{2it}{\lambda} \cdot \mu_j\right)^{-1/2}$$

where $\{\mu_j; j = 1, 2, 3, \cdots\}$ are the eigenvalues of A.

4. **Stochastic processes.** For the sake of completeness we give here a brief indication of the relevance of functional integration to the theory of stochastic processes. Let H be a real Hilbert space of functions $\{h(x) : x \in X\}$ for some domain X, and suppose H possesses the continuous reproducing kernel $K(x, y) : x, y \in X$. If the norm $\|h\|_1 = \sup_{x\in X}|h(x)|$; $\forall h \in H$, is measurable, then H can be completed with respect to this norm to form a Banach space B which is capable of supporting a countably additive Gaussian measure $\mu\{\cdot\}$. The ordinate evaluation functionals are all bounded on B and $\{h(x) : x \in X : h \in B\}$ is a Gaussian stochastic process.

The covariance kernel of the process is easily found. Since $h(x)h(y)$, for fixed $x, y \in X$, is a tame functional based on the two-dimensional space spanned by $K(\cdot, x)$ and $K(\cdot, y)$ we have

$$\begin{aligned} C(x, y) &= \int_B h(x)h(y) \cdot \mu\{dh\} \\ &= \left[\frac{\lambda}{2\pi|G|}\right]^{1/2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_1h_2 \exp\left(-\frac{\lambda}{2}[h_1h_2]G^{-1}\begin{bmatrix} h_1 \\ h_2 \end{bmatrix}\right) \\ &\qquad \cdot dh_1\, dh_2 \end{aligned}$$

where

$$G = \begin{bmatrix} K(x, x) & K(y, x) \\ K(x, y) & K(y, y) \end{bmatrix}.$$

Hence $C(x, y) = K(x, y)/\lambda$.

Conversely, given a finite, positive-definite covariance function $C(x, y)$ on X^2, and a positive dispersion parameter λ, there is a unique Hilbert space of functions on X with reproducing kernel $K(x, y) = \lambda C(x, y)$, (Moore's Theorem). A measure space supporting the Gaussian stochastic process with the given covariance function may then be constructed as described earlier.

For example, consider the Hilbert space H of real, absolutely continuous functions on $[0, 1]$ which vanish at zero and have square-integrable first derivative, with inner product given by

$$(f, g) = \int_0^1 f'(x) \cdot g'(x) \cdot dx; \quad \forall f, g \in H.$$

The norm defined by

$$\|h\|_1 = \sup_{x \in [0,1]} |h(x)|; \quad \forall h \in H,$$

is measurable on H, and completion of H with respect to $\| \cdot \|_1$ leads to the usual Wiener space B of continuous functions on $[0, 1]$.

In particular, taking $F(h) = [(h(x) - h(y))/(x - y)^\alpha]^2$ we find $\int_B F(h) \cdot \mu\{dh\} = \lambda^{-1}|x - y|^{1-2\alpha}$, which approaches zero or infinity as y approaches x, according as α is less than or greater than $\frac{1}{2}$. In other words, although B consists entirely of continuous functions, almost none of these are differentiable anywhere.

5. Applications in numerical analysis.

5.1. *Type of problem considered.* Without wishing to offer a formal definition of the term "numerical analysis", we consider two broad classes of problems with which the numerical analyst finds himself confronted:

(a) Problems with complete information. These include properly posed linear or nonlinear algebraic or differential equations whose true solutions can, in principle, be approximated arbitrarily accurately by means of a sufficiently large number of exact arithmetic operations.

(b) Problems with incomplete information. Many examples arise in connection with the interpretation of experimental measurements. For example, one may be presented with a finite number of ordinate values, measured at given abscissae but known to be subject to observational error, and be required to estimate ordinate values at intermediate abscissae.

We shall be concerned with applying the foregoing mathematical apparatus to the formulation and numerical solution of problems of type (b). However, it is worth noting that some, if not all, of type (a) problems may usefully be treated as if they were of type (b). For example, when performing a digital iteration to a zero of a known function one can never perform more than a finite number of arithmetic operations. At any stage one can ask "On the basis of the information I now have (i.e. a finite number of given and/or computed values), what is the best estimate I can make of the required solution?" Clearly, this is a problem of type (b).

5.2. *Philosophy of approximation.* It is common practice to construct "rules", for example quadrature formulae, for estimating some definite property of a function from a knowledge of other properties. Such a rule is usually expected to give useful output when applied, without special modification, to a wide variety of input functions. Thus, Simpson's rule

$$F_S(h) = (a/3)[f(-a) + 4f(0) + f(a)]$$

is often used to approximate the value $F(f) = \int_{-a}^{a} f(x) \cdot dx$, although it is clear that some $f(\,\cdot\,)$ may be chosen so as to make the error $|R(f)| = |F(f) - F_S(f)|$ arbitrarily large. Of course, rules like this can be tailored to deal accurately with a prescribed, fairly limited class of functions, but a rule designed to be exact for, say, low order polynomials will generally give poor results when applied to functions having singularities near the approximation region, in the complex plane. However, there is practical value in a rule which, like Simpson's rule, is exact for a small class of functions and tolerably good for a much wider class.

Roughly speaking, our approach to problems of type (b) will be as follows:

In any particular problem situation we are given certain specific properties of the solution, e.g. a finite number of ordinate or derivative values at fixed abscissae. If we can assume no more than this basic information we can conclude only that our required solution is a member of that class of functions which possesses the given properties — a tautology which is unlikely to appeal to an experimental scientist! Clearly, we need to be given, or to assume, extra information in order to make more definite statements about the required function.

Typically, we shall assume *general* properties, such as continuity or nonnegativity of the solution and/or its derivatives, and use the given *specific* properties in order to assist in making a selection from the class K of all functions possessing the assumed general properties. We shall choose K either to be a Hilbert space or to be simply related to one.

Golomb and Weinberger (1959) drew attention to the fact that if a best approximating function can be regarded as a member of a Hilbert space, then the specific information referred to above may be interpreted as a set of values of linear or nonlinear, functionals, each of which helps to localise the required function to a more and more limited region of the space. For example, knowledge of the values of a finite number of bounded, linear functionals constrains the required function to lie in a hyperplane, and knowledge of an upper bound on its norm further limits the function to lie in a hyperdisc.

In the present approach, an a priori localisation is achieved effectively by making an assumption about the relative likelihoods of elements of the Hilbert space of possible candidates for the solution to the original problem. Among other things, this permits, at least in principle, the derivation of joint probability density functions for functionals on the space and also allows us to evaluate confidence limits on the estimate of a required functional (in terms of given values of other functionals) without any extra information about the norm of the function in question.

The relationship between the present approach, which might be termed "Functional Estimation", and classical Approximation Theory is illustrated schematically in Figure 5.1. The examples below will help to substantiate and clarify the picture.

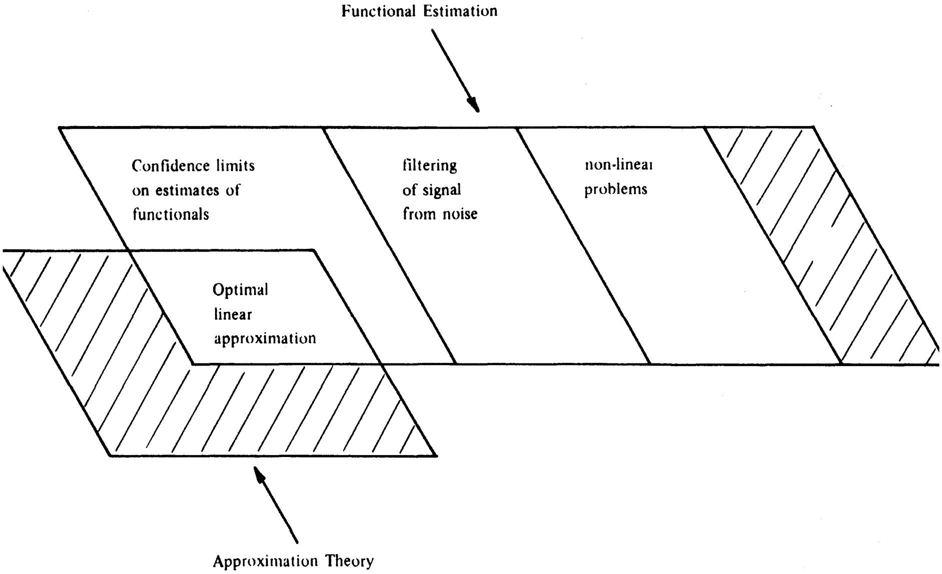
Functional Estimation
Confidence limits on estimates of functionals
filtering of signal from noise
non-lineai problems
Optimal linear approximation
Approximation Theory

Figure 5.1 Relationship between Approximation Theory and Functional Estimation

5.3. *Estimation of linear functionals.* Using the same notation as in §2 we can write the error in a linear approximation rule as

$$Rh = (h, g_0) - \sum_{j=1}^{n} w_j(h, g_j), \quad \forall h \in H.$$

Notice that R is a tame functional on H so it makes sense to write its mean-square value as

$$\begin{aligned} e^2(\underline{w}) &= \int_H |Rh|^2 \cdot \nu\{dh\} \\ &= \int_H \left| \left(h, g_0 - \sum_{j=1}^{n} \bar{w}_j g_j \right) \right|^2 \cdot \nu\{dh\}. \end{aligned}$$

By a previous example (equation (3.7)) we have

$$e^2(\underline{w}) = \|R\|^2/\lambda = \left\| g_0 - \sum_{j=1}^{n} \bar{w}_j g_j \right\|^2 \Big/ \lambda.$$

The mean-square error $e^2(\underline{w})$ is minimised by choosing $\underline{w}$ so as to minimise $\|R\|$, clearly leading to the same result as the usual theory of optimal approximation.

This approach is analogous to the familiar statistical method of Minimum Variance parameter estimation which, for the case of a normal distribution, we would expect to agree with the method of Maximum Likelihood. That this expectation is correct can be seen from another example (relation (3.12)), which gives the joint relative likelihood of the vector

$$\underline{x} = \{x_0, x_1, x_2, \cdots, x_n\} = \{(h, g_0), (h, g_1), (h, g_2), \cdots, (h, g_n)\}$$

as

$$\mathcal{L}(\underline{x}) = \exp\left[-(\lambda/2)\underline{x}' G_{n+1}^{-1} \underline{x} \right],$$

where the $(n + 1)$th order Gram matrix G_{n+1} is given by $[G_{n+1}]_{jk} = (g_j, g_k)$; $j, k = 0, 1, 2, \cdots, n$. It is now only a matter of algebraic manipulation to show that $\mathcal{L}(\underline{x})$ is maximised, subject to given values for $\{x_j; j = 1, 2, \cdots, n\}$ by choosing $x_0 = \sum_{j=1}^{n} w_j x_j$ where the $\{w_j; j = 1, 2, \cdots, n\}$ are the usual optimal weights.

Thus, for the case of the linear approximation problem, we have agreement between classical optimal approximation and functional estimation. The following sections extend the results obtainable by functional estimation beyond the scope of optimal approximation theory.

5.4. *Confidence limits on linear functionals.* We now consider the problem of placing confidence limits upon the optimal estimate of the value of a bounded, linear functional, even in the absence of norm bounds.

Let $\underline{x} = \{x_1, x_2, \cdots, x_n\}$ be a vector of values of bounded, linear functionals on a real Hilbert space, having a joint relative likelihood function

$$\mathcal{L}(\underline{x}) = \lambda^{n/2} \exp\left(-(\lambda/2)\underline{x}'A\underline{x}\right),$$

A being the inverse of the Gram matrix formed from the representers of the bounded, linear functionals in question.

We partition $\underline{x}$ and A as

$$\underline{x} = \begin{array}{cc} \left[\begin{array}{c} \underline{u} \\ \underline{v} \end{array}\right] & \begin{array}{c} \updownarrow m \\ \updownarrow n-m \end{array} \end{array} \quad G^{-1} = A = \begin{array}{cc} \begin{array}{cc} \leftarrow m \rightarrow & \leftarrow n-m \rightarrow \end{array} & \\ \left[\begin{array}{cc} R & S \\ S' & T \end{array}\right] & \begin{array}{c} \updownarrow m \\ \updownarrow n-m \end{array} \end{array} \tag{5.1}$$

$$= \begin{bmatrix} (R - ST^{-1}S')^{-1} & -(R - ST^{-1}S')^{-1}ST^{-1} \\ -(T - S'R^{-1}S)^{-1}S'R^{-1} & (T - S'R^{-1}S)^{-1} \end{bmatrix}^{-1}$$

with the objective of simultaneous estimation of the components of the vector $\underline{v}$ in terms of given values of the components of the vector $\underline{u}$. Let

$$y = \log[\mathcal{L}(\underline{x})] = (n/2)\log\lambda - (\lambda/2)(\underline{u}'R\underline{u} + 2\underline{v}'S'\underline{u} + \underline{v}'T\underline{v}).$$

For a maximum likelihood we must have

$$\frac{\partial y}{\partial \lambda} = \frac{n}{2\lambda} - \frac{1}{2}(\underline{u}'R\underline{u} + 2\underline{v}'S'\underline{u} + \underline{v}'T\underline{v}) = 0,$$

$$\frac{\partial y}{\partial v_j} = -\lambda\left(\sum_{k=1}^{n} S'_{jk}u_k + \sum_{k=1}^{n} T_{jk}v_k\right)$$

$$= 0; \qquad j = 1, 2, \cdots, n - m.$$

These equations are easily solved for the maximum likelihood estimates

$$\hat{\underline{v}} = -T^{-1}S'\underline{u}, \qquad \hat{\lambda} = n/\underline{u}'(R - ST^{-1}S')\underline{u}.$$

Furthermore $\mathcal{L}(\underline{x})$ may be written in the form

$$\begin{aligned} \mathcal{L}(\underline{x}) = \lambda^{n/2}\exp\{-(\lambda/2)[&\underline{u}'(R - ST^{-1}S')\underline{u} \\ &+ (\underline{v}' + \underline{u}'ST^{-1})T(\underline{v} + T^{-1}S'\underline{u})]\} \end{aligned} \tag{5.2}$$

from which it follows that the conditional likelihood of $\underline{v}$, given $\underline{u}$, is

(5.3) $\mathcal{L}(\underline{v} \mid \underline{u}) = \lambda^{(n-m)/2} \cdot \exp\{-(\lambda/2)(\underline{v}' + \underline{u}'ST^{-1})T(\underline{v} + T^{-1}S'\underline{u})\}$

and the marginal likelihood of $\underline{u}$ is simply

(5.4) $\mathcal{L}(\underline{u}) = \lambda^{m/2} \cdot \exp\{-(\lambda/2) \cdot \underline{u}'(R - ST^{-1}S')\underline{u}\}.$

Note that equation (5.3) immediately provides an alternative proof that $-T^{-1}S'\underline{u}$ is a Maximum Likelihood estimator of $\underline{v}$.

Now consider the quadratic forms

$$Q_1 = (\underline{v} + T^{-1}S'\underline{u})'T(\underline{v} + T^{-1}S'\underline{u}), \qquad Q_2 = \underline{u}'(R - ST^{-1}S')\underline{u},$$

which satisfy

$$Q_1 + Q_2 = \underline{x}'A\underline{x}.$$

These expressions may be reformulated as

$$Q_1 = [\underline{u}'\underline{v}'] \begin{bmatrix} ST^{-1} \\ I \end{bmatrix} T [T^{-1}S' \quad I] \begin{bmatrix} \underline{u} \\ \underline{v} \end{bmatrix},$$

$$Q_2 = [\underline{u}'\underline{v}'] \begin{bmatrix} R - ST^{-1}S' & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \underline{u} \\ \underline{v} \end{bmatrix}$$

from which it may be verified that rank$(Q_1) = n - m$ and rank$(Q_2$ $= m$. Thus, by Cochran's theorem (Cochran, 1934), the quantitie Q_1 and Q_2 are distributed as independent χ^2 variates, with $(n - m$ and m degrees of freedom respectively. The statistic

$$q = (m/(n-m))Q_1/Q_2$$

then follows the Fisher $F_{n-m,m}$ distribution law, whose probabilit density function is given by

(5.5) $$f(q) = \frac{\Gamma(n/2)}{\Gamma(m/2)\Gamma((n-m)/2)} \cdot \left(\frac{n-m}{m}\right)^{(n-m)/2} \cdot \frac{q^{(n-m)/2-1}}{\{1 + ((n-m)/m)q\}^{n/2}}; \qquad q > 0.$$

In practice this result may be used as follows: corresponding to an desired confidence level α, we can determine a value q_α from th cumulative table of $f(q)$; thence we can say, with confidence α, tha the required vector $\underline{v}$ lies within the $(n - m)$-dimensional ellipsoi defined by

$$(5.6) \qquad \begin{aligned} &(\underline{v} + T^{-1}S'\underline{u})'T(\underline{v} + T^{-1}S'\underline{u}) \\ &\qquad \leqq ((n-m)/m)q_\alpha \cdot \underline{u}'(R - ST^{-1}S')\underline{u}. \end{aligned}$$

Note. In the complex case Q_1 and Q_2 have $2(n - m)$ and $2m$ real degrees of freedom respectively, and relations (5.5) and (5.6) must be modified accordingly.

Example. Take H to be the Szegö-Hilbert space H of functions analytic within the unit circle C and continuous on C, with inner product defined by

$$(f, g) = \int_C f(z) \cdot \overline{g(z)} \cdot |dz|; \quad \forall f, g \in H.$$

This space possesses the reproducing kernel $K(z, \bar{t}) = 1/2\pi(1 - z\bar{t})$.

Suppose we are given the values

z	$f(z)$
0.0	$1 + 2i$
0.8	$1 + 2i$
$-0.4 - 0.5i$	$1 + 2i$

The above technique permits us to estimate $\hat{f}(0.5i) = 1 + 2i$ with 95% confidence that $|f(0.5i) - \hat{f}(0.5i)| \leqq 1.26$ and 99% confidence that $|f(0.5i) - \hat{f}(0.5i)| \leqq 1.84$. The inclusion of further tabular points with $f = 1 + 2i$ leaves $\hat{f}$ unchanged but narrows the confidence intervals; e.g. given, in addition, the values

z	$f(z)$
$-0.2 + 0.7i$	$1 + 2i$
$0.2 - 0.7i$	$1 + 2i$

we have 95% confidence that $|f(0.5i) - \hat{f}(0.5i)| \leqq 0.34$ and 99% confidence that $|f(0.5i) - \hat{f}(0.5i)| \leqq 0.46$.

The above technique provides a rational basis for dealing with the estimation problem posed in the Introduction (and, of course, with more complicated linear problems).

5.5. *Estimation of bounded linear functionals from noisy data.* Let the linearly independent elements $\{g_j;\ j = 1, 2, \cdots, n\}$, in a real Hilbert space H, be the representers of the bounded linear functionals $\{L_j; j = 1, 2, \cdots, n\}$. We shall suppose that approximate values of the quantities $\underline{s} = \{L_j h;\ j = 1, 2, \cdots, m;\ m < n\}$ are given, and that we wish to estimate the values of the remaining quantities $\underline{t} = \{L_j h; j = m + 1, m + 2, \cdots, n\}$.

The joint distribution of the nth order vector $\underline{x} = \{\underline{s}, \underline{t}\}$ is given by (3.12), and we presume that the given, approximate values $\underline{q}$ result from additive contamination of $\underline{s}$ by a random "noise" vector $\underline{r}$, whose relative likelihood function is assumed to be of the form

$$\mathcal{L}(\underline{r}) = \mu^{m/2} \cdot \exp(-\ (\mu/2) \cdot \underline{r}'B\underline{r}). \tag{5.7}$$

Here B is assumed to be a known matrix, and μ may or may not be known.

This problem is discussed elsewhere (Larkin, 1971) and the following maximum likelihood estimates are found for the unknown quantities:

Partitioning the matrix G as

$$G = \begin{array}{cc} \begin{array}{cc} \leftarrow m \rightarrow & \leftarrow n - m \rightarrow \end{array} & \\ \begin{bmatrix} U & & V \\ & & \\ V' & & W \end{bmatrix} & \begin{array}{c} \uparrow \\ m \\ \downarrow \\ \uparrow \\ n - m \\ \downarrow \end{array} \end{array}$$

we have

$$\left\{ \begin{array}{l} \hat{\underline{s}} = (1 + (\lambda/\mu) \cdot B^{-1}U^{-1})^{-1}\underline{q}, \\ \hat{\underline{r}} = (\lambda/\mu) \cdot B^{-1}U^{-1}\hat{\underline{s}}, \text{ and} \\ \hat{\underline{t}} = \underline{V}'U^{-1}\hat{\underline{s}}, \end{array} \right. \tag{5.8}$$

where, if μ is assumed to be known, $\hat{\lambda}$ satisfies the equation

$$\underline{q}'(U + (\hat{\lambda}/\mu) \cdot B^{-1})^{-1}U(U + (\hat{\lambda}/\mu) \cdot B^{-1})^{-1}\underline{q} = m/(2\hat{\lambda}) \tag{5.9}$$

and, if λ and μ are both unknown, the ratio $\hat{\nu} = \lambda/\mu$ satisfies the equation

$$\frac{\underline{q}'(U + \hat{\nu}B^{-1})^{-1}U(U + \hat{\nu}B^{-1})^{-1}\underline{q}}{\underline{q}'(U + \hat{\nu}B^{-1})^{-1}B^{-1}(U + \hat{\nu}B^{-1})^{-1}\underline{q}} = \hat{\nu}. \tag{5.10}$$

Modifications appropriate to the case of a complex Hilbert space and a singular matrix G^{-1} are also discussed. Notice that equations (5.8) indicate that $\underline{t}$ should be found by optimal approximation from a "smoothed" signal vector $\hat{\underline{s}}$, where λ/μ plays the role of a "smoothing parameter".

5.6. *Simple nonlinear problems.* The most striking feature of functional estimation, as compared with optimal approximation, is its capacity to deal effectively with certain nonlinear problems. Progress in this direction is limited by the difficulty of evaluating function space integrals in closed form, but at least some simple, nonlinear estimation problems turn out to be analytically tractable.

Suppose we are given values of the functionals $\{F_j(h); j = 1, 2, \cdots, n\}$ and wish to estimate the value of $F_0(h)$. In principle, we might find the multivariate distribution of the vector $\underline{F} = \{F_j(h); j = 0, 1, 2, \cdots, n\}$ by inverse Fourier transformation of the $(n + 1)$-dimensional characteristic function given by Theorem 3.3. Substitution of the n known values then leads to the conditional distribution of F_0, given $\{F_j; j = 1, 2, \cdots, n\}$. This distribution comprises all the information we can legitimately infer about F_0, on the basis of the foregoing theory. Unfortunately, this program is impracticable except in certain special cases, so we are led to consider alternative techniques for estimation of F_0.

For example, if the functionals $F_j(L)$; $j = 0, 1, 2, \cdots, n$, are all homogeneous and of the same degree in h, we might consider a linear minimum variance estimator derived as follows:

Take

$$S = \int_B \left| F_0(h) - \sum_{j=1}^{n} w_j F_j(h) \right|^2 \cdot \mu\{dh\}. \tag{5.11}$$

Assuming all the required functional integrals exist, minimisation of S with respect to the weights $\underline{w} = \{w_j; j = 1, 2, \cdots, n\}$ leads to the linear equations

$$C\underline{w} = \underline{d}, \tag{5.12}$$

where

$$C_{jk} = \int_B \overline{F_j(h)} \cdot F_k(h) \cdot \mu\{dh\}$$

and

$$d_j = \int_B F_0(h) \cdot \overline{F_j(h)} \cdot \mu\{dh\}.$$

Thus, if $\underline{\alpha}$ denotes a vector of given values of the $\{F_j(h); j = 1, 2, \cdots, n\}$, we have $\hat{F}_0 = \underline{d}'C^{-1}\underline{\alpha}$, noting that C is Hermitian. The minimal value of S is given by

$$S_{\min} = \int_B |F_0(h)|^2 \cdot \mu\{dh\} - \underline{d}'C^{-1}\underline{d}.$$

Note that the dispersion parameter λ disappears from equations (5.12) under the assumed homogeneity conditions; otherwise λ will appear explicitly in these equations.

Example. Consider the Paley-Wiener-Hilbert space H of real, band-limited functions, square-integrable on the real line, with inner product given by

$$(f, g) = \int_{-\infty}^{\infty} f(t)g(t) \cdot dt, \quad \forall f, g \in H$$

(e.g. de Branges, 1968). This space possesses the reproducing kernel function

$$K(x, y) = \frac{\operatorname{Sin}[a(x-y)]}{\pi(x-y)}, \qquad a \text{ real},$$

where $[-a, a]$ is the support of the Fourier transform of any member of H.

Suppose we are given the values $\underline{\alpha} = \{\alpha_j = h^2(x_j); j = 1, 2, \cdots, n\}$ and wish to estimate the value of $\int_{-b}^{b} h^2(x) \cdot dx$, for some real, fixed b Note that $h(x)h(y)$ is a tame function for fixed x, y and, for a real Hilbert space,

$$\int_H [h(x)h(y)]^2 \cdot \nu\{dh\} = \frac{K(x, x)K(y, y) + 2K^2(x, y)}{\lambda^2}$$
$$= \frac{a^2}{\lambda^2\pi^2}\left\{1 + \frac{2\operatorname{Sin}^2[a(x-y)]}{a^2(x-y)^2}\right\}.$$

Hence, discarding the factor $a^2/\lambda^2\pi^2$ and appealing to the Fubin theorem, we can define

$$C_{jk} = 1 + \frac{2\operatorname{Sin}^2[a(x_j - x_k)]}{a^2(x_j - x_k)^2}; \qquad j, k = 1, 2, \cdots, n,$$

$$d_j = \int_{-b}^{b}\left\{1 + \frac{2\operatorname{Sin}^2[a(x_j - x)]}{a^2(x_j - x)^2}\right\}dx; \quad j = 1, 2, \cdots, n,$$

and use

$$\hat{F}_0(h) = \underline{d}'C^{-1}\underline{\alpha} = \underline{w}'\alpha, \quad \text{say}, \tag{5.13}$$

as an estimator of $\int_{-b}^{b} h^2(x) \cdot dx$.

For illustrative purposes let us choose $n = 3$, $a = 1$, $b = \pi$, $x_1 = -\pi$, $x_2 = 0$, $x_3 = \pi$. This leads to weights

$$w_1 = 1.58467316633354 = w_3, \qquad w_2 = 2.92881509301435.$$

This quadrature rule does not treat constants exactly, since $\sum_{j=1}^{n} u$ $\neq 1$, but this is hardly surprising, since the unit function is not member of this Hilbert space. However, numerical experiments suggest that if

$$x_j = -b + 2b(j-1)/(2n-1); \quad j = 1, 2, \cdots, n, \tag{5.14}$$

then

$$\lim_{n\to\infty} \frac{1}{2b} \sum_{j=1}^{n} w_j = 1$$

and

$$\lim_{b\to 0} \frac{1}{2b} \sum_{j=1}^{n} w_j = 1$$

which indicates that limiting versions of quadrature rule (5.13) may well treat constants exactly.

An alternative approach might be to determine an $\hat{h} \in H$ which maximises

$$\mathcal{L}(\hat{h}) = \exp(-(\lambda/2)\|\hat{h}\|^2)$$

subject to the constraints

$$F_j(\hat{h}) = \alpha_j; \qquad j = 1, 2, \cdots, n.$$

We can then use $F_0(\hat{h})$ as an estimator of F_0. Often this approach will lead to a generalised eigenvalue problem (Larkin, 1969), but for illustrative purposes we consider here a problem similar to that just discussed.

Suppose we are given values of the bounded, linear functionals: $F_j(h) = (h, g_j) = \alpha_j; j = 1, 2, \cdots, n$, and wish to estimate the value of the quadratic functional $F_0(h) = (h, Ah)$, where A is a finite-trace class operator on H. From a previous result (equation (2.4)) we know that $\hat{h} = \underline{\alpha}'G^{-1}\underline{g}$ where $\underline{g} = \{g_1, g_2, \cdots, g_n\}$ and the Gram matrix G is defined by $G_{jk} = (g_j, g_k), \ j, k = 1, 2, \cdots, n$. Hence

$$F_0(\hat{h}) = (\hat{h}, A\hat{h}) = \underline{\alpha}'G^{-1}MG^{-1}\underline{\alpha} \tag{5.15}$$

where the general element of the nth order matrix M is $M_{jk} = (g_j, Ag_k)$; $j, k = 1, 2, \cdots, n$.

We now choose H to be the Hilbert space used in the previous example,

$$F_j(h) = h(x_j); \qquad j = 1, 2, \cdots, n,$$

$$F_0(h) = \int_{-b}^{b} h^2(x) \cdot dx.$$

Thus we have

$$\begin{aligned} F_0(h) &= \int_{-b}^{b} (h(y), K(y, x)) \cdot (h(z), K(z, x)) \cdot dx \\ &= \int_{-b}^{b} (h(y), (h(z), K(y, x)K(z, x))) \cdot dx; \end{aligned}$$

i.e. $F_0(h) = (h, Ah)$ where

$$\begin{aligned} Ah(y) &= \left(h(z), \int_{-b}^{b} K(y, x)K(z, x) \cdot dx \right) \\ &= \int_{-b}^{b} K(y, x)h(x) \cdot dx \in H. \end{aligned}$$

The matrix M is thus given by

$$\begin{aligned} M_{jk} = (g_j, Ag_k) &= \int_{-\infty}^{\infty} g_j(y) \int_{-b}^{b} K(y, x)g_k(x) \cdot dx\, dy \\ &= \int_{-b}^{b} g_j(x)g_k(x) \cdot dx, \end{aligned}$$

i.e. for $j, k = 1, 2, \cdots, n,$

$$M_{jk} = \frac{1}{\pi^2} \int_{-b}^{b} \frac{\mathrm{Sin}[a(x - x_j)] \cdot \mathrm{Sin}[a(x - x_k)]}{(x - x_j)(x - x_k)} \cdot dx, \tag{5.16}$$

and G has the form

$$G_{jk} = \frac{\mathrm{Sin}[a(x_j - x_k)]}{\pi(x_j - x_k)} . \tag{5.17}$$

For fixed $\{x_j; j = 1, 2, \cdots, n\}$ we can use (5.16) and (5.17) to find numerically the matrix $Q = G^{-1}MG^{-1}$ which characterises the quadrature rule (5.15). However, for the equispaced abscissae given by (5.14) it turns out that the elements of Q vary in sign and their absolute magnitudes increase rapidly with n. This feature makes (5.15) unsuitable for practical use, since small changes in the $\{\alpha_j\}$ can result in large changes in $F_0(\hat{h})$.

We thus have three superficially attractive approaches to the problem of approximate quadrature of a nonnegative function whose ordinate values $\{f(x_j); j = 1, 2, \cdots, n\}$ are given at prescribed abscissae, mathematically equivalent to the following procedures:

(a) Find the optimal interpolant to $\{f(x_j); j = 1, 2, \cdots, n\}$ and integrate it over the required range.

(b) Find the optimal interpolant to $\{f^{1/2}(x_j);\ j = 1,\ 2,\ \cdots,\ n\}$ and integrate its square over the required range.

(c) Use the quadrature weights derived by means of the minimum variance technique described earlier.

Although approaches (a) and (b) both avoid the necessity of functional integration, they will both be unreliable in practice—the former because the optimal interpolant may go negative even though the ordinate values may all be positive, and the latter because of intolerable numerical errors arising from the use of inexact arithmetic. Approach (c) suffers from neither of these defects, remaining a serious contender for practical use.

5.7. *The case of unbounded functionals.* We have examined the question of what can be inferred about the value of a bounded, linear functional from a knowledge of the value of another bounded, linear functional, but it is instructive to consider what happens when the norm of one of these tends to infinity.

If g_1 and g_2 are the representers of two bounded, linear functionals on some Hilbert space H, whose numerical values are x_1 and x_2 respectively, the joint distribution of $\underline{x} = \{x_1, x_2\}$ is given by relation (3.12), where

$$G = \begin{bmatrix} \|g_1\|^2 & (g_2, g_1) \\ (g_1, g_2) & \|g_2\|^2 \end{bmatrix},$$

so that

$$G^{-1} = [\|g_1\|^2 \cdot \|g_2\|^2 - |(g_1, g_2)|^2]^{-1} \begin{bmatrix} \|g_2\|^2 & -(g_2, g_1) \\ -(g_1, g_2) & \|g_1\|^2 \end{bmatrix}.$$

The a priori distribution of x_2 is $N(0, \|g_2\|/\sqrt{\lambda})$ while, from equation (5.3), its conditional distribution given x_1 is

$$N\left(x_1 \frac{(g_1, g_2)}{\|g_1\|^2},\ \left[\frac{\|g_2\|^2 - |(g_1, g_2)|^2 \cdot \|g_1\|^{-2}}{\lambda}\right]^{1/2} \right).$$

Thus, as $\|g_1\| \to \infty$, the conditional mean of x_2, given x_1, approaches zero, indicating that a knowledge of the value of an unbounded, linear functional contributes nothing to a knowledge of the value of a bounded, linear functional, except possibly to reduce the variance of its distribution about the a priori mean.

EXAMPLE. Let $H = L_2(0, 1)$ so that

$$(f, g) = \int_0^1 f(t) \cdot g(t) \cdot dt, \quad \forall f, g \in H.$$

Let

$$L_n h = \int_0^1 h(t) \cdot dt = x_n,$$

$$L_j h = \frac{1}{2\epsilon} \int_{t_j - \epsilon}^{t_j + \epsilon} h(t) \cdot dt, \qquad \epsilon \leqq t_j \leqq 1 - \epsilon; j = 1, 2, \cdots, n-1.$$

Thus

$$g_j(t) = \begin{cases} 0; & t < t_j - \epsilon, \\ 1/2\epsilon; & t_j - \epsilon \leqq t \leqq t_j + \epsilon, \qquad j = 1, 2, \cdots, n-1; \\ 0; & t_j + \epsilon < t \leqq 1, \end{cases}$$

$$g_n(t) = 1; \qquad 0 \leqq t \leqq 1.$$

The Gram matrix G is then of the form

$$G = \begin{bmatrix} 1/2\epsilon & & & & 1 \\ & & & \mathbf{0} & 1 \\ & 1/2\epsilon & & & \\ & & & \cdot & \vdots \\ & \mathbf{0} & & \cdot & \\ & & & 1/2\epsilon & 1 \\ 1 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

assuming that the supports of $\{g_j; j = 1, 2, \cdots, n-1\}$ do not overlap, and it may be verified that the optimal approximation to x_n is $\hat{x}_n = 2\epsilon \sum_{j=1}^{n-1} x_j$.

Thus, as $\epsilon \to 0$ and the norms of $\{L_j;\ j = 1, 2, \cdots, n-1\}$ tend tc infinity, a knowledge of the values of $\{L_j h; j = 1, 2, \cdots, n-1\}$ which for numerical purposes is a knowledge of the ordinate values ai the abscissae $\{x_j; j = 1, 2, \cdots, n-1\}$, does not permit us to modify the a priori distribution of $L_n h$. In this case the $\{g_j; j = 1, 2, \cdots$ $n-1\}$ approach orthogonality with g_n as $\epsilon \to 0$, so we cannot ever reduce the a priori variance of $L_n h$ by measuring ordinate values!

Alternatively, this example illustrates the need for restricting consideration to a reproducing kernel Hilbert space when basing estimates upon ordinate values.

References

1. J. H. Ahlberg and E. N. Nilson, *Convergence properties of the spline fit* J. Soc. Indust. Appl. Math. 11 (1963), 95–104. MR 27 #2763.

2. J. H. Ahlberg, E. N. Nilson and J. L. Walsh, *Best approximation and convergence properties of higher-order spline approximations*, J. Math. Mech. 14 (1965), 231-244. MR 35 #5823.

3. ——, *The theory of splines and their applications*, Academic Press, New York, 1967. MR 39 #684.

4. N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc. 68 (1950), 337-404. MR 14, 479.

5. G. Birkhoff and C. W. de Boor, *Error bounds for spline interpolation*, J. Math. Mech. 13 (1964), 827-835. MR 29 #2583.

6. C. W. de Boor, *Best approximation properties of spline functions of odd degree*, J. Math. Mech. 12 (1963), 747-749. MR 27 #3982.

7. L. de Branges, *Hilbert spaces of entire functions*, Prentice-Hall, Englewood Cliffs, N. J., 1968. MR 37 #4590.

8. W. G. Cochran, *The distribution of quadratic forms in a normal system, with applications to the analysis of variance*, Proc. Cambridge Philos. Soc. 30 (1934), 178-191.

9. P. J. Davis, *Interpolation and approximation*, Blaisdell, New York and London, 1963. MR 28 #393.

10. S. F. Edwards, *The statistical mechanics of polymerised material*, Proc. Phys. Soc. 92 (1967).

11. I. M. Gel'fand and A. M. Jaglom, *Integration in functional spaces and its applications in quantum physics*, J. Mathematical Phys. 1 (1960), 48-69. (English transl.) MR 22 #3455.

12. M. Golomb and H. F. Weinberger, *Optimal approximation and error bounds*, Proc. Sympos. on Numerical Approximation (Madison, Wis., 1958), Math. Res. Center, U. S. Army, Univ. of Wisconsin Press, Madison, Wis., 1959, pp. 117-190. MR 22 #12697.

13. L. Gross, *Integration and nonlinear transformations in Hilbert space*, Trans. Amer. Math. Soc. 94 (1960), 404-440. MR 22 #2883.

14. ——, *Measurable functions on Hilbert space*, Trans. Amer. Math. Soc. 105 (1962), 372-390. MR 26 #5121.

15. ——, *Harmonic analysis on Hilbert space*, Mem. Amer. Math. Soc. No. 46 (1963). MR 28 #4304.

16. ——, *Abstract Wiener spaces*, Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), vol. II: Contributions to Probability Theory, part 1, Univ. California Press, Berkeley, Calif., 1967, pp. 31-42. MR 35 #3027.

17. D. C. Handscomb (Editor), *Methods of numerical approximation*, Pergamon Press, New York, 1966.

18. J. C. Holladay, *A smoothest curve approximation*, Math. Tables Aids Comput. 11 (1957), 233-243. MR 20 #414.

19. G. S. Kimeldorf and Grace Wahba, *A correspondence between Bayesian estimation on stochastic processes and smoothing by splines*, MRC Technical Summary Report #967, Oct. 1968. Cf: Ann. Math. Statist. 41 (1970), 495-502. MR 40 #8206.

20. ——, *Spline functions and stochastic processes*, MRC Technical Summary Report #969, Aug. 1969.

21. I. M. Koval'chik, *The Wiener integral*, Uspehi Mat. Nauk 18 (1963), no. 1 (109), 97-134 = Russian Math. Surveys 18 (1963), no. 1, 97-134. MR 36 #5295.

22. V. I. Krylov, *Approximate calculation of integrals,* Fizmatgiz, Moscow, 1959; English transl., Macmillan, New York, 1962. MR **22** #2002; MR **26** #2008.

23. J. Kuelbs, F. M. Larkin and J. Williamson, *Weak probability distributions on reproducing kernel Hilbert spaces,* Rocky Mt. J. Math. **2** (1972), 369–378.

24. F. M. Larkin, *Estimation of a non-negative function,* Nordisk Tidskr. Informationsbehandling **9** (1969), 30–52.

25. ———, *Optimal approximation in Hilbert spaces with reproducing kernel functions,* Math. Comp. **24** (1970), 911–921.

26. ———, *Optimal estimation of bounded linear functionals from noisy data,* Proc. I.F.I.P. Congress, Ljubljana, 1971.

27. E. Mehlum, *A curve-fitting method based on a variational criterion,* Nordisk Tidskr. Informationsbehandling **4** (1964), 213–223. MR **30** #4376.

28. E. Parzen, *Time series analysis papers,* Holden-Day, San Francisco, Calif., 1967. MR **36** #6091.

29. ———, *Statistical inference on time series by RKHS methods,* Department of Statistics, Technical Report #14, Stanford University, Stanford, Calif., 1970.

30. P. Rabinowitz and N. Richter, *Chebyshev-type integration rules of minimum norm,* Math. Comp. **24** (1970), 831–846.

31. A. Sard, *Integral representations of remainders,* Duke Math. J. **15** (1948), 333–345. MR **10**, 197.

32. ———, *Best approximate integration formulas; best approximation formulas,* Amer. J. Math. **71** (1949), 80–91. MR **10**, 576.

33. ———, *Smoothest approximation formulas,* Ann. Math. Statist. **20** (1949), 612–615. MR **12**, 84.

34. ———, *Linear approximation,* Math. Surveys, no. 9, Amer. Math. Soc., Providence, R. I., 1963. MR **28** #1429.

35. M. H. Schultz and R. S. Varga, *L-splines,* Numer. Math. **10** (1967), 345–369. MR **37** #665.

36. I. J. Schoenberg, *Spline interpolation and best quadrature formulae,* Bull. Amer. Math. Soc. **70** (1964), 143–148. MR **28** #394.

37. ———, *On best approximations of linear operators,* Nederl. Akad. Wetensch. Proc. Ser. A **67** = Indag. Math. **26** (1964), 155–163. MR **28** #4284.

38. ———, *On trigonometric spline interpolation,* J. Math. Mech. **13** (1964), 795–825. MR **29** #2589.

39. ———, *Spline functions and the problem of graduation,* Proc. Nat. Acad. Sci. U.S.A. **52** (1964), 947–950. MR **29** #5040.

40. ———, *On interpolation by spline functions and its minimal properties,* Proc. Conference on Approximation Theory (Oberwolfach, Germany, 1963), Birkhäuser, Basel, 1964, pp. 109–129. MR **31** #5015.

41. ———, *On monosplines of least deviation and best quadrature formulae.* I, II, SIAM J. Numer. Anal. **2** (1965), 144–170; ibid. **3** (1966), 321–328. MR **34** #2182; #3170.

42. M. D. Stern, *Optimal quadrature formulae,* Comput. J. **9** (1967), 396–403. MR **35** #3885.

43. A. V. Sul'din, *Wiener measure and its applications to approximation methods.* I, II, Izv. Vysš. Učebn. Zaved. Matematika **1959**, no. 6 (13), 145–158; ibid. **1960**, no. 5 (18), 165–179. MR **28** #722; MR **29** #1482.

44. N. Wiener, *Differential space,* J. Math. Phys. **2** (1923), 131–174.

45. ———, *The average value of a functional,* Proc. London Math. Soc. **22** (1924), 454–467.

46. ———, Proc. London Math. Soc. **55** (1930), 117.

COMPUTING CENTRE, QUEEN'S UNIVERSITY, KINGSTON, ONTARIO, CANADA