Orthogonally Invariant Measures and Best Approximation of Linear Operators

NIELS JUUL MUNCH

Matematisk Institut, Aarhus Universitet,
Ny Munkegade, DK-8000 Aarhus C, Denmark

Communicated by T. J. Rivlin
Received July 27, 1987

This paper studies optimal information and optimal algorithms in Hilbert space for an n-dimensional average case model. The error in approximating a linear operator is the average of some error criterion $E$ with respect to an orthogonally invariant measure. The orthogonally invariant measures are characterized and the problem of best approximation is solved for a wide range of error criteria $E$. In addition, it is shown that adaptation does not help.

1. INTRODUCTION

This paper is concerned with the general problem of estimating the action of a linear bounded operator $A$ on a real, separable Hilbert space $\mathcal{H}$ when only finite information is available. Here information is provided by a map $N$ from $\mathcal{H}$ into the space $\mathbb{R}^n$ of fixed finite dimension $n$. Knowing $Nf, f \in \mathcal{H}$, one seeks the best recovery of $Af$ by means of an algorithm $\varphi$, that is, a map $\varphi: \mathbb{R}^n \rightarrow \mathcal{H}$. In other words, the difference $A - \varphi N$ should be as small as possible in a specified sense.

For a worst case error criterion, this setup has been examined in [4, 5, 8–10] and others. Here we relate to an average error criterion as in [6–8, 12–16]. Assuming $\mu$ to be a Borel probability measure on $\mathcal{H}$ with mean zero and finite second moment $\int_{\mathcal{H}} \|f\|^2 \, d\mu(f)$, the error to be minimized is

$$e(\varphi, N) = \int_{\mathcal{H}} E(Af - \varphi Nf) \, d\mu(f)$$

for some function $E: \mathcal{H} \rightarrow [0, \infty[$. The classical choice for $E$ is $E(f) = \|f\|^2$, which constitutes the average squared error. But also the

* Research supported by the Danish Natural Science Research Council.
probabilistic or hit-and-miss criterion conforms to this framework if we choose $E(f) = 1_{[f, \infty]}(\|f\|)$. It is the striking result of [14] (see also [6, 7, 13, 15, 16]) that for the average squared error and a certain class of “orthogonally invariant” measures $\mu$ adaptive linear information is not superior to non-adaptive linear information. For such $\mu$, possessing a high degree of spatial symmetry, the minimal error $e(\varphi, N)$ can be obtained even within the class of non-adaptive linear information operators $N$ and corresponding linear spline algorithms $\varphi$.

Given the implications of this result it becomes of interest to determine its precise range of validity, i.e., to determine which measures $\mu$ are orthogonally invariant. Examples from [14] are Gaussian measures and, in finite dimensions, measures $\mu$ absolutely continuous with respect to Lebesgue measure $m$ so that the Radon-Nikodym derivative $d\mu/dm$ is rotation invariant in a suitably perturbed inner product. It turns out that these examples are quite characteristic. Denote by $C$ the covariance operator of $\mu$ and assume that the range of $C$ is infinite dimensional. In that case $\mu$ is orthogonally invariant if and only if it has a representation

$$\mu = \int_0^\infty \mu_C \, dv(t),$$

where $\mu_K$ is the Gaussian measure on $\mathcal{H}$ with mean zero and covariance operator $K$ and $v$ is a Borel probability measure on $\mathbb{R}_+$. Further the representation (2) is uniquely determined under the condition $1 = \int_0^\infty t \, dv(t)$ in which case we denote the measure $\mu$ by $\mu_C^r$. One may note from (2) that the projection of $\mu_C^r$ onto any finite dimensional subspace of $\mathcal{H}$ invariably is absolutely continuous w.r.t. Lebesgue measure.

The error criterion (1) has a simple probabilistic interpretation. When $X$ is a second order random variable taking values in $\mathcal{H}$ and $\mu$ is the induced distribution on $\mathcal{H}$ the error $e(\varphi, N)$ is the expected value of $E((A - \varphi N)X)$. Of course Gaussian measures $\mu$ arise from Gaussian variables $X$. But suppose that $X$ is in fact an instance, say $X_1$, of a stochastic process $\{X_t\}_{t \geq 0}$, where each $X_t$ has the distribution $\mu_C$. If the observation time $T$ is subject to noise it may differ from its nominal value $T = 1$, transforming the variable $X = X_1$ into $X = X_T$, where $X_T(\omega) = (X_T(\omega))(\omega)$ for each outcome $\omega$. Under the hypothesis of statistical independence of $\{X_t\}_{t \geq 0}$ and $T$, the distribution of $X = X_T$ is $\mu_C^r$, where $\nu$ is the distribution of $T$. Thus even starting in a purely Gaussian setting it is possible to arrive naturally at measures of the type $\mu_C^r$.

In Section 2 of this paper we prove the structure theorem (2) and derive some ancillary properties of orthogonally invariant measures. As I have recently become aware it appears from a remark in [16, p. 363] that the
connection between orthogonally invariant measures and the "elliptically contoured" measures [3] of the form (2) has been noted previously by Kwapien in private communication. However, since the present approach is quite different from the one in [3] and the results are sharper, I feel it is justified to present this material.

Next, in Section 3, we study the approximation problem for orthogonally invariant measures. For a quite general class of error criteria \( E \) it is proved that adaptive information \( N \) is not more powerful than non-adaptive information and that for a given non-adaptive \( N \) the natural spline algorithm is optimal. In particular there is a linear optimal algorithm. The set of \( E \)'s considered includes the average squared error and the probabilistic error. Consequently our approach unifies and improves previous results for orthogonally invariant measures and the squared error [12, 14, 15] and Gaussian measures and more general criteria [13, 16, (6), (7)]. In addition a number of new results and uniqueness results are obtained.

2. ORTHOGONALLY INVARIANT MEASURES

Let \( \mathcal{H} \) be a real Hilbert space of finite or countable dimension. Consider on \( \mathcal{H} \) a Borel probability measure \( \mu \) with mean zero, finite second moment \( \int_{\mathcal{H}} \|f\|^2 d\mu(f) \), and covariance operator \( C_\mu \) defined by

\[
C_\mu = \int_{\mathcal{H}} (f \otimes f) d\mu(f).
\]

Here the Hilbert-Schmidt operators on \( \mathcal{H} \) are identified with the tensor product \( \mathcal{H} \otimes \mathcal{H} \) through \( (f_1 \otimes f_2) g = (g, f_1) f_2 \). It is assumed that \( C_\mu \) is injective and that \( \mu \) is symmetric, i.e., \( \int_{\mathcal{H}} F(f) d\mu(f) = \int_{\mathcal{H}} F(-f) d\mu(f) \).

Following Wasilkowski and Wozniakowski [14] we define the symmetric measure \( \mu \) to be orthogonally invariant if \( \mu = \mu \circ Q_f^{-1} \) for all \( f \in \mathcal{H} \) normalized so that \( (C_\mu f, f) = 1 \). Here \( Q_f \) is the operator \( Q_f = 2(f \otimes C_\mu f) - I \) which satisfies \( Q_f^2 = I \) provided \( (C_\mu f, f) = 1 \).

Recall that the Fourier transform or characteristic functional \( \hat{\mu} \) of \( \mu \) is the function from \( \mathcal{H} \) into \( \mathbb{C} \) defined by

\[
\hat{\mu}(f) = \int_{\mathcal{H}} \exp(i(g, f)) d\mu(g), \quad f \in \mathcal{H},
\]

and that \( \hat{\mu} \) determines \( \mu \) uniquely [11, pp. 11]. For any non-zero vector \( g \) in \( \mathcal{H} \) denote by \( l_g \) the functional \( l_g(f) = (C_\mu g, g)^{-1/2} (f, g), f \in \mathcal{H} \).

**Proposition 2.1.** For a symmetric measure \( \mu \) the following are equivalent.
(a) The measure \( \mu \) is orthogonally invariant.

(b) All measures \( \mu \circ l_{\frac{1}{n}} \), \( g \in \mathcal{H} \setminus \{0\} \), are equal.

(c) There is a function \( g: \mathbb{R} \to \mathbb{R} \) so that

\[
\hat{\mu}(f) = g((C_\mu f, f)^{1/2}), \quad f \in \mathcal{H}.
\]

(d) There is a twice continuously differentiable, positive definite function \( g_\mu: \mathbb{R} \to \mathbb{R} \) such that

\[
\mu(f) = g_\mu((C_\mu f, f)^{1/2}), \quad f \in \mathcal{H}
\]

and

\[
g_\mu(0) = 1, \quad g_\mu'(0) = 0, \quad g_\mu''(0) = -1.
\]

Further if \( \mu \) is orthogonally invariant and if for some function \( g: \mathbb{R} \to \mathbb{R} \) and self-adjoint operator \( C \) it holds that

\[
\mu(f) = g((Cf, f)^{1/2}), \quad f \in \mathcal{H},
\]

then there is a constant \( \gamma > 0 \) such that

\[
\mu = \gamma^2 C \text{ and } g(s) = g_\mu(\gamma s), \quad s \geq 0.
\]

Proof: (a) \( \Rightarrow \) (c). Assume \( \mu = \int \mathbb{R}^2 \mu_1 \mu_2 \)\( \mu \circ f = \mu_1 \). Let \( g \) be the vector

\[
g = (g_1, g_2), \quad f = (f_1 + f_2, f_1 + f_2, f_1 + f_2).
\]

As \( (C_\mu(f_1 + f_2, f_1 + f_2)) = 2((C_\mu f_1, f_1) + (C_\mu f_2, f_2)) \) one may verify that

\[
\mu(f_1) = \mu(Q_g^{-1}(f_1)) = \mu(Q_g f_1) = \mu(f_2).
\]

(c) \( \Rightarrow \) (a). It is straightforward to verify the relation

\[
\mu(Q_f C_\mu Q_f^* = C_\mu).
\]

Thus

\[
\hat{\mu}(Q_f g) = g((C_\mu Q_f g, Q_f g)^{1/2}) = g((C_\mu g, g)^{1/2}) = \hat{\mu}(g), \quad g \in \mathcal{H}.
\]

(c) \( \Rightarrow \) (b). This equivalence is seen from

\[
\mu \circ l_{R}^{-1}(s) = \mu(s(C_\mu g, g)^{1/2}), \quad s \in \mathbb{R}.
\]

(c) \( \Rightarrow \) (d). Denote the common value of \( \mu \circ l_{g}^{-1} \) by \( \tilde{\mu} \). It is apparent from (3) that \( \tilde{\mu}(s) = g(s), s \geq 0 \). Now (d) is simply the statement that \( g_\mu = \tilde{\mu} \) is the transform of a probability measure with mean zero and second moment one.

To prove the final statement of the proposition assume that

\[
\hat{\mu}(f) = g_{\mu}((C_\mu f, f)^{1/2}) = g((Cf, f)^{1/2}), \quad f \in \mathcal{H}.
\]

Consider any non-zero vector \( g \) in \( \mathcal{H} \) and put \( \alpha^2 = (C_\mu g, g), \beta^2 = (Cg, g). \) Then

\[
\tilde{\mu}(sg) = g_{\mu}(|s| \alpha) = g(|s| \beta).
\]

If \( \beta = 0 \) then \( \tilde{\mu}(R \cdot g) = \{ 1 \} \) and \( \mu \) is con-
centrated on the orthogonal complement of $g$, contradicting the standing assumption that $C_\mu$ is injective. Thus $g(s) = g_\mu(\gamma y), s \geq 0,$ holds with $\gamma = \alpha/\beta$. Since the identity $g_\mu(\gamma \cdot) = g$ can be true for at most one value of $\gamma$ it follows that

$$(C_\mu g, g) = (\gamma^2 C g, g), \quad g \in \mathcal{H}.$$ 

The equality $C_\mu = \gamma^2 C$ is seen by polarization.

As stated in the introduction we denote by $\mu_C$ the Gaussian measure on $\mathcal{H}$ with mean zero and covariance operator $C$. Similarly $\mu_\nu^\vee$ denotes the measure given by

$$\mu_\nu^\vee(\mathcal{B}) = \int_0^\infty \mu_{tC}(\mathcal{B}) \, dv(t)$$

for all Borel sets $\mathcal{B}$.

**Theorem 2.2.** Let $\mu$ be an orthogonally invariant measure on an infinite dimensional, separable real Hilbert space $\mathcal{H}$. Then there is a Borel probability measure $\nu$ on $\mathbb{R}^+$ with $1 = \int_0^\infty t \, dv(t)$ and positive nuclear operator $C = C_\mu$ such that $\mu = \mu_C^\vee$. The pair $(C, \nu)$ is unique.

**Proof.** By the proposition we can express $\mu$ as $\hat{\mu}(f) = g_\mu((C_\mu f, f)^{1/2}), f \in \mathcal{H}$. Since $g_\mu$ is continuous, $\hat{\mu}$ is positive definite, and $C_\mu$ has dense range it follows that the function $g_\mu(\|f\|)$ is positive definite on $\mathcal{H}$. Hence by a famous theorem of Schoenberg [2, p. 152] the function $g_\mu(\sqrt{t})$ for $t \geq 0$ is the Laplace transform $\mathcal{L}\nu$ of a Borel probability measure $\nu$ on $\mathbb{R}^+$. For convenience we express this as

$$g_\mu(t) = \int_0^\infty \exp(-t^2 s/2) \, dv(s), \quad t \geq 0. \quad (4)$$

In turn (4) implies

$$\hat{\mu}(f) = \int_0^\infty \exp(-\frac{1}{2}(C_\mu f, f) t) \, dv(t)$$

$$= \int_0^\infty \mu_{tC_\mu}(f) \, dv(t), \quad f \in \mathcal{H}.$$ 

Hence $\hat{\mu}$ equals the transform of the well defined mixed measure $\mu_\nu^\vee$ and the two must be equal. Since

$$C_\mu = \int_0^\infty (tC_\mu) \, dv(t)$$ 

clearly $1 = \int_0^\infty t \, dv(t)$.
If $\mu = \mu_C^p$ is some other representation one finds

$$\hat{\mu}(f) - g_\mu((C_\mu f, f)^{1/2}) - \tilde{g}((\tilde{C}_f, f)^{1/2}),$$

where $\tilde{g}(\sqrt{2t}) = (\mathcal{L}\tilde{v})(t)$ and $g_\mu(\sqrt{2t}) = (\mathcal{L}v)(t)$, $t \geq 0$. The desired identification $(\tilde{v}, \tilde{C}) = (v, C_\mu)$ follows from combining $\tilde{g}(0) = \int_0^{\infty} t \, d\tilde{v}(t) = -1$ with the proposition above and the injectivity of the Laplace transform.

It is apparent that the projection $\mu \circ p^{-1}$ of $\mu = \mu_C$ onto a finite dimensional subspace of $\mathcal{H}$ is absolutely continuous w.r.t. Lebesgue measure $m$ with a Radon-Nikodym derivative $d(\mu \circ p^{-1})/dm$ which is $\mathcal{C}^\infty$ outside zero. If $v$ vanishes in a neighborhood of zero $d(\mu \circ p^{-1})/dm$ even belongs to the Schwartz space $\mathcal{S}$. In contrast, for any finite dimension, normalized integration over the boundary of the unit ball is an orthogonally invariant measure singular w.r.t. Lebesgue measure. In dimension one this is $\mu = \frac{1}{2}(\delta_1 + \delta_{-1})$ with transform $\hat{\mu}(t) = \cos(t) = \cos((t^2)^{1/2})$ which is not even positive.

The two corollaries to Theorem 2.2 and Proposition 2.1 demonstrate that the Gaussian measures have properties which are quite distinct from those of a general orthogonally invariant measure.

**Corollary 2.3.** If an orthogonally invariant measure $\mu$ is a product measure with respect to a non-trivial orthogonal decomposition $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, then $\mu$ is a Gaussian measure.

**Proof.** Choose non-zero vectors $f_i$ in $\mathcal{H}_i$ and put $\beta_{ij} = (C_\mu f_i, f_j)$. Denote by $G_\mu$ the function $G_\mu(t) = g_\mu(\sqrt{t})$, $t \geq 0$, which by l'Hôpital's rule satisfies

$$\lim_{t \to 0^+} G_\mu'(t) = -\frac{1}{2}. \quad (5)$$

By hypothesis $\hat{\mu}(\lambda_1 f_1 + \lambda_2 f_2) = \hat{\mu}(\lambda_1 f_1) \hat{\mu}(\lambda_2 f_2)$. Consequently, as $\beta_{12}$ is readily shown to be zero,

$$G_\mu(\lambda_1^2 \beta_{11} + \lambda_2^2 \beta_{22}) = G_\mu(\lambda_1^2 \beta_{11}) G_\mu(\lambda_2^2 \beta_{22}), \quad \lambda_i \in \mathbb{R}. \quad (6)$$

In combination with (5) the functional equation (6) implies $G_\mu(t) = \exp(-t/2)$, $t \geq 0$. Thus $\mu = \mu_C$.

For any measure $\lambda$ (on $\mathbb{R}_+$) and positive real number $\alpha$ denote by $\lambda^\alpha$ the measure

$$\int F(x) \, d\lambda^\alpha(x) = \int F(x) \, d\lambda(x).$$

**Corollary 2.4.** Let $\mu$ be the convolution measure $\mu = \mu_1 * \mu_2$, where $\mu_i = \mu_C^p$. 

Then $\mu$ is orthogonally invariant only if either $\mu_i$ are both Gaussian or the covariances $C_i$ are proportional.

**Proof.** The proof is based on (b) of Proposition 2.1. Since $C_\mu = C_1 + C_2$ is known we may set out to determine when $\mu \circ l_g^{-1}, \ g \in \mathcal{K}\setminus\{0\}$, are all equal. Notation will be as in the proof of Proposition 2.1.

Now $\mu \circ l_g^{-1} = (\mu_1 \circ l_g^{-1}) * (\mu_2 \circ l_g^{-1})$ and for

$$a = (C_1 g, g)^{1/2} ((C_1 + C_2) g, g)^{-1/2}$$

one finds that

$$\tilde{\mu} \circ l_g^{-1}(s) = \left(\tilde{\mu}_1 \circ l_g^{-1}\right)(s) \cdot \left(\tilde{\mu}_2 \circ l_g^{-1}\right)(s)$$

$$= \left(\tilde{\mu}_1\right)^a(s) \cdot \left(\tilde{\mu}_2\right)^{1-a^2}((1-a^2)^{1/2} s)$$

Thus the requirement is that the functions

$$g_a(s) = g_1(as) g_2((1-a^2)^{1/2} s)$$

should be independent of the parameter $a$ as it ranges over the closure $I = K^-$ of the set

$$K = \{ \| C_1^{1/2} g \| \cdot \| (C_1 + C_2)^{1/2} g \|^{-1} \ | \ g \in \mathcal{K}\setminus\{0\} \}.$$ 

But $I$ is precisely the set $\{ \| C_1^{1/2}(C_1 + C_2)^{-1/2} f \| \ | \ | f \| = 1 \}^-$ which in turn is identical to the square root $W^{1/2}$ of the numerical range

$$W = \{ ((C_1 + C_2)^{-1/2} C_1 (C_1 + C_2)^{-1/2} f, f) \ | \ | f \| = 1 \}^-.$$

In particular, $I$ is an interval.

In case $I$, and hence $W$, is a singleton set we find by polarization that the $C_i$ are proportional. Otherwise we may differentiate (7) with respect to $a$ in the interior of $I$ to obtain

$$(1-a^2)^{1/2} g_1'(as) g_2((1-a^2)^{1/2} s)$$

$$= a g_1(as) g_2'(((1-a^2)^{1/2} s), \quad a \in I, s \in \mathbb{R}.$$ 

Since $g_i$ are everywhere positive this can be rewritten as

$$(1-a^2) \frac{d}{ds} (\ln g_1(as))$$

$$= a^2 \frac{d}{ds} (\ln g_2((1-a^2)^{1/2} s)), \quad a \in I, s \in \mathbb{R}.$$
and consequently for any fixed \( a \) in \( I \)
\[
g_1(as)^{(1-a^2)} = g_2((1-a^2)^{1/2}s)^2, \quad s \in \mathbb{R}. \tag{8}
\]

On comparison with (7), with the common value of \( g_a \) denoted by \( \tilde{g} \), this yields
\[
g_1(as) = \tilde{g}(s)^2, \quad a \in I, \quad s \in \mathbb{R},
\]
and it follows that the function \( g_1(t) \) equals the function \( \exp(t^2 \ln \tilde{g}(1)) \) through the interval \( I \). Since \( g_1(\sqrt{\cdot}) \) is the Laplace transform of a probability measure it has an analytic continuation to the open right half plane. In turn \( g_1 \) has an analytic extension to the interior of a \( 45^\circ \) cone symmetrically including the positive real axis. Thus from uniqueness of analytic continuation and the condition \( g_1'(0) = -1 \) the identity \( g_1(t) = e^{-t^2/2}, \quad t \in \mathbb{R}_+ \), follows. Hence \( \mu_1 \), and likewise \( \mu_2 \), are Gaussian.

Remarks. (a) To connect Corollary 2.4 with our introductory considerations regarding random variables \( X_T \), let \( Z \) be a sum \( Z = X_T + Y_S \) of two independent variables of this kind. Corollary 2.4 states that if the covariance parameters of \( (X_T)_{t \geq 0} \) and \( (Y_S)_{s \geq 0} \) are not proportional then \( Z \) has an orthogonally invariant distribution only if \( T \) and \( S \) are constants.

(b) One property, however, characteristic of Gaussian measures is preserved for orthogonally invariant measures \( \mu = \mu_C \). When \( \{e_j\}_{j=1}^\infty \) and \( \{\lambda_j\}_{j=1}^\infty \) are the eigenvectors and corresponding eigenvalues of \( C \) the limit
\[
\lim_{n \to +\infty} \frac{1}{n} \sum_{j=1}^{n} (f, e_j)^2/\lambda_j \text{ still exists for } \mu \text{ almost all } f \text{ in } \mathcal{H}.
\]
But it need no longer be equal to one, \( \mu \) a.e. In fact \( \nu \) is equal to \( \mu \circ \rho^{-1} \) and \( \mu_C \), \( t \in \mathbb{R}_+ \), are the conditional measures for \( \{\rho(f) = t\} \), where
\[
\rho(f) = \lim_{n \to +\infty} \frac{1}{n} \sum_{j=1}^{n} \frac{(f, e_j)^2}{\lambda_j}
\]
can be given any value on the set of non-convergence.

3. APPROXIMATION OF LINEAR OPERATORS

This section investigates the approximation of a linear bounded operator \( A : \mathcal{H} \to \mathcal{H} \) with respect to some fixed orthogonally invariant measure \( \mu = \mu_C \). First it is necessary to introduce further definitions and notations.

When the Hilbert space \( \mathcal{H} \) is identified with its own dual space of functionals, an adaptive linear information operator \( N : \mathcal{H} \to \mathbb{R}^n \) is any map of the form \( Nf = (y_i)_{i=1}^{n} \), where \( y_1 = (f, g_1), \ y_2 = (f, g_2(y_1)), \ldots \),
\( y_n = (f, g_n(y_1, ..., y_{n-1})) \), and \( g_i: \mathbb{R}^{i-1} \rightarrow \mathcal{H} \) are measurable functions for \( 1 \leq i \leq n \). Thus the \( i \)th point of evaluation is allowed to depend (measurably) on the previous \((i-1)\) outcomes. The information operator or just information \( N \) is called \textit{non-adaptive} if the \( g_i \) are constant functions, i.e., the points of evaluation have been chosen a priori. Given an error functional \( E \) the \textit{error} of an algorithm \( \phi \) is defined by

\[
e(\phi, N) = \int_{\mathcal{H}} E(Af - \phi Nf) \, d\mu(f)
\]

and the \textit{radius} of an information operator \( N \) is

\[
r(N) = \inf_{\phi} e(\phi, N).
\]

Without essential loss of generality it is assumed that

\[
(Cg_i(y), g_j(y)) = \delta_{ij}
\]

holds for almost all \( y \) in \( \mathbb{R}^n \). For \( y = (y_i)_{i=1}^n \) in \( \mathbb{R}^n \) of course \( g_i(y) \) means \( g_i(y_1, ..., y_{i-1}) \). Also for \( y \) in \( \mathbb{R}^n \) we adopt the notation [13]

\[
m(y) = \sum_{j=1}^{n} y_j Cg_j(y)
\]

\[
\sigma(y) = \sum_{j=1}^{n} g_j(y) \otimes Cg_j(y)
\]

and

\[
S(y) = (I - \sigma(y)) C(I - \sigma(y))^*
\]

The measure \( \mu_C \) is transformed by \( N \) into the measure \( \mu_C^N \) on \( \mathbb{R}^n \). This is readily verified when the \( g_i \) constantly equal suitably normalized eigenvectors for \( C \); the general case then follows from [14, Theorem 4.2]. In [13, Theorem 3.1] it is shown that for \( \mu = \mu_C \) the conditional measure for \( \{Nf = y\} \) is the Gaussian measure \( \mu_{m(y), S(y)} \) with mean \( m(y) \) and covariance \( S(y) \), i.e.,

\[
\mu_C = \int_{\mathbb{R}^n} \mu_{m(y), S(y)} \, d\mu_f(y)
\]

with each \( \mu_{m(y), S(y)} \) supported on \( \{f \mid Nf = y\} \). The next proposition determines the corresponding resolution of an orthogonally invariant \( \mu \) with respect to \( N \).
Denote by $W_s$ the density function $W_s(y) = (2\pi s)^{-n/2} \exp(-\|y\|^2/2s)$, $y \in \mathbb{R}^n$, and by $W^\nu$ the Radon–Nikodym derivative

$$W^\nu(y) = \frac{d\mu^\nu}{dm}(y) = \int_0^\infty W_s(y) \, dv(s).$$

**Proposition 3.1.** It holds that

$$\mu^\nu_c = \int_{\mathbb{R}^n} \mu^\nu \, d\mu^\nu_c(y),$$

where each probability measure

$$\mu^\nu = W^\nu(y)^{-1} \int_0^\infty \mu_{m(y), sS(y)S(y)} W_s(y) \, dv(s)$$

is supported on $\{f \mid Nf = y\}$.

**Proof.** Application of (9) to the covariance operators $C = sC$ and the informations $\mathcal{N}$ given by $\tilde{g}_i = s^{-1/2}g_i$ demonstrates that

$$\mu_{sc} = \int_{\mathbb{R}^n} \mu_{m(s^{1/2}y), sS(y)} \, d\mu_{s}(y)$$

and that each $\mu_{m(s^{1/2}y), sS(y)}$ is supported on $\{f \mid Nf = s^{1/2}y\}$. Thus

$$\mu^\nu_c = \int_0^\infty \int_{\mathbb{R}^n} \mu_{m(s^{1/2}y), sS(y)} \, d\mu_{s}(y) \, dv(s)$$

which after reshuffling, using

$$\int_0^\infty \int_{\mathbb{R}^n} F(s, y) \, d\mu_{s}(y) \, dv(s)$$

$$= \int_{\mathbb{R}^n} \int_0^\infty F(s, s^{-1/2}y) \, W_s(y) \, dv(s) \, dy,$$

becomes (10) and (11). □

In the sequel the following very general class of error functionals is considered. A measurable function $E: \mathcal{H} \to \mathbb{R}_+$ is called an *allowable* error functional if each set

$$\mathcal{B}_t = \{f \in \mathcal{H} \mid E(f) < t\}$$

is convex and balanced. This includes the average squared error and the error in probability. Moreover every convex function $E: \mathcal{H} \to \mathbb{R}_+$ with
$E(0) = 0$ is allowable. We shall refer to $E$ as a standard error function if $\mathcal{B}$ has the form $\mathcal{B} = F(t)\mathcal{B}$, where $\mathcal{B}$ is a bounded, convex, open set containing zero and $F$ is a continuous bijection of $\mathbb{R}_+$. In this case $E$ is given by $E(f) = G(p_\mathcal{B}(f))$ for $G = F^{-1}$ and the continuous Minkowski seminorm $p_\mathcal{B}(f) = \inf\{ t > 0 | f \in t\mathcal{B} \}$. The set of standard functionals includes in particular functions of the type $E(f) = H(\|f\|)$.

**Lemma 3.2.** Let $E$ be an allowable error functional and let $\mu_C$ be a Gaussian measure. Then the function

$$\chi(g) = \int_{\mathcal{H}} E(f-g) \, d\mu_C(f)$$

of $g$ in $\mathcal{H}$ attains its minimum value at $g = 0$. If $\chi(0)$ is finite and $E$ is a standard functional this minimum is unique.

**Proof.** The main tool here is the identity

$$\chi(g) = \int_{\mathcal{H}} E(f) \, d\mu_{g,C}(f) = \int_0^\infty t \, d(\mu_{g,C}(\mathcal{B}_t)).$$

(12)

Optimality of $g = 0$ follows from $\mu_{g,C}(\mathcal{B}_t) \leq \mu_C(\mathcal{B}_t)$, $t \in \mathbb{R}_+$, which holds by the hypothesis on $\mathcal{B}_t$, cf. [13, Lemma 3.11] and [1, Theorem 1].

If conversely, $\chi(0) = \chi(g) < +\infty$ then necessarily $\mu_{g,C}(\mathcal{B}_t) = \mu_C(\mathcal{B}_t)$, $t \in \mathbb{R}_+$. For the standard case $\{ \mathcal{B}_t \}_{t \geq 0}$ is equal to $\{ t\mathcal{B} \}_{t \geq 0}$. From (12) and the symmetry of $\mu_C$

$$\int_{\mathcal{H}} p_\mathcal{B}(f) \, d\mu_C(f) = \int_{\mathcal{H}} p_\mathcal{B}(f-g) \, d\mu_C(f)$$

$$= \int_{\mathcal{H}} \frac{1}{2} (p_\mathcal{B}(f-g) + p_\mathcal{B}(f+g)) \, d\mu_C(f).$$

Combined with the convexity of $p_\mathcal{B}$ this implies

$$2p_\mathcal{B}(f) = p_\mathcal{B}(f-g) + p_\mathcal{B}(f+g), \quad \mu_C \text{ a.e.}$$

(13)

Take a sequence $f_n \to 0$ for which (13) holds. Then by the continuity of $p_\mathcal{B}$ ($\mathcal{B}$ open), $p_\mathcal{B}(g) = 0$, and by the faithfulness of $p_\mathcal{B}$ ($\mathcal{B}$ bounded) $g = 0$.

When $N$ is non-adaptive the constant values of $S(y)$ and $g_*(y)$ are simply denoted $S$ and $g_*$. 

**Theorem 3.3.** Assume that $N$ is non-adaptive information, $\mu = \mu_C^*$ is an
orthogonally invariant measure on $\mathcal{H}$, and $E$ is an allowable error functional. Then

(a) The spline algorithm

$$\varphi^\circ: (y_i)_{i=1}^n \rightarrow \sum_{i=1}^n y_i ACG_i$$

is an optimal algorithm. When $e(\varphi^\circ, N)$ is finite and $E$ is a standard functional $\varphi^\circ$ is a unique optimal algorithm.

(b) $r(N) = \int_{\mathcal{H}} E(f) \, d\mu_{ASA^*}(f)$.

(c) When $E$ is $p$-homogeneous, i.e., $E(\alpha f) = |\alpha|^p E(f)$,

$$r(N) = \left( \int_{0}^{\infty} s^{p/2} \, dv(s) \right) \int_{\mathcal{H}} E(f) \, d\mu_{ASA^*}(f).$$

Proof: (a) Due to Proposition 3.1

$$e(\varphi, N) = \int_{\mathcal{H}} \int_{\mathcal{H}} E(Af - \varphi(y)) \, d\mu^*(f) \, d\mu^*(y), \quad (15)$$

where

$$\int_{\mathcal{H}} E(Af - \varphi(y)) \, d\mu^*(f) = W^*(y)^{-1} \int_{0}^{\infty} W_s(y) \times \int_{\mathcal{H}} E(Af - \varphi(y)) \, d\mu_{m(y),AS}(f) \, dv(s)$$

$$= W^*(y)^{-1} \int_{0}^{\infty} W_s(y) \times \int_{\mathcal{H}} E(f - (\varphi(y) - Am(y))) \times d\mu_{ASA^*}(f) \, dv(s). \quad (16)$$

From this and Lemma 3.2 it is clear that the algorithm $\varphi^\circ(y) - Am(y)$ (for almost all $y$) has the desired properties.

(b) Just combine (15), (16), and (a).

(c) This is a consequence of the general relation

$$\int_{\mathcal{H}} F(f) \, d\mu_{yK}(f) = \int_{\mathcal{H}} F(\gamma f) \, d\mu_K(f).$$
It is seen from 3.3(c) that for $E(f) = p_{\mathcal{B}}(f)^p$ and other $p$-homogeneous functions the approximation problem for $\mu_C^*$ is equivalent to the one for $\mu_C$.

Next we want to consider a restricted class of sets $\mathcal{B}$. But before we do so it is appropriate for us to touch on the problem of optimal information. Denote by $R_n$ the operator

$$R_n = ASA^* = A(I - \sigma) C(I - \sigma)^* A^*$$

and define the $n$th radius of the approximation problem to be

$$r^n = \inf_N r(N).$$

In the next proposition it is tacitly assumed that all eigenvalues of $ACA^*$ are non-degenerate. The general case is similar but more complicated to state.

**Proposition 3.4.** Assume that $E$ is a standard error functional of the form $E(f) = H(\|f\|)$.

Then $r^n = r(N)$, where the information $N$ is given via the $n$ principal eigenvalues and eigenvectors $(\lambda_i, f_i)$ of $ACA^*$ through $\tilde{g}_i = \lambda_i^{-1/2} A^* f_i$. If $N$ is any information then $r(N) = r(\tilde{N})$ if and only if

$$Rg_1 + \cdots + Rg_n = R\tilde{g}_1 + \cdots + R\tilde{g}_n.$$

**Proof.** By 3.3(b) the value of $r(N)$ increases when the eigenvalues of $ASA^*$ increase. Compute

$$R_n = A \left( I - \sum_{i=1}^n g_i \otimes Cg_i \right) \left[ \left( I - \sum_{i=1}^n Cg_i \otimes g_i \right) A^* \right]$$

$$= AC \left( I - \sum_{i=1}^n Cg_i \otimes g_i \right)^2 A^*$$

$$= AC \left( I - \sum_{i=1}^n Cg_i \otimes g_i \right) A^*$$

$$= AC^{1/2} \left( I - \sum_{i=1}^n C^{1/2}g_i \otimes C^{1/2}g_i \right) C^{1/2} A^*.$$

Then $R_n$ is given by

$$R_n = AC^{1/2}(I - P) C^{1/2} A^*,$$

where $P$ is the orthogonal projection onto the linear span of $\{C^{1/2}g_i\}_{i=1}^n$. 

The non-zero eigenvalues of \( R_n = (AC^{1/2}(I - P))(AC^{1/2}(I - P))^* \) equal those of

\[
\tilde{R}_n = (AC^{1/2}(I - P))^* (AC^{1/2}(I - P)) = (I - P) C^{1/2} A^* A C^{1/2}(I - P).
\]

Repeating the argument we note that the (non-zero) eigenvalues of \( C^{1/2} A^* A C^{1/2} \) are \( \{\lambda_i\}_{i=1}^\infty \). Consequently a minimal set of eigenvalues for \( R_n \), namely \( \{\lambda_i\}_{i=n+1}^\infty \), exists and is obtained if and only if

\[
\Re C^{1/2} g_1 + \cdots + \Re C^{1/2} g_n = \Re \eta_1 + \cdots + \Re \eta_n,
\]

where \( \eta_i \) are the \( n \) principal eigenvectors of \( C^{1/2} A^* A C^{1/2} \). However, \( \eta_i \) are proportional to \( C^{1/2} A^* f_i \) and (17) is equivalent to

\[
\Re g_1 + \cdots + \Re g_n = \Re A^* f_1 + \cdots + \Re A^* f_n. \tag{17}
\]

The above proposition, which improves [13, pp. 738-741], is included at this point mainly to emphasize that the directions in \( \mathcal{H} \) determined by the eigenvectors \( \{f_i\}_{i=1}^\infty \) of \( ACA^* \) have a special significance. Thus prepared the reader will hopefully admit to the relevance of the sets \( \mathcal{B} \) in the following corollaries to Theorem 3.3.

**Corollary 3.5.** Let \( E \) be the functional \( E(f) = G(p_B(f)) \), where \( B \) is defined by

\[
B = \left\{ f \in \mathcal{H} \left| \sum_{i=1}^\infty a_i(f, f)^2 < 1 \right. \right\}
\]

for some bounded set \( \{a_i\}_{i=1}^\infty \) of positive numbers and \( G \) is a continuously differentiable bijection of \( \mathbb{R}_+ \). Let \( \theta \) be the real part of the function

\[
\varphi(\lambda) = \prod_{j=n+1}^\infty (1 - 2i\lambda_j a_j)^{-1/2}.
\]

Then

\[
r(N) = \sqrt{\frac{2}{\pi}} \int_0^\infty \int_0^\infty G((ts)^{1/2}) \hat{\theta}(t) dt dv(s), \tag{18}
\]

where \( \hat{\theta} \) denotes the Fourier transform.

**Proof.** First we claim that

\[
\mu_{sA A^*}(t \mathcal{B}) = \frac{1}{\pi} \int_{-\infty}^\infty \frac{\sin(t^2 \lambda)}{\lambda} \theta(s \lambda) d\lambda.
\]
We shan't go into the details of this. The proof is an application of the characteristic function trick that can be found for instance in [17, pp. 66]. Now for $F = G^{-1}$

$$
\frac{d}{dt} (\mu_{\lambda \phi}(F(t) \phi))
$$

$$
= \frac{1}{\pi} \int_{-\infty}^{\infty} 2F'(t) F(t) \cos(F(t)^2 \lambda) \theta(s \lambda) d\lambda
$$

$$
= \sqrt{\frac{2}{\pi}} 2s^{-1}F'(t) F(t) \hat{\theta}(s^{-1}F(t)^2)
$$

$$
= \sqrt{\frac{2}{\pi}} \frac{d}{dt} (s^{-1}F(t)^2) \hat{\theta}(s^{-1}F(t)^2)
$$

and by Theorem 3.3(b)

$$
r(\bar{N}) = \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} \int_{0}^{\infty} t \frac{d}{dt} (s^{-1}F(t)^2) \times \hat{\theta}(s^{-1}F(t)^2) dt dv(s)
$$

$$
= \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} \int_{0}^{\infty} G((ts)^{1/2}) \hat{\theta}(t) dt dv(s).
$$

This is (18).

**COROLLARY 3.6.** For each $E(f) = p_{\lambda}(f)^{2p}$ it holds that

$$
r(\bar{N}) = r(\bar{N}, p_{\lambda}^{2p}) = (-i)^{p} \varphi^{(p)}(0) \int_{0}^{\infty} s^{p} dv(s) . \tag{19}
$$

In particular

$$
r(\bar{N}, p_{\lambda}^{2p}) = \sum_{j=n+1}^{\infty} \lambda_{j} \alpha_{j}
$$

and

$$
r(\bar{N}, p_{\lambda}^{2p}) = \left( 2 \sum_{j=n+1}^{\infty} (\lambda_{j} \alpha_{j})^{2} + \left( \sum_{j=n+1}^{\infty} \lambda_{j} \alpha_{j} \right)^{2} \right) \int_{0}^{\infty} s^{2} dv(s).
$$

**Proof.** From (18)

$$
r(\bar{N}, p_{\lambda}^{2p}) = \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} t^{p} \hat{\theta}(t) dt \int_{0}^{\infty} s^{p} dv(s).
$$

Here the second factor may or may not be finite. Our objective is to deter-
mine the value of the first factor. Since \( \theta(t) = \frac{1}{2}(\varphi(t) + \varphi(-t)) \) it follows that \( \hat{\theta}(t) = \frac{1}{2}(\hat{\varphi}(t) + \hat{\varphi}(-t)) \). Hence

\[
\frac{\sqrt{2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^p \hat{\theta}(t) \, dt
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^p (\hat{\varphi}(t) + \hat{\varphi}(-t)) \, dt
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |t|^p \hat{\varphi}(t) \, dt. \tag{20}
\]

When \( p \) is even this is

\[
\frac{(-i)^p}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\varphi^p)^\wedge(t) \, dt = (i)^p \varphi^{(p)}(0)
\]

and we are done. For odd \( p \) (20) can be rewritten as

\[
\frac{(-i)^p}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \text{sgn}(t) (\varphi^p)^\wedge(t) \, dt. \tag{21}
\]

The function or tempered distribution \( \text{sgn} \) has Fourier transform \( \hat{\text{sgn}} = (-i)^p \sqrt{2/\pi} \text{Vp}(1/t) \), where \( \text{Vp} \) denotes the Cauchy principal value. Thus (20) is equal to

\[
\frac{(-i)^{p+1}}{\pi} \lim_{\varepsilon \to 0^+} \int_{|t| \geq \varepsilon} \frac{\varphi^{(p)}(t)}{t} \, dt. \tag{22}
\]

To estimate this integral we exploit the fact that \( z^{-1/2} \) is an analytic function of \( z \) in the half plane \( \{ z \in \mathbb{C} \mid \text{Re} \, z > 0 \} \). Indeed \( z^{-1/2} = (1/\sqrt{\pi}) \int_{-\infty}^{\infty} e^{-x^2} \, dx \). In turn \( \varphi \) is analytic in the region \( \{ z \in \mathbb{C} \mid \text{Im} \, z > -\gamma \} \), where \( \gamma = (\max_j 2\lambda_j a_j)^{-1} \) and the integral of \( \varphi(z)/z \) along the contour indicated in Fig. 1 is zero for any values of \( \varepsilon \) and \( R \).
Since the integral along the semicircle \( \Gamma_R \) converges to zero it follows that the limit in (22) is in fact

\[
\frac{(-i)^{p+1}}{\pi} \lim_{\varepsilon \to 0^+} \int_0^\pi \frac{\varphi^{(p)}(\varepsilon e^{i\theta})}{\varepsilon e^{i\theta}} i\varepsilon e^{i\theta} \, d\theta = (-i)^p \varphi^{(p)}(0).
\]

Finally, this expression is calculated for the specific values \( p = 1 \) and \( p = 2 \) by use of

\[
\varphi'(z) = A(z) \varphi(z), \tag{23}
\]

where

\[
A(z) = i \sum_{j=n+1}^\infty \frac{\lambda_j a_j}{(1-2iz\lambda_j a_j)}.
\]

Remark. By iterating (23) and using

\[
A^{(k)}(0) = (i)^{k+1} 2^k \cdot k! \sum_{j=n+1}^\infty (\lambda_j a_j)^{k+1}
\]

one may of course generate any desired instance of \((-i)^p \varphi^{(p)}(0)\). But we have not been able to find a closed expression for this.

**Corollary 3.7.** For \( E(f) = \|f\|^{2p} \) the nth radius \( r^n = r^n(\| \cdot \|^{2p}) \) of the approximation problem is

\[
r^n(\| \cdot \|^{2p}) = (-i)^p \varphi^{(p)}(0) \int_0^\infty s^p \, dv(s),
\]

where \( \varphi(z) = \prod_{j=n+1}^\infty (1-2iz\lambda_j)^{-1/2} \). In particular

\[
r^n(\| \cdot \|^2) = \sum_{j=n+1}^\infty \lambda_j
\]

and

\[
r^n(\| \cdot \|^4) = \left(2 \sum_{j=n+1}^\infty \lambda_j^2 + \left(2 \sum_{j=n+1}^\infty \lambda_j \right)^2 \right) \int_0^\infty s^2 \, dv(s).
\]

**Proof.** Combine Corollary 3.6 and Proposition 3.4.

For \( E(f) = \|f\|^p \) and other standard functionals of the form \( E(f) = H(\|f\|) \) we may derive a rather nice expression for the optimal approxima-
tion $\varphi N$. From Theorem 3.3(a) and Proposition 3.4 it follows that $\varphi N$ is optimal if and only if $\mathbb{R}g_1 + \cdots + \mathbb{R}g_n = \mathbb{R}\tilde{g}_1 + \cdots + \mathbb{R}\tilde{g}_n$ and $\varphi$ has the form (14). Again let $P$ denote the projection onto the linear span of $\{\tilde{g}_i\}_{i=1}^n$. Since $\{C^{1/2}g_i\}_{i=1}^n$ is an orthonormal basis in $C^{1/2}P$ one finds, for any $f$ in the domain of $C^{-1/2}$,

$$\varphi^*N(f) = \sum_{i=1}^n (f, g_i) AC g_i$$

$$= AC^{1/2} \sum_{i=1}^n (C^{-1/2}f, C^{1/2}g_i) C^{1/2}g_i$$

$$= AC^{1/2}QC^{-1/2}f,$$

where $Q$ is the projection onto $C^{1/2}P = \text{span}\{\eta_i\}_{i=1}^n$ (cf. the proof of Proposition 3.4). Here the equation $\varphi^*N(f) = AC^{1/2}QC^{-1/2}f$ is independent of the choice of $\{g_i\}_{i=1}^n$. Consequently $AC^{1/2}QC^{-1/2}$ extends to a bounded operator in $\mathcal{H}$ and this operator is the unique optimal value of $\varphi N$. Finally, using $g_i = \tilde{g}_i = \lambda_i^{-1/2}A* f_i$ and the very definition of $f_i$, one finds

$$(\varphi N)^{\text{optimal}}(f) = \sum_{i=1}^n (f, \lambda_i^{-1/2}A* f_i) AC(\lambda_i^{-1/2}A* f_i)$$

$$= \left( \sum_{i=1}^n A* f_i \otimes f_i \right) (f)$$

$$= \left( \sum_{i=1}^n f_i \otimes f_i \right) Af.$$

Thus $(\varphi N)^{\text{opt}}$ is the composition of $A$ and an orthogonal projection of rank $n$.

Finally, in closing this paper, we turn to the problem of adaptive versus non-adaptive information. When $N$ is (adaptive) information let $N_y$, $y \in \mathbb{R}^n$, be the non-adaptive information given by $g_i = g_i(y)$.

The heart of the very elegant proof in [13] that "adaptation doesn't help" is the equality $\mu^r(N) = \mu^r(N_y)$ between conditional measures. It is apparent from (11) that this does not hold generally for non-Gaussian measures $\mu^r$. Nevertheless we have the following.

**Theorem 3.8.** For any allowable error functional $E$ and any information $N$

$$r(N) \geq \inf_{y \in \mathbb{R}^n} r(N_y). \quad (24)$$
Further if $E$ is a standard error functional of the form $E(f) = H(\|f\|)$, $r(N) = r^n$ if and only if
\[
\mathbb{R}g_1(y) + \cdots + \mathbb{R}g_n(y) = \mathbb{R}\tilde{g}_1 + \cdots + \mathbb{R}\tilde{g}_n
\]  \hspace{1cm} (25)
holds for almost all $y$ in $\mathbb{R}^n$.

Proof. Using the results of Proposition 3.1, Lemma 3.2, and Theorem 3.3 compute
\[
e(\varphi, N) = \int_{\mathcal{F}} E(\varphi Nf) \, d\mu(f)
= \int_{\mathbb{R}^n} E(\varphi(y)) \, d\mu^y(f) \, d\mu^y_{\tilde{g}}(y)
= \int_{\mathbb{R}^n} \int_0^\infty W_s(y) \int_{\mathcal{F}} E(\varphi(y))
\times d\mu_{m(y), sN(-1/2)}(f) \, dv(s) \, dy
\ge \int_{\mathbb{R}^n} \int_0^\infty W_s(y) \int_{\mathcal{F}} E(\varphi(f))
\times d\mu_{sN(-1/2)}(f) \, dv(s) \, dy
= \inf_{y \in \mathbb{R}^n} \int_0^\infty \int_{\mathcal{F}} E(\varphi(f)) \, d\mu_{sN(y)}(f) \, dv(s)
= \inf_{y \in \mathbb{R}^n} r(N_y).
\]
This proves (24).

For the final case to be considered it can be read of the above string of calculations that $r(N) = r^n$ if and only if
\[
\varphi(y) = Am(y) \hspace{1cm} (26a)
\]
\[
r^n = r(N_y) \hspace{1cm} (26b)
\]
holds for almost all $y$ in $\mathbb{R}^n$. Combining (26b) with Proposition 3.4 one gets (25). The optimal algorithm is given by
\[
\varphi((y_i)_{i=1}^n) = \sum_{i=1}^n y_i ACg_i(y_1, \ldots, y_{i-1})
\]
for almost all $y = (y_i)_{i=1}^n$ in $\mathbb{R}^n$.  \hspace{1cm} \blacksquare
REFERENCES