Optimal Algorithm for Linear Problems with Gaussian Measures

CUCS-176-85

G. W. Wasilkowski

G.W. Wasilkowski

Department of Computer Science

Columbia University

February, 1984

This research was supported in part by the National Science Foundation under Grant MCS-7823676.

Abstract

We study optimal algorithms for linear problems in two settings: the average case and the probabilistic case settings. We assume that the probability measure is Gaussian. This assumption enables ut to consider a general class of error criteria. We prove that in both settings adaption does not help and a translated spline algorithm is optimal. We also devise optimal information under some additional assumptions concerning the error criterion.

1. Introduction

In this paper we study the optimal reduction of uncertainty for linear problems in two settings: the average case setting and the probabilistic case setting.

By a linear problem we mean the problem of approximating Sf, where S is a linear operator defined on a separable Hilbert space F_1, when only partial information Nf on f is available. This partial information causes uncertainty. In the average case setting the intrinsic uncertainty is measured by the average size of the error of the best possible algorithm that uses N. In the probabilistic case setting it is measured by the probability that the error of the best possible algorithm is small. In this paper we assume that the probability measure on the space F_1 is Gaussian and the difference between Sf and x, the value given by an algorithm, is measured by $E(Sf-x)$, where E is an arbitrary error functional.

The average case setting has been studied in [5,7,8] for rather general class of probability measures assuming however that the error functional is of a special case. Typically it is assumed that $E(Sf-x) = \|Sf-x\|^2$ and $S(F_1)$ is a separable Hilbert space. Here restricting the class of probability measures to Gaussian measures we relax the

assumption concerning the problem and the form of the error functional E. We are able also to study the probabilistic case setting.

The following results are obtained for both the average case and the probabilistic case settings:

1° For every error functional and for every adaptive information N^a there exists nonadaptive information on the same structure as N^a with uncertainty not greater than the uncertainty caused by N^a. Thus adaption does not help.

2° For every error functional and for every nonadaptive information N a translated spline algorithm is optimal. A sufficient condition for the spline algorithm to be optimal is given.

3° Optimal information N^* is exhibited under some additional assumptions concerning the error functional E.

We now comment on the results mentioned above. The result 1° states that adaptive information is not more powerful than nonadaptive information in either setting. A similar result for the average case setting has been established in [5,8]. This is not merely of theoretical interest since adaptive information has several undesirable properties like eg.:

--It has more complicated structure than nonadaptive information

--It is ill-suited for parallel computation, whereas nonadaptive information can be computed very efficiently in parallel.

Since adaptive information does not decrease the uncertainty, it may be replaced in practice by nonadaptive information. We want to stress that many commonly used algorithms use adaptive information.

We comment on the result 2° which states that in both settings a translated spline algorithm φ^* is optimal. (For a similar result for the average case setting see [5,7,9].) Since the spline algorithm is linear, the optimal algorithm φ^* is affine. Hence the cost of evaluating φ^* for given $y = Nf$ is proportional to the cost of evaluating $y = Nf$. This is a desirable property from the complexity point of view.

The result 3° gives us the best information to be used, i.e., information which minimizes the uncertainty in two settings.

We now summarize the contents of the paper. In Section 2 we formulate the problem. In Section 3 we derive some properties of Gaussian measures. These properties will play a key role in the rest of this paper.

In Section 4 we study the average case setting, and we prove that 1°, 2° and 3° hold for that setting. In Section 5 we study the probabilistic case setting, and we prove that 1°, 2° and 3° hold for that setting. In Section 6 we prove that the spline algorithm enjoys one more optimality property. Namely, assuming that the error functional $E(Sf-x) = \|Sf-x\|^2$, the spline algorithm minimizes the variance.

2. Basic Concepts

Our aim is to approximate the solution operator S,

$$S: F_1 \to F_2.$$

We assume that S is linear, F_1 is a separable Hilbert space and F_2 is a linear space, both F_1 and F_2 over the real field. Hence we want to construct an element $x = x(f) \in F_2$ which approximates Sf, $\forall f \in F_1$, with a small error. The error between Sf and X is measured by $E(Sf - x)$, where

$$E: F_2 \to \mathbf{R}_+,$$

is called an error functional. For example, E might be of the form $E(g) = \|g\|^p$ if F_2 is normed. Here we consider a general class of error functionals. The only assumption concerning E is that for every $g \in F_2$, $H(\cdot) \stackrel{df}{=} E(S(\cdot)-g)$ is measurable, i.e., $H^{-1}(B) \in \mathbf{B}(F_1)$ whenever $B \in \mathbf{B}(F_1)$, where $\mathbf{B}(F_1)$ stands for the σ-field of Borel sets from F_1.

To construct $x = x(f)$ we need to know something about f. We assume that our knowledge of f is given by $N^a(f)$. Here N^a is a linear adaptive information operator (for brevity adaptive information), i.e.,

$$N^a(f) = [(f,g_1),(f,g_2(y_1)),\ldots,(f,g_n(y_1,\ldots,y_{n-1}))], \tag{2.1}$$

where $(\cdot,\cdot)$ is the innerproduct in F_1,

$$(2.2) \qquad y_1 = y_1(f) = (f,g_1), \quad y_i = y_i(f)$$

$$= (f,g_i(y_1,\ldots,y_{i-1}), \quad i = 2,3,\ldots,n.$$

For brevity we shall write $g_i(y) = g_i(y_1,\ldots,y_{i-1}) \in F_1$ for every $y = [y_1,\ldots,y_n] \in \mathbb{R}^n$. We assume that $g_i(\cdot)$, as functions of y, are measurable. Without loss of generality we assume that $g_1(y),\ldots,g_n(y)$ are linearly independent for every $y \in \mathbb{R}^n$. By

$$(2.3) \qquad \mathrm{card}(N^a) = n,$$

we mean the cardinality of N^a. Note that in general the ith evaluation $(f,g_i(y_1,\ldots,y_{i-1}))$ depends on the previously computed information $y_1(f),\ldots,y_{i-1}(f)$. That's why N^a is called adaptive. If g_i do not depend on y, $g_i(y) = g_i$, $\forall i$, $\forall y \in \mathbb{R}^n$, then N^a is called <u>nonadaptive</u>. To stress the nonadaptive character of N^a we often write N^{non} instead of N^a. For every adaptive information N^a, by fixing $y \in \mathbb{R}^n$ a priori and letting $g_i \overset{df}{=} g_i(y)$, we obtain a nonadaptive information

$$(2.4) \qquad N_y^{non}(\cdot) = [(\cdot,g_1),\ldots,(\cdot,g_n)]$$

which uses the same evaluations as N^a.

Knowing $N^a(f)$ we construct an approximation $x = x(f)$ by an algorithm φ,

$$x = \varphi(N^a(f)),$$

where by an algorithm φ that uses N^a we mean any mapping

$$\varphi: N^a(F_1) = \mathbb{R}^n \to F_2. \tag{2.5}$$

We are interested in optimal algorithms, that is algorithms with minimal errors. What we mean by the error of an algorithm depends on the setting we are dealing with. In this paper we study two different settings: the average case setting and the probabilistic case setting. We begin with the average case setting.

In the average case setting the error of φ is determined by the average behavior of the error $E(Sf-\varphi(N^a f))$. More precisely, let μ be a Gaussian measure defined on $\mathcal{B}(F_1)$. Then the average error of φ is defined by

$$e^{avg}(\varphi,N^a) = \int_{F_1} E(Sf - \varphi(N^a f))\mu(df) \tag{2.6}$$

and an optimal algorithm φ^* that uses N^a is defined by

$$e^{avg}(\varphi^*,N^a) = r^{avg}(N^a) \stackrel{df}{=} \inf_{\varphi} e^{avg}(\varphi,N^a). \tag{2.7}$$

This means that in the average case setting we are interested in algorithms φ^*, if they exist, whose average error are minimal. In Section 4 we find φ^* for every nonadaptive information N and for adaptive N^a we prove that $r^{avg}(N^a) \asymp r^{avg}(N^{non}_{y^*})$ for some y^*.

We now turn to the probabilistic case setting. In this setting the goodness of φ is measured by the probability of success, i.e., by the probability that the error $E(Sf - \varphi(N^a f))$ of φ is small. More precisely, given $\varepsilon \geq 0$, let

$$\text{Prob}(\varphi, N^a, \varepsilon) = \mu(\{f \in F_1 : E(Sf - \varphi(N^a f)) \leq \varepsilon\}), \tag{2.8}$$

where μ is a Gaussian measure defined on $\mathcal{B}(F_1)$. Then by an optimal algorithm that uses N^a we mean an algorithm φ^* so that

$$\text{Prob}(\varphi^*, N^a, \varepsilon) = \text{Prob}(N^a, \varepsilon) \stackrel{df}{=} \sup_{\varphi} \text{Prob}(\varphi, N^a, \varepsilon). \tag{2.9}$$

In Section 5 we find φ^* for every nonadaptive N^{non}. For adaptive N^a we prove that $\text{Prob}(N^a, \varepsilon) \leq \text{Prob}(N^{non}_{y^*}, \varepsilon)$ for some y^*.

We comment on the definitions (2.6) and (2.8). In order for (2.6) and/or (2.8) to be well defined, $E(S(\cdot) - \varphi(N^a(\cdot)))$, as a function of f, should be measurable.

It is shown in [6] that this assumption is not restrictive since it is possible to extend the definitions (2.6) and (2.8) for every algorithm and prove that for optimal algorithms φ^*, $E(S(\cdot)-\varphi^*(N^a(\cdot)))$ is measurable.

We now recall some basic properties of Gaussian measures. By a Gaussian measure on $\mathcal{B}(F_1)$ we mean a measure μ such that

$$(2.10) \qquad \int_{F_1} \exp\{i(f,x)\}\mu(df) = \exp\{i(a,x)-\tfrac{1}{2}(Ax,x)\}, \quad \forall x \in F_1, \ (i = \sqrt{-1}),$$

where $A: F_1 \to F_1$ is a self-adjoint nonnegative definite operator with finite trace and a is an element of F_1. (The left hand side of (2.10) is called the characteristic functional of μ and is denoted by $\psi_\mu(x)$.) Then the mean element m_μ of μ is given by

$$(2.11) \qquad m_\mu = a$$

and the correlation operator S_μ of μ, by

$$(2.12) \qquad S_\mu = A$$

(see [2,3,4]). Recall that for an arbitrary measure μ, its mean element m_μ is defined by

$$(2.13) \qquad (m_\mu, x) = \int_{F_1} (f,x)\mu(df), \qquad \forall x \in F_1,$$

and its correlation operator S_μ by

$$(2.14) \qquad (S_\mu g,h) = \int_{F_1} (f-m_\mu,g)(f-m_\mu,h)\mu(df),$$

$$\forall g,h \in F_1.$$

Throughout the rest of this paper we shall assume without loss of generality that the mean element m_μ of μ is zero. $m_\mu = 0$, and that the correlation operator S_μ is positive definite.

3. Conditional measure.

In this section we exhibit an important property of the conditional measure for adaptive information. This property will be extensively used in the next sections. We begin with the definition of conditional measure (see [3]).

For an adaptive information operator N^a, let $\mu_1(\cdot, N^a)$ be the probability measure on $\mathbf{B}(\mathbf{R}^n)$ induced by N^a, i.e.,

$$\mu_1(\cdot, A) = \mu((N^a)^{-1}(A)) = \mu(\{f \in F_1 : N^a(f) \in A\}), \quad \forall A \in \mathbf{B}(\mathbf{R}^n). \tag{3.1}$$

Let $\mu_2(\cdot \mid y, N^a), y \in \mathbf{R}^n$, be a family of probability measures on $\mathbf{B}(F_1)$ such that

$$\mu_2(F_1 \mid y, N^a) = \mu_2((N^a)^{-1}(\{y\}) \mid y, N^a) = 1, \quad \text{for almost every } y \in \mathbf{R}^n, \tag{3.2}$$

$$\mu_2(B \mid \cdot, N^a), \text{ as a function of } y, \text{ is } \mu_1(\cdot \mid N^a)\text{-measurable}, \quad \forall B \in \mathbf{B}(F_1), \tag{3.3}$$

and

$$\mu(B) = \int_{\mathbf{R}^n} \mu_2(B \mid y, N^a)\mu_1(dy, N^a), \quad \forall B \in \mathbf{B}(F_1). \tag{3.4}$$

The family $\mu_2(\cdot|y,N^a)$ is called the conditional measure with respect to N^a and y. The existence and uniqueness of μ_2 follows from [3, Th. 8.1].

Let now G be a measurable function, $G: F_1 \to \mathbf{R}_+$. Then

$$(3.5) \qquad \int_{F_1} G(f)\mu(df) = \int_{\mathbf{R}^n}\Big[\int_{V(N^a,y)} G(f)\mu_2(df|y,N^a)\Big]\mu_1(dy,N^a),$$

where $V(N^a,y) = (N^a)^{-1}(\{y\}) = \{f \in F_1: N^a(f) = y\}$ is the set of elements f from F_1 which share the same information, $Nf = y$. The essence of (3.5) is that we first integrate G over all f with fixed information value y, and then over all values y from $\mathbf{R}^n$.

Recall that

$$(3.6) \qquad N^a(f) = [(f,g_1),(f,g_2(y_1)),\ldots,(f,g_n(y_1,\ldots,y_{n-1})],$$

$$y_i = y_i(f) = (f,g_i(y_1,\ldots,y_{i-1})).$$

For brevity we write $g_i(y) = g_i(y_1,\ldots,y_{i-1})$. Without loss of generality we assume that

$$(3.7) \qquad (S_\mu g_i(y), g_j(y)) = \delta_{ij}, \quad \forall y \in \mathbf{R}^n.$$

Let for a fixed $y = [y_1,\ldots,y_n] \in \mathbf{R}^n$,

(3.8) $$m(N^a,y) = \Sigma_{j=1}^{n} y_j S_\mu g_j(y)$$

and

(3.9) $$\sigma_{N^a,y}(\cdot) = \Sigma_{j=1}^{n}(\cdot,g_j(y))S_\mu g_j(y).$$

Then $\sigma_{N^a,y} : F_1 \to F_1$ is linear and $m(N^a,y) = \sigma_{N^a,y}(g)$, for every $g \in V(N^a,y)$, and for every fixed $y \in \mathbb{R}^n$. Of course, $m(N^a,y(f))$ and $\sigma_{N^a,y(f)}$, $y(f) = N^a(f)$, need not be linear in f.

Theorem 3.1: Let N^a be an arbitrary information operator of the form (3.7).

(i) Then the induced probability measure

(3.10) $$\mu_1(\cdot,N^a) = \mu_1(\cdot),$$

where μ_1 is the Gaussian measure on $\mathcal{B}(\mathbb{R}^n)$ with mean element zero and correlation operator identity, i.e.,

(3.11) $$\mu_1(A) = \frac{1}{\sqrt{(2\pi)^n}} \int_A \exp\{-\tfrac{1}{2}(x,x)\}d_n x.$$

(ii) The conditional measure $\mu_2(\cdot\,|y,N^a)$ is the Gaussian measure on $\mathcal{B}(F_1)$ with mean element $m(N^a,y)$ given by (3.6) and correlation operator

(3.12) $$S_{N^a,y} = (I - \sigma_{N^a y})S_\mu(I-\sigma^*_{N^a,y}).$$

Proof: It is shown in [8, Th. 4.2] that there exists a probability measure μ_1 on $\mathcal{B}(\mathbb{R}^n)$ such that $\mu_1(\cdot,N^a) = \mu_1(\cdot)$ for every N^a of the form (3.7). It was shown in [6, Th. 4.2 (i)] that for every nonadaptive N^{non} of the form (3.7), $\mu_1(\cdot,N^{non})$ is the Gaussian measure on $\mathcal{B}(\mathbb{R}^n)$ with mean element zero and correlation operator identity. Since $\mu_1(\cdot,N^a) = \mu_1(\cdot,N^{non}) = \mu_1$, the proof of Theorem 3.1 (i) is completed.

To prove (ii), let $\lambda_2(\cdot|y,N^a)$ be the Gaussian measure on $\mathcal{B}(F_1)$ with mean element $m(N^a,y)$ given by (3.8) and correlation operator $S_{N^a,y}$ given by (3.12). Then, due to (2.10)

$$\int_{F_1} e^{i(f,x)} \lambda_2(df|y,N^a)$$

$$= \exp\{i(m(N^a,y),x) - \frac{1}{2}(S_{N^a,y}x,x)\}.$$

Since $\sigma^*_{N^a,y}(x) = \Sigma_{j=1}^n (x,S_\mu g_j(y))g_j(y)$, we have

$$(S_{N^a,y}x,x) = (S_\mu(x-\sigma^*_{N^a,y}(x)),x-\sigma^*_{N^a,y}(x))$$

$$= (S_\mu x,x) - 2(S_\mu x,\sigma^*_{N^a,y}(x)) + (S_\mu \sigma^*_{N^a,y}(x),\sigma_{N^a,y}(x))$$

$$= (S_\mu x,x) - 2\,\Sigma_{j=1}^n (S_\mu x,g_j(y))^2 + \Sigma_{j=1}^n (S_\mu x,g_j(y))^2$$

$$= (S_\mu x,x) - \Sigma_{j=1}^n (S_\mu x,g_j(y))^2.$$

Hence

$$(3.13)\qquad \int_{F_1} e^{i(f,x)}\,\lambda_2(df|y,N^a) = \exp\{-\tfrac{1}{2}(S_\mu x,x)\}H(x,y),$$

where

$$H(x,y) = \exp\{i(m(N^a,y),x) + \tfrac{1}{2}\,\Sigma_{j=1}^{n}(S_\mu x,g_j(y))^2\}.$$

Due to (3.8)

$$H(x,y) = \exp\{\Sigma_{j=1}^{n}\{iy_j(S_\mu x,g_j(y))+\tfrac{1}{2}(S_\mu x,g_j(y))^2\}$$

$$= \Pi_{j=1}^{n}\,\exp\{iy_j(S_\mu x,g_j(y))+\tfrac{1}{2}(S_\mu x,g_j(y))^2\}.$$

Recall that $g_j(y) = g_j(y_1,\ldots,y_{j-1})$, and that μ_1 is the Gaussian measure. Hence

$$a \overset{df}{=} \int_{\mathbb{R}^n} H(x,y)\mu_1(dy)$$

$$= \frac{1}{\sqrt{(2\pi)^n}} \int_{\mathbb{R}}\cdots\int_{\mathbb{R}}\Pi_{j=1}^{n}\,\exp\{iy_j(S_\mu x,g_j(y_1,\ldots,y_{j-1}))$$

$$+ \tfrac{1}{2}(S_\mu x,g_j(y_1,\ldots,y_{j-1}))\}\exp\{-\tfrac{1}{2}\Sigma_{i=1}^{n}y_j^2\}d[y_1,\ldots,y_n].$$

Observe that

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp\{iy_j(S_\mu x,g_j(y_1,\ldots,y_{j-1}))$$

$$+ \tfrac{1}{2}(S_\mu x,g_j(y_1,\ldots,y_{j-1}))\}\exp\{-\tfrac{1}{2}y_j^2\}dy_j$$

$$= \exp\{\tfrac{1}{2}(S_\mu x,g_j(y_1,\ldots,y_{i-1}))\}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}$$

$$\exp\{iy_j(S_\mu x,g_i(y_1,\ldots,y_{i-1}))\}\exp\{-\tfrac{1}{2}y_j^2\}dy_j$$

$$= \exp\{\tfrac{1}{2}(S_\mu x, g_j(y_1,\ldots,y_{j-n}))\}\exp\{-\tfrac{1}{2}(S_\mu x, g_j(y_1,\ldots,y_{j-n}))\}$$

$$= 1.$$

This yields that $a = 1$ and

$$\int_{\mathbb{R}^n}\int_{F_1} e^{i(f,x)}\lambda_2(df|y,N^a) = \exp\{-\tfrac{1}{2}(S_\mu x,x)\} = \psi_\mu(x),$$

where ψ_μ is the characteristic functional of μ. Since characteristic functional defines measure uniquely and since conditional measure is determined uniquely (up to a set of μ_1-measure zero), $\mu_2(\cdot|y,N^a) = \lambda_2(\cdot|y,N^a)$, $\forall y \in \mathbb{R}^a$, a.e. This proves the theorem. ∎

Theorem 3.1 states that the induced measure $\mu_1(\cdot,N^a)$ does not depend on N^a, it only depends on $n = \mathrm{card}(N^a)$. From (ii) we can easily conclude that for $y \in \mathbb{R}^n$, the conditional measure $\mu_2(\cdot|y,N^a)$ is the same as the conditional measure for the nonadaptive information operator N_y^{non},

$$(3.14) \qquad \mu_2(\cdot|y,N^a) = \mu_2(\cdot|y,N_y^{non}).$$

Furthermore, $\mu_2(\cdot|y,N^a)$ is a translated measure $\mu_2(\cdot|0,N_y^{non})$, i.e.,

$$(3.15) \qquad \mu_2(B|y,N^a) = \mu_2(B-m(N^a,y)|0,N_y^{non}), \quad \forall B \in B(F_1).$$

In particular, if N^{non} is nonadaptive then

$$(3.16) \qquad \mu_2(B|y,N^{non}) = \mu_2(B-m(N^{non},y)|0,N^{non}), \quad \forall B \in B(F_1).$$

We end this section by two lemmas whose proofs, because of their length, are presented in the Appendix.

Lemma 3.1: For every Gaussian measure λ with mean element zero and for every balanced and convex set B,

$$(3.17) \qquad \lambda(B) \geq \lambda(B+h), \quad \forall h \in F_1. \qquad \blacksquare$$

Lemma 3.2: Let λ_1, λ_2 be two Gaussian measures on a separable Hilbert space with mean elements zero and correlation operators S_{λ_1} and S_{λ_2} respectively. Let $\alpha_{1,i}, \alpha_{2,i}, \ldots, (\alpha_{j,i} \geq \alpha_{j+1,i})$ be the eigenvalues of operators S_{λ_i}, $i = 1,2$. If

$$\alpha_{j,1} \leq \alpha_{j,2}, \quad \forall j = 1,2,\ldots$$

then

$$\lambda_1(J(0,\epsilon)) \geq \lambda_2(J(0,\epsilon)), \quad \forall \epsilon \geq 0,$$

where $J(0,\epsilon)$ stands for the ball with center zero and radius ϵ. $\blacksquare$

4. Spline algorithm and adaptive information on the average.

In this section we prove that for every error functional E and for every nonadaptive information, a translated spline algorithm is optimal. We also prove that for every adaptive information N^a there exists nonadaptive information of the same cardinality and whose radius is not greater than the radius of N^a.

Let N^a and φ be given. Recall that the (global) average error of φ is defined by

$$(4.1) \qquad e^{avg}(\varphi, N^a) = \int_{F_1} E(Sf - \varphi(N^a f))\mu(df)$$

and the (global) average radius of N^a, by

$$(4.2) \qquad r^{avg}(N^a) = \inf_{\varphi} e^{avg}(N^a).$$

Hence the global average radius of N^a is the minimal global average error made by any algorithm φ that uses N^a, and the optimal algorithm φ^* that uses N^a is defined so that its error is minimal, i.e.,

$$(4.3) \qquad e^{avg}(\varphi^*, N^a) = r^{avg}(N^a).$$

We now define the concept of the local average error as studied in [6]. Due to (3.5) and Theorem 3.1(i)

$$(4.4) \qquad e^{avg}(\varphi,N^a) = \int_{\mathbb{R}^n} e^{avg}(\varphi,N^a,y)\mu_1(dy),$$

where the local average error $e^{avg}(\varphi,N^a,y)$ is given by

$$(4.5) \qquad e^{avg}(\varphi,N^a,y) = \int E(Sf-\varphi(y))\mu_2(df|y,N^a).$$

Theorem 4.1: For every nonadaptive information N^{non} of the form (3.7) the average radius is given by

$$(4.6) \qquad r^{avg}(N^{non}) = \inf_{g\in F_2} \int_{F_1} E(Sf-g)\mu_2(df|0,N^{non}).$$

Let

$$(4.7) \qquad P = \{g^* \in F_2 : \int_{F_1} E(Sf-g^*)\mu_2(df|0,N^{non}) = r^{avg}(N^{non})\}.$$

An algorithm φ^* that uses N^{non} is optimal iff

$$(4.8) \qquad g(y) \overset{df}{=} \varphi^*(y) - Sm(N^{non},y) \in P, \text{ for almost every } y \in \mathbb{R}^n. \quad \blacksquare$$

Proof: Let φ be an algorithm that uses N^{non}. Consider the local error $e^{avg}(\varphi,N^{non},y)$. Due to (3.16) and linearity of S,

$$e^{avg}(\varphi,N^{non},y) = \int_{F_1} E(S(f+m(N^{non},y)) - \varphi(y))\mu_2(df|0,N^{non})$$

$$= \int_{F_1} E(Sf-(\varphi(y)-Sm(N^{non},y)))\mu_2(df|0,N^{non})$$

$$\geq \inf_{g \in F_2} \int_{F_1} E(Sf-g)\mu_2(df|0,N^{non}) \stackrel{df}{=} H.$$

This proves that

$$r^{avg}(N^{non}) \geq H.$$

To prove that $r^{avg}(N^{non}) = H$ we can assume that H is finite. Then for every $\delta > 0$, there exists $g_\sigma \in F_2$ such that $\int_{F_1} E(Sf-g_\sigma)\mu_2(df|0,N^{non}) \leq H + \delta$. Define $\varphi_\delta(y) = Sm(N^{non},y) + g_\sigma$. Then

$$e^{avg}(\varphi_\delta,N^{non}) = \int_{\mathbb{R}^n} e^{avg}(\varphi_\delta,N^{non},y)\mu_1(dy) \leq H-\delta.$$

Since δ is arbitrary, $r^{avg}(N^{non}) \leq H$ and consequently $r^{avg}(N^{non}) = H$. This proves (4.6). To complete the proof observe that if $H = +\infty$ then every algorithm is optimal and $P = F_2$. Therefore we can assume that $H < +\infty$. If $\varphi^*(y) = Sm(N^{non},y) + g^*(y)$ with $g^*(y) \in P$ for almost every y, then, obviously, $\varphi^*(y) = Sm(N^{non},y) + g^*(y)$ with $g^*(y) + P$ for almost every y, then, obviously, φ^* is optimal. On the other hand, take an arbitrary algorithm φ. Define

$$Y = \{y \in \mathbb{R}^n: g(y) = \varphi(y)-Sm(N^{non},y) \notin P\}.$$

If Y has a positive μ_1 measure, then

$$e^{avg}(\varphi, N^{non}) = \int_Y \int_{F_1} E(Sf - g(y)) \mu_2(df|0, N^{non}) \mu_1(dy)$$

$$+ \int_{\mathbb{R}^n \setminus Y} \int_{F_1} E(Sf - g(y)) \mu_2(df|0, N^{non}) \mu_1(dy)$$

$$> \mu_1(Y) r^{avg}(N^{non}) + \mu_1(\mathbb{R}^n \setminus Y) r^{avg}(N^{non})$$

$$= r^{avg}(N^{non}).$$

Hence φ is not optimal. This completes the proof of Theorem 4.1. ■

Theorem 4.1 states that there exists an optimal algorithm iff the infinum in (4.8) is attained by some element g*. Of course g* need not be unique, but taking any g* satisfying (4.8), the algorithm

$$\varphi^*(\cdot) = \varphi^s(\cdot) + g^*$$

is optimal where

$$\varphi^s(N^{non} f) = \Sigma_{i=1}^n (f, g_i) SS_\mu g_i = Sm(N^{non}, N^{non} f).$$

The algorithm φ^s, called the spline algorithm, is linear. Hence φ^* is an affine mapping, which is a desirable property from the complexity point of view. On the other hand if the infinum in (4.8) is not attained, i.e., $P = \emptyset$, then there is no optimal algorithm. In this case taking g* so that $\int_{F_1} E(Sf - g^*) \mu_2(df|0, N^{non})$ is sufficiently

close to $r^{avg}(N^{non})$, say is not greater than $r^{avg}(N^{non}) + \delta$, the following affine algorithm

$$\varphi^*(\cdot) = \varphi^s(\cdot) + g^*$$

is almost optimal, since

$$e^{avg}(\varphi^*, N^{non}) \le r^{avg}(N^{non}) + \delta.$$

We now prove that adaption does not help on the average. Let N^a be adaptive information of the form (3.7), and let

$$H(y) = \inf_{g \in F_2} \int_{F_2} E(Sf-g)\mu_2(df|0, N_y^{non}).$$

Then, due to (3.15) and (4.6),

$$(4.9) \qquad r^{avg}(N^a) = \int_{\mathbb{R}^n} H(y)\mu_1(dy) = \int_{\mathbb{R}^n} r^{avg}(N_y^{non})^2 \mu_1(dy).$$

Let $y^*, y^* \in \mathbb{R}^n$, be such that

$$(4.10) \qquad r^{avg}(N_{y^*}^{non}) \le r^{avg}(N^a).$$

Observe that such y^* exists. Indeed, $r^{avg}(N_y^{non}) > r^{avg}(N^a)$ for every y would contradict (4.9). Hence we have proven

Theorem 4.2: For every adaptive information N^a there exists $y^* \in \mathbb{R}^n$ such that

$$r^{avg}(N_{y^*}^{non}) \le r^{avg}(N^a). \qquad \blacksquare$$

We now give a sufficient condition on the error

functional E so that the spline algorithm φ^s is optimal. Technically, this means that $0 \in P$.

Theorem 4.3: If E is convex and symmetric (with respect to zero) then for every nonadaptive information N^{non} the spline algorithm φ^s is optimal. ∎

Proof: Although Theorem 4.3 follows immediately from [6], we present its proof for the completeness. Take $g \in F_2$. Then, due to the symmetricity of $\mu_2(\cdot) = \mu_2(\cdot \mid 0, N^{non})$ (i.e., $\mu_2(B) = \mu_2(-B)$, $\forall B \in \mathcal{B}(F_1)$),

$$\int_{F_1} E(Sf-g)\mu_2(df) = \frac{1}{2}\int_{F_1} [E(Sf-g) + E(-Sf-g)]\mu_2(df).$$

Since E is symmetric and convex,

$$\frac{1}{2}[E(Sf-g)+E(Sf-g)] = \frac{1}{2}[E(Sf-g)+E(Sf+g)] \geq E(Sf).$$

Hence

$$\int_{F_1} E(Sf-g)\mu_2(df) \geq \int_{F_1} E(Sf)\mu_2(df), \quad \forall g \in F_2.$$

This proves that $g^* = 0 \in P$ and completes the proof of Theorem 4.2. ∎

Remark 4.1: Optimality of the spline algorithm on the average has been established in [7,8] for orthogonally invariant measures assuming that F_2 is a separable Hilbert space and $E(g) = \|g\|^2$. The same result was obtained in

[5] assuming that F_1 is a finite dimensional space and $E(Sf-\varphi(N^a f)) = \|Sf-\varphi(N^a f)\|^2 \rho((S_\mu f, f))$ for some function ρ.

In this paper, restrictiong the class of probability measures to Gaussian measures we relax the assumptions concerning E and the spaces F_1 and F_2.

We now exhibit an n-th optimal information operator N^* of $\mathrm{card}(N^*) = n$, i.e., N^* satisfying

$$r^{avg}(N^*) \le r^{avg}(N^a), \quad \forall N^a, \ \mathrm{card}(N^a) = n.$$

We find N^* under some additional assumptions on F_2, S and E. Namely, we assume that F_2 is a separable Hilbert space, S is continuous and

$$E(g) = H(\|g\|) \tag{4.11}$$

for some function $H: \mathbb{R}_+ \to \mathbb{R}_+$ which is convex and nondecreasing. Observe that then E is convex and symmetric and therefore the spline algorithm is optimal for every nonadaptive information.

To find N^* we proceed as follows. Let

$$R = SS_\mu S^*: F_2 \to F_2. \tag{4.12}$$

Since S is continuous, R is a nonnegative definite

operator with finite trace. Let $\zeta_1^*, \zeta_2^*, \ldots$ be eigenelements of R corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, i.e.,

$$R\zeta_i^* = \lambda_i \zeta_i^*, \quad (\zeta_i^*, \zeta_j^*) = \delta_{i,j}.$$

Take

$$g_i^* = \frac{1}{\sqrt{\lambda_i}} S^* \zeta_i^*, \quad i = 1,2,\ldots . \tag{4.13}$$

Remark 4.2: The optimal information for the average case setting studied in [7] is derived from the operator K defined by

$$K = S_\mu^{1/2} S^* S S_\mu^{1/2} : F_1 \to F_1.$$

Observe that if η is an eigenvector of K corresponding to an eigenvalue β, $K\eta = \beta\eta$, then letting

$$\zeta = S S_\mu^{1/2} \eta$$

we get

$$R\zeta = S S_\mu S^* S S_\mu^{1/2} \eta = S S_\mu^{1/2} K\eta = \beta S S_\mu^{1/2} \eta = \beta\zeta.$$

Hence the operators K and R have the same eigenvalues. Furthermore η is an eigenvector of K iff $S S_\mu^{1/2}$ is an eigenvector of R. ∎

Define the nonadaptive information operator

$$N^*(f) = [(f,g_1^*),\ldots,(f,g_n^*)]. \tag{4.14}$$

Note that N^* satisfies (3.7).

Theorem 4.4: The information operator N^* defined by (4.14) is nth optimal.

Proof: Due to Theorem 4.2, we need only to prove that

$$r^{avg}(N^*) \leq r^{avg}(N^{non}),$$

for every N^{non} of the form (3.7). Due to Theorems 4.1 and 4.3,

$$r^{avg}(N^{non}) = \int_{F_1} H(\|Sf\|)\mu_2(df|0,N^{non}).$$

If H is constant, $H(x) \equiv c$, then $r^{avg}(N^{non}) = r^{avg}(N^*) = c$ for every N^{non}. Hence without loss of generality we can assume that H is not constant. Then $H(\mathbb{R}_+) = [H(0),+\infty)$. Indeed, convexity of H yields $2H(x) \leq H(0) + H(x)$, $\forall x \in \mathbb{R}_+$. Since H is nondecreasing, $\sup\{H(x):x \in \mathbb{R}_+\}$ $= \lim_{x\to\infty} H(x) = c$. Note that $H(0) < c$. If $c < +\infty$ then $2c \leq H(0) + c < 2c$ which is a contradiction. Hence $H(\mathbb{R}_+) = [H(0),+\infty)$ as claimed. Define

$$\gamma(B,N^{non}) = \mu_2(\{f \in F_1: H(\|Sf\|) \in B\}|0,N^{non}), \tag{4.15}$$

$$\forall B \in \mathcal{B}(H(\mathbb{R}_+)).$$

Then $\gamma(\cdot,N^{non})$ is a probability measure on $\mathcal{B}(H(\mathbb{R}_+))$ and

$$(4.16) \qquad r^{avg}(N^{non}) = \int_{H(0)}^{+\infty} t\gamma(dt,N^{non}).$$

Let $D(\cdot,N^{non})$ be the distribution function for $\gamma(\cdot,N^{non})$, i.e.,

$$(4.17) \qquad D(x,N^{non}) = \int_{H(0)}^{x} \gamma(dt,N^{non}), \quad \forall x \in H(\mathbb{R}_+).$$

We shall prove that

$$(4.18) \qquad D(x,N^{non}) \leq D(x,N^*), \quad \forall x \in H(\mathbb{R}_+), \quad \forall N^{non}.$$

Before presenting the proof of (4.18), we show that (4.18) will complete the proof of Theorem 4.4. For this end, observe that

$$t = \lim_k \Sigma_{i=1}^{k} t_{i,k}\chi_{(a_{i,k},a_{i+1,k}]}(t), \quad \forall t > H(0),$$

for some numbers $H(0) = a_{1,k} < a_{2,k} < \ldots < a_{k,k} < a_{k+1,k} = +\infty$ and $t_i \in (a_{i,k},a_{i+1,k}]$. Hence for every N^{non},

$$\begin{aligned} r^{avg}(N^{non}) &= \lim_k \Sigma_{i=1}^{k} t_{i,k}\gamma((a_{i,k},a_{i+1,k}],N^{non}) \\ &= \lim_k \Sigma_{i=1}^{k} t_{i,k}[D(a_{i+1,k},N^{non}) - D(a_{i,k},N^{non})] \\ &= \lim_k [\Sigma_{i=1}^{k-1}(t_{i,k}-t_{i+1,k})D(a_{i+1,k},N^{non}) + t_{k,k}] \end{aligned}$$

since $D(a_{1,k},N^{non}) = 0$ and $D(a_{k+1,k},N^{non}) = D(+\infty,N^{non}) = 1$. Hence,

$$r^{avg}(N^{non}) - r^{avg}(N^*)$$

$$= \lim_k \Sigma_{i=1}^{k} (t_{i,k} - t_{i+1,k}) [D(a_{i+1,k}, N^{non}) - D(a_{i+1,k}, N^*)]$$

and $t_i - t_{i+1} < 0$ imply that

$$r^{avg}(N^{non}) - r^{avg}(N^*) \geq 0.$$

Hence to complete the proof of Theorem 4.4 it is enough to show that (4.18) holds. Observe that

$$(4.19) \quad D(x, N^{non}) = \mu_2(\{f: H(\|Sf\|) \leq x\} | 0, N^{non})$$

$$= \mu_2(\{f: \|Sf\| \leq H^{-1}(x)\} | 0, N^{non}).$$

Define

$$\lambda(B, N^{non}) = \mu_2(\{f \in F_1: Sf \in B\} | 0, N^{non}),$$

$$\forall B \in \mathcal{B}(F_2).$$

Then $\lambda(\cdot, N^{non})$ is a probability measure on $\mathcal{B}(F_2)$ and, due to (4.19),

$$(4.20) \quad D(x, N^{non}) = \lambda(J(0,z), N^{non}),$$

where now $z = H^{-1}(x)$ and $J(0,z)$ is the ball in F_2 with center zero and radius z. We need the following two lemmas.

Lemma 4.1: For every N^{non}, $\lambda(\cdot, N^{non})$ is the Gaussian measure with mean element zero and correlation operator

$$R_{N^{non}} = S(I-\sigma_{N^{non}})S_{\mu}(I-\sigma^*_{N^{non}})S^*. \quad \blacksquare$$

Proof: Observe that for the characteristic functional $\Psi_{N^{non}}$ of $\lambda(\cdot,N^{non})$ we have

$$\Psi_{N^{non}}(h) = \int_{F_2} \exp\{i(g,h)\}\lambda(dg,N^{non}).$$

Change variables by setting $f = Sg$. Then

$$\begin{aligned}\Psi_{N^{non}}(y) &= \int_{F_1} \exp\{i(f,S^*h)\}\mu_2(df|0,N^{non}) \\ &= \exp\{-\tfrac{1}{2}((I-\sigma_{N^{non}})S_{\mu}(I-\sigma^*_{N^{non}})S^*h,S^*h)\} \\ &= \exp\{-\tfrac{1}{2}(R_{N^{non}}h,h)\}, \qquad \forall h \in F_2.\end{aligned}$$

This completes the proof of Lemma 4.1. $\blacksquare$

Let $\gamma_1,\gamma_2,\cdots$ $(\gamma_i \geq \gamma_{i+1} \geq 0)$ be the eigenvalues of $R_{N^{non}}$. It is easy to check that for N^*, $\lambda_{n+1},\lambda_{n+2},\cdots$ are the dominating eigenvalues of R_{N^*}.

Lemma 4.2:

(4.21) $\quad \lambda_{n+k} \leq \gamma_k, \quad \forall k = 1,2,\ldots . \quad \blacksquare$

Proof (induction on k): For $k = 1$, (4.21) holds trivially. Suppose therefore that (4.21) holds for every $k \leq k_0$. We prove that (4.21) also holds for $k = k_0+1$.

For this end, let $\eta_1,\eta_2,\ldots,\eta_k$ be eigenelements of

$K_{N^{non}}$ corresponding to $\gamma_1, \gamma_2, \ldots, \gamma_k$. Take

$$g = \Sigma_{i=1}^{n+k} x_i \zeta_i^* \in F_2$$

Such that

$$(4.22) \qquad \|g\|^2 = \Sigma_{i=1}^{n+k} x_i^2 = 1,$$

$$(4.23) \qquad \sigma^*_{N^{non}}(S^*g) = 0,$$

and

$$(4.24) \qquad (g, \eta_i) = 0, \quad i = 1,2,\ldots,k_0 = k-1.$$

Since (4.23) and (4.24) are equivalent to a homogeneous system of n+k-1 linear equations with n + k unknowns, such g exists. Furthermore, (4.22) and (4.24) yield that $\gamma_k \geq (R_{N^{non}} g, g)$. Hence, due to (4.23), we get

$$\gamma_k \geq (R_{N^{non}} g, g) = (Rg, g) = \Sigma_{i=1}^{n+k} \lambda_i x_i^2 \geq \lambda_{n+k} \Sigma_{i+1}^{n+k} x_i^2$$

$$= \lambda_{n+k},$$

which completes the proof of Lemma 4.2. ∎

We are ready to complete the proof of Theorem 4.4. Due to Lemmas 4.1, 4.2 and 3.2,

$$\lambda(J(0,z), N^*) \geq \lambda(J(0,z), N^{non}), \quad \forall N^{non}, \quad \forall z \in R_+.$$

Hence (4.20) yields that

$$D(x,N^*) \geq D(x,N^{non}), \quad \forall\, N^{non}, \forall\, x \in \mathbb{R}_+.$$

This completes the proof of (4.18) as well as the proof of Theorem 4.4. ∎

5. Spline algorithm and adaptive information in the probabilistic setting.

In this section we prove that for every error functional E and for every nonadaptive information N, the probability of the fact that the error does not exceed ε, is maximized by a translated spline algorithm. We also prove that adaption does not help in this setting.

Recall that for given $\varepsilon \geq 0$, N^a and φ,

$$\text{(5.1)} \qquad \text{Prob}(\varphi, N^a, \varepsilon) = \mu(\{f \in F_1 : E(Sf - \varphi(N^a f)) \leq \varepsilon\})$$

is the probability of the fact that the error $E(Sf-\varphi(N^a f))$ made by φ is not greater than ε, and

$$\text{(5.2)} \qquad \text{Prob}(N^a, \varepsilon) = \inf_{\varphi} \text{Prob}(\varphi, N^a, \varepsilon).$$

Then $\text{Prob}(N^a, \varepsilon)$ is the maximal probability among all algorithms that use N^a, and the optimal algorithm φ^* that uses N^a is defined so that

$$\text{(5.3)} \qquad \text{Prob}(\varphi^*, N^a, \varepsilon) = \text{Prob}(N^a, \varepsilon).$$

Theorem 5.1: For every nonadaptive information N^{non} of the form (3.7)

$$\text{(5.4)} \qquad \text{Prob}(N^{non}, \varepsilon) = \sup_{g \in F_2} \mu_2(\{f \in F_1 : E(Sf-g) \leq \varepsilon\} | 0, N^{non}).$$

Let

(5.5) $$P = \{g^* \in F_2 : \mu_2(\{f \in F_1 : E(Sf-g^*) \le \varepsilon\} | 0, N^{non}) = Prob(N^{non}, \varepsilon)\}.$$

An algorithm φ^* that uses N^{non} is optimal iff

(5.6) $$g(y) \stackrel{df}{=} \varphi^*(y) - Sm(N^{non}, y) \in P, \text{ for almost every } y \in \mathbb{R}^n.$$ ∎

Proof: The proof of this theorem differs from the proof of Theorem 4.1 only at the beginning. Observe that for every algorithm φ that uses N^{non} we have, due to (3.16) and linearly of S,

$$Prob(\varphi, N^{non}, \varepsilon) = \int_{\mathbb{R}^n} \mu_2(\{f \in F_1 : E(Sf-\varphi(y)) \le \varepsilon\} | y, N^{non}) \mu_1(dy)$$

$$= \int_{\mathbb{R}^n} \mu_2(\{f \in F_1 : E(Sf-(\varphi(y)-Sm(N^{non}, y))) \le \varepsilon\} | 0, N^{non}) \mu_1(dy)$$

$$\le \sup_{g \in F_2} \mu_2(\{f \in F_1 : E(Sf-g) \le \varepsilon\} | 0, N^{non}).$$

Hence using the same reasoning as in the proof of Theorem 4.1 one can easily complete the proof of Theorem 5.1. Therefore we skip this part. ∎

Let N^a be adaptive. Similar as in (4.10), let $y^*, y^* \in \mathbb{R}^n$, be such that

$$\text{(5.7)} \qquad \text{Prob}(N^a, \varepsilon) = \int_{\mathbb{R}^n} \text{Prob}(N_y^{non}, \varepsilon)\mu_1(dy) \le \text{Prob}(N_{y^*}^{non}, \varepsilon).$$

Of course, such y^* exists.

Theorem 5.2: For every adaptive information N^a there exists $y^* \in \mathbb{R}^n$ such that

$$\text{Prob}(N_{y^*}^{non}, \varepsilon) \le \text{Prob}(N^a, \varepsilon). \qquad \blacksquare$$

As in Section 4 we give a sufficient condition on E for the spline algorithm to be optimal, i.e., $g^* = 0 \in P$.

Theorem 5.3: If E is convex and symmetric (with respect to zero) and if $F_2 = S(F_1)$ then for every nonadaptive information N^{non} the spline algorithm φ^s is optimal. $\blacksquare$

Proof: To prove this theorem it is enough to show that

$$\text{(5.8)} \qquad \mu_2(\{f \in F_1 : E(Sf) \le \varepsilon\}) \ge \mu_2(\{f \in F_1 : E(Sf-g) \le \varepsilon\}), \quad \forall g \in F_2,$$

where $\mu_2(\cdot) = \mu_2(\cdot \mid 0, N^{non})$. Let

$$B(g) = \{f \in F_1 : E(Sf-g) \le \varepsilon\}$$

and

$$B = B(0) = \{f \in F_1 : E(Sf) \le \varepsilon\}.$$

Since $F_2 = S(F_1)$, there exists an element $h \in F_1$ such that $Sh = g$. Observe that

$$B(g) \subseteq B + h.$$

Indeed, for $f \in B(g)$ let $\tilde{f} = f-h$. Since $E(S\tilde{f}) = E(S(f-h)) = E(Sf-g) \leq \varepsilon$. Thus $\tilde{f} \in B$ and $f = \tilde{f} + h \in B + h$ as claimed. This means that

$$\mu_2(\{f \in F_1: E(Sf-g) \leq \varepsilon\}) = \mu_2(B(g)) \leq \mu_2(B + h).$$

Hence to prove (5.8) we need only to show that

$$\mu_2(B) \geq \mu_2(B + h), \quad \forall h \in F_1. \tag{5.9}$$

Observe that B is convex and balanced. Indeed, if $f_1, f_2 \in B$ then $E(tf_1+(1-t)f_2) \leq tE(f_1) + (1-t)E(f_2) \leq \varepsilon$, i.e., $tf_1 + (1-t)f_2 \in B$, and if $f \in B$ then $E(-f) = E(f) \leq \varepsilon$, i.e., $-f \in B$. Since μ_2 is a Gaussian measure with mean element zero, Lemma 3.1 completes the proof of Theorem 5.3. ∎

The next theorem is about <u>n-th optimal information</u> N^*. The information N^* of cardinality n is optimal iff

$$\text{Prob}(N^*, \varepsilon) \geq \text{Prob}(N^a, \varepsilon), \quad \forall N^a, \ \text{card}(N^a) = n.$$

<u>Theorem 5.4</u>: Let E be of the form (4.11) and let S be continuous. Then the information N^* defined by (4.14)

is n-th optimal for every $\varepsilon \geq 0$. ∎

Proof: This theorem follows immediately from (4.18). Indeed, due to Theorem 5.2, we need only to consider nonadaptive information N^{non}. But then for every N^{non} and every $\varepsilon \geq 0$,

$$\text{Prob}(N^{non}, \varepsilon) = \mu_2(\{f \in F_1 : H(\|Sf\|) \leq \varepsilon\}) = D(H^{-1}(\varepsilon), N^{non}).$$

Hence (4.18) implies that

$$\text{Prob}(N^*, \varepsilon) \geq \text{Prob}(N^{non}, \varepsilon), \quad \forall N^{non}, \quad \forall \varepsilon \geq 0,$$

which completes the proof. ∎

We end this section by the following problem. For a given set $A \in \mathcal{B}(\mathbb{R}^n)$ let

$$\text{Prob}(\varphi, N^a, \varepsilon, A) = \mu(\{f \in F_1 : E(Sf - \varphi(N^a f)) \leq \varepsilon \wedge N^a f \in A\}). \tag{5.10}$$

We want to find φ^* such that

$$\text{Prob}(N^a, \varepsilon, A) \overset{df}{=} \sup_{\varphi} \text{Prob}(\varphi, N^a, \varepsilon, A) = \text{Prob}(\varphi^*, N^a, \varepsilon, A). \tag{5.11}$$

Observe that $\text{Prob}(\varphi, N^a, \varepsilon, A)$ is the probability that $E(Sf - \varphi(N^a f)) \leq \varepsilon$ under the condition that $N^a(f) \in A$. Of

course, for $A = \mathbb{R}^n$, $\text{Prob}(\varphi, N^a, \varepsilon, A) = \text{Prob}(\varphi, N^a, \varepsilon)$.

For every adaptive information N^a,

$$\text{Prob}(\varphi, N^a, \varepsilon, A) = \int_A \mu_2(\{f \in F_1 : E(Sf - \varphi(y)) \le \varepsilon\} | y, N_y^{non}) \mu_1(dy)$$

$$\le \int_A \sup_{g \in F_2} \mu_2(df \in F_1 : E(Sf - g) \le \varepsilon | 0, N_y^{non}) \mu_1(dy)$$

$$= \int_A \text{Prob}(N_y^{non}, \varepsilon) \mu_1(dy).$$

From this we can conclude

Theorem 5.5:

(i) For every adaptive information N^a there exists $y^* \in \mathbb{R}^n$ such that

$$\text{Prob}(N^a, \varepsilon, A) \le \text{Prob}(N_{y^*}^{non}, \varepsilon, A), \quad \forall E, \quad \forall \varepsilon \ge 0, \forall A \in B(\mathbb{R}^n).$$

(ii) For every nonadaptive information N^{non}

$$\text{Prob}(N^{non}, \varepsilon, A) = \text{Prob}(N^{non}, \varepsilon) \mu_1(A), \quad \forall E, \forall \varepsilon \ge 0$$

$$\forall A \in B(\mathbb{R}^n).$$

In particular, φ^* is optimal independently of A.

(iii) If $F_2 = S(F_1)$ and E is convex and symmetric (with respect to zero) then the spline algorithm φ^s is optimal for every N^{non}, every $\varepsilon \ge 0$ and every $A \in B(\mathbb{R}^n)$.

(iv) If F_2 is a separable Hilbert space, S is continuous and E is of the form (4.11), then N^* defined

by (4.14) is optimal for every $\epsilon \geq 0$ and every $A \in B(\mathbb{R}^n)$. ∎

Theorem 5.5 states that the probability of a small error does not depend on the value $N^{non}f$ of information. This result will be used in a future paper for studying optimal stopping criteria.

5. Variance of spline algorithm.

In previous sections we showed when the spline algorithm φ^s is optimal. Here we exhibit another optimality property of φ^s showing that it minimizes the variance whenever F_2 is a Hilbert space and $E(g) = \|g\|^2$.

Let N^{non} be nonadaptive information and let φ be an algorithm that uses N^{non}. By the variance of φ we mean

$$(6.1) \qquad var(\varphi) = \int_{F_1} \{\|Sf-\varphi(N^{non}f)\|^2 - e^{avg}(\varphi,N^{non})\}^2 \mu(df),$$

where

$$e^{avg}(\varphi,N^{non}) = \int_{F_1} \|Sf-\varphi(N^{non}f)\|^2 \mu(df).$$

Theorem 6.1:

$$(6.2) \qquad var(\varphi^s) = \inf_{\varphi} var(\varphi). \qquad \blacksquare$$

Proof: Let φ be an algorithm. Define $h = \varphi - \varphi^s$, i.e., $\varphi(N^{non}f) = \varphi^s(N^{non}f) + h(N^{non}f)$. Then due to (3.18),

$$\begin{aligned} var(\varphi) &= \int_{\mathbb{R}^n}\int_{F_1} \{\|Sf-\varphi^s(y)-h(y)\|^2 \\ &\qquad - e^{avg}(\varphi,N^{non})\}^2 \mu_2(df|y,N^{non})\mu_1(dy) \\ &= \int_{\mathbb{R}^n}\int_{F_1} \{\|Sf-h(y)\|^2 - e^{avg}(\varphi,N^{non})\}^2 \mu_2(df)\mu_1(dy), \end{aligned}$$

where $\mu_2(\cdot) = \mu_2(\cdot|0,N^{non})$. Observe that $\|Sf-h(y)\|^2$ $= \|Sf\|^2 - 2(f,S^*h(y) + \|h(y)\|^2$. Since mean element of μ_2 is zero, $\int_{F_1}(f,S^*h(y))\mu_2(df) = 0$, and

$$e^{avg}(\phi,N^{non}) = \int_{R^n}\int_{F_1}[\|Sf\|^2-2(f,S^*h(y)) + \|h(y)\|^2]\mu_2(df)\mu_1(dy)$$

$$= \int_{R^n}\int_{F_1}[\|Sf\|^2+\|h(y)\|^2]\mu_2(df)\mu_1(dy)$$

$$= e^{avg}(\phi^s,N^{non}) + \int_{R^n}\|h(y)\|^2\mu_1(dy).$$

Hence

$$var(\phi) = \int_{R^n}\int_{F_1}[\|Sf\|^2-e^{avg}(\phi^s,N^{non})-2(f,S^*h(y))+\|h(y)\|^2 - \int_{R^n}\|h(z)\|^2\mu_1(dz)]^2\mu_2(df)\mu_1(dy).$$

Change the variables by letting f = -f. Then

$$var(\phi) = \frac{1}{2}\int_{R^n}\int_{F_1}[[\|Sf\|^2-e^{avg}(\phi^s,N^{non})-2(f,S^*h(y))+\|h(y)\|^2 - \int_{R^n}\|h(z)\|^2\mu_1(dz)]^2 + [\|Sf\|^2-e^{avg}(\phi^s,N^{non}) + 2(f,S^*h(y))+\|h(y)\|^2 - \int \|h(z)\|^2\mu_1(dz)]^2]\mu_2(df)\mu_1(dy)$$

$$\geq \int_{R^n}\int_{F_1}[\|Sf\|^2-e^{avg}(\phi^s,N^{non})+\|h(y)\|^2 - \int_{R^n}\|h(z)\|^2\mu_1(dz)]^2\mu_2(df)\mu_1(dy),$$

since $\frac{1}{2}[(a+b)^2 + (a-b)^2] \geq a^2$. Hence

$$\text{var}(\varphi) \geq \text{var}(\varphi^s) + 2H_1 + H_2, \tag{6.3}$$

where

$$H_1 = \int_{\mathbb{R}^n}\int_{F_1}\{\|Sf\|^2 - e^{\text{avg}}(\varphi^s, N^{\text{non}})\}\{\|h(y)\|^2 - \int_{\mathbb{R}^n}\|h(z)\|^2\mu_1(dz)\}^2\mu_2(df)\mu_1(dy)$$

and

$$H_2 = \int_{\mathbb{R}^n}\int_{F_1}[\|h(y)\|^2 - \int_{\mathbb{R}^n}\|h(z)\|^2\mu_1(dz)\}^2\mu_2(df)\mu_1(dy).$$

Of course, $H_2 \geq 0$ and therefore

$$\text{var}(\varphi) \geq \text{var}(\varphi^s) + 2H_1. \tag{6.4}$$

We now prove that $H_1 = 0$. Indeed,

$$H_1 = \int_{\mathbb{R}^n}[\{\|h(y)\|^2 - \int_{\mathbb{R}^n}\|h(z)\|^2\mu_1(dz)\}^2 \cdot \int_{F_1}\{\|Sf\|^2 - e^{\text{avg}}(\varphi^s, N^{\text{non}})\}\mu_2(df)]\mu_1(dy)$$

and since $e^{\text{avg}}(\varphi^s, N^{\text{non}}) = \int_{F_1}\|Sf\|^2\mu_2(df)$, see Theorem 4.3 and (4.8), $H_1 = 0$ as claimed. Hence

$$\text{var}(\varphi) \geq \text{var}(\varphi^s), \quad \forall\, \varphi,$$

which completes the proof. ■

We want to stress that the minimal variance of the

spline algorithm strongly depends on the form of E, i.e., $E(g) = \|g\|^2$. For arbitrary E (even convex and symmetric) the spline algorithm need not minimize the variance.

7. Appendix.

We prove Lemmas 3.1 and 3.2. Since these lemmas are well known for finite dimensional spaces, the proofs are mainly to show that the infinite dimensional case can be reduced to a finite dimensional one.

Proof of Lemma 3.1: We prove that (3.17) can be reduced to a problem with a finite dimensional Gaussian measure. Then the Anderson's inequality will complete the proof.

Let $\zeta_1, \zeta_2, \ldots$ be eigenvalues of the covariance operator S_λ, $S_\lambda \zeta_i = \alpha_i \zeta_i$ and $(\zeta_i, \zeta_j) = \delta_{ij}$. Let $X = \ker S_\lambda$ and let $X^\perp$ be the orthogonal complement of X, $F_1 = X^\perp \oplus X$. Then for every $f \in F_1$, $f = f_1 + f_2$, where $f_1 \in X$ and $f_2 \in X^\perp$, and for every $C \in \mathcal{B}(F_1)$

$$\lambda(C) = \lambda^\perp(C \cap X^\perp), \tag{A.1}$$

where $\lambda^\perp$ is the Gaussian measure on $\mathcal{B}(X^\perp)$ with mean element zero and covariance operator $S_{\lambda^\perp} = S_\lambda|_{X^\perp}$ (see [4]). Observe that $B \cap X^\perp$ is convex and balanced and that $(B + h) \cap X^\perp \subset (B \cap X^\perp) + h_2$ ($h = h_1 + h_2$, $h_1 \in X$ and $h_2 \in X^\perp$). Hence, due to (A.1),

$$\lambda(B) = \lambda^\perp(B \cap X^\perp) \text{ and } \lambda(B+h) = \lambda^\perp((B+h) \cap X^\perp) \leq \lambda^\perp((B \cap X^\perp) + h_2).$$

This means that to prove (3.17) we can assume without

loss of generality that $X^{\perp} = F_1$ and $\lambda = \lambda^{\perp}$, i.e., that all eigenvalues of S_λ are positive.

For $k = 1,2,\ldots$ define $P_k: F_1 \to \mathbb{R}^k$,

$$(A.2) \qquad P_k(f) = [(f,\frac{\zeta_1}{\sqrt{\alpha_1}}),\ldots,(f,\frac{\zeta_k}{\sqrt{\alpha_k}})].$$

Observe that for every set $C \in \mathcal{B}(F_1)$, $P_k^{-1}(P_k(C)) \supset P_{k+1}^{-1}(P_{k+1}(C))$ and $\bigcap_{k=1}^{\infty} P_k^{-1}(P_k(C)) = C$. Hence

$$(A.3) \qquad \lambda(C) = \lim_k \lambda(P_k^{-1}(P_k(C))), \qquad \forall C \in \mathcal{B}(F_1).$$

Let λ_k be the probability measure on $\mathcal{B}(\mathbb{R}^k)$ induced by P_k, i.e.,

$$\lambda_k(A) = \lambda(P_k^{-1}(A)), \qquad \forall A \in \mathcal{B}(\mathbb{R}^k).$$

Then (A.3) can be rewritten as

$$(A.4) \qquad \lambda(C) = \lim_k \lambda_k(P_k(C)).$$

Since for every $k = 1,2,\ldots$ the operator P_k is of the form (3.7) then, due to (3.10), λ_k is the Gaussian measure on $\mathcal{B}(\mathbb{R}^k)$ with mean element zero and correlation operator identity. Observe also, that $P_k(B)$ is convex and balanced and that $P_k(B+h) = P_k(B) + P_k(h)$. Hence the Anderson's inequality [1] yields that

(A.5) $$\lambda_k(P_k(B)) \geq \lambda_k(P_k(B+h)), \quad \forall k = 1,2,\ldots .$$

This and (A.4) implies that

$$\lambda(B) \geq \lambda(B+h)$$

which completes the proof of Lemma 3.1. ∎

Proof of Lemma 3.2: Let $\alpha_{j,i}$ be the eigenvalues of S_{λ_i} $(i = 1,2)$, and

(A.6) $$\alpha_{j,1} \leq \alpha_{j,2}, \quad \forall j = 1,2,\ldots .$$

Similarly as in the proof of Lemma 3.1 we can assume that $\alpha_{j,i} > 0$. Then

(A.7) $$\lambda_i(J(0,\epsilon)) = \lim_k A_{i,k}, \quad i = 1,2,$$

where

$$A_{i,k} = \frac{1}{\sqrt{(2\pi)^k}} \int_{B_{i,k}} \exp\{-\tfrac{1}{2} \Sigma_{j=1}^k y_j^2\} d(y_1,\ldots,y_k)$$

and

$$B_{i,k} = \{y \in \mathbb{R}^k : \Sigma_{j=1}^k \alpha_{j,i} y_j^2 \leq \epsilon^2\}.$$

Since $\alpha_{j,1} \leq \alpha_{j,2}$, $\forall j = 1,2,\ldots$, then $B_{2,k} \subset B_{1,k}$ which implies that $A_{1,k} \geq A_{2,k}$, $k = 1,2,\ldots$. This and (A.7) complete the proof of Lemma 3.2. ∎

Acknowledgments

I wish to thank J.F. Traub and H. Woźniakowski for their valuable suggestions concerning this paper.

References

[1] Anderson, T.W., "The Integral of Symmetric Unimodal Function over a Symmetric Convex Set and some Probability Inequalities." Proc. Amer. Math. Soc., Vol. 6 (1955), pp. 170-175.

[2] Kuo, Hui-Hsuing, Gaussian Measures in Banach Spaces, Lecture Notes in Mathematics 463, Springer-Verlag, Berlin, 1975.

[3] Parthasarathy, K.R., Probability Measures on Metric Spaces, Academic Press, New York, 1967.

[4] Skorohad, A.V., Integration in Hilbert Space, Springer-Verlag, New York, 1979.

[5] Traub, J.F., Wasilkowski, G.W. and Woźniakowski, H., "Average Case Optimality for Linear Problems," Dept. of Computer Science Report, Columbia University, 1981. To appear in J.T.C.S. 28 (1984).

[6] Wasilkowski, G.W., "Local Average Error," Dept. of Computer Science Report, Columbia University, 1983.

[7] Wasilkowski, G.W. and Woźniakowski, H., "Average Case Optimal Algorithm in Hilbert Spaces," Dept. of Computer Science Report, Columbia University, 1982.

[8] Wasilkowski, G.W. and Woźniakowski, H., "Can Adaption Help on the Average?" 1983. To appear in Num. Math.