CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Elsevier - Publisher Connector

JOURNAL OF COMPLEXITY 4, 257–276 (1988)

On Adaption with Noisy Information

J. B. KADANE*

Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213; and Center for Advanced Study in the Behavioral Sciences, Stanford, California 94305

G. W. WASILKOWSKI†

Department of Computer Science, University of Kentucky, Lexington, Kentucky 40506

AND

H. WOŹNIAKOWSKI‡

Department of Computer Science, Columbia University, New York, New York, 10027; and Institute for Informatics, University of Warsaw, Warsaw, Poland

Received June 17, 1987

When observations can be made without noise, it is known that adaptive information is no more powerful than nonadaptive information for approximation of linear problems with Gaussian measure. When the noise is additive, independent of the true value, and normal, once again adaption does not help (Theorem 1 in Section 4). However, when those conditions are not satisfied, Examples 1 and 2 of Section 4 show that adaptive information can be much more powerful than nonadaptive information. Finally if orthogonal observations are used with the sample size as well as the number of repetitions fixed, and only the directions of observations are chosen adaptively, then once again adaption does not help (Theorem 2 in Section 5). The issue is analogous to whether sequential designs are more powerful than fixed sample size designs in Bayesian statistics. © 1988 Academic Press, Inc.

*Research supported by the Office of Naval Research under Contract N00014-85-K-0539 and the National Science Foundation under Grant BNS 84-11738.
†Research supported by the National Science Foundation under Grant DCR-86-03674.
‡Research supported by the National Science Foundation under Grant ICT-85-17289.

0885-064X/88 $3.00
Copyright © 1988 by Academic Press, Inc.
All rights of reproduction in any form reserved.

1. INTRODUCTION

We explain the terminology and the problem studied in this paper by using an integration example. Suppose one wants to approximate $S(f) = \int_0^1 f(t)dt$ knowing n values of f at points t_i, $N(f) = [f(t_1), \ldots, f(t_n)]$, and knowing that f belongs to a given class F. If the number of observations, n, and the points t_i are fixed a priori than $N = N^{\text{non}}$ is called *nonadaptive* (or *parallel*) information. If n or the points t_i vary based on previously observed values $f(t_1), \ldots, f(t_{i-1})$ then $N = N^{\text{a}}$ is called *adaptive* (*sequential*) information.

One might expect that adaptive information would be much more powerful than nonadaptive information. That is, an approximation to $S(f)$ based on adaptive points would be much more accurate than an approximation to $S(f)$ based on a comparable number of observations at nonadaptive points. But this is *not* necessarily the case. It was shown in a number of papers that for any adaptive information N^{a} one can find nonadaptive information N^{non} which is as powerful as N^{a}. This result holds for the *worst case*, i.e., when the error of an algorithm is defined by its worst performance assuming that F is a balanced and convex set. This was established in Bakhvalov (1971) for arbitrary linear functionals S and generalized to linear operators S in Gal and Micchelli (1980) and Traub and Woźniakowski (1980). In both cases, N may consist of arbitrary linear functionals, $N(f) = [L_1(f), \ldots, L_n(f)]$.

Adaption also does not help on the *average*. By "on the average" we mean that the error of an algorithm is measured by its average performance according to some probability measure μ. Furthermore N^{non} is constrained to have a number of evaluations, n, roughly the same as the expected number of evaluations in N^{a}. The result holds for linear operators S, $S: F_1 \to F_2$, where F_1 is a separable Banach space and μ is a Gaussian measure defined on the Borel σ-field over F_1. This is proven in Wasilkowski (1986a), where results for more general probability measures, but with restricted notion of adaption, are also cited.

The results that adaption does not help has important implications concerning, for instance, parallel or distributed computations and the design of optimal information.

In the papers cited above *exact* information is assumed. In practice, however, one often has only *noisy* information. For our example, instead of $f(t_i)$ one has $f(t_i) + x_i$, where the noise x_i is a random variable. Therefore in this paper we study the following question: *Does adaption help on the average for noisy information?*

The answer to this question depends very much on the noise x, whose distribution may (or may not) depend on the observation L and the exact value, $y = L(f)$, one tries to observe. The first result of the paper states that

(i) if x has a normal distribution independent of L and y then adaption does not help essentially. (For the precise statement, see Theorem 1, Section 4).

For noise dependent on L and/or y, the situation might be quite different. In Examples 1 and 2 of Section 4 we show that adaption can even be much more effective than nonadaption. In the first example we construct adaptive information with a varying number of evaluations. This number is unbounded, but its expected value is equal to 3. Furthermore, the expected error of an algorithm that uses this information is zero, whereas any algorithm that uses arbitrary nonadaptive information has positive expected error. In the second example we construct information with a fixed number of observations, but the number of repetitions of certain observations varies adaptively. This information admits an algorithm whose expected error is exponentially smaller than the error of any algorithm that uses nonadaptive information with the same number of observations as in the constructed adaptive information. Thus, in both examples the number of observations or the number of repetitions vary but the observed functionals are fixed. In general, in adaptive information one is allowed not only to vary the number of observations or the numbers of repetitions, but also to choose adaptively the form of observation in an arbitrary way.

We believe that the power of adaption occurs only through adaptive choice of sample size or adaptive selection of the number of repetitions, but not through adaptive selection of observations (see Section 5). To support this belief we analyze the power of adaption for a restricted class of adaptive information. That is, we assume the number of evaluations in various orthogonal directions to be fixed, but the directions are adaptively determined. We prove that

(ii) if the distribution of the noise x satisfies the mild assumptions to be stated in (3) and (4) then adaption does not help.

We now comment on relations between the statistical literature and the results of this paper. A general discussion comparing and contrasting the average case setting and Bayesian statistics may be found in Kadane and Wasilkowski (1985). Here, we mention that adaptive information corresponds to sequential experiments in statistics.

The first result in the statistical literature on when the privilege of sequential experimentation is worthless is in Blackwell and Girschick (1954, Thm. 9.3.3, p. 254). They assume that the observations are independent and identically distributed, and that the experiments are charged a constant amount for each observation. They prove that if the Bayes risk is uniformly bounded and depends only on the sample size, the optimal sequential procedure is a fixed-sample-size procedure. The same theorem

is presented in easier notation in DeGroot (1970, Thm. 1, p. 285). While those results are stated for the class of independent and identically distributed observations, it is immediate that they hold as well if the observations are not necessarily identically distributed, but rather are independent given certain design variables.

We stress that in the statistical literature, the observation cost (sample size) and the error are additively combined into one risk function, i.e., Risk = Error + $c \times$ Cost for some constant c. In our approach, we do not combine the error and the observation cost into a risk function, but we relate N^a to N^{non} by comparing their expected errors and expected costs separately. Formally, this corresponds to the following risk function: given an error demand ε, Risk = Cost if the expected value of Error $\leq \varepsilon$, and Risk = $+\infty$ otherwise.

Other papers dealing with conditions under which optimal designs are not sequential are Darling (1972) and Whittle and Lane (1967). A more technical comparison of our results to this literature is given in Section 6.

2. Noisy Observations

Let F_1 be a real separable Banach space. Let F_2 be a real separable Hilbert space whose norm and inner product are denoted by $\|\bullet\|$ and $\langle \bullet, \bullet \rangle$. Consider a continuous linear operator S which maps F_1 into F_2. We wish to approximate Sf for all f from F_1. We assume that we do not know the element f. Instead we can compute (or observe) approximations to $L(f)$ for various continuous linear functionals $L \in F_1^*$. More precisely, we assume that instead of the exact value $y = L(f)$ we know

$$z = y + x = L(f) + x, \tag{1}$$

where the noise x is a random variable with a known probability measure $\eta(\bullet; y, L)$. That is, for any Borel set A of $\mathfrak{R}$,

$$\text{Prob}(x \in A) = \int_A \eta(dt; y, L). \tag{2}$$

Throughout this paper we assume that the probability measure η satisfies two conditions

$$\eta(A; \bullet, \bullet) = \eta(-A; \bullet, \bullet), \qquad \forall A \in \mathbf{B}(\mathfrak{R}), \tag{3}$$

$$\eta(\bullet; y, \bullet) = \eta(\bullet; -y, \bullet), \qquad \forall y \in \mathfrak{R}. \tag{4}$$

Assumption (3) implies that the mean value of the noise is zero. Assumption (4) means that the probability of the noise depends only on the absolute value of y.

We illustrate η by three examples. In each example η is absolutely continuous with respect to Lebesgue measure. The density of $\eta(\bullet; y, L)$ is denoted by $\rho(\bullet; y, L)$

(i) $\rho(t; y, L) = w(t)$ for some nonnegative w. For instance, $w(t) = (1/\sqrt{2\pi\sigma}) \exp(-t^2/(2\sigma))$ corresponds to $\eta(\bullet; y, L)$ being Gaussian (normal $\mathcal{N}(0, \sigma)$). Since w is independent of y and L, the noise x has the same probability whether y and/or $\|L\|$ are large or small. We think that this is a realistic assumption for some applications, but is unrealistic for others.

(ii) $\rho(t; y, L) = (1/\sqrt{2\pi\sigma(y)}) \exp(-t^2/(2\sigma(y)))$, where, for instance, $\sigma(y) = y^2$. This corresponds to a Gaussian probability whose variance depends on the exact value y.

(iii)

$$\rho(t; y, L) = \begin{cases} \dfrac{1}{2\alpha\|L\|(|y| + \delta)}, & \text{if } \dfrac{t}{\|L\|(|y| + \delta)} \in [-\alpha, \alpha], \\ 0, & \text{otherwise.} \end{cases}$$

Here α and δ are positive (small) numbers. This means that the noise x is uniformly distributed in the interval $[-\alpha\|L\|(|y| + \delta), \alpha\|L\|(|y| + \delta)]$. If $|y|$ is large relative to δ then the relative error $|z - y|/|y|$ has, roughly, the uniform distribution on $[-\alpha\|L\|, \alpha\|L\|]$. If $|y|$ is small relative to δ then the absolute error $|z - y|$ has, roughly, the uniform distribution on $[-\alpha\|L\|\delta, \alpha\|L\|\delta]$. Note that the noise x depends on the norm of L. This means that computing $L_c(f)$ instead of $L(f)$ with $L_c = cL$ for $c \in \mathbb{R}$, corresponds to noise x_c which behaves as cx. Such noise may be viewed as an abstraction of rounding errors in floating point arithmetic.

3. Stochastic Adpative Information

Assume that F_1 is equipped with a Gaussian measure μ whose mean element is zero and whose covariance operator S_μ is given. Recall that $S_\mu : F_1^* \to F_1$

$$L_1(S_\mu L_2) = \int_{F_1} L_1(f)L_2(f)\mu(df), \qquad \forall L_1, L_2 \in F_1^*$$

(see, e.g., Kuo (1975), Skorohod (1979), Vakhania (1981)). Without loss of generality we assume that S_μ is positive definite, i.e., $L(S_\mu L) > 0$ unless $L = 0$.

We define (noisy) adaptive information N as follows. Let $L_1 \in F_1^*$. We compute (or observe) $L_1(f)$. Since the computation of $L_1(f)$ involves noise, we observe

$$z_1 = y_1 + x_1, \qquad \text{where } y_1 = L_1(f). \tag{5}$$

As in Section 2, x_1 is a random variable with probability measure $\eta(\bullet; y_1, L_1)$ dependent on y_1 and L_1.

Knowing z_1 we decide whether another observation is desired. If not, z_1 constitutes the information about the element f. Otherwise, we choose another functional $L_2(\bullet; z_1) \in F_1^*$, compute (observe)

$$z_2 = y_2 + x_2, \qquad \text{where } y_2 = L_2(f; z_1), \tag{6}$$

and so on.

More formally, (*noisy*) *adaptive information* N is defined by

$$N(f, \mathbf{x}) = \mathbf{z} = [z_1, z_2, \ldots, z_{n(f,\mathbf{x})}] \tag{7}$$

with $z_i = z_i(f, \mathbf{x}) = y_i + x_i = L_i(f; z_1, \ldots, z_{i-1}) + x_i$. Here $y_i = y_i(f, \mathbf{x}) = L_i(f; z_1, \ldots, z_{i-1})$ is the exact value, and the functional $L_i(\bullet; z_1, \ldots, z_{i-1})$ is chosen based on the previously obtained values $z_1, \ldots, z_{i-1}$. For brevity, we shall often write $L_{i,\mathbf{z}}$ instead of $L_i(\bullet; z_1, \ldots, z_{i-1})$. The number of observations, $n(f, \mathbf{x})$, called the *cardinality* of N at f, is defined via stopping rules, i.e.,

$$n(f, \mathbf{x}) = \min\{k : [z_1(f, \mathbf{x}), \ldots, z_k(f, \mathbf{x})] \in T_k\} \tag{8}$$

for given Borel sets $T_i \subset \mathfrak{R}^i$. (By convention, $\min \emptyset = +\infty$). For simplicity we assume that $n(f, \mathbf{x})$ is finite almost surely.

Note that N defined as above is *adaptive* (*sequential*), since the choice of the ith observation depends on the previously observed values $z_1, z_2, \ldots, z_{i-1}$. Furthermore, the total number of observations, $n(f, \mathbf{x})$, is determined dynamically based on those values. On the other hand, if $n(f, \mathbf{x})$ and the functionals L_i are fixed a priori, we shall say that N is *nonadaptive*.

In this paper we shall relate adaptive to nonadaptive information by comparing their average cardinalities and radii. By the *average cardinality* of N we mean

$$\text{card}^{\text{avg}}(N) = \int_{F_1} \left(\int_{\mathfrak{R}^\infty} n(f, \mathbf{x}) \eta(d\mathbf{x}; \mathbf{y}, N) \right) \mu(df), \tag{9}$$

where μ is the a priori measure on F_1 and $\eta(A; \mathbf{y}, N) = \int_A \prod_{i=1}^k \eta(dx_i; y_i, L_{i,\mathbf{z}})$ for any $A \in \mathbf{B}(\mathfrak{R}^k)$.

To define the average radius, we proceed as follows. Knowing $N(f, \mathbf{x})$ we approximate Sf by $\phi(N(f, \mathbf{x}))$. Here ϕ, called an *algorithm*, is any mapping $\phi: N(F_1 \times \mathfrak{R}^\infty) \to F_2$. The *average error* of ϕ is defined by

$$e^{\text{avg}}(\phi, N) = \int_{F_1} \left(\int_{\mathfrak{R}^\infty} \|Sf - \phi(N(f, \mathbf{x}))\|^2 \eta(d\mathbf{x}; \mathbf{y}, N) \right) \mu(df), \quad (10)$$

and the *average radius* of N is defined by

$$r^{\text{avg}}(N) = \inf_{\phi} e^{\text{avg}}(\phi, N). \quad (11)$$

We assume that ϕ and L_i as a function of $z_1, \ldots, z_{i-1}$ are measurable. This guarantees that (9)–(11) are well defined.

Ideally, one would like to have information N with both $\text{card}^{\text{avg}}(N)$ and $r^{\text{avg}}(N)$ as small as possible. This suggests the following way to compare information N_1 with information N_2. We shall say that N_1 is as powerful as N_2 iff

$$\text{card}^{\text{avg}}(N_1) \le \text{card}^{\text{avg}}(N_2) \quad \text{and} \quad r^{\text{avg}}(N_1) \le r^{\text{avg}}(N_2).$$

In this paper we study when nonadaptive noisy information is as powerful as adaptive noisy information.

4. Adaption Versus Nonadaption

It is known (see, e.g., Wasilkowski (1986a)) that nonadaption is as powerful (or almost as powerful) as adaption for exact information, i.e., when $x_i = 0$ with probability one. This is also true for Gaussian noise, as stated in

THEOREM 1. *Let x_i have $\mathcal{N}(0, \sigma_i)$ distribution with σ_i independent of y_i and L_i. Then for every adaptive N^a:*

(i) *There exists N^* such that*

$$\text{card}^{\text{avg}}(N^*) \le \text{card}^{\text{avg}}(N^a) \quad \text{and} \quad r^{\text{avg}}(N^*) \le r^{\text{avg}}(N^a).$$

Furthermore, N^* has the following form. There exists a number α such that

$$N^*(f, \mathbf{x}) = \begin{cases} N_1(f, \mathbf{x}), & \text{if } z_1 = L_1(f) + x_1 \le \alpha, \\ N_2(f, \mathbf{x}), & \text{otherwise,} \end{cases}$$

where N_1, N_2 are nonadaptive and consist only of certain functionals used by N^a.

(ii) *There exists nonadaptive N^{non} (consisting of linear functionals not necessarily used by N^a) such that*

$$\text{card}^{avg}(N^{non}) \le [\text{card}^{avg}(N^a)] \quad \text{and} \quad r^{avg}(N^{non}) \le r^{avg}(N^a).$$

Sketch of the Proof. Letting $\tilde{F}_1 = F_1 \times \mathfrak{R}^\infty$, $S(f, \mathbf{x}) = Sf$ and $\tilde{L}_i((f, \mathbf{x}); z_1, \ldots, z_{i-1}) = L_i(f; z_1, \ldots, z_{i-1}) + x_i$, we get an equivalent problem with exact information. Since the a priori measure μ on F_1 is Gaussian, the joint probability on $\tilde{F}_1$ is also Gaussian. Hence the results of Wasilkowski (1986a) for exact information imply the existence of N^* and N^{non} with the properties stated in the theorem. ■

Theorem 1 holds for more general problems with an arbitrary linear space F_2 and with $\|Sf - \phi(N(f, \mathbf{x}))\|^2$ replaced by any loss of the form $P(Sf - \phi(N(f, \mathbf{x})))$ with $P: F_2 \to \mathfrak{R}_+$ such that $P(S(\bullet) - g)$ is Borel measurable for any fixed $g \in F_2$.

Although N^* in Theorem 1 (i) need not be nonadaptive in a strict sense, it is only "mildly" adaptive, since it is equal either to N_1 or to N_2. Furthermore, if β is the probability that $N^* = N_1$, i.e., β is the probability that $L_1(f) + x_1 \le \alpha$, then

$$\text{card}^{avg}(N_1) \le \text{card}^{avg}(N^*) \quad \text{and} \quad r^{avg}(N_1) \le 2r^{avg}(N^*) \quad \text{if } \beta \ge \tfrac{1}{2},$$

and

$$\text{card}^{avg}(N_2) \le 2\,\text{card}^{avg}(N^*) \quad \text{and} \quad r^{avg}(N_2) \le r^{avg}(N^*) \quad \text{if } \beta \le \tfrac{1}{2}.$$

As stated in Section 2 (i), white noise seems to us too restrictive for some applications. Hence one would like to have similar results for more general classes of noise distributions. However, in general, adaption can be much more powerful than nonadaption as exhibited in the following two examples. The distribution of noise in these examples is discrete. We chose this for simplicity. The same could be achieved with a continuous distribution of noise.

EXAMPLE 1. Let $F_1 = F_2 = \mathfrak{R}$, $S = I$, and $\mu = \mathcal{N}(0, I)$. Let $x_i = -1$ or $x_i = +1$, each with probability $\frac{1}{2}$. Consider adaptive N^a which consists of repetitive observations of $L(f) = f$, i.e., $N(f, \mathbf{x}) = [z_1, \ldots, z_{n(f,\mathbf{x})}]$, $z_i = f + x_i$, with the folowing stopping rule: $n(f, \mathbf{x}) = \min\{i \ge 2 : z_{i-1} \ne z_i\}$. As always, $\min \emptyset = +\infty$. Note that for every f, $n(f, \mathbf{x}) = i$ with probability $2^{-(i-1)}$. Then the algorithm $\phi(N^a(f, \mathbf{x})) = (z_{i-1} + z_i)/2$ is equal to f, and therefore has the average error zero. Hence

$$r^{avg}(N^a) = 0.$$

The average cardinality of N^a is given by

$$\operatorname{card}^{\operatorname{avg}}(N^a) = \sum_{i=2}^{\infty} \frac{i}{2^{i-1}} = 3.$$

Consider now nonadaptive N_k^{non} consisting of k repetitive noisy observations of $L(f) = f$. Then f can be recovered exactly only with probability $1 - 2^{-(k-1)}$ (when two observations are different). Hence

$$\begin{aligned} r^{\operatorname{avg}}(N_k^{\text{non}}) &= \frac{1}{2^k} \inf_{\phi} \int_{F_1} ((f - \phi(f + 1))^2 + (f - \phi(f - 1))^2)\mu(df) \\ &= \frac{1}{2^k\sqrt{2\pi}} \int_{\Re} \inf_{x \in \Re} ((z - 1 - x)^2 e^{-(z-1)^2/2} \\ &\quad + (z + 1 - x)^2 e^{-(z+1)^2/2}) dz. \end{aligned}$$

The last infimum is attained for

$$x = x(z) = \frac{(z - 1)e^{-(z-1)^2/2} + (z + 1)e^{-(z+1)^2/2}}{e^{-(z-1)^2/2} + e^{-(z+1)^2/2}},$$

i.e., the optimal algorithm $\phi^*(z, \ldots, z) = x(z)$, and

$$r^{\operatorname{avg}}(N_k^{\text{non}}) = \frac{4}{2^k\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{e^{-(z+1)^2/2}}{1 + e^{-2x}} dz.$$

Hence information N_k^{non} of cardinality k has positive average radius, whereas information N^a solves the problem exactly with average cardinality equal to 3.

EXAMPLE 2. Let $F_1 = F_2 = \Re^2$ be equipped with the l_2 norm, i.e., $f = [f_1, f_2]$ and $\|f\|^2 = f_1^2 + f_2^2$. Consider $Sf = [f_1, f_2]$. Let $\mu = \mathcal{N}(0, I)$ and let the noise of observing $G_i(f) = f_i$ be so that $x_i = -1$ or $x_i = +1$, each with probability $\frac{1}{2}$. Consider adaptive N^a with fixed cardinality, $n(f, \mathbf{x}) = n$, such that $L_1 = L_2 = G_1$, $L_i = G_1$ if $z_1 = \ldots = z_{i-1}$ and $L_i = G_2$ otherwise. In a fashion similar to the method used in Example 1, one can show that

$$\begin{aligned} r^{\operatorname{avg}}(N^a) &= b\left(2^{-(n-2)} + \frac{2^{-(n-2)}}{b} + (n - 2)2^{-(n-5)}\right) \\ &= b2^{-n}\left(2 + \frac{4}{b} + 2^5(n - 2)\right), \end{aligned}$$

where $b = (4/\sqrt{2\pi}) \int_{-\infty}^{+\infty} ((e^{-(z+1)^2/2})/(1 + e^{-2x}))dz$.

On the other hand, if N^{non} consists of n_1 noisy observations of G_1 and $(n - n_1)$ noisy observations of G_2 with n_1 fixed, then

$$r^{\text{avg}}(N^{\text{non}}) = b(2^{-n_1} + 2^{-(n-n_1)}),$$

which is minimized for $n_1^* = n/2$. This means that any nonadaptive information of cardinality n has average radius satisfying

$$r^{\text{avg}}(N^{\text{non}}) \geq 2b2^{-n/2}.$$

Hence adaption is exponentially more powerful than nonadaption.

In the above examples we exhibited adaptive information N^a which was more powerful than nonadaptive information. In Example 1, we constructed N^a by taking advantage of varying cardinality. In Example 2, $n(f, \mathbf{x})$ was fixed, but we adaptively changed the number of repetitions of the functional G_1. Thus, in both examples adaption was more powerful than nonadaption because of using either varying cardinality or varying the number of repetitions of certain nonadaptive functionals. In the next section we show that, in a sense to be made precise, these are the only causes for adaption to be more powerful than nonadaption.

5. Adaptive Choice of Observations Does Not Help

In this section we assume that the cardinality and the number of repetitions are fixed. We permit the observations (or equivalently, the functionals) to be chosen adaptively. We shall show that this adaptive choice of observations does not help.

Stated precisely, we assume that for given k, $n_1, \ldots, n_k$, $\sum_{i=k}^{k} n_i = n$,

$$N(f, \mathbf{x}) = \mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_k], \tag{12}$$

where

$$\mathbf{z}_i = [z_{i,1}, \ldots, z_{i,n_i}],\ z_{i,j} = L_i(f; \mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) + x_{i,j},$$
$$1 \leq i \leq k,\ 1 \leq j \leq n_i, \tag{13}$$

and the functionals $L_{1,\mathbf{z}}, \ldots, L_{k,\mathbf{z}}$ are μ-orthonormal for every fixed $\mathbf{z}$,

$$L_{i,\mathbf{z}}(S_\mu L_{j,\mathbf{z}})$$
$$= \int_{F_1} L_i(g; \mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) L_j(g; \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}) \mu(dg) = \delta_{i,j}. \tag{14}$$

Remark 1. The notion of fixed repetition numbers n_i requires us to distinguish between the functionals $L_{i,\mathbf{z}}$, $i = 1, \ldots, k$. One could hope that it would be enough to assume that $L_{i,\mathbf{z}} \neq L_{j,\mathbf{z}}$ for $i \neq j$. This assumption is, however, too weak. Indeed, consider once more adaptive information N^a from Example 2, with L_i replaced by $\tilde{L}_i = L_i + \varepsilon^{i-1}G_2$ for sufficiently small ε. Let N_ε^a consist of single observations of $\tilde{L}_1, \ldots, \tilde{L}_n$. For small ε, N_ε^a and N^a are practically the same, though the first information has fixed repetition numbers ($n_i = 1$) whereas the second one has varying repetition numbers. Hence the assumption $L_{i,\mathbf{z}} \neq L_{i,\mathbf{z}}$ does not lead to a meaningful notion of fixed repetition numbers.

Our definition (14) of fixed repetition numbers requires μ-orthogonality of $L_{i,\mathbf{z}}$. Observe that this holds for Examples 1 and 2. We have chosen this definition to simplify the analysis. We stress that this choice is not unique. Furthermore, Theorem 2, which we present below, need not be true for different notions of fixed repetition numbers.

Finally, we add that μ-orthogonality is not restrictive for exact information. Indeed, we can always fulfill (14) by taking a suitable linear combination of $L_{i,\mathbf{z}}$. This can be done, for instance, by applying the Gram–Schmidt reorthogonalization process.

We are ready to state

THEOREM 2. *For any adaptive N^a of the above form, there exists a vector $\mathbf{z}^* \in \mathbb{R}^n$ such that*

$$r^{\text{avg}}(N^{\text{non}}_{\mathbf{z}^*}) \leq r^{\text{avg}}(N^a).$$

Here $N^{\text{non}}_{\mathbf{z}^*}$ stands for nonadaptive information obtained from N^a by replacing $\mathbf{z} = z(f, \mathbf{x})$ by $\mathbf{z}^*$ in the functionals used by N^a.

We prove Theorem 2 assuming that for every $y \in \mathbb{R}$ and every $L \in F_1^*$ the probability of the noise, $\eta(\bullet; y, L)$, is absolutely continuous with respect to Lebesgue measure, and its density is denoted by $\rho(\bullet; y, L)$. This assumption is without loss of generality, and is made to simplify the notation. In the sequel, $\mu_k = \mathcal{N}(0, I)$ on $\mathbb{R}^k$, i.e., for any Borel set $A \subset \mathbb{R}^k$.

$$\mu_k(A) = \frac{1}{(2\pi)^{k/2}} \int_A e^{-\|\mathbf{y}\|^2/2}\, d_k\mathbf{y}. \tag{15}$$

Let N be adaptive information of the form (12). We need a few lemmas.

LEMMA 1. *For every algorithm ϕ*

$$e^{\text{avg}}(\phi, N) = \int_{F_1} \|Sf\|^2 \mu(df) - R(\phi, N) \tag{16}$$

with

$$R(\phi, N) = \int_{\mathfrak{R}^n} \int_{\mathfrak{R}^k} \left(2 \sum_{i=1}^{k} y_i \langle SS_\mu L_{i,\mathbf{z}}, \phi(\mathbf{z})\rangle - \|\phi(\mathbf{z})\|^2\right) \times \rho(\mathbf{x}; \mathbf{y}, N)\, d_n\mathbf{x}\, \mu_k(d\mathbf{y}), \tag{17}$$

where $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]$ *and* $\mathbf{z}_i = [y_i + x_{i,1}, \ldots, y_i + x_{i,n_i}]$.

Proof. Observe that

$$e^{\text{avg}}(\phi, N) = \int_{F_1} \int_{\mathfrak{R}^n} \|Sf - \phi(\mathbf{z})\|^2 \rho(\mathbf{x}; \mathbf{y}, N)\, d_n\mathbf{x}\, \mu(df)$$

$$= \int_{F_1} \int_{\mathfrak{R}^n} (\|Sf\|^2 - 2\langle Sf, \phi(\mathbf{z})\rangle + \|\phi(\mathbf{z})\|^2) \times \rho(\mathbf{x}; \mathbf{y}, N)\, d_n\mathbf{x}\, \mu(df).$$

Since $\int_{\mathfrak{R}^n} \rho(\mathbf{x}; \mathbf{y}, N)\, d_n\mathbf{x} = 1$, we have

$$e^{\text{avg}}(\phi, N) = \int_{F_1} \|Sf\|^2 \mu(df) - R(\phi, N),$$

where

$$R(\phi, N) = \int_{F_1} \int_{\mathfrak{R}^k} (2\langle Sf, \phi(\mathbf{z})\rangle - \|\phi(\mathbf{z})\|^2) \rho(\mathbf{x}; \mathbf{y}, N)\, d_n\mathbf{x}\, \mu(df).$$

Thus, it is enough to show that $R(\phi, N)$ satisfies (17). Changing the order of integration we have

$$R(\phi, N) = \int_{\mathfrak{R}^k} \int_{F_1} (2\langle Sf, \phi(\mathbf{z})\rangle - \|\phi(\mathbf{z})\|^2) \rho(\mathbf{x}; \mathbf{y}, N)\, \mu(df)\, d_n\mathbf{x}.$$

For fixed $\mathbf{x}$, let $N_{\mathbf{x}}(f) = [L_{1,\mathbf{z}}(f), \ldots, L_{k,\mathbf{z}}(f)]$. Recall that $L_{i,\mathbf{z}}(f) = L_i(\bullet; \mathbf{z}_1, \ldots, \mathbf{z}_{i-1})$. Then $N_{\mathbf{x}}$ is exact (noise-free) adaptive information.

Let $\mu_1(A, N_{\mathbf{x}}) = \mu(N_{\mathbf{x}}^{-1}(A))$ for all Borel sets A of $\mathfrak{R}^k$. It is known (see, e.g., Lee and Wasilkowski (1986), Wasilkowski (1986b), Wasilkowski and Woźniakowski (1984)) that $\mu_1(\bullet, N_{\mathbf{x}})$ does not depend on $N_{\mathbf{x}}$ and is equal to the Gaussian measure μ_k of (15).

For any Borel set B of F_1 we have

$$\mu(B) = \int_{\mathfrak{R}^k} \mu_2(B|\mathbf{y}, N_{\mathbf{x}}) \mu_k(d\mathbf{y}), \tag{18}$$

where $\mu_2(\bullet|\mathbf{y}, N_{\mathbf{x}})$ is the probability measure concentrated on $N_{\mathbf{x}}^{-1}(\mathbf{y})$ (see Parthasarathy (1967, Thm.8.1, p. 147)). From the papers cited above, it follows that

$$m_{\mathbf{y}} = \sum_{i=1}^{k} y_i S_\mu(L_{i,\mathbf{z}}) \tag{19}$$

is the mean element of the measure $\mu_2(\bullet|\mathbf{y}, N_{\mathbf{x}})$. Hence

$$\int_{F_1} \langle Sf, \phi(\mathbf{z})\rangle \mu_2(df|\mathbf{y}, N_{\mathbf{x}}) = \sum_{i=1}^{k} y_i \langle SS_\mu L_{i,\mathbf{z}}, \phi(\mathbf{z})\rangle.$$

From (18) and (19) we have

$$\begin{aligned} R(\phi, N) &= \int_{\mathfrak{R}^n} \int_{\mathfrak{R}^k} \int_{F_1} (2\langle Sf, \phi(\mathbf{z})\rangle - \|\phi(\mathbf{z})\|^2) \mu_2(df|\mathbf{y}, N_{\mathbf{x}}) \\ &\quad \times \rho(\mathbf{x}; \mathbf{y}, N) \mu_k(\mathbf{y})\, d_n\mathbf{x} \\ &= \int_{\mathfrak{R}^n} \int_{\mathfrak{R}^k} \left(2 \sum_{i=1}^{k} y_i \langle SS_\mu L_{i,\mathbf{x}}, \phi(\mathbf{z})\rangle - \|\phi(\mathbf{z})\|^2\right) \\ &\quad \times \rho(\mathbf{x}; \mathbf{y}, N) \mu_k(\mathbf{y})\, d_n\mathbf{x}. \end{aligned}$$

Hence $R(\phi, N)$ satisfies (17) as claimed. ■

We now exhibit an *optimal* algorithm, i.e., an algorithm ϕ^* such that $e^{\text{avg}}(\phi^*, N) = r^{\text{avg}}(N)$. Keeping in mind that $\rho(\mathbf{x}; \mathbf{y}, N)$ is the density of $\eta(\mathbf{x}; \mathbf{y}, N)$ and that μ_k is given by (15), we change variables in (17) by setting $z_{i,j} = y_i + x_{i,j}$. Then

$$\begin{aligned} R(\phi, N) &= (2\pi)^{-k/2} \int_{\mathfrak{R}^n} \int_{\mathfrak{R}^k} \left(2 \sum_{i=1}^{k} y_i \langle SS_\mu L_{i,\mathbf{x}}, \phi(\mathbf{z})\rangle - \|\phi(\mathbf{z})\|^2\right) \\ &\quad \times \prod_{i=1}^{k} \left(e^{-y_i^2/2} \prod_{j=1}^{n_i} \rho(z_{i,j} - y_i; y_i, L_{i,\mathbf{z}})\right) d_k\mathbf{y}\, d_n\mathbf{z}. \end{aligned} \tag{20}$$

Define

$$\begin{aligned} &\nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) \\ &\quad = \frac{1}{\sqrt{2\pi}} \int_{\mathfrak{R}} e^{-y_i^2/2} \prod_{j=1}^{n_i} \rho(z_{i,j} - y_i; y_i, L_i(\bullet; \mathbf{z}_1, \ldots, \mathbf{z}_{i-1}))\, dy_i. \end{aligned} \tag{21}$$

Then

$$\int_{\mathfrak{R}^{n_i}} \nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})\, d_{n_i}\mathbf{z}_i = 1. \tag{22}$$

This means that $\nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})$ is the density function of a probability measure on $\mathfrak{R}^{n_i}$. Define

$$\lambda_i(y_i|\mathbf{z}_1, \ldots, \mathbf{z}_i) = \frac{(2\pi)^{1/2}e^{-y_i^2/2}}{\nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})} \times \prod_{j=1}^{n_i} \rho(z_{i,j} - y_i;\, y_i,\, L_i(\bullet;\, \mathbf{z}_1, \ldots, \mathbf{z}_{i-1})). \tag{23}$$

Since $\int_{\mathfrak{R}} \lambda_i(y_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})dy_i = 1$, $\lambda_i(\bullet|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})$ is the density of a probability measure. We rewrite (20) using (21) and (23):

$$R(\phi, N) = \int_{\mathfrak{R}^n} \left(2\sum_{i=1}^{k} \langle SS_\mu L_{i,\mathbf{z}},\, \phi(\mathbf{z})\rangle \int_{\mathfrak{R}} y_i\lambda_i(y_i|\mathbf{z}_1, \ldots, \mathbf{z}_i)\, dy_i - \|\phi(\mathbf{z})\|^2\right) \prod_{i=1}^{k} \nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})\, d_n\mathbf{z}. \tag{24}$$

Let

$$H_i(\mathbf{z}) = H_i(\mathbf{z}_1, \ldots, \mathbf{z}_i) = \int_{\mathfrak{R}} y_i\lambda_i(y_i|\mathbf{z}_1, \ldots, \mathbf{z}_i)\, dy_i. \tag{25}$$

Define the algorithm

$$\phi^*(\mathbf{z}) = \sum_{i=1}^{k} H_i(\mathbf{z})SS_\mu L_{i,\mathbf{z}}. \tag{26}$$

We comment on the implementation of (26). The functionals $L_{i,\mathbf{z}}$ are given by the noisy adaptive information N. The elements $SS_\mu L_{i,\mathbf{z}}$ are determined by the problem being solved. Observe that for nonadaptive information these elements do not depend on $\mathbf{z}$. In any case, in order to compute $\phi^*(\mathbf{z})$ we have to compute $H_i(\mathbf{z})$ given by (25). The difficulty of computing $H_i(\mathbf{z})$ depends on the density function ρ of the noise. For some ρ it is relatively easy to compute $H_i(\mathbf{z})$. Then $\phi^*(\mathbf{z})$ can also be relatively easy to compute.

LEMMA 2. *The algorithm ϕ^* defined by (26) is optimal, i.e.,*

$$e^{\text{avg}}(\phi^*, N) = r^{\text{avg}}(N) = \int_{F_1} \|Sf\|^2 \mu(df) - R(\phi^*, N), \tag{27}$$

where

$$R(\phi^*, N) = \int_{\mathfrak{R}^n} \|\phi^*(\mathbf{z})\|^2 \nu(\mathbf{z})\, d_n\mathbf{z}$$

$$\nu(\mathbf{z}) = \prod_{i=1}^{k} \nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_i^{-1}).$$

Proof. From (24) to (26) we get

$$\begin{aligned} R(\phi, N) &= \int_{\mathfrak{R}^n} (2\langle \phi^*(\mathbf{z}), \phi(\mathbf{z})\rangle - \|\phi(\mathbf{z})\|^2)\nu(\mathbf{z})\, d_n\mathbf{z} \\ &= \int_{\mathfrak{R}^n} (\|\phi^*(\mathbf{z})\|^2 - \|\phi^*(\mathbf{z}) - \phi(\mathbf{z})\|^2)\nu(\mathbf{z})\, d_n\mathbf{z} \\ &\le \int_{\mathfrak{R}^n} \|\phi^*(\mathbf{z})\|^2 \nu(\mathbf{z})\, d_n\mathbf{z} = R(\phi^*, N). \end{aligned}$$

This and (16) yields $e^{\text{avg}}(\phi^*, N) \le e^{\text{avg}}(\phi^*, N)$ as claimed. ∎

We now establish some properties of ν_i and H_i.

LEMMA 3. *For $i = 1, \ldots, k$ and all vectors $\mathbf{z}_1, \ldots, \mathbf{z}_i$*

$$\nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) = \nu_i(-\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) \tag{28}$$

$$H_i(\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}_i) = -H_i(\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, -\mathbf{z}_i). \tag{29}$$

Proof. Changing y_i to $-y_i$ in (21), we get

$$\nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) = \frac{1}{\sqrt{2\pi}} \int_{\mathfrak{R}} e^{-y_i^2/2} \prod_{j=1}^{n_i} \rho(z_{i,j} + y_i; -y_i, L_{i,\mathbf{z}})\, dy_i.$$

Due to (3) and (4) we know that

$$\rho(z_{i,j} + y_i; -y_i, L_{i,\mathbf{z}}) = \rho(-z_{i,j} - y_i; -y_i, L_{i,\mathbf{z}}),$$

which yields (28). In a similar fashion one can prove that

$$\lambda_i(y_i|\mathbf{z}_1, \ldots, \mathbf{z}_i) = \lambda_i(-y_i|\mathbf{z}_1, \ldots, \mathbf{z}_i).$$

This and (22) yield (29). ∎

Define

$$G_i(\mathbf{z}) = G_i(\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) = \int_{\mathfrak{R}^{n_i}} H_i^2(\mathbf{z}_i, \ldots, \mathbf{z}_i)\nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})\, d_{n_i}\mathbf{z}_i. \tag{30}$$

LEMMA 4.

$$R(\phi^*, N) = \int_{\mathfrak{R}^n} \left(\sum_{i=1}^{k} \|SS_\mu L_{i,\mathbf{z}}\|^2 G_i(\mathbf{z})\right) \times \left(\prod_{i=1}^{k} \nu_i(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})\right) d_n\mathbf{z} = \sum_{i=1}^{k} \int_{\mathfrak{R}^n} \|SS_\mu L_{i,\mathbf{x}}\|^2 G_i^2(\mathbf{z})\nu(\mathbf{z})\, d_n\mathbf{z}. \tag{31}$$

Proof. From (26) we have

$$\|\phi^*(\mathbf{z})\|^2 = \sum_{i=1}^{k} \|SS_\mu L_{i,\mathbf{z}}\|^2 H_i^2(\mathbf{z}) + \sum_{i\neq j} \langle SS_\mu L_{i,\mathbf{z}}, SS_\mu L_{j,\mathbf{z}}\rangle H_i(\mathbf{z})H_j(\mathbf{z}).$$

For $j > i$ we have

$$\begin{aligned} A_{i,j} &= \int_{\mathfrak{R}^n} \langle SS_\mu L_{i,\mathbf{z}}, SS_\mu L_{j,\mathbf{z}}\rangle H_i(\mathbf{z})H_j(\mathbf{z})\nu(\mathbf{z})\, d_n\mathbf{z} \\ &= \int_{\mathfrak{R}^{n_1}} \cdots \int_{\mathfrak{R}^{n_{j-1}}} \langle SS_\mu L_i(\bullet, \mathbf{z}_1, \ldots, \mathbf{z}_{i-1}), \\ &\quad SS_\mu L_j(\bullet; \mathbf{z}_1, \ldots, \mathbf{z}_{j-1})\rangle H_i(\mathbf{z}_1, \ldots, \mathbf{z}_i) \\ &\quad \times \left(\int_{\mathfrak{R}^{n_j}} H_j(\mathbf{z}_1, \ldots, \mathbf{z}_j)\nu_j(\mathbf{z}_j|\mathbf{z}_1, \ldots, \mathbf{z}_{j-1})\, d_{n_j}\mathbf{z}_j\right) \\ &\quad \times \prod_{p=1}^{j-1} \nu_p(\mathbf{z}_p|\mathbf{z}_1, \ldots, \mathbf{z}_{p-1})\, d_{n_1}\mathbf{z}_1 \cdots d_{n_{j-1}}\mathbf{z}_{j-1}. \end{aligned}$$

Due to Lemma 3, the integral over $\mathfrak{R}^{n_j}$ is zero. Thus, $A_{i,j} = 0$. This yields

$$R(\phi^*, N) = \int_{\mathfrak{R}^n} \|\phi^*(\mathbf{z})\|^2\nu(\mathbf{z})\, d_n\mathbf{z} = \sum_{i=1}^{k} \int_{\mathfrak{R}^n} \|SS_\mu L_{i,\mathbf{z}}\|^2 H_i^2(\mathbf{z})\nu(\mathbf{z})\, d_n\mathbf{z}$$

$$= \sum_{i=1}^{k} \int_{\mathfrak{R}^{n_1}} \cdots \int_{\mathfrak{R}^{n_i}} \|SS_\mu L_i(\mathbf{z}_1, \ldots, \mathbf{z}_{i-1})\|^2 H_i^2(\mathbf{z}_1, \ldots, \mathbf{z}_i)$$

$$\times \prod_{p=1}^{i} \nu_p(\mathbf{z}_p|\mathbf{z}_1, \ldots, \mathbf{z}_{p-1})\, d_{n_1}\mathbf{z}_1 \cdots d_{n_i}\mathbf{z}_i,$$

due to (22). Thus,

$$R(\phi^*, N) = \sum_{i=1}^{k} \int_{\mathfrak{R}^{n_1}} \cdots \int_{\mathfrak{R}^{n_{i-1}}} \|SS_\mu L_{i,\mathbf{z}}\|^2 G_i(\mathbf{z})$$

$$\times \prod_{p=1}^{i} \nu_p(\mathbf{z}_p|\mathbf{z}_1, \ldots, \mathbf{z}_{p-1})\, d_{n_1}\mathbf{z}_1 \cdots d_{n_{i-1}}\mathbf{z}_{i-1}$$

$$= \sum_{i=1}^{k} \int_{\mathfrak{R}^n} \|SS_\mu L_{i,\mathbf{z}}\|^2 G_i^2(\mathbf{z})\nu(\mathbf{z})\, d_n\mathbf{z},$$

due to (22). Hence (31) is proven. ■

We are ready to prove Theorem 2. From Lemma 4 we have

$$R(\phi^*, N) = \int_{\mathfrak{R}^n} G(\mathbf{z})\nu(\mathbf{z})\, d_n\mathbf{z}, \tag{32}$$

where $G(\mathbf{z}) = \sum_{i=1}^{k} \|SS_\mu L_{i,\mathbf{z}}\|^2 G_i^2(\mathbf{z})$. Observe that there exists $\mathbf{z}^* = [\mathbf{z}_1^*, \ldots, \mathbf{z}_k^*] \in \mathfrak{R}^n$ such that

$$R(\phi^*, N) \leq G(\mathbf{z}^*). \tag{33}$$

Indeed, if $R(\phi^*, N) > G(\mathbf{z})$, $\forall \mathbf{z} \in \mathfrak{R}^n$, then

$$R(\phi^*, N) = \int_{\mathfrak{R}^n} R(\phi^*, N)\nu(\mathbf{z})\, d_n\mathbf{z} > \int_{\mathfrak{R}^n} G(\mathbf{z})\nu(\mathbf{z})\, d_n\mathbf{z} = R(\phi^*, N),$$

which is a contradiction. Here we used the fact that $\int_{\mathfrak{R}^n} \nu(\mathbf{z}) d_n\mathbf{z} = 1$ and that $R(\phi^*, N)$ is finite,

$$R(\phi^*, N) \leq \int_{F_1} \|Sf\|^2 \mu(df) \leq \|S\|^2 \int_{F_1} \|f\|^2 \mu(df) < \infty.$$

Thus, (33) holds.

Define $L_1^* = L_1$, $L_i^* = L_i(\bullet; \mathbf{z}_1, \ldots, \mathbf{z}_{i-1})$. Let

$$N^{\text{non}}_{\mathbf{z}^*}(f, \mathbf{x}) = [L_1^*(f) + x_{1,1}, \ldots, L_1^*(f) + x_{1,n_1}, \ldots, L_k^*(f) + x_{k,1}, \ldots, L_k^*(f) + x_{k,n_k}] \tag{34}$$

be the *nonadaptive* noisy information. The $x_{i,j}$ are random variables with density function equal to $\rho(\bullet; y_i, L_i^*)$. We prove that

$$r^{\text{avg}}(N^{\text{non}}_{\mathbf{z}^*}) = \int_{F_1} \|Sf\|^2 \mu(df) - G(\mathbf{z}^*). \tag{35}$$

Indeed, let ν_i^*, H_i^*, and G_i^* be defined by (21), (25), and (30) for the nonadaptive information $N^{\text{non}}_{\mathbf{z}^*}$. Then

$$\nu_i^*(\mathbf{z}_i|\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) = \nu_i^*(\mathbf{z}_i|\mathbf{z}_1^*, \ldots, \mathbf{z}_{i-1}^*),$$
$$H_i^*(\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}_i) = H_i(\mathbf{z}_1^*, \ldots, \mathbf{z}_{i-1}^*, \mathbf{z}_i),$$
$$G_i^*(\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}) = G_i^*(\mathbf{z}_1^*, \ldots, \mathbf{z}_{i-1}^*).$$

From (32) we conclude that for an optimal algorithm ϕ^* using $N^{\text{non}}_{\mathbf{z}^*}$ we have $R(\phi^*, N^{\text{non}}_{\mathbf{z}^*}) = G(\mathbf{z}^*)$. Then (27) of Lemma 2 yields (35).

We return to (33). Due to (33) and (35) we have

$$r^{\text{avg}}(N) = \int_{F_1} \|Sf\|^2 \mu(df) - R(\phi^*, N)$$
$$\geq \int_{F_1} \|Sf\|^2 \mu(df) - G(\mathbf{z}^*) = r^{\text{avg}}(N^{\text{non}}_{\mathbf{z}^*}).$$

Thus,

$$r^{\text{avg}}(N) \geq r^{\text{avg}}(N^{\text{non}}_{\mathbf{z}^*}),$$

which completes the proof of Theorem 2. ■

6. A Statistical Example

In this section, we give further details on the classical statistical problem of optimal sequential design for a class of normal linear models. To do so, we switch to statistical language and notation. The sample size n is fixed in advance, but the placement of the observations is permitted to depend on past observations. Nontheless, it is the case that for this example, the optimal sequential design ignores past observations.

Suppose we can observe a vector $y = [y_1, y_2, \ldots, y_n]^T$ such that

$$y = X^T\theta + e,$$

where $X = (X_1, X_2, \ldots, X_n)$ is the $k \times n$ design matrix and each X_i is a k-dimensional column vector, $\theta = [\theta_1, \theta_2, \ldots, \theta_k]^T$ is a vector of k unknown parameters, and $e|t \sim \mathcal{N}(0, tI)$ is the n-dimensional random vector of observations having a normal distribution with mean vector zero and precision matrix tI. Suppose that the prior on θ and t is such that the conditional distribution of θ given t is normal $\mathcal{N}(\theta_0, tR)$, where R is a specified $k \times k$ matrix. The posterior conditional distribtion of $\theta|y$, t is a normal with mean $\theta_1 = (R + XX^T)^{-1}(Xy + R\theta_0)$ and precision matrix $t(R + XX^T)$. If a particular linear combination $c^T\theta$ of θ_i's is of interest and squared error loss is appropriate, the optimal estimate is $c^T\theta_1$, and the posterior risk is the expected variance of $c^T\theta_1$, that is $c^T(R + XX^T)^{-1} cE_{t|y}(t^{-1})$, where $E_{t|y}(t^{-1})$ is the posterior mean of t^{-1}. Suppose first that the prior on t is such that t is known. Then the sequentially optimal choice of X would be a choice that minimizes $c^T(R + XX^T)^{-1}c$. Note that such X does not depend on y, so a fixed sample-size procedure is optimal in this case. This is an application of the (extended) Blackwell–Girshick Theorem (see Section 1). See Chaloner (1984) for methods of finding such an optimal design. Furthermore, since the optimal design does not depend on t either, it is sequentially optimal whatever prior is taken on t.

REFERENCES

BAHKVALOV, N. S. (1971), On the optimality of linear methods for operator approximation in convex classes of functions, *USSR Comput. Math. Math. Phys.* **11,** 244–249.

BLACKWELL, D., AND GIRSHICK, M. A. (1954), "Theory of Games and Statistical Decisions," Wiley, New York; also Dover, New York (1979).

CHALONER, K. (1984), Optimal Bayesian experimental designs for linear models, *Ann. Statist.* **12,** 283–300.

DARLING, D. (1972), When is a fixed number of observations optimal?" *in* "Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability," Vol. IV, pp. 33–35.

DEGROOT, M. H. (1970), "Optimal Statistical Decisions," McGraw–Hill, New York.

GAL, S., AND MICCHELLI, C. A. (1980), Optimal sequential and non-sequential procedures for evaluating a functional, *Appl. Anal.* **10,** 105–120.

KADANE, J. B., AND WASILKOWSKI, G. W. (1985), Average case ε-complexity: A Bayesian view, *in* "Bayesian Statistics" (J. M. Bernardo *et al.*, Eds.), Vol. 2, pp. 361–375, North-Holland, New York.

KUO, HUI-HSUING (1975), "Gaussian Measures in Banach Spaces," Lecture Notes in Math., Vol. 463, Springer-Verlag, New York.

LEE, D., AND WASILKOWSKI, G. (1986), Approximation of linear functionals on a Banach space with a Gaussian measure, *J. Complexity* **2,** 12–43.

PARTHASARATHY, K. R. (1967), "Probability Measures on Metric Spaces," Academic Press, New York.

SKOROHOD, A. V. (1979), "Integration in Hilbert Space," Springer-Verlag, New York.

TRAUB, J. F., AND WOŹNIAKOWSKI, H. (1980), "A General Theory of Optimal Algorithms," Academic Press, New York.

VAKHANIA, N. N. (1981), "Probability Distributions on Linear Spaces," North-Holland, New York.

WASILKOWSKI, G. W. (1986a), Information of varying cardinality, *J. Complexity* **2,** 204–228.

WASILKOWSKI, G. W. (1986b), Optimal algorithms for linear problems with Gaussian measures, *Rocky Mountain J. Math.* **16,** 727–749.

WASILKOWSKI, G. W., AND WOŹNIAKOWSKI, H. (1984), Can adaption help on the average? *Numer. Math.* **44,** 169–190.

WHITTLE, P., AND LANE, R. J. D. (1967), A class of situations in which a sequential estimation procedure is non-sequential, *Biometrika* **54,** 229–234.