Noisy Information and Computational Complexity

L. Plaskota
Institute of Applied Mathematics and Mechanics
University of Warsaw

Contents

Chapter 1

Overview

In the process of doing scientific computations we always rely on some *information.* A typical situation in practice is that this information is contaminated by errors. We say that it is *noisy.* Sources of noise include:

- previous computations,
- inexact measurements,
- transmission errors,
- arithmetic limitations,
- adversary's lies.

Problems with noisy information have always attracted a considerable attention of researchers in many different scientific fields: statisticians, engineers, control theorists, economists, applied mathematicians. There is also a vast literature, especially in statistics, where noisy information is analyzed from different perspectives.

In this monograph, noisy information is studied in the context of the computational complexity of solving mathematically posed problems.

The computational complexity focuses on the intrinsic difficulty of problems as measured by the minimal amount of time, memory, or elementary operations necessary to solve them. *Information–based complexity* (IBC) is a branch of computational complexity that deals with problems for which the available information is:

- *partial*,
- *noisy*,
- *priced.*

Information being *partial* means that the problem is not uniquely determined by the given information. Information is *noisy* since it may be contaminated by some errors. Finally, information is *priced* since we must pay for getting it. These assumptions distinguish IBC from *combinatorial complexity*, where information is complete, exact, and free.

Since information is partial and noisy, only approximate solutions are possible. One of the main goals of IBC is finding the *complexity* of the problem, i.e., the intrinsic cost of computing an approximation with given accuracy. Approximations are obtained by algorithms that use some information. These solving the problem with minimal cost are of special importance and called *optimal*.

Partial, noisy and priced information is typical of many problems arising in different scientific fields. These include, for instance, signal processing, control theory, computer vision, and numerical analysis. As a rule, a digital computer is used to perform scientific computations. A computer can only make use of a finite set of numbers. Usually, these numbers cannot be exactly entered into the computer memory. Hence, problems described by infinitely many parameters can be "solved" using only partial and noisy information.

The theory of optimal algorithms for solving problems with partial information has a long history. It can be traced back to the late forties when Kiefer, Sard and Nikolskij wrote pioneering papers. A systematic and uniform approach to such kind of problems was first presented by J.F. Traub and H. Woźniakowski in the monograph *A General Theory of Optimal Algorithms*, Academic Press, 1980. This was an important stage in the development of the theory of IBC.

The monograph was followed then by *Information, Uncertainty, Complexity*, Addison-Wesley, 1983, and *Information-Based Complexity*, Academic Press, 1988, both authored by J.F. Traub, G.W. Wasilkowski, and H. Woźniakowski. Computational complexity of approximately solved problems is also studied in the books: *Deterministic and Stochastic Error Bounds in Numerical Analysis* by E. Novak, Springer Verlag, 1988, and *The Computational Complexity of Differential and Integral Equations* by A.G. Werschulz, Oxford University Press, 1991.

Relatively few IBC papers study noisy information. One reason is the technical difficulty of the analysis of noisy information. A second reason is that even if we are primarily interested in noisy information, the results on exact information establish a benchmark. All negative results for exact information are also applicable for the noisy case. On the other hand, it is not clear whether positive results for exact information have a counterpart for noisy information.

In the mathematical literature, the word “noise” is used mainly by statisticians and means a random error that occurs for experimental observations. We also want to study deterministic error. Therefore by noise, we mean random or deterministic error. Moreover, in our model, the source of the information is not important. We may say that “information is observed” or that it is “computed”.

We also stress that the case of exact information is not excluded, neither in the model nor in most results. Exact information is obtained as a special case by setting the noise level to zero. This permits us to study the dependence of the results on the noise level, and to compare the noisy and exact information cases.

In general, optimal algorithms and problem complexity depend on the *setting.* The setting is specified by the way the error and cost of an algorithm are defined. If the error and cost are defined by their worst performance, we have the *worst case setting.* The *average case setting* is obtained when the average performance of algorithms is considered. In this monograph, we study the worst and average case settings as well as mixed settings and asymptotic setting. Other settings such as probabilistic and randomized settings will be the topic of future research.

Despite the differences, the settings have certain features in common. For instance, algorithms that are based on smoothing splines are optimal, independent of the setting. This is a very desirable property, since it shows that such algorithms are universal and robust.

Most of the research presented in this monograph has been done over the last 5–6 years by different people, including the author. Some of the results have not been previously reported. The references to the original results are given in Notes and Remarks at the end of each section. Clearly, the author does not pretend to cover the whole subject of noisy information in one monograph. Only these topics are presented that are typical of IBC, or are needed for the complexity analysis. Many problems are still open. Some of these are indicated in the text.

The monograph consists of six chapters. We start with the worst case setting in Chapter 2. Chapter 3 is devoted to the average case setting. Each of these two settings is studied following the same scheme. We first look for the best algorithms that use fixed information. Then we allow the information to vary and seek optimal information. Finally, complexity concepts are introduced and complexity results are presented for some particular problems. Chapters 4 and 5 are devoted to the mixed settings, while Chapter 6 to the asymptotic setting.

Each chapter consists of several sections, each followed by Notes and Remarks, and Exercises. A preview of the results is presented in the introduction of each chapter.

Chapter 2

Worst case setting

2.1 Introduction

In this chapter we study the worst case setting. We shall present already known results as well as we show some new results. As already mentioned in the Overview, precise information about what is known and what is new can be found in Notes and Remarks.

Our major goal is to obtain tight complexity bounds for the approximate solution of linear continuous problems that are defined on infinite dimensional spaces. We first explain what is to be approximated and how an approximation is obtained. That is, we carefully introduce the fundamental concepts of solution operator, noisy information and algorithm. A special attention is devoted to information which is most important in our analysis. Information is, roughly speaking, what we know about the problem to be solved. A crucial assumption is that information is *noisy*, i.e., it is not given exactly, but with some error.

Since information is usually partial (i.e., many elements share the same information) and noisy, it is impossible to solve the problem exactly. We have to be satisfied with only approximate solutions. They are obtained by algorithms that use information as data. In the worst case setting, the error of an algorithm is given by its worst performance over all problem elements and possible information. A sharp lower bound on the error is given by a quantity called a *radius of information.* We are obviously interested in algorithms with the minimal error. Such algorithms are called optimal.

In Sections 2.4 to 2.6 we study optimal algorithms and investigate whether they can be linear or affine. In many cases the answer is positive. This is the case for approximation of linear functionals and approximation of operators that act between spaces endowed by Hilbert seminorms, assuming that information is linear with noise bounded in a Hilbert seminorm. The optimal linear algorithms are based on the well known smoothing splines. This confirms a common opinion that smoothing splines are a very good practical tool for constructing approximations. We show that in some special cases smoothing splines are closely related to the least squares and regularization algorithms.

When using smoothing splines or regularization, a good choice of the smoothing or regularization parameters becomes an important question. Often special methods, such as cross validation, are developed to find them. We show how to choose the smoothing and regularization parameters optimally in the worst case setting, and how this choice depends on the noise level and the domain of the problem. It turns out that in some cases the regularization parameter is independent of the noise level provided that a bound on the noise is sufficiently small.

In Sections 2.7 and 2.8 we allow not only algorithms but also information to vary. We assume that information is obtained by successive noisy observations (or computations) of some functionals. The choice of functionals and noise bounds depend on us. We stress that we do not exclude the case when errors coming from different observations are correlated. This allows us also to model information where the noise of information is bounded, say, in a Hilbert norm.

With varying information, it is important to know whether adaption can lead to better approximations than nonadaption. We give sufficient conditions under which adaption is not better than nonadaption. These conditions are satisfied, for instance, if linear information with noise bounded in a norm is used.

Then we study the optimal choice of observations with given precisions. This is in general a difficult problem. Therefore we establish complete results only for two classes of problems. The first class consists of approximating compact operators acting between Hilbert spaces where the noise is bounded in the weighted Euclidean norm. In particular, it turns out that in this case the error of approximation can be arbitrarily reduced by using observations with fixed precisions. This does not hold for noise bounded in the supremum norm. When using this norm, to decrease the error of approximation, we have to perform observations with higher precisions. We stress that observations with noise bounded in the supremum norm seem to be most often used in practice. Exact formulas for the minimal errors are in this case obtained for approximation of Lipschitz functions based on noisy function values.

In Section 2.9 we present the model of computation and define the ε–complexity of a problem as the minimal cost needed to obtain an approximation with the (worst case) error at most ε. In the worst case setting, the cost of approximation is measured by the worst performance of an algorithm over all elements of the problem. In general, the cost of successive observations depends on their precisions. However, the model also covers the case when observations with a given, fixed precision are only allowed.

The complexity results are obtained using previously established results on optimal algorithms, adaption and optimal information. We first give tight general bounds on the ε–complexity. It turns out that if the optimal algorithms are linear then in many cases the cost of combining information is much less than the cost of gaining it. In such a case, the problem complexity is roughly equal to the *information complexity* which is defined as the minimal cost of obtaining information that guarantees approximation within the error ε. This is the reason why we are so much interested in existence of optimal linear algorithms.

In the last section we specify the general complexity results to some special problems. First, we consider approximation of compact operators in Hilbert spaces where information is linear with noise bounded in the weighted Euclidean norm. We show sharp upper and lower complexity bounds. We also investigate how the complexity depends on the cost assigned to each precision.

Next, we derive the ε–complexity for approximation and integration of Lipschitz functions. For a fixed positive bound on the noise, the complexity is infinite for sufficiently small ε. To make the complexity finite for all positive ε, we have to allow observations with arbitrary precisions. Then the ε–complexity is roughly attained by information that uses observations of function values at equidistant points with the same precision which is proportional to ε.

Finally, we consider approximation of smooth multivariate functions in a Banach space. We assume that the noise of successive observations is bounded in the absolute or relative sense. We show that in both cases the ε–complexity is roughly the same and is achieved by polynomial interpolation based on data about function values at equispaced points, and with a noise bound proportional to ε.

2.2 Information, algorithm, approximation

Let F be a linear space and G a normed space, both over the reals. Let

$$S : F \to G$$

be a mapping, called a *solution operator*. We are mainly interested in linear S. However, for the general presentation of the basic concepts we do not have to put any restrictions on S. We wish to approximate elements $S(f)$ for f belonging to a set $E \subset F$. An approximation is constructed based *only* on some noisy information about f. We now explain precisely how the noisy information and the approximation are obtained.

An *information operator* (or simply *information*) is a mapping

$$\mathbb{N} : F \to 2^Y,$$

where Y is a set of finite real sequences, $Y \subset \bigcup_{n=1}^{\infty} \mathbb{R}^n$. That is, $\mathbb{N}(f)$ is a subset of Y. We assume that $\mathbb{N}(f)$ is nonempty for all $f \in F$. Any element $y \in \mathbb{N}(f)$ will be called *information about* f. Note that knowing y, we conclude that f is a member of the set $\{ f_1 \in F \mid y \in \mathbb{N}(f_1) \}$. This yields some information about the element f and justifies the names for $\mathbb{N}$ and y.

If the set $\mathbb{N}(f)$ has exactly one element for all $f \in F$, information $\mathbb{N}$ is called *exact*. In this case, $\mathbb{N}$ will be identified with the operator $N : F \to Y$, where $N(f)$ is the unique element of $\mathbb{N}(f)$. If there exists f for which $\mathbb{N}(f)$ has at least two elements, we say that $\mathbb{N}$ is *noisy*.

Knowing the information y about f, we combine it to get an approximation. More precisely, the approximation is produced by an *algorithm* which is given as a mapping

$$\varphi : Y \to G.$$

The algorithm takes the obtained information as data. Hence, the approximation to $S(f)$ is $\varphi(y)$ where y is information about f. The *error of approximation* is defined by the difference $\|S(f) - \varphi(y)\|$ where $\|\cdot\|$ is the norm in the space G.

We illustrate the concepts of noisy information and algorithm by three simple examples.

Example 2.1 Suppose we want to approximate a real number (parameter) f based on its perturbed value y, $|y - f| \le \delta$. This corresponds to $F = G = \mathbb{R}$ and $S(f) = f$. The information is of the form

$$\mathbb{N}(f) = \{ y \in \mathbb{R} \mid \ |y - f| \le \delta \}$$

with $Y = \mathbb{R}$. For $\delta = 0$, we have exact information, $N(f) = f$, and for $\delta > 0$ we have noisy information. An algorithm φ is a mapping $\varphi : \mathbb{R} \to \mathbb{R}$. For instance, it may be given as $\varphi(y) = y$.

Example 2.2 Suppose we want to approximate a smooth function based on noisy function values at n points. This can be modeled as follows.

Let F be the space of two-times continuously differentiable real functions $f : [0,1] \to \mathbb{R}$. We approximate $f \in F$ in the norm of the space $G = \mathcal{L}_2(0,1)$. That is, $S(f) = f$. For $t_i \in [0,1]$, the information operator is given by

$$\mathbb{N}(f) \;=\; \left\{ y \in \mathbb{R}^n \;\Big|\; \sum_{i=1}^{n} (y_i - f(t_i))^2 \le \delta^2 \right\}.$$

Knowing y corresponds to n noisy observations of $f(t_i)$, $1 \le i \le n$. An example of the algorithm is provided by the smoothing spline. For a given parameter $\gamma \ge 0$, it is defined as the function $\varphi_\gamma(y)$ which minimizes the functional

$$\Gamma_\gamma(f, y) \;=\; \gamma \cdot \int_0^1 (f''(t))^2 \, dt \,+\, \sum_{i=1}^{n} (y_i - f(t_i))^2$$

over all $f \in F$.

Example 2.3 Let F be as in Example 2.2 or another "nice" class of smooth functions. The problem now is to approximate the integral of f based on noisy function values $f(t_i)$ with different precisions. That is, the solution operator is given as

$$S(f) \;=\; \int_0^1 f(t) \, dt \,,$$

and information is defined as

$$\mathbb{N}(f) \;=\; \{\, y \in \mathbb{R}^n \mid \; |y_i - f(t_i)| \le \delta_i, \, 1 \le i \le n \,\}.$$

An example of the algorithm is a quadrature formula $\varphi(y) = \sum_{i=1}^{n} a_i \, y_i$. □

In all the above examples, information operators belong to a common class. This class is defined in the following way.

An *extended seminorm* in a linear space X is a functional $\|\cdot\|_X : X \to [0,+\infty]$, such that the set $X_1 = \{\, x \in X \mid \|x\|_X < +\infty \,\}$ is a linear subspace, and $\|\cdot\|_X$ is a seminorm on X_1. That is,

(a) $\|\alpha x\|_X \;=\; |\alpha|\, \|x\|_X, \qquad \forall\, \alpha \in \mathbb{R},\ \forall\, x \in X_1,$

(b) $\|x_1 + x_2\|_X \;\le\; \|x_1\|_X + \|x_2\|_X, \qquad \forall\, x_1, x_2 \in X_1.$

We say that an information operator is *linear with uniformly bounded noise*, iff it is of the form

$$\mathbb{N}(f) \;=\; \{\, y \in \mathbb{R}^n \mid \quad \|y - N(f)\|_Y \le \delta \,\}, \qquad \forall\, f \in F, \tag{2.1}$$

where $N : F \to Y = \mathbb{R}^n$ is a linear operator, $\|\cdot\|_Y$ is an extended seminorm in $\mathbb{R}^n$, and $\delta \ge 0$.

For instance, in Example 2.2 we have

$$N(f) \;=\; [\, f(t_1), f(t_2), \ldots, f(t_n) \,].$$

As the extended seminorm $\|\cdot\|_Y$ we may take the Euclidean norm, $\|x\|_Y = \|x\|_2 = (\sum_{i=1}^n x_i^2)^{1/2}$. In Example 2.3 the operator N is as above, and

$$\|x\|_Y \;=\; \max_{1 \le i \le n} \frac{|x_i|}{\delta_i}$$

(with the convention that $a/(+\infty) = 0,\ a/0 = +\infty,\ 0/0 = 0$), and $\delta = 1$.

Observe that for any linear information with uniformly bounded noise, the extended seminorm $\|\cdot\|_Y$ and the parameter δ are not determined uniquely. In particular, replacing $\|\cdot\|_Y$ for $\delta > 0$ by $\|x\|_Y' = \|x\|_Y/\delta$, and for $\delta = 0$ by

$$\|x\|_Y' \;=\; \begin{cases} 0 & \|x\|_Y = 0, \\ +\infty & \|x\|_Y > 0, \end{cases}$$

we can always set δ to be 1. However, we prefer to have a parameter δ (and the norm independent of δ) since it can be often interpreted as a *noise level.* The smaller δ, the smaller the noise. If $\|\cdot\|_Y$ is a norm and δ goes to zero, then noisy information approaches exact information.

We now characterize linear information with uniformly bounded noise. Suppose that a subset B of a linear space X is convex (i.e., $x, y \in B$ implies $\alpha x + (1-\alpha) y \in B$ for all $\alpha \in [0,1]$), and balanced (i.e., $x \in B$ iff $-x \in F$). Let

$$p_B(x) \;=\; \inf\{\, t > 0 \mid \quad x/t \in B \,\}, \qquad x \in X.$$

Lemma 2.1 *The functional p_B is an extended seminorm on X.*

Proof Indeed, let $p_B(x), p_B(y) < +\infty$ and $\alpha \in \mathbb{R}$. Then, for $\alpha = 0$ we have $p_B(\alpha x) = 0 = \alpha p_B(x)$, and for $\alpha \neq 0$ we have

$$\begin{aligned} p_B(\alpha x) &= \inf\{\, t > 0 \mid \quad \alpha x/t \in B \,\} \\ &= \inf\{\, |\alpha| t > 0 \mid \quad x/t \in B \,\} \;=\; |\alpha|\, p_B(x). \end{aligned}$$

We now check the triangle inequality. If $x/t, y/u \in B$, then from the convexity of B we obtain

$$\frac{x+y}{t+u} \;=\; \frac{t}{t+u} \cdot \frac{x}{t} + \frac{u}{t+u} \cdot \frac{y}{u} \in B.$$

Hence,

$$\begin{aligned} p_B(x) + p_B(y) &= \inf\{\, t > 0 \mid \quad x/t \in B \,\} + \inf\{\, u > 0 \mid \quad y/u \in B \,\} \\ &\geq \inf\{\, t + u > 0 \mid \quad (x+y)/(t+u) \in B \,\} \\ &= p_B(x+y). \end{aligned}$$

Thus the set $X_1 = \{\, x \in X \mid p_B(x) < \infty \,\}$ is a linear subspace, on which p_B is a seminorm, which means that p_B is an extended seminorm on X. □

We also observe that

$$\{\, x \in X \mid p(x) < 1 \,\} \;\subset\; B \;\subset \{\, x \in X \mid p(x) \leq 1 \,\}.$$

Moreover, if B is a closed [1] subset of $\mathbb{R}^n$ then $B = \{\, x \in \mathbb{R}^n \mid p(x) \leq 1 \,\}$.

Now, let the set $B \subset \mathbb{R}^n$ be convex, balanced and closed. Consider the information operator of the form

$$\mathbb{N}(f) \;=\; \{\, N(f) + x \mid \quad x \in B \,\}, \tag{2.2}$$

where $N : F \to \mathbb{R}^n$ is a linear mapping. Then, setting $\|x\|_Y = \delta \cdot p(x)$ we have that $\mathbb{N}$ is linear with noise bounded uniformly by δ in the extended seminorm $\|\cdot\|_Y$. On the other hand, if information $\mathbb{N}$ is of the form (2.1) then it can be expressed by (2.2) with $B = \{\, x \in \mathbb{R}^n \mid \|x\|_Y \leq \delta \,\}$. Thus, we have proved the following fact.

[1] Recall that in $\mathbb{R}^n$ all norms are equivalent. Therefore, if B is closed with respect to a particular norm then B is also closed with respect to all norms in R^n.

Corollary 2.1 *The classes of information (2.2) and linear information with uniformly bounded noise are equivalent.* □

Clearly, not all information operators of interest can be expressed by (2.1).

Example 2.4 Suppose we have a vector $f = [f_1, f_2, \dots, f_n] \in \mathbb{R}^n$ with $|f_i| \le 1$, $\forall i$, which we store in computer memory using floating point arithmetic with t mantissa bits. Then the difference between the exact f_i and stored data y_i satisfies $|y_i - f_i| \le 2^{-t}\,|f_i|$. The vector y can be interpreted as noisy information about f where

$$\mathbb{N}(f) \;=\; \{\, y \in \mathbb{R}^n \mid \quad |y_i - f_i| \le 2^{-t}|f_i|,\ 1 \le i \le n\,\}.$$

In this case, $\mathbb{N}(0) = \{0\}$ is a singleton which is not true for $\mathbb{N}(f)$ with $f \ne 0$. Hence, the noise of information is *not* uniformly bounded.

Notes and Remarks

NR 2.1 A more concept of solution operator may be found in Traub *et al.* [107].

NR 2.2 For the exact information case, the formulation presented here corresponds to the formulation given in Traub *et al.* [108]. The concept of noisy information is, however, slightly different than this given in Traub *et al.* [108, Chap.12].

NR 2.3 The problem of approximating an operator $S : F \to G$ by noisy or exact information can be formulated in terms of approximating multi–valued operators by single–valued operators. Indeed, let the multi–valued operator be given as $\mathbb{S} : Y_0 \to 2^G$ with $Y_0 = \bigcup_{f\in E} \mathbb{N}(f)$ and

$$\mathbb{S}(y) \;=\; \{\, S(f) \mid \quad f \in E,\ y \in \mathbb{N}(f)\,\}.$$

Then $\mathbb{S}(y)$ is approximated by $\varphi(y)$, where $\varphi : Y_0 \to G$ is an arbitrary single-valued operator. This approach is presented in, e.g., Arestov [1] or Magaril–Il'yaev and Osipenko [52].

NR 2.4 The functional $p_B(x)$ is called the Minkowski functional (or gauge function) corresponding to the set B, see e.g., Wilansky [126].

2.3 Radius and diameter of information

Let $\mathbb{N} : F \to 2^Y$ be a given information operator. The *worst case error* (or simply *error*) of an algorithm $\varphi : Y \to G$ (that uses information $\mathbb{N}$) over the set $E \subset F$ is defined as

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi) \;=\; \sup_{f \in E} \sup_{y \in \mathbb{N}(f)} \|S(f) - \varphi(y)\|. \tag{2.3}$$

Our aim is to minimize the error (2.3) with respect to all algorithms φ. An algorithm φ_{opt} for which

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{opt}}) \;=\; \inf_{\varphi} \; \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi),$$

is called *optimal*.

It turns out that the problem of optimal algorithm is tightly related to the concepts of radius and center of a set. We recall that the *radius* of a set $A \subset G$ is given as

$$r(A) \;=\; \inf_{g \in G} \sup_{a \in A} \|a - g\|.$$

If for some $g_A \in G$ we have $\sup_{a \in A} \|a - g_A\| \;=\; r(A)$, then g_A is called a *center* of A.

Denote $Y_0 \;=\; \bigcup_{f \in E} \mathbb{N}(f)$. For $y \in Y_0$, let

$$E(y) \;=\; \{\, f \in E \mid \quad y \in \mathbb{N}(f)\,\}$$

be the set of all elements f which are in E and share the same information y. Finally, let

$$A(y) \;=\; \{\, S(f) \mid \quad f \in E(y)\,\}$$

be the set of solution elements with information y. A *radius of information* $\mathbb{N}$ is defined as

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sup_{y \in Y_0} r(A(y)).$$

Clearly, the radius $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ depends not only on information $\mathbb{N}$ but also on the solution operator S and the set E. If necessary, we will indicate this dependence and write, for instance, $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; S, E)$ or $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E)$.

It turns out that the radius of information yields the minimal error of algorithms. Namely, we have

Theorem 2.1 *For any information operator* $\mathbb{N}$*,*

$$\inf_{\varphi} \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi) \;=\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}).$$

The optimal algorithm exists if and only if $r(A(y)) = \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ *implies that* $A(y)$ *has a center. In particular, if for any* y *there exists a center* g_y *of the set* $A(y)$ *then the algorithm*

$$\varphi_{\mathrm{ctr}}(y) \;=\; g_y$$

is optimal.

Proof For any algorithm φ, its error can be rewritten as

$$\begin{aligned} \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi) &= \sup_{y \in Y_0} \sup_{f \in E(y)} \|S(f) - \varphi(y)\| \\ &= \sup_{y \in Y_0} \sup_{g \in A(y)} \|g - \varphi(y)\|. \end{aligned}$$

Hence, using the definition of the radius of a set, we obtain

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi) \;\ge\; \sup_{y \in Y_0} r(A(y)) \;=\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}),$$

and consequently

$$\inf_{\varphi} \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi) \;\ge\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}).$$

To prove the inverse inequality, it suffices to observe that for any $\delta > 0$ it is possible to select elements $\varphi_\delta(y)$, $y \in Y_0$, such that

$$\sup_{f \in E(y)} \|S(f) - \varphi_\delta(y)\| \;\le\; r(A(y)) \,+\, \delta.$$

For the algorithm φ_δ we have

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_\delta) \;\le\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \,+\, \delta.$$

Since δ is arbitrary, $\inf_\varphi \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi) \le \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$.

To prove the second part of the theorem, suppose that each set $A(y)$ with $r(A(y)) = \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ has a center g_y. Then, for any $y \in Y_0$ we can choose an element $\tilde{g}_y \in G$ such that

$$\sup_{a \in A(y)} \|a - \tilde{g}_y\| \;\le\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$$

(if $r(A(y)) = \mathrm{rad}^{\mathrm{wor}}(N)$ then $\tilde{g}_y = g_y$). An optimal algorithm is given as $\varphi_{\mathrm{opt}}(y) = \tilde{g}_y$.

On the other hand, if for some $y_0 \in Y_0$ we have $r(A(y_0)) = \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ and the set $A(y_0)$ has no center, then for any algorithm we have

$$\begin{aligned} \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi) \;&\geq\; \sup_{f \in E(y_0)} \|S(f) - \varphi(y_0)\| \\ &>\; r(A(y_0)) \;=\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}). \end{aligned}$$

This shows that an optimal algorithm does not exist. □

The algorithm φ_{ctr} defined in the above theorem is called *central.* The central algorithm (if it exists) has even stronger properties than the usual optimal algorithm. Indeed, φ_{ctr} is optimal not only with respect to the set E, but also with respect to each $E(y)$. Namely, for any $y \in Y_0$ we have

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{ctr}}; E(y)) \;=\; \inf_{\varphi} \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi; E(y)) \;=\; r(A(y)).$$

Together with the notion of a radius, it is convenient to introduce the notion of a diameter of information $\mathbb{N}$. Recall first that the diameter of a set A is given as

$$d(A) \;=\; \sup_{a_{-1}, a_1 \in A} \|a_1 - a_{-1}\|.$$

We also recall that for any set A we have

$$r(A) \;\leq\; d(A) \;\leq\; 2 \cdot r(A). \tag{2.4}$$

Example 2.5 Let a set $A \subset G$ be centrosymmetric. That is, there exists an element $a^* \in G$ such that the condition $a \in A$ implies $2\,a^* - a \in A$. Then a^* is the center of A and

$$d(A) \;=\; 2 \cdot r(A) \;=\; 2 \cdot \sup\{\, \|a - a^*\| \;|\; \; a \in A\,\}$$

Indeed, using the triangle inequality we obtain

$$\begin{aligned} r(A) \;&\geq\; \inf_{g \in G} \sup_{a \in A} \frac{1}{2}\,(\, \|g - a\| + \|g - (2a^* - a)\|\,) \\ &\geq\; \inf_{g \in G} \sup_{a \in A} \|a - a^*\| \;=\; \sup_{a \in A} \|a - a^*\|, \end{aligned}$$

which shows that a^* is a center. To prove the remaining equality, observe that

$$d(A) \geq \sup_{a \in A} \|a - (2a^* - a)\| \;=\; 2 \sup_{a \in A} \|a - a^*\|. \quad \square$$

A *diameter of information* $\mathbb{N}$ is defined as

$$\mathrm{diam}(\mathbb{N}) \;=\; \sup_{y\in Y_0} d(A(y)).$$

Observe that in view of the equality

$$\begin{aligned} & d(A(y)) \\ =\; & \sup\{\, \|S(f_1)-S(f_{-1})\| \;|\; f_{-1},f_1\in F_0,\, y\in\mathbb{N}(f_{-1})\cap\mathbb{N}(f_1)\,\}, \end{aligned}$$

the diameter of information can be rewritten as

$$\mathrm{diam}(\mathbb{N}) \;=\; \sup \|S(f_1)-S(f_{-1})\|,$$

where the supremum is taken over all $f_{-1},f_1\in E$ such that $\mathbb{N}(f_{-1})\cap\mathbb{N}(f_1)\neq\emptyset$. Thus, roughly speaking, $\mathrm{diam}(\mathbb{N})$ measures the largest distance between two elements in $S(E)$ which cannot be distinguished with respect to information.

The diameter of information is tightly related to the radius. although its definition is independent of the notion of an algorithm. Namely, in view of (2.4), we have the following fact.

Theorem 2.2 *For any information* $\mathbb{N}$*,*

$$\mathrm{diam}(\mathbb{N}) \;=\; c\cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$$

where $c=c(\mathbb{N})\in[1,2]$*.* □

In general, c depends on information and the set E. However, in some cases it turns out to be an absolute constant.

Example 2.6 Let S be a functional, i.e., the range space $G=\mathbb{R}$. Then, for any set $A\subset\mathbb{R}$ we have $d(A)=2\,r(A)$ and the center of A is $(\sup A+\inf A)/2$. Hence, for any information $\mathbb{N}$ the constant c in Theorem 2.2 is equal to 2. □

The relation between the radius and diameter of information allows us to show "almost" optimality of an important class of algorithms. An algorithm φ_{int} is called *interpolatory* iff for all $y\in Y_0$

$$\varphi_{\mathrm{int}}(y) \;=\; S(f_y),$$

for an element $f_y \in E(y)$.

Since $S(f_y)$ is a member of $A(y)$, for any $f \in E(y)$ we have

$$\|S(f) - \varphi_{\rm int}(y)\| \;=\; \|S(f) - S(f_y)\| \;\le\; d(A(y)) \;\le\; \mathrm{diam}(\mathbb{N}).$$

This yields the following fact.

Corollary 2.2 *For any interpolatory algorithm $\varphi_{\rm int}$ we have*

$$\mathrm{e}^{\rm wor}(\mathbb{N}, \varphi_{\rm int}) \;\le\; 2 \cdot \mathrm{rad}^{\rm wor}(\mathbb{N}). \quad \Box$$

In some important cases, the diameter of information can be expressed in a simple way. For a set $A \subset F$, let

$$\mathrm{bal}(A) \;=\; (A - A)/2 \;=\; \{\,(a_1 - a_{-1})/2 \mid \quad a_{-1}, a_1 \in A\,\}.$$

Observe that the set $\mathrm{bal}(A)$ is balanced, i.e., it is centrosymmetric with the center zero. It is also convex for convex A. Obviously, $\mathrm{bal}(A) = A$ for convex and balanced A.

Lemma 2.2 *Let the solution operator S be linear. Let $\mathbb{N}$ be an information operator with $Y = \mathbb{R}^n$ satisfying*

$$\mathbb{N}(f_1) \cap \mathbb{N}(f_{-1}) \neq \emptyset \quad \textit{for} \quad f_{-1}, f_1 \in E \;\Longrightarrow\; 0 \;\in\; \mathbb{N}\left(\frac{f_1 - f_{-1}}{2}\right) \tag{2.5}$$

and

$$\begin{array}{lcl} h \in \mathrm{bal}(E),\; 0 \in \mathbb{N}(h) & \Longrightarrow & \exists\, f_{-1}, f_1 \in E,\; \textit{such that}\; \mathbb{N}(f_1) \cap \mathbb{N}(f_{-1}) \neq \emptyset \\ & & \textit{and}\;\; h = (f_1 - f_{-1})/2\,. \end{array} \tag{2.6}$$

Then

$$\mathrm{diam}(\mathbb{N}) \;=\; 2 \cdot \sup\{\, \|S(h)\| \mid \quad h \in \mathrm{bal}(E),\; 0 \in \mathbb{N}(h)\,\}. \tag{2.7}$$

If, in addition, the set E is convex and balanced, then

$$\begin{aligned} \mathrm{diam}(\mathbb{N}) \;&=\; 2 \cdot \sup\{\, \|S(h)\| \mid \quad h \in E,\, 0 \in \mathbb{N}(h)\,\} \\ &=\; d(A(0)) \;=\; 2 \cdot r(A(0)), \end{aligned} \tag{2.8}$$

where $A(0) \;=\; \{\, S(h) \mid h \in E,\, 0 \in \mathbb{N}(h)\,\}$.

Proof The first part of the lemma follows directly from (2.5), (2.6), and linearity of S. The assumption (2.5) yields the upper bound and (2.6) yields the lower bound on $\mathrm{diam}(\mathbb{N})$ in (2.7). Since for convex and balanced set E we have $\mathrm{bal}(E) = E$, the first equality in (2.8) is also valid.

To prove the remaining two equalities in (2.8), we first show that the set $A(0)$ is balanced. Indeed, let $h \in E$, $0 \in \mathbb{N}(h)$. Then, from (2.6) we have $h = (f_1 - f_{-1})/2$, where $f_{-1}, f_1 \in E$ and $\mathbb{N}(f_{-1}) \cap \mathbb{N}(f_1) \neq \emptyset$. Using (2.5) we get $0 \in \mathbb{N}(\,(f_{-1} - f_1)/2\,) = \mathbb{N}(-h)$. Hence, $S(h) \in A(0)$ implies $-S(h) = S(-h) \in A(0)$.

To complete the proof it suffices to observe that the set $A(0)$ is centrosymmetric with the center zero and use the fact proven in Example 2.5. □

Lemma 2.2 yields the following theorem which is the main result of this section.

Theorem 2.3 *Let S be a linear operator. Let information $\mathbb{N}$ be linear with uniformly bounded noise,*

$$\mathbb{N}(f) \;=\; \{\, y \in \mathbb{R}^n \mid \quad \|y - N(f)\|_Y \;\le\; \delta \,\}.$$

If the set E is convex then

$$\mathrm{diam}(\mathbb{N}) \;=\; 2 \cdot \sup \{\, \|S(h)\| \mid \quad h \in b(E),\; \|N(h)\| \le \delta \,\}.$$

Proof It suffices to check the assumptions of Lemma 2.2. Indeed, if $\|y - N(f_i)\|_Y \le \delta$, for $i = -1, 1$, then also $\|0 - N(f_1 - f_{-1})/2\|_Y \le \delta$, which shows (2.5). To show (2.6), let $h = (f_1 - f_{-1})$ with $f_1, f_{-1} \in E$ and $0 \in \mathbb{N}(h)$, i.e., $\|N(f_1 - f_{-1})/2\|_Y \le \delta$. Then for $y = N(f_{-1} + f_1)/2$ we have $\|y - N(f_i)\|_Y \le \delta$, as claimed. □

A larger class of information for which Lemma 2.2 holds consists of information operators $\mathbb{N} : F \to 2^Y$, such that $Y = \mathbb{R}^n$ and the graph

$$\mathrm{gr}(\mathbb{N}; E) \;=\; \{\, (f, y) \in F \times \mathbb{R}^n \mid \quad f \in E,\; y \in \mathbb{N}(f) \,\}$$

is a convex and balanced set. This fact is left as E 2.8.

Notes and Remarks

NR 2.5 Abstractly, the concept of an optimal algorithm can be introduced as follows. Let R be a relation defined on the Cartesian product of algorithms. For two algorithms we write $\varphi_1 \prec \varphi_2$ iff $(\varphi_1, \varphi_2) \in R$ and say that φ_1 is *not worse* than φ_2 (or that φ_2 is *not better* than φ_1). An algorithm φ_{opt} is *optimal* iff

$$\varphi_{\mathrm{opt}} \;\prec\; \varphi, \qquad \forall \varphi.$$

In this section we use the (worst case) error criterion. It corresponds to the relation

$$\varphi_1 \;\prec\; \varphi_2 \quad \Longleftrightarrow \quad \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_1) \;\le\; \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_2).$$

If the relation is defined as

$$\varphi_1 \;\prec\; \varphi_2 \quad \Longleftrightarrow \quad \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_1; E(y)) \;\le\; \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_2; E(y)), \quad \forall y \in Y_0,$$

then only the central algorithm (if it exists) turns out to be optimal.

NR 2.6 The notions of the radius and diameter of information were introduced in Traub and Woźniakowski [109]. The formula for $\mathrm{diam}(N)$ in the case of linear information with noise bounded in a seminorm and convex and balanced set E, was first shown by Micchelli and Rivlin [59]. They used the fact that the radius of noisy information is equal to the radius of some appropriately chosen exact information; see also E 2.7.

Exercises

E 2.1 Give an example of information $\mathbb{N}$ and a set E for which:
1. Optimal algorithm does not exist.
2. Optimal algorithm does exist, but central algorithm does not.

E 2.2 Show that the set of all optimal algorithms is convex.

E 2.3 Prove the inequalities

$$r(A) \;\le\; d(A) \;\le\; 2 \cdot r(A),$$

for an arbitrary set A.

E 2.4 Let $1 \le c \le 2$.
1. Find a set A for which $d(A) \;=\; c \cdot r(A)$, with $r(A) \in (0, +\infty)$.
2. Find information $\mathbb{N}$ and a set E, such that

$$\mathrm{diam}(\mathbb{N}) \;=\; c \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$$

and $r(\mathbb{N}) \in (0, +\infty)$.

E 2.5 Let $S : F \to G$ be an arbitrary solution operator. Show that for any information operator $\mathbb{N}$ and any convex set $E \subset F$ we have

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E) \;=\; c \cdot \sup_{f_1, f_2 \in E} \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; [f_1, f_2]\,),$$

where $c \in [1, 2]$. Moreover, if S is a functional then $c = 1$. (Here $[f_1, f_2] = \{\, \alpha f_1 + (1-\alpha) f_2 \mid 0 \le \alpha \le 1 \,\}$.)

E 2.6 Let the solution operator $S : F \to G$ be linear. Let E be a balanced and convex set, and let information $\mathbb{N}$ be linear with noise bounded uniformly in a norm $\|\cdot\|_Y$. Suppose there exists an operator $A : Y \to F$ such that for any $f \in E$ and $y \in \mathbb{N}(f)$ we have $f - A(y) \in \{\, h \in E \mid \; \|N(h)\|_Y \le \delta \,\}$. Show that then the algorithm $\varphi(y) = S(A(y)\,)$, $\forall y$, is optimal.

E 2.7 Let the solution operator $S : F \to G$, information $\mathbb{N} : F \to 2^Y$ with $Y = \mathbb{R}^n$, and set E be given. Define the space $\tilde{F} = F \times Y$, solution operator $\tilde{S} : \tilde{F} \to G$, exact information operator $\tilde{N} : \tilde{F} \to Y$, and set $\tilde{E} \subset \tilde{F}$ as

$$\begin{aligned} \tilde{S}(f, y) &= S(f), \\ \tilde{N}(f, y) &= y, \\ \tilde{E} &= \{\, (f, y) \mid \; f \in E,\; y \in \mathbb{N}(f) \,\}. \end{aligned}$$

Show that for any algorithm $\varphi : Y \to G$ we have

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi; S, E) \;=\; \tilde{\mathrm{e}}^{\mathrm{wor}}(\tilde{N}, \varphi; \tilde{S}, \tilde{E})$$

where the second quantity stands for the error of φ over $\tilde{E}$, for approximating $\tilde{S}(f, y)$ based on exact information $y = \tilde{N}(f)$.

E 2.8 Show that information whose graph $\mathrm{gr}(N; E)$ is convex and balanced satisfies the conditions (2.5) and (2.6). *of Lemma 2.2.*

E 2.9 Let

$$\mathbb{N}(f) \;=\; \{\, y \in \mathbb{R}^n \mid \;\; (y - N(f)) \in B \,\},$$

where $N : F \to \mathbb{R}^n$ is linear and B is a given set of $\mathbb{R}^n$. Show that the graph $\mathrm{gr}(\mathbb{N}; E)$ is convex (and balanced) if both sets B and E are convex (and balanced).

2.4 Affine algorithms for linear functionals

In this section we deal with the case when

- the solution operator S is a linear functional.

We are especially interested in finding optimal linear or affine algorithms.

2.4.1 Existence of optimal affine algorithms

Since now the space $G = \mathbb{R}$, we have

$$\mathrm{diam}(\mathbb{N}) \;=\; 2 \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sup_{y \in Y_0} \left(\sup A(y) - \inf A(y) \right),$$

where $Y_0 = \bigcup_{f \in E} \mathbb{N}(f)$, $A(y) = \{\, S(f) \,|\, f \in E,\, y \in \mathbb{N}(f)\,\}$. The algorithm $\varphi(y) = (\sup A(y) + \inf A(y)\,)/2$ is optimal and also central. We now ask if there exists an optimal algorithm which is linear or affine. It is easily seen that, in general, this is not true.

Example 2.7 Let $F = \mathbb{R}^2$ and

$$E \;=\; \{\, f = (f_1, f_2) \in \mathbb{R}^2 \;|\quad f_2 = f_1^3,\ |f_1| \le 1\,\}.$$

Then the set E is balanced but not convex. Let $S(f) = f_2$ and $\mathbb{N}(f) = \{f_1\}$. In this case the problem can be solved exactly. However, the only optimal algorithm, $\varphi_{\mathrm{opt}}(y) = y^3$, is nonlinear. □

Restricting properly the class of problems, it is however possible to show the positive result. In what follows, we assume that $Y = \mathbb{R}^n$ and $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) < +\infty$.

Theorem 2.4 *Let S be a linear functional. If the graph $\mathrm{gr}(\mathbb{N}; E)$ of the information operator $\mathbb{N}$ is convex then there exists an optimal affine algorithm. If, in addition, $\mathrm{gr}(\mathbb{N}, E)$ is balanced then any optimal affine algorithm is linear.*

Proof Suppose first that $\mathrm{gr}(\mathbb{N}, E)$ is a convex set. Let $r = \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$. If $r = 0$ then each set $A(y)$, $y \in Y_0$, has exactly one element which we denote by a_y. Let $y_0 \in Y_0$. The functional $\varphi_1(y) = a_{y+y_0} - a_{y_0}$ is linear on its convex domain $Y_0 - y_0$ and can be extended to a linear functional φ_2 defined on Y. Letting $\varphi(y) = \varphi_2(y - y_0) + a_{y_0}$ we obtain an optimal affine algorithm.

Let $r > 0$. Consider the set

$$A \;=\; \{\, (y, S(f)) \in \mathbb{R}^{n+1} \;|\quad f \in E,\, y \in \mathbb{N}(f)\,\}.$$

Since $\mathrm{gr}(\mathbb{N}, E)$ is convex, A is also convex. Then the set $A_1 = \mathrm{bal}(A) = (A - A)/2$ is convex and balanced. Let

$$p(u) \;=\; \inf\{\, t > 0 \;|\quad u/t \in A_1\,\}, \qquad u \in \mathbb{R}^{n+1}.$$

We show that for $u = (0, g) \in A_1$, $g > 0$, we have $p(u) = g/r$. Indeed, Lemma 2.2 yields

$$\begin{aligned} r &= \sup\{\, |S(h)| \mid \quad h \in \mathrm{bal}(E),\, 0 \in \mathbb{N}(h)\,\} \\ &= \sup\{\, \alpha \in \mathbb{R} \mid \quad (0, \alpha) \in A_1\,\}. \end{aligned}$$

Hence, the infimum over all $t > 0$ such that $(0, g/t) \in A_1$ is equal to g/r.

Recall that $p(u)$ is a seminorm on the linear space $P = \{u \in \mathbb{R}^{n+1} | \, p(u) < +\infty\,\}$. Let $P_0 = \{\, u \in \mathbb{R}^{n+1} \mid p(u) = 0\,\}$ and $P_1 = \{\,(0, g) \in \mathbb{R}^{n+1} \mid g \in \mathbb{R}\,\}$. Since $P_1 \cap P_0 = \{0\}$, the space P can be decomposed as $P = P_0 \oplus P_0^{\perp}$ where $P_1 \subset P_0^{\perp}$. Define on P_1 the linear functional ξ_1 as $\xi_1(u) = p(u) = g/r$ where $u = (0, g)$. Since $p(u)$ is a norm on $P_0^{\perp}$, from the classical Hahn-Banach theorem it follows that ξ_1 can be extended to a functional ξ_2 which is defined on $P_0^{\perp}$ and satisfies $\xi_2(u) = \xi_1(u)$ for $u \in P_1$, and $\xi_2(u) \le p(u)$ for all $u \in P_0^{\perp}$.

For $u = u_0 + u_0^{\perp} \in P$ with $u_0 \in P_0$, $u_0^{\perp} \in P_0^{\perp}$, we now define $\xi(u) = \xi_2(u_0^{\perp})$. We claim that the functional ξ has two properties:

(i) $\quad \xi(u) = p(u), \quad \forall u \in P_1,$

(ii) $\quad \xi(u) \le p(u), \quad \forall u \in P.$

As (i) is obvious, it remains to show (ii). Let $u = u_0 + u_0^{\perp}$ and $t > 0$ be such that $u/t \in A_1$. Let $0 < \alpha < 1$ and $\beta = -\alpha/(1-\alpha)$. Since $p(u_0) = 0$, we have $\beta u_0/t \in A_1$, and from convexity of A_1 it follows that $\alpha u_0^{\perp}/t = \alpha u/t + (1-\alpha)\beta u_0/t \in A_1$. Since t and α can be arbitrarily close to $p(u)$ and 1, respectively, we obtain $p(u_0^{\perp}) \le p(u)$. Hence, $\xi(u) = \xi_2(u_0^{\perp}) \le p(u_0^{\perp}) \le p(u)$, and (ii) follows.

For $(y, g) \in P$, $y \in \mathbb{R}^n$, $g \in \mathbb{R}$, the functional ξ can be represented as $\xi(y, g) = \varphi_1(y) + \gamma(g)$ where $\varphi_1(y) = \xi(y, 0)$ and $\gamma(g) = \xi(0, g) = g/r$. As $u \in A_1$ yields $p(u) \le 1$, we have $A_1 \subset P$. Hence, for any $f_i \in E$, $y_i \in \mathbb{N}(f_i)$, $i = -1, 1$,

$$\begin{aligned} &\xi\left(\frac{y_1 - y_{-1}}{2}, \frac{S(f_1) - S(f_{-1})}{2}\right) \\ &= \varphi_1\left(\frac{y_1 - y_{-1}}{2}\right) + \frac{1}{2r}(\,S(f_1) - S(f_{-1})\,) \;\le\; 1. \end{aligned}$$

Setting $\varphi_2 = -r\varphi_1$ we get from the last inequality that

$$S(f_1) - \varphi_2(y_1) - r \;\le\; S(f_{-1}) - \varphi_2(y_{-1}) + r.$$

It now follows that there exists a number $a \in \mathbb{R}$ such that for all f_i and $y_i \in \mathbb{N}(f_i)$, $i = -1, 1$, it holds

$$S(f_1) - \varphi_2(y_1) - r \;\le\; a \le\; S(f_{-1}) - \varphi_2(y_{-1}) + r.$$

Setting $\varphi_{\mathrm{aff}}(y) = \varphi_2(y) + a$ we finally obtain

$$|\, S(f) - \varphi_{\mathrm{aff}}(y) \,| \le r, \qquad f \in E,\, y \in \mathbb{N}(f).$$

Thus the affine algorithm φ_{aff} is optimal.

Suppose now that $\mathrm{gr}(\mathbb{N}, E)$ is not only convex but also balanced. Then from Lemma 2.2 we have $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = r(A(0))$. Since in this case the set $A(0)$ is balanced, its center is equal to zero and for any optimal algorithm φ we have $\varphi(0) = 0$. Hence, any optimal affine algorithm is linear. □

The fact that S is a functional together with Theorem 2.4 yields an interesting property of the radius of information. Assume that E is convex and that the information is linear with noise bounded in a (not necessarily Hilbert) norm $\|\cdot\|_Y$,

$$\mathbb{N}(f) \;=\; \{\, y \in \mathbb{R}^n \mid \;\; \|y - N(f)\|_Y \le \delta \,\}.$$

Let $r(\delta)$ be the radius of $\mathbb{N}$. Then we have the following fact.

Lemma 2.3 *The function $K(\delta)$ defined by*

$$K(\delta) \;=\; \frac{r(\delta) \,-\, r(0)}{\delta}, \qquad \delta > 0,$$

is nonincreasing and bounded. In particular, the derivative $r'(0^+)$ exists.

Proof We first show that $K(\delta)$ is nonincreasing. Let $0 < \gamma < \delta$. For $\varepsilon > 0$, let $h_0, h_\delta \in \mathrm{bal}(E)$ be such that $N(h_0) = 0$, $S(h_0) \ge r(0) - \varepsilon$, and $\|N(h_\delta)\|_Y \le \delta$, $S(h_\delta) \ge r(\delta) - \varepsilon$. Let $h_\gamma = h_0 + (\gamma/\delta)(h_\delta - h_0)$. Then $h_\gamma \in \mathrm{bal}(E)$ and $\|N(h_\gamma)\|_Y \le \gamma$. Hence,

$$\begin{aligned} r(\gamma) \;&\ge\; S(h_\gamma) \;=\; S(h_0) \,+\, \frac{\gamma}{\delta}\,(\, S(h_\delta) - S(h_0)\,) \\ &\ge\; r(0) \,+\, \gamma\,\frac{r(\delta) - r(0)}{\delta} \,-\, \varepsilon\left(1 + \frac{\gamma}{\delta}\right). \end{aligned}$$

Letting $\varepsilon \to 0$, we obtain the desired inequality $K(\gamma) \ge K(\delta)$.

We now prove that $K(\delta)$ is bounded. To this end, let $\varphi_{\rm aff}$ be the optimal affine algorithm for $\delta = 0$. Then $\varphi_{\rm lin}(y) = \varphi_{\rm aff}(y) - \varphi_{\rm aff}(0)$ is a linear functional whose norm

$$\|\varphi_{\rm lin}\|_Y \;=\; \sup_{\|x\|_Y\le 1} |\varphi_{\rm lin}(x)|$$

is finite. For any $f \in E$ and $y \in \mathbb{N}(f)$ we have

$$\begin{aligned} |\,S(f)-\varphi_{\rm aff}(y)\,| \;&\le\; |\,S(f)-\varphi_{\rm aff}(N(f)\,)\,| \,+\, |\,\varphi_{\rm aff}(y)-\varphi_{\rm aff}(N(f)\,)\,| \\ &\le\; r(0) \,+\, \delta\,\|\varphi_{\rm lin}\|_Y. \end{aligned}$$

Taking the supremum over f and y we get $K(\delta) \le \|\varphi_{\rm lin}\|_Y$. □

Observe now that if $r'(0^+) = 0$ then $r(\delta) \equiv$ const. This means that information is useless, $r(\delta) = \sup\{\, S(h) \mid h \in {\rm bal}(E)\,\}$, and the optimal algorithm is constant. This and Lemma 2.3 yield the following theorem.

Theorem 2.5 *For an arbitrary linear functional S and the noise bounded uniformly in a norm by δ, the radius $r(\delta)$ of noisy information is either constant or converges to the radius $r(0)$ of exact information linearly in $\delta \to 0^+$, i.e.,*

$$r(\delta) \;=\; r(0) \,+\, \delta\cdot r'(0^+) \,+\, o(\delta).$$

2.4.2 The case of Hilbert noise

We now construct all optimal affine algorithms for an important class of problems. Namely, we assume that the set E is convex and information is linear with noise uniformly bounded in a Hilbert norm, i.e.,

$$\mathbb{N}(f) \;=\; \{\, y\in\mathbb{R}^n \mid \quad \|y-N(f)\|_Y \le \delta\,\} \tag{2.9}$$

where $\delta > 0$ and the norm $\|\cdot\|_Y$ is induced by an inner product $\langle\cdot,\cdot\rangle_Y$. Clearly, in this case the graph ${\rm gr}(\mathbb{N},E)$ is convex and an optimal affine algorithm exists.

We also assume that the radius $r = {\rm rad}^{\rm wor}(\mathbb{N})$ is finite and is attained. That is, there exists $h^* = (f_1^*-f_{-1}^*)/2 \in {\rm bal}(E)$ with $f_{-1}^*, f_1^* \in E$, such that $\|N(h^*)\|_Y \le \delta$ and $r = S(h^*)$. We shall see later that the latter assumption is not restrictive.

For two elements $f_{-1}, f_1 \in F$, let $I = I(f_{-1},f_1)$ denote the interval $I = \{\,\alpha f_{-1} + (1-\alpha)f_1 \mid 0 \le \alpha \le 1\,\}$. It is clear that if $f_{-1}, f_1 \in E$

then $I(f_{-1},f_1) \subset E$ and $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N};I) \le \mathrm{rad}^{\mathrm{wor}}(\mathbb{N};E)$. Furthermore, for $I^* = I(f^*_{-1},f^*_1)$ we have

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N};E) \;=\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N};I^*)$$

(compare with E 2.5). Hence, the problem of approximating $S(f)$ for f belonging to the one dimensional subset $I^* \subset E$ is as difficult as the original problem of approximating $S(f)$ for $f \in E$. We shall say, for brevity, that I^* is the *hardest one–dimensional subproblem* contained in the original problem E. In particular, we have that any algorithm optimal for E is also optimal for I^*.

The latter observation yields a method of finding *all* optimal affine algorithms. Namely, it suffices to find all such algorithms for I^* and then check which of them do not increase the error when taken over the whole set E. In the sequel, we follow this approach.

Observe first that if $\|N(h^*)\|_Y < \delta$ then the only optimal affine algorithm is constant, $\varphi(y) = S(f_0)$ where $f_0 = (f^*_1 + f^*_{-1})/2$. Indeed, let $y = N(f_0)+x$ where $\|x\|_Y \le \delta - \|N(h^*)\|_Y$. Then y is noisy information for any $f \in I^*$ and therefore $\varphi_{\mathrm{aff}}(y) = S(f_0)$. Hence, φ_{aff} is constant on a nontrivial ball. Its unique affine extension on $\mathbb{R}^n$ is $\varphi_{\mathrm{aff}} \equiv S(f_0)$.

In what follows, we assume that $\|N(h^*)\|_Y = \delta$.

Lemma 2.4 *For the hardest one–dimensional subproblem $I^* = [f^*_{-1}, f^*_1]$, all optimal affine algorithms are given as*

$$\varphi_{\mathrm{aff}}(y) \;=\; S(f_0) \,+\, d\cdot\langle\, y - N(f_0), w\,\rangle_Y, \tag{2.10}$$

where $w = N(h^*)/\|N(h^*)\|_Y$ and $d = c\,r/\delta$, for any $c \in [0,1]$.

Proof Let $y_0 = N(f_0)$ and $w^* = N(h^*)$. For $y_\alpha = y_0 + \alpha w^*$, $\alpha \in \mathbb{R}$, the set of all elements which are in the interval $S(I^*)$ and cannot be distinguished with respect to information y_α is given by $S(I^*) \cap B(S(f_0) + \alpha r, r)$, where $B(a,\tau)$ is the ball with center a and radius τ. From this it follows that for any optimal affine algorithm φ_{aff} we have

$$\varphi_{\mathrm{aff}}(y_\alpha) \;=\; S(f_0) \,+\, c\,\alpha\, r \tag{2.11}$$

where $0 \le c \le 1$. Since $\alpha = \langle y_\alpha - y_0, w\rangle_Y/\delta$, (2.11) can be rewritten as

$$\varphi_{\mathrm{aff}}(y_\alpha) \;=\; S(f_0) \,+\, c\cdot\frac{r}{\delta}\cdot\langle\, y_\alpha - y_0, w\,\rangle_Y. \tag{2.12}$$

We now show that for any $c \in [0,1]$, the formula (2.12) is valid not only for y_α, but for all $y \in \mathbb{R}^n$. To this end, it is enough to show that for any $y = y_0 + x$, where $\|x\|_Y \le \delta$, $\langle x, w\rangle_Y = 0$, we have $\varphi_{\mathrm{aff}}(y) = \varphi_{\mathrm{aff}}(y_0) = S(f_0)$. Indeed, let $\varphi_{\mathrm{aff}}(y) = S(f_0) + a$, where (without loss of generality) $a > 0$. Then $\varphi_{\mathrm{aff}}(y_0 + \varepsilon x) = S(f_0) + \varepsilon a$. Since $y_0 + \varepsilon x$ is noisy information for $f_\varepsilon = f_0 - h^* \sqrt{1 - \varepsilon^2 \|x\|_Y^2 / \delta^2}$, we obtain

$$\begin{aligned} \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{aff}}; I^*) &\ge \varphi_{\mathrm{aff}}(y_0 + \varepsilon x) - S(f_\varepsilon) \\ &= \varepsilon a + r\sqrt{1 - \varepsilon^2 \|x\|_Y^2/\delta^2}. \end{aligned}$$

For small $\varepsilon > 0$, the last expression is greater than r, which contradicts the assumption that the algorithm φ_{aff} is optimal. This completes the proof. □

The question now is as follows: for what values of d the affine algorithm (2.10) (which is optimal for the hardest one–dimensional subproblem I^*) is optimal for the original problem E?

To give an answer, we first evaluate the error $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{aff}}; E)$ of the algorithm (2.10). For any $f \in E$ and $y = N(f) + x \in \mathbb{N}(f)$, we have

$$\begin{aligned} S(f) - \varphi_{\mathrm{aff}}(y) &= S(f) - S(f_0) - d\,\langle N(f) - y_0, w\rangle_Y - d\,\langle x, w\rangle_Y \\ &= S(f) - \varphi_{\mathrm{aff}}(N(f)) - d\,\langle x, w\rangle_Y. \end{aligned}$$

Hence,

$$\sup_{\|x\|_Y \le \delta} |S(f) - \varphi_{\mathrm{aff}}(y)| = |S(f) - \varphi_{\mathrm{aff}}(N(f))| + d\,\delta. \qquad (2.13)$$

We also have

$$S(f_1^*) - \varphi_{\mathrm{aff}}(N(f_1^*)) = -(\,S(f_{-1}^*) - \varphi_{\mathrm{aff}}(N(f_{-1}^*))\,) = r - d\delta. \qquad (2.14)$$

From (2.13) and (2.14) it follows that the necessary and sufficient condition for the algorithm (2.10) to be optimal for the set E is that for all $f \in E$

$$S(f_{-1}^*) - \varphi_{\mathrm{aff}}(N(f_{-1}^*)) \le S(f) - \varphi_{\mathrm{aff}}(N(f)) \le S(f_1^*) - \varphi_{\mathrm{aff}}(N(f_1^*)).$$

Using the formula for φ_{aff} these two inequalities can be rewritten as

$$\begin{aligned} S(f_1^*) - S(f) &\ge d \cdot \langle N(f_1^*) - N(f), w\rangle_Y, &(2.15)\\ S(f_{-1}^*) - S(f) &\le d \cdot \langle N(f_{-1}^*) - N(f), w\rangle_Y. &(2.16) \end{aligned}$$

We now show that (2.15) and (2.16) are equivalent to

$$S(h^*) \,-\, S(h) \;\geq\; d\cdot\langle\, N(h^*) - N(h), w\,\rangle_Y, \qquad \forall\, h \in \mathrm{bal}(E). \tag{2.17}$$

Indeed, let (2.15) and (2.16) hold. Then, for any $h = (f_1 - f_{-1})/2$, $f_i \in E$, we have

$$\begin{aligned} S(h^*) - S(h) &= \frac{1}{2}\left((S(f_1^*) - S(f_1)) \,-\, (S(f_{-1}^*) - S(f_{-1}))\right) \\ &\geq \frac{1}{2}d\,\left(\langle N(f_1^* - f_1), w\rangle_Y - \langle N(f_{-1}^* - f_{-1}), w\rangle_Y\right) \\ &= d\,\langle\, N(h^*) - N(h), w\,\rangle_Y. \end{aligned}$$

Suppose now that (2.17) holds. Let $f \in E$. Then, for $h = (f - f_{-1}^*)/2 \in \mathrm{bal}(E)$ we have

$$\begin{aligned} S(f_1^*) - S(f) &= 2\,(\,S(h^*) - S(h)\,) \;\geq\; 2d\,\langle N(h^*) - N(h), w\rangle_Y \\ &= d\,\langle\, N(f_1^*) - N(f), w\,\rangle_Y \end{aligned}$$

which shows (2.15). Similarly, taking $h = (f_1^* - f)/2$ we obtain (2.16).

Thus the number d should be chosen in such a way that (2.17) holds. This condition has a nice geometrical interpretation. Namely, for $\gamma > 0$, let

$$r(\gamma) \;=\; \sup\{\, S(h) \mid \quad h \in \mathrm{bal}(E),\; \|N(h)\|_Y \leq \gamma\,\}$$

be the radius of information $\mathbb{N}$ with the noise level δ replaced by γ.

Lemma 2.5 *The condition (2.17) holds if and only if the line with the slope d passing through $(\delta, r(\delta)\,)$ lies above the graph of $r(\gamma)$, i.e.,*

$$r(\gamma) \;\leq\; r(\delta) \,+\, d\,(\gamma - \delta), \qquad \forall\, \gamma > 0. \tag{2.18}$$

Proof Observe first that (2.18) can be rewritten as

$$S(h^*) - S(h) \;\geq\; d\,(\,\|N(h^*)\|_Y - \|N(h)\|_Y\,), \qquad \forall\, h \in \mathrm{bal}(E). \tag{2.19}$$

Indeed, if (2.18) holds then for any $h \in \mathrm{bal}(E)$, $\gamma = \|N(h)\|_Y$, we have

$$\begin{aligned} S(h^*) - S(h) &\geq r(\delta) - r(\gamma) \;\geq\; d\,(\delta - \gamma) \\ &= d\,(\|N(h^*)\|_Y - \|N(h)\|_Y). \end{aligned}$$

Let (2.19) holds. Then for any $\gamma > 0$ and $\varepsilon > 0$ there is $h_\varepsilon \in \mathrm{bal}(E)$ such that $\|N(h_\varepsilon)\|_Y \le \gamma$ and $S(h_\varepsilon) \ge r(\gamma) - \varepsilon$. Hence,

$$\begin{aligned} r(\delta) &= S(h^*) \ \ge\ S(h_\varepsilon) + d\,(\|N(h^*)\|_Y - \|N(h_\varepsilon)\|_Y) \\ &\ge\ r(\gamma) - \varepsilon + d\,(\delta - \gamma). \end{aligned}$$

Letting $\varepsilon \to 0^+$ we get (2.18).

Thus, it suffices to show that (2.17) is equivalent to (2.19). Indeed, since

$$\begin{aligned} \langle\, N(h^*) - N(h), w\,\rangle_Y &= \|N(h^*)\|_Y - \frac{\langle N(h), N(h^*)\rangle_Y}{\|N(h^*)\|_Y} \\ &\ge \|N(h^*)\|_Y - \|N(h)\|_Y, \end{aligned}$$

the condition (2.17) implies (2.19).

We now show that (2.17) follows from (2.19). Let $h \in \mathrm{bal}(E)$,

$$S(h^*) - S(h) \ =\ d\,\langle\, N(h^*) - N(h), w\,\rangle_Y + a. \tag{2.20}$$

For $0 < \tau \le 1$, let $h_\tau = (1-\tau)h^* + \tau h = h^* - \tau(h^* - h)$. Then $h_\tau \in \mathrm{bal}(E)$ and from (2.20) we have

$$S(h^*) - S(h_\tau) \ =\ \tau\,(S(h^*) - S(h)) \ =\ \tau\, d\,\langle\, N(h^*) - N(h), w\,\rangle_Y + \tau\, a. \tag{2.21}$$

We also have

$$\begin{aligned} \|N(h_\tau)\|_Y^2 &= \|N(h^*) - \tau(\,N(h^*) - N(h)\,)\|_Y^2 \\ &= (\,\|N(h^*)\|_Y - \tau\,\langle\, N(h^*) - N(h), w\,\rangle_Y\,)^2 + O(\tau^2), \end{aligned}$$

as $\tau \to 0^+$. Hence,

$$\|N(h^*)\|_Y - \|N(h_\tau)\|_Y \ =\ \tau\,\langle\, N(h^*) - N(h), w\,\rangle_Y + O(\tau^2). \tag{2.22}$$

Combining (2.21) and (2.22) with (2.19) we now obtain $\tau\, a \ge O(\tau^2)$, which means that a is nonnegative. This together with (2.20) proves (2.17). □

We summarize our analysis in the following theorem.

Theorem 2.6 *Let $\mathbb{N}$ be information (2.9) with the noise level $\delta > 0$. Let $h^* = (f_1^* - f_{-1}^*)/2$, $f_1^*, f_{-1}^* \in E$, be such an element that*

$$S(h^*) \ =\ \sup\{\, S(h) \mid \quad h \in \mathrm{bal}(E),\ \|N(h)\| \le \delta\,\}.$$

Then all optimal affine algorithms are given by

$$\varphi_{\mathrm{aff}}(y) \;=\; g_0 \,+\, d \cdot \langle\, y - y_0, w \,\rangle_Y,$$

where $g_0 = S(f_1^* + f_{-1}^*)/2$, $y_0 = N(f_1^* + f_{-1}^*)/2$, $w = N(h^*)/\|N(h^*)\|_Y$ *(or* $w = 0$ *for* $N(h^*) = 0$*), and* d *satisfies*

$$r(\gamma) \le r(\delta) + d(\gamma - \delta), \qquad \forall\, \gamma \ge 0. \quad \Box$$

We stress that Theorem 2.6 gives *all* optimal affine algorithms. In particular, another choice of h^* leads to the same optimal affine algorithms. Note also that if $\|N(h^*)\|_Y < \delta$ then $d = 0$ and $\varphi_{\mathrm{aff}} \equiv S(f_0)$.

We now briefly discuss the case when the hardest one–dimensional subproblem does not exist. Then we can extract a sequence $\{h_i\} \subset \mathrm{bal}(E)$ such that $\|N(h_i)\|_Y \le \delta$, $\forall i$, and $\lim_{i\to\infty} S(h_i) = r(\delta)$. As $\{N(h_i)\}$ is a bounded set of $\mathbb{R}^n$, it has an attraction point w^*.

Suppose first that $\|w^*\|_Y = \delta$. In this case we let $w = w^*/\delta$ and d as in Theorem 2.6. Using the technique from the proof of Lemma 2.5 and some approximation arguments, we can show that for all $h \in \mathrm{bal}(E)$,

$$r(\delta) \,-\, S(h) \;\ge\; d \cdot (\delta - \langle N(h), w\rangle_Y), \quad \forall\, h \in \mathrm{bal}(E),$$

which corresponds to inequality (2.17). Hence, $S(h) - d\langle N(h), w\rangle_Y \le r(\delta) - \delta d$, or equivalently,

$$S(f_{-1}) \,-\, d\,\langle N(f_{-1}), w\rangle_Y \,-\, r(\delta) \;\le\; S(f_1) \,-\, d\,\langle N(f_1), w\rangle_Y \,+\, r(\delta),$$

for all $f_{-1}, f_1 \in E$. Letting $g = \sup_{f\in E} S(f) - d\langle N(f), w\rangle_Y - r(\delta)$, we obtain $|S(f) - d\langle y, w\rangle_Y - g| \le r(\delta)$, $\forall\, f \in E$, $y \in \mathbb{N}(f)$. This means that the algorithm

$$\varphi_{\mathrm{aff}}(y) \;=\; g \,+\, d \cdot \langle y, w\rangle_Y$$

is optimal.

If $\|w^*\|_Y < \delta$ then $r(\gamma)$ is constant for $\gamma > \|w^*\|_Y$. Hence, the optimal affine algorithm is also a constant, $\varphi_{\mathrm{aff}} \equiv \sup_{f\in E} S(f) - r(\delta)$.

So far we have not covered the case $\delta = 0$. It turns out, however, that exact information can be treated as the limiting case. Indeed, let $\varphi_\delta = g_\delta + d_\delta\langle\cdot, w_\delta\rangle_Y$ be the optimal affine algorithm for $\delta > 0$. Let w_0 be an attraction point of $\{w_\delta\}$ as $\delta \to 0^+$. As $\lim_{\delta\to 0} d_\delta = r'(0^+)$ and

$$S(h) \,-\, d_\delta\langle N(h), w_\delta\rangle_Y \;\le\; r(\delta) \,-\, \delta\, d_\delta, \quad \forall\, h \in \mathrm{bal}(E),$$

letting $\delta \to 0^+$ we obtain

$$S(h) \,-\, r'(0^+)\langle N(h), w_0\rangle_Y \;\le\; r(0), \qquad \forall\, h \in \mathrm{bal}(E).$$

Hence, for $g_0 = \sup_{f\in E} S(f) - r'(0^+)\langle N(f), w_0\rangle_Y - r(0)$ we have $|S(f) - r'(0^+)\langle N(f), w_0\rangle_Y - g_0| \le r(0)$, $\forall f \in E$, and the algorithm

$$\varphi_0(y) \;=\; r'(0^+)\cdot\langle y, w_0\rangle_Y \,+\, g_0$$

is optimal. (See also E 2.17 for another construction.)

We end this section by a simple illustration of Theorem 2.6.

Example 2.8 Let F be a linear space of Lipschitz functions $f : [0,1] \to \mathbb{R}$ that satisfy $f(0) = f(1)$. Let

$$E \;=\; \{\, f \in F \mid \quad |f(x_1) - f(x_2)| \le |x_1 - x_2|,\ \forall\, x_1, x_2 \,\}.$$

We want to approximate the integral of f, i.e.,

$$S(f) \;=\; \int_0^1 f(t)\,dt.$$

Noisy information is given by perturbed evaluations of function values at equidistant points, $y = [y_1, \ldots, y_n] \in \mathbb{R}^n$, where $y_i \;=\; f(i/n) + x_i$, $1 \le i \le n$, and the noise $\|x\|_2 \;=\; (\,\sum_{i=1}^n x_i^2\,)^{1/2} \;\le\; \delta$.

Since S is a functional and the set E is convex and balanced, Theorem 2.3 yields

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sup\left\{ \int_0^1 f(t)\,dt \;\middle|\; \quad f \in E, \quad \sum_{i=1}^n f^2(i/n) \;\le \delta^2 \right\}.$$

The last supremum is attained for

$$h^*(t) \;=\; \frac{\delta}{\sqrt{n}} + \frac{1}{2n} - \left|t - \frac{2i-1}{2n}\right|, \qquad \frac{i-1}{n} \le t \le \frac{i}{n},\ 1 \le i \le n,$$

and

$$r(\delta) \;=\; \frac{\delta}{\sqrt{n}} + \frac{1}{4n}$$

(compare also with Theorem 2.5. Hence, $w \;=\; (1, 1, \ldots, 1)/\sqrt{n}$ and $d \;=\; n^{-1/2}$. The unique optimal linear algorithm is the well known arithmetic mean

$$\varphi_{\mathrm{lin}}(y) \;=\; \frac{1}{n}\sum_{i=1}^n y_i.$$

Note that in this case optimal linear algorithm is independent of the noise level δ. However, its error does depend on δ.

Notes and Remarks

NR 2.7 The problem of existing optimal linear or affine algorithms for approximating linear functionals has a long history. The first positive result on this subject is due to Smolyak [96] who considered the exact information case and convex and balanced set E; see also Bakhvalov [4]. His results was then generalized by Sukharev [101] to the case of only convex set E. Noisy case was considered by, e.g., Marchuk and Osipenko [55], Micchelli and Rivlin [59]. The proof of Theorem 2.4 is taken from Magaril–Il'yaev and Osipenko [52] where even a more general result is given; see E 2.11.

NR 2.8 We want to stress that Theorem 2.4 does not hold when the solution operator S is linear but not a functional. Examples (for exact information) are provided by Micchelli and Rivlin [59], Packel [69], Werschulz and Woźniakowski [125]; see also Traub *et al.*[108, Sect.5.5 of Chap.4].

NR 2.9 The dependence of the radius on the noise level δ was studied in Kacewicz and Kowalski [29] for the solution operator S being a functional, and in Kacewicz and Kowalski [30] for arbitrary linear S. They showed, in particular, that if E is the unit ball with respect to a Hilbert seminorm and S is a functional, then $r(\delta) = r(0) + \delta\,\|\varphi_{\mathrm{lin}}\|_Y + o(\delta)$ where φ_{lin} is as in the proof of Lemma 2.3. The general result of Theorem 2.5 seems however to be new.

NR 2.10 Optimality of the affine algorithms defined in Theorem 2.6 was shown by Donoho [12]. The idea of using in the proof the hardest one–dimensional subproblems belongs to him. We additionally showed that those are *all* optimal affine algorithms. The results in the case when the radius is not attained as well as the formulas for the optimal affine algorithm in exact information case are new.

NR 2.11 Optimal algorithms for noise bounded in the uniform norm rather than in the Hilbert norm are in general unknown. We mention one special result which has been recently obtained by Osipenko [68].

Let F be a separable Hilbert space with a complete orthonormal system $\{e_i\}_{i\ge 1}$. For $f \in F$, let $f_j = \langle f, e_j\rangle_F$ be the ith Fourier coefficient of f. Consider the problem of approximating a functional $S = \langle \cdot, s\rangle_F$ for f from the unit ball of F, based on noisy values of the Fourier coefficients, $y_i = f_i + x_i$, where $|x_i| \le \delta_i$, $1 \le i \le n$. Osipenko showed, in particular, that the optimal linear algorithm φ_{opt} is in this case given as follows. Let $\lambda \in (0, \|s\|_F]$ be the (existing) solution of

$$\|s\|_F^2 - \sum_{j=1}^n (|s_j|^2 - \lambda^2\delta_j^2)_+ - \lambda^2 \;=\; 0.$$

Then

$$\varphi_{\mathrm{opt}}(y) \;=\; \sum_{j=1}^n (1 - \lambda\delta_j|s_j|^{-1})_+ s_j y_j$$

and the radius equals

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \lambda + \sum_{j=1}^{n} \delta_j(|s_j| - \lambda\delta_j)_+ .$$

Exercises

E 2.10 Show that an optimal affine (linear) algorithm for a functional S exists if the set

$$\tilde{A} \;=\; \{\, (y, S(f)\,) \mid \quad y \in \mathbb{N}(f),\; f \in E\,\}$$

is convex (convex and balanced).

E 2.11 (Magaril–Il'yaev and Osipenko) Let $\mathrm{c}(A)$ $(\,\mathrm{cb}(A)\,)$ be the smallest convex (convex and balanced) set which contains A. Show that an optimal affine (linear) algorithm exists iff

$$\mathrm{rad}^{\mathrm{wor}}(\,c(\mathbb{N}), c(E)\,) = \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E) \qquad (\,\mathrm{rad}^{\mathrm{wor}}(\,cb(\mathbb{N}), cb(E)\,) = \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E)\,).$$

E 2.12 Suppose that the radius is attained for two elements $h_1^*, h_2^* \in \mathrm{bal}(E)$ such that $N(h_1^*) \neq N(h_2^*)$. Show that then the only optimal affine algorithm is constant, $\varphi_{\mathrm{aff}} \equiv (\sup_{f\in E} S(f) + \inf_{f\in E} S(f))/2$.

Use this result to show the formula for h^* in Example 2.8.

E 2.13 Consider the problem of estimating a real parameter f from the interval $I = [-\tau, \tau] \subset \mathbb{R}$, based on the data y such that $|y - f| \le \delta$. Show that in this case the radius is equal to $\min\{\tau, \delta\}$ and the optimal affine algorithm is given as

$$\varphi_{\mathrm{aff}}(y) \quad = \quad \left\{ \begin{array}{lll} y & \delta < \tau, & \\ c\,y & \delta = \tau & (\,0 \le c \le 1\,), \\ 0 & \delta > \tau. & \end{array} \right.$$

E 2.14 Let $\mathbb{N}$ be linear information with noise bounded uniformly by $\delta \ge 0$ in a Hilbert space norm. Let $f_{-1}, f_1 \in F$ be such that $\|N(f_1 - f_{-1})\|_Y > 2\delta$. Show that for the interval $I = [f_{-1}, f_1]$ we have

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; I) \;=\; \frac{|S(f_1) - S(f_{-1})|}{\|N(f_1) - N(f_{-1})\|_Y} \cdot \delta$$

and the only optimal affine algorithm is given as

$$\varphi_{\mathrm{aff}}(y) \;=\; S(f_0) \;+\; \frac{S(f_1) - S(f_{-1})}{\|N(f_1) - N(f_{-1})\|_Y} \cdot \langle\, y - N(f_0), w\,\rangle_Y ,$$

where $f_0 = (f_{-1} + f_1)/2$ and $w \;=\; (\,N(f_1) - N(f_{-1})\,)/\|N(f_1) - N(f_{-1})\|_Y$.

E 2.15 Let E be a convex set. Show that the radius

$$r(\delta) \;=\; \sup\{\, S(h) \mid \quad h\in \mathrm{bal}(E),\; \|N(h)\|_Y \le \delta\,\}$$

is a concave and subadditive function of δ.

E 2.16 Why does the number d in Theorem 2.6 exist? For $0\le a\le b<+\infty$, give an example where the set of all such d forms the closed interval $[a,b]$.

E 2.17 (Bakhvalov) Let E be a convex and balanced set. Consider approximation of a linear functional S for $f\in E$, based on *exact* linear information $y = [L_1(f),\ldots,L_n(f)]$ where the functionals L_i are linearly independent on span E. Let

$$r_k(x) \;=\; \sup\,\{\, S(h) \mid \quad h\in E,\; L_k(h)=x,\; L_j(h)=0,\; i\ne k\,\}.$$

Assuming that the derivatives $r_k'(0)$ exist for all $1\le k\le n$, show that the algorithm $\varphi(y)=\sum_{j=1}^n r_j'(0)\,y_j$ is a unique optimal linear.

E 2.18 Show that if the solution operator S is linear but not a functional then the assertion of Thorem 2.5 in no longer true.

E 2.19 Show an example of a balanced but *not* convex set E such that for some linear functional S and some linear information $\mathbb{N}$ with noise bounded uniformly in a Hilbert space norm we have $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})=0$, but the error of any affine algorithm is infinite.

E 2.20 Find an optimal linear algorithm for the integration problem of Example 2.8 when the function values are observed at arbitrary, not necessarily equidistant, points.

2.5 Optimality of spline algorithms

In this section we assume that:

- S is an arbitrary linear operator.
- E is the unit ball in an extended seminorm $\|\cdot\|_F$,

$$E \;=\; \{\, f\in F \mid \quad \|f\|_F\le 1\,\}.$$

- Information is linear with uniformly bounded noise,

$$\mathbb{N}(f) \;=\; \{\, y\in\mathbb{R}^n \mid \quad \|y-N(f)\|_Y\le\delta\,\}, \qquad \delta>0.$$

As explained in Section 2.2, the second assumption is roughly equivalent to the fact that E is a convex and balanced set. The assumption $\delta > 0$ is not restrictive. If $\delta = 0$ then changing the extended seminorm $\|\cdot\|_Y$ properly we can make δ positive.

Due to Theorem 2.3, in this case we have

$$\mathrm{diam}(\mathbb{N}) \;=\; 2\cdot\sup\{\, \|S(h)\| \mid \quad \|h\|_F \le 1,\ \|N(h)\|_Y \le \delta \,\}. \tag{2.23}$$

Optimal or almost optimal algorithms can be now constructed using the so–called splines. We shall see that sometimes spline algorithms turn out to be not only optimal but also linear.

2.5.1 Splines and smoothing splines

Let $\rho \ge 1$. For information $y \in \{\, N(f)+x \mid f \in F,\ \|x\|_Y \le \delta \,\}$, an *(ordinary) spline* is an element $\mathbf{s}_o(y) \in F$ defined by the following two conditions:

1. $y \in \mathbb{N}(\mathbf{s}_o(y)\,)$,
2. $\|\mathbf{s}_o(y)\|_F \;\le\; \rho\cdot\inf\{\, \|f\|_F \mid \quad y \in \mathbb{N}(f)\,\}$.

Hence, $\mathbf{s}_o(y)$ is the element whose extended seminorm does not exceed ρ times the minimal value of $\|f\|_F$ among all f that share the same information y. Note that for $\rho > 1$, the spline $\mathbf{s}_o(y)$ always exists, but it is not determined uniquely.

An *(ordinary) spline algorithm* is given as

$$\varphi_o(y) \;=\; S(\mathbf{s}_o(y)\,).$$

Theorem 2.7 *For the spline algorithm φ_o, we have*

$$\|S(f)-\varphi_o(y)\| \;\le\; c(f)\cdot\mathrm{diam}(\mathbb{N}), \qquad \forall f \in F,\ \forall y \in \mathbb{N}(f),$$

where $c(f) \;=\; \max\{\, 1,\ \frac{1+\rho}{2}\,\|f\|_F\,\}$. *Hence,*

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi_o) \;\le\; \frac{1+\rho}{2}\cdot\mathrm{diam}(\mathbb{N}).$$

Proof For $f \in F$ and information y such that $\|y - N(f)\|_Y \le \delta$, we have $\|N(f-\mathbf{s}_o(y))\|_Y \;\le\; \|N(f)-y\|_Y + \|y - N(\mathbf{s}_o(y))\|_Y \;\le\; 2\,\delta$. Hence, for $\|f-\mathbf{s}_o(y)\|_F \le 2$, we get from (2.23)

$$\|S(f)-\varphi_o(y)\| \;=\; \|S(f-\mathbf{s}_o(y)\,)\| \;\le\; \mathrm{diam}(\mathbb{N}).$$

On the other hand, for $\|f - \mathbf{s}_o(y)\|_F > 2$, we have

$$\begin{aligned}\|S(f) - \varphi_o(y)\| &= \|f - \mathbf{s}_o(y)\|_F \cdot \left\| S\left(\frac{f - \mathbf{s}_o(y)}{\|f - \mathbf{s}_o(y)\|_F}\right)\right\| \\ &\leq \frac{1}{2}\,(\,\|f\|_F + \|\mathbf{s}_o(y)\|_F) \cdot \mathrm{diam}(\mathbb{N}) \\ &\leq \frac{1+\rho}{2}\,\|f\|_F \cdot \mathrm{diam}(\mathbb{N}).\end{aligned}$$

Combining both cases and the fact that $f \in E$ implies $\|f\|_F \leq 1$, we obtain the theorem. □

Thus the error of the spline algorithm with $\rho \cong 1$ is, roughly speaking, at most twice as large as the optimal error. This is not very surprising since φ_o is close to the interpolatory algorithm. For an arbitrary ρ, an additional advantage of φ_o is that it is the spline algorithm for any set $E = \{\, f \in F |\ \|f\|_F \leq b\,\}$, $b > 0$. Indeed, the definition of φ_o is independent on b. Hence, φ_o preserves almost optimal properties for any such a set. Unfortunately, as illustrated below, the ordinary spline algorithm is usually not linear, even when $\|\cdot\|_F$ and $\|\cdot\|_Y$ are Hilbert space norms.

Example 2.9 Consider the problem of approximating a real parameter $f \in E = [-a, a]$ from information $y \in \mathbb{N}(f) = \{\, f + x \,|\ |x| \leq \delta\,\}$. Then the ordinary spline algorithm with $\rho = 1$ is given as

$$\varphi_o(y) \;=\; \begin{cases} y - \delta & y > \delta, \\ 0 & |y| \leq \delta, \\ y + \delta & y < -\delta. \end{cases}$$

For $\delta > 0$, this is *not* a linear algorithm. □

We now turn to smoothing spline algorithms. The idea is to minimize not only the norm of f in the definition of a spline element, but also the noise $y - N(f)$. In general, a *smoothing spline algorithm* φ_* is given in the following way.

Let $\|(\cdot,\cdot)\|_*$ be an extended seminorm in the linear space $F \times \mathbb{R}^n$, and let $\rho \geq 1$. A *smoothing spline* is an element $\mathbf{s}_*(y) \in F$ satisfying

$$\|\,(\mathbf{s}_*(y), y - N(\mathbf{s}_*(y))\,)\,\|_* \;\leq\; \rho \cdot \inf_{f \in F}\, \|\,(f, y - N(f)\,)\,\|_*.$$

Then

$$\varphi_*(y) \;=\; S(\mathbf{s}_*(y))$$

is a *smoothing spline algorithm.*

Consider first the case when the extended seminorm $\|(\cdot,\cdot)\|_*$ is given as

$$\|(f,x)\|_* \;=\; \max\{\,\|f\|_F, \|x\|_Y/\delta\,\}, \qquad f\in F,\ x\in\mathbb{R}^n.$$

In this case, we write $\|(\cdot,\cdot)\|_* = \|(\cdot,\cdot)\|_\infty$ and $\varphi_* = \varphi_\infty$.

Theorem 2.8 *For the smoothing spline algorithm φ_∞, we have*

$$\|S(f)-\varphi_\infty(y)\| \;\le\; \frac{1+\rho}{2}\cdot\max\{\,\|f\|_F, \|y-N(f)\|_Y/\delta\,\}\cdot\mathrm{diam}(\mathbb{N}),$$

for all $f\in F$ and $y\in\mathbb{N}(f)$. Hence,

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi_\infty) \;\le\; \frac{1+\rho}{2}\cdot\mathrm{diam}(\mathbb{N}).$$

Proof Let $f\in F$ and $y\in\mathbb{R}^n$ be such that $\|y-N(f)\|_Y\le\delta$. Consider first the case when $\|f-\mathbf{s}_\infty(y)\|_F=0$ and $\|N(f-\mathbf{s}_\infty(y)\,)\|_Y=0$. Then for any c we have $f_c \;=\; c\,(f-\mathbf{s}_\infty(y)\,)\in E$ and zero is noisy information for f_c. Since also $\|S(f_c)-\varphi_\infty(0)\| \;=\; |c|\cdot\|S(f-\mathbf{s}_\infty(y)\,)\|$, we obtain $\|S(f)-\varphi_\infty(y)\| \;=\; 0$, or $\mathrm{diam}(\mathbb{N}) \;=\; +\infty$. In both cases the theorem holds.

Assume now that $\max\{\,\|f-\mathbf{s}_\infty(y)\|_F,\ \|N(f-\mathbf{s}_\infty(y)\,)\|_Y\}>0$. Then

$$\begin{aligned}
\|S(f)-\varphi_\infty(y)\| \;&=\; \|(f-\mathbf{s}_\infty(y), N(f-\mathbf{s}_\infty(y))\,)\|_\infty \\
&\quad\cdot\left\|S\left(\frac{f-\mathbf{s}_\infty(y)}{\|(f-\mathbf{s}_\infty(y), N(f-\mathbf{s}_\infty(y))\,)\|_\infty}\right)\right\| \\
&\le\; (\|(f,y-N(f))\|_\infty + \|(\mathbf{s}_\infty(y), y-N(\mathbf{s}_\infty(y)))\|_\infty) \\
&\quad\cdot\sup_{\|(h,N(h)\,)\|_\infty\le 1}\|S(h)\| \\
&\le\; \frac{1+\rho}{2}\,\|(f,y-N(f)\,)\|_\infty\,\mathrm{diam}(N).
\end{aligned}$$

Since for $f\in E$ and $y\in\mathbb{N}(f)$ we have $\max\{\|f\|_F, \|y-\mathbb{N}(f)\|_Y/\delta\}\;\le 1$, the smoothing spline algorithm φ_∞ is almost interpolatory and the upper bounds on the errors $e(\mathbb{N},\varphi_\infty)$ and $e(\mathbb{N},\varphi_o)$ are the same. This completes the proof. $\square$

The advantage of the smoothing spline algorithms is that in some cases the extended seminorm $\|(\cdot,\cdot)\|_*$ can be chosen in such a way that φ_* becomes not only (almost) optimal, but also linear. This holds when F and Y are Hilbert type spaces. Because of its importance, we devote a special attention to this case.

2.5.2 α–smoothing splines

We now additionally assume that

- $\|\cdot\|_F$ and $\|\cdot\|_Y$ are Hilbert extended seminorms.

This means that on the linear subspaces $F' = \{\, f \in F \mid \|f\|_F < +\infty \,\}$ and $Y' = \{\, y \in \mathbb{R}^n \mid \|y\|_Y < +\infty \,\}$, the functionals $\|\cdot\|_F$ and $\|\cdot\|_Y$ are seminorms induced by some semi–inner products $\langle\cdot,\cdot\rangle_F$ and $\langle\cdot,\cdot\rangle_Y$, respectively. Moreover, F' and Y' are complete with respect to $\|\cdot\|_F$ and $\|\cdot\|_Y$.

Let $0 \le \alpha \le 1$. For $f \in F$ and $y \in \mathbb{R}^n$, define

$$\Gamma_\alpha(f,y) \;=\; \alpha \cdot \|f\|_F^2 \,+\, (1-\alpha)\cdot \delta^{-2}\|y - N(f)\|_Y^2.$$

We use above the convention $a \cdot (+\infty) = +\infty$, $\forall a \ge 0$. Observe that $\Gamma_\alpha(f,y)$ represents a trade–off between the seminorm of f and fidelity of $N(f)$ to the data y. This trade–off is controlled by the parameter α. Let

$$\Gamma_\alpha(y) \;=\; \inf_{f\in F} \Gamma_\alpha(f,y).$$

Then the set Y_1 of all y for which $\Gamma_\alpha(y) < +\infty$ is a linear subspace of $\mathbb{R}^n$, and

$$Y_1 \;=\; \{\, N(f) + x \mid \quad \|f\|_F < +\infty,\ \|x\|_Y < +\infty \,\}.$$

Also, $\Gamma_\alpha(y) \le 1$ if y is noisy information for some $f \in E$, $y \in \mathbb{N}(E)$.

An *α–smoothing spline* is an element $\mathbf{s}_\alpha(y) \in F$ for which

$$\Gamma_\alpha(\mathbf{s}_\alpha(y), y) \;=\; \Gamma_\alpha(y).$$

Hence, the α–smoothing spline is a special instance of a smoothing spline (with $\rho = 1$) when the extended seminorm $\|(\cdot,\cdot)\|_* = \|(\cdot,\cdot)\|_\alpha$ in $F \times \mathbb{R}^n$ is induced by the semi–inner product

$$\langle\, (f_1,x_1),(f_2,x_2)\,\rangle_\alpha \;=\; \alpha\,\langle\, f_1,f_2\,\rangle_F \,+\, (1-\alpha)\,\delta^{-2}\langle\, x_1,x_2\,\rangle_Y.$$

An *α–smoothing spline algorithm* is defined as

$$\varphi_\alpha(y) \;=\; S(\mathbf{s}_\alpha(y)).$$

We first give a sufficent condition for existing and interpretation of the α–smoothing splines.

Lemma 2.6 *Assume that the operator N is closed, i.e., $\|f_i\|_F \to 0$ and $\|N(f_i) - y\|_Y \to 0$ imply $y = 0$. Then*

(i) An α–smoothing spline exists for any $y \in \mathbb{R}^n$.

(ii) Let $y \in Y_1$. Then $\mathbf{s}_\alpha(y)$ is an α–smoothing spline for y if and only if $\Gamma_\alpha(\mathbf{s}_\alpha(y), y) < +\infty$ and

$$\alpha\,\langle \mathbf{s}_\alpha(y), f\rangle_F + \,(1-\alpha)\,\delta^{-2}\langle N(\mathbf{s}_\alpha(y)\,) - y, N(f)\rangle_Y \;=\; 0, \tag{2.24}$$

for all $f \in F$ for which $\Gamma_\alpha(f, 0) < +\infty$.

(iii) For all $y \in Y_1$, the α–smoothing spline is defined uniquely if and only if

$$\Gamma_\alpha(f, N(f)) \;>\; 0, \qquad \forall\, f \neq 0.$$

(iv) There exist smoothing splines $\mathbf{s}_\alpha(y)$, such that the mapping $y \longrightarrow \mathbf{s}_\alpha(y)$, $y \in \mathbb{R}^n$, is linear.

Proof In the proof we write, for brevity, f_y instead of $\mathbf{s}_\alpha(y)$.

(i) If $\Gamma_\alpha(y) = +\infty$, any element of F is a smoothing spline. Assume thus that $\Gamma_\alpha(y)$ is finite. Then $f_y \in F$ is an α–smoothing spline if and only if $(f_y, N(f_y)\,)$ is an element of the subspace $V = \{(f, N(f)\,)\,|\, f \in F\,\} \subset F \times \mathbb{R}^n$, closest to $(0, y)$ with respect to the extended seminorm $\|\cdot\|_\alpha$. Hence, for existence of an α–smoothing spline it suffices to show that the subspace V is closed with respect to $\|(\cdot,\cdot)\|_\alpha$. Indeed, if $(f_i, N(f_i)\,) \to (f, y)$ then $\|f_i - f\|_F \to 0$ and $\|N(f_i) - y\|_Y \to 0$. Since N is closed, we obtain $y = N(f)$ which means that $(f, y) = (f, N(f)\,) \in V$.

(ii) If $\|(\cdot,\cdot)\|_\alpha$ is a Hilbert norm then the element $(f_y, N(f_y)) - (0, y)$ is orthogonal to V. This means that

$$\alpha\,\langle\, f_y, f\,\rangle_F \,+\, (1-\alpha)\,\delta^{-2}\,\langle\, N(f_y) - y, N(f)\,\rangle_Y \;=\; 0,$$

for all $f \in F$. We can easily convinced ourselves that the same holds for $\|(\cdot,\cdot)\|_\alpha$ being an extended seminorm. (The prove goes exactly as for the

Hilbert norm.) We add the condition $\Gamma_\alpha(f,0) = \|(f,N(f))\|_\alpha^2 < +\infty$ only to make the semi–inner product in (2.24) well defined.

Suppose now that $\Gamma_\alpha(f_y,y) < +\infty$ and (2.24) hold. Let f be such that $\Gamma_\alpha(f,y) < +\infty$. Then $\|(f_y,N(f_y)-y)\|_\alpha < +\infty$, $\|(f,N(f)-y)\|_\alpha < +\infty$, which forces that the element $(f-f_y,N(f-f_y)) = (f,N(f)-y) - (f_y,N(f_y)-y)$ has also finite extended seminorm $\|\cdot\|_\alpha$, or equivalently, $\Gamma_\alpha(f-f_y,0) < +\infty$. Since, in addition, this element is in V, from the orthogonality condition (2.24) we obtain

$$\begin{aligned}\Gamma_\alpha(f,y) &= \|(f,N(f)-y)\|_\alpha^2\\ &= \|(f_y,N(f_y)-y) + (f-f_y,N(f-f_y))\|_\alpha^2\\ &= \|(f_y,N(f_y)-y)\|_\alpha^2 + \|(f-f_y,N(f-f_y))\|_\alpha^2\\ &\geq \Gamma_\alpha(f_y,y).\end{aligned}$$

This means that f_y is an α–smoothing spline.

(iii) The orthogonal projection on the subspace V is determined uniquely iff $\|\cdot\|_\alpha$ is an extended norm on V. This in turn is equivalent to $\Gamma_\alpha(f,N(f)) > 0$, for $f \neq 0$, as claimed.

(iv) From (2.24) it follows that smoothing splines are linear on the subspace Y_1. That is, if $\mathbf{s}_\alpha(y_1)$, $\mathbf{s}_\alpha(y_2)$ are α–smoothing splines for $y_1,y_2 \in Y_1$ then $\gamma_1\,\mathbf{s}_\alpha(y_1)+\gamma_2\,\mathbf{s}_\alpha(y_2)$ is an α–smoothing spline for $\gamma_1\,y_1+\gamma_2\,y_2$. Hence, $\mathbf{s}_\alpha(y)$ can be chosen in such a way that the mapping $y \to \mathbf{s}_\alpha(y)$, $y \in \mathbb{R}^n$, is linear. □

We now turn to the error of the α–smoothing spline algorithm.

Lemma 2.7 *For any $f \in E$ and $y \in \mathbb{N}(f)$*

$$\begin{aligned}\|S(f)-\varphi_\alpha(y)\| \;&\leq\; \sqrt{1-\Gamma_\alpha(y)} \qquad (2.25)\\ &\quad \sup\{\,\|S(h)\| \;\mid\; \alpha\,\|h\|_F^2 + (1-\alpha)\,\delta^{-2}\|N(h)\|_Y^2 \leq 1\,\}.\end{aligned}$$

In particular, if $\alpha \in (0,1)$ then

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi_\alpha) \;\leq\; c(\alpha)\cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$$

where $c(\alpha) \;=\; \max\{\,\alpha^{-1/2},(1-\alpha)^{-1/2}\,\}$.

Proof Lemma 2.6(ii) yields

$$\begin{aligned} & \|(f,N(f))-(0,y)\|_\alpha^2 \\ = \; & \|(f,N(f))-(\mathbf{s}_\alpha(y),N(\mathbf{s}_\alpha(y))\,)\|_\alpha^2 \\ & \quad +\|(\mathbf{s}_\alpha(y),N(\mathbf{s}_\alpha(y))\,)-(0,y)\,\|_\alpha^2 \\ = \; & \alpha\,\|f-\mathbf{s}_\alpha(y)\|_F^2 \;+\;(1-\alpha)\,\delta^{-2}\|N(f-\mathbf{s}_\alpha(y))\|_Y^2 \;+\;\Gamma_\alpha(y). \end{aligned}$$

We also have

$$\|(f,N(f))-(0,y)\|_\alpha^2 \;=\; \alpha\|f\|_F^2+(1-\alpha)\delta^{-2}\|y-N(f)\|_Y^2 \;\le\; 1.$$

Hence, setting $h \;=\; f-\mathbf{s}_\alpha(y)$, we obtain

$$\alpha\,\|h\|_F^2 \;+\;(1-\alpha)\,\delta^{-2}\|N(h)\|_Y^2 \;\le\; 1\;-\;\Gamma_\alpha(y)$$

and (2.25) follows.

To show the second inequality of the lemma, observe that the condition $\alpha\|h\|_F^2+(1-\alpha)\delta^{-2}\|N(h)\|_Y^2 \;\le\; 1$ implies $\|h\|_F\le\alpha^{-1/2}$ and $\|N(h)\|_Y\le\delta\,(1-\alpha)^{-1/2}$. Hence,

$$\begin{aligned} & \sup\{\,\|S(h)\|\;|\quad \alpha\|h\|_F^2+(1-\alpha)\delta^{-2}\|N(h)\|_Y^2\le 1\} \\ \le\; & c(\alpha)\;\cdot\;\sup\{\,\|S(h)\|\;|\quad \|h\|_F\le 1,\;\|N(h)\|_Y\le\delta\,\}. \end{aligned}$$

The last supremum is equal to the half of diameter of information $\mathbb{N}$ on the set E. The lemma now follows from the fact that $(1/2\,\mathrm{diam}(\mathbb{N})\le\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$. □

Thus the error of the α-apline algorithms is at most $c(\alpha)$ larger than the minimal error. In particular, one can take $\min_\alpha c(\alpha)=c(1/2)=\sqrt{2}$. Then

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi_{1/2})\;\le\;\sqrt{2}\cdot\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}).$$

In some cases, the parameter α can be chosen is such a way that the α–smoothing spline algorithm is strictly optimal. Clearly, this is true if $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})=+\infty$. For $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})=0$, it follows from Lemma 2.7 that the algorithm φ_α is optimal for any $0<\alpha<1$.

Assume that $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})\in(0,+\infty)$. Let $\mathrm{conv}(A)$ be the convex hull of A. For $a\;=\;(a_1,a_2,\ldots,a_m)\in\mathbb{R}^m$, let $\|a\|_\infty\;=\;\max_{1\le i\le m}|a_i|$. Then we have the following theorem.

Theorem 2.9 *Let* $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \in (0,+\infty)$*. Let the set*

$$A \;=\; \Big\{ \Big(\|h\|_F^2,\, \|N(h)\|_Y^2/\delta^2 \Big) \in \mathbb{R}^2 \;\Big|\; h \in F,\, \|S(h)\| \ge 1 \Big\}$$

satisfy

$$\inf_{a\in A} \|a\|_\infty \;=\; \inf_{a\in \mathrm{conv}(A)} \|a\|_\infty. \tag{2.26}$$

Then

$$\begin{aligned} \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;&=\; \frac{1}{2}\,\mathrm{diam}(\mathbb{N}) \\ &=\; \sup\{\, \|S(h)\| \;|\; \|h\|_F \le 1,\, \|N(h)\|_Y \le \delta \,\} \end{aligned}$$

and there exists $0 \le \alpha^* \le 1$ *such that the* α^**–smoothing spline algorithm is optimal. Furthermore,*

$$\|S(f) - \varphi_{\alpha^*}(y)\| \;\le\; \sqrt{1-\Gamma_{\alpha^*}(y)}\cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}), \quad \forall f\in E,\, \forall y \in \mathbb{N}(f).$$

Proof For $r = \sup\{\, \|S(h)\| \;|\; \|h\|_F \le 1,\, \|N(h)\|_Y \le 1\}$ we have $r \in (0,+\infty)$. Since $r = 0.5\cdot\mathrm{diam}(\mathbb{N}) \le \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$, it follows from (2.25) that the sufficient condition for the α–smoothing spline algorithm to be optimal is that

$$r \;\ge\; \sup\{\, \|S(h)\| \;|\; \alpha\|h\|_F^2 + (1-\alpha)\delta^{-2}\|N(h)\|_Y^2 \le 1\,\},$$

or equivalently

$$\begin{aligned} &\inf_{\|S(h)\|\ge r} \max\{\, \|h\|_F^2,\, \|N(h)\|_Y^2/\delta^2 \,\} \\ \le\;& \inf_{\|S(h)\|\ge r} \alpha\,\|h\|_F^2 \,+\, (1-\alpha)\,\delta^{-2}\|N(h)\|_Y^2. \end{aligned} \tag{2.27}$$

For $a = (a_1,a_2)$, $b = (b_1,b_2) \in \mathbb{R}^2$, let $\langle a,b\rangle_2 = a_1b_1 + a_2b_2$ be the ordinary inner product in $\mathbb{R}^2$. Then (2.27) can be rewritten as

$$\inf_{a\in A} \|a\|_\infty \;\le\; \inf_{a\in A} \langle \beta, a\rangle_2 \tag{2.28}$$

where $\beta = \beta(\alpha) = (\,\alpha, 1-\alpha\,)$.

We now show that there exists $\alpha = \alpha^*$ for which (2.28) holds. Assume first that the set A is convex. Let $\gamma = \inf_{a\in A}\|a\|_\infty$ and $a^* = (a_1^*, a_2^*) \in \overline{A}$ be a point, for which $\|a^*\|_\infty = \gamma$. Clearly, a^* is a boundary point of $\overline{A}$.

It follows from convex analysis that there exists a line $\langle \beta, a\rangle_2 = c$ passing through a^* and separating A from the convex (and disjoint with the interior of A) set $\{\, a \in \mathbb{R}^2 \mid \|a\|_\infty \le \gamma\,\}$. That is, $\langle \beta, a^*\rangle = c$ and $\langle \beta, a\rangle_2 \ge c$ for $a \in A$, $\langle \beta, a\rangle_2 \le c$ for $\|a\|_\infty \le \gamma$. Observe that β_1, β_2 and c can be chosen all nonnegative. Let β be normalized in such a way that $\beta_1 + \beta_2 = 1$. Then $c = \gamma$. Indeed, this is clear if $a_1^* = a_2^* = \gamma$. For $a_1^* < a_2^* = \gamma$ (or $a_2^* < a_1^* = \gamma$) it is enough to note that then $\beta = (0,1)$ (or $\beta = (1,0)$). Hence, we have obtained that $\inf_{a\in A}\langle \beta, a\rangle_2 \ge \gamma$ and (2.28) follows with $\alpha^* = \beta_1$.

If the set A is not convex, we take β constructed as above for the convex set $\mathrm{conv}(A)$. From the condition (2.26) we obtain

$$\begin{aligned} \inf_{a\in A} \|a\|_\infty &= \inf_{a\in c(A)} \|a\|_\infty \\ &\le \inf_{a\in c(A)} \langle \beta, a\rangle_2 \;\le\; \inf_{a\in A} \langle \beta, a\rangle_2, \end{aligned}$$

as claimed. The proof of the theorem is complete. □

Theorem 2.9 applies in the case when the range space G of the solution operator S is a Hilbert space. Indeed, this follows from the following lemma.

Lemma 2.8 *Let $\|\cdot\|_i$, $i = 0,1,2$, be three extended Hilbert space seminorms on the same linear space X. Then the set*

$$A \;=\; \left\{ \left(\|x\|_1^2, \|x\|_2^2 \right) \in \mathbb{R}^2 \;\middle|\; 1 \le \|x\|_0^2 < +\infty \right\}$$

satisfies

$$\inf_{a\in A} \|a\|_\infty \;=\; \inf_{a\in c(A)} \|a\|_\infty.$$

Proof For $x \in X$, we denote $a(x) = (\|x\|_1^2, \|x\|_2^2)$. Let $\gamma = \inf_{a\in A} \|a\|_\infty$.

Suppose that the lemma is not true. Then there exist two different points $a(x), a(y) \in A$ such that $\|x\|_0 = \|y\|_0 = 1$, and for some u from the interval $[a(x), a(y)]$ we have $\|u\|_\infty < \gamma$. We show that this is impossible. More precisely, we show that there exists a continuous curve $C \subset A$ joining $a(x)$ with $a(y)$ and passing through the interval $[0, u]$ at some u'. Then $u' \in A$ and $\|u'\|_\infty \le \|u\|_\infty < \inf_{a\in A} \|a\|_\infty$, which is a contradiction.

Since $a(x) = a(-x)$, we can assume that $\langle x, y\rangle_0 \ge 0$. Let $L = \{\, a \in \mathbb{R}^2 \mid \langle w, a\rangle = c\,\}$ be the line passing through $a(x)$ and $a(y)$. ($\langle \cdot, \cdot\rangle$ is here the ordinary inner product in $\mathbb{R}^2$.) Since $\|u\|_\infty < \min\{\|a(x)\|_\infty, \|a(y)\|_\infty\}$, L

passes through the half-lines $\{(t,0) \mid t \geq 0\}$ and $\{(0,t) \mid t \geq 0\}$. We consider two cases.
1. $\|x-y\|_0 > 0$.
Denote $x(t) = t\,x + (1-t)\,y$ and $u(t) = a(\,x(t)/\|x(t)\|_0\,)$. Since $\langle x, y \rangle_0 \geq 0$, the 0-seminorm of $x(t)$ is positive. Then $\{\,u(t) \mid\ -\infty < t < +\infty\,\}$ is a continuous curve in A with $\lim_{t\to\pm\infty} u(t) = a(\,(x-y)/\|x-y\|_0) \in A$. Since the quadratic polynomial $Q(t) = \|x(t)\|_0^2(\,\langle w, u(t)\,\rangle_2 - c\,)$ vanishes for $t = 0, 1$, L divides the curve into two curves which lay on the opposite sides of L and join $a(x)$ with $a(y)$. One of them passes through $[0, u]$.
2. $\|x-y\|_0 = 0$.
In this case $\|x(t)\|_0 = 1$, for all $t \in \mathbb{R}$. Hence, $\lim_{t\to\pm\infty} u(t)/t^2 = a(x-y) \neq 0$. Using this and the argument about zeros of the polynomial $Q(t)$, we conclude that the curve $\{\,u(t) \mid 0 \leq t \leq 1\,\}$ passes through $[0, u]$. □

We have shown that there exists an optimal linear smoothing spline algorithm provided that $\|\cdot\|_F$, $\|\cdot\|_Y$ and $\|\cdot\|$ are all Hilbert extended seminorms. This was a consequence of the fact that for some $\alpha = \alpha^*$ we have the equality

$$\begin{aligned} & \sup\{\,\|S(h)\| \mid \quad \|h\|_F \leq 1,\ \|N(h)\|_Y \leq \delta\,\} \\ = \quad & \sup\{\,\|S(h)\| \mid \quad \alpha\,\|h\|_F^2 + (1-\alpha)\delta^{-2}\|N(h)\|_Y^2 \leq 1\,\}. \end{aligned}$$

In particular, α^* should be chosen in such a way that the right hand side of this equality is minimal. Hence, we have the following corollary.

Corollary 2.3 *The optimal value of α is given as*

$$\alpha^* \;=\; arg \min_{0\leq\alpha\leq 1} \sup\{\,\|S(h)\| \mid \quad \alpha\|h\|_F^2 + (1-\alpha)\delta^{-2}\|N(h)\|_Y^2 \leq 1\,\}. \qquad \square$$

The same ideas can be used to prove optimality of smoothing spline algorithms in another case. As before, we assume that $\|\cdot\|_F$ is a Hilbert space seminorm, but relax this requirement for the seminorm $\|\cdot\|_Y$. That is, we let

$$\mathbb{N}(f) \;=\; \{\,y = [y_1, \ldots, y_n] \in \mathbb{R}^n \mid \quad |y_i - L_i(f)| \leq \delta_i,\ 1 \leq i \leq n\,\},$$

where L_i are linear functionals and $0 \leq \delta_i \leq +\infty$, $1 \leq i \leq n$. We also assume that the solution operator S is a linear functional.

For $\beta = (\beta_1, \ldots, \beta_n, \beta_{n+1}) \in \mathbb{R}^{n+1}$ such that $\beta_i \geq 0$ and $\sum_{i=1}^{n+1} \beta_i = 1$, we define an extended seminorm on $F \times \mathbb{R}^n$ as

$$\Gamma_\beta(f,y) \;=\; \beta_{n+1}\, \|f\|_F^2 \,+\, \sum_{i=1}^n \beta_i\, \delta_i^{-2}\, |y_i - L_i(f)|^2.$$

A *β–smoothing spline algorithm* is then $\varphi_\beta(y) \;=\; S(\mathbf{s}_\beta(y)\,)$ where $\mathbf{s}_\beta(y)$ minimizes $\Gamma_\beta(f,y)$ over all $f \in F$.

Due to Theorem 2.4, an optimal linear algorithm exists. It turns out, that it can be interpreted as a β–smoothing spline algorithm.

Theorem 2.10 *If S is a linear functional then*

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sup\,\{\, S(h) \mid \quad \|h\| \leq 1,\ |L_i(h)| \leq \delta_i,\ 1 \leq i \leq n\,\}$$

and there exists β^* such that the β^*–smoothing spline algorithm is optimal. Furthermore, for any $f \in F$ and $y \in \mathbb{N}(f)$

$$|\, S(f) - \varphi_{\beta^*}(y)\,| \;\leq\; \sqrt{1 - \Gamma_{\beta^*}(y)} \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}).$$

Proof As in the proof of Lemma 2.7 we can show that

$$|\, S(f) - \varphi_\beta(y)\,| \;\leq\; \sqrt{1 - \Gamma_\beta(y)}$$
$$\sup\,\{\, S(h) \mid \quad \beta_{n+1}\|h\|_F^2 + \sum_{i=1}^n \beta_i\, \delta_i^{-2} |L_i(h)|^2 \;\leq\; 1\,\}.$$

Repeating a corresponding part of the proof of Theorem 2.9 we obtain a sufficient condition for the β–smoothing spline algorithm to be optimal. Namely, the set

$$B \;=\; \left\{ \left(\, \|x\|_1^2, \ldots, \|x\|_{n+1}^2 \,\right) \in \mathbb{R}^{n+1} \mid \quad S(x) \geq 1 \right\},$$

where $\|x\|_i = |L_i(x)|/\delta_i,\ 1 \leq i \leq n,\ \|x\|_{n+1} = \|x\|_F$, must satisfy

$$\inf_{b \in B}\ \|b\|_\infty \;=\; \inf_{b \in c(B)}\ \|b\|_\infty. \tag{2.29}$$

But, (2.29) follows from the fact, that for any $b \in c(B)$ there exists $\tilde{b} \in B$ with all $n+1$ components not greater than the corresponding components of b. Indeed, if $b = \sum_{j=1}^m c_j\, b_j$, where $b_j \;=\; (\|x_j\|_i^2,\ 1 \;\leq\; i \;\leq\; n+1\,) \;\in\; B$,

$\sum_{j=1}^{m} c_j = 1$, $c_j \geq 0$, then one can take $\tilde{b} = (\| \sum_{j=1}^{m} c_j x_j \|_i^2,\ 1 \leq i \leq n+1\,)$. Direct calculations show that

$$\left\| \sum_{j=1}^{m} c_j \, x_j \right\|_i^2 \;\leq\; \sum_{j=1}^{m} c_j \, \|x_j\|_i^2, \qquad 1 \leq i \leq n+1.$$

Notes and Remarks

NR 2.12 The theory of splines traces back to the end of fifties when researchers found out many interesting things about polynomial splines (see the next section). We cite only Golomb and Weinberger [17], Schoenberg [91] [90], and Schoenberg and Greville [92]. Since that time splines have been well known and studied from different viewpoints in approximation theory,numerical analysis and statistics. A general approach to spline algorithms in the worst case setting with exact information and most relations between splines and optimal error algorithms were presented in Traub and Woźniakowski [109, Chap.4]. The general definition of smoothing splines and Theorem 2.8 seem to be new.

NR 2.13 Optimality of ordinary splines in the worst case setting was shown by Kacewicz and Plaskota [33], while optimality of α–smoothing splines by Melkman and Micchelli [57]. The proof of Theorem 2.10 is based on the latter paper (see also Micchelli [58]). Lemma 2.6 and Theorem 2.10 seem however to be new.

NR 2.14 For optimality of the α–smoothing spline algorithm, it is essential that G is a Hilbert space; see Melkmann and Micchelli [57] for a counterexample.

Exercises

E 2.21 Give an example showing that the estimate $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_o) \leq (1+\rho)/2 \cdot \mathrm{diam}(\mathbb{N})$ in Theorem 2.7 is sharp.

E 2.22 Show that the definitions of the ordinary spline $\mathbf{s}_o(y)$ for $\rho = 1$ and α–smoothing spline $\mathbf{s}_\alpha(y)$ with $0 < \alpha < 1$ coincide, if information $\mathbb{N}$ is exact and linear.

E 2.23 Let $\|(\cdot,\cdot)\|_*$ be an extended seminorm on the space $F \times \mathbb{R}^n$, such that for some $0 < d_1 \leq d_2 < +\infty$ we have

$$d_1 \, \|(f,x)\|_\infty \;\leq\; \|(f,x)\|_* \;\leq\; d_2 \, \|(f,x)\|_\infty, \quad \forall\, f \in F,\ x \in \mathbb{R}^n.$$

Show that then for the smoothing spline algorithm φ_* we have

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_*) \;\leq\; \frac{1+\rho}{2} \, \frac{d_2}{d_1} \, \mathrm{diam}(\mathbb{N}).$$

E 2.24 Let $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) < +\infty$ and $0 < \alpha < 1$. Show that then $S(f_1) = S(f_2)$, for any α–smoothing splines f_1, f_2 corresponding to information y such that $\Gamma_\alpha(y) < +\infty$. That is, the α–smoothing spline algorithm is defined uniquely on the subspace $\{\, y \in \mathbb{R}^n \mid \Gamma_\alpha(y) < +\infty \,\}$.

E 2.25 Let $\|\cdot\|_0$ and $\|\cdot\|_1$ be two extended seminorms on a linear space, and let $0 < r_0, r_1 < +\infty$. Show that

$$\sup_{\|h\|_1 \le r_1} \|h\|_0 \;=\; r_0 \quad\Longleftrightarrow\quad \inf_{\|h\|_0 \ge r_0} \|h\|_1 \;=\; r_1 .$$

E 2.26 Prove that the set A in Lemma 2.8 is convex.

E 2.27 Let $\|\cdot\|_F$ and $\|\cdot\|_Y$ be Hilbert extended seminorms, and let $0 < \alpha < 1$. Define the set

$$E \;=\; \{\, f \in F \mid \ \ \|f\|_F^2 \le 1/\alpha \,\},$$

and the information operator

$$\mathbb{N}(f) \;=\; \left\{ y \in \mathbb{R}^n \;\middle|\; \ \|y - N(f)\|_Y^2 \le \frac{1 - \alpha\,\|f\|_F^2}{1-\alpha} \right\}$$

where $N : F \to \mathbb{R}^n$ is a linear operator. Show that in this case the α–smoothing spline algorithm is optimal for any linear solution operator S and

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sup\{\, \|S(h)\| \mid \ \ \alpha\,\|h\|_F^2 + (1-\alpha)\,\|N(h)\|_Y^2 \le 1 \,\}.$$

E 2.28 Suppose that the solution operator S in Theorem 2.10 is not a functional. Show that then there exists a linear algorithm with error not larger than $\sqrt{n+1}\cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$, where n is the number of functionals in $\mathbb{N}$.

E 2.29 Give an example of a problem for which the smoothing spline element does not exist.

2.6 Special splines

In this section we consider several special cases. We first show explicit formulas for the α–smoothing spline and optimal choice of α in the case when F is a Hilbert space. Then we prove that regularization and classical polynomial splines lead to algorithms that are optimal in the worst case setting. Finally, we consider splines in reproducing kernel Hilbert spaces.

2.6.1 The Hilbert case with optimal α

Let F and G be separable Hilbert spaces and let $S : F \to G$ be a linear and continuous operator. Let E be the unit ball in F. Suppose that for an (unknown) element $f \in E$ we observe data

$$y \;=\; N(f) \,+\, x,$$

where $N : F \to Y = \mathbb{R}^n$ is a continuous linear operator,

$$N \;=\; [\,\langle\cdot, f_1\rangle_F, \langle\cdot, f_2\rangle_F, \ldots, \langle\cdot, f_n\rangle_F\,],$$

and $f_i \in F$, $1 \le i \le n$. The noise x is bounded in an extended Hilbert norm of $\mathbb{R}^n$, $\|x\|_Y \le \delta$ with $\delta > 0$. That is, $\|\cdot\|_Y$ is given by $\|x\|_Y = \sqrt{\langle x, x\rangle_Y}$,

$$\langle x_1, x_2\rangle_Y \;=\; \begin{cases} \langle \Sigma^{-1} x_1, x_2\rangle_2 & x_1, x_2 \in \Sigma(\mathbb{R}^n), \\ +\infty & \text{otherwise,} \end{cases}$$

where the operator (matrix) $\Sigma : \mathbb{R}^n \to \mathbb{R}^n$ is symmetric and nonnegative definite, $\Sigma = \Sigma^* \ge 0$. Note that $\langle\cdot,\cdot\rangle_Y$ is a well defined inner product on $\Sigma(\mathbb{R}^n)$ since $\Sigma y_1 = \Sigma y_2 = x_1$ implies $\langle y_1, x_2\rangle_2 = \langle y_2, x_2\rangle_2$, for all $x_1, x_2 \in \Sigma(\mathbb{R}^n)$.

We first show formulas for the α–smoothing spline. For $\alpha = 1$ we obviously have $\mathbf{s}_\alpha \equiv 0$.

Lemma 2.9 *Let $0 \le \alpha < 1$. The quantity $\Gamma_\alpha(y) = \inf_{f \in F} \Gamma_\alpha(f, y)$ is finite if and only if $y \in Y_1 = N(F) + \Sigma(\mathbb{R}^n)$. For $y \in Y_1$, the smoothing spline is given as*

$$\mathbf{s}_\alpha(y) \;=\; \sum_{j=1}^{n} z_j\, f_j,$$

where $z \in Y_1$ is the solution of the linear system

$$(\,\gamma\,\Sigma \,+\, G_N\,)\, z \;=\; y,$$

with the matrix

$$G_N \;=\; \{\,\langle\, f_i, f_j\,\rangle_F\}_{i,j=1}^{n}$$

and $\gamma = \alpha(1-\alpha)^{-1}\delta^2$. Moreover, $\Gamma_\alpha(y) = \Gamma_\alpha(\mathbf{s}_\alpha(y), y) = \langle y, z\rangle_2$.

For $\alpha > 0$ and $y \in Y_1$, the α–smoothing spline is defined uniquely.

Proof If $y \notin Y_1$ then for all $f \in F$ we have $y - N(f) \notin \Sigma(\mathbb{R}^n)$ and $\|y - N(f)\|_Y = +\infty$. Hence, $\Gamma_\alpha(f, y) = +\infty$.

Assume that $y \in Y_1$. Then any $f \in F$ can be decomposed as $f = \sum_{j=1}^n \beta_j f_j + f^\perp$, where $f^\perp$ is orthogonal to $\mathrm{span}\{f_1, \ldots, f_n\}$. (Note that this decomposition need not be unique.) We have $\|f\|_F^2 = \langle G_N\beta, \beta\rangle_2 + \|f^\perp\|_F^2$ and $\|y - N(f)\|_Y^2 = \langle \Sigma^{-1}(y - G_N\beta), (y - G_N\beta)\,\rangle_2$. Hence,

$$\Gamma_\alpha(y) \;=\; \inf_\beta \quad \gamma\,\langle G_N\beta, \beta\rangle_2 \;+\; \langle \Sigma^{-1}(y - G_N\beta), (y - G_N\beta)\rangle_2.$$

Denoting by P the orthogonal projection in $\mathbb{R}^n$ onto the subspace $\Sigma(\mathbb{R}^n)$ with respect to $\langle\cdot,\cdot\rangle_2$, we obtain

$$\begin{aligned} & \gamma\,\langle G_N\beta, \beta\rangle_2 \;+\; \langle \Sigma^{-1}(y - G_N\beta), (y - G_N\beta)\rangle_2 \\ =\;& \gamma\,\langle G_N\beta, \beta\rangle_2 \;+\; \langle \Sigma^{-1}P(y - G_N\beta), (y - G_N\beta)\rangle_2 \\ =\;& \langle A\beta, \beta\rangle_2 \;-\; 2\,\langle b, \beta\rangle_2 \;+\; c, \end{aligned}$$

where

$$A \;=\; G_N\left(\gamma I + \Sigma^{-1}PG_N\right), \quad b \;=\; G_N\Sigma^{-1}Py, \quad c \;=\; \langle \Sigma^{-1}Py, y\rangle_2.$$

Clearly, $A = A^* > 0$. It is well known that $\langle A\beta, \beta\rangle_2 - 2\langle b, \beta\rangle_2$ is minimized for any β satisfying $A\beta = b$, i.e.,

$$G_N\left(\gamma I + \Sigma^{-1}PG_N\right)\,\beta \;=\; G_N\Sigma^{-1}Py. \tag{2.30}$$

In particular, (2.30) holds for $\beta = z$. Furthermore, for $f_y = \sum_{j=1}^n z_j f_j$ we have $\Gamma_\alpha(f_y, y) = \langle z, y\rangle_2$.

To prove the uniqueness of $\mathbf{s}_\alpha$ in the case $\alpha \neq 0$, it suffices to show that if (2.30) holds for two different $\beta^{(1)}$ and $\beta^{(2)}$, then $f^{(1)} = f^{(2)}$ where $f^{(1)} = \sum_{j=1}^n \beta_j^{(1)} f_j$ and $f^{(2)} = \sum_{j=1}^n \beta_j^{(2)} f_j$. Indeed, let $\beta = \beta^{(1)} - \beta^{(2)}$. Then $A\beta = 0$ and

$$\langle A\beta, \beta\rangle_2 \;=\; \gamma\,\langle G_N\beta, \beta\rangle_2 + \langle G_N\Sigma^{-1}PG_N\beta, \beta\rangle_2 \;=\; 0.$$

Since $G_N\Sigma^{-1}PG_N$ is nonnegative definite, we obtain $\langle G_N\beta, \beta\rangle_2 = \|f^{(1)} - f^{(2)}\|_F^2 = 0$ which means that $f^{(1)} = f^{(2)}$, as claimed. □

We note that Lemma 2.9 says, in particular, that the smoothing spline is in the space spanned by the elements f_i which form information N. To find $\mathbf{s}_\alpha$, it suffices to solve a linear system of equations with the Gram matrix G_N.

Corollary 2.4 *For $0 \le \alpha < 1$, the α–smoothing spline algorithm is given as*

$$\varphi_\alpha(y) \;=\; \sum_{j=1}^{n} z_j S(f_j), \qquad y \in Y_1,$$

where $z \in Y_1$ satisfies $(\gamma\Sigma + G_N)z = y$ and $\gamma = \alpha(1-\alpha)^{-1}\delta^2$. □

We pass to the optimal choice of α. Recall that the algorithm φ_α with $\alpha = 1/2$ gives error at most $\sqrt{2}$ times larger than the minimal error, and that there exists α^* for which φ_{α^*} is optimal.

Consider first the case when Σ is positive definite, $\Sigma > 0$. Then $\|\cdot\|_Y$ is a Hilbert norm and there exists the operator N^* adjoint to N with respect to the inner product $\langle\cdot,\cdot\rangle_Y$ in $\mathbb{R}^n$. That is,

$$\langle N(f), y\rangle_Y \;=\; \langle f, N^*(y)\rangle_F, \qquad \forall f \in F, y \in \mathbb{R}^n.$$

Lemma 2.10 *Let $\Sigma > 0$. Then*

$$\alpha^* \;=\; \arg \min_{0\le\alpha\le 1} \max \left\{ \lambda \;\middle|\; \lambda \in Sp(\, SA_\alpha^{-1}S^*\,) \right\},$$

where

$$A_\alpha \;=\; \alpha\, I \;+\; \frac{1-\alpha}{\delta^2}\, N^*N$$

and $Sp(\cdot)$ is the spectrum of an operator. Furthermore,

$$\begin{aligned} \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\alpha^*}) &= \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \\ &= \max \left\{ \sqrt{\lambda} \;\middle|\; \lambda \in Sp(SA_{\alpha^*}^{-1}S^*) \right\}. \end{aligned}$$

Note that for $\alpha = 0$, the operator $A_\alpha = \delta^{-2}N^*N$ may be not one-to-one. In this case, if $\ker N \not\subset \ker S$ then we formally set $\max\{\,\lambda \mid \lambda \in Sp(SA_0^{-1}S^*)\,\} = +\infty$. If $\ker N \subset \ker S$ then we treat $A_0 = \delta^{-2}N^*N$ as an operator acting in the space $V = (\ker N)^\perp$. Since $S(\ker N) = \{0\}$, we have $S^*(V) \subset V$. Hence, $SA_0^{-1}S^* : V \to V$ is a well defined self–adjoint and nonnegative definite operator.

Proof Due to Corollary 2.3, the optimal $\alpha = \alpha^*$ minimizes

$$\sup\{\, \|Sh\| \;\mid\; \alpha\|h\|_F^2 + (1-\alpha)\delta^{-2}\|Nh\|_Y^2 \le 1\,\}, \tag{2.31}$$

and $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ is equal to the minimal value of (2.31). In our case, (2.31) can be rewritten as

$$\begin{aligned} & \sup\{\, \|SA_\alpha^{-1/2}(A_\alpha^{1/2}h)\| \mid \ \|A_\alpha^{1/2}h\|_F \le 1\,\} \\ = \ & \sup\{\, \|SA_\alpha^{-1/2}h\| \mid \ \|h\|_F \le 1\,\} \\ = \ & \max\{\, \sqrt{\lambda} \mid \ \ \lambda \in Sp(SA_\alpha^{-1}S^*)\,\} \end{aligned}$$

(this holds also for $\alpha = 0$). This completes the proof. □

From Lemma 2.10 we can derive the following more specific theorem about α^*.

Theorem 2.11 *Let $\{\xi_j\}_{j\ge1}$ be a complete orthonormal basis of eigenelements of the operator N^*N. Let η_j be the corresponding eigenvalues,*

$$N^*N\xi_j \ = \ \eta_j\xi_j, \qquad j \ge 1.$$

Then the optimal α^* is the minimizer of

$$\psi(\alpha) \ = \ \sup_{\|g\|=1} \sum_{j\ge1} \frac{\langle S(\xi_j), g\rangle^2}{\alpha + \delta^{-2}\eta_j(1-\alpha)}, \qquad 0 \le \alpha \le 1,$$

and $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \sqrt{\psi(\alpha^*)}$. *In particular, if S is a functional then*

$$\psi(\alpha) \ = \ \sum_{j\ge1} \frac{S^2(\xi_j)}{\alpha + \delta^{-2}\eta_j(1-\alpha)}.$$

Proof Due to Lemma 2.27, the optimal α^* minimizes the inner product $\langle SA_\alpha^{-1}S^*g, g\rangle$ over all g with $\|g\| = 1$. Observe that

$$A_\alpha^{-1}\xi_j \ = \ (\alpha + (1-\alpha)\eta_j\delta^{-2})^{-1}\xi_j$$

and $S^*g = \sum_{j\ge1}\langle S\xi_j, g\rangle\xi_j$. This and orthonormality of $\{\xi_j\}$ yield

$$\langle SA_\alpha^{-1}S^*g, g\rangle \ = \ \langle A_\alpha^{-1}S^*g, S^*g\rangle_F \ = \ \sum_{j\ge1} \frac{\langle S\xi_j, g\rangle^2}{\alpha + \delta^{-2}\eta_j(1-\alpha)}.$$

If S is a functional then $g \in \{-1, 1\}$ and $\langle S\xi_j, g\rangle^2 = S^2\xi_j$, which completes the proof. □

For singular Σ, Lemma 2.10 and Theorem 2.11 should be modified as follows. Let $F_1 = \{\, f \in F \mid N(f) \in \Sigma(\mathbb{R}^n)\,\}$. Let $S_1 : F_1 \to G$ and $N_1 : F_1 \to \Sigma(\mathbb{R}^n)$ be defined by $S_1(f) = f$ and $N_1(f) = N(f)$, $f \in F_1$. Since $\|\cdot\|_Y$ is a Hilbert norm in $\Sigma(\mathbb{R}^n)$, the adjoint operator $N_1^* : \Sigma(\mathbb{R}^n) \to F_1$ exists. Lemma 2.10 and Theorem 2.11 hold with S and N replaced by S_1 and N_1. For instance, if information is exact then $\Sigma \equiv 0$, $F_1 = \ker N$, $N_1^* N_1 \equiv 0$, and $A_\alpha = \alpha I$. The optimal α is then $\alpha^* = 1$. That is, $\mathbf{s}_{\alpha^*}(y)$ is the element of the set $\{\, f \in F \mid N(f) = y\,\}$ with the minimal norm, and

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\alpha^*}) \;=\; \sup\{\, \|S(h)\| \mid \quad h \in \ker N\,, \|h\| \le 1\,\}.$$

We now specialize the formulas for the optimal α^*, $\gamma^* = \alpha^*(1-\alpha^*)^{-1}\delta^2$, and for $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ assuming that S is a compact operator and S^*S and N^*N possess a common orthonormal basis of eigenvectors. That is, we assume that there exists in F an orthonormal basis $\{\xi_i\}_{i=1}^d$ ($d = \mathrm{dim} F \le +\infty$), such that

$$S^*S\xi_i \;=\; \lambda_i \xi_i \qquad \text{and} \qquad N^*N\xi_i \;=\; \eta_i \xi_i,$$

where $\lambda_1 \ge \lambda_2 \ge \ldots \ge 0$ are the dominating eigenvalues of S^*S and η_i are eigenvalues of N^*N. (If $d < +\infty$ then we formally set $\lambda_i = \eta_i = 0$ for $i > d$.)

In this case, $\{S\xi_i/\|S\xi_i\|\}$, $\xi_i \notin \ker S$, is an orthonormal basis in $S(F)$ of eigenelements of the operator $SA_\alpha^{-1}S^*$, and the corresponding eigenvalues are

$$\tilde{\lambda}_i(\alpha) \;=\; \frac{\lambda_i}{\alpha + \delta^{-2}\eta_i(1-\alpha)}\,, \qquad i \ge 1.$$

Hence, to find the optimal α and the radius of $\mathbb{N}$ we have to minimize $\max_{i\ge1} \tilde{\lambda}_i(\alpha)$ over all $\alpha \in [0,1]$.

Let $1 = p_1 < p_2 < \cdots < p_k$ be the finite sequence of integers defined (uniquely) by the following condition. For any i, p_{i+1} is the smallest integer such that $p_{i+1} > p_i$ and $\lambda_{p_i}/\eta_{p_i} < \lambda_{p_{i+1}}/\eta_{p_{i+1}}$ (here $a/0 = +\infty$ for $a > 0$ and $0/0 = 0$). If such an p_{i+1} does not exist then $k = i$. It is easy to see that then for any α

$$\max_{i\ge1} \tilde{\lambda}_i(\alpha) \;=\; \max_{1\le i\le k} \tilde{\lambda}_{p_i}(\alpha).$$

Next, let

$$P_1 \;=\; \{\, p_i \mid \quad 1 \le i \le k,\ \delta^2 < \eta_{p_i}\,\}, \tag{2.32}$$

$$P_2 \;=\; \{\, p_i \mid \quad 1 \le i \le k,\ \delta^2 \ge \eta_{p_i}\,\}. \tag{2.33}$$

Observe that for any $i \in P_1$, $\tilde{\lambda}_i(\alpha)$ is an increasing function of α, while for $j \in P_2$ it is nonincreasing. Hence, in the case $P_1 = \emptyset$ we have $\alpha^* = 1$

and $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \sqrt{\lambda_1}$, while for $P_2 = \emptyset$ we have $\alpha^* = 0$ and $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \delta\sqrt{\max_i \lambda_i/\eta_i}$.

Suppose that both sets P_1 and P_2 are nonempty. Then

$$\min_{0\le\alpha\le 1} \max_{1\le j\le k} \tilde{\lambda}_{i_j}(\alpha) \;=\; \max_{i\in P_1, j\in P_2} \beta_{ij},$$

where $\beta_{ij} = \tilde{\lambda}_i(\alpha_{ij})$ and α_{ij} are such that $\tilde{\lambda}_i(\alpha_{ij}) = \tilde{\lambda}_j(\alpha_{ij})$. The optimal $\alpha^* = \alpha_{st}$ where s, t are chosen in such a way that β_{st} minimizes β_{ij} over $i \in P_1$ and $j \in P_2$. Furthermore, $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \sqrt{\beta_{st}}$.

Noting that

$$\alpha_{ij} \;=\; \frac{\lambda_j\eta_i - \lambda_i\eta_j}{\lambda_i(\delta^2-\eta_j) + \lambda_j(\eta_i - \delta^2)}$$

and

$$\beta_{ij} \;=\; \lambda_j \;+\; \left(\frac{\delta^2-\eta_j}{\eta_i-\eta_j}\right)(\lambda_i - \lambda_j),$$

we obtain the following corollary.

Corollary 2.5 *Suppose that the operators S^*S and N^*N have a common basis of eigenelements with the corresponding eigenvalues λ_i and η_i. Let the sets P_1 and P_2 be defined by (2.32) and (2.33).*

(i) If $P_1 = \emptyset$ then $\alpha^* = 1$ and $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \sqrt{\lambda_1}$.

(ii) If $P_2 = \emptyset$ then $\alpha^* = 0$ and

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \delta\cdot\sqrt{\max_{i\ge 1}\frac{\lambda_i}{\eta_i}}\,.$$

(iii) If both P_1 and P_2 are nonempty then

$$\alpha^* \;=\; \frac{\lambda_t\eta_s - \lambda_s\eta_t}{\lambda_s(\delta^2-\eta_t) + \lambda_t(\eta_s-\delta^2)}$$

and

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sqrt{\lambda_t \;+\; \left(\frac{\delta^2-\eta_t}{\eta_s-\eta_t}\right)(\lambda_s-\lambda_t)},$$

where

$$(s,t) \;=\; arg \max_{(i,j)\in P_1\times P_2} \lambda_j \;+\; \left(\frac{\delta^2-\eta_j}{\eta_i-\eta_j}\right)(\lambda_i-\lambda_j). \quad \Box$$

Suppose now that δ^2 is small,

$$0 \;<\; \delta^2 \;\le\; \max_{1\le i\le t} \eta_i$$

where $t = \min\{\, i \ge 1 \mid \eta_i = 0\,\}$. Let $s = \arg\max_{1\le i\le t-1}(\lambda_i - \lambda_t)/\eta_i$. Then Corollary 2.5 yields the following formulas:

$$\alpha^* \;=\; \frac{\lambda_t}{\lambda_t + \frac{\delta^2}{\eta_s}(\lambda_s - \lambda_t)}$$

and

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sqrt{\lambda_t + \frac{\delta^2}{\eta_s}(\lambda_s - \lambda_t)}.$$

Observe that $\alpha^* \to 1$ as $\delta^2 \to 0^+$. For $\alpha^* \ne 1$ (which holds when $\lambda_s > \lambda_t$) the parameter

$$\gamma^* \;=\; \frac{\eta_s \lambda_t}{\lambda_s - \lambda_t}$$

is constant. This means that the optimal algorithm is independent of the noise level, provided that δ is small (Actually, the same algorithm is optimal also for exact information, see E 2.32.) This nice property is however not preserved in general, as shown in E 2.33.

2.6.2 Least squares and regularization

Consider the Hilbert case of Section 2.6.1 with $\|\cdot\|_Y$ being a Hilbert norm. In this case, as an approximation to $S(f)$ one can take

$$\varphi_{\mathrm{ls}}(y) = S(u_{\mathrm{ls}}(y)\,),$$

where $u_{\mathrm{ls}}(y)$ is the solution of the *least squares* problem. It is defined by the equation

$$\|y - N(u_{\mathrm{ls}}(y)\,)\|_Y \;=\; \min_{f\in F} \|y - N(f)\|_Y,$$

or equivalently, by $N(u_{\mathrm{ls}}(y)) \;=\; P_N y$ where $P_N : \mathbb{R}^n \to \mathbb{R}^n$ is the orthogonal projection of y onto $Y_N = N(F)$ (with respect to the inner product $\langle\cdot,\cdot\rangle_Y$). This is in turn equivalent to the fact that $u_{\mathrm{ls}}(y)$ is the solution of the normal equations

$$N^*N\,f \;=\; N^*y. \tag{2.34}$$

Indeed, if (2.34) holds then $\langle y - Nf, Nf\rangle_Y \;=\; \langle N^*y - N^*Nf, f\rangle_F \;=\; 0$, which means that $Nf \;=\; P_N y$. On the other hand, since $N^*(Y_N^\perp) \;=\; \{0\}$, for any solution h of the least squares problem we have $N^*Nh = N^*P_N y = N^*y$, as claimed.

The algorithm φ_{ls} is called the *(generalized) least squares algorithm.* For finite dimensional problems with small noise level, the least squares turn out to be optimal. Namely, we have the following theorem. $\dim F = \dim N(F) = d < +\infty$.

Theorem 2.12 *Let* $\dim F = \dim N(F) = d < +\infty$*. Let* $\overline{g} \in G$ *be such that* $\|\overline{g}\| = 1$ *and*

$$\|S(N^*N)^{-1}S^*\overline{g}\,\| \;=\; \|S(N^*N)^{-1}S^*\|\,.$$

Then for sufficiently small noise level δ*,*

$$\delta^2 \cdot \Big\langle\, S(N^*N)^{-2}S^*\overline{g},\, \overline{g}\,\Big\rangle \;\leq\; \|\,S(N^*N)^{-1}S^*\,\|, \tag{2.35}$$

the (generalized) least squares φ_{ls} *is an optimal algorithm. Furthermore,*

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi_{\mathrm{ls}}) \;=\; \delta\cdot\sqrt{\|S(N^*N)^{-1}S^*\|}\,.$$

Proof We can assume that $\|S(N^*N)^{-1}S^*\| > 0$ since otherwise $S \equiv 0$ and the theorem is trivially true. Let $\overline{h} = (N^*N)^{-1}S^*\overline{g}$ and $h = \overline{h}/\|\overline{h}\|_F$. Observe that then $\|S(\overline{h})\| \;=\; \|S(N^*N)^{-1}S^*\|$ and

$$\begin{aligned}\|N(\overline{h})\|_Y &= \langle N(N^*N)^{-1}S^*\overline{g}, N(N^*N)^{-1}S^*\overline{g}\rangle_Y^{1/2}\\ &= \langle S(N^*N)^{-1}S^*\overline{g},\overline{g}\rangle^{1/2} \;=\; \|S(N^*N)^{-1}S^*\|^{1/2}.\end{aligned}$$

Hence, the condition (2.35) gives $\delta \leq \|N(h)\|_Y$ and consequently

$$\begin{aligned}\mathrm{rad}^{\mathrm{wor}}(\mathbb{N};E) &\geq \mathrm{rad}^{\mathrm{wor}}(\mathbb{N},[-h,h])\\ &= \sup\{\,\|S(f)\|\;|\;\; f\in[-h,h],\ \|N(f)\|_Y\leq\delta\,\}\\ &= \delta\,\frac{\|S(h)\|}{\|N(h)\|} \;=\; \delta\,\|S(N^*N)^{-1}S^*\|^{1/2}.\end{aligned}$$

On the other hand, for any f the least squares algorithm gives

$$\begin{aligned}&\sup_{\|x\|_Y\leq\delta}\|S(f)-\varphi_{\mathrm{ls}}(N(f)+x)\,\| \;=\; \sup_{\|x\|_Y\leq\delta}\|SN^{-1}P_Nx\|\\ &=\; \delta\,\|SN^{-1}P_N\| \;=\; \delta\,\|S(N^*N)^{-1}S^*\|^{1/2},\end{aligned}$$

and the theorem follows.

Example 2.10 Consider the case where $F = G$ and S is the identity operator, $S = I$. That is, we want to approximate f from the unit ball of F. In this case, the condition (2.35) takes the form $\delta \leq \lambda_{\min}^{1/2}$ where $\lambda_{\min}$ is the minimal eigenvalue of N^*N. For such δ we have $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \delta\,\lambda_{\min}^{-1/2}$. On the other hand, for $\delta \geq \lambda_{\min}^{1/2}$ we have $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \geq \delta$ and the error δ is achieved by the zero algorithm. Hence,

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \min\left\{ 1,\, \frac{\delta}{\sqrt{\lambda_{\min}}} \right\}.$$

For small noise level, $\delta \leq \lambda_{\min}^{1/2}$, the least squares algorithm is optimal. Otherwise information is useless – zero is the best approximation.

We also note that if the unit ball E is replaced by the whole space F, then φ_{ls} is optimal unconditionally. □

Unfortunately, in general, the least squares algorithm can be arbitrarily bad. For instance, for the simple one–dimensional problem of Example 2.9 we have $\varphi_{\mathrm{ls}}(y) = y$. Hence, the error of φ_{ls} equals $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{ls}}) = \delta$, while $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \min\{a, \delta\}$. Consequently,

$$\frac{\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{ls}})}{\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})} \to +\infty, \qquad \text{as} \qquad \frac{a}{\delta} \to 0.$$

Observe also that the solution of (2.34) is in general not unique and therefore the least squares algorithm φ_{ls} is not uniquely determined.

A simple modification of the least squares relies on *regularization* of the normal equations (2.34). That is, instead of (2.34) we solve "perturbed" linear equations

$$(\,\omega I \,+\, N^*N\,)\, f \;=\; N^*\, y, \tag{2.36}$$

where $I : F \to F$ is the identity operator and $\omega > 0$ is a *regularization parameter*. Then the solution $u_\omega(y)$ of (2.36) exists and is unique for any y. Moreover, it turns out that for a properly chosen parameter ω the *regularization algorithm* $S(u_\omega(y)\,)$ is optimal. Indeed, we have the following fact.

Lemma 2.11 *For $0 < \alpha < 1$, the α–smoothing spline is the regularized solution, i.e.,*

$$\mathbf{s}_\alpha(y) \;=\; u_\omega(y) \;=\; (\,\omega I \,+\, N^*N\,)^{-1}\, N^*\, y$$

where $\omega = \alpha(1-\alpha)^{-1}\delta^2$. *Or, equivalently,* u_ω *is the* α*–smoothing spline with* $\alpha = \omega(\omega+\delta^2)^{-1}$.

Proof Let $\alpha \in (0,1)$. Define the Hilbert space $\tilde{F} = F \times \mathbb{R}^n$ with the extended norm

$$\|(f,y)\|^2 \;=\; \omega\,\|f\|_F^2 \,+\, \|y\|_Y^2,$$

where $\omega = \alpha(1-\alpha)^{-1}\delta^2$. We know from Lemma 2.6 that $\mathbf{s}_\alpha(y)$ is the α–smoothing spline iff

$$\|(0,y) - (\mathbf{s}_\alpha(y)\,)\| \;=\; \min_{f\in F} \|(0,y) - (f, N(f)\,)\|.$$

As in (2.34) we can show that $\mathbf{s}_\alpha(y)$ is the solution of

$$\tilde{N}^*\tilde{N}f \;=\; \tilde{N}^*\tilde{y},$$

where the information operator $\tilde{N} : F \to \tilde{F}$ is defined as $\tilde{N}(f) = (f, N(f)\,)$, and $\tilde{y} = (0,y)$. Since $\tilde{N}^*\tilde{y} = N^*y$ and $\tilde{N}^*\tilde{N} = \omega\,I + N^*N$, the lemma follows. □

Thus, the well known regularization leads to the smoothing spline algorithms. It is interesting that the optimal value of the regularization parameter ω is the same as optimal γ in Theorem 2.9.

Example 2.11 Let $F = G$ with the complete orthonormal basis $\{\xi_i\}_{i\ge 1}$. Let

$$S(f) \;=\; \sum_{i=1}^{\infty} \beta_i \langle f, \xi_i\rangle_F \xi_i,$$

$\beta_1 \ge \beta_2 \ge \cdots \ge 0$, and let information consist of noisy evaluations of the Fourier coefficients, i.e.,

$$N(f) \;=\; [\,\langle f,\xi_1\rangle_F, \ldots, \langle f, \xi_n\rangle_F\,]$$

and $\mathbb{N}(f) = \{\,N(f) + x \mid \|x\|_2 \le \delta\,\}$ ($\|\cdot\|_2$ stands for the Euclidean norm) We also assume, for simplicity, that $\delta \le 1$. Due to Corollary 2.5, in this case the optimal α is

$$\alpha^* \;=\; \frac{\beta_{n+1}^2}{\beta_{n+1}^2 + \delta^2(\beta_1^2 - \beta_{n+1}^2)}.$$

Hence, for $\beta_1 = \beta_{n+1}$ the zero algorithm is optimal, while for $\beta_{n+1} = 0$ we obtain optimality of the least squares algorithm. Let $\beta_1 > \beta_{n+1} > 0$. Then the regularization algorithm with the parameter

$$\omega^* \;=\; \frac{\beta_{n+1}}{\beta_1 - \beta_{n+1}}$$

is an optimal algorithm.

Observe that for $\delta \to 0^+$ we have $\alpha^* \to 1$. However, the regularization parameter ω^* is constant. This seems to contradict the intuition that the smaller δ^2, the smaller ω^*.

Clearly, we can always apply the algorithm $\varphi_\omega(y) = S(u_\omega(y))$ with $\omega = \delta^2$. Then $\omega \to 0^+$ as $\delta^2 \to 0^+$ and $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_\omega) \le \sqrt{2}\,\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$.

2.6.3 Polynomial splines

In this section we recall the classical result that polynomial splines are also α–smoothing splines.

Let $a < b$ and let knots $a \le t_1 < t_2 < \cdots \le t_m \le b$ be given. A *polynomial spline* of order r ($r \ge 1$) corresponding to the knots t_i is a function $\mathbf{p} : [a,b] \to \mathbb{R}$ satisfying the following conditions:

(a) $\mathbf{p} \in \Pi_{2r-1}$ on each interval $[a,t_1], [t_m,b], [t_i,t_{i+1}],\ 1 \le i \le m-1$,

(b) $\mathbf{p}$ has continuous $(2r-2)$nd derivative on $[a,b]$,

Here Π_k is the space of polynomials of degree at most k. A polynomial spline is called *natural* if additionally

(c) $\mathbf{p}^{(i)}(a) = 0 = \mathbf{p}^{(i)}(b),\ \ r \le i \le 2r-2$, and also $\mathbf{p}^{2r-1}(a) = 0$ if $a < t_1$, and $\mathbf{p}^{2r-1}(b) = 0$ if $t_m < b$.

If instead of (c) we have

(c') $\mathbf{p}^{(i)}(a) = \mathbf{p}^{(i)}(b),\ 1 \le i \le 2r-2$, and in the case $a < t_1$, $t_m < b$ also $\mathbf{p}^{(2r-1)}(a) = \mathbf{p}^{(2r-1)}(b)$,

then the polynomial spline is called *periodic*.

Let $W_r(a,b)$ be the Sobolev space of functions f defined on $[a,b]$ that have absolutely continuous $(r-1)$st derivative and rth derivative is square integrable,

$$W_r(a,b) \;=\; \{\, f : [a,b] \to \mathbb{R} \mid \ \ f^{(r-1)} \text{ is abs. cont., } f^{(r)} \in \mathcal{L}_2(a,b) \,\}.$$

Similarly, let $\tilde{W}_r(a,b)$ be the space of functions from $W_r(a,b)$ that can be extended to $(b-a)$–periodic functions on $\mathbb{R}$ with $r-1$ continuous derivative,

$$\tilde{W}_r(a,b) \;=\; \{\, f \in W_r(a,b) \mid \quad f^{(i)}(a) = f^{(i)}(b),\ 0 \le i \le r-1\,\}.$$

Note that natural and periodic polynomial splines belong to $W_r(a,b)$ and $\tilde{W}_r(a,b)$, respectively. They are also α–smoothing splines in these spaces, provided that information is given by noisy function values at t_i's. This fact follows from the following two well known lemmas. For completeness, we add the proofs.

Lemma 2.12 *Let a function $f \in W_r(a,b)$ (or $f \in \tilde{W}_r(a,b)$) vanish at t_i,*

$$f(t_i) \;=\; 0, \qquad 1 \le i \le m.$$

Then for any natural (periodic) polynomial spline $\mathbf{p}$ *of order* r *we have*

$$\int_a^b f^{(r)}(x)\mathbf{p}^{(r)}(x)\,dx \;=\; 0.$$

Proof Integrating by parts we get

$$\int_a^b f^{(r)}(x)\mathbf{p}^{(r)}(x)\,dx \;=\; f^{(r-1)}(x)\mathbf{p}^{(r)}(x)\Big|_a^b \;-\; \int_a^b f^{(r-1)}(x)\mathbf{p}^{(r+1)}(x)\,dx.$$

Observe that $f^{(r-1)}(x)\mathbf{p}^{(r)}(x)|_a^b = 0$, no matter if we have periodic or nonperiodic case. Proceeding in this way we obtain

$$\begin{aligned}\int_a^b f^{(r)}(x)\mathbf{p}^{(r)}(x)\,dx \;&=\; -\int_a^b f^{(r-1)}(x)\mathbf{p}^{(r+1)}(x)\,dx \\ &=\; f^{(r-2)}(x)\mathbf{p}^{(r+1)}(x)\Big|_a^b \;-\; \int_a^b f^{(r-2)}(x)\mathbf{p}^{(r+2)}(x)\,dx \\ &=\; \cdots \;=\; (-1)^i \int_a^b f^{(r-i)}(x)\mathbf{p}^{(r+i)}(x)\,dx \\ &=\; \int_a^b f'(x)\mathbf{p}^{(2r-1)}(x)\,dx. \qquad (2.37)\end{aligned}$$

The function $\mathbf{p}^{(2r-1)}$ is piecewise constant. Denote $t_0 = a$, $t_{m+1} = b$, and by p_i the value of $\mathbf{p}^{(2r-1)}$ on the interval (t_i, t_{i+1}), $0 \le i \le m$ (for $a = t_1$ we set $p_0 = p_1$, and for $t_m = b$ we set $p_m = p_{m-1}$). Then (2.37) equals

$$\sum_{i=0}^{m} p_i\,(\,f(t_{i+1}) - f(t_i)\,) \;=\; p_m f(b) - p_0 f(a) \;=\; 0,$$

as claimed.

Lemma 2.13 *Let $f \in W_r(a,b)$ (or $f \in \tilde{W}_r(a,b)$). Then there exists a natural (periodic) polynomial spline $\mathbf{p}_f$ of order r such that*

$$\mathbf{p}_f(t_i) \;=\; f(t_i), \qquad 1 \le i \le m.$$

The spline $\mathbf{p}_f$ is determined uniquely for all $m \ge r$ (for all $m \ge 1$). Moreover,

$$\int_a^b (\mathbf{p}_f^{(r)}(x)\,)^2\,dx \;\le\; \int_a^b (f^{(r)}(x)\,)^2\,dx.$$

Proof In the nonperiodic case and $m < r$ we can take as $\mathbf{p}_f$ any polynomial p of degree at most $r-1$ satisfying $p(t_i) = f(t_i)$, $\forall i$. Therefore we can assume in the nonperiodic case that $m \ge r$.

We first show that $\mathbf{p} \equiv 0$ is the unique spline that vanishes at t_i, $1 \le i \le m$. Indeed, if for a natural (periodic) spline is $\mathbf{p}(t_i) = 0$, $\forall\, i$, then by Lemma 2.12 with $f = \mathbf{p}$ we have $\int_a^b (\,\mathbf{p}^{(r)}(x)\,)^2\,dx \;=\; 0$. Thus, $\mathbf{p}^{(r)} \equiv 0$, and $\mathbf{p}$ is a polynomial of degree at most $r-1$ that vanishes at m different points. In the nonperiodic case we have $m \ge r$ which means that $\mathbf{p} \equiv 0$. In the periodic case, $\mathbf{p}$ satisfies $\mathbf{p}^{(i)}(a) = \mathbf{p}^{(i)}(b)$, $0 \le i \le r-1$. Then, it must be of the form

$$\mathbf{p}(x) \;=\; \sum_{i=0}^{r-1} \beta_i (x-a)^i \;=\; \sum_{i=0}^{r-1} \beta_i (x-b)^i.$$

This in turn means that $\mathbf{p}$ is a constant polynomial, that vanishes at at least one point. Hence $p \equiv 0$, as claimed.

Observe now that to find all the coefficients of the (natural or periodic) spline that interpolates f, we have to solve a quadratic system of linear equations. The necessary and sufficient condition for the system to have a unique solution is that zero is the only solution of the homogeneous system. This is however the case since the homogeneous system corresponds to $f \equiv 0$.

To show the second part of the lemma, observe that by Lemma 2.12 (with f replaced by $f - \mathbf{p}_f$) we have $\int_a^b \mathbf{p}_f^{(r)}(x)(\,f^{(r)}(x) - \mathbf{p}_f^{(r)}(x)\,)\,dx \;=\; 0$. Hence,

$$\begin{aligned}\int_a^b (f^{(r)}(x)\,)^2\,dx \;&=\; \int_a^b (\mathbf{p}_f^{(r)}(x)\,)^2\,dx + \int_a^b (\,f^{(r)}(x) - \mathbf{p}_f^{(r)}(x)\,)^2\,dx \\ &\ge\; \int_a^b (\mathbf{p}_f^{(r)}(x)\,)^2\,dx,\end{aligned}$$

which completes the proof. □

Now, let $F = W_r(a,b)$ (or $F = \tilde{W}_r(a,b)$) with the seminorm $\|\cdot\|_F$ which is generated by the semi–inner product

$$\langle f_1, f_2\rangle_F \;=\; \int_a^b f_1^{(r)}(x)\, f_2^{(r)}(x)\, dx.$$

We consider the problem with an arbitrary linear solution operator $S : F \to G$, and with information of the form

$$\mathbb{N}(f) \;=\; \{\, y \in \mathbb{R}^n \mid \quad \|y - N(f)\|_Y \le \delta \,\},$$

where

$$N(f) \;=\; [\,\underbrace{f(t_1),\ldots,f(t_1)}_{k_1},\ldots,\underbrace{f(t_m),\ldots,f(t_m)}_{k_m}\,],$$

$\sum_{i=1}^m k_i = n$, and $\|\cdot\|_Y$ is a Hilbert norm in $\mathbb{R}^n$.

Theorem 2.13 *Let $\mathbf{p}_y$ be the natural (periodic) polynomial spline of order r minimizing*

$$\Gamma_\alpha(\mathbf{p}, y) \;=\; \alpha \int_a^b (\mathbf{p}^{(r)}(x)\,)^2\, dx \;+\; \frac{1-\alpha}{\delta^2}\, \|y - N(\mathbf{p})\|_Y^2.$$

Then $\mathbf{p}_y$ is the α–smoothing spline.

Proof It follows from Lemma 2.6 (i) that the α–smoothing spline $\mathbf{s}_\alpha(y)$ exists. We choose $\mathbf{p}$ to be the natural (periodic) polynomial spline of order r satisfying $\mathbf{p}(t_i) = \mathbf{s}_\alpha(y)(t_i)$, $1 \le i \le m$. By Lemma 2.13 we have $\|\mathbf{p}\|_F \le \|\mathbf{s}_\alpha(y)\|_F$. This means that $\Gamma_\alpha(\mathbf{p}, y) \le \Gamma_\alpha(\mathbf{s}_\alpha(y), y)$ and $\mathbf{p}$ is the α–smoothing spline. □

Thus the search for the α–smoothing spline can be restricted to the (finite dimensional) subspace of polynomial splines.

We conclude that for $\alpha = 1/2$ the algorithm $\varphi_{1/2}(y) = S(\mathbf{p}_y)$ is at most $\sqrt{2}$ times worse than optimal. If G is a Hilbert space then

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sup\left\{ \|S(f)\| \;\Big|\; \int_a^b (f^{(r)}(x)\,)^2 \le 1,\; \|y - N(f)\|_Y \le \delta \right\}.$$

However, the optimal value of α is in this case not known, even for such a problem as integration.

2.6.4 Splines in r.k.h.s.

In this section we consider smoothing splines in function spaces where function evaluations are continuous functionals. Such spaces have a nice characterization which we briefly recall.

Let $\mathcal{T}$ be a given set of indices, e.g., $\mathcal{T} = [0,1]$, and let $R : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ be a given symmetric and nonnegative definite function [2]. It is known that then there exists a uniquely determined Hilbert space H_R of functions $f : \mathcal{T} \to \mathbb{R}$, such that for any $t \in \mathcal{T}$, $f \to f(t)$ is a continuous linear functional whose representer is $L_t = R(\cdot, t)$. That is,

$$f(t) \;=\; \langle f, L_t \rangle_R, \qquad f \in H_R,$$

where $\langle \cdot, \cdot \rangle_R$ is the inner product in H_R.

The space H_R is called a *reproducing kernel Hilbert space* with *reproducing kernel* R, or r.k.h.s. with r.k. R, for brevity. It consists of all linear combinations of the functions L_t, $t \in \mathcal{T}$, and their limits with respect to the norm $\| \cdot \|_R = \sqrt{\langle \cdot, \cdot \rangle_R}$ where

$$\left\langle \sum_{i=1}^{n} \alpha_i L_{t_i}, \sum_{j=1}^{k} \beta_j L_{s_j} \right\rangle_R \;=\; \sum_{i=1}^{n} \sum_{j=1}^{k} \alpha_i \beta_j R(t_i, s_j).$$

On the other hand, with any Hilbert space H of functions $f : \mathcal{T} \to \mathbb{R}$ possessing the property that the functionals $f \to f(t)$ are continuous, we can associate a uniquely determined r.k. R, so that $H = H_R$ is an r.k.h.s. Namely,

$$R(s,t) \;=\; \langle L_s, L_t \rangle_H, \qquad s, t \in \mathcal{T},$$

where $L_t \in H$ is the representer of function evaluation at t. Hence, there exists a one–to–one correspondence between symmetric and nonnegative functions and Hilbert spaces in which function evaluations are continuous functionals.

Example 2.12 Let $a < b$ and $r \geq 1$. Define the separable Hilbert space $W_r^0(a,b)$ as

$$\begin{aligned} W_r^0(a,b) \;&=\; \{\, f : [a,b] \to \mathbb{R} \mid \;\; f^{(r-1)}\text{–absolutely continuous}, \\ &\qquad f^{(i)}(a) = 0,\, 0 \leq i \leq r-1,\, f^{(r)} \in \mathcal{L}_2(a,b) \,\}, \end{aligned}$$

[2] This means that for any $n \geq 1$ and $t_i \in \mathcal{T}$, $1 \leq i \leq n$, the matrix $\{R(t_i, t_j)\}_{i,j=1}^{n}$ is symmetric and nonnegative definite.

with the inner product $\langle f_1, f_2\rangle_{W_r} = \int_a^b f_1^{(r)}(u) f_2^{(r)}(u)\,du$. Then $W_r^0(a,b)$ is an r.k.h.s. with r.k.

$$R(s,t) \;=\; R_{r-1}(s,t) \;=\; \int_a^b G_{r-1}(s,u)\,G_{r-1}(t,u)\,du,$$

where

$$G_{r-1}(t,u) \;=\; \frac{(t-u)_+^{r-1}}{(r-1)!}$$

and $x_+ = \max\{x, 0\}$.

Indeed, applying r times the formula $f(t) = \int_a^t f'(u)\,du$ we obtain

$$f(t) \;=\; \int_a^b \frac{(t-u)_+^{r-1}}{(r-1)!}\, f^{(r)}(u)\,du \;=\; \int_a^b G_{r-1}(t,u)\, f^{(r)}(u)\,du, \quad f \in W_r^0. \tag{2.38}$$

Hence, $f \to f(t)$ is a continuous functional,

$$|f(t)| \;\le\; \sqrt{\int_a^b |G_{r-1}(t,u)|^2 du}\;\cdot\; \|f\|_{W_r}.$$

Letting in (2.38) $f = L_t$ (the representer of evaluation at t), we get that $L_t^{(r)}(u) = G_{r-1}(t,u)$ and

$$R(s,t) \;=\; \langle L_s, L_t\rangle_{W_r} \;=\; \int_a^b G_{r-1}(t,u)\,G_{r-1}(s,u)\,du,$$

as claimed.

In particular, for $r = 1$ we have $R_0(s,t) = \min\{s,t\}$. □

The fact that any r.k.h.s. is determined by its r.k. R allows to write the formulas for the α–smoothing spline in terms of R. Namely, using Lemma 2.9 we immediately obtain the following theorem.

Theorem 2.14 *Let $F = H$ be an r.k.h.s. with r.k. $R : \mathcal{T}\times\mathcal{T} \to \mathbb{R}$. Let information $y = N(f) + x$ where*

$$N(f) \;=\; [\,f(t_1), f(t_2), \ldots, f(t_n)\,],$$

$t_i \in \mathcal{T}$, $1 \le i \le n$, and $\|x\|_2 \le \delta$. Let the matrix

$$R_{t_1 \ldots t_n} \;=\; \{\, R(t_i, t_j)\,\}_{i,j=1}^n.$$

Then the α–smoothing spline is given as

$$\mathbf{s}_\alpha(y) \;=\; \sum_{j=1}^n z_j R(t_j,\cdot),$$

where $(\gamma I + R_{t_1\ldots t_n})z = y$ *and* $\gamma = \alpha(1-\alpha)^{-1}\delta^2$. □

Observe that the values of $\mathbf{s}_\alpha(y)$ at t_i, $1 \le i \le n$, are equal to $N(\mathbf{s}_\alpha(y)\,) = R_{t_1\ldots t_n}z = y - \gamma z$. Hence, the spline is the function from $\mathrm{span}\{\, R(t_i,\cdot) \mid 1 \le i \le n\}$ that interpolates data $\{t_i, w_i\}_{i=1}^n$, where $w_i = y_i - \gamma z_i$ are obtained by smoothing the original data y.

In the end, we discuss the case when $H = W_r^0 = W_r^0(0,1)$ is the r.k.h.s. of Example 2.12. Since in this case $L_t^{(r)}(s) = G_{r-1}(t,s)$, the representer of evaluation at t,

$$L_t(s) \;=\; (-1)^r\,\frac{(t-s)_+^{2r-1}}{(2r-1)!} \;+\; \sum_{j=0}^{r-1}(-1)^j\,\frac{t^{r+j}s^{r-j-1}}{(r+j)!(r-j-1)!},$$

is a polynomial spline. We have that $\mathbf{s}_\alpha(y)$ is the unique polynomial spline of order r corresponding to the knots t_i, $1 \le i \le n$, that satisfies the linear boundary conditions $\mathbf{s}_\alpha^{(i)}(0) = 0$, $0 \le i \le r-1$, and $\mathbf{s}_\alpha^{(r)}(1) = 0$, and interpolates data $\{t_i, w_i\}_{i,j=1}^n$.

Notes and Remarks

NR 2.15 The formulas for $\mathbf{s}_\alpha(y)$ and the optimal choice of the smoothing parameter of Section 2.6.1 seem to be new.

NR 2.16 Regularization was originally proposed by Tikhonov as a method of "solving" problems that are ill–posed, see e.g. Tikhonov [105] and Tikhonov and Arsenin [106].

Ill–posed problems in the worst case setting were studied by Werschulz [123] (see also Werschulz [124]). He proved, in particular, that if the solution operator S is unbounded then it cannot be approximated with finite error.

NR 2.17 The worst case optimality of the least squares algorithm in the case when $F = G = \mathbb{R}^n$, $S = I$ and E is the whole space, $E = \mathbb{R}^n$, was shown by Kacewicz *et al.* [31].

NR 2.18 The minimal norm properties of polynomial splines presented in Section 2.6.3 were first noticed by Schoenberg [90], see also e.g. Greville [21]. To polynomial splines is devoted the monograph of Steckin and Subbotin [99].

NR 2.19 The smoothing splines and, in particular, polynomial splines are commonly studied in numerical analysis and statistics in the context of smoothing experimental data. Continuous as well as discrete problems are considered. The main question there is how to choose the smoothing (regularization) parameter. To this end, special methods are developed. They include discrepancy principle, ordinary and generalized cross–validation, or methods based on so–called L–curve, see e.g., Golub *et al.* [18], Hansen [22], Lawson and Hanson [45], Morozow [60], Wahba [116] and references there.

We note that in most of those methods the choice of smoothing (regularization) parameter depends also on the data y. Hence, the resulting algorithms for approximating $S(f)$ are usually nonlinear in y. Moreover, in the sense of the worst case setting they can be far away from optimal; see E 2.36.

NR 2.20 Reproducing kernel Hilbert spaces are studied in Aronszajn [2], Parzen [72] [73], Vakhania *et al.* [113], Wahba [116].

NR 2.21 In the multivariate case, r.k.h.s. may be defined as tensor products of r.k.h.s.'s in the univariate case. More precisely, let H_i be the r.h.h.s. of functions $f : [a_i, b_i] \to \mathbb{R}$, and let R_i be its r.k., $1 \le i \le d$. Then the tensor product H of H_i's, $H = \bigotimes_{i=1}^d H_i$, is the r.k.h.s. of functions $f : \times_{i=1}^d [a_i, b_i] \to \mathbb{R}$ with r.k. $R(s,t) = \prod_{i=1}^d R_i(s_i, t_i)$, where $s = (s_1, \ldots, s_d), t = (t_1, \ldots, t_d) \in [0,1]^d$.

For instance, if H_i is the space $W_{r_i}^0 = W_{r_i}^0(0,1)$ of Example 2.12, the tensor product space H is given as follows. Let

$$D^{i_1 \ldots i_d} f \;=\; \frac{\partial^{i_1 + \cdots + i_d}}{\partial x_1^{i_1} \ldots \partial x_d^{i_d}}\, f.$$

Then

$$\begin{aligned} H &= W_{r_1 \ldots r_d}^{0 \ldots 0} \\ &= \{\, f : [0,1]^d \to \mathbb{R} \mid \quad & D^{r_1 - 1 \ldots r_d - 1} f\text{–abs. cont., } D^{r_1 \ldots r_d} f \in \mathcal{L}_2((0,1)^d), \\ & & D^{i_1 \ldots i_d} f(t) = 0,\ 0 \le i_j \le r_j - 1,\ 1 \le j \le d, \\ & & \text{when one of the components of } t \text{ is zero } \} \end{aligned}$$

with the inner product

$$\langle f_1, f_2 \rangle_{W_{r_1 \cdots r_d}} \;=\; \int_{[0,1]^d} (D^{r_1 \cdots r_d} f_1)(t)(D^{r_1 \cdots r_d} f_2)(t)\, dt.$$

Exercises

E 2.30 Suppose one approximates a vector $f \in E \subset \mathbb{R}^n$ based on information $y = f + x$, where $x \in B$ and $B \subset \mathbb{R}^n$ is a convex, balanced and bounded set. Let $h \in \mathbb{R}^n$ be such that $h \in \overline{B}$ and

$$\|h\|_2 \;=\; r(B) \;=\; \sup_{x\in B}\, \|x\|_2.$$

Show that if the interval $[-h,h]$ is a subset of $\overline{E}$ then the identity algorithm, $\varphi(y) = y$, is optimal and its error equals $r(B)$.

E 2.31 One may modify the least–squares algorithm as $\varphi_{mod}(y) = S(\mathbf{s}_{mod}(y)\,)$, where $\mathbf{s}_{mod}(y) \in E$ and

$$\|y - N(\mathbf{s}_{mod}(y)\,)\|_Y \;=\; \inf_{f\in E} \|y - N(f)\|_Y,$$

i.e., the minimization is taken over the set E instead of the whole space F. Show that

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{mod}) \;\le\; 2\cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}),$$

but this algorithm is in general not linear.

E 2.32 Consider the problem of Corollary 2.5 with $\lambda_s > \lambda_t$. Show that if information is exact, $\delta = 0$, then the algorithm $\varphi_\gamma(y) = \sum_{j=1}^n z_j S(f_j)$ where $(\gamma\Sigma + G_N)z = y$, is optimal for any

$$\gamma \in \left[0,\,,\, \frac{\eta_s\,\lambda_t}{\lambda_s - \lambda_t}\right].$$

E 2.33 Let S and N be linear functionals on $F = \mathbb{R}^2$, i.e., $S = \langle\cdot, s\rangle_2$ and $N = \langle\cdot, v\rangle_2$ for some $s, v \in \mathbb{R}^2$. Consider approximation of $S(f)$ for f from the unit ball based on information $y = N(f) + x$ where $|x| \le \delta$. What is the optimal value of the regularization parameter γ^* in this case? In particular, show that if s and v are linearly independent and not orthogonal, then γ^* is positive, but it tends to zero linearly with $\delta \to 0^+$.

E 2.34 Let F be a Hilbert space with an inner product $\langle\cdot,\cdot\rangle_F$. Let the set E be the ellipsoid

$$E \;=\; \{\, f \in F \mid \;\; \langle Bf, f\rangle_F \;\le\; 1\,\},$$

where $B : F \to F$ is a self–adjoint and positive definite operator. Let the information operator be defined as

$$\mathbb{N}(f) \;=\; \{\, N(f) + x \mid \;\; \|Dx\|_2 \le \delta\,\},$$

where $N : F \to \mathbb{R}^n$ and $D : \mathbb{R}^n \to \mathbb{R}^n$ are linear and continuous. Show that then the α–smoothing spline (if it exists) is the solution of the linear system

$$(\,\omega\, B \;+\; N^*D^*DN\,)\, f \;=\; N^*D^*\, y$$

where $\omega = \alpha(1-\alpha)^{-1}\delta^2$.

E 2.35 Show that if the operator B in the previous exercise is compact then the α–smoothing spline does not necessarily exist.

E 2.36 The smoothing parameter α_{ocv} given by the ordinary cross validation (ocv) is determined by the following condition:

$$\alpha_{ocv} \;=\; \alpha_{ocv}(y) \;=\; \arg \min_{0\le\alpha\le1} \|y - N(\mathbf{s}_\alpha(y)\,)\|_Y .$$

Show that the algorithm $\varphi_{ocv}(y) = S(\mathbf{s}_{\alpha_{ocv}})(y)$ is in general not optimal and the ratio $\mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi_{ocv})/\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ can be arbitrarily large.

E 2.37 Show that natural (periodic) polynomial splines are also ordinary splines for the problem of Theorem 2.13.

E 2.38 Find the natural (periodic) polynomial spline of order 1 that minimizes $\Gamma_\alpha(\mathbf{p},y)$ in the case when $\alpha = 1/2$ and $\|\cdot\|_Y$ is the Euclidean norm.

E 2.39 Let $\mathcal{T} \subset \mathbb{R}$. Find the r.k.h.s. H_R for $R(s,t) = \delta_{st}$, $s,t \in \mathcal{T}$.

E 2.40 Show that in an r.k.h.s. the functionals $f \to f(t_i)$, $1 \le i \le n$, are linearly dependent iff $\det\{\,R(t_i,t_j)\,\}_{i,j=1}^n = 0$.

E 2.41 Let H be an r.k.h.s. with positive r.k. R. Show that then the interpolation problem: find $f \in \mathrm{span}\{\,R(t_1,\cdot),\ldots,R(t_n,\cdot)\,\}$ such that $f(t_i) = y_i$, $1 \le i \le n$, has unique solution.

E 2.42 Show that the following functions are reproducing kernels:

$$\begin{aligned}
R(s,t) &= 1 - |s-t|, \qquad s,t \in [0,1],\\
R(s,t) &= \exp\{-|s-t|\}, \qquad s,t \in [0,1],\\
R(s,t) &= (\,\|s\|_2 + \|t\|_2 - \|s-t\|_2\,)/2, \qquad s,t \in [0,1]^d.
\end{aligned}$$

E 2.43 Let the functions $R_i : [0,1]\times[0,1] \to \mathbb{R}$, $1 \le i \le d$, be symmetric and nonnegative definite. Show that then the function $R : [0,1]^d \times [0,1]^d \to \mathbb{R}$, $R(s,t) = \prod_{i=1}^d R_i(s_i,t_i)$ also is symmetric and nonnegative definite.

E 2.44 Let $0 < t_1 < t_2 < \cdots < t_n$ and let $M = \{\min\{t_i,t_j\}\,\}_{i,j=1}^n$. Show that $M = M^* > 0$, and that the inverse M^{-1} is given as

$$\begin{bmatrix}
\frac{1}{t_1}+\frac{1}{t_2-t_1}, & \frac{-1}{t_2-t_1}, & & & \\
\frac{-1}{t_2-t_1}, & \frac{1}{t_2-t_1}+\frac{1}{t_3-t_2}, & \frac{-1}{t_3-t_2}, & & 0\\
& \frac{-1}{t_3-t_2}, & \frac{1}{t_3-t_2}+\frac{1}{t_4-t_3}, & \ddots & \\
& & \ddots & \ddots & \frac{-1}{t_n-t_{n-1}}\\
& 0 & & \frac{-1}{t_n-t_{n-1}}, & \frac{1}{t_n-t_{n-1}}
\end{bmatrix}.$$

2.7 Varying information

So far the information operator $\mathbb{N}$ was regarded as given and we had been looking for the optimal algorithm. In this section, we assume that not only the algorithm but also the information operator can vary.

2.7.1 Nonadaptive and adaptive information

Let Λ be a given class of functionals over the space F. We assume that we can collect information about f only by noisy observations of functionals L_i at f, where L_i belong to Λ. Each such an observation is performed with some precision δ_i which also can vary.

More specifically, a (noisy) *nonadaptive* information operator $\mathbb{N}$ is determined by an exact information operator $N : F \to \mathbb{R}^n$ of the form

$$N(f) \;=\; [\,L_1(f), L_2(f), \ldots, L_n(f)\,], \qquad \forall\, f \in F,$$

where the functionals $L_i : F \to \mathbb{R}$ belong to Λ, and by a *precision vector*

$$\Delta \;=\; [\,\delta_1, \delta_2, \ldots, \delta_n\,],$$

where $\delta_i \ge 0$, $1 \le i \le n$. When using N and Δ we obtain information $y = N(f) + x$, where the noise x is known to belong to a given set $B = B(\Delta, N(f))$ of $\mathbb{R}^n$. That is, the nonadaptive information operator $\mathbb{N}$ is identified with the pair N, Δ, i.e., $\mathbb{N} = \{N, \Delta\}$, and it is formally given as $\mathbb{N} : F \to 2^Y$, where $Y = \mathbb{R}^n$ and

$$\mathbb{N}(f) \;=\; \{\, y \in \mathbb{R}^n \mid \quad x = (y - N(f)\,) \in B(\Delta, N(f)\,)\,\}. \tag{2.39}$$

We may consider, for instance,

$$B(\Delta, N(f)\,) \;=\; B(\Delta) \;=\; \{\, x \in \mathbb{R}^n \mid \quad |x_i| \le \delta_i,\ 1 \le i \le n\,\}, \tag{2.40}$$

which means that for each i the value of $L_i(f)$ is observed with absolute error δ_i, $|y_i - L_i(f)| \le \delta_i$. This definition of $B(\Delta, N(f))$ seems to be most natural. However, we can also have a more complicated dependence of the noise on the precision vector. Namely,

$$B(\Delta, N(f)\,) \;=\; B(\Delta) \;=\; \left\{ x \in \mathbb{R}^n \,\Big|\, \quad \sum_{i=1}^{n} \frac{x_i^2}{\delta_i^2} \le 1 \right\}, \tag{2.41}$$

which corresponds to noise bounded in the weighted Euclidean norm, or more generally,

$$B(\Delta, N(f))\,) \;=\; B(\Delta) \;=\; \left\{ x \in \mathbb{R}^n \;\Big|\; \sum_{i=1}^n \frac{|x_i|^p}{\delta_i^p} \le 1 \right\}, \quad p \ge 1. \qquad (2.42)$$

Contrary to (2.40), the bound on noise x_i coming from the ith observation now depends on $x_1, \ldots, x_{i-1}$. Namely, $|x_i| \le (1 - \sum_{j=1}^{i-1} |x_j|^p/\delta_i^p)^{1/p}$. In particular, if $|x_1| = \delta_1$ then the next observations are performed exactly, $x_2 = \cdots = x_n = 0$.

In (2.40) to (2.42) the noise is independent of the exact information $N(f)$. The noise depends on $N(f)$ when, for instance, the relative error is considered. In this case,

$$B(\Delta, N(f)\,) \;=\; \{\, x \in \mathbb{R}^\infty \;|\; \; |x_i| \le \delta_i\, |L_i(f)|,\, 1 \le i \le n \,\}. \qquad (2.43)$$

In general, we assume that the sets $B(\Delta, z\,)$, for $\Delta, z \in \mathbb{R}^n$ and $n \ge 1$, satisfy the following conditions.

1. $$B(0, z) \;=\; \{0\}.$$

2. If $0 \le \delta_i \le \delta_i'$, $1 \le i \le n$, then
$$B\left([\delta_1, \ldots, \delta_n], z\right) \;\subset B\left([\delta_1', \ldots, \delta_n'], z\right).$$

3. Let $z^n = [z_1, \ldots, z_n]$, $\Delta^n = [\delta_1, \ldots, \delta_n]$, and $z^{n+1} = [z_1, \ldots, z_n, z_{n+1}]$, $\Delta^{n+1} = [\delta_1, \ldots, \delta_n, \delta_{n+1}]$. Then
$$B(\Delta^n, z^n\,) \;=\; \{\, x \in \mathbb{R}^n \;|\; \; \exists a \in \mathbb{R},\; [x, a] \in B(\Delta^{n+1}, z^{n+1}\,)\,\}. \qquad (2.44)$$

The first condition means that the zero precision vector corresponds to the exact information. The second condition says that we decrease the noise by decreasing the precisions δ_i. The third condition indicates a relation between the noise of successive observations. It states that from the nth observation we can pass to the $(n+1)$st observation. Indeed, suppose that there is a noise vector $x^n = (y^n - N^n(f)) \in B(\Delta^n, N^n(f)\,)$ that cannot be extended to $[x^n, a] \in B(\Delta^{n+1}, N^{n+1}(f)\,)$. This means that then the noisy observation of $L_{n+1}(f)$ is impossible. Similarly, suppose that for some $x^{n+1} = (y^{n+1} - N^{n+1}(f)) \in B(\Delta^{n+1}, N^{n+1}(f))$ we have $x^n = (y^n - N^n(f)) \notin B(\Delta^n, N^n(f))$. Then y^n is not information about f, although this vector comes from the first n observations.

We left as an exercise to show that all the three conditions are satisfied by the noise defined by (2.40) to (2.43).

We also admit a more general class of *adaptive* information where decisions about successive observations are made based on previously obtained values y_i. The effect of adaption can be obtained by adaptive choice of:

- information functionals L_i, or
- precisions δ_i, or
- the number n of observations.

Formally, a (noisy) adaptive information operator $\mathbb{N} : F \to 2^Y$ is determined by a family $N = \{N_y\}_{y\in Y}$ of exact information operators of the form

$$N_y \;=\; [\,L_1(\cdot), L_2(\cdot; y_1), \ldots, L_{n(y)}(\cdot; y_1, \ldots, y_{n(y)-1})\,],$$

where for all $1 \le i \le n$ and $y_1, \ldots, y_{i-1}$ the functionals $L_i(\cdot; y_1, \ldots, y_{i-1}) \in \Lambda$, and by a family $\Delta = \{\Delta_y\}_{y\in Y}$ of precision vectors,

$$\Delta_y \;=\; [\,\delta_1, \delta_2(y_1), \ldots, \delta_{n(y)}(y_1, \ldots, y_{n(y)-1})\,].$$

($n(y)$ denotes here the length of y.) We also assume that the set Y (the range of $\mathbb{N}$) satisfies the following condition:

$$\text{for any } [y_1, y_2, \ldots] \in \mathbb{R}^\infty \quad \begin{array}{l}\text{there exists exactly one index } n\\ \text{such that } [y_1, \ldots, y_n] \in Y.\end{array} \tag{2.45}$$

For the noisy adaptive information operator $\mathbb{N} = \{N, \Delta\}$ we have

$$\mathbb{N}(f) \;=\; \{\, y \in Y \mid \quad x = (y - N_y(f)) \in B(\Delta_y, N_y(f))\,\}.$$

The essence of the above definition is as follows. At the ith step of gaining information, we observe a noisy value y_i of $L_i(f; y_1, \ldots, y_{i-1})$ with precision $\delta_i(y_1, \ldots, y_{i-1})$. Then we check whether the condition $[y_1, \ldots, y_i] \in Y$ [3] is satisfied. If the answer is "yes", the observations are terminated and $[y_1, \ldots, y_i]$ is noisy information about f. Otherwise we proceed to the $(i{+}1)$st

[3] This is what in practical computations is often called the *termination criterion* or *stopping rule.*

step. Note that (2.45) assures that the observations will be terminated after a finite number of steps. The resulting information y about f satisfies $(y - N_y(f)) \in B(\Delta_y, N_y(f))$, as though we used nonadaptive information $\{N_y, \Delta_y\}$.

Clearly, any nonadaptive information $\mathbb{N} = \{N, \Delta\}$ of the form (2.39) can be considered as adaptive since then $Y = \mathbb{R}^n$, $N_y = N$, and $\Delta_y = \Delta$, $\forall y \in Y$.

To stress what kind of information we deal with, we sometimes add the superscripts "ad" and "non", and write $\mathbb{N}^{\mathrm{ad}}$ and $\mathbb{N}^{\mathrm{non}}$ for adaptive and nonadaptive information, respectively.

2.7.2 When does adaption not help?

It is clear that adaptive information has a much richer structure than nonadaptive information and therefore should usually lead to better approximations. This is however not always true. We shall give a sufficient condition under which adaption does not help much.

For an (in general adaptive) information operator $\mathbb{N} : F \to 2^Y$, let $Y_0 = \bigcup_{f \in E} \mathbb{N}(f) \subset Y$ be the set of all possible information values. For $y \in Y_0$, let

$$A_{\mathbb{N}}(y) \;=\; \{\, S(f) \mid \quad f \in E,\; y \in \mathbb{N}(f)\,\}.$$

We shall say that $f^* \in E$ is a *κ–hard element* iff for any nonadaptive information operator $\mathbb{N} = \{N, \Delta\}$ of the form (2.39) we have

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;\le\; \kappa \cdot r(\, A_{\mathbb{N}}(N(f^*))\,).$$

(Recall that $r(B)$ is the usual radius of the set B, see Section 2.3.)

Suppose that the κ–hard element f^* exists. Let $\mathbb{N}^{\mathrm{ad}}$ be an arbitrary adaptive information operator corresponding to a set Y, family

$$N_y \;=\; [\,L_1(\cdot), L_2(\cdot; y_1), \ldots, L_{n(y)}(\cdot; y_1, \ldots, y_{n(y)-1})\,]$$

and precisions Δ_y, $y \in Y$. Let $y^* \in Y$ be given as $y_1^* = L_1(f^*)$ and $y_i^* = L_i(f^*; y_1^*, \ldots, y_{i-1}^*)$ for $2 \le i \le n^*$, where n^* is the length of y^*, i.e., the minimal n for which $[y_1^*, \ldots, y_n^*] \in Y$. Define a nonadaptive information operator $\mathbb{N}^{\mathrm{non}} = \{N^{\mathrm{non}}, \Delta^{\mathrm{non}}\}$ where

$$N^{\mathrm{non}} \;=\; N_{y^*} \;=\; [\,L_1(\cdot), L_2(\cdot; y_1^*), \ldots, L_{n^*}(\cdot; y_1^*, y_2^*, \ldots, y_{n^*-1}^*)\,]$$

and

$$\Delta^{\rm non} \;=\; \Delta_{y^*} \;=\; [\,\delta_1, \delta_2(y_1^*), \ldots, \delta_{n^*}(y_1^*, \ldots, y_{n^*-1}^*)].$$

It turns out that nonadaptive information $\mathbb{N}^{\rm non}$ is almost as good as adaptive information $\mathbb{N}^{\rm ad}$.

Theorem 2.15

$${\rm rad}^{\rm wor}(\mathbb{N}^{\rm non}) \;\le\; \kappa \cdot {\rm rad}^{\rm wor}(\mathbb{N}^{\rm ad}).$$

Proof Observe that $A_{\mathbb{N}^{\rm non}}(y^*) = A_{\mathbb{N}^{\rm ad}}(y^*)$. Hence,

$$\begin{aligned} {\rm rad}^{\rm wor}(\mathbb{N}^{\rm non}) \;&\le\; \kappa \cdot r(A_{\mathbb{N}^{\rm non}}(y^*)\,) \\ &=\; \kappa \cdot r(A_{\mathbb{N}^{\rm ad}}(y^*)\,) \;\le\; \kappa \cdot {\rm rad}^{\rm wor}(\mathbb{N}^{\rm ad}), \end{aligned}$$

as claimed. □

The meaning of Theorem 2.15 is evident when information uses only adaptive choice of functionals, and the precision vector Δ and the number n of functionals are fixed. Then the existence of the κ–hard element suffices for the adaptive information to be no more than κ times better than some nonadaptive information that uses the same number n of functionals with the same precision Δ.

The κ–hard element exists for some important problems. Consider first the case of linear solution operator S with convex and balanced set $E \subset F$. We assume that the class Λ consists of some linear functionals, and that the set $B(\Delta, z)$ is the unit ball in an extended seminorm $\|\cdot\|_\Delta$ (which can depend on Δ),

$$B(\Delta, z) \;=\; B(\Delta) \;=\; \{\,x \in \mathbb{R}^n \mid \;\; \|x\|_\Delta \le 1\,\}, \quad \Delta, z \in \mathbb{R}^n,\; n \ge 1. \tag{2.46}$$

Observe that then any nonadaptive information is linear with noise bounded uniformly in an extended seminorm. Lemma 2.2 and Theorem 2.3 yield that for any nonadaptive information $\mathbb{N}$ using n observations we have

$${\rm rad}^{\rm wor}(\mathbb{N}) \;\le\; 2 \cdot r(A_{\mathbb{N}}(\underbrace{0, \ldots, 0}_{n})\,).$$

Hence, the zero element of F is the κ–hard element with $\kappa = 2$. If S is a functional, or if E is a ball in a Hilbert extended seminorm, $\|\cdot\|_\Delta$ is a Hilbert extended seminorm, and G is a Hilbert space, then we can even take $\kappa = 1$ since in these cases ${\rm diam}(\mathbb{N}) = 2\,{\rm rad}^{\rm wor}(\mathbb{N})$.

Corollary 2.6 *Suppose that the set E is convex and balanced, the solution operator S is linear, the class Λ consists of some linear functionals, and $B(\Delta, z)$ is of the form (2.46). Then for any adaptive information $\mathbb{N}^{\mathrm{ad}} = \{N_y, \Delta_y\}_{y \in Y}$ we have*

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{non}}) \ \le \ 2 \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{ad}}),$$

where $\mathbb{N}^{\mathrm{non}}$ is the nonadaptive information constructed as in Theorem 2.15 for $f^* = 0$. If, additionally, S is a linear functional or if we have the Hilbert case, then $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{ad}}) \le \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{non}})$. □

We now show an example in which f^* exists and is not zero, although E is a convex and balanced set. Let F be a normed space and let E be the unit ball in F. Consider a nonadaptive linear information operator with noise bounded in the relative sense,

$$\mathbb{N}(f) \ = \ \{\, y \in \mathbb{R}^n \mid \ \ |y_i - L_i(f)| \le \delta_i \cdot |L_i(f)|, \ 1 \le i \le n \,\}, \qquad (2.47)$$

where $0 \le \delta_i < 1$ and $\|L_i\|_F \le 1$, $1 \le i \le n$. Then for any linear solution operator S and for any f with $\|f\|_F \le 1$, we have

$$\begin{aligned} r(A_{\mathbb{N}}(N(f)\,)\,) \ &\ge \ \frac{1}{2}\, d(A_{\mathbb{N}}(N(f)\,)\,) \\ &\ge \ \frac{1}{2}\, \sup\{\, \|S(f+h) - S(f-h)\| \mid \ \ \|h\|_F \le 1 - \|f\|_F, \\ &\qquad |L_i(h)| \le \delta_i |L_i(f \pm h)|, \ 1 \le i \le n \,\} \\ &\ge \ \sup\,\{\, \|S(h)\| \mid \ \ \|h\| \le 1 - \|f\|, \\ &\qquad |L_i(h)| \le \frac{\delta_i}{1+\delta_i} |L_i(f)|, \ 1 \le i \le n \Big\} \\ &\ge \ \min\{1 - \|f\|_F, \ a(f)/2\} \\ &\qquad \sup\{\, \|S(h)\| \mid \ \ \|h\|_F \le 1, \ |L_i(h)| \le \delta_i, \ 1 \le i \le n \,\}, \end{aligned}$$

where $a(f) = \min_{1 \le i \le n} |L_i(f)|$. The last supremum is equal to the half of the diameter of the same linear information, but with noise bounded in the absolute sense. Since $\|L_i\|_F \le 1$, the inequality $|y_i - L_i(f_1)| \le \delta_i |L_i(f_1)|$ implies $|y_i - L_i(f_1)| \le \delta_i$. Hence, the diameter of information with noise bounded in the absolute sense is not smaller than diameter of information bounded in the relative sense. We obtain

$$r(A_{\mathbb{N}}(N(f)\,)\,) \ \ge \ \frac{1}{2} \min\left\{ 1 - \|f\|_F, \ \frac{1}{2}\, a(f) \right\} \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}).$$

Suppose now that F is the space of functions $f : [0,1] \to \mathbb{R}$ with rth continuous derivative. Let

$$\|f\|_F \;=\; \max_{1\le i\le r}\; \sup_{0\le t\le 1} |f^{(i)}(t)|. \tag{2.48}$$

The class Λ consists of functionals of the form $L(f) = f(t)$, for some $t \in [0,1]$. Then, taking $f^* \equiv 2/3$ we have $1/2 \cdot \min\{1 - \|f^*\|_F,\, a(f^*)/2\} = 1/6$. Hence, f^* is the 6–hard element and adaption cannot be much better than nonadaption. That is,

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{non}}) \;\le\; 6 \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{ad}})$$

with $\mathbb{N}^{\mathrm{non}} = \{N_{2/3}, \Delta_{2/3}\}$.

We left to the reader to verify that in this case zero is not the κ–hard element (see E 2.47).

Notes and Remarks

NR 2.22 A similar general model with varying noisy information, but with fixed noise level is considered in Traub *et al.* [107, Chap.4]. The method of showing when adaption does not help is adopted from that book.

NR 2.23 We have shown that for convex and balanced set E, and for linear information with noise bounded in an extended seminorm, adaptive information can be at most twice better than nonadaptive information. It has long been an open problem whether adaption helps at all. An example of a problem (with exact information) where adaption helps a little was given by Kon and Novak [39] [40].

NR 2.24 Korneichuk [41], Novak [64] [65] [62] considered the problem of adaption (and n-widths) for convex but nonbalanced sets. For such sets adaptive information can be significantly better than nonadaptive information. An example is given in E 2.48.

NR 2.25 The fact that adaption may be not much better than nonadaption in the case of relative perturbations was noticed by Kacewicz and Plaskota [32].

NR 2.26 Adaptive and nonadaptive information are also frequently called sequential and parallel, or active and passive, respectively.

Exercises

E 2.45 Show that if $B(\Delta, N(f)\,) = B(\Delta)$ is the unit ball in an extended seminorm $\|\cdot\|_\Delta$ then (2.44) is equivalent to the following condition. Let $\Delta = [\delta_1, \ldots, \delta_n]$ and

$\Delta' = [\delta_1, \ldots, \delta_n, \delta_{n+1}]$. Then

$$\|x\|_\Delta \;=\; \min_{a\in\mathbb{R}} \|[x,a]\|_{\Delta'}, \qquad \forall x \in \mathbb{R}^n.$$

E 2.46 Show that for any information operators N^n, $N^{n+1} = [N^n, L_{n+1}]$, and precision vectors Δ^n, $\Delta^{n+1} = [\Delta^n, \delta_{n+1}]$, we have

$$\mathrm{rad}^{\mathrm{wor}}(N^n, \Delta^n) \;\le\; \mathrm{rad}^{\mathrm{wor}}(N^{n+1}, \Delta^{n+1}).$$

E 2.47 Let F be the space of functions $f : [0,1] \to \mathbb{R}$ with continuous rth derivative and with the norm (2.48). Let E be the unit ball in F. Show an example of a solution operator that the following holds. For any $\kappa < +\infty$ there is an information operator $\mathbb{N}$ of the form (2.47) such that $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) > \kappa \cdot r(A_{\mathbb{N}}(0))$.

E 2.48 (Novak) Let

$$E \;=\; \left\{ f \in \mathbb{R}^\infty \;\middle|\; f_i \ge 0,\, \sum_{i=1}^{\infty} f_i \le 1,\, f_k \ge \max\{f_{2k}, f_{2k+1}\} \right\}.$$

Consider approximation of $f \in E$ in the l_∞-norm from exact information of the form $N(f) = [f_{j_1}, \ldots, f_{j_n}]$.
1. Show that the radius of nonadaptive information using n observations is minimal for $N_n(f) = [f_1, \ldots, f_n]$, and

$$\mathrm{rad}^{\mathrm{wor}}(N_n) \;\approx\; \frac{1}{\log_2(n+1)}, \qquad \text{as} \quad n \to \infty.$$

2. Find adaptive information N_n^{ad} that uses exactly n observations of f_j for which

$$\mathrm{rad}^{\mathrm{wor}}(N_n^{\mathrm{ad}}) \;\le\; \frac{1}{n+3}.$$

2.8 Optimal information

Suppose that n and the precision vector $\Delta = [\delta_1, \delta_2, \ldots, \delta_n]$ are given. Then it makes sense to ask for the minimal error that can be achieved when noisy observations of n functionals from the class Λ with precisions δ_i are used. We formalize this issue in the following way.

Let $\mathcal{N}_n$ be the class of exact information operators consisting of n functionals, i.e., $N \in \mathcal{N}_n$ iff

$$N \;=\; [\,L_1, L_2, \ldots, L_n\,],$$

for some $L_i \in \Lambda$, $1 \le i \le n$. Let $\mathrm{rad}^{\mathrm{wor}}(N, \Delta)$ denote the radius of noisy information $\mathbb{N}$ corresponding to N and precision vector Δ.

The *minimal radius* corresponding to the precision vector Δ is given as

$$\mathrm{r}_n^{\mathrm{wor}}(\Delta) \;=\; \inf_{N\in\mathcal{N}_n} \mathrm{rad}^{\mathrm{wor}}(N,\Delta).$$

If for some $N_\Delta \in \mathcal{N}_n$ is

$$\mathrm{r}_n^{\mathrm{wor}}(\Delta) \;=\; \mathrm{rad}^{\mathrm{wor}}(N_\Delta,\Delta)$$

then N_Δ is called an *optimal information.*

We shall find the minimal radius and optimal information in two special cases: for linear problems defined in Hilbert spaces, and for approximation and integration of Lipschitz functions.

2.8.1 Linear problems in Hilbert spaces

We assume that F and G are separable Hilbert spaces and that the solution operator $S : F \to G$ is compact. The set E is the unit ball in F. The class Λ of permissible information functionals consists of all linear functionals with norm bounded by 1,

$$\Lambda \;=\; \left\{ L\text{–linear functional} \;\middle|\; \|L\|_F \;=\; \sup_{\|f\|_F=1} |L(f)| \;\le\; 1 \right\}.$$

We also assume that the observation noise is bounded in the weighted Euclidean norm, $\|x\|_Y \;=\; \big(\sum_{i=1}^\infty x_i^2/\delta_i^2\big)^{1/2}$. Hence, for given $\Delta = [\delta_1,\ldots,\delta_n]$ and $N \;=\; [\,L_1,\ldots,L_n\,] \in \mathcal{N}_n$, a vector $y \in \mathbb{R}^n$ is noisy information about $f \in F$ iff

$$\sum_{i=1}^n \frac{1}{\delta_i^2}\,(\,y_i - L_i(f)\,)^2 \;\le\; 1.$$

To cover the exact information case, we also allow $\delta_i = 0$. In this case, we formally set $x_i/\delta_i = 0$ for $x_i = 0$, and $x_i/\delta_i = +\infty$ otherwise. Note that if all δ_i's are equal, $\delta_i = \delta$, then $\sum_{i=1}^n (y_i - L_i(f)\,)^2 \le \delta^2$, i.e., the noise is bounded by δ in the Euclidean norm.

Before stating a theorem about optimal information, we first introduce a necessary notation. Let $d =$ dim F. Let $\{\xi_i\}_{i=1}^d$ be a complete orthonormal in F system of eigenelements of the operator S^*S. Let λ_i be the corresponding eigenvalues,

$$S^*S\,\xi \;=\; \lambda_i\,\xi_i.$$

Since S is compact, we can assume without loss of generality that λ_i's are ordered, $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. We consider the sequence $\{\lambda_i\}$ to be infinite by setting, if necessary, $\lambda_i = 0$ for $i > d$. Similarly, $\xi_i = 0$ for $i > d$. For $d = +\infty$ or $d < +\infty$ we have $\lim_{i\to\infty} \lambda_i = 0$.

We also need the following important lemma.

Lemma 2.14 *Let the nonincreasing sequences $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n \geq 0$ and $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_n \geq 0$ be such that*

$$\sum_{i=r}^{n} \eta_i \leq \sum_{i=r}^{n} \beta_i, \qquad 1 \leq r \leq n\,,$$

and $\sum_{i=1}^n \eta_i = \sum_{i=1}^n \beta_i$. *Then there exists a real matrix $W = \{w_{ij}\}_{i,j=1}^n$ for which*

$$\sum_{s=1}^{n} w_{is}^2 \;=\; \beta_i \qquad \textit{and} \qquad \sum_{s=1}^{n} w_{si} w_{sj} \;=\; \eta_i \delta_{ij}\,,$$

for all $1 \leq i, j \leq n$ (δ_{ij} stands for the Kronecker delta).

Proof We shall construct the matrix W using induction on n. For $n = 1$ we obviously have $\eta_1 = \beta_1$ and $w_{11} = \sqrt{\eta_1}$. Let $n \geq 2$. If $\eta_i = \beta_i$, $1 \leq i \leq n$, then $W = \operatorname{diag}\{\sqrt{\eta_1}, \ldots, \sqrt{\eta_n}\}$. Otherwise there is an index s, $1 \leq s \leq n-1$, such that $\eta_s > \beta_s \geq \eta_{s+1}$. Set $\bar\eta = \eta_s + \eta_{s+1} - \beta_s > 0$. Let $U \in \mathbb{R}^{(n-1)\times(n-1)}$ be the required matrix for the sequences $\beta_1 \geq \cdots \geq \beta_{s-1} \geq \beta_{s+1} \geq \cdots \geq \beta_n$ and $\eta_1 \geq \cdots \geq \eta_{s-1} \geq \bar\eta \geq \eta_{s+2} \geq \cdots \geq \eta_n$. Let $u_i \in \mathbb{R}^{n-1}$ be the columns of U, $1 \leq i \leq n-1$. Let

$$\begin{aligned} a \;&=\; \left(\frac{\eta_{s+1}(\eta_s - \beta_s)}{\bar\eta(\eta_s - \eta_{s+1})}\right)^{1/2}, & b \;&=\; (1-a^2)^{1/2}\,, \\ c \;&=\; \left(\frac{\eta_s(\beta_s - \eta_{s+1})}{(\eta_s - \eta_{s+1})}\right)^{1/2}, & d \;&=\; -(1-c^2)^{1/2}\,. \end{aligned}$$

Elementary calculations show that the desired matrix is $W = \{w_1, \ldots, w_n\}$, $w_i \in \mathbb{R}^n$, $1 \leq i \leq n$, where

$$\begin{aligned} w_i \;&=\; (\,u_i^T, 0\,)^T, \quad \text{for } i \neq s, s+1, \\ w_s \;&=\; (\,a u_s^T, c\,)^T, \\ w_{s+1} \;&=\; (\,b u_s^T, d\,)^T \end{aligned}$$

(the superscript "T" denotes transposition). □

For the precision vector $\Delta = [\delta_1, \ldots, \delta_n]$, we assume without loss of generality that

$$0 \;=\; \delta_1 \;=\; \cdots \;=\; \delta_{n_0} \;<\; \delta_{n_0+1} \;\le\; \cdots \;\le \delta_n.$$

(If all δ_i's are positive then $n_0 = 0$.) It turns out that the following minimization problem plays a crucial role in finding the optimal information N_Δ.

Problem (MP) *Minimize*

$$\Omega(\alpha;\, \eta_{n_0+1}, \ldots, \eta_n) \;=\; \max_{n_0+1 \le i \le n+1} \frac{\lambda_i}{\alpha + (1-\alpha)\,\eta_i}$$

over all $0 \le \alpha \le 1$ *and* $\eta_{n_0+1} \ge \cdots \ge \eta_{n+1} = 0$ *satisfying*

$$\sum_{i=r}^{n} \eta_i \;\le\; \sum_{i=r}^{n} \frac{1}{\delta_i^2}\,, \qquad n_0 + 1 \le r \le n\,, \tag{2.49}$$

and $\sum_{i=n_0+1}^{n} \eta_i \;=\; \sum_{i=n_0+1}^{n} \delta_i^{-2}$ *(as before,* $a/0 = +\infty$ *for* $a > 0$ *and* $0/0 = 0$*, by convention).*

Theorem 2.16 *Let* α^* *and* $\eta^*_{n_0+1} \ge \cdots \ge \eta^*_n$ *be the solution of* (MP)*. Then the minimal radius*

$$\mathrm{r}_n^{\mathrm{wor}}(\Delta) \;=\; \sqrt{\Omega(\alpha^*; \eta^*_{n_0+1}, \ldots, \eta^*_n)}.$$

Furthermore, the optimal information is given as

$$N_\Delta \;=\; [\;\langle\,\cdot\,,\xi_1\rangle_F\,, \ldots, \langle\,\cdot\,,\xi_{n_0}\rangle_F, \langle\,\cdot\,,\xi^*_{n_0+1}\rangle_F, \ldots, \langle\,\cdot,\xi^*_n\rangle_F\;],$$

where

$$\xi^*_{n_0+i} \;=\; \delta_{n_0+i} \sum_{j=1}^{n-n_0} w_{ij}\xi_{n_0+j},$$

and $W = \{w_{ij}\}_{i,j=1}^{n-n_0}$ *is the matrix from Lemma 2.14 applied for*

$$\eta_i \;=\; \eta^*_{n_0+i} \qquad \text{and} \qquad \beta_i \;=\; \frac{1}{\delta^2_{n_0+i}}\,,$$

$1 \le i \le n - n_0\,.$

Proof Consider first the case when all δ_i's are positive, $n_0 = 0$. Let

$$N \;=\; [\, \langle \cdot\,, f_1\rangle_F, \langle \cdot\,, f_2\rangle_F, \ldots, \langle \cdot\,, f_n\rangle_F \,]$$

with $\|f_i\|_F \le 1$, be an arbitrary information operator. In fact, we can assume that $\|f_i\|_F \;=\; 1$, $1 \le i \le n$, since multiplying f_i by a constant larger than one we can only increase the accuracy of noisy information. Due to Lemma 2.10, the radius of N is the minimal norm of the operator $SA_\alpha^{-1/2}$ with respect to all $\alpha \in [0,1]$. In our case,

$$A_\alpha \;=\; \alpha\, I \;+\; (1-\alpha)\, N^*N,$$

where $N^* : Y \to F,\;\; N^*(y) \;=\; \sum_{i=1}^n y_i f_i/\delta_i^2$. Then

$$\|\, SA_\alpha^{-1/2}\,\|_F^2 \;=\; \sup_{h\ne 0} \frac{\|SA_\alpha^{-1/2}h\|_F^2}{\|h\|_F^2} \;=\; \sup_{h\ne 0} \frac{\langle S^*Sh, h\rangle_F}{\langle A_\alpha h, h\rangle_F}\,.$$

Taking $h = \xi_i$, $1 \le i \le n$, we obtain

$$\|\, SA_\alpha^{-1/2}\|_F^2 \;\ge\; \max_{1\le i\le n} \frac{\lambda_i}{\langle A_\alpha \xi_i, \xi_i\rangle_F}\,. \tag{2.50}$$

For $d > n$, we get an additional lower bound. Namely, since the operator N is at most n–dimensional, there exists a nonzero element $h_0 \in \mathrm{span}\{\xi_1, \ldots, \xi_{n+1}\}$ such that $N(h_0) = 0$. Hence,

$$\|\, SA_\alpha^{1/2}\,\|_F^2 \;\ge\; \frac{\langle S^*Sh_0, h_0\rangle_F}{\langle A_\alpha h_0, h_0\rangle_F} \;\ge\; \frac{\lambda_{n+1}}{\alpha}. \tag{2.51}$$

Let $\eta_1 \ge \eta_2 \ge \ldots \ge \eta_n \ge 0 = \eta_{n+1}$ be the eigenvalues of N^*N. Then η_i's are also eigenvalues of the operator $NN^* : Y \to Y$ whose matrix (in the versor basis $\{e_i\}$) is

$$M \;=\; \left\{ \delta_j^{-2} \langle f_i, f_j\rangle_F \right\}_{i,j=1}^n\,.$$

Since $\tilde{e}_i = \delta_i e_i$, $1 \le i \le n$, is an orthonormal basis in Y, for $1 \le r \le n$ we have

$$\sum_{i=r}^n \eta_i \;\le\; \sum_{i=r}^n \langle M\tilde{e}_i, \tilde{e}_i\rangle_Y \;=\; \sum_{i=r}^n \frac{1}{\delta_i^2}\,, \tag{2.52}$$

and $\sum_{i=1}^n \eta_i \;=\; \sum_{i=1}^n \delta_i^{-2}$.

Taking together (2.50), (2.51) and (2.52), we obtain the following lower bound on $\mathrm{r}_n^{\mathrm{wor}}(\Delta)$,

$$(\mathrm{r}_n^{\mathrm{wor}}(\Delta))^2 \;\geq\; \min \max_{1\leq i\leq n+1} \frac{\lambda_i}{\alpha+(1-\alpha)\,\eta_i}\,, \tag{2.53}$$

where the minimum is taken over all $\alpha \in [0,1]$ and over all η_i's satisfying (2.49). To complete the proof of the lower bound, observe that the minimum in (2.53) is attained for some η_i^*'s satisfying $\eta_1^* \geq \cdots \geq \eta_n^*$.

We now show that the lower bound (2.53) is attained for the information operator N_Δ. To this end, it suffices to show that all ξ_i's are the eigenelements of the operator $N_\Delta^* N_\Delta$ and that the corresponding eigenvalues are η_i^*. Indeed, we have

$$\begin{aligned} N_\Delta^* N_\Delta\,\xi_i &= \sum_{s=1}^n \delta_s^{-2}\,\langle \xi_i, \xi_s^*\rangle_F \xi_s^* \;=\; \sum_{s=1}^n \left\langle \xi_i, \sum_{t=1}^n w_{st}\xi_t \right\rangle_F \left(\sum_{j=1}^n w_{sj}\xi_j \right) \\ &= \sum_{s=1}^n \sum_{j=1}^n w_{si}w_{sj}\xi_j \;=\; \sum_{j=1}^n \left(\sum_{s=1}^n w_{si}w_{sj} \right) \xi_j \\ &= \sum_{j=1}^n \eta_i^* \delta_{ij}\xi_j \;=\; \eta_i^*\,\xi_i\,. \end{aligned}$$

Since

$$\|\langle\,\cdot\,,\xi_i^*\rangle\|_F \;=\; \|\xi_i^*\|_F^2 \;=\; \delta_i^2 \sum_{j=1}^n w_{ij}^2 \;=\; 1,$$

N_Δ is also a permissible information operator, $N_\Delta \in \mathcal{N}_n$. This completes the proof of the case $n_0 = 0$.

Suppose now that not all δ_i's are positive, $n_0 \geq 1$. Then for any $N = [L_1, \ldots, L_n] \in \mathcal{N}_n$ we have

$$\begin{aligned} \mathrm{rad}^{\mathrm{wor}}(N,\Delta) &= \sup\{\, \|S(h)\| \;|\; \|h\|_F \leq 1,\ \|N(h)\|_Y \leq 1 \,\} \\ &= \sup\{\, \|S_1(h)\| \;|\; \|h\|_F \leq 1,\ \|N_1(h)\|_{Y_1} \leq 1 \,\}, \end{aligned}$$

where $F_1 = \{\, f \in F \,|\, L_i(f) = 0,\ 1 \leq i \leq n_0 \,\}$, $S_1 : F_1 \to G$ is the restriction of S to the space F_1, $S_1 = S_{|F_1}$, information operator $N_1 = [L_{n_0+1}, \ldots, L_n]$, and $\|\cdot\|_{Y_1}$ is the extended seminorm on $\mathbb{R}^{n-n_0}$ defined as $\|x\|_{Y_1} = \|[0,x]\|_Y$. It is known that the dominating eigenvalues $\lambda_1' \geq \lambda_2' \geq \ldots$ of the operator $S_1^* S_1 : F_1 \to F_1$ satisfy $\lambda_i' \geq \lambda_{n_0+i}$, $\forall i \geq 1$. Moreover, for $L_j = \langle\,\cdot\,,\xi_j\rangle_F$,

$1 \le i \le n_0$, we have $\lambda_i' = \lambda_{n_0+i}$, $\forall i \ge 1$. Thus, we obtain the desired result by reducing our problem to that of finding optimal $N_1 \in \mathcal{N}_{n-n_0}$ for approximation of S_1 from data $y \in \mathbb{R}^{n-n_0}$ satisfying $\|y - N_1(f)\|_{Y_1} \le 1$. □

Thus, to construct the optimal information N_Δ, we first have to solve the minimization problem (MP) and then find the matrix W. Solution of (MP) will be given below. The matrix W can be found following the construction from the proof of Lemma 2.14. Note that the optimal approximation φ_Δ is given by the α^*–smoothing spline where α^* comes from the solution of (MP).

We now show how to solve the problem (MP). For $0 \le \alpha \le 1$ and $n_0 \le q \le r \le n$, define the following two auxiliary problems:

Problem $\mathrm{P}_\alpha(\mathrm{q,r})$ *Minimize*

$$\Omega^\alpha_{qr}(\eta_{q+1}, \ldots, \eta_r) \;=\; \max_{q+1 \le i \le r} \frac{\lambda_i}{\alpha + (1-\alpha)\eta_i}$$

over all $\eta_{q+1} \ge \cdots \ge \eta_r \ge 0$ *satisfying* $\sum_{i=q+1}^r \eta_i = \sum_{i=q+1}^r \delta_i^{-2}$.

Problem P(q) *Minimize*

$$\Omega_q(\alpha; \eta_{q+1}, \ldots, \eta_n) \;=\; \max_{q+1 \le i \le n+1} \frac{\lambda_i}{\alpha + (1-\alpha)\eta_i}$$

over all $0 \le \alpha \le 1$ *and* $\eta_{q+1} \ge \cdots \ge \eta_{n+1} = 0$ *satisfying* $\sum_{i=q+1}^n \eta_i = \sum_{i=q+1}^n \delta_i^{-2}$.

Consider first the problem $\mathrm{P}_\alpha(\mathrm{q,r})$. If $\alpha = 1$ then $\Omega_\alpha \equiv \lambda_1$. Let $0 \le \alpha < 1$. Then the solution $\eta^* = (\eta^*_{q+1}, \ldots, \eta^*_r)$ of $\mathrm{P}_\alpha(\mathrm{q,r})$ can be obtained as follows. Let $\gamma = \gamma(\alpha) = \alpha/(1-\alpha)$. Let $k = k(\alpha; q, r)$ be the largest integer satisfying $q+1 \le k \le r$ and

$$\lambda_k \;\ge\; \frac{\gamma \sum_{j=q+1}^k \lambda_j}{\gamma\,(k-q) \;+\; \sum_{j=q+1}^r \delta_j^{-2}}\;. \tag{2.54}$$

Then

$$\eta_i^* \;=\; \frac{\gamma\,(k-q) \;+\; \sum_{j=q+1}^r \delta_j^{-2}}{\sum_{j=q+1}^k \lambda_j}\,\lambda_i \;-\; \gamma, \qquad q+1 \le i \le k, \tag{2.55}$$

and $\eta_i^* = 0$ for $k+1 \le i \le r$. Furthermore,

$$\Omega_{qr}^{\alpha}(\eta^*) \;=\; \frac{\sum_{j=q+1}^{k} \lambda_j}{\alpha\,(k-q) \,+\, (1-\alpha)\,\sum_{j=q+1}^{r} \delta_j^{-2}}\,. \tag{2.56}$$

We now pass to the solution of P(q). Let α_i, $i \ge q+1$, be defined in such a way that we have equality in (2.54) when k and γ are replaced by i and $\gamma_i = \alpha_i/(1-\alpha_i)$, respectively. Such α_i exists only for $i \ge s = \min\{\, j \mid \lambda_j < \lambda_{q+1}\,\}$. Then $\alpha_i = \gamma_i/(1+\gamma_i)$, where

$$\gamma_i \;=\; \frac{\lambda_i\,\sum_{j=q+1}^{n} \delta_j^{-2}}{\sum_{j=q+1}^{i-1}(\lambda_j - \lambda_i)}\,.$$

Setting $\alpha_i = 1$ for $i < s$, we have $1 = \alpha_{q+1} \ge \alpha_{q+2} \ge \cdots$ and the solution η^i of the problem $\mathrm{P}_\alpha(\mathrm{q},\mathrm{r})$ with $r = n$ satisfies $\eta_{q+1}^i \ge \cdots \ge \eta_i^i = 0$. Since in addition the right hand side of (2.56) is a monotone function of α, we obtain that

$$\min_{\alpha,\eta}\, \Omega_q(\alpha;\eta) \;=\; \min_{q+1\le i\le n+1}\, \Omega_q(\alpha_i,\eta^i).$$

Providing some further calculations we finally arrive at the following formulas for the solution (α^*,η^*) of P(q). Let

$$k \;=\; k(q) \;=\; \min\left\{ n,\, q + \left\lfloor \sum_{j=q+1}^{n} \delta_j^{-2} \right\rfloor \right\}. \tag{2.57}$$

If $\lambda_{q+1} = \lambda_{k+1}$ then $\alpha^* = 1$ and $\Omega_q(\alpha^*;\cdot) \equiv \lambda_{q+1}$. If $\lambda_{q+1} > \lambda_{k+1}$ then $\alpha^* = \gamma^*/(1+\gamma^*)$, where

$$\gamma^* \;=\; \frac{\lambda_{k+1}\,\sum_{j=q+1}^{n} \delta_j^{-2}}{\sum_{j=q+1}^{k}(\lambda_j - \lambda_{k+1})}$$

and η^* is given by (2.55) with $\gamma = \gamma^*$. Furthermore,

$$\Omega_q(\alpha^*;\eta^*) \;=\; \lambda_{k+1} \,+\, \frac{\sum_{j=q+1}^{k}(\lambda_j - \lambda_k)}{\sum_{j=q+1}^{n} \delta_j^{-2}}\,.$$

We shall say that the solution $\eta^* = (\eta_{q+1}^*,\ldots,\eta_r^*)$ of $\mathrm{P}_\alpha(\mathrm{q},\mathrm{r})$ is *acceptable* iff

$$\sum_{j=s}^{r} \eta_j^* \;\le\; \sum_{j=s}^{r} \frac{1}{\delta_j^2}, \qquad \text{for all } q+1 \le s \le r. \tag{2.58}$$

Similarly, the solution (α^*, η^*) of P(q) is acceptable iff (2.58) holds with $r = n$.

Let the number p, $0 \le p < n$, and the sequence $n = n_{p+1} > n_p > \cdots > n_0 \ge 0$ be defined (uniquely) by the conditions:

$$\begin{aligned} n_p &= \min\{\, s \ge n_0 \mid \text{ solution of } (\mathrm{P(s)}) \text{ is acceptable} \,\}, && (2.59) \\ n_i &= \min\{\, s \ge n_0 \mid \text{ solution of } (\mathrm{P}_{\alpha^*}(\mathrm{s}, \mathrm{n_{i+1}})) \text{ is acceptable} \,\}, && (2.60) \end{aligned}$$

$0 \le i \le p-1$, where α^* comes from the solution of $(\mathrm{P(n_p)})$.

Theorem 2.17 *Let p, the sequence $n_0 < n_1 < \cdots < n_{p+1} = n$ and α^* be defined by (2.59) and (2.60). Then the solution of the problem* (MP) *is given by α^* and*

$$\eta^* \;=\; (\,\eta^{(0)}, \eta^{(1)}, \ldots, \eta^{(p)}\,),$$

where $\eta^{(p)}$ and $\eta^{(i)}$ are solutions of $(\mathrm{P(n_p)})$ *and* $(\mathrm{P}_{\alpha^*}(\mathrm{n_i}, \mathrm{n_{i+1}}))$, *$0 \le i \le p-1$, respectively.*

Proof Let $k = k(n_p)$. Due to the definition of n_i we have $\eta_1^* \ge \cdots \ge \eta_k^* \ge \eta_{k+1}^* = 0$ and the maximal value of $\lambda_j/(\,\alpha^* + (1-\alpha^*)\eta_j^*)$, $n_0 + 1 \le j \le n$, is attained for $j = n_p + 1$. The definition of n_p yields in turn that $\eta^{(p)}$ are the last $(n - n_p)$ components of the solution of (MP) and that α^* is optimal. This completes the proof. □

As a consequence of this theorem we obtain the following corollary.

Corollary 2.7 *Let n_p and $k = k(n_p)$ be defined by (2.59) and (2.57), respectively. Then*

$$\mathrm{r}_n^{\mathrm{wor}}(\Delta) \;=\; \sqrt{\lambda_{k+1} + \frac{\sum_{j=n_p+1}^{k} (\lambda_j - \lambda_{k+1})}{\sum_{j=n_p+1}^{n} \delta_j^{-2}}}\,. \quad \square$$

Observe that we always have $\sqrt{\lambda_{n+1}} \le \mathrm{r}_n^{\mathrm{wor}}(\Delta) \;\le\; \sqrt{\lambda_1}$. The lower bound is achieved if for instance δ_i's are zero, i.e., if we deal with exact information. The upper bound is achieved, $\mathrm{r}_n^{\mathrm{wor}}(\Delta) = \sqrt{\lambda_1}$, if for instance $\sum_{i=1}^n \delta_i^{-2} \;\le\; 1$, see also E 2.52. In this case, information is useless.

Let us now consider the case when all δ_i's are constant, $\delta_i = \delta$, and $0 < \delta \le 1$. That is, noisy information satisfies

$$\sqrt{\sum_{i=1}^{n} (\, y_i - L_i(f)\,)^2} \;\le\; \delta.$$

Then the solution of $(P(0))$ is acceptable and $k = k(0) = n$. Hence, the formula for the radius reduces to

$$\mathrm{r}_n^{\mathrm{wor}}(\Delta) \;=\; \mathrm{r}_n^{\mathrm{wor}}(\delta) \;=\; \sqrt{\lambda_{n+1} \,+\, \frac{\delta^2}{n}\sum_{j=1}^{n}(\lambda_j - \lambda_{n+1})}\,. \tag{2.61}$$

If $\lambda_1 = \cdots = \lambda_{n+1}$ then $\mathrm{r}_n^{\mathrm{wor}}(\Delta) = \sqrt{\lambda_1}$ and the zero approximation is optimal. For $\lambda_1 > \lambda_{n+1}$ we have $\gamma^* = \delta^{-2}\gamma^{**}$ where

$$\gamma^{**} \;=\; \frac{n\lambda_{n+1}}{\sum_{j=1}^{n}(\lambda_j - \lambda_{n+1})},$$

and the optimal η_i^* are $\eta_i^* = \delta^{-2}\eta_i^{**}$ with

$$\eta_i^{**} \;=\; \frac{n(\lambda_i - \lambda_{n+1})}{\sum_{j=1}^{n}(\lambda_j - \lambda_{n+1})}, \qquad 1 \le i \le n.$$

The optimal information $N_n = [\langle\cdot,\xi_1^*\rangle_F,\ldots,\langle\cdot,\xi_n^*\rangle_F]$ is given by Theorem 2.16 with the matrix W constructed for $\eta_i = \eta_i^{**}$ and $\beta_i = 1$, $1 \le i \le n$. The optimal algorithm is $\varphi_n(y) = \sum_{j=1}^{n} z_j S(f_j)$ where $(\gamma_n I + G_{N_n})\,z = y$ and the parameter $\gamma_n = \gamma^{**}$. We stress that neither optimal information nor optimal algorithm depend on the noise level δ.

We now comment on the minimal radius $\mathrm{r}_n^{\mathrm{wor}}(\delta)$. If we fix n and tend with the noise level δ to zero then $\mathrm{r}_n^{\mathrm{wor}}(\delta)$ approaches the minimal radius of exact information, $\mathrm{r}_n^{\mathrm{wor}}(0) = \sqrt{\lambda_{n+1}} > 0$. For $\mathrm{r}_n^{\mathrm{wor}}(0) > 0$ we have

$$\mathrm{r}_n^{\mathrm{wor}}(\delta) \;-\; \mathrm{r}_n^{\mathrm{wor}}(0) \;\approx\; \frac{\delta^2}{2\,n\sqrt{\lambda_{n+1}}}\sum_{j=1}^{n}(\lambda_j - \lambda_{n+1}),$$[4]

while for $\mathrm{r}_n^{\mathrm{wor}}(0) = 0$ we have

$$\mathrm{r}_n^{\mathrm{wor}}(\delta) \;-\; \mathrm{r}_n^{\mathrm{wor}}(0) \;=\; \mathrm{r}_n^{\mathrm{wor}}(\delta) \;=\; \delta\sqrt{\frac{1}{n}\sum_{j=1}^{n}\lambda_j}\,.$$

[4] The symbol "≈" denotes here the *strong* equivalence of functions. We write $\psi_1(\delta) \approx \psi_2(\delta)$ iff $\lim_{\delta\to 0^+}\psi_1(\delta)/\psi_2(\delta) = 1$, as $\delta \to 0^+$.

Hence, for $\mathrm{r}_n^{\mathrm{wor}}(0) > 0$ the convergence is quadratic, and for $\mathrm{r}_n^{\mathrm{wor}}(0) = 0$ it is linear in δ.

Consider now the case where the noise level δ is fixed and $n \to +\infty$. The formula (2.61) can be rewritten as

$$\mathrm{r}_n^{\mathrm{wor}}(\delta) \;=\; \sqrt{\lambda_{n+1}(1-\delta^2) \,+\, \frac{\delta^2}{n}\sum_{j=1}^n \lambda_j}.$$

The compactness of S^*S implies $\lim_j \lambda_j \;=\; 0$. Hence, λ_{n+1} as well as $n^{-1}\sum_{j=1}^n \lambda_j$ converge to zero with n, and consequently,

$$\lim_{n\to+\infty} \mathrm{r}_n^{\mathrm{wor}}(\delta) \;=\; 0.$$

This result should not be a surprise since for noise bounded in the Euclidean norm we can obtain the value of any functional L at f with arbitrarily small error. Indeed, repeating k times observations of $L(f)$ we obtain information $y_1,\ldots,y_k$ such that $\sum_{i=1}^k (y_i - L(f))^2 \le \delta^2$. Hence, for large k most of the y_i's are very close to $L(f)$, and for $\|f\|_F \le 1$ the least squares approximation, $k^{-1}\sum_{i=1}^k y_i$, converges uniformly to $L(f)$.

Observe also that $\mathrm{r}_n^{\mathrm{wor}} \ge \delta\lambda_1/\sqrt{n}$. Thus, for $S \not\equiv 0$ the radius cannot tend to zero faster than $\delta/\sqrt{n}$.

To see more precisely how $r_n(\delta)$ depends on the eigenvalues λ_j, suppose that

$$\lambda_j \;\asymp\; \left(\frac{\ln^s j}{j}\right)^p, \qquad \text{as} \quad j \to +\infty,$$

where $p > 0$ and $s \ge 0$. Such a behavior of the eigenvalues is typical of some multivariate problems defined in a tensor product spaces; see NR 2.30. In this case, for $\delta > 0$ we have

$$\mathrm{r}_n^{\mathrm{wor}}(\delta) \;\asymp\; \begin{cases} \delta\,\left(\frac{\ln^s n}{n}\right)^{p/2} & 0<p<1, \\ \delta\,\left(\frac{\ln^{s+1} n}{n}\right)^{1/2} & p=1, \\ \delta\,\frac{1}{\sqrt{n}} & p>1, \end{cases} \tag{2.62}$$

where the constants in the "$\asymp$" [5] notation do not depend on δ. Since

$$\mathrm{r}_n^{\mathrm{wor}}(0) \;=\; \left(\frac{\ln^s(n+1)}{n+1}\right)^{p/2},$$

[5] For two sequences, we write $a_n \asymp b_n$ iff there exist constants $0 < c_1 \le c_2 < +\infty$ such that for all n, $c_1\,a_n \le b_n \le c_2 a_n$. Such sequences are said to be *weakly* equivalent.

we conclude that for $p < 1$ the radius of noisy information essentially behaves as the radius of exact information, while for $p > 1$ the radius of noisy information essentially behaves as $\delta/\sqrt{n}$. Hence, in the presence of noise, the best speed of convergence is $\delta/\sqrt{n}$.

2.8.2 Approximation and integration of Lipschitz functions

In this section we deal with noise bounded in the sup–norm. We assume that F is the space of Lipschitz functions $f : [0,1] \to \mathbb{R}$, and consider the following two solution operators on F:

- *Function approximation.*

It is defined by the solution operator $\mathrm{App} : F \to C([0,1])$,

$$\mathrm{App}(f) \;=\; f, \qquad f \in F,$$

where $C([0,1])$ is the space of continuous functions $f : [0,1] \to \mathbb{R}$ with the norm

$$\|f\| \;=\; \|f\|_\infty \;=\; \max_{0 \le t \le 1} |f(t)|.$$

- *Integration*

The solution operator is given by $\mathrm{Int} : F \to \mathbb{R}$,

$$\mathrm{Int}(f) \;=\; \int_0^1 f(t)\,dt.$$

The set $E \subset F$ is assumed to be the set of functions for which the Lipschitz constant is 1,

$$E \;=\; \{\, f : [0,1] \to \mathbb{R} \mid \quad |f(t_1) - f(t_2)| \le |t_1 - t_2|,\; 0 \le t_1, t_2 \le 1 \,\}.$$

Observe that E is the unit ball of F with respect to the seminorm

$$\|f\|_F \;=\; \sup_{0 \le t_1 < t_2 \le 1} \frac{|f(t_1) - f(t_2)|}{|t_1 - t_2|}.$$

Information $y \in \mathbb{R}^n$ about $f \in F$ is obtained by noisy observations of the function values at some points $t_i \in [0,1]$ where

$$|y_i - f(t_i)| \;\le\; \delta_i, \qquad 1 \le i \le n.$$

That is, exact information is now of the form

$$N(f) \;=\; [\,f(t_1), f(t_2), \ldots, f(t_n)\,], \qquad f \in F, \tag{2.63}$$

and the noise $x = y - N(f)$ belongs to the set $B(\Delta, z) = B(\Delta) = \{\, x \in \mathbb{R}^n \mid |x_i| \le \delta_i,\ 1 \le i \le n \,\}$. Hence, the noise is bounded uniformly in the "weighted" sup–norm.

For a given precision vector $\Delta = [\delta_1, \ldots, \delta_n] \in \mathbb{R}^n$, we want to choose t_i's in such a way as to minimize the radius $\mathrm{r}_n^{\mathrm{wor}}(S; N, \Delta)$, for $S \in \{\mathrm{App}, \mathrm{Int}\}$. We assume without loss of generality that

$$0 \le \delta_1 \le \delta_2 \le \cdots \le \delta_n.$$

We have the following theorem.

Theorem 2.18 *Let k be the largest integer such that $1 \le k \le n$ and*

$$\delta_k \;\le\; \frac{1}{k}\left(\frac{1}{2} + \sum_{j=1}^{k} \delta_j\right).$$

Then the minimal radius

$$\mathrm{r}_n^{\mathrm{wor}}(\mathrm{App}; \Delta) \;=\; \frac{1}{k}\left(\frac{1}{2} + \sum_{j=1}^{k} \delta_j\right)$$

and

$$\mathrm{r}_n^{\mathrm{wor}}(\mathrm{Int}; \Delta) \;=\; \frac{1}{k}\left(\frac{1}{2} + \sum_{j=1}^{k} \delta_j\right)^2 \;-\; \sum_{j=1}^{k} \delta_j^2.$$

Furthermore, the optimal points t_i^* are for both problems given as

$$t_i^* \;=\; \frac{2i-1}{k}\left(\frac{1}{2} + \sum_{j=1}^{k} \delta_j\right) \;-\; 2\left(\sum_{j=1}^{i-1} \delta_j\right) \;-\; \delta_i, \qquad \textit{for } 1 \le i \le k,$$

and t_i^*–arbitrary for $k+1 \le i \le n$.

Proof Consider first the approximation problem, $S = \mathrm{App}$. Let an exact information operator N of the form (2.63) be given. Then we have

$$\begin{aligned} \mathrm{rad}^{\mathrm{wor}}(\mathrm{App}; N, \Delta) \;&=\; \frac{1}{2} \cdot \mathrm{diam}(\mathrm{App}; N, \Delta) \\ &=\; \sup\{\, \|f\|_\infty \mid \ \ |f(t_i)| \le \delta_i,\ 1 \le i \le n \,\}. \end{aligned} \tag{2.64}$$

Indeed, for $y \in \bigcup_{f\in E} \mathbb{N}(f)$, define the functions

$$\begin{aligned} f_y^+(t) &= \sup\{\, f(t) \mid \quad f\in E,\, |y_i - f(t_i)| \le \delta_i,\, 1\le i\le n\,\} \\ &= \min_{1\le i\le n} (\, y_i + \delta_i + |t-t_i|\,), \end{aligned}$$

$$\begin{aligned} f_y^-(t) &= \inf\{\, f(t) \mid \quad f\in E,\, |y_i - f(t_i)| \le \delta_i,\, 1\le i\le n\,\} \\ &= \max_{1\le i\le n} (\, y_i - \delta_i - |t-t_i|\,). \end{aligned}$$

Then $f_y^+, f_y^- \in E$ and for any $t\in[0,1]$ and $f_1, f_2 \in E$ such that $y \in \mathbb{N}(f_1) \cap \mathbb{N}(f_2)$, we have $|f_1(t) - f_2(t)| \le f_y^+(t) - f_y^-(t)$. Hence, $f_y = (f_y^+ + f_y^-)/2 \in E$ is the center of the set $A_{\mathbb{N}}(y)$ of functions from E that share information y, and $r(A_{\mathbb{N}}(y)) = 1/2 \cdot d(A_{\mathbb{N}}(y))$. Consequently, $\mathrm{rad}^{\mathrm{wor}}(\mathrm{App}; N, \Delta) = 0.5\,\mathrm{diam}(\mathrm{App}; N, \Delta)$. The second equality in (2.64) follows from the definition of the diameter.

The formula for f_y^+ with $y=0$ yields

$$\mathrm{rad}^{\mathrm{wor}}(\mathrm{App}; N, \Delta) \;=\; \|f_0^+\|_\infty \;=\; \max_{0\le t\le 1}\,\min_{1\le i\le n} (\,\delta_i + |t-t_i|\,).$$

Thus $\mathrm{rad}^{\mathrm{wor}}(\mathrm{App}; N, \Delta)$ is a continuous function of $t_1, \ldots, t_n$ defined on a compact set $[0,1]^n$. Therefore the minimal radius $\mathrm{r}_n^{\mathrm{wor}}(\mathrm{App}; \Delta)$ is attained. Using some geometrical arguments we get that $\mathrm{r}_n^{\mathrm{wor}}(\mathrm{App}; \Delta)$ is attained for t_i's satisfying the following system of equations:

$$\left\{ \begin{array}{rcll} A &=& t_1 + \delta_1, & \\ A &=& \frac{t_i - t_{i-1}}{2} + \frac{\delta_{i-1}+\delta_i}{2}, & 2\le i\le m, \\ A &=& \delta_m + (1-t_m), & \end{array} \right.$$

where $A = \mathrm{rad}^{\mathrm{wor}}(\mathrm{App}; N, \Delta)$ and m is the largest integer such that $1 \le m \le n$ and $\delta_m \le A$. Solving this system we obtain the desired result.

We now turn to the integration problem, $S = \mathrm{Int}$. Since Int is a functional, we have

$$\mathrm{rad}^{\mathrm{wor}}(\mathrm{Int}; N, \Delta) \;=\; \sup\left\{ \int_0^1 f(t)\,dt \;\middle|\quad f\in E,\, |f(t_i)| \le \delta_i,\, 1\le i\le n \right\}.$$

Using again some geometrical arguments we obtain that the optimal t_i's are the same as for function approximation. Hence, the formulas for integration can be obtained by integrating the function f_0^+ constructed for $t_i = t_i^*$. It is given as

$$f_0^+(t) \;=\; \delta_i + |t - t_i^*|, \qquad |t - t_i^*| \le A - \delta_i,\ 1 \le i \le k.$$

This completes the proof. □

Assume now that all observations are performed with the same precision, $\Delta = [\underbrace{\delta, \ldots, \delta}_{n}]$. Then the formulas of Theorem 2.18 take the following form:

$$\mathrm{r}_n^{\mathrm{wor}}(\mathrm{App}; \Delta) \;=\; \mathrm{r}_n^{\mathrm{wor}}(\mathrm{App}; \delta) \;=\; \frac{1}{2n} + \delta \tag{2.65}$$

and

$$\mathrm{r}_n^{\mathrm{wor}}(\mathrm{Int}; \Delta) \;=\; \mathrm{r}_n^{\mathrm{wor}}(\mathrm{Int}; \delta) \;=\; \frac{1}{4n} + \delta. \tag{2.66}$$

The optimal points are $t_i^* = (2i-1)/(2n)$, $1 \le i \le n$. The reader can also check that for $S \in \{\mathrm{App}, \mathrm{Int}\}$, the optimal algorithm is in this case given by $\varphi(y) = S(\mathbf{p}(y))$, where $\mathbf{p}(y)$ in the natural spline of degree 1 such that $\mathbf{p}(t_i^*) = y_i$, $1 \le i \le n$.

For both problems we have $\mathrm{r}_n^{\mathrm{wor}}(\delta) = \mathrm{r}_n^{\mathrm{wor}}(0) + \delta$. Thus the error of any algorithm is always greater than δ, no matter how many observations have been performed. Actually, this is not a coincidence, but a common property of problems with noise bounded in the sup–norm. Namely, consider a general problem with linear S, the set E being the unit ball in a seminorm $\|\cdot\|_F$, and the noise satisfying $|x_i| \le \delta$, $\forall i$.

Lemma 2.15 *Suppose there exists an element $h^* \in F$ such that $h^* \notin \ker S$ and*

$$|L(h^*)| \;\le\; 1 \qquad \textit{for all } L \in \Lambda.$$

Then for any $n \ge 1$ we have

$$\mathrm{r}_n^{\mathrm{wor}}(\delta) \;\ge\; \min\{\,\delta, 1/\|h^*\|_F\} \cdot \|S(h^*)\|$$

($1/0 = +\infty$).

Proof For

$$h_\delta \;=\; \begin{cases} \delta\, h^* & \delta \le 1/\|h^*\|_F, \\ h^*/\|h^*\|_F & \delta > 1/\|h^*\|_F, \end{cases}$$

we have $\|h_\delta\|_F \le 1$ and $|L(h_\delta)| \le \delta$, $\forall L \in \Lambda$. Hence, for any exact information $N = [L_1, \ldots, L_n]$ with $L_i \in \Lambda$, $1 \le i \le n$,

$$\begin{aligned} \mathrm{rad}^{\mathrm{wor}}(N, [\underbrace{\delta, \ldots, \delta}_{n}]) &\ge \frac{1}{2}\,\mathrm{diam}(N, [\underbrace{\delta, \ldots, \delta}_{n}]) \\ &= \sup\{\, \|S(h)\| \mid \quad h \in E,\ |L_i(h)\| \le \delta,\ 1 \le i \le n \,\} \\ &\ge \|S(h_\delta)\| \ = \ \min\{\delta, 1/\|h^*\|_F\}\, \|S(h^*)\|, \end{aligned}$$

as claimed. □

For the problems App and Int, the lemma holds with $h^* \equiv 1$. Then $\|h^*\|_F = 0$ and $\mathrm{r}_n^{\mathrm{wor}}(\delta,) \ge \delta$. The formulas (2.65) and (2.66) show that this is the "worst" possible choice of h^*. (Actually, we have $\mathrm{r}_n^{\mathrm{wor}}(\delta) > \delta$, see also E 2.59.)

Lemma 2.15 also applies for the problem considered in Section 2.8.1, i.e., when F and G are Hilbert spaces, E is the unit ball in E, the solution operator $S : F \to G$ is compact, and Λ is the class of continuous linear functionals with norm bounded by 1. Taking h^* as the eigenvector corresponding to the largest eigenvalue of the operator S^*S, we obtain

$$\mathrm{r}_n^{\mathrm{wor}}(\delta) \ \ge \ \min\{1, \delta\} \cdot \|S\|_F.$$

(Here also the inequality "$\le$" can ce replaced by "$<$".) This observation should be contrasted to the case of noise bounded by δ in the Euclidean norm where the radius always converges to zero as $n \to +\infty$, see (2.61).

Notes and Remarks

NR 2.27 Some parts of Section 2.8.1 (e.g. Lemma 2.14) have been taken from Plaskota [81] where the corresponding problem in the average case setting was solved, see also Section 3.8.1. The other results of Section 2.8 are new.

NR 2.28 In the case of exact linear information, the minimal radius $\mathrm{r}_n^{\mathrm{wor}} = \mathrm{r}_n^{\mathrm{wor}}(0)$ is closely related to the Gelfand *n–widths* and *s–numbers* of the classical approximation theory. If one allows only linear algorithms, there are relations to the Kolmogorow n–widths. These relations are discussed in Mathe [56], Novak[63], Traub and Woźniakowski [109, Sect.6 of Chap.2 and Sect.5 of Chap.3], Traub *et al.* [108, Sect.5 of Chap.4], and Kowalski *et al.* [43]. The survey of the theory of n–widths and many other references are presented in Pinkus [75].

We note that for noisy information such closed relations do not hold any longer. Indeed, as we convinced ourselves, the minimal radius $\mathrm{r}_n^{\mathrm{wor}}(\delta)$ may for $\delta > 0$ tend do zero arbitrarily slower than $\mathrm{r}_n^{\mathrm{wor}}(0)$ (or it may not converge at all), and consequently the ratio of the n–width and $\mathrm{r}_n^{\mathrm{wor}}(\delta)$ may be arbitrary.

NR 2.29 For exact data, optimal nonadaptive information that uses n observations is usually called the nth optimal information and its radius the nth minimal radius, see e.g. Traub *et al.*[108, Sect.5.3 of Chap.4]. If observations are always performed with the same precision δ, the notion of nth minimal radius and nth optimal information can be in a natural way carried over to the noisy case.

NR 2.30 We now present an example of a problem for which the results of Section 2.8.1 can be applied.

Let $d \geq 1$ and $r_i \geq 1$, $1 \leq i \leq d$. Let $F = W^{0\ldots 0}_{r_1 \ldots r_d}$ be the r.k.h.s. defined in NR 2.21. That is, F is the r.k.h.s. of multivariate functions $f : [0,1]^d \to \mathbb{R}$ with r.k. $R = \bigotimes_{i=1}^n R_{r_i}$ where R_{r_i} is the r.k. of the r_i–fold Wiener measure. Define the solution operator as $S : F \to G = \mathcal{L}_2([0,1]^d)$, $S(f) = f$. That is, we want to approximate functions in the $\mathcal{L}_2$–norm.

It is known (see e.g. Papageorgiou and Wasilkowski [70]) that in this case the eigenvalues of the operator S^*S satisfy

$$\lambda_j \asymp \left(\frac{\ln^{k-1} j}{j} \right)^{2r}$$

where $r = \min\{r_1, \ldots, r_d\}$ and k is the number of indices i for which $r_i = r$. Observe that the exponent $2r \geq 2$. Hence, due to (2.62) we have $\mathrm{r}_n^{\mathrm{wor}}(\delta) \asymp \delta/\sqrt{n}$ for $\delta > 0$, and $\mathrm{r}_n^{\mathrm{wor}}(0) \asymp (\ln^{k-1} n/n)^r$.

NR 2.31 It would be interesting to know the radius and optimal information in the Hilbert case for restricted class Λ of permissible functionals. For instance, for the problem of NR 2.30 it is natural to assume that only noisy function values can be observed. Much is known in the exact information case, see e.g., Lee and Wasilkowski [48], Woźniakowski [127] [128], Wasilkowski and Woźniakowski [122]. Unfortunately, finding optimal noisy information turns out to be a very difficult problem. Some results can be obtained from the average case analysis of Chapter 3, see NR 3.22.

Optimal information in the Hilbert case and for noise bounded in the sup–norm is also unknown, even when $\delta_i = \delta$, $\forall i$.

Exercises

E 2.49 Let F and G be normed spaces and let the solution operator $S : F \to G$ be continuous and linear. Let $\|\cdot\|_Y$ be a norm in $\mathbb{R}^n$. Consider the problem of approximating $S(f)$ for $f \in E$ (E–arbitrary), based on information $y \in \mathbb{R}^n$ such that $\|y - N(f)\|_Y \leq \delta$, where $N = [L_1, \ldots, L_n]$ consists of continuous linear functionals from a class Λ. Let $\mathrm{r}_n^{\mathrm{wor}}(\delta)$ be the minimal radius of information consisting of n functionals. Show that if Λ satisfies

$$L \in \Lambda \implies c\,L \in \Lambda, \quad \forall c \in \mathbb{R},$$

then $\mathrm{r}_n^{\mathrm{wor}}(\delta) = \mathrm{r}_n^{\mathrm{wor}}(0)$.

E 2.50 Show that at least $(n-2)(n-1)/2$ elements of the matrix W from Lemma 2.14 are zero.

E 2.51 How will the formula for $\mathrm{r}_n^{\mathrm{wor}}(\Delta)$ in the Hilbert case change if E is the ball of radius b and the class Λ consists of functionals whose norm is bounded by M?

E 2.52 Consider the minimal radius in the Hilbert case. Let $n_0 = \max\{\, 1 \le i \le n \mid \delta_i = 0 \,\}$. Let

$$\begin{aligned} s &= \min\{\, 1 \le i \le n+1 \mid \quad \lambda_i = \lambda_{n+1} \,\} \\ t &= \max\{\, 1 \le i \le n+1 \mid \quad \lambda_i = \lambda_1 \,\}. \end{aligned}$$

Show that $\mathrm{r}_n^{\mathrm{wor}}(\Delta) = \sqrt{\lambda_{n+1}}$ iff $s \le n_0+1$, and $\mathrm{r}_n^{\mathrm{wor}}(\Delta) = \sqrt{\lambda_1}$ iff $\sum_{j=n_0+1}^{n} \delta_j^{-2} \le n-t$.

E 2.53 Let δ be such that

$$\frac{1}{\delta^2} \le \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\delta_i^2}.$$

Show that then in the Hilbert case $\mathrm{r}_n^{\mathrm{wor}}(\underbrace{\delta, \ldots, \delta}_{n}) \le \mathrm{r}_n^{\mathrm{wor}}(\delta_1, \ldots, \delta_n)$.

E 2.54 Let η_i', $n_0+1 \le i \le n$, minimize

$$\Omega'(\eta_{n_0+1}, \ldots, \eta_n) = \max_{n_0+1 \le i \le n+1} \frac{\lambda_i}{1+\eta_i}$$

over all $\eta_{n_0+1} \ge \cdots \ge \eta_n \ge 0$ satisfying (2.49). Let N_Δ' be the information operator constructed as in Theorem 2.16 with η_i^* replaced by η_i'. Show that in the Hilbert case $\mathrm{rad}^{\mathrm{wor}}(N_\Delta', \Delta) \le \sqrt{2} \cdot \mathrm{r}_n^{\mathrm{wor}}(\Delta)$.

E 2.55 Show that if in the Hilbert case the solution operator S is not compact then the radius $\mathrm{r}_n^{\mathrm{wor}}(\delta)$ does not converge to zero with n. Moreover, the optimal information does not necessarily exist.

E 2.56 Discuss the existence of optimal linear algorithms for the problems App and Int, for an arbitrary precision vector Δ.

E 2.57 When are the optimal points t_i^* in Theorem 2.18 determined uniquely?

E 2.58 Let $\Delta = [\delta_1, \ldots, \delta_n]$ and $\delta = \left(\sum_{i=1}^{n} \delta_i\right)/n$. Show that for $S \in \{\mathrm{App}, \mathrm{Int}\}$ we have $\mathrm{r}_n^{\mathrm{wor}}(S;\delta) \ge \mathrm{r}_n^{\mathrm{wor}}(S;\Delta)$.

E 2.59 Suppose that the element h^* in Lemma 2.15 satisfies

$$\max\{\, \|S(h^*)+g\|, \|S(h^*)-g\|\,\} \;>\; \|S(h^*)\|, \quad \forall\, g \neq 0.$$

Show that if, in addition, $\mathrm{r}_n^{\mathrm{wor}}(0) > 0$, $\forall n \geq 1$, then

$$\mathrm{r}_n^{\mathrm{wor}}(\delta) \;>\; \delta\,\|S(h^*)\|, \qquad \forall \delta < 1/\|h^*\|_F.$$

Apply this result to the concrete problems considered in this section.

E 2.60 Let F be the class of functions $f : [0,1] \to \mathbb{R}$ for which the rth derivatives exist and are Lipschitz functions. Let

$$E \;=\; \{\, f \in E \mid \quad |f^{(r)}(t_1) - f^{(r)}(t_2)| \leq |t_1 - t_2|\,\}.$$

Consider the solution operator $S : F \to C([0,1])$ given as $S(f) = f^{(k)}$ where $0 \leq k \leq r$. Show that if information consists of noisy function values with noise bounded in the sup–norm, then

$$\mathrm{r}_n^{\mathrm{wor}}(\delta) \;\geq\; 2\,\delta\,\frac{r!}{(r-k)!}, \qquad 1 \leq k \leq r,$$

and $\mathrm{r}_n^{\mathrm{wor}}(\delta) \geq \delta$ for $k = 0$.

2.9 Complexity

Up to now we have analyzed only the error of algorithms. It is clear that in practical computations we are interested not only in the error but also in the cost of obtaining approximations. In this section, we explain what we mean by the cost of approximation and discuss the concept of complexity. Then we derive some general bounds on complexity of a problem. We assume that we are given:

- The solution operator $S : F \to G$ where F is a linear space and G is a normed space.
- The set $E \subset F$ of elements f for which we want to construct approximations of $S(f)$.
- The class Λ of permissible information functionals.
- The sets $B(\Delta, z)$ of all possible values of noise corresponding to the precision vector Δ and exact information z.

2.9.1 Computations over the space G

In order to be able to analyze the cost of obtaining approximations, we first present our *model of computation.* Roughly speaking, this model is based on the following two postulates. Namely, we assume that we can gain information about f by noisy observations of functionals at f. Then, using some primitive permissible operations, the information can be combined to get an approximation. These primitive operations are: arithmetic operations and comparisons over the reals, linear operations over the space G, and logical operations over the Boolean values.

To describe the computational process leading to obtaining an approximation, we shall use the concept of a *program.* To define the program precisely, we adopt notation from the programming language *Pascal.*

Any program consists of two main parts:

- description of objects that are to be used, and
- description of actions that are to be performed.

The objects used in programs are called *constants* and *variables.* A constant is a fixed real number, an element of the space G, or a Boolean value ('true' or 'false'). A variable can assume an arbitrary value from a given nonempty set T. This set determines the *type* of the variable. We have three basic types: *real* (i.e., $T=\mathbb{R}$), *G–type* ($T=G$), and *Boolean* ($T=\{\text{'true'},\text{'false'}\}$). We also allow T to be a subset of one from the basic types. The description of constants and variables is called a *declaration.*

The actions are described by *statements.* We have two *simple* statements: information statement, assignment statement, and three *structured* statements: compouned statement, conditional statement, repetitive statement. We now define all the statements in turn.

- The information statement

$$\mathcal{I}\,(\,\mathbf{y};\,L,\delta\,)$$

where $\mathbf{y}$ is a real variable, $L\in\Lambda$, and $\delta\geq 0$. This statement describes the action which allows to gain data – the noisy value of $L(f)$ where $f\in F$ is the (unknown) element for which we want to compute an approximation. We "ask" for this noisy value. The "answer" is a real number y which is then assigned to the variable $\mathbf{y}$. We assume that by the ith question we obtain a value y_i such that

$$[y_1,\dots,y_i] \in B(\,[\delta_1,\dots,\delta_i],\ [L_1(f),\dots,L_i(f)]\,).$$

- The assignment statement

$$\mathbf{v}\ :=\ \mathcal{E}$$

where $\mathbf{v}$ is a variable and $\mathcal{E}$ is an *expression.* This is the second fundamental statement. It specifies that the value of the expression $\mathcal{E}$ be evaluated for the current values of variables, and that this value be assigned to the variable $\mathbf{v}$.

Expressions are constructs denoting rules of computation for evaluating values of functions of some variables using only permissible *primitive operations* over the reals, elements of the space G, and Boolean values, which are represented by the constants and current values of variables. The primitive operations are as follows:

arithmetic operations over the reals $\mathbb{R}$: sign inversion ($x \mapsto -x$), addition ($(x,y) \mapsto x+y$), subtraction ($(x,y) \mapsto x-y$), multiplication ($(x,y) \mapsto x*y$), division ($(x,y) \mapsto x/y$, $y \neq 0$),

comparisons over $\mathbb{R}$: equality ($(x,y) \mapsto x=y$), inequality ($(x,y) \mapsto x \neq y$), ordering ($(x,y) \mapsto x<y$, $(x,y) \mapsto x \le y$),

linear operations over the space G: sign inversion ($g \mapsto -g$), addition ($(g_1,g_2) \mapsto g_1+g_2$), subtraction $(g_1,g_2) \mapsto g_1-g_2$), multiplication by a scalar ($(x,g) \mapsto x*g$),

Boolean operations: negation ($b \mapsto$ not b), union ($(b_1,b_2) \mapsto b_1$ or b_2), conjunction ($(b_1,b_2) \mapsto b_1$ and b_2).

To be more precise, an expression is a single constant or variable, or it is the construct of the form $f(w)$ or $f(w,z)$, where f stands for a primitive operation (of one or two arguments), and w,z are already defined expressions. In the following three examples, $a,b \in \mathbb{R}$ and $h \in G$ are constants, $\mathbf{x},\mathbf{y}$ are real variables, and $\mathbf{g}$ is a variable of the type G.

$$\begin{aligned} (\mathbf{x}-a)(b-\mathbf{x})/(b-a) &\qquad \text{(real expression)},\\ \mathbf{g}+\mathbf{y}h &\qquad \text{(G–type expression)},\\ (\mathbf{y}<a) \text{ or } (\mathbf{y}\le 3) &\qquad \text{(Boolean expression)}. \end{aligned}$$

Those were the simple statements. The structured statements contain other statements in their definitions.

- The compound statement

$$\textbf{begin} \quad \mathcal{S}_1;\ \mathcal{S}_2;\ \dots\ ;\mathcal{S}_n \quad \textbf{end}$$

where $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_n$ are statements. It specifies the successive execution of $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_n$.

- The conditional statement

$$\textbf{if} \quad \mathcal{E} \quad \textbf{then} \quad \mathcal{S}_1 \quad \textbf{else} \quad \mathcal{S}_2$$

where $\mathcal{E}$ is a Boolean expression and $\mathcal{S}_1, \mathcal{S}_2$ are statements or '*empty*' (i.e., do nothing). This corresponds to the following action. First the value of $\mathcal{E}$ is evaluated. If this value is 'true', the action $\mathcal{S}_1$ is executed. If 'false', we perform $\mathcal{S}_2$.

- The repetitive statement

$$\textbf{while} \quad \mathcal{E} \quad \textbf{do} \quad \mathcal{S}$$

where $\mathcal{E}$ is a Boolean expression and $\mathcal{S}$ is a statement. The action described by this statement relies on repetitive execution of $\mathcal{S}$ until the value of $\mathcal{E}$ is 'false'. If $\mathcal{E}$ is 'false' at the beginning, $\mathcal{S}$ is not executed at all.

Summarizing, the program consists of one declaration and one compound statement. We make an additional assumption that the set of variables contains a special variable **g** of the G–type. An approximation is obtained by executing the program. The result of computation – approximation to $S(f)$ – is the value of the variable **g**. The initial values of all variables are undefined. In order that the result be always well defined, we assume that for any data y a finite number of simple statements is executed including at least one assignment to **g**.

2.9.2 Cost and complexity, general bounds

We now present our notion of cost and complexity. The basic assumption is that we must charge for gaining information and performing any primitive operation.

1. Obtaining noisy value of $L(f)$ with precision δ costs $\mathrm{c}(\delta)$ where $\mathrm{c} : [0, +\infty) \to [0, +\infty)$ is a given *cost function.* It is nonnegative, nonincreasing, and positive for sufficiently small but positive δ.

Observe that the cost of obtaining a single datum d depends on the precision δ. The smaller δ, the larger the cost. Theoretically, c can assume even infinite values. However, $\mathrm{c}(\delta) = +\infty$ will mean that the precision δ cannot be used. Examples of cost functions include $\mathrm{c}(\delta) = \delta^{-2}$ or $\mathrm{c}(\delta) = \max\{0, \log_2 1/\delta\}$. The function

$$\mathrm{c}(\delta) \;=\; \begin{cases} +\infty & 0 \le \delta < \delta_0, \\ c_0 & \delta \ge \delta_0, \end{cases} \tag{2.67}$$

corresponds to the case when any observation is performed with fixed precision δ_0. In particular, taking $\delta_0 = 0$ we obtain the exact information case.

2. We assign the following costs to the primitive operations:

arithmetic operations over $\mathbb{R}$	–	1,	
comparisons over $\mathbb{R}$	–	1,	
linear operations over G	–	g	$(\mathrm{g} \ge 1)$,
Boolean operations	–	0.	

Let $\mathcal{P}$ be a program. The total cost of executing $\mathcal{P}$ is given by the sum of costs of gaining information (information cost) and performing all primitive operations (combinatory cost). Observe that the total cost depends only on the obtained information (data) y. We denote this cost by $\mathrm{cost}(\mathcal{P}; y)$. The *(worst case) cost* of computing an approximation using the program $\mathcal{P}$ is given as

$$\mathrm{cost}^{\mathrm{wor}}(\mathcal{P}) \;=\; \sup \left\{ \mathrm{cost}(\mathcal{P}; y) \;\middle|\; y \in \bigcup_{f \in E} \mathbb{N}(f) \right\}.$$

We now define the complexity of an algorithm. Observe first that for any program $\mathcal{P}$ there exist a unique (in general adaptive) information operator $\mathbb{N} = \{N, \Delta\}$ and algorithm φ with the following property. For any $f \in E$, the possible information obtained by executing $\mathcal{P}$ is in the set $\mathbb{N}(f)$, and for information $y \in \mathbb{N}(f)$ the program gives the approximation $\varphi(y)$. We say that $\mathcal{P}$ is a *realization* of the algorithm φ using information $\mathbb{N}$. It is clear that not all algorithms φ using some information $\mathbb{N}$ can be realized.

On the other hand, if an algorithm has at least one realization then it has also many other realizations. We are interested in such realizations $\mathcal{P}$ that have minimal $\mathrm{cost}^{\mathrm{wor}}(\mathcal{P})$. This minimal cost will be called an *algorithm complexity* and denoted by $\mathrm{comp}^{\mathrm{wor}}(\mathbb{N},\varphi)$. That is,

$$\mathrm{comp}^{\mathrm{wor}}(\mathbb{N},\varphi) \;=\; \inf\{\,\mathrm{cost}^{\mathrm{wor}}(\mathcal{P}) \mid \quad \mathcal{P} \text{ is a realization of } \varphi \text{ using } \mathbb{N}\,\}.$$

(If φ and $\mathbb{N}$ cannot be realized then $\mathrm{comp}^{\mathrm{wor}}(\mathbb{N},\varphi) = +\infty$.) Observe that $\mathrm{comp}^{\mathrm{wor}}(\mathbb{N},\varphi)$ is independent of any realization, i.e., this is the property of the operators $\mathbb{N}$ and φ only.

We are now ready to define the problem complexity. Let $\varepsilon \ge 0$. Suppose that we want to compute approximations to $S(f)$ for $f \in E$ with the (worst case) error not exceeding ε. An *ε–complexity* of this problem, $\mathrm{Comp}^{\mathrm{wor}}(\varepsilon)$, is defined as the minimal $\mathrm{cost}^{\mathrm{wor}}(\mathcal{P})$ over all programs $\mathcal{P}$ which allow to compute approximations with error at most ε. Clearly, such approximations can be computed only when the corresponding information and algorithm satisfy $\mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi) \le \varepsilon$. Hence,

$$\mathrm{Comp}^{\mathrm{wor}}(\varepsilon) \;=\; \inf\{\,\mathrm{comp}^{\mathrm{wor}}(\mathbb{N},\varphi) \mid \quad \mathbb{N},\varphi \text{ such that } \mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi) \le \varepsilon\,\}.$$

Information $\mathbb{N}_\varepsilon$ and an algorithm φ_ε with $\mathrm{comp}^{\mathrm{wor}}(\mathbb{N}_\varepsilon,\varphi_\varepsilon) = \mathrm{Comp}^{\mathrm{wor}}(\varepsilon)$ and $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}_\varepsilon,\varphi_\varepsilon) \le \varepsilon$, will be called *$\varepsilon$–complexity optimal*, or simply *optimal* if it is known from the context which optimality concept is considered.

Clearly, the ε–complexity depends not only on ε but also on the other parameters of the problem. Therefore we shall sometimes write $\mathrm{Comp}^{\mathrm{wor}}(S;\varepsilon)$, $\mathrm{Comp}^{\mathrm{wor}}(S,\mathrm{c};\varepsilon)$, etc. To make the notation shorter, the superscript "wor" in $\mathrm{comp}^{\mathrm{wor}}$ and $\mathrm{Comp}^{\mathrm{wor}}$ will be usually omitted.

We now give some useful general bounds on the ε–complexity that will be used in the next section. To this end, we first define several auxiliary concepts.

For an information operator $\mathbb{N} = \{N_y, \Delta_y\}_{y\in Y}$ where

$$\Delta_y = [\delta_1, \delta_2(y_1), \ldots, \delta_n(y_1, \ldots, y_{n(y)-1})],$$

the complexity of $\mathbb{N}$ is given as

$$\mathrm{comp}(\mathbb{N}) \;=\; \sup_{y\in Y} \sum_{i=1}^{n(y)} \mathrm{c}(\delta_i(y_1,\ldots,y_{i-1})\,).$$

($n(y)$ denotes the length of y.) Note that for nonadaptive $\mathbb{N}$ we obviously have $\mathrm{comp}(\mathbb{N}) = \sum_{i=1}^n \mathrm{c}(\delta_i)$. An *$\varepsilon$–information complexity* is defined as

$$\begin{aligned}\mathrm{IComp}(\varepsilon) \;=\; \inf\{\,\mathrm{comp}(\mathbb{N}) \;\; | \;\; & \mathbb{N}\text{–adaptive, and there exists } \varphi \\ & \text{such that } \mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi) \le \varepsilon\,\}.\end{aligned}$$

Hence, $\mathrm{IComp}(\varepsilon)$ is the minimal complexity of information from which it is possible (at least theoretically) to obtain approximation with error at most ε. We also define a corresponding quantity for nonadaptive information as

$$\begin{aligned}\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \;=\; \inf\{\,\mathrm{comp}(\mathbb{N}) \;\; | \;\; & \mathbb{N}\text{–nonadaptive, and there exists } \varphi \\ & \text{such that } \mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi) \le \varepsilon\,\}.\end{aligned}$$

We introduced in Section 2.7 the concept of a κ–hard element. We shall say that $f^* \in F$ is a *κ–strongly hard* element iff for any nonadaptive information $\mathbb{N} = \{N, \Delta\}$ there exists an algorithm φ such that

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N},\varphi) \;\le\; \kappa \cdot r(A_{\mathbb{N}}(N(f^*)\,)).$$

Note that if f^* is a κ_1–hard element then it is a κ–strongly hard element for any $\kappa > \kappa_1$. If for any nonadaptive information there exists an optimal algorithm, then f^* is also the κ_1–strongly hard element.

We have the following bounds on the ε-complexity.

Theorem 2.19 *(i) If the κ–strongly hard element exists then*

$$\mathrm{Comp}(\varepsilon) \;\ge\; \mathrm{IComp}^{\mathrm{non}}(\kappa\,\varepsilon).$$

(ii) Let $\rho \ge 1$. Suppose that there exists a nonadaptive information $\mathbb{N}_\varepsilon$ using $n(\varepsilon)$ observations, and a linear algorithm φ_ε such that

$$\mathrm{comp}(\mathbb{N}_\varepsilon) \;\le\; \rho\cdot\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \quad \textit{and} \quad \mathrm{e}^{\mathrm{wor}}(\mathbb{N}_\varepsilon,\varphi_\varepsilon) \le \varepsilon.$$

Then

$$\mathrm{Comp}(\varepsilon) \;\le\; \rho\cdot\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \;+\; (2\,n(\varepsilon)-1)\,\mathrm{g}.$$

Proof (i) Let $\mathbb{N}^{\mathrm{ad}}$ be an arbitrary, in general adaptive, information with radius $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{ad}}) \le \varepsilon$. Let $\mathbb{N}^{\mathrm{non}}$ be the nonadaptive information from

Theorem 2.15 corresponding to $\mathbb{N}^{\mathrm{ad}}$. Then $\mathrm{comp}(\mathbb{N}^{\mathrm{non}}) \le \mathrm{comp}(\mathbb{N}^{\mathrm{ad}})$ and there is an algorithm φ such that

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{non}}, \varphi) \;\le\; \kappa \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^{\mathrm{ad}}) \;\le\; \kappa\,\varepsilon.$$

Hence, $\mathrm{Comp}(\varepsilon) \ge \mathrm{IComp}^{\mathrm{non}}(\kappa\,\varepsilon)$.

(ii) The algorithm φ_ε is of the form $\varphi_\varepsilon(y) = \sum_{i=1}^{n(\varepsilon)} y_i\, g_i$. Hence, it requires at most $2\,n(\varepsilon)-1$ linear operations in the space G to compute $\varphi_\varepsilon(y)$. Therefore,

$$\begin{aligned} \mathrm{comp}(\mathbb{N}_\varepsilon, \varphi_\varepsilon) &\le \mathrm{comp}(\mathbb{N}_\varepsilon) + (2\,n(\varepsilon) - 1)\,\mathrm{g} \\ &\le \rho \cdot \mathrm{IComp}^{\mathrm{non}}(\varepsilon) + (2\,n(\varepsilon) - 1)\,\mathrm{g}, \end{aligned}$$

as claimed. □

As a consequence of this theorem we obtain the following corollary.

Corollary 2.8 *If the assumptions of Theorem 2.19 are fulfilled and, additionally,*

$$\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \;=\; O(\,\mathrm{IComp}^{\mathrm{non}}(\kappa\,\varepsilon)\,) \quad \textit{and} \quad n(\varepsilon) \;=\; O(\,\mathrm{IComp}^{\mathrm{non}}(\varepsilon)\,),$$

then

$$\mathrm{Comp}(\varepsilon) \;\asymp\; \mathrm{IComp}^{\mathrm{non}}(\varepsilon), \qquad \textit{as} \quad \varepsilon \to 0^+.$$ [6] □

Hence, for problems satisfying the assumptions of Corollary 2.8, the ε–complexity, $\mathrm{Comp}(\varepsilon)$, is essentially equal to $\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$. Note that the condition $\mathrm{IComp}^{\mathrm{non}}(\varepsilon) = O(\mathrm{IComp}^{\mathrm{non}}(\kappa\,\varepsilon))$ means that $\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$ does not increase too fast as $\varepsilon \to 0$. It is satisfied if, for instance, $\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$ behaves polynomially in $1/\varepsilon$. The condition $n(\varepsilon) = O(\mathrm{IComp}^{\mathrm{non}}(\varepsilon))$ holds if the information operators $\mathbb{N}_\varepsilon$ use observations with costs bounded uniformly from below by a positive constant. Obviously, this is the case for $\mathrm{c}(\delta) \ge c_0 > 0$, $\forall\,\delta \ge 0$, since then $n(\varepsilon) \le \mathrm{IComp}^{\mathrm{non}}(\varepsilon)/c_0$.

It turns out that $\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$ is closely related to the minimal radius of information. To see this, let

$$\mathrm{R}(T) \;=\; \inf\left\{ \mathrm{r}_n^{\mathrm{wor}}(\delta_1, \ldots, \delta_n) \;\middle|\; n \ge 1,\ \sum_{i=1}^{n} \mathrm{c}(\delta_i) \le T \right\}$$

[6] For two functions, $a(\varepsilon) \asymp b(\varepsilon)$ as $\varepsilon \to 0^+$ means the *weak* equivalence of functions. That is, there exist $\varepsilon_0 > 0$ and $0 < K_1 \le K_2 < +\infty$ such that $K_1 \le a(\varepsilon)/b(\varepsilon) \le K_2$ for all $\varepsilon \le \varepsilon_0$.

be the Tth *minimal radius.* Let

$$\overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon) \;=\; \inf\{\, T \mid \quad \mathrm{R}(T) \le \varepsilon\,\}.$$

Lemma 2.16 *We have*

$$\lim_{\alpha\to 0^+} \overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon-\alpha) \;\ge\; \mathrm{IComp}^{\mathrm{non}}(\varepsilon) \;\ge\; \overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon).$$

Proof Let $0 < \alpha < \varepsilon$ and $\beta > 0$. Then $\mathrm{R}(\overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon-\alpha)+\beta) \le \varepsilon-\alpha$ and there are information $\mathbb{N}_\beta$ and an algorithm φ_β such that $\mathrm{comp}(\mathbb{N}_\beta) \le \overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon-\alpha)+\beta$ and $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}_\beta,\varphi_\beta) \le \varepsilon$. Then $\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \le \mathrm{comp}(\mathbb{N}_\beta)$. Letting $\beta \to 0^+$ we get $\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \le \overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon-\alpha)$. The first inequality in the lemma now follows from the fact that the function $\overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon)$ is nonincreasing and therefore the limit exists.

To show the second inequality, for $\beta > 0$ we take information $\mathbb{N}_\beta$ and algorithm φ_β such that $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}_\beta,\varphi_\beta) \le \varepsilon$ and $\mathrm{comp}(\mathbb{N}_\beta) \le \mathrm{IComp}^{\mathrm{non}}(\varepsilon)+\beta$. Hence, $\overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon) \le \mathrm{IComp}^{\mathrm{non}}(\varepsilon)+\beta$. Since this holds for arbitrary β, we obtain $\overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon) \le \mathrm{IComp}^{\mathrm{non}}(\varepsilon)$. □

Thus, $\mathrm{IComp}^{\mathrm{non}}$ is, roughly speaking, the inverse function to the Tth minimal radius $\mathrm{R}(T)$. Note that the minimal radii $\mathrm{r}_n^{\mathrm{wor}}(\delta_1,\ldots,\delta_n)$ have already been known for some problems, see Section 2.8. For those problems, $\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$ can be evaluated using Lemma 2.16.

For fixed precision (2.67), the complexity of information $\mathbb{N}_\varepsilon$ from Theorem 2.19 equals $\mathrm{comp}(\mathbb{N}_\varepsilon) = c_0\, n(\varepsilon)$. Hence, in this case the bounds in Theorem 2.19 can be rewritten as

$$c_0\, n(\kappa\,\varepsilon) \;\le\; \mathrm{Comp}(\varepsilon) \;\le\; n(\varepsilon)(c_0+2\,\mathrm{g}) \;-\; 2\,\mathrm{g}.$$

Observe also that the ε–information complexity of a problem with a given cost function c can be bounded from above by $\mathrm{IComp}(\varepsilon)$ of the same problem, but with fixed precision. Indeed, it suffices to set δ_0 in (2.67) to be such that $\mathrm{c}(\delta_0) > 0$, and $c_0 = \mathrm{c}(\delta_0)$. On the other hand, if only the cost function c is bounded from below by a positive constant c_0, the ε–information complexity is not smaller than $\mathrm{IComp}(\varepsilon)$ of the same problem with exact information and with the cost function $\mathrm{c} \equiv c_0$.

Notes and Remarks

NR 2.32 In the case of exact information or information with fixed noise level, our model of computation corresponds to that of Traub *et al.* [107, Chap.5] and [108, Chap.3]. As far as we know, the model with cost dependent on the noise level was first studied by Kacewicz and Plaskota [32].

NR 2.33 Some researchers define a model of computation using the concept of a *machine*. The most known is the *Universal Turing Machine* which can be used to study discrete problems. Another example is the *Unlimited Register Machine* (URM) discussed in Cutland [10]. Machines and complexity over the reals (without oracle) are presented in Blum, Shub, and Smale [6]. Recently, Novak [66] introduced the real number URM and showed how it can be used to study complexity of problems with partial information. For related models, see also Ko [38] and Schönhage [93].

NR 2.34 We use a rather simple version of the model of computation. It is based on the following conviction: if the solution element is in a space G then operations which define this space should be permitted. In our case G is a linear space over the reals. As the real ring is an ordered set in which addition and multiplication are defined, and a linear space in a set in which addition and multiplication by scalars are defined, we assume that arithmetic operations and comparisons of real numbers as well as linear operations in G are permitted.

Sometimes useful generalizations of the model are possible. We now give one example. Suppose that G is a Cartesian product of some other linear spaces over $\mathbb{R}$, $G = G_1 \times G_2 \times \cdots \times G_s$. For instance, $G = \mathbb{R}^s = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{s}$. Or, if G is a space of functions $g : \mathbb{R}^d \to \mathbb{R}^s$ then any element $g \in G$ can be represented as $g(x) = (g_1(x), \ldots, g_s(x))$ with $g_i : \mathbb{R}^d \to \mathbb{R}$. In these cases it is natural to assume that we can perform linear operations over each "coordinate" G_i. Clearly, we would also be able to perform linear operations over G (via the identification $g = (g_1, \ldots, g_s) \in G$, $g_i \in G_i$) with the cost $\mathrm{g} = \sum_{i=1}^{s} \mathrm{g}_i$, where g_i is the cost of linear operations over G_i.

However, for our purpose any such a "generalization" is not necessary. As it will turn out, for problems considered in this book the complexity essentially equals the information complexity. Hence, using a "more powerful" model leads to similar results.

NR 2.35 The assumption that we can use arbitrary elements of $\mathbb{R}$ or G corresponds in practice to the fact that *precomputation* is possible. This may be sometimes too idealized assumption. For instance, even if we know theoretically that the optimal algorithm is linear, $\varphi_{\mathrm{opt}}(y) = \sum_{i=1}^{n} y_i g_i$, the elements g_i can be sometimes not known exactly, or can be very difficult to precompute. We believe however that such examples are exceptional.

We also stress that the "precomputed" elements *can* depend on ε. One may assume that precomputing is independent of ε. This leads to another, also interesting model, in which one wants to have a "good" single program which allows to produce an ε–approximation to $S(f)$ for any $\varepsilon > 0$. Some examples on this can be found in Kowalski [42], Novak [66], and Paskov [74].

NR 2.36 Clearly, our model assumes also other idealizations. One of them is that the cost of observing noisy value of a functional depends on the noise level only. That is, we neglect the dependence on the element for which information is obtained. Errors that may occur when the value of $\varphi(y)$ is computed, are also neglected.

NR 2.37 One can argue that the assumption that linear operations over G are allowed is not much realistic when $\dim G = +\infty$. In practice usually digital computers are used to perform calculations, and they can only manipulate with bits. This is certainly true. On the other hand, the computers have been successfully used for solving some very complicated problems including, in particular, continuous problems which require at least computations over the reals. This paradox is possible only because the computer arithmetic (which is in fact discrete) can very well imitate the computations in the real number model. Similarly, by using an appropriate discrete model, we can make computations over an arbitrary linear space G possible.

This point can also be expressed in the following way. Even if it is true that the real world is discrete in nature, it is often more convenient (and simpler!) to use a continuous model to describe, study, and to understand some phenomena. We believe that the same applies to scientific computations.

NR 2.38 We consider a *sequential* model of computation, where only one instruction can be performed at each step. It would also be interesting to study a *parallel* model, see, e.g., Heinrich and Kern [23], Kacewicz [28], Nemirowski [61].

NR 2.39 We note that Theorem 2.19 is not always true, even if the problem is linear and linear information is used. That is, there are problems, for which the ε–complexity is much larger (or even infinite) than ε–information complexity, see e.g. Wasilkowski and Woźniakowski [120].

NR 2.40 Information about the programming language Pascal can be found, e.g., in Jensen and Wirth [26].

Exercises

E 2.61 Show that if the conditional and repetitive statements were not allowed then only algorithms using nonadaptive information would be realizable and the cost of computing $\varphi(y)$ would be independent of y.

E 2.62 Give an example of an algorithm φ and information $\mathbb{N}$ that cannot be realized.

E 2.63 Let $\mathbb{N}$ be an information operator with $Y = \mathbb{R}^n$, and let φ be an algorithm of the form

$$\varphi(y) \;=\; \sum_{i=1}^{n} q_i(y)\, g_i,$$

where $g_i \in G$ and q_i are some real rational functions of n variables $y_1, \ldots, y_n$. Show that then there exists a realization of φ using $\mathbb{N}$.

E 2.64 Let φ_1 and φ_2 be two algorithms that use the same information $\mathbb{N}$. Show that if $\varphi_2 = A\varphi_1$ where $A : G \to G$ is a linear transformation then $\mathrm{comp}(\mathbb{N}, \varphi_2) \le \mathrm{comp}(\mathbb{N}, \varphi_1)$. If, in addition, A is one-to-one then $\mathrm{comp}(\mathbb{N}, \varphi_1) = \mathrm{comp}(\mathbb{N}, \varphi_2)$.

E 2.65 Give an example of a problem for which optimal information is nonadaptive and the upper bound in Theorem 2.19 is not sharp, i.e., $\mathrm{Comp}(\varepsilon) < \mathrm{IComp}^{non}(\varepsilon) + (2n(\varepsilon) - 1)\mathrm{g}$.

E 2.66 Show that Lemma 2.16 will hold if we replace $\overline{\mathrm{IComp}}^{\mathrm{non}}$ by

$$\overline{\overline{\mathrm{IComp}}}^{\mathrm{non}}(\varepsilon) \;=\; \inf\left\{ \sum_{i=1}^{n} \mathrm{c}(\delta_i) \;\Big|\; n \ge 1,\ \mathrm{r}_n^{\mathrm{wor}}(\delta_1, \ldots, \delta_n) \le \varepsilon \right\}.$$

2.10 Complexity of special problems

In this section, we derive the ε–complexity for several classes of problems. To this end, we use the general bounds given in the previous section. A special attention will be devoted to the dependence of $\mathrm{Comp}(\varepsilon)$ on the cost function.

2.10.1 Linear problems in Hilbert spaces

We begin with the problem defined in Section 2.8.1. That is, we assume that S is a compact operator acting between separable Hilbert spaces F and G, and E is the unit ball in F. The class Λ of permissible information functionals consists of all continuous linear functionals with norm bounded by 1. The noise x satisfies $\sum_{i=1}^{n} x_i^2/\delta_i^2 \le 1$ where n is the length of x and $[\delta_1, \ldots, \delta_n]$ is the precision vector used.

In this case, it is convenient to introduce the function

$$\tilde{\mathrm{c}}(x) \;=\; \mathrm{c}(x^{-2}), \qquad 0 < x < +\infty.$$

We assume that $\tilde{c}$ is concave or convex.

We first show how in general the Tth minimal radius can be evaluated. As in Section 2.8.1, we denote by ξ_j, $j \geq 1$, the orthonormal basis of eigenvectors of the operator S^*S, and by λ_j the corresponding eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots$. We shall also use the function Ω which was defined in Section 2.8.1,

$$\Omega \;=\; \Omega(\alpha;\, \eta_1, \ldots, \eta_n) \;=\; \max_{1 \leq i \leq n+1} \frac{\lambda_i}{\alpha + (1-\alpha)\, \eta_i}$$

(with the convention $\lambda_i/0 = +\infty$ for $\lambda_i > 0$ and $0/0 = 0$).

Lemma 2.17 *The Tth minimal radius is equal to*

$$\mathrm{R}(T) \;=\; \inf \sqrt{\Omega(\alpha; \eta_1, \ldots, \eta_n)}\,,$$

where the infimum is taken over all $0 \leq \alpha \leq 1$, n, and $\eta_i \geq 0$, $1 \leq i \leq n$, satisfying
(a1) for $\tilde{c}$–concave

$$\sum_{i=1}^{n} \tilde{c}(\eta_i) \leq T,$$

(b1) for $\tilde{c}$–convex

$$n\, \tilde{c} \left(\frac{1}{n} \sum_{i=1}^{n} \eta_i \right) \leq T.$$

Moreover, if the infimum is attained for some n^* and $\eta^* = (\eta_1^*, \ldots, \eta_{n^*}^*)$, then

$$\mathrm{R}(T) \;=\; \mathrm{rad}^{\mathrm{wor}}(\{N_T, \Delta_T\})$$

where
(a2) for $\tilde{c}$–concave

$$\Delta_T \;=\; \left[1/\sqrt{\eta_1^*}, \ldots, 1/\sqrt{\eta_{n^*}^*} \right], \qquad N_T \;=\; [\langle \cdot, \xi_1 \rangle_F, \ldots, \langle \cdot, \xi_{n^*} \rangle_F],$$

(b2) for $\tilde{c}$–convex

$$\Delta_T \;=\; [\, \underbrace{1/\sqrt{\eta_0^*}, \ldots, 1/\sqrt{\eta_0^*}}_{n^*} \,], \qquad N_T \;=\; [\, \langle \cdot, \xi_1^* \rangle_F, \ldots, \langle \cdot, \xi_{n^*}^* \rangle_F],$$

where $\eta_0^* \;=\; 1/n^* \sum_{i=1}^{n^*} \eta_i^*$ and ξ_i^*'s are as in Theorem 2.16 with $\delta_i = \sqrt{1/\eta_0^*}$, $\forall\, i$.

Proof We first prove (a1) and (b1). Let the function $\tilde{\mathrm{c}}$ be concave. Then for any n and $\eta_1, \ldots, \eta_n$ satisfying (2.49) we have

$$\sum_{i=1}^{n} \tilde{c}(\eta_i) \ \leq \ \sum_{i=1}^{n} \tilde{c}(1/\delta_i^2). \tag{2.68}$$

Denoting by $\eta^*(\delta)$ the vector minimizing Ω over all $0 \leq \alpha \leq 1$ and η satisfying (2.49), we obtain from Theorem (2.16) and (2.68) that

$$\begin{aligned} \mathrm{R}^2(T) &= \inf \left\{ (\mathrm{r}_n^{\mathrm{wor}}(\delta_1, \ldots, \delta_n))^2 \ \Big| \ n \geq 1, \ \sum_{i=1}^{n} c(\delta_i^2) \leq T \right\} \\ &= \inf \, \{ \, \Omega(\alpha; \eta^*(\delta) \,) \mid \ 0 \leq \alpha \leq 1, \ n \geq 1, \\ &\qquad\qquad \left. \delta = (\delta_1, \ldots, \delta_n), \ \sum_{i=1}^{n} \tilde{\mathrm{c}}(1/\delta_i^2) \leq T \right\} \\ &= \inf \, \{ \, \Omega(\alpha; \eta) \mid \ 0 \leq \alpha \leq 1, \ n \geq 1, \\ &\qquad\qquad \left. \eta = (\eta_1, \ldots, \eta_n), \ \sum_{i=1}^{n} \tilde{\mathrm{c}}(\eta_i) \leq T \right\}. \end{aligned}$$

Let $\tilde{\mathrm{c}}$ be convex. Then for any n and $\eta_1, \ldots, \eta_n$ we have

$$\sum_{i=1}^{n} \tilde{\mathrm{c}}(\eta_i) \ \geq \ n \, \tilde{c}(\eta_0)$$

where $\eta_0 = 1/n \sum_{i=1}^{n} \eta_i$. Since for $\delta_i^2 = 1/\eta_0, \ 1 \leq i \leq n,$ the condition (2.49) holds for any $\eta_1 \geq \cdots \geq \eta_n$, we obtain

$$\begin{aligned} \mathrm{R}^2(T) &= \inf \left\{ (\mathrm{r}_n^{\mathrm{wor}}(\underbrace{\delta, \ldots, \delta}_{n}))^2 \ \Big| \ n \geq 1, \ n \, c(\delta) \leq T \right\} \\ &= \inf \, \{ \, \Omega(\alpha; \eta) \mid \ 0 \leq \alpha \leq 1, \ n \geq 1 \\ &\qquad\qquad \left. \eta = (\eta_1, \ldots, \eta_n), \quad n \, \tilde{\mathrm{c}} \left(\frac{1}{n} \sum_{i=1}^{n} \eta_i \right) \leq T \right\}. \end{aligned}$$

To show (a2) and (b2) it is enough to apply (a1), (a2), and Theorem 2.16. □

We now comment on the above lemma. Observe first that the Tth minimal radius depends only on the cost function and eigenvalues of the operator S^*S. The second remark is about information for which $\mathrm{R}(T)$ is achieved.

In the case of convex $\tilde{\mathrm{c}}$, optimal information uses observations with fixed precision δ. This is no longer true for concave functions $\tilde{\mathrm{c}}$. However, in this case we can restrict ourselves to observations of the functionals $\langle\cdot,\xi_i\rangle_F$.

To give an illustration of how Lemma 2.17 can be used to evaluate $\mathrm{Comp}(\varepsilon)$, suppose that the cost function is given by

$$\mathrm{c}(\delta) \;=\; \mathrm{c_{lin}}(\delta) \;=\; \begin{cases} \delta^{-2} & \delta > 0, \\ +\infty & \delta = 0. \end{cases}$$

This cost function possesses an interesting property. Namely, the error of approximating the value of a functional from several observations depends only on the total cost of observations and not on the number of them and precisions used. Indeed, if we observe n times the value $L(f)$ with accuracy $\Delta = [\delta_1, \ldots, \delta_n]$, then the minimal error of approximating $L(f)$ is $\left(\sum_{i=1}^n \delta_i^{-2}\right)^{-1/2} = \sum_{i=1}^n \mathrm{c_{lin}}(\delta_i)$.

Note that in this case the function $\tilde{\mathrm{c}}_{\mathrm{lin}}(x) = x$. Hence, it is convex and concave. After some calculations we obtain

$$\mathrm{R}(c_{\mathrm{lin}};T)^2 \;=\; \lambda_{n+1} \,+\, \frac{1}{T}\sum_{j=1}^{n}(\lambda_j - \lambda_{n+1}) \tag{2.69}$$

where $n = n(T) = \lfloor T \rfloor$. Observe now that for $0 \le T \le 1$ we have $\mathrm{R}(\mathrm{c_{lin}};T)^2 = \mathrm{R}(0)^2 = \lambda_1$, while for $T \ge c_0$, $\mathrm{R}(\mathrm{c_{lin}};T)^2$ is linear on each interval $[n,\, n+1]$ and $\mathrm{R}(c_{\mathrm{lin}};n)^2 = 1/n\sum_{j=1}^n \lambda_j$, $j \ge 1$. Hence, the Tth minimal radius is a continuous function of T and $\lim_{T\to\infty}\mathrm{R}(\mathrm{c_{lin}};T) = 0$. Moreover, since $\lambda_1 \ge \lambda_2 \ge \cdots \to 0$, for sufficiently large T, $T > \min\{\, j \mid \lambda_j < \lambda_1\}$, it is also decreasing. We obtain from Lemma 2.16 that for small ε

$$\mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon) \;=\; \mathrm{R}^{-1}(\varepsilon) \;\approx\; \min\left\{ n \ge 1 \;\middle|\; \frac{1}{n}\sum_{j=1}^{n}\lambda_j \le \varepsilon^2 \right\}.[7]$$

To get the ε–complexity of our problem, we can use Theorem 2.19. We have that 0 is the 1–strongly hard element and therefore $\mathrm{IComp}(\mathrm{c_{lin}};\varepsilon) = \mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon)$. Furthermore, $\mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon)$ can be achieved by information that uses $n(\varepsilon) = \lfloor \mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon)\rfloor$ observations, and there exists

[7] $a(\varepsilon) \approx b(\varepsilon)$ means the *strong* equivalence of functions, i.e., $\lim_{\varepsilon\to 0^+} a(\varepsilon)/b(\varepsilon) = 1$ $(0/0 = \infty/\infty = 1)$.

an optimal linear algorithm. Hence,

$$\mathrm{Comp}(\mathrm{c_{lin}};\varepsilon) \;\asymp\; \min\left\{ n\geq 1 \;\Big|\; \frac{1}{n}\sum_{j=1}^{n}\lambda_j \leq \varepsilon^2 \right\}.$$

It turns out that the cost function c_{lin} is in some sense “worst” possible. Namely, we have the followinf fact.

Lemma 2.18 *Let* c *be an arbitrary cost function. Let δ_0 be such that* $\mathrm{c}(\delta_0)<+\infty$. *Then*

$$\mathrm{Comp}(\mathrm{c};\varepsilon) \;\leq\; M\cdot\mathrm{Comp}(\mathrm{c_{lin}},\varepsilon), \qquad \forall\,\varepsilon>0,$$

where $M=M(\mathrm{c},\delta_0)=\lceil 2\delta_0^2\rceil(\,\mathrm{c}(\delta_0)+2\mathrm{g})$.

Proof Since $\mathrm{R}(\mathrm{c_{lin}};T)^2\geq\lambda_1/\max\{1,T\}$, we have

$$\mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon)\;\left\{\begin{array}{ll} =0 & \varepsilon\geq\sqrt{\lambda_1}, \\ >1 & \varepsilon<\sqrt{\lambda_1}. \end{array}\right.$$

In the first case zero is the best approximation and the lemma is true.

Let $\varepsilon<\sqrt{\lambda_1}$. Let $\mathbb{N}$ be such an information operator that $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})=\varepsilon$ and $\mathrm{comp}(\mathrm{c_{lin}};\mathbb{N})=\mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon)$. We can assume that $\mathbb{N}$ uses $n=\lfloor\mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon)\rfloor$ observations with the same precision δ satisfying $\delta^{-2}=\mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon)/n$. Let $k=k(\delta_0)=\lceil 2\delta_0^2\rceil$. Consider the information operator $\tilde{\mathbb{N}}$ which repeats k times observations of the same functionals as in $\mathbb{N}$, but with precisions $\tilde\delta$ where $\tilde\delta^{-2}=\delta^{-2}/k$. We obviously have $\mathrm{rad}^{\mathrm{wor}}(\tilde{\mathbb{N}})=\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ and

$$\begin{aligned}\mathrm{comp}(\mathrm{c};\tilde{\mathbb{N}}) \;&=\; k\,n\,\tilde{\mathrm{c}}\left(\frac{\mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon)}{k\,n}\right)\\ &\leq\; k\,n\,\tilde{\mathrm{c}}(2/k) \;\leq\; k\,\mathrm{c}(\delta_0)\,\mathrm{IComp}^{non}(\mathrm{c_{lin}};\varepsilon).\end{aligned}$$

Since the optimal algorithm $\tilde\varphi$ for information $\tilde{\mathbb{N}}$ is linear, we finally obtain

$$\begin{aligned}\mathrm{Comp}(\mathrm{c};\varepsilon) \;&\leq\; \mathrm{comp}(\mathrm{c};\tilde{\mathbb{N}},\tilde\varphi) \;\leq\; k\,\mathrm{c}(\delta_0)\mathrm{Comp}(\mathrm{c_{lin}};\varepsilon)\;+\;(2\,k\,n-1)\mathrm{g}\\ &\leq\; k\,(\mathrm{c}(\delta_0)+2\,g)\,\mathrm{Comp}(\mathrm{c_{lin}};\varepsilon),\end{aligned}$$

as claimed. □

Lemma 2.18 can be used for deriving complexity for some other cost functions. For instance, consider the case of fixed positive precision where the cost function $\mathrm{c}_{\mathrm{fix}}(\delta) = c_0$ for $\delta \geq \delta_0$, and $\mathrm{c}_{\mathrm{fix}}(\delta) = +\infty$ for $\delta < \delta_0$, with $c_0, \delta_0 > 0$. Since $\mathrm{c}_{\mathrm{fix}}(\delta) \geq c_0\, \delta_0^2\, \mathrm{c}_{\mathrm{lin}}(\delta)$, we have $\mathrm{IComp}^{non}(\mathrm{c}_{\mathrm{fix}}; \varepsilon) \geq c_0\, \delta_0^2\, \mathrm{IComp}^{non}(\mathrm{c}_{\mathrm{lin}}; \varepsilon)$. Hence, $\mathrm{c}_{\mathrm{fix}}$ is also the "worst" cost function and $\mathrm{Comp}(\mathrm{c}_{\mathrm{fix}}; \varepsilon) \asymp \mathrm{Comp}(\mathrm{c}_{\mathrm{lin}}; \varepsilon)$.

We now consider the exact information case where the cost function is constant, e.g., $\mathrm{c}_{\mathrm{exa}} \equiv 1$. In this case $\mathrm{r}_n^{\mathrm{wor}}(0) = \sqrt{\lambda_{n+1}}$. Hence, $\mathrm{R}(\mathrm{c}_{\mathrm{exa}}; T) = \sqrt{\lambda_{n+1}}$ where $n = n(T) = \lfloor T \rfloor$, and

$$\mathrm{Comp}(\mathrm{c}_{\mathrm{exa}}; \varepsilon) \;\asymp\; \min \left\{ n \geq 1 \;\middle|\; \lambda_{n+1} \leq \varepsilon^2 \right\}.$$

Clearly, $\mathrm{Comp}(\mathrm{c}_{\mathrm{exa}}; \varepsilon)$ gives a lower bound for complexity corresponding to a cost function that is bounded from below by a positive constant. That is, if $\mathrm{c}(\delta) \geq c_0 > 0$ for all $\delta \geq 0$, then

$$\mathrm{Comp}(\mathrm{c}; \varepsilon) \;\geq\; c_0 \cdot \mathrm{Comp}(\mathrm{c}_{\mathrm{exa}}; \varepsilon).$$

Let us see more exactly how the complexity depends on the cost function c and eigenvalues λ_j. Consider

$$\mathrm{c}_q(\delta) \;=\; \begin{cases} (1+\delta^{-2})^q & \delta > 0, \\ +\infty & \delta = 0, \end{cases}$$

where $q \geq 0$. Note that for $q \geq 1$ the function $\tilde{\mathrm{c}}_q$ is convex, while for $0 < q \leq 1$ it is concave. The case $q = 0$ corresponds to the exact information.

Since for all q we have $\mathrm{Comp}(q; \varepsilon) = O(\mathrm{Comp}(1; \varepsilon))$, we can restrict ourselves to $0 \leq q \leq 1$. To obtain the formula for the Tth minimal radius we set, for simplicity, $\alpha = 1/2$ in the α–smoothing spline algorithm and use Lemma 2.17. The minimum $\min_{1 \leq i \leq n+1} \lambda_i/(1+\eta_i)$ over all $\eta_1 \geq \cdots \geq \eta_n \geq \eta_{n+1} = 0$ such that $\sum_{i=1}^n (1+\eta_i)^q \leq T$, is attained at

$$\eta_j \;=\; \frac{T^{1/q}}{\left(\sum_{i=1}^n \lambda_i^q\right)^{1/q}} \cdot \lambda_j \,-\, 1, \qquad 1 \leq j \leq n,$$

where n is the largest integer satisfying

$$\sum_{j=1}^n \lambda_j^q \;\leq\; \lambda_n^q\, T.$$

Furthermore,

$$\mathrm{R}(q;T)^2 \;=\; b \cdot \left(\frac{1}{T}\sum_{i=1}^{n} \lambda_i^q\right)^{1/q}$$

where $1/2 \le b \le 1$. Assume now that the eigenvalues of the operator S^*S satisfy

$$\lambda_j \;\asymp\; \left(\frac{\ln^s j}{j}\right)^p$$

with $p > 0$ and $s \ge 0$. As we know, this corresponds to function approximation in tensor product spaces, see NR 2.30. In this case we have

$$\mathrm{R}(q,p,s;T) \;\asymp\; \mathrm{R}(1,pq,s;T)^{1/q}.$$

The formulas for $\mathrm{R}(1,pq,s;T)$ can be derived based on (2.69). We obtain that for all q

$$\mathrm{R}(q,p,s;T) \;\asymp\; \begin{cases} \left(\frac{1}{T}\right)^{1/\tilde{q}} & p\,\tilde{q} > 1, \\ \left(\frac{\ln^{s+1} T}{T}\right)^{p} & p\,\tilde{q} = 1, \\ \left(\frac{\ln^{s} T}{T}\right)^{p} & 0 \le p\,\tilde{q} < 1, \end{cases}$$

as $T \to +\infty$, where $\tilde{q} = \min\{1, q\}$. This together with Lemma 2.16 and Corollary 2.8 gives the ε–complexity. Namely,

Theorem 2.20

$$\mathrm{Comp}^{\mathrm{wor}}(q,p,s;\varepsilon) \;\asymp\; \begin{cases} \left(\frac{1}{\varepsilon}\right)^{2\tilde{q}} & p\,\tilde{q} > 1, \\ \left(\frac{1}{\varepsilon}\right)^{2/p} \ln^{s+1}\left(\frac{1}{\varepsilon}\right) & p\,\tilde{q} = 1, \\ \left(\frac{1}{\varepsilon}\right)^{2/p} \ln^{sp}\left(\frac{1}{\varepsilon}\right) & 0 \le p\,\tilde{q} < 1, \end{cases}$$

as $\varepsilon \to 0$. □

We see that the complexity is determined by the value of $p\,\tilde{q}$. More precisely, suppose first that $p > 1$. Then for $pq > 1$ we have $\mathrm{Comp}(q,p,s;\varepsilon) \asymp \mathrm{Comp}(1,p,s;\varepsilon)$, while for $pq < 1$ we have $\mathrm{Comp}(q,p,s;\varepsilon) \asymp \mathrm{Comp}(0,p,s;\varepsilon)$. If $p < 1$ then $\mathrm{Comp}(q,p,s;\varepsilon) \asymp \mathrm{Comp}(0,p,s;\varepsilon) \asymp \mathrm{Comp}(1,p,s;\varepsilon)$, for all $q \ge 0$. This means, roughly speaking, that the ε–complexity may behave in at most two different ways – as for exact information, or as for the "worst" cost function $\mathrm{c_{lin}}(\delta) = 1/\delta^2$. Moreover, if the eigenvalues λ_j tend to zero sufficiently slowly ($p < 1$), then the ε–complexity behaves independently of the cost function.

2.10.2 Approximation and integration of Lipschitz functions

We pass to approximation, App, and integration, Int, of real valued Lipschitz functions $f : [0,1] \to \mathbb{R}$, based on noisy values of f at some points. The noise $x = y - N_y(f)$ is assumed to be bounded in the weighted sup–norm, $x \in B(\Delta_y)$, where

$$B(\Delta_y) \;=\; \{\, x \in \mathbb{R}^n \mid \quad |x_i| \le \delta_i(y_1, \ldots, y_{i-1}),\, 1 \le i \le n(y)\,\}.$$

These problems were precisely defined in Section 2.8.2.

Theorem 2.21 *Let the cost function* c *be convex. Then*

$$\mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon) \;=\; \inf_{0\le\delta<\varepsilon} \mathrm{c}(\delta)\,\left\lceil \frac{1}{2(\varepsilon-\delta)} \right\rceil$$

and

$$\inf_{0\le\delta<2\varepsilon} \mathrm{c}(\delta)\,\left\lceil \frac{1}{2(2\varepsilon-\delta)} \right\rceil \;\le\; \mathrm{IComp}^{\mathrm{non}}(\mathrm{Int};\varepsilon) \;\le\; \inf_{0\le\delta<\varepsilon} \mathrm{c}(\delta)\,\left\lceil \frac{1}{4(\varepsilon-\delta)} \right\rceil .$$

Furthermore,

$$\mathrm{Comp}(\mathrm{App};\varepsilon) \;\asymp\; \mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon)$$

and

$$\mathrm{Comp}(\mathrm{Int};\varepsilon) \;\asymp\; \mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\alpha(\varepsilon)\,\varepsilon),$$

where $\alpha(\varepsilon) \in [1,2]$.

Proof For both problems we have

$$\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \;=\; \inf \left\{ \sum_{i=1}^n \mathrm{c}(\delta_i) \;\middle|\; \mathrm{r}_n^{\mathrm{wor}}(\delta_1, \ldots, \delta_n) \le \varepsilon \right\},$$

where $\mathrm{r}_n^{\mathrm{wor}}(\delta_1, \ldots, \delta_n)$ is the minimal radius of information using observations with precisions δ_i. The formulas for $\mathrm{r}_n^{\mathrm{wor}}(\delta_1, \ldots, \delta_n)$ are given in Theorem 2.18.

Consider first the function approximation problem. Observe that for δ_i's such that $\sum_{i=1}^n \delta_i = A$, the radius is minimized at $\delta_i = \delta = A/n$, $\forall i$. Due to convexity of c we also have $\sum_{i=1}^n \delta_i \ge n\,\mathrm{c}(\delta)$. Hence,

$$\begin{aligned} \mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon) \;&=\; \inf \left\{ n\,\mathrm{c}(\delta) \;\middle|\; n \ge 1,\, \frac{1}{2n} + \delta \le \varepsilon \right\} \\ &=\; \inf_{0\le\delta<\varepsilon} \mathrm{c}(\delta)\,\left\lceil \frac{1}{2(\varepsilon-\delta)} \right\rceil . \end{aligned}$$

We now turn to the integration. The upper bound for $\mathrm{IComp}^{\mathrm{non}}(\mathrm{Int};\varepsilon)$ can be obtained by setting again $\delta_i = \delta$, $\forall i$. Then,

$$\begin{aligned}\mathrm{IComp}^{\mathrm{non}}(\mathrm{Int};\varepsilon) &\le \inf\left\{ n\,\mathrm{c}(\delta) \mid \quad n \ge 1,\ \frac{1}{4n} + \delta \le \varepsilon \right\} \\ &= \inf_{0\le\delta<\varepsilon} \mathrm{c}(\delta) \left\lceil \frac{1}{4(\varepsilon-\delta)} \right\rceil .\end{aligned}$$

To get the lower bound, we first observe that for all $n \ge 1$ and $A \ge 0$, the maximum

$$M(A,n) = \max\left\{ \sum_{i=1}^{n} \delta_i^2 \;\Big|\; \sum_{i=1}^{n} \delta_i = A,\ \delta_j \le \frac{1}{n}\left(\frac{1}{2}+A\right),\ \forall j \right\}$$

is attained at

$$\delta_j^* = \begin{cases} \frac{1}{n}\left(\frac{1}{2}+A\right) & 1 \le j \le k = \lfloor \frac{nA}{1/2+A} \rfloor, \\ A - \frac{k}{n}\left(\frac{1}{2}+A\right) & j = k+1, \\ 0 & k+2 \le j \le n, \end{cases}$$

and therefore

$$M(A,n) = \sum_{i=1}^{n} (\delta_i^*)^2 \le \left(\frac{nA}{1/2+A}\right)\left(\frac{1/2+A}{n}\right)^2 = \frac{A}{n}\left(\frac{1}{2}+A\right).$$

This yields that for all δ_i's such that $\sum_{i=1}^{n} \delta_i = A$, we have

$$\mathrm{r}_n^{\mathrm{wor}}(\delta_1,\ldots,\delta_n) \ge \frac{1}{n}\left(\frac{1}{2}+A\right)^2 - \frac{A}{n}\left(\frac{1}{2}+A\right) = \frac{1}{4n} + \frac{A}{2n}.$$

Hence,

$$\begin{aligned}\mathrm{IComp}^{\mathrm{non}}(\mathrm{Int};\varepsilon) &\ge \inf\left\{ n\,\mathrm{c}(\delta) \;\Big|\; \frac{1}{4n} + \frac{\delta}{2} \le \varepsilon \right\} \\ &= \inf_{0\le\delta<2\varepsilon} \mathrm{c}(\delta) \left\lceil \frac{1}{2(2\varepsilon-\delta)} \right\rceil .\end{aligned}$$

To prove the remaining part of the theorem, observe that the just proven bounds for $\mathrm{IComp}^{\mathrm{non}}(\mathrm{Int};\varepsilon)$ yield

$$\mathrm{IComp}^{\mathrm{non}}(\mathrm{App};2\,\varepsilon) \le \mathrm{IComp}^{\mathrm{non}}(\mathrm{Int};\varepsilon) \le \mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon). \tag{2.70}$$

Moreover, $\mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon)$ and the upper bound for $\mathrm{IComp}^{\mathrm{non}}(\mathrm{Int};\varepsilon)$ are attained by information which uses $n(\varepsilon)$ observations of the same precision $\delta(\varepsilon)$, and $\delta(\varepsilon) \longrightarrow 0$ as $\varepsilon \longrightarrow 0$. For such information the linear algorithms based on natural splines of order 1 interpolating data, are optimal. Hence, for sufficiently small ε, we have $\mathrm{c}(\delta(\varepsilon)) \geq \mathrm{c}_0$ for some $c_0 > 0$, and consequently $n(\varepsilon) \leq \mathrm{IComp}^{\mathrm{non}}(\varepsilon)/c_0$. The formulas for $\mathrm{Comp}(\varepsilon)$ now follow from (2.70) and Corollary 2.8. □

Clearly, if $\mathrm{IComp}^{\mathrm{non}}(\mathrm{App};2\,\varepsilon) \asymp \mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon)$ (which holds when $\mathrm{c}(\delta)$ tends to infinity not too fast as $\delta \longrightarrow 0$, see E.2.70), then the factor $\alpha(\varepsilon)$ in Theorem 2.21 can be omitted.

To give concrete examples, suppose that the cost function is given as

$$\mathrm{c}_q(\delta) \;=\; \delta^{-q}, \qquad \delta > 0,$$

where $q > 0$, and $\mathrm{c}_q(0) = +\infty$. Then we have

$$\mathrm{IComp}^{\mathrm{non}}(\mathrm{App}, q;\varepsilon) \;\approx\; \left(\frac{1}{\varepsilon}\right)^{q+1} \frac{(q+1)^{q+1}}{2\,q^q}.$$

The best information uses $n(\varepsilon) \approx (1+q)/(2\varepsilon)$ observations with precision $\delta(\varepsilon) \approx q(1+q)^{-1}\varepsilon$. Thus

$$\mathrm{Comp}(\mathrm{App}, q;\varepsilon) \asymp \mathrm{Comp}(\mathrm{Int}, q;\varepsilon) \asymp \mathrm{IComp}^{\mathrm{non}}(\mathrm{App}, q;\varepsilon). \tag{2.71}$$

Note that letting $q \longrightarrow 0$ we get results for exact information with $\mathrm{c} \equiv 1$.

For $\mathrm{c}(\delta) = \max\{\,0, \log_2 \delta^{-1}\}$ for $\delta > 0$, and $\mathrm{c}(0) = +\infty$, we have in turn that

$$\mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon) \;\approx\; \frac{\log_2(1/\varepsilon)}{2\,\varepsilon},$$

and $n(\varepsilon) \approx 1/(2\varepsilon)$, $\delta(\varepsilon) \approx \varepsilon/(\ln \varepsilon^{-1})$. Clearly, (2.71) also holds.

Observe that for any cost function we have the following bounds:

$$\mathrm{c}(\varepsilon)\,\lceil 1/(2\varepsilon)\rceil \;\leq\; \mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon) \;\leq\; \mathrm{c}(\varepsilon/2)\,\lceil 1/\varepsilon\rceil.$$

Hence, the ε–complexity tends to infinity roughly as $\mathrm{c}(\varepsilon)/\varepsilon$, as $\varepsilon \longrightarrow 0$. This means, in particular, that for problems with fixed noise level the complexity is infinite, if only ε is sufficiently small. Actually, this is the consequence of Lemma 2.15 which says that the radius of information cannot be arbitrarily small. We now translate that general result to the language of complexity.

We consider the general linear solution operator $S : F \to G$. The set E is the unit ball of F with respect to a seminorm $\|\cdot\|_F$. We wish to approximate $S(f)$ from noisy, possibly adaptive, observations of linear functionals from a class Λ, with noise bounded always by $\delta_0 > 0$. Recall that this corresponds to a cost function which assumes $+\infty$ on the interval $[0, \delta_0)$.

Theorem 2.22 *Suppose there exists an element $h^* \in F$ such that $h^* \notin \ker S$ and*

$$|L(h^*)| \;\le\; 1, \qquad \textit{for all}\, L \in \Lambda.$$

Then for all $\varepsilon < \min\{\, \delta_0,\ \|h^*\|_F^{-1}\}\ \|S(h^*)\|$ we have

$$\mathrm{Comp}(S;\varepsilon) \;=\; +\infty.$$

Proof It was shown in Lemma 2.15 that for any nonadaptive information $\mathbb{N}$ that uses observations with noise not smaller than δ_0 we have $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \ge \min\{\delta_0, \|h^*\|_F^{-1}\}\ \|S(h^*)\|$. Repeating the same proof we find that the same bound holds for any adaptive information $\mathbb{N}$. Hence, the theorem follows. □

Recall that for the problems App and Int we can take $h^* \equiv 1$. Then, Theorem 2.22 says that $\mathrm{Comp}(\mathrm{App};\varepsilon) = \mathrm{Comp}(\mathrm{Int};\varepsilon) = +\infty$, for all $\varepsilon < \delta_0$ (actually, this is true also for $\varepsilon = \delta_0$). For approximation of a compact operator S in Hilbert spaces of Section 2.10.1, we find that the ε–complexity is infinite if only $\varepsilon < \min\{\,1, \delta_0\}\ \|S\|_F$.

2.10.3 Multivariate approximation in a Banach space

For the problems App and Int the ε–complexity is achieved by nonadaptive information that uses observations with precisions $\delta_i \asymp \varepsilon$. In this section we show that this nice result can be generalized to a class of problems where the noise of information is bounded in the absolute or relative sense. The results will be oriented towards approximation of multivariate functions from noisy data about function values. This particular problem will be defined later. Now, we consider a general problem.

We assume that F is a linear space equipped with a norm $\|\cdot\|_F$ and G is a normed space. The solution operator $S : F \to G$ is linear. We want to approximate $S(f)$ for all f from the ball $\|f\|_F \le 1$, based on noisy observations of some functionals $L \in \Lambda$ at f. Two types of information noise are considered: absolute and relative. More precisely, exact information is given as

$$N(f) \;=\; [\,L_1(f),\ldots,L_n(f))\,]$$

where $L_i \in \Lambda$, $1 \le i \le n$. In the case of noise bounded in the *absolute* sense, the obtained information $y \in \mathbb{R}^n$ satisfies

$$|y_i - L_i(f)| \;\le\; \delta_i,$$

while for noise bounded in the *relative* sense we have

$$|y_i - L_i(f)| \;\le\; \delta_i \cdot |L_i(f)|,$$

$1 \le i \le n$. We stress that the functionals L_i, precisions δ_i, and the number n of observations can depend adaptively on y_j. That is, we deal in general with adaptive information.

To distinguish the absolute and relative noise, we shall sometimes use the subscripts "abs" and "rel".

We start with the analysis of noise bounded in the absolute sense. Let d_n be the minimal diameter of nonadaptive information that uses n exact observations,

$$d_n \;=\; \inf\{\,\mathrm{diam}(N,0)\mid \quad N=[L_1,\ldots,L_n],\, L_i\in\Lambda,\, 1\le i\le n\,\}.$$

We also let $d_0 = 2\,\|S\|_F = 2\,\sup_{\|h\|_F\le 1}\|S(h)\|$. Define the number

$$n^*(\varepsilon) \;=\; \min\{\,n\ge 0\mid \quad d_n(0)\le 2\,\varepsilon\,\}$$

$(\min\emptyset = +\infty)$.

We first show a lower bound on $\mathrm{Comp}_{\mathrm{abs}}(\varepsilon)$. To this end, assume that the following condition is satisfied. There exists a constant K, $0<K<+\infty$, such that

$$|L(h)| \;\le\; K\cdot\|S(h)\|, \qquad \text{for all } \; L\in\Lambda \quad \text{and} \quad h\in F. \tag{2.72}$$

Lemma 2.19 *Suppose that the κ–strongly hard element exists. If the condition (2.72) is satisfied then*

$$\mathrm{Comp}_{\mathrm{abs}}(\varepsilon) \;\ge\; n^*(\kappa\,\varepsilon)\cdot \mathrm{c}(K\,\kappa\,\varepsilon).$$

Proof We first show that

$$\mathrm{IComp}^{\mathrm{non}}_{\mathrm{abs}}(\varepsilon) \;\geq\; n^*(\varepsilon)\cdot \mathrm{c}(K\,\varepsilon). \tag{2.73}$$

If $\varepsilon \geq \|S\|_F$ then the zero approximation is optimal. Hence, $n^*(\varepsilon) = 0$ and (2.73) follows.

Let $\varepsilon < \|S\|_F$. Let $\mathbb{N} = \{N, \Delta\}$ where $N = [L_1, \dots, L_n]$ and $\Delta = [\delta_1, \dots, \delta_n]$, $0 \leq \delta_1 \leq \cdots \leq \delta_n$, be an arbitrary information operator with the radius $\mathrm{rad}^{\mathrm{wor}}_{\mathrm{abs}}(\mathbb{N}) \leq \varepsilon$. Let

$$k \;=\; \max\left\{\, i \leq n \;\Big|\; \delta_i \leq \frac{1}{2} K\,\mathrm{diam}_{\mathrm{abs}}(\mathbb{N}) \,\right\}.$$

(If $\delta_1 > K\,\mathrm{diam}_{\mathrm{abs}}(\mathbb{N})/2$ then $k = 0$.) We claim that

$$k \;\geq\; n^*(\varepsilon). \tag{2.74}$$

To show this, it suffices that for information $\mathbb{N}' = \{N', \Delta'\}$ where $N' = [L_1, \dots, L_k]$ and $\Delta' = [\delta_1, \dots, \delta_k]$ (or for information $\mathbb{N}' \equiv \{0\}$ if $k = 0$), we have $\mathrm{diam}_{\mathrm{abs}}(\mathbb{N}') \leq 2\,\varepsilon$.

Indeed, suppose to the contrary that $\mathrm{diam}_{\mathrm{abs}}(\mathbb{N}') > 2\,\varepsilon$. Then there is $h \in F$ such that $\|h\|_F \leq 1$, $|L_i(h)| \leq \delta_i$, $1 \leq i \leq k$, and $2\,\|S(h)\| > 2\,\varepsilon \geq \mathrm{diam}_{\mathrm{abs}}(\mathbb{N})$. Let

$$h' \;=\; \min\left\{1,\, \frac{\delta_{k+1}}{K\,\|S(h)\|}\right\}\cdot h.$$

Then $\|h'\|_F \leq 1$, and for all $k+1 \leq j \leq n$ it holds

$$|L_j(h')| \;\leq\; K\cdot\|S(h')\| \;\leq\; \min\{\,K\,\|S(h)\|, \delta_j\,\} \;=\; \delta_j.$$

Since also for $1 \leq i \leq k$ we have $|L_i(h')| \leq |L_i(h)| \leq \delta_i$,

$$\begin{aligned}\mathrm{diam}_{\mathrm{abs}}(\mathbb{N}) \;\geq\; 2\,\|S(h')\| \;&=\; 2\,\min\left\{1,\, \frac{\delta_{k+1}}{K\,\|S(h)\|}\right\}\,\|S(h)\| \\ &=\; \min\left\{2\,\|S(h)\|,\, 2\frac{\delta_{k+1}}{K}\right\} \;>\; \mathrm{diam}_{\mathrm{abs}}(\mathbb{N}),\end{aligned}$$

which is a contradiction. Hence, $\mathrm{diam}_{\mathrm{abs}}(\mathbb{N}') < 2\,\varepsilon$. Since information $\mathbb{N}'$ uses k observations, (2.74) follows.

Observe that (2.74) also yields $k \geq 1$. Hence, we have

$$\begin{aligned}\mathrm{comp}(\mathbb{N}) \;&=\; \sum_{i=1}^{n} \mathrm{c}(\delta_i) \;\geq\; \sum_{i=1}^{k} \mathrm{c}(\delta_i) \;\geq\; k\cdot\mathrm{c}\left(\frac{1}{2}K\,\mathrm{diam}(\mathbb{N})\right) \\ &\geq\; k\,\mathrm{c}(K\varepsilon) \;\geq\; n^*(\varepsilon)\cdot\mathrm{c}(K\varepsilon).\end{aligned}$$

Since $\mathbb{N}$ was arbitrary, the proof of (2.73) is complete.

Now, Theorem 2.19 together with (2.73) yields

$$\mathrm{Comp}_{\mathrm{abs}}(\varepsilon) \;\geq\; \mathrm{IComp}^{\mathrm{non}}(\kappa\,\varepsilon) \;\geq\; n^*(\kappa\,\varepsilon)\cdot \mathrm{c}(\kappa\varepsilon)$$

which proves the lemma. □

To show an upper bound, we assume that for any $n \geq 1$ and $\delta > 0$, there exists information $\mathbb{N}$ that uses n observations with the precision vector $\Delta = [\underbrace{\delta, \ldots, \delta}_{n}]$, and a linear algorithm φ, such that

$$\mathrm{e}^{\mathrm{wor}}_{\mathrm{abs}}(\mathbb{N},\varphi) \;\leq\; M\cdot(d_n+\delta). \tag{2.75}$$

Here M is an absolute positive constant independent of n and δ.

Lemma 2.20 *If the condition (2.75) is satisfied then*

$$\mathrm{Comp}_{\mathrm{abs}}(\varepsilon) \;\leq\; n^*(m\varepsilon)\,(\,\mathrm{c}(m\varepsilon)+2\,g)$$

where $m = (3M)^{-1}$.

Proof Let $\delta = m\varepsilon$ and $n = n^*(\delta)$. Let $\mathbb{N}$ be such information that it uses n observations with precisions δ, and let φ be such an algorithm that $\mathrm{e}^{\mathrm{wor}}_{\mathrm{abs}}(\mathbb{N},\varphi) \leq M(d_n+\delta)$. Since $d_n \leq 2\,\delta$, we have

$$\mathrm{e}^{\mathrm{wor}}_{\mathrm{abs}}(\mathbb{N},\varphi) \;\leq\; 3\,\delta\,M \;\leq\; 3\,M\,m\,\varepsilon \;=\; \varepsilon.$$

Hence,

$$\begin{aligned} \mathrm{Comp}_{\mathrm{abs}}(\varepsilon) \;&\leq\; \mathrm{comp}(\mathbb{N},\varphi) \;\leq\; n\,\mathrm{c}(\delta) \,+\, (2n-1)g \\ &\leq\; n\,(\mathrm{c}(\delta)+2g) \;=\; n\,(m\,\varepsilon))\;(\,\mathrm{c}\,(m\,\varepsilon)\,+\,2g)\,, \end{aligned}$$

as claimed. □

The upper and lower bounds on $\mathrm{Comp}_{\mathrm{abs}}(\varepsilon)$ give the following theorem.

Theorem 2.23 *Assume that for any* $\alpha > 0$,

$$n^*(\alpha\,\varepsilon) \;\asymp\; n^*(\varepsilon) \qquad \textit{and} \qquad \mathrm{c}(\alpha\,\varepsilon) \;\asymp\; \mathrm{c}(\varepsilon),$$

as $\varepsilon \to 0^+$. *If the conditions (2.72) and (2.75) are satisfied and the* κ*–strongly hard element exists, then*

$$\mathrm{Comp}_{\mathrm{abs}}(\varepsilon) \;\asymp\; n^*(\varepsilon)\cdot\mathrm{c}(\varepsilon), \qquad \textit{as} \quad \varepsilon \to 0^+.$$

Furthermore, optimal information uses $n \asymp n^*(\varepsilon)$ *observations with the same precision* $\delta \asymp \varepsilon$. □

This theorem has a very useful interpretation. It states that the ε–complexity is proportional to the cost $\mathrm{c}(\varepsilon)$ of obtaining the value of a functional with precision ε, and to the complexity in exact information case with $\mathrm{c} \equiv 1$. We stress that Theorem 2.23 applies *only* to problems for which $\mathrm{c}(\varepsilon)$ and $n^*(\varepsilon)$ tend to infinity at most polynomially in $1/\varepsilon$, as $\varepsilon \to 0^+$. It seems also worthwhile to mention that we obtained the complexity results without knowing exact formulas for the minimal radii $\mathrm{r}_n^{\mathrm{wor}}(\Delta)$.

We now pass to the case of relative noise. We assume that all functionals $L \in \Lambda$ satisfy $\|L\|_F \leq 1$, and that there exists h_0 such that $\|h_0\|_F \leq 1$ and

$$\inf_{L\in\Lambda} |L(h_0)| \;=\; A \;>\; 0. \tag{2.76}$$

Theorem 2.24 *Suppose that the assumptions of Theorem 2.23 and the condition (2.76) are satisfied. Then*

$$\mathrm{Comp}_{\mathrm{rel}}(\varepsilon) \;\asymp\; \mathrm{Comp}_{\mathrm{abs}}(\varepsilon) \;\asymp\; n^*(\varepsilon)\cdot \mathrm{c}(\varepsilon).$$

Proof Observe first that if $|y_i - L_i(f)| \leq \delta_i |L_i(f)|$ then $|y_i - L_i(f)| \leq \delta_i \|L_i\|_F \|f\|_F \leq \delta_i$. This means that for any information $\mathbb{N}$ and f with $\|f\|_F \leq 1$, we have $\mathbb{N}_{\mathrm{rel}}(f) \subset \mathbb{N}_{\mathrm{abs}}(f)$. Hence,

$$\mathrm{e}_{\mathrm{rel}}^{\mathrm{wor}}(\mathbb{N},\varphi) \;\leq\; \mathrm{e}_{\mathrm{abs}}^{\mathrm{wor}}(\mathbb{N},\varphi), \quad \forall\,\mathbb{N},\,\forall\,\varphi,$$

and consequently

$$\mathrm{Comp}_{\mathrm{rel}}(\varepsilon) \;\leq\; \mathrm{Comp}_{\mathrm{abs}}(\varepsilon).$$

This shows the upper bound for $\mathrm{Comp}_{\mathrm{rel}}(\varepsilon)$.

As in Section 2.7.2, we can show that for any adaptive information $\mathbb{N}$ we have

$$\mathrm{rad}_{\mathrm{rel}}^{\mathrm{wor}}(\mathbb{N}) \;\geq\; \frac{1}{2}\,\min\,\{1-\|\alpha h_0\|_F, \alpha A/2\}\cdot \mathrm{rad}_{\mathrm{abs}}^{\mathrm{wor}}(\mathbb{N}).$$

Taking $\alpha = (1+A/2)^{-1}$ we obtain

$$\mathrm{rad}_{\mathrm{rel}}^{\mathrm{wor}}(\mathbb{N}) \;\geq\; \frac{1}{2}\frac{A}{A+2}\,\mathrm{rad}_{\mathrm{abs}}^{\mathrm{wor}}(\mathbb{N}).$$

Hence, for any $B > 2(A+2)/A$ we have

$$\begin{aligned}\mathrm{Comp}_{\mathrm{rel}}(\varepsilon) \;&\geq\; \mathrm{IComp}_{\mathrm{rel}}^{\mathrm{non}}(\varepsilon) \;\geq\; \mathrm{IComp}_{\mathrm{abs}}^{\mathrm{non}}(B\varepsilon)\\ &\asymp\; \mathrm{IComp}_{\mathrm{abs}}^{\mathrm{non}}(\varepsilon) \;\asymp\; \mathrm{Comp}_{\mathrm{abs}}(\varepsilon),\end{aligned}$$

which shows the lower bound for $\mathrm{Comp}_{\mathrm{rel}}(\varepsilon)$ and completes the proof. □

Thus, under some assumptions, the cases of relative and absolute noise are (almost) equivalent. We note that such an equivalence does not always hold. For instance, for the problems App and Int of Section 2.10.2 and information $\mathbb{N}$ using n observations of function values with precision $\delta \in (0,1)$, we have $\mathrm{rad}^{\mathrm{wor}}_{\mathrm{rel}}(\mathbb{N}) = +\infty$. Indeed, the vector $y = [\underbrace{a, \ldots, a}_{n}]$, $a > 0$, is noisy information about $f_1 \equiv a/(1-\delta)$ and $f_{-1} \equiv a/(1+\delta)$. We also have $f_1, f_{-1} \in E$. Hence, for $S \in \{\mathrm{App}, \mathrm{Int}\}$

$$\begin{aligned} \mathrm{rad}^{\mathrm{wor}}_{\mathrm{rel}}(\mathbb{N}) &\ge \frac{1}{2}\|S(f_1) - S(f_{-1})\| \\ &= a\,\frac{2\delta}{1-\delta^2} \longrightarrow +\infty, \qquad \text{as } a \to +\infty. \end{aligned}$$

(See also E 2.74.)

Multivariate approximation

We now apply the obtained results to a concrete problem. We consider approximation of multivariate functions from noisy data.

Let $F = F^r_s$ be the space of all real valued functions defined on the s–dimensional unit cube $D = [0,1]^s$ that possess all partial continuous derivatives of order r, $r \ge 1$. The norm in F^r_s is given as

$$\|f\|_F = \max_{0 \le k_1 + \cdots + k_s = i \le r} \sup_{t \in D} \left| \frac{\partial^i f(t)}{(\partial x^1)^{k_1} \ldots (\partial x^s)^{k_s}} \right|, \quad f \in F,$$

where $t = [t^1, \ldots, t^s]$. Information about f is given by noisy values of f at some points, i.e., exact information is of the form

$$N(f) = [\, f(t_1), f(t_2), \ldots, f(t_n)\,],$$

where $t_i \in D$, $1 \le i \le n$. We want to approximate $S(f) = f$ in the sup–norm. That is, formally $S : F \to G$ where G is the space of continuous functions $f : D \to \mathbb{R}$ with the norm

$$\|g\| = \|g\|_\infty = \sup_{t \in D} |g(t)|.$$

We shall show that the assumptions of Theorems 2.23 and 2.24 are satisfied. Clearly, there exists the κ–strongly hard element, for all $\kappa > 2$. Since for any $t \in D$ and $f \in F_s^r$ we have

$$|f(t)| \;\le\; \|f\|_\infty \;\le\; \|f\|_F,$$

the condition (2.72) holds with $K = 1$, and $\|L\|_F \le 1$ for any functional L of the form $L(f) = f(t)$. It is also easily seen that (2.76) is satisfied with the function $f \equiv 1$ and $A = 1$. Hence, it remains to show (2.75). We do it in two steps.

Lemma 2.21 *For the multivariate approximation we have*

$$d_n \;\ge\; \gamma \cdot n^{-r/s}$$

where γ is positive and independent of n.

Proof Let $\psi : \mathbb{R} \to \mathbb{R}$ to be an arbitrary nonzero function such that
(i) $\psi(x) = 0$, for all $|x| \ge 1/2$, and
(ii) the rth derivative $\psi^{(r)}$ exists and is continuous.
Let $\Psi : \mathbb{R}^s \to \mathbb{R}$,

$$\Psi(t) \;=\; \alpha\, \psi(t^1) \cdots \psi(t^s)$$

where $\alpha \neq 0$ is chosen in such a way that $\|\Psi\|_F \le 1$.

Let $n \ge 1$ and let $\mathbb{N}(f) = [\,f(t_1), \ldots, f(t_n)]$, $t_i \in D$, be an arbitrary exact nonadaptive information. Define $m \ge 1$ in such a way that $(m/2)^s \le n < m^s$, and the set $\mathcal{K} \subset D$ of m^s points,

$$\mathcal{K} \;=\; \left\{ x = [x^1, \ldots, x^s] \in \mathbb{R}^s \;\Big|\; x^j = \frac{2i_j - 1}{m},\, 1 \le i_j \le m,\, 1 \le j \le s \right\}.$$

The set $\mathcal{K}$ determines the collection of m^s functions

$$\Psi_x(t) \;=\; m^{-r}\Psi(m(t - x)\,), \qquad x \in \mathcal{K}.$$

They are linearly independent and, moreover, they have mutually different supports. Since $n < m^s$, there exist real coefficients β_x, $x \in \mathcal{K}$, not all equal to zero, such that the function

$$f_N \;=\; \sum_{x \in \mathcal{K}} \beta_x \Psi_x$$

is in $\ker N$. We can also assume $\max_{x\in\mathcal{K}}|\beta_x| = 1$ so that $\|f_N\|_F \le 1$. Hence,

$$d_n \;\ge\; \|S(f_N)\| \;=\; \|f_N\|_\infty \;=\; m^{-r}\|\Psi\|_\infty \;=\; n^{-r/s}\,2^{-r}\alpha\,\|\psi\|_\infty^s,$$

and the lemma holds with $\gamma = 2^{-r}\alpha\|\psi\|_\infty^s$. $\Box$

We now exhibit exact nonadaptive information N_n and a linear algorithm φ_n such that

$$\mathrm{e}^{\mathrm{wor}}(N_n, [\underbrace{\delta,\ldots,\delta}_{n}], \varphi_n) \;\le\; M(n^{-r/s} + \delta),$$

for all δ and $n \ge (r-1)^s$, where M is independent of n and δ.

Assume first that $r \ge 2$. Let $n \ge r^s$. Let $k \ge 1$ be the largest integer such that $(k(r-1)+1)^s = m \le n$. Information N_n consists of function evaluations at m equispaced points, i.e., $N(f) = \{f(t)\}_{t\in\mathcal{K}}$, where

$$\mathcal{K} \;=\; \left\{ t = [t^1,\ldots,t^s] \;\Big|\; t^i = \frac{i_j}{k(r-1)},\; 0 \le i_j \le k(r-1),\; 1 \le i \le s \right\}.$$

Let $h = 1/k$. Divide the cube $D = [0,1]^s$ onto k^s subcubes

$$D_{i_1\ldots i_k} \;=\; \times_{j=1}^s[(i_j-1)h, i_jh],$$

$1 \le i_j \le k$, $1 \le j \le s$. Observe that each subcube contains exactly r^s points from $\mathcal{K}$. For given information $y = \{y_x\}_{x\in\mathcal{K}}$ about $f \in F_s^r$, the approximation $\varphi_n(y)$ is given as such a function $w = w_y$ that

(i) on each subcube w is a polynomial of the form

$$w(t) \;=\; \sum a_{i_1\ldots i_s}(t^1)^{i_1}\cdots(t^s)^{i_s},$$

where the summation is taken over all $0 \le i_j \le r-1$, $1 \le j \le s$,

(ii) w interpolates the data y, i.e.,

$$w(x) \;=\; y_x, \qquad \forall x \in \mathcal{K}.$$

Note that w exists and is determined uniquely for any information y. Moreover, w depends linearly on y.

Lemma 2.22 *If $|y_x - f(x)| \le \delta$, $\forall x \in \mathcal{K}$, then for all $t \in D$ we have*

$$|f(t) - w(t)| \;\le\; \frac{h^r}{r!}\left(\sum_{i=0}^{s-1} A^i\right) + \delta\, A^s$$

where

$$A \;=\; \sup_{0\le x\le r-1} \sum_{i=0}^{r-1} \left| \prod_{j=0,\neq i}^{r-1} \frac{x-j}{i-j} \right| .$$

Proof We prove the lemma by induction on s. We assume without loss of generality that $t\in[0,h]^s$.

Let $s=1$. Let w_f be the polynomial of degree at most $r-1$ that interpolates f at the points $t_i=ih/(r-1)$, $0\le i\le r-1$, i.e.,

$$w_f(x) \;=\; \sum_{i=0}^{r-1} f(t_i) \left(\prod_{j=0,\neq i}^{r-1} \frac{x-t_j}{t_i-t_j} \right) .$$

From the well known formula for the error of interpolation we get

$$\begin{aligned} f(t)-w(t) \;&=\; (\,f(t)-w_f(t)\,) + (\,w_f(t)-w(t)\,) \\ &=\; \frac{f^{(r)}(u(t)\,)}{r!} \prod_{i=0}^{r-1}(t-t_i) \,+\, \sum_{i=0}^{r-1}(y_i-f(t_i)\,) \left(\prod_{j=0,\neq i}^{r-1} \frac{t-t_j}{t_i-t_j} \right) \end{aligned}$$

where $0\le u(t)\le h$. Hence,

$$|f(t)-w(t)| \;\le\; (r!)^{-1}h^r \,+\, \delta\,A.$$

Let $s>1$. Let $t=[t^1,\ldots,t^s]\in[0,h]^s$. Consider the function of one variable $f_{t^1\ldots t^{s-1}}(x)=f(t^1,\ldots,t^{s-1},x)$, and the corresponding polynomial $w_{t^1\ldots t^{s-1}}(x)=w(t^1,\ldots,t^{s-1},x)$, $0\le x\le h$. From the inductive assumption it follows that for all $x=ih/(r-1)$, $0\le i\le r-1$, we have

$$|f_{t^1\ldots t^{s-1}}(x) - w_{t_1\ldots t^{s-1}}(x)| \;\le\; \frac{h^r}{r!}\left(\sum_{i=0}^{s-2} A^i\right) \,+\, \delta\,A^{s-1}.$$

As in the case $s=1$, let w_f be the polynomial of one variable that interpolates $f_{t^1\ldots t^{s-1}}$ at $x=ih/(r-1)$, $0\le i\le r-1$. Then we have

$$\begin{aligned} |f(t)-w(t)| \;&\le\; |f_{t^1\ldots t^{s-1}}(t^s)-w_f(t^s)| + |w_f(t^s)-w_{t^1\ldots t^{s-1}}(t^s)| \\ &\le\; \frac{h^r}{r!} + \left(\frac{h^r}{r!}\left(\sum_{i=0}^{s-2} A^i\right) + \delta A^{s-1} \right) A \\ &=\; \frac{h^r}{r!}\left(\sum_{i=0}^{s-1} A^i\right) \,+\, \delta\,A^s, \end{aligned}$$

as claimed. □

Observe now that $h \approx (r-1)\, n^{-1/s}$. Thus Lemmas 2.21 and 2.22 yield $d_n \asymp n^{-r/s}$, and there exists a positive constant M such that for any information y about f

$$\|f - \varphi_n(y)\|_\infty \;\leq\; M\,(\,d_n + \delta\,),$$

This is true also for $r = 1$ since then we can use the same information and algorithm as for $r = 2$.

We summarize our analysis in the following theorem.

Theorem 2.25 *If the cost function* c *satisfies*

$$\mathrm{c}(\alpha\,\delta) \;\asymp\; \mathrm{c}(\delta), \qquad \forall \alpha > 0,$$

then for the multivariate approximation we have

$$\mathrm{Comp}_{\mathrm{abs}}(\varepsilon) \;\asymp\; \mathrm{Comp}_{\mathrm{rel}}(\varepsilon) \;\asymp\; \mathrm{c}(\varepsilon)\cdot\varepsilon^{-s/r}.$$

Furthermore, optimal information uses $n = \varepsilon^{-s/r}$ equidistant observations with precision $\delta \asymp \varepsilon$, and piecewise polynomial approximation is the optimal algorithm.

Notes and Remarks

NR 2.41 Sections 2.10.1 and 2.10.2are original, while Section 2.10.3 is based on Kacewicz and Plaskota [32].

NR 2.42 Using results from the average case analysis of Chapter 3, one can also obtain some complexity results for integration in the r.k.h.s $W_r^0(0,1)$ when only observations of function values are allowed. See NR 3.30 for details.

NR 2.43 Approximation of smooth functions of a single variable in the case of noise bounded in the absolute sense by a fixed constant, was also studied by Lee *et al.* [47]. In that paper the complexity of information is measured by the memory needed to store it. Consequently, the ε–information complexity can be interpreted as the minimal amount of memory sufficient to store information, from which it is possible to recover a function with given accuracy. With such an interpretation, the case of absolute noise corresponds to the fact that the *fixed* point representation of y_i is used with roughly $\max\{\,0, \log_2 1/\delta_i\}$ bits. The relative noise in turn corresponds to the *floating* point representation using the same number of bits. In both cases the cost function is $\mathrm{c}(\delta) = \max\{0, \log_2 1/\delta\,\}$. A more detailed discussion on this subject can be found in Kacewicz and Plaskota [32].

NR 2.44 The techniques used in the proofs of Lemmas 2.21 and 2.22 are well known and often applied to evaluate diameters of problems with exact information, and also some n–widths in approximation theory (see also NR 2.28). The fact that for multivariate approximation we have $\mathrm{r}_n^{\mathrm{wor}}(\delta) \le M(d_n + \delta)$ can be derived from Babenko [3].

Exercises

E 2.67 Consider the Hilbert case. Suppose that the cost function c satisfies the following condition: there exists x_0 and $d > 0$ such that $x\mathrm{c}(x) > d$, $\forall x \ge x_0$. Prove that then

$$\mathrm{IComp}^{\mathrm{non}}(\mathrm{c};\varepsilon) \;\ge\; \frac{d}{x_0} \cdot \mathrm{IComp}^{\mathrm{non}}\left(\mathrm{c}_0; \sqrt{1+\frac{1}{x_0}}\;\varepsilon\right), \quad \forall \varepsilon > 0.$$

Show also that if the condition is not satisfied then $\mathrm{IComp}^{\mathrm{non}}(\mathrm{c};\varepsilon) = 0$, $\forall \varepsilon > 0$.

E 2.68 Let the cost function $\mathrm{c}_{\mathrm{ln}}(\delta) = \ln(1+\delta^{-2})$. Prove that then in the Hilbert case we have

$$\mathrm{R}(T)^2 \;=\; d \cdot \left(\frac{\prod_{j=1}^n \lambda_j}{e^T}\right)^{1/n},$$

where $n = n(T)$ is the largest integer for which $\lambda_n \ge (\prod_{j=1}^n \lambda_j)^{1/n} e^{-T/n}$, and $1/2 \le d \le 1$.

E 2.69 Use the previous exercise to show that for $\lambda_j = j^{-p}$, $j \ge 1$, we have $\mathrm{Comp}(\mathrm{c}_{\mathrm{ln}};\varepsilon) \asymp \mathrm{Comp}(\mathrm{c}_0;\varepsilon)$, for all $p > 0$, while for $\lambda_j = e^{-j}$ we have $\mathrm{Comp}(\mathrm{c}_{\mathrm{ln}};\varepsilon) \asymp \ln(1/\varepsilon)^2$, and $\mathrm{Comp}(\mathrm{c}_0;\varepsilon) \asymp \ln(1/\varepsilon)$.

E 2.70 Show that the equivalence $\mathrm{IComp}^{\mathrm{non}}(\mathrm{App};2\,\varepsilon) \asymp \mathrm{IComp}^{\mathrm{non}}(\mathrm{App};\varepsilon)$ holds iff $\mathrm{c}(\delta)$ tends to infinity not faster than polynomially in $1/\delta$, as $\delta \to 0$.

E 2.71 Show that Theorem 2.23 can be applied for the problems App and Int.

E 2.72 (Kacewicz and Plaskota) Let $F = F_s^r$ be the space defined as in the multivariate approximation problem. Let Λ be the class of functionals $L : F \to \mathbb{R}$ of the form

$$L(f) \;=\; \frac{\partial^i f(t)}{(\partial x^1)^{k_1} \ldots (\partial x^s)^{k_s}}, \qquad \text{for some} \quad t \in [0,1]^s,$$

where $k_1, \ldots, k_s$ and i are certain integers such that $0 \le k_1 + \cdots + k_s = i \le k$, where $0 \le k \le r$. Show that then

$$\sup_{\|f\|_F \le 1} \inf_{L \in \Lambda} |L(f)| \;\ge\; e^{-\min\{s,k\}}\,(\min\{1, k/s\})^k$$

(with the convention that $0^0 = 1$).

E 2.73 (Kacewicz and Plaskota) Show that if the space F and the class Λ are as in the previous exercise, then for any solution operator S, nonadaptive information N, and precision vector Δ, we have

$$\frac{\mathrm{diam}_{\mathrm{abs}}(N,\Delta)}{1+2\,e^{\min\{k,s\}}\,(\min\{1,k/s\})^{-k}} \;\le\; \mathrm{diam}_{\mathrm{rel}}(N,\Delta) \;\le\; \mathrm{diam}_{\mathrm{abs}}(N,\Delta).$$

E 2.74 (Kacewicz and Plaskota) For given p, $1 \le p < +\infty$, define $F = \{\, f \in \mathbb{R}^\infty \mid \|f\|_p < +\infty \,\}$, where $\|f\|_p = (\sum_{i=1}^\infty |f_i|^p)^{1/p}$, $f = [f_1, f_2, \ldots] \in \mathbb{R}^\infty$. For $\|f\|_p \le 1$, we approximate values $S(f)$ of the operator $S: F \to F$,

$$S(f) \;=\; [\,\alpha_i f_1, \alpha_2 f_2, \ldots],$$

where $\alpha_i = 2^{(1-i)/p}$, $i \ge 1$. Exact information is given as $N_n(f) = [f_1, f_2, \ldots, f_n]$. Show that

$$\sup_{\|f\|_p \le 1} \; \min_{1 \le i \le n} \, |f_i| \;=\; n^{-1/p},$$

and that for $\delta_i = \alpha_{n+1}/\alpha_i$, $1 \le i \le n$, $\Delta_n = [\delta_1, \ldots, \delta_n]$, we have

$$\frac{\mathrm{diam}_{\mathrm{rel}}(N_n,\Delta_n)}{\mathrm{diam}_{\mathrm{abs}}(N_n,\Delta_n)} \;\le\; n^{-1/p}.$$

E 2.75 Theorem 2.23 cannot be applied if instead of the multivariate approximation problem, $S(f) = f$, the multivariate integration, $S(f) = \int_D f(t)dt$, is considered. Why?

Chapter 3

Average case setting

3.1 Introduction

This chapter is devoted to the average case setting. In the average case setting, we are interested in the *average* performance of the error and cost of algorithms. The material is organized similarly to the worst case setting of Chapter 2. That is, we first deal with optimal algorithms, then we analyze the optimal information, and finally, we present some complexity results.

To study the average performance of the error and/or cost, we have to assume some probability distribution μ on the space F of the problem elements as well as some distribution of the information noise. The latter assumption means that information is corrupted with *random* noise. Basically, we consider Gaussian distributions (measures) which seem to be most natural and are most often used in practice.

In Section 3.2, we give a general formulation of the average case setting. We also introduce the concept of the (average) radius of information which, similarly to the worst case, provides a sharp lower bound on the (average) error of algorithms.

Then we pass to linear problems with Gaussian measures. These are problems where the solution operator is linear, μ is a Gaussian measure, and information is linear with Gaussian noise. In Section 3.3, we recall what a Gaussian measure on a Banach space is and list some of its important properties. In Sections 3.4 to 3.6 we study optimal algorithms. Formulas for the optimal algorithm and radius of information are presented in Section 3.8.1. The optimal algorithm turns out to be linear and unique. In Section 3.5, we specialize the obtained results to the solution operator being a linear functional. In particular, we show that in this case the problem is as difficult as an appropriately chosen one dimensional subproblem.

As we know, in the worst case setting optimal algorithms are smoothing splines with appropriately chosen parameter α. It turns out that a similar fact holds in the average case. More precisely, in Section 3.6 we show that for linear problems with Gaussian measures, the optimal algorithm can be interpreted as 1/2–smoothing spline algorithm. This smoothing spline corresponds to Hilbert norms $\|\cdot\|_H$ and $\|\cdot\|_Y$ which are induced by the distribution μ on F and by the distribution of noise, correspondingly. Note that, unlike in the worst case, the optimal parameter $\alpha = 1/2$ is constant and the optimal regularization parameter γ equals the variance σ^2 of noise.

The fact that smoothing splines are optimal in the worst and average cases enables us to establish a correspondence between the both settings. Namely, the optimal algorithm for the average case is almost optimal for a corresponding problem in the worst case, where the set $E \subset F$ is the unit ball in $\|\cdot\|_H$ and the noise is bounded uniformly in the norm $\|\cdot\|_Y$. Moreover, for approximating a linear functional, the corresponding worst and average radii of the same linear information differ only by a factor of $\sqrt{2}$.

Next, we allow information to vary. In Section 3.7 we carefully define nonadaptive and adaptive information. Then we show that for linear problems with Gaussian measures, adaptive information cannot reduce the minimal average error given by nonadaptive information. That is, adaption does not help with respect to the error.

The problem of optimal information is studied in Section 3.8. Using a similar technique to that from the worst case with Hilbert norms, in Section 3.8.1 we show the optimal selection of functionals forming information, for n independent observations with given variances σ_i^2 of noise. The formulas for optimal information and minimal radius are given in terms of σ_i^2's and eigenvalues of the correlation operator of the a priori Gaussian distribution on the space G. It turns out that for independent observations with the same variances, the minimal radius converges to zero with n, but not faster than $\sigma/\sqrt{n}$. We also show relations between optimal information in the average case and the corresponding worst case settings. We construct information which is almost optimal for both settings.

Tight bounds on the minimal error for function approximation and integration on the Wiener space, are found in Section 3.8.2. For these problems, independent noisy observations of function values with the same variances are assumed. Observations at equidistant points turn out to be almost optimal.

In the last two sections we study the average case complexity. In Section 3.9 we show the second theorem on adaption. It says that, under some assumptions, adaption cannot help not only with respect to the error but also with respect to the average cost of information. That is, for any adaptive information there exists nonadaptive information whose radius and cost are not (much) larger than the radius and cost of the adaptive information. This holds when the minimal cost of information with radius not greater than $\sqrt{\varepsilon}$ is a semiconvex function of ε. This fact and linearity of optimal algorithms imply, similarly to the worst case, that the ε–complexity essentially equals the information complexity.

In Section 3.10 we apply the general complexity results to two special problems. We first consider a linear continuous problem with Gaussian measures and with information consisting of functionals bounded by 1 in a norm induced by the measure μ. We find sharp bounds on the ε–complexity dependent on the cost function. We note that the situation here reminds that from the worst case setting with Hilbert norms.

Finally, we show some complexity bounds for function approximation and integration on the Wiener space, based on information about noisy function values.

3.2 Information and its radius

Let $S : F \to G$, where F is a linear space and G is a normed space, be a given solution operator. As in the worst case setting, we wish to find approximations to $S(f)$ for $f \in F$. Basically, the approximations are constructed as before, i.e., by means of an algorithm that uses some information. However, we now assume that the elements $f \in F$ as well as information values y are distributed randomly, according to some probability measures.

More specifically, we assume that the space F is equipped with a probability measure μ defined on a σ–field of F, with respect to which S is a measurable mapping. The measure μ shows the probability of occurrence of elements $f \in F$. Observe that the exact solution $S(f)$ can also be viewed as a random variable distributed according to the measure $\nu = \mu S^{-1}$,

$$\nu(B) \;=\; \mu(S^{-1}(B)\,) \;=\; \mu\,(\{\, f \in F \mid \quad S(f) \in B\}),$$

for all Borel sets B of G.

An *information operator* assigns to any $f \in F$ a probability measure π_f on a set Y of real sequnces. This measure shows how often certain values

$y \in Y$ occur when gaining information about f. Formally, an information operator as a mapping

$$\mathbb{N} : \; F \;\rightarrow\; P_Y$$

where P_Y denotes all probability distributions on the Borel sets of Y. The Borel structure on Y is given in a natural way. That is, $B \subset Y$ is a Borel set iff $B \cap \mathbb{R}^n$ are Borel sets of $\mathbb{R}^n$. In particular, $Y \cap \mathbb{R}^n$ must be Borel sets of $\mathbb{R}^n \;\; \forall n$. We additionally assume that the mapping $f \rightarrow \pi_f(B)$ is measurable for any measurable set $B \subset Y$.

Noisy information about $f \in F$ is any vector $y \in Y$ which is a realization of the random variable distributed according to $\pi_f = \mathbb{N}(f)$.

If π_f is a Dirac measure for any $f \in F$ a.e. (almost everywhere), i.e., if there are elements $N(f) \in Y$ such that for any B

$$\pi_f(B) \;=\; \begin{cases} 0 & N(f) \notin B, \\ 1 & N(f) \in B, \end{cases}$$

then information is called *exact.* In this case the vector $y = N(f)$ is observed with probability one. Otherwise, information is *noisy.*

We now give two examples.

Example 3.1 Suppose we want to approximate a one dimensional random variable f with normal distribution, $f \sim \mathcal{N}(0, \lambda)$ where $\lambda > 0$, based on information $y \sim \mathcal{N}(f, \sigma^2)$, $\sigma^2 \geq 0$. In this case $F = G = \mathbb{R}$ and $S(f) = f$. The measure μ on F is defined as

$$\mu(B) \;=\; \frac{1}{\sqrt{2\pi\lambda}} \int_B e^{-x^2/(2\lambda)} \, dx, \quad \forall B\text{– Borel set of } \mathbb{R}.$$

Furthermore, the information operator $\mathbb{N} : \mathbb{R} \rightarrow P_{\mathbb{R}}$ is given as $\mathbb{N}(f) = \mathcal{N}(f, \sigma^2)$. That is, for $\sigma^2 > 0$ the noisy information y about f is distributed according to the measure

$$\pi_f(B) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \int_B e^{-(y-f)^2/(2\sigma^2)} \, dy,$$

while for $\sigma^2 = 0$ we have $\pi_f(B) = 0$ if $f \notin B$, and $\pi_f(B) = 1$ otherwise. Hence, for $\sigma^2 = 0$ information is exact, while for $\sigma^2 > 0$ it is noisy.

Example 3.2 Suppose we wish to approximate the value of the integral $S(f) = \int_0^1 f(t) \, dt$ of a continuous function $f : [0, 1] \rightarrow \mathbb{R}$. Information is given by independent noisy observations of f at n points. That is, in the ith observation we obtain $y_i = f(t_i) + x_i$ where the noise $x_i \sim \mathcal{N}(0, \sigma^2)$, $1 \le i \le n$, and $\sigma^2 > 0$. This corresponds to $Y = \mathbb{R}^n$ and $\mathbb{N}(f) = \pi_f$ where

$$\pi_f(B) \;=\; (2\,\pi\,\sigma^2)^{-n/2} \int_{\mathbb{R}^n} \exp\left\{ -\frac{1}{2\,\sigma^2} \sum_{i=1}^n (y_i - f(t_i))^2 \right\} dy_1\, dy_2 \ldots dy_n.$$

As μ we take the classical *Wiener measure*, $\mu = w$. We recall that the Wiener measure is defined on the σ–field of Borel sets of the space

$$F \;=\; \{\, f : [0,1] \to \mathbb{R} \mid \quad f\text{–continuous},\; f(0) = 0 \,\},$$

with the supremum norm, $\|f\| = \sup_{x \in [0,1]} |f(x)|$. It is uniquely determined by the following condition. Let $m \ge 1$ and let B be a Borel set of $\mathbb{R}^m$. Let $B^{t_1 \ldots t_m} = \{\, f \in F \mid (\, f(t_1), \ldots, f(t_m)\,) \in B \,\}$ where $0 < t_1 < t_2 < \cdots < t_n \le 1$. Then

$$\begin{aligned} w(B^{t_1 \ldots t_m}) \;&=\; \{\, (2\pi)^n t_1 (t_2 - t_1) \ldots (t_n - t_{n-1}) \,\}^{-1/2} \\ &\qquad \int_B \exp\left\{ -\frac{1}{2} \left(\frac{x_1^2}{t_1} + \frac{(x_2 - x_1)^2}{t_2 - t_1} + \cdots + \frac{(x_n - x_{n-1})^2}{t_n - t_{n-1}} \right) \right\} \\ &\qquad\qquad dx_1\, dx_2 \ldots dx_n. \quad \Box \end{aligned}$$

Let $\mathbb{N} : F \to P_Y$ be a given information operator. For an algorithm $\varphi : Y \to G$, we define its error as the square root of the average performance of the difference $\|S(f) - \varphi(y)\|^2$, over all $f \in F$ and $y \in Y$. Hence, the *average case error* of φ is given as

$$\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi) \;=\; \sqrt{\int_F \int_Y \|S(f) - \varphi(y)\|^2 \, \pi_f(dy)\, \mu(df)}\,.$$

In order that the error be well defined, we consider only such algorithms φ that the mapping $y \to \varphi(y)$ is measurable with respect to the a priori distribution μ_1 on Y. This distribution is given as

$$\mu_1(B) \;=\; \int_F \pi_f(B)\, \mu(df), \qquad \forall B\text{–measurable set of } Y.$$

Note that we can equivalently say that the error is taken with respect to the a priori distribution $\tilde{\mu}$ of elements (f, y) in the product space $F \times Y$,

$$\tilde{\mu}(B) \;=\; \int_F \pi_f(B_f)\, \mu(df), \quad \forall\, B\text{–measurable set of } F \times Y,$$

where $B_f = \{ y \in Y \mid (f,y) \in B \}$.

Let $\mathbb{N}$ be a given information operator. Our first aim will be to minimize the error $\mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi)$ over all algorithms φ. As usually, an algorithm φ_{opt} for which

$$\mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi_{\mathrm{opt}}) \;=\; \inf_{\varphi} \mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi).$$

is called *optimal.*

We assume that there exists a unique (up to a set of μ_1–measure zero) family $\{\mu_2(\cdot|y)\}_{y\in Y}$ of probability measures that satisfy the following conditions:

(i) $\mu_2(\cdot|y)$ are probability measures on the σ–field of F, $\forall y$ a.e.,

(ii) the maps $y \to \mu_2(B|y)$ are μ_1–measurable for all $y \in Y$, and

(iii) $\tilde\mu(B) = \int_Y \mu_2(B_y|y)\,\mu_1(dy)$, $\forall B$, $B_y = \{ f \in F \mid (f,y) \in B \}$.

Such a family is called a *regular conditional probability distribution.* It exists under some mild assumptions, e.g., if F is a separable Banach space and $Y = \mathbb{R}^n$; see NR 3.3. We interpret $\mu_2(\cdot|y)$ as the a posteriori (or conditional) distribution on F, after information y has been observed.

The most important for us will be the property (iii). It says that the measure $\tilde\mu$ can be equivalently defined by the right hand side of (iii). Hence, the error of an algorithm φ that uses information $\mathbb{N}$ can be rewritten as

$$\mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi) \;=\; \sqrt{\int_Y \int_F \|S(f)-\varphi(y)\|^2\,\mu_2(df|y)\,\mu_1(dy)}\,. \tag{3.1}$$

For a probability measure ω on G, let

$$r(\omega) \;=\; \inf_{a\in G} \sqrt{\int_G \|g-a\|^2\,\omega(dg)}\,.$$

We call $r(\omega)$ a *radius* of the measure ω. An element $g_\omega \in G$ is a *center* of ω iff $r(\omega) = \sqrt{\int_G \|g-g_\omega\|^2\omega(dg)}$.

Example 3.3 Suppose that the measure ω is centrosymmetric. That is, there exists an element $g^* \in G$ such that for any measurable set $B \subset G$

it holds $\omega(B) = \omega(\{ 2\,g^* - g \mid g \in B\})$. Then g^* is the center of ω and $r(\omega) = \sqrt{\int_G \|g - g^*\|^2\,\omega(dg)}$. Indeed, since

$$\|x+y\|^2 + \|x-y\|^2 \;\geq\; \frac{1}{2}\,(\,\|x+y\| + \|x-y\|\,)^2 \;\geq\; 2\,\|x\|^2,$$

for any $a \in G$ we have

$$\begin{aligned} \int_G \|g-a\|^2\,\omega(dg) \;&=\; \int_G \|2g^* - g - a\|^2\,\omega(dg) \\ =\;& \frac{1}{2}\int_G \|(g^*-p) + (g^*-a)\|^2 + \|(g-g^*) - (g^*-a)\|^2 \\ \geq\;& \int_G \|g-g^*\|^2\,\omega(dg). \quad \Box \end{aligned}$$

For $y \in Y$, define the measures $\nu_2(\cdot|y) = \mu_2(S^{-1}(\cdot)|y)$. That is, $\nu_2(\cdot|y)$ is the a posteriori distribution of the elements $S(f)$ after information y has been observed. Assuming the mapping $y \to r(\nu_2(\cdot|y)\,)$ is μ_1–measurable, an *(average) radius of information* $\mathbb{N}$ is given as

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; \sqrt{\int_Y (\,r(\nu_2(\cdot|y))\,)^2\,\mu_1(dy)}\,.$$

Hence, $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$ is the average radius of the conditional distributions in G.

Lemma 3.1 *If the space G is separable then the function*

$$\psi(y) = \inf_{a\in G}\int_G \|a-g\|^2\,\nu_2(dg|y), \qquad y \in Y,$$

is μ_1–measurable.

Proof It suffices to show that the set

$$B \;=\; \{\,y \in Y \mid \quad \psi(y) \geq a\,\}$$

is μ_1–measurable for any $a \in \mathbb{R}$. Let $\psi(x,y) = \int_G \|x-g\|^2\,\nu_2(dg|y)$, $x \in G$, $y \in Y$. Then ψ is continuous with respect to x and measurable with respect to y, and $\psi(y) = \inf_{x\in G}\psi(x,y)$. Choosing a countable and dense in G set A, we obtain

$$\begin{aligned} B \;&=\; \{\,y \in Y \mid \quad \forall x \in G,\ \psi(x,y) \geq a\,\} \\ &=\; \{\,y \in Y \mid \quad \forall x \in A,\ \psi(x,y) \geq a\,\} \\ &=\; \bigcap_{x\in A}\{\,y \in Y \mid \quad \psi(x,y) \geq a\,\}. \end{aligned}$$

Hence, $B(a)$ is a countable intersection of measurable sets. This implies that B is also measurable. □

From now on we assume that the space G is separable. As we have just shown, separability of G makes the radius of information well defined. We are ready to show the main result of this section.

Theorem 3.1 *For any information operator $\mathbb{N}$ we have*

$$\inf_{\varphi} \mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi) \;=\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}).$$

If $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) < +\infty$ *then a necessary and sufficient condition for existence of the optimal algorithm is that for all $y \in Y$ a.e., there exists a center g_y of the measure $\nu_2(\cdot|y)$. In particular, the algorithm*

$$\varphi_{\mathrm{ctr}}(y) \;=\; g_y$$

is optimal.

Proof We can assume that $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) < +\infty$. Then the set $A = \{y \in Y \mid r(\nu_2(\cdot|y)) = +\infty\}$ is of μ_1–measure zero. Let $\psi(x,y)$ be as in the proof of Lemma 3.1. We already mentioned that ψ is continuous with respect to x. We also have $\sup_{x\in G}\psi(x,y) = +\infty$. Indeed, there is $t > 0$ such that the ball $B_t = \{\, g \in G \mid \|g\| \le t\,\}$ has positive $\nu_2(\cdot|y)$–measure. Then, for $x \in G$ such that $\|x\| > t$, we have

$$\psi(x,y) \;\ge\; \int_{B_t} \|x-g\|^2\,\nu_2(dg|y) \;\ge\; \nu_2(B_t|y)\cdot(\,\|x\| - t\,)^2,$$

and consequently $\psi(x,y) \to +\infty$ as $x \to +\infty$.

Thus, for fixed $y \in Y \setminus A$ the function $\psi(x,y)$ assumes all values from the interval $(\,r(\nu_2(\cdot|y)),\, +\infty)$. Hence, for any $\varepsilon > 0$ we can find an element $a_y \in G$ such that

$$\psi(a_y, y) \;=\; \int_G \|g - a_y\|^2\,\nu_2(dg|y) \;=\; (\,r(\,\nu_2(\cdot|y)\,)\,)^2 \,+\, \varepsilon^2. \tag{3.2}$$

We now define $\varphi_\varepsilon(y) = a_y$ for $y \in Y \setminus A$, and $\varphi_\varepsilon(y) = 0$ for $y \in A$. Then the algorithm φ_ε is μ_1–measurable and due to (3.1) and (3.2) we have

$$\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_\varepsilon) \;=\; \sqrt{\int_Y (\,r(\nu_2(\cdot|y)\,)^2\,\mu_1(dy) \,+\, \varepsilon^2} \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \,+\, \varepsilon.$$

On the other hand, for an arbitrary algorithm φ we have

$$\begin{aligned}
(\mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi))^2 &= \int_Y\int_F \|S(f)-\varphi(y)\|^2\,\mu_2(df|y)\,\mu_1(dy)\\
&= \int_Y\int_G \|g-\varphi(y)\|^2\,\nu_2(dg|y)\,\mu_1(dy)\\
&\ge \int_Y (r(\nu_2(\cdot|y)))^2\,\mu_1(dy) \;=\; (\mathrm{rad}^{\mathrm{ave}}(\mathbb{N},\varphi))^2,
\end{aligned}$$

which proves the first part of the theorem.

Let φ be such an algorithm that $\mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$. Let $\psi_1(y) = \int_G \|g-\varphi(y)\|^2\,\nu_2(dg|y)$ and $\psi_2(y) = (r(\nu_2(\cdot|y)))^2$. Then $\psi_1(y) \ge \psi_2(y)$, $\forall y \in Y$, and $\int_Y \psi_1(y)\,\mu_1(dy) = \int_Y \psi_2(y)\,\mu_1(dy)$. It is a very known fact that this can hold if and only if $\psi_1(y) = \psi_2(y)$, for all y a.e. Since the last equality means that $\varphi(y)$ is a center of $\nu_2(\cdot|y)$, the proof is complete. □

Following the terminology of the worst case setting we can call φ_{ctr} a central algorithm. We see that unlike in the worst case, in the average case setting optimal algorithm may differ from the central algorithm only on a set of μ_1–measure zero.

In some cases, the optimal algorithms turn out to be mean elements of conditional distributions. Recall that m_ω is the mean element of a measure ω defined on a separable Banach space G iff for any continuous linear functional $L: G \to \mathbb{R}$ it holds $\int_G L(g)\,\omega(dg) = L(m_\omega)$. We also recall that ν is the a priori distribution of $S(f) \in G$, $\nu = \mu S^{-1}$.

Lemma 3.2 *Let G be a separable Hilbert space and let $m(y)$ be the mean elements of the measures $\nu_2(\cdot|y)$, $y \in Y$. Then the unique (up to a set of μ_1–measure zero) central algorithm is $\varphi_{\mathrm{ctr}}(y) = m(y)$ and*

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; \mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi_{\mathrm{ctr}}) \;=\; \sqrt{\int_G \|g\|^2\,\nu(dg) \;-\; \int_Y \|m(y)\|^2\,\mu_1(dy)}\;.$$

Proof For any $y \in Y$ and $a \in G$ we have

$$\begin{aligned}
\int_G \|g-a\|^2\,\nu_2(dg|y) &= \|a\|^2 \;-\; 2\,\langle a, m(y)\rangle \;+\; \int_G \|g\|^2\,\nu_2(dg|y)\\
&= \|a-m(y)\|^2 \;+\; \int_G \|g\|^2\,\nu_2(dg|y) \;-\; \|m(y)\|^2.
\end{aligned}$$

The minimum of this is attained only at $a = m(y)$. Hence, $\varphi_{\mathrm{opt}}(y) = m(y)$ $\forall y$ a.e., and

$$\begin{aligned}(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}))^2 &= (\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{opt}}))^2 \\ &= \int_Y \int_G \|g\|^2 \, \nu_2(dg|y) \, \mu_1(dy) \;-\; \int_Y \|m(y)\|^2 \, \mu_1(dy).\end{aligned}$$

To complete the proof, observe that $\int_G \|g\|^2 \nu_2(dg|y) = \int_F \|S(f)\|^2 \mu_2(df|y)$, and consequently

$$\int_Y \int_G \|g\|^2 \, \nu_2(dg|y) \, \mu_1(dy) \;=\; \int_F \|S(f)\|^2 \, \mu(df) \;=\; \int_G \|g\|^2 \, \nu(dg). \quad \Box$$

Notes and Remarks

NR 3.1 Modulo some details, main results of this section have been adopted from Traub *et al.* [Sect.2,3 of Chap.6][108] (see also Wasilkowski [117]), where exact information is considered.

NR 3.2 We assume that the algorithm is a measurable mapping. One can allow arbitrary algorithms and define the error $\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi)$ as the upper integral, as in the papers cited in NR 3.1 (see also Novak [63] where even nonmeasurable S and N are allowed). As it will turn out, for problems considered in this monograph optimal algorithms in both cases are the same.

NR 3.3 A general theorem on existence of the regular conditional probability distribution reads as follows. Let X and Y be two separable Banach spaces, and let ω be a probability measure on Borel sets of X. Let $\psi : X \to Y$ be a measurable mapping and $\omega_1 = \omega\psi^{-1}$. Then there exists a family of probability measures $\{\omega_2(\cdot|y)\}_{y \in Y}$ such that:

(i) $\omega_2(\psi^{-1}(y)|y) = 1$, $\forall y$ a.e.,

(ii) for any Borel set B the mapping $y \to \omega_2(B|y)$ is measurable, and

(iii) $\omega(B) = \int_Y \omega_2(B|y) \, \omega_1(dy)$.

Moreover, any other family satisfying (i)–(iii) may differ from $\{\mu_2(\cdot|y)\}_{y \in Y}$ only on a set of μ_1–measure zero. For a proof, see Parthasarathy [71] or Varadarajan [114].

Observe that that theorem tells about decomposition of the measure ω with respect to the "exact" mapping ψ. The "noisy" version can be derived as follows. We set $X = F \times Y$, $\omega = \tilde{\mu}$, and $\psi(f, y) = y$, $\forall f \in F$, $\forall y \in Y$. Then $\omega\psi^{-1} = \mu_1$. Hence, there exists a family $\{\tilde{\mu}_2(\cdot|y)\}_{y \in Y}$ of measures defined on $F \times Y$, such that $\tilde{\mu}_2(\cdot|y)$ is concentrated on $F \times \{y\}$ a.e., the maps $y \to \tilde{\mu}_2(B|y)$ are measurable and

$\tilde{\mu}(B) = \int_Y \tilde{\mu}_2(B|y)\mu_1(dy)$. Letting $\mu_2(\cdot|y) = \tilde{\mu}_2(\cdot \times \{y\}|y)$, $\forall y \in Y$, we obtain that $\mu_2(\cdot|y)$ are concentrated on F a.e., the maps $\mu_2(B|\cdot)$ are measurable and

$$\tilde{\mu}(B) \;=\; \int_Y \tilde{\mu}_2(B|y)\,\mu_1(dy) \;=\; \int_Y \mu_2(B_y|y)\,\mu_1(dy),$$

as claimed. We also note that if F is a Banach space and $Y = \mathbb{R}^n$, then $F \times Y$ is also a Banach space and the regular conditional distribution exists.

NR 3.4 In the exact information case and linear information N, the radius of N is closely related to average widths, see e.g., Magaril–Il'yaev [51], Maiorov [53] [54], Sun and Wang [104].

Exercises

E 3.1 Give an example of a measure ω for which
1. The center does not exist.
2. The center is not unique.

E 3.2 A diameter of a measure ω on G is defined as

$$d(\omega) \;=\; \sqrt{\int_G\int_G \|g_1 - g_2\|^2\,\omega(dg)\,\omega(dg)}\,.$$

Consequently, a diameter of information $\mathbb{N}$ is given as

$$\mathrm{diam}^{\mathrm{ave}}(\mathbb{N}) \;=\; \sqrt{\int_Y (d(\mu_2(\cdot|y)\,)^2\,\mu_1(dy)}\;.$$

Show that $r(\omega) \le d(\omega) \le 2\cdot r(\omega)$ and $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \le \mathrm{diam}^{\mathrm{ave}}(\mathbb{N}) \le 2\cdot \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$.

E 3.3 Let the space G of the previous exercise be a separable Hilbert space. Show that then $d(\omega) = \sqrt{2}\cdot r(\omega)$ and $\mathrm{diam}^{\mathrm{ave}}(\mathbb{N}) = \sqrt{2}\cdot \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$.

E 3.4 Let $F = \mathbb{R}^m$ and let μ be the weighted Lebesgue measure,

$$\mu(A) \;=\; \int_A \alpha(f)\,d_m f,$$

for some positive $\alpha : \mathbb{R}^m \to \mathbb{R}_+$ such that $\int_{\mathbb{R}^m} \alpha(f)\,d_m f = 1$, where d_m is the m–dimensional Lebesgue measure. Consider the information operator $\mathbb{N}$ with $Y = \mathbb{R}^n$, given as $\mathbb{N}(f) = \pi_f$,

$$\pi_f(B) \;=\; \int_B \beta(y - N(f))\,d_n y,$$

where $N : \mathbb{R}^m \to \mathbb{R}^n$, $\beta : \mathbb{R}^n \to \mathbb{R}_+$ and $\int_{\mathbb{R}^n} \beta(y)\,d_n y = 1$. Show that in this case

$$\mu_1(B) \;=\; \int_B \gamma(y)\,d_n y \qquad \text{and} \qquad \mu_2(A|y) \;=\; \frac{1}{\gamma(y)}\int_A \alpha(f)\,\beta(y - N(f))\,d_m f,$$

where $\gamma(y) = \int_{\mathbb{R}^m} \alpha(f)\beta(y - N(f))\,d_m f$, $\forall y \in Y$.

E 3.5 Let the solution operator $S : F \to G$, measure μ on F and information $\mathbb{N}(f) = \pi_f$ with $Y = \mathbb{R}^n$ be given. Define the space $\tilde{F} = F \times Y$, solution operator $\tilde{S} : \tilde{F} \to G$, measure $\tilde{\mu}$ on $\tilde{F}$ and exact information operator $\tilde{N} : F \to Y$ as

$$\begin{aligned} \tilde{S}(f,y) &= S(f), \\ \tilde{\mu}(B) &= \int_F \pi_f(B_f)\, \mu(df), \\ \tilde{N}(f,y) &= y. \end{aligned}$$

Show that for any algorithm $\varphi : Y \to G$ we have

$$\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi; S, \mu) \;=\; \tilde{\mathrm{e}}^{\mathrm{ave}}(\tilde{N}, \varphi; \tilde{S}, \tilde{\mu}).$$

(The second quantity stands for the average error of φ with respect to $\tilde{\mu}$, for approximating $\tilde{S}(f,y)$ based on exact information $y = \tilde{N}(f,y)$.)

3.3 Gaussian measures on Banach spaces

In our study a crucial role will play Gaussian measures defined on Banach spaces. In this section, we recall what a Gaussian measure is and cite these properties of Gaussian measures that will be needed later.

3.3.1 Basic properties

Assume first that F is a finite dimensional space, $F = \mathbb{R}^d$, $d < +\infty$. A *Gaussian measure* μ on $\mathbb{R}^d$ is uniquely defined by its *mean element* $m \in \mathbb{R}^d$ and *correlation operator* (matrix) $\Sigma : \mathbb{R}^d \to \mathbb{R}^d$ which is symmetric and nonnegative definite, $\Sigma = \Sigma^* \geq 0$. If $m = 0$ and Σ is positive definite, $\Sigma > 0$, then

$$\mu(B) \;=\; \frac{1}{(2\,\pi)^{d/2}(\det \Sigma)^{1/2}} \int_B \exp\left\{-\frac{1}{2}\langle \Sigma^{-1} f, f\,\rangle_2\right\} df. \qquad (3.3)$$

(Here df stands for the Lebesgue measure on $\mathbb{R}^d$). In the case $m \neq 0$ and/or singular Σ, the Gaussian measure μ is concentrated on $m + X_1$ where $X_1 = \Sigma(X)$ and given as follows. Let $\Sigma_1 : X_1 \to X_1$, $\Sigma_1(x) = \Sigma(x)\ \forall x \in X_1$, and let $d_1 = \dim X_1$. Then, for any $B = m + B_1$ where B_1 is a Borel subset of X_1, the measure $\mu(B_1)$ is given by the right hand side of (3.3) with Σ, d and B replaced by Σ_1, d_1 and B_1, respectively, and with the Lebesgue measure df in X_1.

If $m = 0$ and $\Sigma = I$ is the identity then μ is called the *standard d–dimensional* Gaussian distribution.

Let μ be a Gaussian measure on $\mathbb{R}^d$. Then for any $x, x_1, x_2 \in \mathbb{R}^d$ we have $\int_{\mathbb{R}^n} \langle x, f\rangle_2 \, \mu(df) = \langle x, m\rangle_2$ and

$$\int_{\mathbb{R}^d} \langle x_1, f-m\rangle_2 \langle x_2, f-m\rangle_2 \, \mu(df) \;=\; \langle \Sigma x_1, x_2\rangle_2 \;=\; \langle \Sigma x_2, x_1\rangle_2 .$$

Consider now a more general case when F be a separable Banach space. A Borel measure μ on F is Gaussian iff for any n and any continuous linear mapping $N : F \to \mathbb{R}^n$, the measure $\mu_N = \mu N^{-1}$ given as

$$\mu(N^{-1}(B)\,) \;=\; \mu\{\, f \in F \mid \quad N(f) \in B\,\}, \quad \forall B\text{–Borel set of } \mathbb{R}^n,$$

is Gaussian.

As in the finite dimensional case, any Gaussian measure μ defined on a separable Banach space F is determined by its mean element $m_\mu \in F$ and correlation operator $C_\mu : F^* \to F$. [1] They are defined as

$$L(m_\mu) \;=\; \int_F L(f)\, \mu(df), \qquad \forall\, L \in F^*,$$

and

$$L_1(C_\mu L_2) \;=\; \int_F L_1(f - m_\mu) L_2(f - m_\mu)\, \mu(df), \qquad \forall\, L_1, L_2 \in F^*.$$

That is, for any mapping $N(f) = [L_1(f), \ldots, L_n(f)]$ where $L_i \in F^*$, $1 \le i \le n$, the Gaussian measure μN^{-1} has mean element $m = N(m_\mu)$ and correlation matrix $\Sigma = \{L_i(C_\mu L_j)\}_{i,j=1}^n$.

The correlation operator is always symmetric, $L_1(C_\mu L_2) = L_2(C_\mu L_1)$, and nonnegative definite, $L(C_\mu L) \ge 0$. It is positive definite, i.e., $L(C_\mu L) > 0 \;\forall L \ne 0$, iff μ has full support, $\operatorname{supp} \mu = F$. In general, μ is concentrated on the hyperplane $m_\mu + \overline{C_\mu(F^*)}$.

Suppose that F is a separable Hilbert space. Then $C_\mu : F^* = F \to F$ is the correlation operator of a Gaussian measure on F iff it is symmetric, nonnegative definite and has a finite trace, i.e.,

$$\operatorname{trace}(C_\mu) \;=\; \int_F \|f\|^2 \, \mu(df) \;=\; \sum_{i=1}^{\infty} \langle C_\mu \eta_i, \eta_i\rangle \;<\; +\infty,$$

[1] For $F = \mathbb{R}^d$ or, more generally, for F being a Hilbert space we have $F^* = F$. Then C_μ can be considered as an operator in F, $C_\mu : F \to F$.

where η_i, $i \geq 1$, is a complete orthonormal system in F.

The complete characterization of correlation operators of Gaussian measures on Banach spaces is not known. However, in this case we have the following fact. Let C_μ be the correlation operator of a Gaussian measure on F. Let $a \in F$ and let $C' : F^* \to F$ satisfy the following conditions: C' is symmetric, $L_1(C'L_2) = L_2(C'L_1)$, and $0 \leq L(C'L) \leq L(C_\mu L)$, $\forall L_1, L_2, L \in F^*$. Then there exists a (unique) Gaussian measure on F with mean element a and correlation operator C'.

The *characteristic functional* of a measure μ is given as $\psi_\mu : F^* \to \mathbf{C}$,

$$\psi_\mu(L) \;=\; \int_F e^{i\,L(f)}\,\mu(df) \qquad\qquad (i = \sqrt{-1}).$$

Any measure is uniquely determined by its characteristic functional. If μ is Gaussian with mean m_μ and correlation operator C_μ then

$$\psi_\mu(L) \;=\; \exp\left\{ i\,L(m_\mu) - \frac{1}{2}\,L(C_\mu L) \right\}.$$

The correlation operator C_μ generates the μ-semi-inner product on the space F^*. It is defined as $\langle\cdot,\cdot\rangle_\mu : F^* \times F^* \to \mathbb{R}$,

$$\begin{aligned} \langle L_1, L_2\rangle_\mu &= L_1(C_\mu L_2) \;=\; L_2(C_\mu L_1) \\ &= \int_F L_1(f)\,L_2(f)\,\mu(df), \qquad L_1, L_2 \in F^*. \end{aligned}$$

We denote by $\|\cdot\|_\mu$ the corresponding semi-norm, $\|L\|_\mu = \sqrt{\langle L, L\rangle_\mu}$. If $\operatorname{supp}\mu = F$ then C_μ is one-to-one and $\langle\cdot,\cdot\rangle_\mu$ is an inner product and $\|\cdot\|_\mu$ is a norm. The space F^* with the norm $\|\cdot\|_\mu$ is complete only if $\dim F < +\infty$. μ–orthogonality in F^* means orthogonality with respect to $\langle\cdot,\cdot\rangle_\mu$.

3.3.2 Gaussian measures as abstract Wiener spaces

We noticed that any Gaussian measure μ is determined by its mean and correlation operator. Sometimes it is convenient to define μ in another way.

Let H be a separable Hilbert space. For any (cylindrical) set $B \subset H$ of the form $B = \{\, g \in H \mid P(g) \in A \,\}$, where P is the orthogonal projection in H onto a finite dimensional subspace of H and A is a Borel set in $P(H)$, we let

$$\mu'(B) \;=\; \frac{1}{(\sqrt{2\pi})^n} \int_A e^{-\|g\|_H^2/2}\,dg \tag{3.4}$$

where $n = \dim P(H)$ and dg is the Lebesgue measure on $P(H)$. That is, μ' is the standard *weak* distribution on the algebra of cylindrical sets. Note that μ' is an additive measure but, in case $\dim H = +\infty$, it cannot be extended to a σ–additive measure on the Borel σ-field of F. Let $\|\cdot\|_F$ be another norm on H which is weaker than the original norm $\|\cdot\|_H$, i.e., $\|\cdot\|_F \le K\|\cdot\|_H$ for some constant $K > 0$. Let F be the closure of H with respect to $\|\cdot\|_F$. It turns out that if $\|\cdot\|_F$ possesses some additional properties (it is in some sense measurable, see NR 3.8), then there exists a unique σ–additive measure μ defined on the Borel sets of F, such that the following holds. For any n and continuous linear functionals $L_i \in F^*$, $1 \le i \le n$, we have

$$\begin{aligned} &\mu(\{\, f \in F \mid \quad (L_1(f), \dots, L_n(f)\,) \in B\,\}\,) \\ = \;& \mu'(\,\{g \in H \mid \quad (\langle g_{L_1}, g\rangle_H, \dots, \langle g_{L_n}, g\rangle_H\,) \in B\,\}\,), \end{aligned}$$

for all Borel sets $B \subset \mathbb{R}^n$. Here g_L is the representer of L in H, i.e., $L(f) = \langle g_L, f\rangle_H$ for $f \in H$ or, in other words, $g_L = e^*L$ where $e : H \to F$ is the continuous embedding. The pair $\{H, F\}$ is called an *abstract Wiener space.*

Observe that for the measure μ constructed as above we have

$$\int_F L(f)\,\mu(df) \;=\; (2\pi\|g_L\|_H^2)^{-1/2}\int_R x\,\exp\{-x^2/(2\|g_L\|_H^2)\}\,dx \;=\; 0$$

and

$$L_1(C_\mu L_2) \;=\; \int_F L_1(f)L_2(f)\,\mu(df) \;=\; \langle g_{L_1}, g_{L_2}\rangle_H$$

$\forall L, L_1, L_2 \in F^*$. Hence, μ is the zero mean Gaussian measure with positive definite correlation operator $C_\mu(L) = g_L$. Moreover, $C_\mu(F^*) \subset H \subset \overline{C_\mu(F^*)} = F$.

Such an extension of μ' to a Gaussian measure μ always exists and is not unique. For instance, we can take $\|g\|_F = \sqrt{\langle Ag, g\rangle_H}$ where $A : H \to H$ is an arbitrary symmetric, positive definite operator with finite trace. Then the resulting space F is a separable Hilbert space and the correlation operator of μ is given by the continuous extension of A to the operator $A : F \to F$.

On the other hand, for any separable Banach space F equipped with a zero mean Gaussian measure μ, there exists a unique separable Hilbert space H, such that $C_\mu(F^*) \subset H \subset \overline{C_\mu(F^*)}$ and $\{H, F_1\}$ with $F_1 = \text{supp}\,\mu = \overline{C_\mu(F^*)}$ is an abstract Wiener space. The space H is given as follows. Let $H_0 = C_\mu(F^*)$. For $f_2, f_2 \in H_0$, we define $\langle f_1, f_2\rangle_H = \langle L_1, L_2\rangle_\mu$ where L_i are arbitrary functionals satisfying $C_\mu L_i = f_i$, $i = 1, 2$. Since $\langle f_1, f_2\rangle_H$ does not depend on the choice of L_i, it is a well defined inner product on H_0. Then H is the closure of H_0 with respect to the norm $\|\cdot\|_H = \sqrt{\langle\cdot,\cdot\rangle_H}$. Clearly, (3.4) also holds.

Thus any zero mean Gaussian measure on a separable Banach space can be viewed as an abstract Wiener space. And, of course, vice versa.

In the end, consider the case when H in an abstract Wiener space $\{H, F\}$ is an r.k.h.s. with r.k. $R : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ (see Section 2.6.4). Suppose that the functionals $L_t(f) = f(t)$, $f \in F$, are continuous in F for all $t \in \mathcal{T}$. Then we always have

$$L_t(C_\mu L_s) \;=\; \langle R_t, R_s\rangle_H \;=\; R(t,s) \qquad \forall s,t \in \mathcal{T},$$

no matter what norm $\|\cdot\|_F$ has been used. Therefore the reproducing kernel R is also called the *covariance kernel.* μ is determined uniquely by its covariance kernel.

Example 3.4 Let $r \ge 0$. Let $H = W^0_{r+1}$ be the reproducing kernel Hilbert space of Example 2.12 with $(a,b) = (0,1)$. That is,

$$\begin{aligned} W^0_{r+1} \;=\; \{\, f : [0,1] \to \mathbb{R} \;\;|\;\; & f^{(r)}\text{–abs. cont.}, \\ & f^{(i)} = 0,\ 0 \le i \le r,\ f^{(r+1)} \in \mathcal{L}_2([0,1]) \,\} \end{aligned}$$

Let $\|f\|_{C_r} = \sup_{0\le x\le 1} |f^{(r)}(x)|$. Then $\|f\|_{C_r} \le \|f\|_{W_{r+1}}$. The space W^0_{r+1} can be completed with respect to the norm $\|\cdot\|_{C_r}$. The resulting space is a separable Banach space,

$$C^0_r \;=\; \{\, f : [0,1] \to \mathbb{R} \mid \quad f^{(r)}\text{–continuous},\ f^{(i)}(0) = 0,\ 0 \le i \le r \,\}.$$

Then $\{W^0_{r+1}, C^0_r\}$ is an abstract Wiener space. That is, μ constructed based on the weak distribution on W^0_r is a well defined Gaussian measure on the Borel sets of C^0_r. In the case $r = 0$ we obtain the *classical Wiener measure* w of Example 3.2, where the covariance kernel $R(s,t) = \int_0^1 G_0(s,u)G_0(t,u)\,du = \min\{s,t\}$. For arbitrary r, μ is called the *r–fold Wiener measure* and denoted by w_r. □

The name for w_r is justified by the following property. For a Borel set $B \subset C^0_r$, let $D^r(B) = \{\, f^{(r)} \mid f \in B \,\}$. Then $w_r = wD^r$. To see this, observe that $\tilde{w}_r = wD^r$ is a well defined Borel measure on C_r^0. Since $w = w_0$ is uniquely determined by its covariance kernel $R(s,t) = \min\{s,t\}$, $\tilde{w}_r$ is uniquely determined by the equation $\int_{C_r^0} f^{(r)}(s) f^{(r)}(t)\, w_r(df) = \min\{s,t\}$, $s,t \in [0,1]$. On the other hand, the representer of the functional $f^{(r)}(t)$ in W_{r+1}^0 is given as $G_r(t,\cdot)$. Hence,

$$\begin{aligned}\int_{C_r^0} f^{(r)}(s) f^{(r)}(t)\, w_r(df) &= \int_0^1 G_r^{(r)}(s,u)\, G_r^{(r)}(t,u)\, du \\ &= \int_0^1 G_0(s,u)\, G_0(t,u)\, du \;=\; \min\{s,t\},\end{aligned}$$

and $\tilde{w}_r = w_r$.

Notes and Remarks

NR 3.5 For references about Gaussian measures on separable Hilbert and Banach spaces see, e.g., Kuo [44], Parthasarathy [71], Skorohod [95], Vakhania [112], Vakhania *et al.* [113].

NR 3.6 Let F^{all} be the space of all functions $f : [a,b] \to \mathbb{R}$. Then any function $f(t)$, $a \le t \le b$, can be viewed as a realization of the stochastic process corresponding to a covariance kernel $R(s,t)$, $a \le s,t \le b$. For instance, the process corresponding to the kernel $R(s,t) = \min\{s,t\}$ is called a *Brownian motion.* The reader interested in stochastic processes is referred to, e.g., Gikhman and Skorohod [16].

NR 3.7 Some interesting things about Gaussian measures on the space $C([0,1])$ can be found in Parthasarathy [71]. In particular, he gives a sufficient condition for covariance kernel $R : [0,1]^2 \to \mathbb{R}$ to determine a unique probability measure μ on $C([0,1])$. Namely, it suffices that there exist constants α, β, $K > 0$, such that for all $t_1, t_2 \in [0,1]$

$$\int_{\mathbb{R}^2} |x_1 - x_2|^\alpha \, \mu_{t_1 t_2}(dx) \;\le\; K\, |t_1 - t_2|^{1+\beta}.$$

Here $x = (x_1, x_2)$ and $\mu_{t_1 t_2}$ is the Gaussian measure in $\mathbb{R}^2$ with correlation matrix $\{R(t_i,t_j)\}_{i,j=1}^2$.

NR 3.8 A norm $\|\cdot\|_F$ in a Hilbert space H is called *measurable* iff for any $\varepsilon > 0$ there exists a finite dimensional orthogonal projection P_0, such that for any finite dimensional orthogonal projection $P \perp P_0$ it holds

$$\mu'(\{\, g \in H \mid \; \|Pg\|_F > \varepsilon \,\}) \;\le\; \varepsilon.$$

If $\|\cdot\|_F$ is measurable then the weak measure μ' can be extended to a measure μ defined on the closure F of H with respect to $\|\cdot\|_F$.

NR 3.9 Gaussian measures as abstract Wiener spaces are studied e.g. in Kuo [44].

NR 3.10 In the case of multivariate functions, Gaussian measures may be defined based on Gaussian distributions on univariate functions. An example is provided by the *Wiener sheet measure* which is given as follows.

Let $d \ge 1$ and $r_i \ge 0$, $1 \le i \le d$. Let F be the Banach space of functions $f : [0,1]^d \to \mathbb{R}$ that are r_i times continuously differentiable with respect to the ith variable,

$$\begin{aligned} F \;=\; \mathbf{C}^{0\dots0}_{r_1\dots r_d} \;&=\; \big\{\, f:[0,1]^d \to \mathbb{R} \;\big|\; D^{r_1\cdots r_d} f\text{–cont.}, \\ &\qquad (D^{r_1\cdots r_d} f)(t) = 0,\; 0 \le i_j \le r_j,\; 1 \le j \le d, \\ &\qquad \text{when at least one } t_i \text{ is zero } \big\}, \end{aligned}$$

with the norm $\|f\| = \sup_{t\in[0,1]^d} |(D^{r_1\cdots r_d} f)(t)|$. The Wiener sheet measure on F is defined as

$$w_{r_1\dots r_d}(B) \;=\; w_{0\dots0}(D^{r_1\cdots r_d}(B)\,) \qquad \forall B\text{–Borel set in } F,$$

where $w_{0\dots0}$ is the classical Wiener measure on $\mathbf{C}^{0\dots0}_{0\dots0}$. Its covariance kernel is given as

$$R_{0\dots0}(s,t) \;=\; \int_{\mathbf{C}^{0\dots0}_{0\dots0}} f(s)f(t) \;=\; \prod_{j=1}^{d} \min\{s_j, t_j\}$$

where $s = (s_1, \dots, s_d)$, $t = (t_1, \dots, t_d)$.

It is easy to see that $w_{r_1\dots r_d}$ is the zero mean Gaussian measure with covariance kernel

$$R_{r_1\dots r_d}(s,t) \;=\; \prod_{j=1}^{d} R_{r_j}(s_j, t_j),$$

where R_{r_j} is the covariance kernel of the r_j–fold Wiener measure on $\mathbf{C}^0_{r_j}$. Hence, the associated with $w_{r_1\dots r_d}$ abstract Wiener space is $\{W^{0\dots0}_{r_1+1\dots r_d+1}, \mathbf{C}^{0\dots0}_{r_1\dots r_d}\}$ where $W^{0\dots0}_{r_1+1\dots r_d+1}$ is the r.k.h.s. defined in NR 2.21.

Anther example of a Gaussian distribution on multivariate functions is the *isotropic Wiener measure* (or the Brownian motion in Lévy's sense) which is defined on the space $\mathbf{C}([0,1]^d)$. Its mean is zero and covariance kernel is given as

$$R(s,t) \;=\; \frac{\|s\|_2 + \|t\|_2 - \|s-t\|_2}{2} \qquad s,t \in [0,1]^d,$$

see e.g. Ciesielski [9] for more details.

Exercises

E 3.6 Let H be a separable Hilbert space. Let e_i, $i \ge 1$, be a complete orthonormal system in H, and let $P_n : H \to \mathbb{R}^n$, $n \ge 1$, be defined as $P_n(x) = \{\langle x, e_i\rangle\}_{i=1}^n$. Prove that there is no such a Gaussian measure μ on H that for any n, μP_n^{-1} is ithe zero mean n–dimensional Gaussian measure with identity correlation operator.

E 3.7 The space l_2 can be treated as the space of functions $f : \{1, 2, \ldots\} \to \mathbb{R}$, such that $\|f\|^2 = \sum_{i=1}^{\infty} f^2(i) < +\infty$. Show that $R(i,j) = \lambda_i \delta_{ij}$, $i,j \geq 1$, is the covariance kernel of a Gaussian measure on l_2 iff $\sum_{i=1}^{\infty} \lambda_i < +\infty$.

E 3.8 Let H be a separable Hilbert space and let $\|\cdot\|_F$ be a norm equivalent to $\|\cdot\|_H$, i.e., $K_1\|\cdot\|_F \leq \|\cdot\|_H \leq K_2\|\cdot\|_F$ for some $0 < K_1 \leq K_2 < +\infty$. Show that $\{F, H\}$ is an abstract Wiener space if and only if $\dim H < +\infty$.

E 3.9 Let $\{F, H\}$ be an abstract Wiener space and let μ be the associated with it Gaussian measure. Show that $\mu(H) = 0$.

E 3.10 Show that the r–fold Wiener measures w_r satisfy $w_r = w_s D^{r-s}$, where D^k is the differential operator of order k, and $r \geq s \geq 0$.

E 3.11 Let R_r be the covariance kernel of the r–fold Wiener measure. Show that

$$R_r(s,t) \;=\; \int_0^s \int_0^t R_{r-1}(u_1, u_2)\, du_1\, du_2.$$

3.4 Linear problems with Gaussian measures

We start the study of linear problems with Gaussian measures. The final goal of this section is to give general formulas for the optimal algorithm and radius of information.

We assume that

- F is a separable Banach space, G is a separable Hilbert space, and the solution operator $S : F \to G$ is continuous and linear.
- The a priori distribution μ on F is a zero mean Gaussian measure.

We also assume that the information values y are distributed according to some Gaussian measure. More precisely, we assume that $Y = \mathbb{R}^n$ and there exists a continuous linear operator $N : F \to \mathbb{R}^n$,

$$N(f) \;=\; [\,L_1(f), L_2(f), \ldots, L_n(f)\,], \qquad f \in F,$$

where $L_i \in F^*$, $1 \leq i \leq n$, as well as a matrix $\Sigma : \mathbb{R}^n \to \mathbb{R}^n$, $\Sigma = \Sigma^* \geq 0$, such that

$$\mathbb{N}(f) \;=\; \mathcal{N}(\,N(f), \Sigma), \qquad \forall\, f \in F. \tag{3.5}$$

Here $\mathcal{N}(N(f),\Sigma)$ stands for the n–dimensional Gaussian (normal) distribution with mean $N(f)$ and correlation matrix (operator) Σ. In other words, information y about f is obtained by noisy observation of the value $N(f)$ of a linear mapping N, $y = N(f) + x$, and the noise x is a zero mean Gaussian random variable.

Sometimes we shall write $\mathbb{N}(f) = \mathcal{N}(N(f), \sigma^2\Sigma)$ to stress that the noise level depends also on a parameter σ^2. In particular, for $\sigma^2 = 0$ (or for $\Sigma \equiv 0$) we obtain exact information.

Information (3.5) will be called *linear with Gaussian noise.* Note that information with Gaussian noise seems to be most often used in practice.

3.4.1 Induced and conditional distributions

In this section, we give formulas for induced and conditional distributions. They are necessary to find the optimal algorithm and radius of information.

The following lemma is well known. For completeness, we provide it with a proof.

Lemma 3.3 *Let ω be a Gaussian measure on F with the mean element m_ω and correlation operator C_ω. Then the measure ωS^{-1} is also Gaussian. The mean element of ωS^{-1} is $S(m_\omega)$, and the correlation operator equals $S(C_\omega S^*)$ where $S^* : G = G^* \to F^*$ is the adjoint operator to S, i.e., $S^*(g) = \langle S(\cdot), g\rangle$.*

Proof Indeed, the characteristic functional of ωS^{-1} is given as

$$\begin{aligned}\psi_{\omega S^{-1}}(g) &= \int_G e^{i\langle x,g\rangle}\,\omega S^{-1}(dx)\\ &= \int_F e^{i\langle S(f),g\rangle}\,\omega(df) \;=\; \int_F e^{i(S^*g)(f)}\,\omega(df)\\ &= e^{i(S^*g)(m_\omega)-\frac{1}{2}(S^*g)(C_\mu(S^*g))} \;=\; e^{i\langle S(m_\omega),g\rangle-\frac{1}{2}\langle SC_\mu(S^*g),g\rangle}.\end{aligned}$$

Hence, $S(m_\omega)$ is the mean element and $S(C_\omega S^*)$ is the correlation operator of ω. □

Define the matrix

$$G_N \;=\; \{\langle L_j, L_k\rangle_\mu\}_{j,k=1}^n.$$

Clearly, G_N is symmetric and nonnegative definite. Let $Y_1 = (\Sigma + G_N)(\mathbb{R}^n)$. Then for any $y \in Y_1$ there is exactly one element $z \in Y_1$ satisfying $(\Sigma + G_N)z = y$.

We need the following simple fact.

Lemma 3.4 *For any $L \in F^*$ we have $N(C_\mu L) \in Y_1$.*

Proof Indeed, any $L \in F^*$ can be decomposed as $L = L_0 + \sum_{j=1}^n \alpha_j L_j$, where $L_0 \perp_\mu \operatorname{span}\{L_1, \ldots, L_n\}$. Then $N(C_\mu L) = G_N(\alpha)$, $\alpha = (\alpha_1, \ldots, \alpha_n)$. Since both matrices Σ and G_N are symmetric and nonnegative definite, we have $G_N(\mathbb{R}^n) \subset (\Sigma + G_N)(\mathbb{R}^n) = Y_1$, and $N(C_\mu L) \in Y_1$. □

We now show formulas for the regular conditional distribution. Recall that the distribution of information y on $\mathbb{R}^n$ is denoted by μ_1, and the conditional distribution on F with respect to y is denoted by $\mu_2(\cdot|y)$.

Theorem 3.2 *For the linear information with Gaussian noise, μ_1 is a zero mean Gaussian measure and its correlation matrix is $C_{\mu_1} = \Sigma + G_N$. Furthermore, the conditional measure $\mu_2(\cdot|y)$, $y \in Y_1$, is also Gaussian. Its mean element equals*

$$m(y) \;=\; \sum_{j=1}^n z_j \,(C_\mu L_j)$$

where $z = z(y) = (z_1, \ldots, z_n) \in Y_1$ satisfies $(\Sigma + G_N)\, z = y$. The correlation operator of $\mu_2(\cdot|y)$ is independent of y and given as

$$C_{\mu_2}(L) \;=\; C_\mu(L) \;-\; m(\,N(C_\mu L)\,), \qquad L \in F^*.$$

The lemma needs an explanation in the case when the matrix $\Sigma + G_N$ is singular. Then the measure μ_1 is concentrated on Y_1, i.e., $\mu_1(Y_1) = 1$. Hence, it suffices to know the conditional measure $\mu_2(\cdot|y)$ for $y \in Y_1$. We also note that due to Lemma 3.4, the element $m(N(C_\mu L))$ in the definition of C_{μ_2} is well defined.

Proof The characteristic functional of the measure μ_1 is given as ($a \in \mathbb{R}^n$ and $i = \sqrt{-1}$)

$$\begin{aligned}\psi_{\mu_1}(a) \;&=\; \int_{\mathbb{R}^n} e^{i\,\langle y,a\rangle_2}\,\mu_1(dy) \;=\; \int_F \int_{\mathbb{R}^n} e^{i\,\langle y,a\rangle_2}\,\pi_f(dy)\,\mu(df)\\ &=\; \int_F \exp\left\{ i\langle N(f), a\rangle_2 - \frac{1}{2}\langle \Sigma a, a\rangle_2 \right\}\,\mu(df).\end{aligned}$$

Since for the functional $L_a(\cdot) = \langle N(\cdot), a\rangle_2$ we have $L_a(C_\mu L_a) = \langle G_N a, a\rangle_2$,

$$\psi_{\mu_1}(a) \;=\; \exp\left\{ -\frac{1}{2}\langle(\Sigma + G_N)a, a\rangle_2 \right\}.$$

Hence, μ_1 is the zero mean Gaussian measure with correlation matrix $\Sigma + G_N$.

We now pass to the conditional distribution. For $y \in Y_1$, let $\mu_2'(\cdot|y)$ be the Gaussian measure on F with the mean $m'(y) = \sum_{j=1}^n z_j(C_\mu L_j)$, $(\Sigma + G_N)\, z = y$, and correlation operator $C'(\cdot) = C_\mu(\cdot) - m'(NC_\mu(\cdot))$. Observe that $\mu_2'(\cdot|y)$ are well defined Gaussian measures. Indeed, for $y \in Y_1$ we have

$$\begin{aligned} L(m'(y)) &= L\left(\sum_{j=1}^n z_j C_\mu L_j\right) = \langle (\Sigma + G_N)^{-1} y, NC_\mu L\rangle_2 \\ &= \langle y, (\Sigma + G_N)^{-1} NC_\mu L\rangle_2. \end{aligned}$$

Hence, for any $L, L' \in F^*$

$$\begin{aligned} L(C'L') &= L(C_\mu L') - L(m'(NC_\mu L')) \\ &= L(C_\mu L') - \langle NC_\mu L', (\Sigma + G_N)^{-1} NC_\mu L\rangle_2 \\ &= L'(C_\mu L) - \langle NC_\mu L, (\Sigma + G_N)^{-1} NC_\mu L'\rangle_2 \\ &= L'(C'L), \end{aligned}$$

and $0 \le L(C'L) \le L(C_\mu L)$.

We need to show that the characteristic functional of the measure $\tilde\mu$ is equal to the characteristic functional of the measure $\tilde\mu'$ defined as

$$\tilde\mu'(B) = \int_Y \mu_2'(B_y|y)\, \mu_1(dy), \quad B\text{–Borel set of } \tilde F = F \times \mathbb{R}^n.$$

To this end, let $\tilde L \in \tilde F^*$. Then there are $L \in F^*$ and $w \in \mathbb{R}^n$ such that $\tilde L(\tilde f) = L(f) + \langle y, w\rangle_2$, $\forall \tilde f = (f, y) \in \tilde F$. We have

$$\begin{aligned} \psi_{\tilde\mu'}(\tilde L) &= \int_{\mathbb{R}^n} \left(\int_F \exp\{i(L(f) + \langle y, w\rangle_2)\} \mu_2'(df|y) \right) \mu_1(dy) \\ &= \int_{\mathbb{R}^n} \exp\{i\langle y, w\rangle_2\} \left(\int_F \exp\{iL(f)\}\, \mu_2'(df|y) \right) \mu_1(dy) \\ &= \int_{\mathbb{R}^n} \exp\Big\{ i\,(\langle y, w\rangle_2 + L(m'(y))\,) \\ &\qquad - \frac{1}{2}(L(C_\mu L) - L(m'(NC_\mu L))\,) \Big\}\, \mu_1(dy). \end{aligned}$$

Recall that for $y \in Y_1$ we have $L(m'(y)) = \langle y, (\Sigma + G_N)^{-1} NC_\mu L\rangle_2$. Hence, $L(m'(NC_\mu L)) = \langle NC_\mu L, (\Sigma + G_N)^{-1} NC_\mu L\rangle_2$, and

$$\psi_{\tilde\mu'}(\tilde L) = \exp\left\{ -\frac{1}{2}(\, L(C_\mu L) - \langle NC_\mu L, (\Sigma + G_N)^{-1} NC_\mu L\rangle_2\,) \right\}$$

$$\begin{aligned}
&\int_{\mathbb{R}^n} \exp\{\, i\langle y, w + (\Sigma + G_N)^{-1} N C_\mu L\rangle_2 \,\}\, \mu_1(dy) \\
&= \exp\left\{ -\frac{1}{2} (\, L(C_\mu L) - \langle N C_\mu L, (\Sigma + G_N)^{-1} N C_\mu L\rangle_2 \,) \right\} \\
&\quad \exp\left\{ -\frac{1}{2} (\ \langle (\Sigma + G_N) w, w\rangle_2 + \langle N C_\mu L, (\Sigma + G_N) N C_\mu L\rangle_2 \right. \\
&\qquad + 2\, \langle w, N C_\mu L\rangle_2 \ )\} \\
&= \exp\left\{ -\frac{1}{2} (\ L(C_\mu L) + 2\langle w, N C_\mu L\rangle_2 + \langle (\Sigma + G_N) w, w\rangle_2 \ ) \right\}.
\end{aligned}$$

On the other hand, for the characteristic functional $\psi_{\tilde{\mu}}$ of the measure $\tilde{\mu}$ we have

$$\begin{aligned}
\psi_{\tilde{\mu}}(\tilde{L}) &= \int_F \int_{\mathbb{R}^n} \exp\{\, i(L(f) + \langle y, w\rangle_2) \,\}\, \pi_f(dy)\, \mu(df) \\
&= \int_F \exp\{iL(f)\} \left(\int_{\mathbb{R}^n} \exp\{\, i\langle y, w\rangle_2 \}\, \pi_f(dy) \right) \mu(df) \\
&= \exp\ \left\{ -\frac{1}{2} \langle (\Sigma + G_N) w, w\rangle_2 \right\} \\
&\qquad \int_F \exp\{\, i(L(f) + \langle N(f), w\rangle_2) \,\}\, \mu(df) \\
&= \exp\ \left\{ -\frac{1}{2} (\ L(C_\mu L) + 2\langle w, N C_\mu L\rangle_2 + \langle (\Sigma + G_N) w, w\rangle_2 \ ) \right\}.
\end{aligned}$$

Thus $\psi_{\tilde{\mu}} = \psi_{\tilde{\mu}'}$ which completes the proof.

3.4.2 Optimal algorithms

We are now ready to give formulas for the optimal algorithm and radius of information. They can be easily found using Lemma 3.2.

Indeed, this lemma states that φ_{opt} is determined uniquely (up to a set of y of μ_1–measure zero), and that $\varphi_{\mathrm{opt}}(y)$ $(y \in Y_1)$ is the mean element of the measure $\nu_2(\cdot|y) = \mu_2(S^{-1}(\cdot)|y)$. Using Lemma 3.3 we find that $\varphi_{\mathrm{opt}}(y) = S(m(y)\,)$ where $m(y)$ is the mean element of $\mu_2(\cdot|y)$. Due to Theorem 3.2, we have $m(y) = \sum_{j=1}^n z_j(C_\mu L_j)$ where $z = (\Sigma + G_N)^{-1} y \in Y_1$. Furthermore, for the radius of information $\mathbb{N}$ we have

$$\begin{aligned}
(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}))^2 &= (\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{opt}}))^2 \\
&= \int_G \|g\|^2\, \nu(dg) \ - \int_Y \|S(m(y)\,)\|^2\, \mu_1(dy) \\
&= \mathrm{trace}\,(S C_\mu S^*) \ - \ \mathrm{trace}\,(\,(Sm)(\Sigma + G_N)(Sm)^*\,),
\end{aligned}$$

where $Sm : Y_1 \to G$, $(Sm)(y) = S(m(y)\,)$, and $(Sm)^* : G \to Y_1$ is the adjoint operator to Sm. Observe now that $(Sm)^* = (\Sigma + G_N)^{-1} NC_\mu(S^*g)$. Indeed, for any $y \in Y_1$ and $g \in G$ we have

$$\begin{aligned}
\langle Sm(y), g\rangle &= \left\langle \sum_{j=1}^n z_j S(C_\mu L_j),\, g \right\rangle \;=\; \sum_{j=1}^n z_j \langle S(C_\mu L_j), g\rangle \\
&= \sum_{j=1}^n z_j (S^*g)(C_\mu L_j) \;=\; \sum_{j=1}^n z_j L_j(C_\mu S^* g) \\
&= \langle z, NC_\mu(S^*g)\rangle_2 \;=\; \left\langle y, (\Sigma + G_n)^{-1} NC_\mu(S^*g)\right\rangle_2 \\
&= \langle y, (Sm)^* g\rangle.
\end{aligned}$$

Thus $(Sm)(\Sigma + G_N)(Sm)^* g = Sm(NC_\mu(S^*g)\,)$, $\forall g \in G$.

We summarize this in the following theorem.

Theorem 3.3 *For the linear information $\mathbb{N}$ with Gaussian noise the optimal algorithm is linear and equals*

$$\varphi_{\mathrm{opt}}(y) \;=\; \sum_{j=1}^n z_j\, S(C_\mu L_j), \qquad y \in Y_1,$$

where $z = z(y) \in Y_1$ satisfies $(\Sigma + G_N)\, z = y$. Furthermore,

$$\begin{aligned}
\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) &= \mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{opt}}) \\
&= \sqrt{\mathrm{trace}\,(SC_\mu S^*) \;-\; \mathrm{trace}\,(\,Sm(NC_\mu S^*)\,)}\,. \quad \square
\end{aligned}$$

The above formulas are rather complicated. They can be simplified if we assume a special form of Σ and N. Namely, suppose that information consists of independent observations of n functionals which are μ–orthonormal. This corresponds to diagonal matrix Σ, $\Sigma = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$, and the assumption that $N = [L_1, \ldots, L_n]$ where $\langle L_i, L_j\rangle_\mu = \delta_{ij}$ (the Kronecker delta).

In this case, the Gram matrix G_N is the identity and $\Sigma + G_N = \mathrm{diag}\{1 + \sigma_1^2, \ldots, 1 + \sigma_n^2\}$. Hence, $S(m(y)\,) = \sum_{j=1}^n (1 + \sigma_j^2)^{-1} y_j S(C_\mu L_j)$, $y \in \mathbb{R}^n$. If we replace y above by $N(C_\mu L)$, $L \in F^*$, then

$$S(m(NC_\mu L)\,) \;=\; \sum_{j=1}^n \frac{\langle L, L_j\rangle_\mu}{1 + \sigma_j^2},$$

so that for $g \in G$ we have

$$\begin{aligned}\langle S(m(NC_\mu(S^*g))\,),\,g\rangle &= \sum_{j=1}^n \frac{\langle S^*g, L_j\rangle_\mu}{1+\sigma_j^2}\langle S(C_\mu L_j), g\rangle \\ &= \sum_{j=1}^n \frac{\langle S(C_\mu L_j), g\rangle^2}{1+\sigma_j^2}.\end{aligned}$$

Choosing an orthonormal basis $\{g_i\}_{i=1}^\infty$ in G, we obtain

$$\mathrm{trace}\,(Sm(NC_\mu S^*)\,) \;=\; \sum_{i=1}^\infty\sum_{j=1}^n \frac{\langle S(C_\mu L_j), g_i\rangle^2}{1+\sigma_j^2}$$

$$=\; \sum_{j=1}^n \frac{1}{1+\sigma_j^2}\sum_{i=1}^\infty \langle S(C_\mu L_j), g_i\rangle^2 \;=\; \sum_{j=1}^n \frac{\|S(C_\mu L_j)\|^2}{1+\sigma_j^2}\,.$$

Thus, we have the following corollary.

Corollary 3.1 *Let the functionals L_j be orthonormal, $\langle L_i, L_j\rangle_\mu = \delta_{ij}$, $1 \le i,j \le n$. If the observations of successive L_j are independent, $\Sigma =$ diag$\{\sigma_1^2,\ldots,\sigma_n^2\}$, then the optimal algorithm*

$$\varphi_{\mathrm{opt}}(y) \;=\; \sum_{j=1}^n \frac{y_j}{1+\sigma_j^2}\, S(C_\mu L_j)$$

and the radius of information

$$(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}))^2 \;=\; \mathrm{trace}\,(SC_\mu S^*) \;-\; \sum_{j=1}^n \frac{\|S(C_\mu L_j)\|^2}{1+\sigma_j^2}\,. \quad \Box$$

It turns out that the assumptions of Corollary 3.1 are not restrictive. More precisely, we now show that using some linear transformation, any linear information with Gaussian noise, $\mathbb{N}(f) = \mathcal{N}(N(f), \Sigma)$, can be translated to other information $\mathbb{M}$ which is as powerful as $\mathbb{N}$ and consists of independent observations of μ–orthonormal functionals.

Suppose first that the matrix Σ is nonsingular. Denote by $\tilde{L}_i$ the functionals which form the operator $\Sigma^{-1/2}N$, i.e., $\Sigma^{-1/2}N = [\tilde{L}_1,\ldots,\tilde{L}_n]$. Let $G = \{\langle\tilde{L}_i,\tilde{L}_j\rangle_\mu\}_{i,j=1}^n$, and let $\{q^{(i)}\}_{i=1}^n$ be the orthonormal basis of eigenvectors of the matrix G, $Gq^{(i)} = \eta_i q^{(i)}$ where $\eta_1 \ge \cdots \ge \eta_m > 0 = \eta_{m+1} =$

$\ldots = \eta_n$. Letting Q to be the (orthogonal) $n \times n$ matrix of vectors $q^{(i)}$, and D_1 to be the $m \times n$ diagonal matrix $\mathrm{diag}\{\eta_1^{-1/2}, \ldots, \eta_m^{-1/2}\}$, we define $M = D_1 Q^* \Sigma^{-1/2} N : F \to \mathbb{R}^m$.

The problem of approximating $S(f)$ from the data $y = N(f) + x$ where $x \sim \mathcal{N}(0, \Sigma)$, can be translated to the problem of approximating $S(f)$ from $y' = D_1 Q^* \Sigma^{-1/2} y = M(f) + x'$, where $x' \sim \mathcal{N}(0, \mathrm{diag}\{\eta_1^{-1}, \ldots, \eta_m^{-1}\})$ and the functionals in M are μ–orthonormal. Indeed, if $M = [K_1, \ldots, K_m]$ then $K_i = \sum_{j=1}^m \eta_j^{-1/2} q_j^{(i)} \tilde{L}_j$ and

$$\begin{aligned} \langle K_i, K_j \rangle_\mu &= \eta_i^{-1/2} \eta_j^{-1/2} \sum_{s,t=1}^n q_s^{(i)} q_t^{(j)} \langle \tilde{L}_i, \tilde{L}_j \rangle_\mu \\ &= \eta_i^{-1/2} \eta_j^{-1/2} \langle G q^{(i)}, q^{(j)} \rangle_2 \;=\; \delta_{ij}. \end{aligned}$$

The random variable y' is Gaussian with mean $M(f)$ and correlation matrix $\Sigma' = (D_1 Q^* \Sigma^{-1/2}) \Sigma (D_1 Q^* \Sigma^{-1/2})^* = \mathrm{diag}\{\eta_1^{-1}, \ldots, \eta_m^{-1}\}$.

We now show that the information operator $\mathbb{M} = \mathcal{N}(M(\cdot), D)$ is as powerful as $\mathbb{N}$ i.e., $\mathrm{rad}^{\mathrm{ave}}(\mathbb{M}) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$. To this end, it suffices to show that the conditional measures in F with respect to information $\mathbb{N}$ and $\mathbb{M}$ have the same correlation operator. This in turn holds iff the functionals $\sum_{j=1}^n z_j L_j$, where $(\Sigma + G_N) z = N C_\mu L$, and $\sum_{j=1}^n z_j' K_j$, where $(D + I) z' = M C_\mu L$, coincide for all $L \in F^*$. Indeed, straightforward calculations show that z and z' satisfy $Q^* \Sigma^{1/2} z = (\,D^{1/2} z', \underbrace{0, \ldots, 0}_{n-m})$. Hence,

$$\begin{aligned} \sum_{j=1}^m z_j' K_j(f) &= \langle z', M(f) \rangle_2 \;=\; \langle z', D_1 Q^* \Sigma^{-1/2} N(f) \rangle_2 \\ &= \langle Q^* \Sigma^{1/2} z, Q^* \Sigma^{1/2} N(f) \rangle_2 \;=\; \langle \Sigma^{-1/2} Q Q^* \Sigma^{-1/2} z, N(f) \rangle_2 \\ &= \sum_{j=1}^n z_j L_j(f), \end{aligned}$$

as claimed.

Consider now the case when Σ is singular, $\mathrm{rank}(\Sigma) = k < n$. Then there exists a nonsingular and symmetric matrix V such that $V \Sigma V = \mathrm{diag}\{\underbrace{0, \ldots, 0}_{n-k}, \underbrace{1, \ldots, 1}_{k}\}$. Let $VN = [\tilde{L}_1, \ldots, \tilde{L}_n]$. We can assume that the functionals $\tilde{L}_i$ and $\tilde{L}_j$ for $1 \le i \le n-k < j \le n$ are μ–orthogonal since otherwise $\tilde{L}_j$'s can be replaced by their μ–orthogonal projections onto $(\mathrm{span}\{\tilde{L}_1, \ldots, \tilde{L}_{n-k}\})^\perp$. Let $N_0 = [L'_1, \ldots, L'_{n-k}]$, $N_1 = [L'_{n-k+1}, \ldots, L'_n]$. Let D_0 be the zero matrix in $\mathbb{R}^{n-k}$ and let D_1 be the identity matrix in $\mathbb{R}^k$. Now we can use the already known procedure to transform $\mathbb{N}_0 = \mathcal{N}(N_0(\cdot), D_0)$ and $\mathbb{N}_1 = \mathcal{N}(N_1(\cdot), D_1)$ to equivalent information $\mathbb{M}_0$ and $\mathbb{M}_1$, where $\mathbb{M}_0$ is exact and both consist of independent observations of μ–orthonormal functionals. Then $\mathbb{M} = [\mathbb{M}_0, \mathbb{M}_1]$ also consists of independent observations of μ–orthonormal functionals, and $\mathbb{M}$ is equivalent to $\mathbb{N}$.

Let us see how the optimal algorithm and radius depend on the accuracy of information. As explained above, we can assume without loss of generality that $\Sigma = \sigma^2 D$ where $D = \mathrm{diag}\{\eta_1, \ldots, \eta_n\}$ and $\langle L_i, L_j \rangle_\mu = \delta_{ij}$, $1 \le i, j \le n$. Let $r(\sigma^2)$ be the radius of information $\mathbb{N}_\sigma = \mathcal{N}(N(\cdot), \sigma^2 D)$, and let φ_σ be the optimal algorithm for $\mathbb{N}_\sigma$. Then

$$\varphi_\sigma(y) \;=\; \varphi_0(y) \;-\; \sigma^2 \sum_{j=1}^n \frac{\eta_j}{1+\sigma^2 \eta_j}\, \langle S, L_j \rangle_\mu, \qquad y \in \mathbb{R}^n,$$

and

$$r^2(\sigma^2) \;=\; r^2(0) \;+\; \sigma^2 \sum_{j=1}^n \frac{\eta_j}{1+\sigma^2\eta_j}\, \|S(C_\mu L_j)\|^2.$$

Hence, for $r(0) > 0$ we have

$$r(\sigma^2) - r(0) \;\approx\; \sigma^2 \cdot \frac{\sum_{j=1}^n \eta_j \|S(C_\mu L_j)\|^2}{2\left(\mathrm{trace}(S C_\mu S^*) - \sum_{j=1}^n \|S(C_\mu L_j\|^2\right)^{1/2}},$$

while for $r(0) = 0$ we have

$$r(\sigma^2) - r(0) \;=\; r(\sigma^2) \;\approx\; \sigma \cdot \sqrt{\sum_{j=1}^n \eta_j \|S(C_\mu L_j)\|^2},$$

as $\sigma \to 0^+$. Thus, if $r(0) = 0$ then the radius of noisy information converges to the radius of exact information linearly in σ. Otherwise we have quadratic convergence. For S being a functional, this stands in contrast to results of the worst case setting where we always have linear convergence of $r(\delta)$ to $r(0)$; see Theorem 2.5.

Notes and Remarks

NR 3.11 In the case $\Sigma = \sigma^2 I$, the conditional distribution of a Gaussian measure was evaluated in Plaskota [78]. For exact information, $\Sigma = 0$, see Traub *et al.* [108].

NR 3.12 It is worthwhile to mention that the space G in Theorem 3.3 need not be a Hilbert space. That is, the algorithm $\varphi(y) = S(m(y))$, where $m(y)$ is the mean element of $\mu_2(\cdot|y)$, is optimal also when G is a separable Banach space. Indeed, observe that in this case the measure $\nu_2(\cdot|y) = \mu_2(S^{-1}(\cdot)|y)$ remains Gaussian with the mean element $S(m(y))$ (see E 3.14). Any Gaussian measure is centrosymmetric with respect to its mean element (see e.g. Vakhania *et al.* [113]). Hence, due to Example 3.3, the element $S(m(y))$ is the center of $\nu_2(\cdot|y)$ and the algorithm $\varphi(y) = S(m(y))$ is optimal.

Exercises

E 3.12 Prove that $N(\overline{C_\mu(F^*)}) = G_N(\mathbb{R}^n)$.

E 3.13 Show that the measure $\tilde{\mu}$ defined on $F\times\mathbb{R}^n$ is Gaussian. The mean element of $\tilde{\mu}$ is zero and the correlation operator is given as

$$C_{\tilde{\mu}}(\tilde{L}) \;=\; \big(\, C_\mu(L) \,+\, C_\mu(\,\langle N(\cdot), w\,\rangle_2),\; N(C_\mu L) \,+\, (\sigma^2\Sigma + G_N)\, w \,\big) \;\in\; F\times\mathbb{R}^n$$

where $\tilde{L}(f,y) = L(f) + \langle y, w\rangle_2$, $f\in F$, $y\in\mathbb{R}^n$.

E 3.14 Let F and G is separable Banach spaces and let $S: F\to G$ be a continuous linear operator. Let ω be a Gaussian measure on F with mean element m_ω and correlation operator C_ω. Show that then the measure ωS^{-1} on G is also Gaussian. Its mean element is $S(m_\omega)$ and correlation operator $C_{\omega S^{-1}}(L) = S(C_\mu(LS)\,)$, $L\in G^*$.

E 3.15 Suppose that the functionals L_j, $1\le j\le n$, are orthonormal and that $\Sigma = \mathrm{diag}\{\sigma_1^2,\dots,\sigma_n^2\}$. Let $P_N: F^*\to F^*$ be the μ–orthogonal projection onto the subspace $V = \mathrm{span}\{L_1,\dots,L_n\}$, and let $D: V\to V$ be defined by $D(L_j) = (1+\sigma_j^2)L_j$, $1\le j\le n$. Show that then the correlation operator C_{μ_2} of the conditional distribution $\mu_2(\cdot|y)$ can be rewritten as $C_{\mu_2} = C_\mu(I - D^{-1}P_N)$. Hence, for small σ_j^2 the operator C_{μ_2} is roughly the superposition of the "almost" μ–orthogonal projection onto $V^\perp$, and C_μ.

E 3.16 Show that $\varphi_{\mathrm{opt}}(y) = \varphi_0(y - \Sigma z)$ where $(\Sigma + G_N)z = y$ and φ_0 is the optimal algorithm for exact information ($\Sigma\equiv 0$).

E 3.17 Let $S: F\to G$ and $\mathbb{N}$ be given solution operator and linear information with Gaussian noise. Let μ_m be a Gaussian measure on F with the mean element m, not necessarily equal to zero. Let φ_m and $\mathrm{rad}_m^{\mathrm{ave}}(\mathbb{N})$ denote the optimal algorithm and radius of information with respect to μ_m. Show that for all $m\in F$ we have $\mathrm{rad}_m^{\mathrm{ave}}(\mathbb{N}) = \mathrm{rad}_0^{\mathrm{ave}}(\mathbb{N})$ and $\varphi_m(y) = S(m) + \varphi_0(y - N(m))$.

E 3.18 Let $\mathbb{N}$ be linear information with Gaussian noise, $\mathbb{N}(f) = \mathcal{N}(N(f), \Sigma)$, with $Y = \mathbb{R}^n$. Let $B : \mathbb{R}^n \to \mathbb{R}^n$ be a linear mapping. Show that for information $\mathbb{N}'$ given as $\mathbb{N}'(f) = \mathcal{N}(BN(f), B\Sigma B^*)$, we have $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \le \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}')$. If B is nonsingular then $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}')$ and the corresponding optimal algorithms satisfy $\varphi'_{\mathrm{opt}}(y) = \varphi_{\mathrm{opt}}(B^{-1}y)$.

3.5 The case of linear functionals

In this section we make an additional assumption that

- the solution operator S is a continuous linear functional.

In this case, the formulas for φ_{opt} and $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$ obtained in Theorem 3.3 can be expressed in a simple way. Namely, we have $\varphi_{\mathrm{opt}}(y) = \sum_{j=1}^n z_j S(C_\mu L_j) = \sum_{j=1}^n z_j L_j(C_\mu S) = \langle z, N(C_\mu S)\rangle_2$ where $(\Sigma + G_N)z = y$, or equivalently,

$$\varphi_{\mathrm{opt}}(y) \;=\; \langle\, y,\, w\,\rangle_2$$

where w satisfies $(\Sigma + G_N)w = N(C_\mu S)$. To find the radius, observe that $S(C_\mu S^*) = \|S\|_\mu^2$ and $Sm(NC_\mu S) = \langle w, N(C_\mu S)\rangle_2$. Hence,

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; \sqrt{\|S\|_\mu^2 \,-\, \langle w,\, N(C_\mu S)\rangle_2}.$$

For independent observations of μ–orthonormal functionals, i.e., when $\Sigma = D = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$ and $\langle L_i, L_j\rangle_\mu = \delta_{ij}$, we have

$$\varphi_{\mathrm{opt}}(y) \;=\; \sum_{j=1}^n y_j \frac{\langle S, L_j\rangle_\mu}{1+\sigma_j^2}$$

and

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; \sqrt{\|S\|_\mu^2 \,-\, \sum_{j=1}^n \frac{\langle S, L_j\rangle_\mu^2}{1+\sigma_j^2}}.$$

Let P_N be the μ–orthogonal projection onto $V = \mathrm{span}\{L_1, \ldots, L_n\}$. Then $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) = \|S - (I+D)^{-1}P_N S\|_\mu$. In particular, for exact information $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$ is the μ–distance between the functional S and the linear subspace spanned by the functionals L_j, and $\varphi_{\mathrm{opt}}(N(f)) = (P_N S)(f)$.

In Section 2.4.2, we noticed that in the worst case setting the problem of approximating a functional based on given information is as difficult as the hardest one–dimensional subproblem contained in the original problem.

Exercise E 2.5 of the same chapter shows that this actually holds for an arbitrary linear solution operator. We shall see that in the average case a positive result can be shown only if S is a functional.

For a functional $K \in F^*$ with $\|K\|_\mu > 0$, let $P_K : F \to F$ be given by

$$P_K(f) \;=\; f \,-\, \frac{K(f)}{\|K\|_\mu^2}\, C_\mu K\,.$$

That is, P_K is the projection onto $\ker K$ with $\ker P_K = \mathrm{span}\{C_\mu K\}$. The a priori Gaussian measure μ on F can be decomposed as

$$\mu \;=\; \int_{\ker K} \mu_K(\,\cdot\,|g)\,\mu P_K^{-1}(dg)$$

where $\mu_K(\,\cdot\,|g)$ is the conditional measure on F given $g = P_K(f)$. Clearly, $\mu_K(\,\cdot\,|g)$ is concentrated on the line

$$P_K^{-1}(g) \;=\; \{\, g \,+\, \alpha\, C_\mu K \;|\quad \alpha \in \mathbb{R}\,\}\,.$$

We also formally allow K with $\|K\|_\mu = 0$. In this case we set $P_0(f) = f$ $\forall\, f$. Hence, $\mu P_K^{-1} = \mu$ and $\mu_0(\cdot|g)$ is the Dirac measure concentrated in $\{g\}$, $\forall\, g \in F$.

Any functional K determines a *family* of one–dimensional subproblems. This family is indexed by $g \in \ker K$ and given as follows. For $g \in \ker K$, the subproblem relays on minimizing the average error

$$\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi; \mu_K(\cdot|g)\,) \;=\; \sqrt{\int_F \int_{\mathbb{R}^n} |S(f) - \varphi(y)|^2\, \pi_f(dy)\, \mu_K(df|g)}$$

over all algorithms φ. Thus, in the subproblem we use additional information that $g = P_K(f)$ or, in other words, that $f = g + \alpha C_\mu K$ for some $\alpha = \alpha(f) \in \mathbb{R}$.

Lemma 3.5 *Let $\|K\|_\mu > 0$. Then for all g a.e. the measure $\mu_K(\,\cdot\,|g)$ is Gaussian with mean $m(g) = g$ and correlation operator*

$$A_K(L) \;=\; \frac{\langle L, K\rangle_\mu}{\|K\|_\mu^2}\, C_\mu K.$$

Proof We shall use the fact that any Gaussian measure ω is uniquely determined by its characteristic functional ψ_ω. For $\omega = \mu P_K^{-1}$ we have

$$\begin{aligned}\psi_\omega(L) &= \int_{\ker K} \exp\{iL(g)\}\, \mu P_K^{-1}(dg) \;=\; \int_F \exp\{iL(P_K f)\}\, \mu(df) \\ &= \exp\left\{ -\frac{1}{2}\, (L\, P_K(C_\mu(LP_K)\,)\,) \right\}.\end{aligned}$$

Hence, the measure μP_K^{-1} is zero mean Gaussian and its correlation operator is given as

$$C_\omega(L) \;=\; P_K(C_\mu(LP_K)) \;=\; C_\mu L \;-\; \frac{\langle K, L\rangle_\mu}{\|K\|_\mu^2}\, C_\mu K\,.$$

Now, let $\mu'_K(\,\cdot\,|g)$ be the Gaussian measure with mean g and correlation operator A_K. Then the characteristic functional of the measure $\mu' \,=\, \int_{\ker K} \mu'_K(\,\cdot\,|g)\, \mu P_K^{-1}(dg)$ is given as

$$\begin{aligned}\psi_{\mu'}(L) &= \int_{\ker K}\int_F \exp\{iL(f)\}\, \mu'_K(\,df\,|g)\, \mu P_K^{-1}(dg) \\ &= \int_{\ker K} \exp\left\{ i\,L(g) \;-\; \frac{\langle K, L\rangle_\mu^2}{2\,\|K\|_\mu^2} \right\} \mu P_K^{-1}(dg) \\ &= \exp\left\{ -\frac{\langle K, L\rangle_\mu^2}{2\,\|K\|_\mu^2} \right\} \int_{\ker K} \exp\{iL(g)\}\, \mu P_K^{-1}(dg) \\ &= \exp\left\{ -\frac{1}{2}\langle L, L\rangle_\mu \right\}.\end{aligned}$$

This shows that $\mu \,=\, \mu'$. Since conditional measures are determined uniquely (up to a set of measure zero), the lemma follows. □

Since the measures $\mu_K(\cdot|g)$ have the same correlation operator for all $g \in \ker K$, the radius $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu_K(\cdot|g)\,)$ does not depend on g (compare with E 3.17). We denote it by $\mathrm{rad}_K^{\mathrm{ave}}(\mathbb{N})$. It is clear that

$$\mathrm{rad}_K^{\mathrm{ave}}(\mathbb{N}) \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}). \tag{3.6}$$

Indeed, we have

$$\begin{aligned}(\mathrm{rad}_K^{\mathrm{ave}}(\mathbb{N}))^2 &= \int_{\ker K} (\mathrm{e}^{\mathrm{ave}}(\mathbb{N}; \mu_K(\cdot|g))\,)^2\, \mu P_K^{-1}(dg) \\ &= \int_{\ker K} \inf_\varphi\, (\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi; \mu_K(\cdot|g))\,)^2\, \mu P_K^{-1}(dg)\end{aligned}$$

$$
\begin{aligned}
&\le \inf_{\varphi} \int_{\ker K} \left(\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi; \mu_K(\cdot|g))\,\right)^2 \, \mu P_K^{-1}(dg) \\
&= \inf_{\varphi} \left(\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi)\right)^2 \;=\; \left(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})\right)^2 .
\end{aligned}
$$

We now prove that for a special choice of K we have equality in (3.6).

Theorem 3.4 *Consider the family of one–dimensional subproblems determined by the functional*

$$K^* \;=\; S \,-\, \langle w, N(\cdot)\rangle_2 \;=\; S \,-\, \sum_{j=1}^n w_j L_j$$

where $(\Sigma + G_N)w = N(C_\mu S)$*. Then, for all* g *a.e., the algorithm* $\varphi_{\mathrm{opt}}(y) = \langle y, w\rangle_2$ *is optimal for the subproblem determined by* $g \in \ker K^*$*. Furthermore,*

$$\mathrm{rad}_{K^*}^{\mathrm{ave}}(\mathbb{N}) \;=\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}).$$

Proof If $\|K^*\|_\mu = 0$ then $S(f) = \langle w, N(f)\rangle_2$, $\forall f \in \overline{C_\mu(F^*)}$. In this case, the measure $\mu_{K^*}(\cdot|g)$ is concentrated in $\{g\}$ and any algorithm with the property $\varphi(N(g)) = S(g)$ is optimal for the subproblem indexed by g. As for $g \in \overline{C_\mu(F^*)}$ we have $\varphi_{\mathrm{opt}}(N(g)) = \langle w, N(g)\rangle_2 = S(g)$, the algorithm φ_{opt} is optimal for any subproblem a.e. Moreover, we have $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) = \sqrt{S(C_\mu K^*)} = 0 = \mathrm{rad}_{K^*}^{\mathrm{ave}}(\mathbb{N})$.

Assume that $K^* \neq 0$. Let ω be the zero mean Gaussian measure with correlation operator $A = A_{K^*}$, where A_{K^*} is defined in Lemma 3.5. We need to show that the algorithm $\varphi_{\mathrm{opt}} = \langle\,\cdot\,, w\rangle_2$ is optimal if the average error over f is taken with respect to ω, i.e.,

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \omega) \;=\; \inf_{\varphi} \mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi; \omega) \;=\; \mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{opt}}; \omega)$$

where $\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi; \omega) = \sqrt{\int_F \int_{\mathbb{R}^n} \|S(f) - \varphi(y)\|^2 \, \pi_f(dy) \, \omega(df)}$.

Due to Theorem 3.3, the optimal algorithm with respect to ω is given by $\varphi_\omega(y) = \sum_{j=1}^n z_j S(AL_j)$, where $(\Sigma + H_N)z = y$, $H_N = \{L_i(AL_j)\}_{i,j=1}^n$, and $z, y \in (\Sigma + H_N)(\mathbb{R}^n)$. We have $L_i(AL_j) = \langle K^*, L_i\rangle_\mu \langle K^*, L_j\rangle_\mu / \|K\|_\mu^2$ and $S(AL_j) = \langle K^*, S\rangle_\mu \langle K^*, L_j\rangle_\mu / \|K^*\|_\mu^2$. Hence, setting $a = N(C_\mu K^*)$ we obtain

$$\varphi_\omega(y) \;=\; \frac{\langle K^*, S\rangle_\mu}{\|K^*\|_\mu^2} \langle\, z,\, a\,\rangle_2 \tag{3.7}$$

where

$$\Sigma z + \frac{\langle a, z\rangle_2}{\|K^*\|_\mu^2}\, a \;=\; y. \tag{3.8}$$

Observe now that

$$\begin{aligned} a \;=\; N(C_\mu K^*) \;&=\; N(C_\mu S) \;-\; \sum_{j=1}^n w_j N(C_\mu L_j) \\ &=\; N(C_\mu S) \;-\; G_N w \;=\; \Sigma\, w. \end{aligned}$$

This and (3.8) yield $\langle y, w\rangle_2 = \langle \Sigma z, w\rangle_2 + \langle a, z\rangle_2 \langle a, w\rangle_2 / \|K^*\|_\mu^2 = \langle z, a\rangle_2 (1 + \langle \Sigma w, w\rangle_2 / \|K^*\|_\mu^2)$, so that

$$\langle z, a\rangle_2 \;=\; \frac{\|K^*\|_\mu^2\, \langle y, w\rangle_2}{\|K^*\|_\mu^2 + \langle \Sigma w, w\rangle_2}. \tag{3.9}$$

We also have

$$\begin{aligned} \langle S, K^*\rangle_\mu \;&=\; \|S\|_\mu^2 \;-\; \langle w, N(C_\mu S)\rangle_2 \qquad (3.10) \\ &=\; (\, \|S\|_\mu^2 - 2\, \langle w, N(C_\mu S)\rangle_2 + \langle G_N w, w\rangle_2\,) \\ &\qquad + (\, \langle w, (\Sigma + G_N) w\rangle_2 - \langle G_N w, w\rangle_2\,) \\ &=\; \|K^*\|_\mu^2 \;+\; \langle \Sigma w, w\rangle_2. \end{aligned}$$

Taking together (3.9), (3.10) and (3.7) we finally obtain

$$\varphi_\omega(y) \;=\; \frac{\langle S, K^*\rangle_\mu}{\|K^*\|_\mu^2 + \langle \Sigma w, w\rangle_2}\, \langle y, w\rangle_2 \;=\; \langle y, w\rangle_2,$$

as claimed.

Now, let ω_g be the Gaussian measure with mean $g \in \ker K^*$ and correlation operator A. Then, due to E 3.17, the optimal algorithm for ω_g is given as

$$\begin{aligned} \varphi_g(y) \;&=\; S(g) + \langle y - N(g), w\rangle_2 \;=\; S(g) - \langle N(g), w\rangle_2 + \langle y, w\rangle_2 \\ &=\; K^*(g) \;+\; \langle y, w\rangle_2 \;=\; \langle y, w\rangle_2 \;=\; \varphi_{\mathrm{opt}}(y). \end{aligned}$$

Since $\mu_{K^*}(\cdot|g) = \omega_g$ $\forall g$ a.e., the algorithm φ_{opt} is optimal for all subproblems almost everywhere.

To prove the equality $\mathrm{rad}^{\mathrm{ave}}_{K^*}(\mathbb{N}) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$, observe that

$$\begin{aligned}
(\mathrm{rad}^{\mathrm{ave}}_{K^*}(\mathbb{N})\,)^2 &= \int_{\ker K^*} (\mathrm{rad}^{\mathrm{ave}}(\mathbb{N};\mu_{K^*}(\cdot|g)))^2 \;\mu P_{K^*}^{-1}(dg) \\
&= \int_{\ker K^*} \int_F \left(\int_{\mathbb{R}^n} \|S(f) - \varphi_{\mathrm{opt}}(y)\|^2 \,\pi_f(dy) \right) \\
&\qquad\qquad\qquad\qquad \mu_{K^*}(df|g)\,\mu P_{K^*}^{-1}(dg) \\
&= \int_F \int_{\mathbb{R}^n} \|S(f) - \varphi_{\mathrm{opt}}(y)\|^2 \,\pi_f(dy)\,\mu(df) \\
&= (\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})\,)^2.
\end{aligned}$$

This completes the proof of the theorem. □

Thus we have shown that there exists a family of one–dimensional subproblems which are as difficult as the original problem. In words, this result can be interpreted as follows: approximation of $S(f)$ based on information $y \in \mathbb{N}(f)$ is as difficult as approximation of $S(f)$ based on y and the additional information that f is in the line $\{g + \alpha\, C_\mu K^* \mid \alpha \in \mathbb{R}\}$.

We summarize our analysis in the following corollary.

Corollary 3.2 *Let the solution operator S be a functional. Then, for any information $\mathbb{N}$, we have*

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; \sup_{K \in F^*} \,\mathrm{rad}^{\mathrm{ave}}_K(\mathbb{N}) \;=\; \mathrm{rad}^{\mathrm{ave}}_{K^*}(\mathbb{N})$$

where the functional K^* is given by $K^*(f) = S(f) - \varphi_{\mathrm{opt}}(N(f)\,)$, $f \in F$. □

If S is not a functional then Corollary 3.2 is no longer true. An example is moved to E 3.22.

Notes and Remarks

NR 3.13 The main result of this section was obtained by Plaskota [82].

Exercises

E 3.19 Suppose we want to estimate a real random variable f, $f \sim \mathcal{N}(0,\lambda)$, $\lambda > 0$, based on the data $y = f + x$ where $x \sim \mathcal{N}(0,\sigma^2)$. Show that in this case the radius equals

$$r(\sigma^2) \;=\; \sqrt{\frac{\sigma^2\,\lambda}{\sigma^2+\lambda}},$$

and the optimal algorithm

$$\varphi_{\mathrm{opt}}(y) \;=\; \frac{\lambda}{\sigma^2+\lambda}\,y, \qquad y \in \mathbb{R}.$$

E 3.20 Consider the problem of the previous exercise but with information $y = [y_1, \ldots, y_n]$ where $y_i \sim \mathcal{N}(f, \sigma_i^2)$ and $\sigma_j^2 > 0$, $1 \le i \le n$. Show that the radius of information is given as

$$r(\sigma_1^2, \ldots, \sigma_n^2) \;=\; \sqrt{\frac{\lambda}{1+\lambda\sum_{i=1}^n 1/\sigma_i^2}}$$

and

$$\varphi_{\mathrm{opt}}(y) \;=\; \frac{\lambda}{1+\lambda\sum_{i=1}^n 1/\sigma_i^2}\,\sum_{i=1}^n \frac{y_i}{\sigma_i^2}.$$

Hence, n observations of f with variances σ_i^2 is as good as one observation of f with the variance $\sigma^2 = (\sum_{i=1}^n 1/\sigma_i^2)^{-1}$.

E 3.21 Consider the one–dimensional linear problem with the correlation operator $C_\mu(L) = \lambda L(f_0) f_0$, $\forall L \in F^*$, where $\lambda > 0$ and $f_0 \in F$. Let $g_0 = S(f_0) \in \mathbb{R}$ and $y_0 = N(f_0) \in \mathbb{R}^n$. Show that for $y_0 \in \Sigma(\mathbb{R}^n)$ we have

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; |g_0|\sqrt{\frac{\lambda}{1+\lambda\,\langle\Sigma^{-1}y_0, y_0\rangle_2}}, \qquad \varphi_{\mathrm{opt}}(y) \;=\; g_0\,\frac{\lambda\,\langle\Sigma^{-1}y_0, y\rangle_2}{1+\lambda\,\langle\Sigma^{-1}y_0, y_0\rangle_2},$$

while for $y_0 \notin \Sigma(\mathbb{R}^n)$ we have $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) = 0$ and

$$\varphi_{\mathrm{opt}}(y) \;=\; g_0\,\frac{\langle Py_0, y\rangle_2}{\langle Py_0, y_0\rangle_2},$$

where P is the orthogonal projection in $\mathbb{R}^n$ onto $(\Sigma(\mathbb{R}^n))^\perp$.

E 3.22 Let $F = G = \mathbb{R}^d$ and let S be the identity. Let μ be the standard Gaussian measure on $\mathbb{R}^d$, $\mu = \mathcal{N}(0, I)$. Consider information $\mathbb{N}$ consisting of $n < d$ noisy observations. Show that $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \ge d - n$, while for any functional $K \in F^*$ we have $\mathrm{rad}_K^{\mathrm{ave}}(\mathbb{N}) \le 1$. Hence,

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;\ge\; (d-n)\sup_{K\in F^*}\,\mathrm{rad}_K^{\mathrm{ave}}(\mathbb{N})$$

and any one–dimensional subproblem is $d-n$ times easier than the original problem.

3.6 Optimal algorithms as smoothing splines

Recall that in the worst case setting we defined an α–smoothing spline algorithm as $\varphi_\alpha(y) = S(\mathbf{s}_\alpha(y)\,)$, where $\mathbf{s}_\alpha(y)$ is the α–smoothing spline element. It minimizes the functional

$$\Gamma_\alpha(f, y) \;=\; \alpha \cdot \|f\|_F^2 \,+\, (1-\alpha)\cdot \delta^{-2}\|y - N(f)\|_Y^2,$$

where $\|\cdot\|_F$ and $\|\cdot\|_Y$ are Hilbert extended seminorms. Moreover, for a properly chosen α, the algorithm φ_α turns out to be optimal; see Section 2.5.2.

In this section, we show that optimal algorithms in the average case setting can also be viewed as smoothing spline algorithms. We use the fact that Gaussian measures can be equivalently defined as abstract Wiener spaces.

3.6.1 A general case

We consider the linear problem of Section 3.4. That is, the measure μ on F is zero mean Gaussian. Information $\mathbb{N}$ is linear with Gaussian noise, $\mathbb{N}(f) = \mathcal{N}(N(f), \sigma^2\Sigma)$ where $\sigma^2 > 0$. The operator N consists of functionals $L_i \in F^*$,

$$N \;=\; [\,L_1, L_2, \ldots, L_n\,].$$

Let H be the associated with μ separable Hilbert space, so that the pair $(H, \overline{C_\mu(F^*)})$ is an abstract Wiener space. Recall that $C_\mu(F^*) \subset H \subset \overline{C_\mu(F^*)}$, see Section 3.3.2. Let $\|\cdot\|_Y$ be the extended norm in $\mathbb{R}^n$ defined as

$$\|x\|_Y \;=\; \begin{cases} \sqrt{\langle \Sigma^{-1}x, x\rangle_2} & x \in \Sigma(\mathbb{R}^n), \\ +\infty & x \notin \Sigma(\mathbb{R}^n). \end{cases}$$

We denote by $\mathbf{s}(y) \in H$ the smoothing spline that minimizes

$$\Gamma(f, y) \;=\; \|f\|_H^2 \,+\, \frac{1}{\sigma^2}\,\|y - N(f)\|_Y^2$$

over all $f \in H$. For instance, in the case of independent observations, $\Sigma = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$ and $\sigma^2 = 1$, $\mathbf{s}(y)$ is the minimizer of

$$\|f\|_H^2 + \sum_{j=1}^n \frac{1}{\sigma_j^2}(y_j - L_j(f)\,)^2$$

(with the convention that $0/0 = 0$). As usually, the smoothing spline algorithm is given as

$$\varphi_{\mathrm{spl}}(y) \;=\; S(\mathbf{s}(y)\,), \qquad y \in \mathbb{R}^n.$$

Let $f_j = C_\mu L_j \in H$, $1 \le j \le n$. Then f_j is the representer of L_j in H and for all $f \in H$ we have

$$N(f) \;=\; [\,\langle f_1, f\rangle_H, \langle f_2, f\rangle_H, \ldots, \langle f_n, f\rangle_H].$$

Applying Lemma 2.9 we immediately obtain that $\Gamma(y) = \inf_{f\in F}\Gamma(f,y)$ is finite if and only if $y \in \Sigma(\mathbb{R}^n) + N(F)$, or equivalently, $y \in Y_1 = (\sigma^2\Sigma + G_N)(\mathbb{R}^n)$. For $y \in Y_1$, the smoothing spline is unique and given as

$$\mathbf{s}(y) \;=\; \sum_{j=1}^{n} z_j f_j$$

where $z \in Y_1$ satisfies $(\sigma^2\Sigma + G_N)z = y$. Comparing this with Theorem 3.2 we obtain that $\mathbf{s}(y)$ is the mean element $m(y)$ of the conditional distribution on F. Hence, $\varphi_{\mathrm{spl}}(y) = S(\mathbf{s}(y)) = S(m(y)\,)$ is the optimal algorithm.

Theorem 3.5 *The smoothing spline algorithm φ_{spl} is optimal.* □

Thus, in the average case setting, optimal algorithms are smoothing spline algorithms. Observe that, unlike in the worst case, this time we have no problems with the optimal choice of the parameters α or $\gamma = \alpha(1-\alpha)^{-1}\sigma^2$. Namely, we always have $\alpha^* = 1/2$ and $\gamma^* = \sigma^2$.

3.6.2 Special cases

The formulas for α–smoothing splines in some special cases were given in Section 2.6. Clearly, they can be applied to obtain optimal algorithms in the average case. It suffices to set $\alpha = 1/2$ and replace mechanically δ^2 and γ by σ^2, and the norm $\|\cdot\|_F$ by $\|\cdot\|_H$. Therefore we now devote more attention only to the regularization and least squares.

Consider the linear problem of Section 3.6.1 with positive definite matrix Σ. Then $\|\cdot\|_Y$ is a Hilbert norm. We denote by Y the Hilbert space of vectors from $\mathbb{R}^n$ with the inner product $\langle\cdot,\cdot\rangle_Y = \langle\Sigma^{-1}(\cdot),\cdot\rangle_2$. Let $N_H : H \to Y$ be the restriction of $N : F \to Y$ to the subspace $H \subset F$, i.e., $N_H(f) = N(f)$, $\forall f \in H$. Let $N_H^* : Y \to H$ be the adjoint operator to N_H. That is, N_H^*

is defined by $\langle N_H(f), y\rangle_Y = \langle f, N_H^*(y)\rangle_H$, $\forall f \in H$, $\forall y \in Y$. Similarly, we define the operators $S_H : H \to G$ and $S_H^* : G \to H$.

Recall that the regularized approximation of $S(f)$ is given as $\varphi_\gamma(y) = S(u_\gamma(y))$ where $u_\gamma(y) \in H$ is the solution of the equation

$$(\gamma I_H + N_H^* N_H) f \;=\; N_H^* y.$$

Here $\gamma > 0$ is the regularization parameter and I_H is the identity in H; compare with Section 2.6.2. In view of Lemma 2.11 and Theorem 3.5, we immediately obtain the following fact.

Corollary 3.3 *The regularized solution*

$$u_\gamma(y) \;=\; (\gamma I_H + N_H^* N_H)^{-1} N_H^* y$$

is the smoothing spline, $u_\gamma(y) = \mathbf{s}(y)$, if and only if the regularization parameter $\gamma = \sigma^2$. Hence, the algorithm $\varphi_{\sigma^2}(y) = S(u_{\sigma^2}(y)\,)$ is optimal. □

We now derive a formula for the radius. Let $\{\xi_i\}_{i\ge 1}$ be the complete orthonormal in H basis of eigenelements of $N_H^* N_H$, and let η_i's be the corresponding eigenvalues, $N_H^* N_H \xi_i = \eta_i \xi_i$. We assume without loss of generality that $\eta_1 \ge \cdots \ge \eta_k > 0$ where $k = \dim N(F)$.

Lemma 3.6 *We have*

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;=\; \sqrt{\sigma^2 \cdot \sum_{i=1}^{k} \frac{\|S(\xi_i)\|^2}{\sigma^2 + \eta_i} \;+\; \sum_{j=k+1}^{\dim H} \|S(\xi_j)\|^2}\,.$$

Proof Observe that for any continuous linear operator $A : F \to H_1$ where H_1 is a Hilbert space, we have $A_H^* h = C_\mu(A^* h)$, $A^* h = \langle A(\cdot), h\rangle_{H_1}$, $\forall h \in H_1$. This and Theorem 3.3 yield

$$\begin{aligned}
&(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}))^2 \;=\; \mathrm{trace}(SC_\mu S^*) \;-\; \mathrm{trace}(Sm(NC_\mu S^*)\,) \\
=\;& \mathrm{trace}(SC_\mu S^*) \;-\; \mathrm{trace}(\,S(\sigma^2 I_H + N_H^* N_H)^{-1} N_H^*(NC_\mu S^*)\,) \\
=\;& \mathrm{trace}(\,S_H S_H^*) \;-\; \mathrm{trace}(S_H(\sigma^2 I_H + N_H^* N_H)^{-1} N_H^* N_H S_H^*\,) \\
=\;& \sum_{i\ge 1} \|S_H(\xi_i)\|^2 \;-\; \sum_{i\ge 1} \|S_H(\,(\sigma^2 I_H + N_H^* N_H)^{-1} N_H^* N_H)^{1/2} \xi_i\|^2 \\
=\;& \sigma^2 \cdot \sum_{i=1}^{k} \frac{\|S(\xi_i)\|^2}{\sigma^2 + \eta_i} \;+\; \sum_{j\ge k+1} \|S(\xi_j)\|^2,
\end{aligned}$$

as claimed. □

We note that if the matrix Σ is singular then the formula for $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$ in Lemma 3.6 holds with H replaced by the space $H_1 = \{ f \in H \mid N(f) \in \Sigma(\mathbb{R}^n) \}$ with the norm $\|\cdot\|_{H_1} = \|\cdot\|_H$ (compare with the proof of Theorem 2.10). In the special case, when the operators $S_H^* S_H$ and $N_H^* N_H$ possess a common orthonormal basis of eigenelements and $S_H^* S_H \xi_i = \lambda_i \xi_i$, we have

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; \sqrt{\sigma^2 \sum_{i=1}^{k} \frac{\lambda_i}{\sigma^2 + \eta_i} \;+\; \sum_{j=1}^{\dim H} \lambda_j}\;.$$

Let us now consider the (generalized) least squares algorithm φ_{ls}, as applied to a problem with $F = \mathbb{R}^d$. We assume that the correlation operator C_μ of the Gaussian measure μ on $\mathbb{R}^d$ is positive definite. Information about f is given as $y = N(f) + x$, where $\dim N(\mathbb{R}^d) = d$ and $x \sim \mathcal{N}(0, \sigma^2\,\Sigma)$. Recall that $\varphi_{\mathrm{ls}} = S(N^*N)^{-1}N^*$, or equivalently, $\varphi_{\mathrm{ls}} = SN^{-1}P_N$ where P_N is the orthogonal projection onto $N(\mathbb{R}^d)$ with respect to $\langle\cdot,\cdot\rangle_Y$.

As the optimal value of the regularization parameter is $\gamma = \sigma^2$, the least squares are optimal only for exact information, $\sigma^2 = 0$. However, it turns out that for small noise level δ, this algorithm is very close to optimal. Namely, we have the following theorem.

Theorem 3.6 *For the (generalized) least squares algorithm* φ_{ls} *we have*

$$\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{ls}}) \;=\; \sigma \cdot \sqrt{\mathrm{trace}(S(N^*N)^{-1}S^*)} \;\approx\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}), \qquad \textit{as} \quad \sigma^2 \to 0^+.$$

Proof The formula for $\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{ls}})$ follows from the fact that for any f

$$\begin{aligned} \int_{\mathbb{R}^d} \|S(f) - \varphi_{\mathrm{ls}}(N(f)+x)\|^2\,\pi(dx) \;&=\; \int_{\mathbb{R}^d} \|SN^{-1}P_N(x)\|^2\,\pi(dx) \\ =\;\; \sigma^2\,\mathrm{trace}(\,(SN^{-1})(SN^{-1})^*) \;&=\; \sigma^2\,\mathrm{trace}(S(N^*N)^{-1}S^*). \end{aligned}$$

Since $N^* = C_\mu^{-1}N_H^*$ and $S^* = C_\mu^{-1}S_H^*$, we can equivalently write

$$\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{ls}}) \;=\; \sigma^2\,\mathrm{trace}(\,S_H(N_H^*N_H)^{-1}S_H^*\,).$$

Denote, as before, by ξ_i, η_i the eigenpairs of the operator $N_H^* N_H$. Letting $\sigma^2 \to 0^+$ and using Lemma 3.6 we obtain

$$\begin{aligned}\sigma^2 \,\mathrm{trace}(S_H(N_H^* N_H)^{-1} S_H^*) &= \sigma^2 \sum_{i=1}^{d} \frac{\|S(\xi_i)\|^2}{\eta_i} \\ &\approx \sigma^2 \sum_{i=1}^{d} \frac{\|S(\xi_i)\|^2}{\sigma^2 + \eta_i} \;=\; (\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}))^2,\end{aligned}$$

which completes the proof.

3.6.3 A correspondence theorem

The fact that smoothing spline algorithms are optimal in the worst and average case settings enables us to show a correspondence between both settings. Namely, consider the following two problems.

P1: Approximate $S(f)$ for $f \in E \subset F$, based on information $y = N(f) + x \in \mathbb{R}^n$ where $x \in \Sigma(\mathbb{R}^n)$ and $\|x\|_Y = \sqrt{\langle \Sigma^{-1} x, x \rangle_2} \le \delta$.

P2: Approximate $S(f)$, where $f \in F$ is distributed according to a zero mean Gaussian measure μ on F, based on information $y = N(f) + x \in \mathbb{R}^n$ such that $x \sim \mathcal{N}(0, \sigma^2 \Sigma)$.

Then we have the following correspondence theorem.

Theorem 3.7 *Suppose that $\{H, F\}$ is the abstract Wiener space corresponding to the measure μ, and that the set E is the unit ball of H. If $\delta^2 = \sigma^2$ then the algorithm $\varphi_{\mathrm{spl}}(y) = S(\mathbf{s}(y))$ is optimal for the problem (P2) in the average case,* $\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{spl}}; \mu) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu)$*, and almost optimal for the problem (P1) in the worst case,* $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{spl}}; E) \le \sqrt{2} \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{spl}}; E)$*. Furthermore,*

$$\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E) \;\le\; \sqrt{2} \cdot \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu).$$

If S is a functional then

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu) \;\le\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E) \;\le \sqrt{2} \cdot \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu).$$

Proof Optimality or almost optimality of φ_{spl} follows from Theorem 3.5, Lemma 2.7, and the fact that $\mathbf{s}(y)$ is the 1/2–smoothing spline for (P1).

To obtain the formulas for the radii, we proceed as follows. Assume first that $\Sigma > 0$. Let $\{f_i\}_{i=1}^{\dim H}$ be the complete and orthonormal in H basis of eigenelements of the operator $N_H^* N_H : H \to H$, $N_H^* N_H f_i = \eta_i f_i$, $i \geq 1$. Due to Lemma 3.6 we have

$$(\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\mathrm{spl}}))^2 \;=\; (\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}))^2 \;=\; \sigma^2 \sum_{i=1}^{\dim H} \frac{\|S(f_i)\|^2}{\sigma^2 + \eta_i}. \tag{3.11}$$

On the other hand, from Lemma 2.7 and Theorem 2.10 we have

$$\begin{aligned} (\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{spl}}))^2 &\leq 2\,\delta^2\, \|S(\delta^2 I_H + N_H^* N_H)^{-1/2}\|^2 \\ &\leq 2\,(\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}))^2. \end{aligned}$$

Note that any operator $T : H \to G$ satisfies $\|T\|^2 \leq \mathrm{trace}(T^* T)$, and if T is a functional then $\|T\|^2 = \mathrm{trace}(T^* T)$. This and (3.11) yield

$$\begin{aligned} (\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{spl}}))^2 &\leq 2\,\delta^2 \sum_{i=1}^{\dim H} \|S(\delta^2 I_H + N_H^* N_H)^{-1/2} f_i\|^2 \\ &= 2\,\delta^2 \sum_{i=1}^{\dim H} \frac{\|S(f_i)\|^2}{\delta^2 + \eta_i} \;=\; 2\,(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}))^2, \end{aligned}$$

which proves $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \leq \sqrt{2}\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$. If S is a functional then

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; \frac{1}{\sqrt{2}}\,\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_{\mathrm{spl}}) \;\leq\; \frac{1}{\sqrt{2}}\,(\sqrt{2}\,\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})\,) \;=\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}),$$

as claimed.

If Σ is singular then we repeat the proof with H replaced by $H_1 = \{\, f \in H \mid N(f) \in \Sigma(\mathbb{R}^n)\,\}$.

Notes and Remarks

NR 3.14 Optimality of spline algorithms in the average case setting and for exact information was shown in Traub *et al.* [108, Sect.5.4 of Chap.6]. Optimality properties of smoothing splines in reproducing kernel Hilbert spaces and for $\Sigma = \sigma^2 I$ are well known in Bayesian statistics. We mention only Kimeldorf and Wahba [37] and Wahba [116] where many other references can be found. The general result of Theorem 3.5 (together with Lemma 2.9) is however new.

NR 3.15 The correspondence theorem is well known in the case of exact information, $\Sigma \equiv 0$, and solution operator S being a functional. Then the algorithm φ_{spl} is optimal in both settings and $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$. The generalization of these results to the noisy case and arbitrary S seem to be new.

Exercises

E 3.23 Show that for exact information, $\Sigma \equiv 0$, the optimal algorithms in the worst and average case settings are the same and given as $\varphi_{\mathrm{opt}}(y) \;=\; S(\mathbf{s}(y)\,)$, $y \in N(F)$, where $\mathbf{s}(y) \in H$ is such an element that $N(\mathbf{s}(y)\,) = y$ and $\|\mathbf{s}(y)\|_H = \inf\{\,\|f\|_H \mid f \in H,\ N(f) = y\,\}$. Moreover, if S is a functional then $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$.

E 3.24 Consider the approximation of a parameter $f \in \mathbb{R}$ based on information $y = f + x$, where
(a) $|f| \le 1$ and $|x| \le \delta$,
(b) $f \sim \mathcal{N}(0,1)$ and $x \sim \mathcal{N}(0,\sigma^2)$.
Let $r^w(\gamma)$ and $r^a(\gamma)$ be the worst and average radii of information for the problems (a) and (b) with $\delta^2 = \gamma^2$ and $\sigma^2 = \gamma^2$, respectively. Show that

$$\frac{r^w(\gamma)}{r^a(\gamma)} \;=\; \begin{cases} (1+\gamma^2)^{1/2} & 0 \le \gamma \le 1, \\ (1+\gamma^{-2})^{1/2} & \gamma > 1. \end{cases}$$

That is, the ratio $r^w(\gamma)/r^a(\gamma)$, $\gamma \ge 0$, assumes all values from the interval $[1,\sqrt{2}]$.

E 3.25 Suppose that the solution operator S in Theorem 3.7 is finite dimensional, i.e., $\dim S(F) = d < +\infty$. Show that then

$$d^{-1}\cdot \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;\le\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}) \;\le\; \sqrt{2}\cdot \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}).$$

E 3.26 Show that the inequality $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \le \mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ in Theorem 3.7 does not hold any longer if S is not a functional. Even more, the ratio $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})/\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$ can be arbitrarily large.

3.7 Varying information

With this section we start the study of varying information. Basically, we assume that information can be obtained as in the worst case setting. The only difference is in the interpretation of noise which is now random.

3.7.1 Nonadaptive and adaptive information

A nonadaptive information operator $\mathbb{N}$ is uniquely determined by exact information $N : F \to \mathbb{R}^n$,

$$N(f) \;=\; [\,L_1(f), L_2(f), \ldots, L_n(f)\,], \qquad \forall\, f \in F,$$

where L_i's are continuous linear functionals, and by a precision vector $\Sigma = [\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2]$ where $\sigma_i^2 \ge 0$, $1 \le i \le n$. Given N and Σ, the nonadaptive noisy information operator $\mathbb{N} = \{N, \Sigma\}$ is given as

$$\mathbb{N}(f) \;=\; \mathcal{N}(N(f), \Sigma\,)$$

where $\mathcal{N}(N(f), \Sigma\,)$ is the n–dimensional Gaussian measure with mean $N(f)$ and diagonal correlation matrix $\mathrm{diag}\{\sigma_1^2, \dots, \sigma_n^2\}$. This means that the successive observations are independent and the noise of ith observation has normal distribution, $x_i = y_i - L_i(f) \sim \mathcal{N}(0, \sigma_i^2)$.

We shall use the same letter Σ to denote the precision vector $\Sigma = [\sigma_1^2, \dots, \sigma_n^2]$ as well as the matrix $\mathrm{diag}\{\sigma_1^2, \dots, \sigma_n^2\}$.

We now define adaptive information. As in the worst case, we assume that the set Y of possible information values satisfies the following condition:

$$\text{for any } (y_1, y_2, \dots\,) \in \mathbb{R}^\infty \qquad \begin{array}{l}\text{there exists exactly one index } n \\ \text{such that } (y_1, y_2, \dots, y_n\,) \in Y.\end{array}$$

An adaptive information operator $\mathbb{N}$ is determined by a family $N = \{N_y\}_{y \in Y}$ of exact nonadaptive information operators,

$$N_y \;=\; [\,L_1(\cdot), L_2(\cdot; y_1), \dots, L_{n(y)}(\cdot; y_1, \dots, y_{n(y)-1})\,],$$

where $L_i(\cdot; y_1, \dots, y_{n-1}) \in F^*$, $1 \le i \le n$, and $n(y)$ is the length of y, and by a family $\Sigma = \{\Sigma_y\}_{y \in Y}$ of precision vectors,

$$\Sigma_y \;=\; [\,\sigma_1^2, \sigma_2^2(y_1), \dots, \sigma_{n(y)}^2(y_1, \dots, y_{n(y)-1})\,].$$

To complete the definition, we have specify the measures $\pi_f = \mathbb{N}(f)$ on Y, for $f \in F$. They are defined as follows. For $m \ge 1$, let

$$W_m \;=\; \{\,(y_1, \dots, y_m) \in \mathbb{R}^m \mid \quad (y_1, \dots, y_j) \notin Y,\ 1 \le j \le m-1\,\}$$

($W_1 = \mathbb{R}$). In words, $y \in W_m$ iff $y \in \mathbb{R}^m$ and it can be extended to a vector belonging to Y. Note that W_m are measurable. Indeed, so is W_1, and for arbitrary $m \ge 2$ we have $W_m = (W_{m-1} \setminus Y_{m-1}) \times \mathbb{R}$.

Assuming the maps $L_i(f; \cdot) : \mathbb{R}^{i-1} \to \mathbb{R}$ and $\sigma_i(\cdot) : \mathbb{R}^{i-1} \to \mathbb{R}_+$ are Borel measureable, we define on W_m measures $\omega_m = \omega_{m,f}$ as follows. Let $\mathcal{G}(\cdot | t, \sigma^2)$ be the one–dimensional Gaussian measure with mean $t \in \mathbb{R}$ and variance $\sigma^2 \ge 0$. Then

$$\begin{aligned} \omega_1(B) \;&=\; \mathcal{G}\left(B \middle| L_1(f), \sigma_1^2\right), \\ \omega_{m+1}(B) \;&=\; \int_{W_m \setminus Y_m} \mathcal{G}\left(B^{(t)} \middle| L_{m+1}(f; t), \sigma_{m+1}^2(t)\right)\, \omega_m(dt) \end{aligned}$$

where $t \in \mathbb{R}^m$ and $B^{(t)} = \{ u \in \mathbb{R} \,|\, (t,u) \in B \}$. The measure π_f is now given as

$$\pi_f(\cdot) \;=\; \sum_{m=1}^{\infty} \omega_m(\,\cdot\, \cap Y_m).$$

Lemma 3.7 *π_f is a well defined probability measure on Y.*

Proof The σ-field on Y is generated by cylindrical sets of the form

$$B \;=\; \left(\bigcup_{i=1}^{m-1} B_i \right) \cup \{ y \in Y \,|\, y^m \in A_m \}$$

where A_m is a Borel set of W_m and y^m is the vector consisting of the first m components of y, $m \geq 1$. For any such a set, we let

$$\tilde{\pi}_f(B) \;=\; \sum_{i=1}^{m-1} \omega_i(B_i) \,+\, \omega_m(A_m).$$

Observe that $\tilde{\pi}_f(B)$ is well defined since it does not depend on the representation of B. Indeed, representation of the same set B with m replaced by $m+1$ is given as

$$B \;=\; \left[\left(\bigcup_{i=1}^{m-1} B_i \right) \cup (A_m \cap Y_m) \right] \cup \left\{ y \in Y \,|\, y^{m+1} \in (A_m \setminus Y_m) \times \mathbb{R} \right\}.$$

Then

$$\begin{aligned} & \sum_{i=1}^{m-1} \omega_i(B_i) + \omega_m(A_m \cap Y_m) + \omega_{m+1}((A_m \setminus Y_m) \times \mathbb{R}) \\ = \; & \sum_{i=1}^{m-1} \omega_i(B_i) + \omega_m(A_m \cap Y_m) + \omega_m(A_m \setminus Y_m) \\ = \; & \sum_{i=1}^{m-1} \omega_i(B_i) + \omega_m(A_m). \end{aligned}$$

$\tilde{\pi}_f$ is an additive measure defined on cylidrical sets. Hence, $\tilde{\pi}_f$ can be uniquely extended to a σ-additive measure defined on the Borel sets of Y. As $\tilde{\pi}(Y) = \omega_1(W_1) = 1$, this is a probability measure.

Now, for any $B = \bigcup_{i=1}^{\infty} B_i$ where $B_i = B \cap Y_i$, we have

$$\tilde{\pi}_f(B) \;=\; \lim_{m\to\infty} \tilde{\pi}_f\left(\bigcup_{i=1}^{m} B_i\right) \;=\; \lim_{m\to\infty} \sum_{i=1}^{m} \omega_i(B_i) \;=\; \pi_f(B).$$

Thus $\pi_f = \tilde{\pi}_f$ and π_f is well defined. □

We note that π_f possesses the following property. Let $(y_1, \ldots, y_{m-1})$ be such that $(y_1, \ldots, y_j) \notin Y_j$, $1 \le j \le m-1$. Then the distribution of y_m given $(y_1, \ldots, y_{m-1})$ is Gaussian with mean $L_m(f; y_1, \ldots, y_{m-1})$ and variance $\sigma_m^2(y_1, \ldots, y_{m-1})$.

Clearly, nonadaptive information can be treated also as adaptive information. Then $Y = \mathbb{R}^n$, $L_i(\cdot; y_1, \ldots, y_{i-1}) = L_i$ and $\sigma_i^2(y_1, \ldots, y_{i-1}) = \sigma_i^2$ are independent of $y_1, \ldots, y_{i-1}$.

3.7.2 Adaption versus nonadaption, I

Let $\mathbb{N} = \{\mathbb{N}_y\}_{y\in Y}$ be an arbitrary adaptive information operator. We know from Section 2.7.2 that it is often possible to select $y^* \in Y$ in such a way that the worst case radius of nonadaptive information $\mathbb{N}_{y^*}$, $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}_{y^*})$, is not much larger than $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$; see Theorem 2.15. Our aim now is to show a similar result in the average case setting, for linear problems with Gaussian measures. That is, we assume that F is a separable Banach space, G is a separable Hilbert space, and the solution operator $S : F \to G$ is continuous linear. The measure μ is zero mean Gaussian with correlation operator $C_\mu : F^* \to F$.

Recall that for $y = (y_1, \ldots, y_n) \in Y$, the nonadaptive information $\mathbb{N}_y = \{N_y, \Sigma_y\}$ is given as

$$N_y \;=\; [\, L_{1,y}, L_{2,y}, \ldots, L_{n,y}\,]$$

and

$$\Sigma_y \;=\; [\, \sigma_{1,y}^2, \sigma_{2,y}^2, \ldots, \sigma_{n,y}^2\,]$$

where, for brevity, $L_{i,y} = L_i(\cdot; y_1, \ldots, y_{i-1})$ and $\sigma_{i,y}^2 = \sigma_i^2(y_1, \ldots, y_{i-1})$, $1 \le i \le n$. Recall also that μ_1 denotes the a priori distribution of information y in Y. Clearly, we have

$$\mu_1 \;=\; \int_F \pi_f(\cdot)\, \mu(df),$$

and μ_1 is in general not Gaussian, even when $Y = \mathbb{R}^n$. For any $f \in F$, the measure π_f is supported on $Y_{1,f} = \{\, y \in Y \mid y \in N_y(f) + \Sigma_y(\mathbb{R}^{n(y)})\,\}$. Hence,

μ_1 is supported on $Y_1 = \{\, y \in Y \mid y \in N_y(F_1) + \Sigma_y(\mathbb{R}^{n(y)})\}$ where $F_1 = \overline{C_\mu(F^*)} = \mathrm{supp}\ \mu$, or equivalently, $Y_1 = \{\, y \in Y \mid y \in (\Sigma_y + G_{N_y})(\mathbb{R}^{n(y)})\}$.

We need a theorem about the conditional distribution of a Gaussian measure with respect to adaptive information.

Theorem 3.8 *For adaptive information $\mathbb{N} = \{N_y, \Sigma_y\}_{y \in Y}$, the conditional distribution $\mu_2(\cdot|y)$, $y \in Y_1$, is Gaussian. Its mean element is given as*

$$m(y) \;=\; \sum_{j=1}^{n(y)} z_j (C_\mu L_{j,y}),$$

where z is the solution of $(\Sigma_y + G_{N_y})z = y$, $\Sigma_y = diag\{\sigma_{1,y}^2, \ldots, \sigma_{n(y),y}^2\}$, $G_{N_y} = \{\langle L_{i,y}, L_{j,y}\rangle_\mu\}_{i,j=1}^{n(y)}$, and $n(y)$ is the length of y. The correlation operator of $\mu_2(\cdot|y)$ is given as

$$C_{\mu_2,y}(L) \;=\; C_\mu(L) \;-\; m(\,N_y(C_\mu L)\,), \qquad L \in F^*.$$

Proof We first give a proof for adaptive information $\mathbb{N}$ with $Y = \mathbb{R}^n$.

Let $\tilde{F} = F \times \mathbb{R}^n$, and let $\tilde{\mu}$ be the joint probability on $\tilde{F}$,

$$\tilde{\mu}(B) \;=\; \int_F \pi_f(B^{(f)})\,\mu(df),$$

$B^{(f)} = \{\, y \in Y \mid (f,y) \in B\,\}$. For $B \subset \tilde{F}$, let χ_B be the characteristic function of B, i.e., $\chi_B(\tilde{f}) = 1$ for $\tilde{f} \in B$, and $\chi_B(\tilde{f}) = 0$ for $\tilde{f} \notin B$. We denote by $\mu_1(\cdot|\mathbb{M})$ the a priori distribution of information values with respect to (adaptive or nonadaptive) information $\mathbb{M}$, and by $\mu_2(\cdot|y,\mathbb{M})$ the conditional distribution on F given $y \in Y = Y(\mathbb{M})$.

Due to Theorem 3.2, we have to show that $\mu_2(\cdot|y,\mathbb{N}) = \mu_2(\cdot|y,\mathbb{N}_y)$. To this end, we shall use induction on n.

If $n = 1$ then any adaptive information is also nonadaptive and the proof is obvious.

Suppose $n > 1$. Let $\mathbb{N}^{n-1}$ be the adaptive information consisting of noisy evaluations of the $(n-1)$ first functionals from $\mathbb{N}$. For $y \in \mathbb{R}^n$, we write $y = (y^{n-1}, y_n)$ where $y^{n-1} \in \mathbb{R}^{n-1}$ and $y_n \in \mathbb{R}$. Then we have

$$\tilde{\mu}(B) \;=\; \int_F \int_{\mathbb{R}^n} \chi_B(f,y)\,\pi_f(dy)\,\mu(df)$$

$$
\begin{aligned}
&= \int_F \int_{\mathbb{R}^{n-1}} \chi_B(f,y)\, \mathcal{G}\left(dy_n \middle| L_{n,y^{n-1}}(f), \sigma^2_{y^{n-1}}\right)\, \omega_{n-1,f}(dy^{n-1})\, \mu(df) \\
&= \int_{\mathbb{R}^{n-1}} \int_F \int_{\mathbb{R}} \chi_B(f,y)\, \mathcal{G}\left(dy_n \middle| L_{n,y^{n-1}}(f), \sigma^2_{y^{n-1}}\right) \\
&\qquad\qquad \mu_2(df|y^{n-1}, \mathbb{N}^{n-1})\, \mu_1(dy^{n-1}, \mathbb{N}^{n-1}) \;=\; (*).
\end{aligned}
$$

Due to the inductive assumption and Theorem 3.2, $\mu_2(\cdot|y^{n-1}, \mathbb{N}^{n-1})$ can be interpreted as the conditional distribution on F with respect to the nonadaptive information $\mathbb{N}^{n-1}_{y^{n-1}}$. Hence, denoting by ρ the distribution of y_n given y^{n-1}, and using decomposition of $\mu_2(\cdot|y^{n-1}, N^{n-1})$ with respect to y_n, we have

$$
\int_F \int_{\mathbb{R}} \mathcal{G}\left(dy_n \middle| L_{n,y^{n-1}}(f), \sigma^2_{y^{n-1}}\right)\; \mu_2(df|y^{n-1}, \mathbb{N}^{n-1}) \;=\; \int_{\mathbb{R}} h(y)\, \rho(dy_n)
$$

where $h(y) = \int_F \chi_B(f,y)\, \mu_2(df|y, \mathbb{N}_y)$. As a consequence, we obtain

$$
\begin{aligned}
(*) \;&=\; \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} h(y)\, \rho(dy_n)\, \mu_1(dy^{n-1}|\mathbb{N}^{n-1}) \\
&=\; \int_{\mathbb{R}^{n-1}} \int_{\mathbb{R}} \int_F h(y)\, \mu_2(df|y, \mathbb{N}_y)\, \rho(dy_n)\, \mu_1(dy^{n-1}|\mathbb{N}^{n-1}) \\
&=\; \cdots \;=\; \int_F \int_{\mathbb{R}^n} h(y)\, \pi_f(dy)\, \mu(df) \;=\; \int_{\mathbb{R}^n} h(y)\, \mu_1(dy) \\
&=\; \int_{\mathbb{R}^n} \int_F \chi_B(f,y)\, \mu_2(df|y, \mathbb{N}_y)\, \mu_1(dy).
\end{aligned}
$$

On the other hand, $\tilde{\mu}(B) = \int_{\mathbb{R}^n} \int_F \chi_B(f,y) \mu_2(df|y, \mathbb{N}) \mu_1(dy)$. Thus

$$
\mu_2(\cdot|y, \mathbb{N}) \;=\; \mu_2(\cdot|y, N_y), \qquad \forall y \text{ a.e.},
$$

as claimed.

In the general case ($Y \neq \mathbb{R}^n$), we have

$$
\begin{aligned}
\tilde{\mu}(B) \;&=\; \int_F \int_Y \chi_B(f,y)\, \pi_f(dy)\, \mu(df) \\
&=\; \sum_{m=1}^{\infty} \int_F \int_{Y_m} \chi_B(f,y)\, \omega_{m,f}(dy)\, \mu(df) \\
&=\; \sum_{m=1}^{\infty} \int_{Y_m} \int_F \chi_B(f,y)\, \mu_2(df|y, \mathbb{N}_y)\, \mu_1(dy) \\
&=\; \int_Y \int_F \chi_B(f,y)\, \mu_2(df|y, \mathbb{N}_y)\, \mu_1(dy).
\end{aligned}
$$

The proof is complete. □

Thus, the conditional distribution $\mu_2(\cdot|y)$ with respect to adaptive information $\mathbb{N}$ is equal to the conditional distribution with respect to nonadaptive information operator $\mathbb{N}_y$ and the same y. This should be intuitively clear. Indeed, in both cases, the information y is obtained by using the same functionals $L_{i,y}$ and precisions $\sigma_{i,y}^2$.

From Theorem 3.8 we obtain almost immediately the following result corresponding to Theorem 2.15 of the worst case.

Theorem 3.9 *For any adaptive information* $\mathbb{N} = \{\mathbb{N}_y\}_{y\in Y}$*, there exists* $y^* \in Y$ *such that for the nonadaptive information* $\mathbb{N}_{y^*}$ *we have*

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_{y^*}) \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}).$$

Proof There exists $y^* \in Y_1$ such that

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;=\; \sqrt{\int_Y (\, r(\nu_2(\cdot|y))\,)^2\,\mu_1(dy)} \;\ge\; r(\nu_2(\cdot|y^*)\,) \tag{3.12}$$

where, as in Section 3.2, $\nu_2(\cdot|y) = \mu_2(S^{-1}(\cdot)|y)$ and $r(\nu_2(\cdot|y))$ is the radius of $\nu_2(\cdot|y)$. From Theorem 3.2 we know that the conditional measures $\mu_2(\cdot|y,\mathbb{N}_{y^*})$ have the same correlation operator. Hence, $r(\nu_2(\cdot|y,\mathbb{N}_{y^*})) = r(\nu_2(\cdot|y^*,\mathbb{N}_{y^*}))$ $\forall y$ a.e. This, Theorem 3.8, and (3.12) yield

$$\begin{aligned}\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_{y^*}) \;&=\; \sqrt{\int_{\mathbb{R}^{n(y^*)}} r(\nu_2(\cdot|y,\mathbb{N}_{y^*}))\,\mu_1(dy|\mathbb{N}_{y^*})} \;=\; r(\nu_2(\cdot|y^*,\mathbb{N}_{y^*}))\\ &=\; r(\nu_2(\cdot|y^*)) \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}). \quad \Box\end{aligned}$$

Note that Theorem 3.8 does not say anything about the construction of y^*. Actually, y^* can assume arbitrary values (compare with E.3.28). Thus the situation differs from that in the worst case where in the linear case we have $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}_0) \le 2\,\mathrm{rad}^{\mathrm{wor}}(\mathbb{N})$.

Notes and Remarks

NR 3.16 Adaptive information with fixed, but not necessarily Gaussian observation noise was studied by Kadane *et al.* [36]. They give examples that adaption can generally be much more powerful than nonadaption and show, under some additional assumptions, a result corresponding to Theorem 3.8; see also E 3.29.

NR 3.17 The adaptive information with varying noise in the average case setting was presented in Plaskota [83].

Exercises

E 3.27 Let $\mathbb{N} = \{N, \Sigma\}$ be such adaptive information that $\sigma_i^2(y_1, \ldots, y_{i-1}) > 0$, for all i and $y_1, \ldots, y_{i-1}$. Show that then the measure π_f is given as

$$\begin{aligned}\pi_f(B) \;&=\; \sum_{m=1}^{\infty} (2\,\pi)^{-m/2} \int_{B_m} (\sigma_1 \sigma_2(t_1) \cdots \sigma_m(t_1, \ldots, t_{m-1})\,)^{-1} \\ &\qquad \exp\left\{ -\frac{1}{2} \sum_{i=1}^{m} \frac{(\, t_i - L_i(f; t_1, \ldots, t_{i-1})\,)^2}{\sigma_i^2(t_1, \ldots, t_{i-1})} \right\} dt_m dt_{m-1} \ldots dt_1.\end{aligned}$$

E 3.28 Let $y \in \mathbb{R}^n$. Give an example of adaptive information $\mathbb{N}$ with $y \in Y$, such that
(a) y is the only element for which $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_y)$.
(b) $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) = 0$, but $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_y) > 0$.

E 3.29 Let (Kadane, Wasilkowski, Woźniakowski) $F = \mathbb{R}^2$ be equipped with the Euclidean norm and standard Gaussian measure μ, and let S be the identity in F. Consider adaptive information $\mathbb{N}$ with $Y = \mathbb{R}^n$, consisting of noisy observations of n adaptively chosen functionals L_i,

$$L_i(f) \;=\; L_i(f; y_1, \ldots, y_{i-1}) \;=\; \begin{cases} f_1 & y_1 = y_2 = \cdots = y_{i-1}, \\ f_2 & \text{otherwise}, \end{cases}$$

with noise x_i such that $x_i = -1$ or $x_i = 1$ with probability 1/2. Show that

$$\lim_{n \to +\infty} \frac{\inf_{y \in \mathbb{R}^n} \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_y)}{\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})} \;=\; +\infty.$$

3.8 Optimal information

In this section we study the minimal radius and optimal (nonadaptive) information. Recall that they are defined as follows. Let $\mathcal{N}_n$ be the class of exact nonadaptive information operators N consisting of n functionals from a given class Λ, $N = [L_1, \ldots, L_n]$, $L_i \in \Lambda$. Then the minimal (average) radius corresponding to a precision vector $\Sigma = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$ is given as

$$\mathrm{r}_n^{\mathrm{ave}}(\Sigma) \;=\; \inf_{N \in \mathcal{N}_n} \mathrm{rad}^{\mathrm{ave}}(N, \Sigma).$$

Information $N_\Sigma \in \mathcal{N}_n$ is optimal iff

$$\mathrm{r}_n^{\mathrm{ave}}(\Sigma) \;=\; \mathrm{rad}^{\mathrm{ave}}(N_\Sigma, \Sigma).$$

We shall consider a general problem with Gaussian measures and also function approximation and integration on the classical Wiener space.

3.8.1 Linear problems with Gaussian measures

We start with the general problem with a continuous linear solution operator $S : F \to G$ and a zero mean Gaussian measure μ on F. The class Λ of permissible information functionals consists of functionals whose μ–norm is bounded by 1,

$$\Lambda \;=\; \Lambda^{\mathrm{all}} \;=\; \left\{ L \in F^* \;\middle|\; \|L\|_\mu \;=\; \sqrt{L(C_\mu L)} \le 1 \right\}.$$

Observe that Λ^{all} can be equivalently defined as

$$\Lambda^{\mathrm{all}} \;=\; \left\{ L \in F^* \;\middle|\; \|L_H\|_H \;=\; \sup_{\|f\|_H=1} |L(f)| \le 1 \right\}$$

where H is the associated with μ Hilbert space, so that $\{H, \overline{C_\mu(F)}\}$ is an abstract Wiener space. The precision vector is $\Sigma = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$ where, without loss of generality,

$$0 \;=\; \sigma_1^2 \;=\; \cdots \;=\; \sigma_{n_0}^2 \;<\; \sigma_{n_0+1}^2 \;\le\; \cdots \;\le\; \sigma_n^2$$

(if all σ_i's are nonzero then $n_0 = 0$).

We shall see that the method of finding optimal information is in this case similar to that used in Section 2.8.1, where the problem of optimal information in the worst case for a compact solution operator and noise bounded in the weighted Euclidean norm was studied.

Let $\nu = \mu S^{-1}$ be the a priori distribution on the space G, induced by the measure μ and the operator S. Then ν is zero mean Gaussian with correlation operator $C_\nu = S C_\mu S^* : G \to G$ where $S^* : G \to F^*$ is the adjoint operator to S, $S^* g = \langle S(\cdot), g\rangle \;\forall g \in G$. Moreover, C_ν is self-adjoint, nonnegative definite, and has finite trace.

Let $\{\xi_i\}_{i=1}^{\dim G} \subset G$ be the complete and orthonormal system of eigenelements of C_ν. Let $\lambda_1 \ge \lambda_2 \ge \lambda_3 \ge \cdots \ge 0$ be the corresponding eigenvalues,

$C_\nu\xi_i = \lambda_i\xi_i$. We consider the sequence $\{\lambda_i\}$ to be infinite by setting, if necessary, $\lambda_i = 0$ for $i > \dim G$. For $\lambda_i > 0$, define the functionals

$$K_i^* \;=\; \lambda_i^{-1/2} S^*\xi_i \;=\; \lambda_i^{-1/2}\langle S(\cdot), \xi_i\rangle.$$

(For $\lambda_i = 0$ we formally set $K_i^* = 0$.) Clearly, the functionals K_i^* are μ-orthonormal,

$$\begin{aligned}\langle K_i^*, K_j^*\rangle_\mu &= (\lambda_i\lambda_j)^{-1/2}\langle S_H^*\xi_i, S_H^*\xi_j\rangle_H \\ &= (\lambda_i\lambda_j)^{-1/2}\langle S_H S_H^*\xi_i, \xi_j\rangle \;=\; \delta_{ij}.\end{aligned}$$

Since $C_\nu = S_H S_H^*$, λ_i's are also the dominating eigenvalues of the compact operator $S_H^* S_H : H \to H$, and the corresponding orthonormal in H eigenelements are $\xi_{H,i} = C_\mu K_i^* \in H$, $i \geq 1$.

Recall that in the worst case setting the problem of optimal information was related to some minimization problem. The corresponding problem in the average case is as follows:

Problem (MP) *Minimize*

$$\Omega(\eta_{n_0+1}, \ldots, \eta_n) \;=\; \sum_{i=n_0+1}^{n} \frac{\lambda_i}{1+\eta_i} \tag{3.13}$$

over all $\eta_{n_0+1} \geq \cdots \geq \eta_n \geq 0$ *satisfying*

$$\sum_{i=r}^{n} \eta_i \;\leq\; \sum_{i=r}^{n} \frac{1}{\sigma_i^2}, \qquad n_0 + 1 \leq r \leq n, \tag{3.14}$$

and $\sum_{i=n_0+1}^{n} \eta_i = \sum_{i=n_0+1}^{n} \sigma_i^{-2}$.

Theorem 3.10 *Let* $\eta_{n_0+1}^* \geq \cdots \geq \eta_n^*$ *be the solution of* (MP). *Then*

$$\mathrm{r}_n^{\mathrm{ave}}(\Sigma) \;=\; \sqrt{\Omega(\eta_{n_0+1}^*, \ldots, \eta_n^*) \;+\; \sum_{i=n+1}^{\infty} \lambda_i}\,.$$

Furthermore, the optimal information is given as

$$N_\Sigma \;=\; [\,K_1^*, \ldots, K_{n_0}^*, L_{n_0+1}^*, \ldots, L_n^*\,],$$

where

$$L^*_{n_0+i} \;=\; \sigma_{n_0+i} \sum_{j=1}^{n-n_0} w_{ij} K^*_{n_0+j},$$

and $W^* = \{w_{ij}\}_{i,j=1}^{n-n_0}$ *is the matrix from Lemma 2.14 applied for*

$$\eta_i \;=\; \eta^*_{n_0+i} \qquad \textit{and} \qquad \beta_i \;=\; \frac{1}{\sigma^2_{n_0+i}}\,,$$

$1 \le i \le n - n_0\,.$

Proof Assume first that all σ_i^2 are positive, $n_0 = 0$. Let $N \;=\; [\,L_1, \ldots, L_n\,]$ be an arbitrary information from $\mathcal{N}_n$. We can assume that $\|L_i\|_\mu \;=\; 1$, $1 \le i \le n$.

We start with the lower bound on $\mathrm{rad}^{\mathrm{ave}}(N, \Sigma)$. Let the matrix

$$G \;=\; \Sigma^{-1/2} G_N \Sigma^{-1/2} \;=\; \{(\sigma_i\sigma_j)^{-1}\langle L_i, L_j\rangle_\mu\}_{i,j=1}^n,$$

and let $\{q^{(i)}\}_{i=1}^n$ be the orthonormal basis of eigenvectors of G, $Gq^{(i)} = \eta_i q^{(i)}$ where $\eta_1 \;\ge\; \cdots \;\ge\; \eta_m \;>\; 0 \;=\; \eta_{m+1} \;=\; \cdots \;=\; \eta_n$. We know from Section 3.4.2 that the radius of $\mathbb{N} = \{N, \Sigma\}$ is equal to the radius of information $\mathbb{M} = \{M, \Sigma'\}$ where M consists of m μ–orthonormal functionals K_i,

$$K_i \;=\; \frac{1}{\eta_i} \sum_{j=1}^n \frac{q_j^{(i)}}{\sigma_j}\, L_j, \qquad 1 \le i \le m,$$

and $\Sigma' = \mathrm{diag}\{\eta_1^{-1}, \ldots, \eta_m^{-1}\}$. That is,

$$(\mathrm{rad}^{\mathrm{ave}}(N, \Sigma)\,)^2 \;=\; (\mathrm{rad}^{\mathrm{ave}}(M, \Sigma')\,)^2 \;=\; \mathrm{trace}(SC_\mu S^*) \;-\; \sum_{i=1}^m \frac{\|SC_\mu K_i\|^2}{1 + \eta_i^{-1}}.$$

It is a well known fact that for any orthonormal in H elements f_i, $1 \le i \le k$, it holds $\sum_{i=1}^k \langle S_H^* S_H f_i, f_i\rangle_H \;\le\; \sum_{i=1}^k \lambda_i$. Since for $f_i \;=\; C_\mu K_i$ is $\langle f_i, f_j\rangle_H = \delta_{ij}$, we have

$$\sum_{i=1}^k \|S(C_\mu K_i)\|^2 \;=\; \sum_{i=1}^k \|S_H f_i\|^2 \;=\; \sum_{i=1}^k \langle S_H^* S_H f_i, f_i\rangle_H \;\le\; \sum_{i=1}^k \lambda_i.$$

This and $\eta_1 \ge \cdots \ge \eta_m$ yield

$$\sum_{i=1}^m \frac{\eta_i}{1 + \eta_i} \|S(C_\mu K_i)\|^2 \;\le\; \sum_{i=1}^n \frac{\eta_i}{1 + \eta_i} \lambda_i$$

which implies the following lower bound on $\mathrm{rad}^{\mathrm{ave}}(N,\Sigma)$:

$$(\mathrm{rad}^{\mathrm{ave}}(N,\Sigma))^2 \;\ge\; \sum_{i=1}^{\infty}\lambda_i \,-\, \sum_{j=1}^{n}\frac{\eta_j}{1+\eta_j}\,\lambda_j \;=\; \Omega(\eta_1,\ldots,\eta_n) \,+\, \sum_{j=n+1}^{\infty}\lambda_j\,.$$

Observe that for all $1\le r\le n$ we also have

$$\sum_{i=r}^{n}\eta_i \;\le\; \sum_{i=r}^{n}\langle Ge_i,e_i\rangle_2 \;=\; \sum_{i=r}^{n}\frac{1}{\sigma_i^2},$$

(e_i stands for the ith versor), and $\sum_{i=1}^n \eta_i = \sum_{i=1}^n \sigma_i^{-2}$. Thus we finally obtain

$$\mathrm{r}_n^{\mathrm{ave}}(\Sigma) \;\ge\; \sqrt{\Omega(\eta_1^*,\ldots,\eta_n^*) \,+\, \sum_{i=n+1}^{\infty}\lambda_i\,}\,. \tag{3.15}$$

We now show that $\mathrm{rad}^{\mathrm{ave}}(N_\Sigma,\Sigma)$ is equal to the right hand side of (3.15). Indeed, since

$$\langle L_i^*,L_j^*\rangle_\mu \;=\; \left\langle \sigma_i\sum_{s=1}^{n}w_{is}K_s^*,\sigma_j\sum_{t=1}^{n}w_{jt}K_t^*\right\rangle_\mu \;=\; \sigma_i\sigma_j\sum_{s=1}^{n}w_{ij}w_{js}, \tag{3.16}$$

the corresponding matrix $G=WW^*$, the eigenvectors $q^{(i)}$ of G are the columns $w^{(i)}$ of W, and $Gw^{(i)} = \eta_i^* w^{(i)}$, $1\le i\le n$. Furthermore, for $1\le i\le m$, the functionals K_i equal

$$\begin{aligned} K_i \;&=\; \frac{1}{\eta_i}\sum_{s=1}^{n}q_s^{(i)}\sigma_s^{-1}\left(\sigma_s\sum_{j=1}^{n}w_{sj}K_j^*\right)\\ &=\; \frac{1}{\eta_i}\sum_{s=1}^{n}w_{si}\sum_{j=1}^{n}w_{sj}K_j^* \;=\; \frac{1}{\eta_i}\sum_{j=1}^{n}K_j^*\sum_{s=1}^{n}w_{si}w_{sj}\\ &=\; \frac{1}{\eta_i}\eta_iK_i^* \;=\; K_i^*. \end{aligned}$$

Hence,

$$\begin{aligned} \mathrm{rad}^{\mathrm{ave}}(N_\Sigma,\Sigma) \;&=\; \sqrt{\sum_{i=1}^{\infty}\lambda_i \,-\, \sum_{j=1}^{m}\frac{\eta_i^*}{1+\eta_i^*}\|S(C_\mu K_j)\|^2}\\ &=\; \sqrt{\sum_{i=1}^{\infty}\lambda_i \,-\, \sum_{j=1}^{n}\frac{\eta_j^*}{1+\eta_j^*}\lambda_j} \;=\; \sqrt{\Omega(\eta_1^*,\ldots,\eta_n^*) \,+\, \sum_{j=n+1}^{\infty}\lambda_j}. \end{aligned}$$

Since, due to (3.16), we also have $\|L_i^*\|_\mu = \sigma_i^2 \sum_{s=1}^n w_{is}^2 = 1$, information N_Σ is in $\mathcal{N}_n$. This completes the proof of the case $n_0 = 0$.

Suppose now that $n_0 \geq 1$. Then $N = [N^{(0)}, N^{(1)}]$ and $\Sigma = [\Sigma^{(0)}, \Sigma^{(1)}]$, where $N^{(0)} = [L_1, \dots, L_{n_0}]$, $N^{(1)} = [L_{n_0+1}, \dots, L_n]$, and $\Sigma^{(0)} = [\sigma_1^2, \dots, \sigma_{n_0}^2]$, $\Sigma^{(1)} = [\sigma_{n_0+1}^2, \dots, \sigma_n^2]$. The a posteriori Gaussian measure on F with respect to information $N^{(0)}$ (which is exact, $\Sigma^{(0)} = 0$) has the correlation operator $C_{\mu,N^{(0)}} = C_\mu(I - P_{N^{(0)}})$, where $P_{N^{(0)}} : F^* \to F^*$ is the μ–orthogonal projection onto span$\{L_1, \dots, L_{n_0}\}$. For the dominating eigenvalues $\tilde{\lambda}_i$ of $SC_{\mu,N^{(0)}}S^*$, which is the correlation operator of the a posteriori measure on G with respect to $N^{(0)}$, we have $\tilde{\lambda}_i \geq \lambda_{n_0+i}$, $\forall i \geq 1$. Moreover, if $N^{(0)} = N^* = [K_1^*, \dots, K_{n_0}^*]$ then $\tilde{\lambda}_i = \lambda_{n_0+i}$, and the corresponding eigenelements are $\tilde{\xi}_i = \xi_{n_0+i}$, $\forall i \geq 1$. Hence, we obtain the desired result by reducing our problem to that of finding optimal $N^{(1)}$, where the precision is $\Sigma^{(1)}$ and the a priori distribution on F is Gaussian with correlation operator $C_\mu(I - P_{N^*})$. □

We now give an explicit formula for the solution of the minimization problem (MP) as well as for the minimal radius $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$. For $n_0 \leq q < r \leq n$, define the following auxiliary minimization problem

Problem $\mathrm{P(q,r)}$ *Minimize*

$$\Omega_{qr}(\eta_{q+1}, \dots, \eta_r) \;=\; \sum_{j=q+1}^{r} \frac{\lambda_j}{1 + \eta_j}$$

over all $\eta_{q+1} \geq \cdots \geq \eta_r \geq 0$ *satisfying* $\sum_{i=q+1}^r \eta_i = \sum_{i=q+1}^r \sigma_i^{-2}$.

The solution $\eta^* = (\eta_{q+1}^*, \dots, \eta_r^*)$ of $\mathrm{P(q,r)}$ is as follows. Let $k = k(q,r)$ be the largest integer satisfying $q + 1 \leq k \leq r$ and

$$\frac{\sum_{j=q+1}^k \lambda_j^{1/2}}{\sum_{j=q+1}^r \sigma_j^{-2} + (k - q)} \;\leq\; \lambda_k^{1/2}. \tag{3.17}$$

Then

$$\eta_i^* \;=\; \frac{\sum_{j=q+1}^r \sigma_j^{-2} + (k - q)}{\sum_{j=q+1}^k \lambda_j^{1/2}} \cdot \lambda_i^{1/2} \;-\; 1 \quad \text{for } q + 1 \leq i \leq k, \tag{3.18}$$

and $\eta_i^* = 0$ for $k+1 \le i \le r$. Furthermore,

$$\Omega_{qr}(\eta^*) \;=\; \frac{\left(\sum_{j=q+1}^{k} \lambda_j^{1/2}\right)^2}{\sum_{j=q+1}^{r} \sigma_j^{-2} + (k-q)} \;+\; \sum_{j=k+1}^{r} \lambda_j \,.$$

We shall say that the solution $\eta^* = (\eta_{q+1}, \dots, \eta_r^*)$ of P(q, r) is *acceptable* iff

$$\sum_{j=s}^{r} \eta_j^* \;\le\; \sum_{j=s}^{r} \frac{1}{\sigma_j^2}\,, \qquad \text{for all } q+1 \le s \le r\,.$$

Let the number p, $0 \le p < n$, and the sequence $0 \le n_0 < n_1 < \cdots < n_p < n_{p+1} = n$ be defined (uniquely) by the condition

$$n_i \;=\; \min\{\, s \ge n_0 \,|\, \text{ solution of } (\mathrm{P(s, n_{i+1})} \text{ is acceptable } \}, \qquad (3.19)$$

for all $0 \le i \le p$.

Theorem 3.11 *Let p and the sequence $n_0 < n_1 < \cdots < n_{p+1} = n$ be defined by (3.19). Then the optimal η^* is given as*

$$\eta^* \;=\; (\eta^{(0)}, \eta^{(1)}, \dots, \eta^{(p)})$$

where $\eta^{(i)} = (\eta_{n_i+1}^*, \dots, \eta_{n_{i+1}}^*)$ is the solution of $\mathrm{P(n_i, n_{i+1})}$, $0 \le i \le p$.

Proof Let $t = \max\{n_0 + 1 \le i \le n \,|\, \eta_i^* > 0\}$. For $n_0 + 1 \le i \le t$, the function

$$\psi_i(\tau) = \Omega(\eta_{n_0+1}^*, \dots, \eta_{i-2}^*, \eta_{i-1}^* + \eta_i^* - \tau, \tau, \eta_{i+1}^*, \dots, \eta_n^*)$$

is continuous, convex, and attains the minimum at τ_0 such that $\lambda_{i-1}(1 + \eta_{i-1}^* + \eta_i^* - \tau_0)^{-2} = \lambda_i (1 + \tau_0)^{-2}$. From this and from the definition of (MP) it follows that

$$\frac{\lambda_{i-1}}{(1 + \eta_{i-1}^*)^2} \;\le\; \frac{\lambda_i}{(1 + \eta_i^*)^2}\,. \qquad (3.20)$$

Moreover, if $\lambda_{i-1}(1 + \eta_{i-1}^*)^{-2} < \lambda_i (1 + \eta_i^*)^{-2}$ then $\sum_{j=i}^{n} \eta_j^* = \sum_{j=i}^{n} \sigma^{-2}$. If $t < n$ then, using the same argument with $i = t+1$, we find that

$$\frac{\lambda_t}{(1 + \eta_t^*)^2} \ge \lambda_{t+1}. \qquad (3.21)$$

Let $m_1 < \cdots < m_s$ be the sequence of all indices i, $n_0 < i < t$, for which $\lambda_i (1 + \eta_i^*)^{-2} < \lambda_{i+1} (1 + \eta_{i+1}^*)^{-2}$. Set $m_0 = n_0$ and $m_{s+1} = n$. From (3.20) it follows that $\sum_{j=m_i+1}^{m_{i+1}} \eta_j^* = \sum_{j=m_i+1}^{m_{i+1}} \sigma_j^{-2}$, $0 \le i \le s$. This and (3.21) yield that the numbers $\eta^*_{m_i+1}, \dots, \eta^*_{m_{i+1}}$ are the solution of $(\mathrm{P}(\mathrm{m_i}, \mathrm{m_{i+1}}))$ for all $0 \le i \le s$. To complete the proof, it is now enough to show that the sequences $\{m_i\}_{i=0}^{s+1}$ and $\{n_i\}_{i=0}^{p+1}$ are the same, i.e., $\{m_i\}_{i=0}^{q+1}$ satisfies (3.19). Indeed, suppose to the contrary that for some i there is j_0, $0 \le j_0 < m_i$, such that the solution $\tilde\eta^*_{j_0+1}, \dots, \tilde\eta^*_{m_{i+1}}$ of $(\mathrm{P}(\mathrm{j_0}, \mathrm{m_{i+1}}))$ is acceptable. Then

$$\sum_{j=m_i+1}^{m_{i+1}} \tilde\eta_j^* \;\le\; \sum_{j=m_i+1}^{m_{i+1}} \frac{1}{\sigma_j^2} \;=\; \sum_{j=m_i+1}^{m_{i+1}} \eta_j^* \,.$$

From this and the formulas (3.17), (3.18) we get $\tilde\eta_j^* \le \eta_j^*$ for all $m_i + 1 \le j \le m_{i+1}$. Similarly, for $j_0 \le j \le m_i$ we have

$$\frac{\lambda_j}{(1+\eta_j^*)^2} \;<\; \frac{\lambda_{m_i+1}}{(1+\eta^*_{m_i+1})^2} \;\le\; \frac{\lambda_{m_i+1}}{(1+\tilde\eta^*_{m_i+1})^2} \;=\; \frac{\lambda_j}{(1+\tilde\eta^*_j)^2}\,,$$

and consequently $\tilde\eta_j^* < \eta_j^*$. Hence,

$$\sum_{j=j_0+1}^{m_{i+1}} \frac{1}{\sigma_j^2} \;=\; \sum_{j=j_0+1}^{m_{i+1}} \tilde\eta_j^* \;<\; \sum_{j=j_0+1}^{m_{i+1}} \eta_j^*\,,$$

which is a contradiction. □

Knowing optimal η^*, we can write an explicit formula for $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$.

Corollary 3.4 *Let p and the sequence $\{n_i\}_{i=0}^{p+1}$ be defined by (3.19), and let $k = k(n_p, n)$ be given by (3.17). Then the minimal radius $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$ equals*

$$\sqrt{\sum_{i=0}^{p-1} \frac{\left(\sum_{j=n_i+1}^{n_{i+1}} \lambda_j^{1/2}\right)^2}{\sum_{j=n_i+1}^{n_{i+1}} \sigma_j^{-2} + (n_{i+1}-n_i)} + \frac{\left(\sum_{j=n_p+1}^{k} \lambda_j^{1/2}\right)^2}{\sum_{j=n_p+1}^{n} \sigma_j^{-2} + (k-n_p)} + \sum_{j=k+1}^{\infty} \lambda_j}\,.$$

□

As we see, the formula for the minimal radius given in terms of the eigenvalues λ_i and precisions σ_i, $1 \le i \le n$, is rather complicated. Let, for simplicity, all σ_i's be nonzero. Then we have the following bounds on $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$:

$$\sqrt{\frac{\left(\sum_{i=1}^{k} \lambda_i^{1/2}\right)^2}{\sum_{i=1}^{n} \sigma_i^{-2} + k} + \sum_{i=k+1}^{\infty} \lambda_i} \;\le\; \mathrm{r}_n^{\mathrm{ave}}(\Sigma) \;\le\; \sqrt{\sum_{i=1}^{n} \frac{\lambda_i}{\sigma_i^{-2}+1} + \sum_{i=n+1}^{\infty} \lambda_i}, \tag{3.22}$$

where k is the largest integer satisfying $1 \le k \le n$ and

$$\frac{\sum_{j=1}^{k} \lambda_j^{1/2}}{\sum_{j=1}^{n} \sigma_j^{-2} + k} \;\le\; \lambda_k^{1/2}\,.$$

Clearly, we always have $\sqrt{\sum_{i=n+1}^{\infty} \lambda_i} \le \mathrm{r}_n^{\mathrm{ave}}(\Sigma) \le \sqrt{\sum_{i=1}^{\infty} \lambda_i}$.

Let us see what happens when all n observations are performed with the same precisions, $\sigma_i^2 = \sigma^2 > 0$, $\forall i$. It is easy to verify that then

$$\mathrm{r}_n^{\mathrm{ave}}(\Sigma) \;=\; \mathrm{r}_n^{\mathrm{ave}}(\sigma^2) \;=\; \sqrt{\sigma^2 \cdot \frac{\left(\sum_{i=1}^{k} \lambda_i^{1/2}\right)^2}{n + \sigma^2\, k} \;+\; \sum_{j=1}^{\infty} \lambda_j}, \tag{3.23}$$

where $k = k(\sigma^2, n)$ is the largest integer satisfying $1 \le k \le n$ and

$$\sigma^2 \cdot \frac{\sum_{j=1}^{k} \lambda_j^{1/2}}{n + \sigma^2\, k} \;\le\; \lambda_k^{1/2}\,.$$

The optimal information $N_{n,\sigma} = [L_1^*, \dots, L_n^*]$ is given by Theorem 3.10 with

$$\eta_i^* \;=\; \frac{n\,\sigma^{-2} + k}{\sum_{j=1}^{k} \lambda_j^{1/2}} \cdot \lambda_i^{1/2} \;-\; 1, \qquad 1 \le i \le k,$$

and $\eta_i^* = 0$ for $k+1 \le i \le n$. The optimal algorithm is the smoothing spline (or regularized) algorithm with $\gamma = \sigma^2$.

Let us look at the behavior of $\mathrm{r}_n^{\mathrm{ave}}(\sigma^2)$. Suppose first that $\sigma^2 \to 0^+$. Then, for $\mathrm{r}_n^{\mathrm{ave}}(0) = \sqrt{\sum_{i=n+1}^{\infty} \lambda_i} > 0$, we have

$$\mathrm{r}_n^{\mathrm{ave}}(\sigma^2) \;-\; \mathrm{r}_n^{\mathrm{ave}}(0) \;\approx\; \frac{\sigma^2\,\left(\sum_{i=1}^{n} \lambda_i^{1/2}\right)^2}{2\,n\,(1+\sigma^2)\sqrt{\sum_{j=1}^{\infty} \lambda_j}},$$

while for $\mathrm{r}_n^{\mathrm{ave}}(0) = 0$ we have

$$\mathrm{r}_n^{\mathrm{ave}}(\sigma^2) \;-\; \mathrm{r}_n^{\mathrm{ave}}(0) \;\approx\; \frac{\sigma\,\sum_{i=1}^{n} \lambda_i^{1/2}}{\sqrt{n + \sigma^2\, m}}$$

where m is the largest integer such that $\lambda_m > 0$.

For fixed σ^2 and $n \to +\infty$, the radius converges to zero, but not faster than $\sigma/\sqrt{n}$. Suppose that the eigenvalues λ_j satisfy

$$\lambda_j \asymp \left(\frac{\ln^s j}{j}\right)^p, \qquad \text{as} \quad j \to +\infty, \tag{3.24}$$

where $p > 1$ and $s \geq 0$. Providing some calculations we find that for $\sigma^2 > 0$

$$\left(\mathrm{r}_n^{\mathrm{ave}}(\sigma^2)\right)^2 \asymp \begin{cases} \sigma^2 \frac{\ln^{sp} n}{n^{p-1}} & 1 < p < 2, \\ \sigma^2 \frac{\ln^{2(s+1)} n}{n} & p = 2, \\ \sigma^2 \frac{1}{n} & p > 2, \end{cases}$$

where the constants in the "$\asymp$" notation do not depend on σ^2. For exact information we have

$$(\mathrm{r}_n^{\mathrm{ave}}(0))^2 \asymp \frac{\ln^{sp} n}{n^{p-1}}.$$

Hence, for $1 < p < 2$ the radius of noisy information behaves as $\mathrm{r}_n^{\mathrm{ave}}(0)$, while for $p > 2$ it achieves the best possible rate of convergence $\sigma/\sqrt{n}$.

Note that the eigenvalues (3.24) correspond to the function approximation in $\mathcal{L}_2((0,1)^d)$ with respect to the Wiener sheet measure, see NR 3.20.

We now devote some attention to the already mentioned relations between the optimal information problem considered in this section and the optimal information problem of Section 2.8.1.

Consider the pair of problems defined as in Section 3.6.3. That is, assume that $\{H, F\}$ is an abstract Wiener space and μ is the associated with it zero mean Gaussian measure. We wish to approximate values $S(f)$ of a continuous linear operator $S : F \to G$, based on noisy information $y = N(f)+x$ where $N = [L_1, \ldots, L_n]$ and the functionals $\|L_i\|_\mu = \|(L_i)_H\|_H \leq 1$. We know from Theorem 3.7 that for fixed N the optimal algorithms in the worst and average cases are (almost) the same. We now want to see whether a similar result holds with respect to optimal information. More precisely, suppose we want to choose information $N \in \mathcal{N}_n$ and algorithm φ as to minimize:

P1: Worst case error $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi)$ over the class $E = \{\, f \in H \mid \|f\|_H \leq 1\}$ and noise $\|x\|_Y = \sqrt{\langle \Sigma^{-1}x, x\rangle_2} \leq \delta$,

P2: Average case error $\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi)$ over the Gaussian measure μ and noise $x \sim \mathcal{N}(0, \sigma^2\Sigma)$,

where Σ is an $n \times n$ diagonal matrix.

Observe first that in both problems the minimal radii are determined by n dominating eigenvalues of the operator $S_H^* S_H$, and the optimal functionals L_i^* are linear combinations of the corresponding functionals K_i^*, $1 \le i \le n$. Furthermore, in the case $\Sigma = I$ and $\delta^2 = \sigma^2 > 0$, the radii $\mathrm{r}_n^{\mathrm{wor}}(\delta)$ and $\mathrm{r}_n^{\mathrm{ave}}(\sigma^2)$ decrease to zero as $n \to +\infty$, but convergence cannot be faster than $n^{-1/2}$.

It turns out that even stronger correspondence between both problems holds, similar to that of Theorem 3.7. Namely, assume additionally that $\delta^2 = \sigma^2$ and that $0 = \gamma_1^2 = \cdots = \gamma_{n_0}^2 < \gamma_{n_0+1}^2 \le \cdots \le \gamma_n^2$ are the diagonal elements of the matrix Σ. Let α^w and $\eta_{n_0+1}^w \ge \cdots \ge \eta_n^w \ge \eta_{n+1}^w = 0$ be the solution of the minimization problem (MP) of Section 2.8.1 with $\delta_i = \gamma_i$, and let $\eta_1^a \ge \cdots \ge \eta_n^a \ge 0$ be the solution of the minimization problem (MP) of the present section with $\sigma_i = \gamma_i$. Next, let information N^* be given as in Theorem 3.10 (or in Theorem 2.16) with $\sigma_i^2 = \gamma_i^2/2$ (or $\delta_i = \gamma_i/\sqrt{2}$) and $\eta_i^* = (\eta_i^w + \eta_i^a)/2$, $n_0 + 1 \le i \le n$. Observe that N^* is well defined since for any $n_0 + 1 \le r \le n$ we have

$$\sum_{i=r}^{n} \eta_i^* \;=\; \sum_{i=r}^{n} (\eta_i^w + \eta_i^a)/2 \;\le\; \sum_{i=r}^{n} 1/\gamma_i^2,$$

and the assumptions of Lemma 2.14 are satisfied. Also, $N^* \in \mathcal{N}_n$. We have the following theorem.

Theorem 3.12 *For information N^* and the spline algorithm φ_{spl} we have*

$$\mathrm{e}^{\mathrm{wor}}\left(\{N^*, \Delta\}, \varphi_{\mathrm{spl}}\right) \;\le\; 2 \cdot \mathrm{r}_n^{\mathrm{wor}}(\Delta)$$

and

$$\mathrm{e}^{\mathrm{ave}}\left(\{N^*, \Sigma\}, \varphi_{\mathrm{spl}}\right) \;\le\; \sqrt{2} \cdot \mathrm{r}_n^{\mathrm{ave}}(\Sigma)$$

where $\Delta = [\gamma_1, \ldots, \gamma_n]$ *and* $\Sigma = [\gamma_1^2, \ldots, \gamma_n^2]$.

Proof Indeed, the formulas for the worst and average case errors of the algorithm φ_{spl} using information N^* can be obtained as in the proofs of Theorem 2.16 and Theorem 3.10, respectively. Hence,

$$\begin{aligned}
\left(\mathrm{e}^{\mathrm{wor}}\left(\{N^*, \Delta\}, \varphi_{\mathrm{spl}}\right)\right)^2 \;&\le\; \max_{n_0+1 \le i \le n+1} \frac{2\,\lambda_i}{1 + \eta_i^*} \;\le\; \max_{n_0+1 \le i \le n+1} \frac{2\,\lambda_i}{1 + \eta_i^w/2} \\
&\le\; 4 \cdot \max_{n_0+1 \le i \le n+1} \frac{\lambda_i}{\alpha^w + (1 - \alpha^w)\eta_i^w} \\
&=\; 4 \cdot (\mathrm{r}_n^{\mathrm{wor}}(\Delta))^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
(\mathrm{e}^{\mathrm{ave}}(\{N^*,\Sigma\},\varphi_{\mathrm{spl}}))^2 &= \sum_{i=n_0+1}^{n} \frac{\lambda_i}{1+\eta_i^*} + \sum_{j=n+1}^{\infty} \lambda_j \\
&\le \sum_{i=n_0+1}^{n} \frac{\lambda_i}{1+\eta_i^a/2} + \sum_{j=n+1}^{\infty} \lambda_j \\
&\le 2\cdot\left(\sum_{i=n_0+1}^{n} \frac{\lambda_i}{1+\eta_i^a} + \sum_{j=n+1}^{\infty} \lambda_j\right) \\
&= 2\cdot(\mathrm{r}_n^{\mathrm{ave}}(\Sigma))^2,
\end{aligned}$$

as claimed. □

We stress that Theorem 3.12 does not say anything about a correspondence between $\mathrm{r}_n^{\mathrm{wor}}(\Delta)$ and $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$. Due to Theorem 3.7, we have $\mathrm{r}_n^{\mathrm{wor}}(\Delta) \le \sqrt{2}\,\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$. However, the ratio $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)/\mathrm{r}_n^{\mathrm{wor}}(\Delta)$ can be arbitrarily large. For instance, consider $\Delta = [\underbrace{\delta,\ldots,\delta}_{n}]$, $\Sigma = [\underbrace{\sigma^2,\ldots,\sigma^2}_{n}]$, and the eigenvalues as in (3.24). Then, using results of this and Section 2.8.1, for $\sigma^2 = \delta^2 > 0$ we obtain

$$\frac{\mathrm{r}_n^{\mathrm{ave}}(\sigma^2)}{\mathrm{r}_n^{\mathrm{wor}}(\delta)} \asymp \begin{cases} n^{1-p/2}\ln^{sp/2} n & 1<p<2, \\ \ln^{s+1} n & p=2, \\ 1 & p>2, \end{cases}$$

as $n \to +\infty$. For exact information we have $\mathrm{r}_n^{\mathrm{ave}}(0)/\mathrm{r}_n^{\mathrm{wor}}(0) \asymp \sqrt{n}$.

3.8.2 Approximation and integration on the Wiener space

In this section, we study optimal information for approximation and integration of continuous scalar functions, based on noisy observations at n points. More precisely, we let F to be the space of functions defined as

$$F \;=\; \mathbf{C}^0 \;=\; \{\, f:[0,1]\to\mathbb{R} \mid \quad f\text{–continuous}, f(0)=0\,\},$$

equipped with the classical Wiener measure w. Recall that w is uniquely determined by its mean element zero and the covariance kernel

$$R(s,t) \;=\; \int_{\mathbf{C}^0} f(s)f(t)\,w(df) \;=\; \min\{s,t\}, \qquad 0\le s,t\le 1.$$

The solution operator corresponding to the function approximation is given as

$$\mathrm{App}:\ \mathbf{C}^0 \to \mathcal{L}_2(0,1), \qquad \mathrm{App}(f) = f, \quad \forall f \in \mathbf{C}^0,$$

while the integration operator is defined as

$$\mathrm{Int}:\ \mathbf{C}^0 \to \mathbb{R}, \qquad \mathrm{Int}(f) = \int_0^1 f(x)\,dx, \quad \forall f \in \mathbf{C}^0.$$

We assume that $\Lambda = \Lambda^{\mathrm{std}}$. That is, information about $f \in \mathbf{C}^0$ is supplied by n noisy values of f at arbitrary points from $[0,1]$,

$$N(f) \;=\; [\, f(t_1), f(t_2), \ldots, f(t_n)\,] \tag{3.25}$$

where $0 = t_0 \le t_1 \le t_2 \le \cdots \le t_n \le 1$. The information noise is assumed to be white, $\Sigma = [\underbrace{\sigma^2, \ldots, \sigma^2}_{n}]$.

We start with a remark about the radius of information. Let $S \in \{\mathrm{App}, \mathrm{Int}\}$, the variance σ^2, and information operator N of the form (3.25) be given. We know that the radius $\mathrm{rad}^{\mathrm{ave}}(S, N, \sigma^2)$ is attained by the algorithm $\varphi_{\mathrm{opt}}(y) = S(m(y))$, $y \in \mathbb{R}^n$, where $m(y)$ is the mean of the conditional measure $w(\cdot|y)$ with respect to the observed vector y. Furthermore,

$$\mathrm{rad}^{\mathrm{ave}}(S, N, \sigma^2) \;=\; \left(\int_{\mathbf{C}^0} \|S(f)\|^2 \, w(df|0) \right)^{1/2}. \tag{3.26}$$

Let $R_N : [0,1]^2 \to \mathbb{R}$ denote the covariance kernel function of the conditional measure $w(\cdot|y)$ (it is independent of y). Applying (3.26) and the Fubini theorem we obtain

$$\begin{aligned}
\left(\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, N, \sigma^2)\right)^2 \;&=\; \int_{\mathbf{C}^0} \left(\int_0^1 f^2(x)\,dx \right) w(df|0) \\
&=\; \int_0^1 \left(\int_{\mathbf{C}^0} f^2(x)\, w(df|0) \right) dx \\
&=\; \int_0^1 R_N(x,x)\,dx
\end{aligned} \tag{3.27}$$

and

$$\begin{aligned}
\left(\mathrm{rad}^{\mathrm{ave}}(\mathrm{Int}, N, \sigma^2)\right)^2 \;&=\; \int_{\mathbf{C}^0} \left(\int_0^1 f(x)\,dx \right)^2 w(df|0) \\
&=\; \int_{\mathbf{C}^0} \left(\int_0^1 \int_0^1 f(s) f(t)\,ds\,dt \right) w(df|0) \\
&=\; \int_0^1 \int_0^1 R_N(s,t)\,ds\,dt.
\end{aligned} \tag{3.28}$$

Hence, the problem of finding the minimal error $\mathrm{r}_n^{\mathrm{ave}}(S,\sigma^2)$ can be reduced to minimizing (3.27) for function approximation and (3.28) for integration, over all N of the form (3.25).

We shall find the optimal information in three steps. We first give formulas for R_N. Then, using these formulas, we estimate the radius of information N_n consisting of observations at equidistant points. Finally, we present lower bounds on $\mathrm{r}_n^{\mathrm{ave}}(S,\sigma^2)$, $S \in \{\mathrm{App}, \mathrm{Int}\}$, from which it will follow that information N_n is almost optimal.

Covariance kernel of the conditional distribution

Recall that the mean element $m(y)$ of $w(\cdot|y)$ is the natural linear spline interpolating data $\{t_i, z_i\}_{i=0}^n$ where z_i are obtained by smoothing the original data $\{y_i\}_{i=1}^n$, see Section 2.6.4.

We now find formulas for the covariance kernel function R_N of $w(\cdot|y)$. To this end, we first define sequences $\{a_i\}_{i=0}^n$, $\{c_i\}_{i=0}^n$, $\{d_i\}_{i=0}^n$, and $\{b_i\}_{i=1}^n$, as follows.

$$\begin{aligned} a_0 &= c_0 = 0, \\ c_i &= \frac{\sigma^2\,(t_i - a_{i-1})}{\sigma^2 + (t_i - a_{i-1})}, \\ a_i &= t_i - c_i, \qquad\qquad i = 1,2,\ldots,n. \end{aligned} \tag{3.29}$$

$$\begin{aligned} d_n &= c_n, \\ b_i &= a_{i-1} + \frac{(t_i - a_{i-1})^2}{(t_i - a_{i-1}) - d_i}, \\ d_{i-1} &= \frac{(t_{i-1} - a_{i-1})(b_i - t_{i-1})}{b_i - a_{i-1}}, \qquad i = n, n-1, \ldots, 1. \end{aligned} \tag{3.30}$$

(To make these and next formulas well defined for all σ's and t_i's, we use the convention that $0/0 = 0$.) Note that the numbers b_i, $b_i \ge t_i$, are defined in such a way that for the parabola

$$p_i(t) = \frac{(b_i - t)(t - a_{i-1})}{b_i - a_{i-1}}$$

we have $p_i(t_{i-1}) = d_{i-1}$. If information is exact, $\sigma^2 = 0$, then $a_i = b_i = t_i$, while for $\sigma^2 > 0$ we have

$$0 \le a_{i-1} \le t_{i-1} \le t_i \le b_i, \qquad 1 \le i \le n.$$

Theorem 3.13 *The covariance kernel function R_N is given as*

$$R_N(s,t) \;=\; \begin{cases} s - a_n & t_n \le s \le t \le 1, \\ \frac{(s-a_{i-1})(b_i-t)}{b_i-a_{i-1}} & t_{i-1} \le s \le t \le t_i, \quad 1 \le i \le n, \\ \frac{s-a_{i-1}}{t_i-a_{i-1}}\, R_N(t_i,t) & t_{i-1} \le s \le t_i < t, \quad 1 \le i \le n, \end{cases}$$

where a_i, $0 \le i \le n$, and b_i, $1 \le i \le n$, are defined by (3.29) and (3.30).

Proof We start the proof with the following observation. Let μ be a Gaussian measure on a separable Banach space F of functions $f : [0,1] \to \mathbb{R}$, whose covariance kernel is K_0. Let $L(f) = f(u)$, $\forall f \in F$, where $0 \le u \le 1$. Then the conditional distribution of μ with respect to information operator $N = L : F \to \mathbb{R}$ and variance σ^2 is Gaussian with covariance kernel

$$K_1(s,t) \;=\; K_0(s,t) \;-\; \frac{K_0(s,u)\,K_0(t,u)}{K_0(u,u)+\sigma^2}, \qquad 0 \le s,t \le 1. \tag{3.31}$$

Indeed, from the general formulas for conditional distributions given in Section 3.4.1, it follows that $K_1(s,t) = K_0(s,t) - (\,m_1(K_0(s,u))\,)(t)$ where $(m_1(y))(t) = (\sigma^2 + K_0(u,u))^{-1}K_0(t,u)y$, $\forall\, y \in \mathbb{R}$, $0 \le t \le 1$. This gives (3.31).

We now use (3.31) to prove the theorem by induction with respect to the number n of observations. Clearly, the theorem holds for $n = 0$. Assume that the theorem holds for some n, $n \ge 0$. We shall show that it holds also for any information operator consisting of $n+1$ function values,

$$N_1(f) = [f(t_1), \dots, f(t_n), f(t_{n+1})],$$

$0 = t_0 \le t_1 \le \cdots \le t_{n+1} \le 1$. That is, we show that the function $\tilde{R} : [0,1]^2 \to \mathbb{R}$,

$$\tilde{R}(s,t) \;=\; \begin{cases} s - \tilde{a}_{n+1} & t_{n+1} \le s \le t \le 1, \\ \frac{(s-\tilde{a}_{i-1})(\tilde{b}_i-t)}{\tilde{b}_i-\tilde{a}_{i-1}} & t_{i-1} \le s \le t \le t_i, \quad 1 \le i \le n+1, \\ \frac{s-\tilde{a}_{i-1}}{t_i-\tilde{a}_{i-1}}\tilde{R}(t_i,t & t_{i-1} \le s \le t_i \le t, \quad 1 \le i \le n+1, \end{cases}$$

where $\{\tilde{a}_i\}_{i=0}^{n+1}$ and $\{\tilde{b}_i\}_{i=1}^{n+1}$ are defined by (3.29), (3.30), for information operator N_1, is equal to the covariance kernel R_1 of the measure $w(\cdot|0, N_1)$. To this end, let the information operator $N(f) = [f(t_1), \dots, f(t_n)]$, and let

R_0 be the covariance kernel of $w(\cdot|0,N)$. Then (3.31) is valid with $u = t_{n+1}$ and

$$\tilde{a}_i \;=\; a_i, \qquad 0 \le i \le n. \tag{3.32}$$

We have three cases:

1. $t_{n+1} \le s \le t \le 1$.
Then, from (3.31), (3.32), and from the inductive assumption we get

$$\begin{aligned} R_1(s,t) &= (s-a_n) - \frac{(t_{n+1}-a_n)^2}{\sigma^2 + (t_{n+1}-a_n)} \\ &= s - \left[t_{n+1} - \frac{\sigma^2 (t_{n+1}-a_n)}{\sigma^2 + (t_{n+1}-a_n)} \right] \;=\; s - \tilde{a}_{n+1} \;=\; R_2(s,t). \end{aligned}$$

2. $t_{i-1} \le s \le t \le t_i, \quad 1 \le i \le n+1$.
To show that for such s,t is $R_2(s,t) = R_1(s,t)$, we use induction on i, $i = n+1, n, \ldots, 1$. For $i = n+1$ we have

$$\begin{aligned} R_1(s,t) &= (s-a_n) - \frac{(s-a_n)(t-a_n)}{\sigma^2 + (t_{n+1}-a_n)} \;=\; \frac{(s-a_n)(\sigma^2 + t_{n+1} - t)}{\sigma^2 + t_{n+1} - a_n} \\ &= \frac{(s-\tilde{a}_n)(\tilde{b}_{n+1}-t)}{\tilde{b}_{n+1}-\tilde{a}_n} \;=\; R_2(s,t). \end{aligned}$$

Suppose that $1 \le i \le n$. Then, from (3.31) we obtain

$$R_1(s,t) \;=\; \frac{(s-a_{i-1})(b_i-t)}{b_i-a_{i-1}} - \frac{(s-a_{i-1})(t-a_{i-1})}{(t-a_{i-1})^2} \cdot \frac{R_0^2(t_i,t_{n+1})}{\sigma^2 + R_0(t_{n+1},t_{n+1})} \tag{3.33}$$

and

$$R_1(t_i,t_i) \;=\; \frac{(t_i-a_{i-1})(b_i-t_i)}{b_i-a_{i-1}} - \frac{R_0^2(t_i,t_{n+1})}{\sigma^2 + R_0(t_{n+1},t_{n+1})}. \tag{3.34}$$

On the other hand, we have

$$R_1(t_i,t_i) \;=\; \frac{(\tilde{b}_{i+1}-t_i)(t_i-\tilde{a}_i)}{\tilde{b}_{i+1}-\tilde{a}_i} \;=\; \frac{(\tilde{b}_i-t_i)(t_i-a_{i-1})}{\tilde{b}_i-a_{i-1}}. \tag{3.35}$$

Combining (3.33), (3.34), (3.35) and providing some elementary calculations, we finally get

$$R_1(s,t) \;=\; \frac{(s-a_{i-1})(b_i-t)}{b_i-a_{i-1}} + \frac{(s-a_{i-1})(t-a_{i-1})}{(t_i-a_{i-1})^2}$$

$$
\begin{aligned}
&\times \left[\frac{(t_i - a_{i-1})(b_i - t_i)}{b_i - a_{i-1}} - \frac{(\tilde{b}_i - t_i)(t_i - a_{i-1})}{\tilde{b}_i - a_{i-1}}\right] \\
= \quad &\cdots = \frac{(s - a_{i-1})(\tilde{b}_i - t)}{\tilde{b}_i - a_{i-1}} = R_2(s,t).
\end{aligned}
$$

3. $t_{i-1} \le s \le t_i \le t, \quad 1 \le i \le n+1.$
   In this case,

$$
\begin{aligned}
R_1(s,t) \quad = \quad & \frac{s - a_{i-1}}{t_i - a_{i-1}} R_0(t_i, t) - \frac{(s - a_{i-1}) R_0(t_i, t_{n+1})}{t_i - a_{i-1}} \\
& \cdot \frac{R_0(t, t_{n+1})}{\sigma^2 + R_0(t_{n+1}, t_{n+1})} \quad = \quad \frac{s - \tilde{a}_{i-1}}{t_i - \tilde{a}_{i-1}} R_1(t_i, t).
\end{aligned}
$$

This completes induction on n and the proof of the theorem. □

Note that in the case of exact information, $\sigma^2 = 0$, the formulas for R_N reduce to

$$
R_N(s,t) \;=\; \begin{cases} s - t_n, & t_n \le s \le t \le 1, \\ \frac{(s - t_{i-1})(t_i - t)}{t_i - t_{i-1}}, & t_{i-1} \le s \le t \le t_i, \quad 1 \le i \le n, \\ 0, & \text{otherwise}. \end{cases}
$$

Equidistant points

We now consider information consisting of observations at equidistant points. That is, we assume that

$$
N(f) \;=\; N_n(f) \;=\; [f(t_1^*), f(t_2^*), \ldots, f(t_n^*)], \qquad \forall f \in \mathbf{C}^0,
$$

where $t_i^* = i/n, \quad 1 \le i \le n$. Such information is of practical importance since obtaining function values at equidistant points is usually much easier than obtaining them at any other points.

Using (3.36) and (3.27), (3.28), we find that

$$
\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, N_n, 0) \;=\; \frac{1}{\sqrt{6n}}
$$

and

$$
\mathrm{rad}^{\mathrm{ave}}(\mathrm{Int}, N_n, 0) \;=\; \frac{1}{2\sqrt{3}\,n}.
$$

Theorem 3.14 *For $\sigma^2 > 0$ we have*

$$\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, N_n, \sigma^2) \;\approx\; \left(\frac{\sigma^2}{4n}\right)^{1/4}$$

and

$$\mathrm{rad}^{\mathrm{ave}}(\mathrm{Int}, N_n, \sigma^2) \;\approx\; \left(\frac{\sigma^2}{n}\right)^{1/2}$$

as $n \to +\infty$.

The proof will be based on the following two lemmas. Let the sequences $\{c_i^*\}_{i=0}^n$ and $\{d_i^*\}_{i=0}^n$ be defined by (3.29) and (3.30) for the information operator N_n. That is,

$$c_0^* \;=\; 0, \qquad c_i^* \;=\; \frac{\sigma^2(c_{i-1}^* + 1/n)}{\sigma^2 + c_{i-1}^* + 1/n}, \qquad 1 \le i \le n,$$

$$d_n^* \;=\; c_n^*, \qquad d_{i-1}^* \;=\; d_i^* \left(\frac{c_{i-1}^*}{c_{i-1}^* + 1/n}\right)^2 + \frac{c_{i-1}^*/n}{c_{i-1}^* + 1/n}, \qquad n \ge i \ge 1.$$

Lemma 3.8 *(i)*

$$0 \;\le\; c_i^* \;<\; \frac{\sigma}{\sqrt{n}}\,, \qquad \forall i \ge 0.$$

(ii) Let $0 < \alpha < 1$ and $K > \alpha/(1-\alpha^2)$. Then for sufficiently large n we have

$$c_i^* \;\ge\; \frac{\alpha\sigma}{\sqrt{n}}\,, \qquad \forall i \ge K\sigma\sqrt{n}.$$

Proof Observe that the function

$$\xi(x) \;=\; \frac{\sigma^2(x + 1/n)}{\sigma^2 + x + 1/n} - x \;=\; -\frac{x^2 + x/n - \sigma^2/n}{\sigma^2 + x + 1/n}, \qquad x \ge 0,$$

is decreasing and attains zero at

$$g \;=\; \frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{1}{4\sigma^2 n}} - \frac{1}{2n}\,.$$

Moreover, if $0 \le x < g$ then $x + \xi(x) < g$. Hence, the sequence $\{c_i^*\}$ is increasing and $c_i^* < g < \sigma n^{-1/2}$, $\forall i \ge 0$, which proves (i).

To show (ii) observe that for $c_s^* < h < g$ we have

$$c_s^* \;=\; \sum_{j=0}^{s-1} c_{j+1}^* - c_j^* \;=\; \sum_{j=0}^{s-1} \xi(c_j^*) \;\geq\; s\xi(c_s^*) \;\geq\; s\xi(h).$$

Hence, for any $0 < h < g$ and $s \geq 0$

$$c_s^* \;\geq\; \min\{\, h, s\xi(h)\,\} \;=\; \min\left\{\, h,\, -\frac{s(h^2 + h/n - \sigma^2/n)}{\sigma^2 + h + 1/n} \,\right\}. \tag{3.36}$$

Setting $h = \alpha\sigma n^{-1/2}$ and using (3.36) we get that the inequality $c_i^* \geq \alpha\sigma n^{-1/2}$ is satisfied for

$$i \;\geq\; \frac{\alpha\sigma\sqrt{n}\,(1 + \alpha(\sigma\sqrt{n})^{-1} + (\sigma\sqrt{n})^{-2})}{1 - \alpha^2 - \alpha(\sigma\sqrt{n})^{-1}} \;\approx\; \frac{\alpha}{1-\alpha^2}\,\sigma\sqrt{n}.$$

Hence, (ii) follows.

Lemma 3.9 *(i) Let $0 < \alpha < 1$ and $K > \alpha/(1-\alpha^2)$. Then for sufficiently large n*

$$d_i^* \;>\; \frac{\alpha\sigma}{2\sqrt{n}}\,, \qquad \forall i \geq K\sigma\sqrt{n}.$$

(ii) Let $\beta > 1$ and $L > -\frac{1}{2}\ln(\beta - 1)$. Then for sufficiently large n

$$d_{n-i}^* \;\leq\; \frac{\beta\sigma}{2\sqrt{n}}\,, \qquad \forall i \geq L\sigma\sqrt{n}.$$

Proof Let

$$A_i \;=\; \left(\frac{c_i^*}{c_i^* + 1/n}\right)^2, \qquad B_i \;=\; \frac{c_i^*/n}{c_i^* + 1/n}\,, \qquad 0 \leq i \leq n.$$

Let α_1 be such that $\alpha < \alpha_1 < 1$ and $\alpha/(1-\alpha^2) < \alpha_1/(1-\alpha_1^2) < K$. Due to Lemma 3.8, for large n and $i \geq K\sigma\sqrt{n}$ we have $c_i^* \geq \alpha_1\sigma/\sqrt{n}$. Hence, for such i and n

$$\begin{aligned} d_i^* \;&=\; A_i d_{i+1} \,+\, B_i \;\geq\; A d_{i+1}^* \,+\, B \\ &\geq\; \cdots \;\geq\; A^{n-i} d_n^* + B \sum_{j=0}^{n-i-1} A^j \;=\; A^{n-i}\left(c_n^* - \frac{B}{1-A}\right) \,+\, \frac{B}{1-A}, \end{aligned}$$

where

$$A \;=\; \left(\frac{\alpha_1\sigma\sqrt{n}}{1+\alpha_1\sigma\sqrt{n}}\right)^2, \qquad B \;=\; \frac{\alpha_1\sigma}{\alpha_1\sigma n+\sqrt{n}}\,.$$

Since $c_n^* \approx \sigma/\sqrt{n}$ and

$$\frac{B}{1-A} \;=\; \frac{\alpha_1\sigma}{2\sqrt{n}}\cdot\frac{2+2\alpha_1\sigma\sqrt{n}}{1+2\alpha_1\sigma\sqrt{n}} \;\approx\; \frac{\alpha_1\sigma}{2\sqrt{n}},$$

(i) is proved.

To show (ii), let

$$C \;=\; \left(\frac{\sigma\sqrt{n}}{1+\sigma\sqrt{n}}\right)^2, \qquad D \;=\; \frac{\sigma}{\sigma n+\sqrt{n}}\,.$$

Then, due to Lemma 3.8(i), we have $A_i \le C$ and $B_i \le D$, $\forall i$. Hence,

$$\begin{aligned} d_{n-i} \;&\le\; C^i\left(d_n^* - \frac{D}{1-C}\right) + \frac{D}{1-C} \\ &\le\; \frac{\sigma}{2\sqrt{n}}\cdot\frac{2+2\sigma\sqrt{n}}{1+2\sigma\sqrt{n}}\left(1+\frac{\sigma\sqrt{n}}{1+\sigma\sqrt{n}}\left(1+\frac{1}{\sigma\sqrt{n}}\right)^{-2i}\right). \end{aligned}$$

Since for $i \ge L\sigma\sqrt{n}$ we have

$$\left(1+\frac{1}{\sigma\sqrt{n}}\right)^{-2i} \;\le\; e^{-2L} \;<\; \beta-1,$$

(ii) follows.

Proof of Theorem 3.14 It follows from Theorem 3.13 that for $t_{i-1}^* \le t \le t_i^*$ we have

$$\min\{\,d_{i-1}^*, d_i^*\,\} \;\le\; R_N(t,t) \;\le\; \max\{d_{i-1}^*, d_i^*\,\} \,+\, \frac{1}{4n}\,. \tag{3.37}$$

Consider first the approximation problem. Let $0<\alpha<1<\beta$ and K, L be as in Lemma 3.9. Using (3.27) and (3.37) we obtain that for sufficiently large n

$$\begin{aligned} &\left(\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, N_n, \sigma^2)\right)^2 \\ &=\; \int_0^{1-\frac{L\sigma}{\sqrt{n}}-\frac{1}{n}} R_N(t,t)\,dt \;+\; \int_{1-\frac{L\sigma}{\sqrt{n}}-\frac{1}{n}}^1 R_N(t,t)\,dt \end{aligned}$$

$$\begin{aligned} &\leq \left(1-\frac{L\sigma}{\sqrt{n}}-\frac{1}{n}\right)\left(\frac{\beta\sigma}{2\sqrt{n}}+\frac{1}{4n}\right)+\left(\frac{L\sigma}{\sqrt{n}}+\frac{1}{n}\right)\left(\frac{\sigma}{\sqrt{n}}+\frac{1}{4n}\right) \\ &\approx \frac{\beta\sigma}{2\sqrt{n}}\,. \end{aligned} \tag{3.38}$$

On the other hand,

$$\begin{aligned} \left(\mathrm{rad}^{\mathrm{ave}}(\mathrm{App},N_n,\sigma^2)\right)^2 &\geq \int_{\frac{K\sigma}{\sqrt{n}}+\frac{1}{n}}^1 R_N(t,t)\,dt \\ &\geq \left(1-\frac{K\sigma}{\sqrt{n}}-\frac{1}{n}\right)\frac{\alpha\sigma}{2\sqrt{n}} \approx \frac{\alpha\sigma}{2\sqrt{n}}\,. \end{aligned} \tag{3.39}$$

Since (3.38) and (3.39) hold for arbitrary α and β satisfying $0<\alpha<1<\beta$, (a) follows.

We now turn to the integration problem. Let $0\leq s\leq t\leq 1$, $t_{j-1}^*\leq s\leq t_j^*$, $t_{i-1}^*\leq t\leq t_i^*$, $1\leq j\leq i\leq n$. Due to Theorem 3.13 we have

$$\begin{aligned} &R_N(s,t) \\ &= \begin{cases} \frac{s-a_{j-1}^*}{t-a_{j-1}^*}\cdot R_N(t,t) & j=i, \\ \frac{s-a_{j-1}^*}{t_j^*-a_{j-1}^*}\cdot\frac{t_{i-1}^*-a_{i-1}^*}{t-a_{i-1}^*}\cdot R_N(t,t) & j=i+1, \\ \frac{s-a_{j-1}^*}{t_j^*-a_{j-1}^*}\cdot\frac{t_{i-1}^*-a_{i-1}^*}{t-a_{i-1}^*}\cdot\prod_{k=j+1}^{i-1}\frac{t_{k-1}^*-a_{k-1}^*}{t_k^*-a_{k-1}^*}\cdot R_N(t,t) & \text{otherwise,} \end{cases} \end{aligned} \tag{3.40}$$

where the sequence $\{a_i^*\}_{i=0}^n$ is defined by (3.29) for information operator N_n. From Lemma 3.8(i) it follows that

$$\frac{t_i^*-a_i^*}{t_{i+1}^*-a_i^*} = \frac{c_i^*}{c_k^*+1/n} \leq \gamma = \frac{\sigma\sqrt{n}}{1+\sigma\sqrt{n}}\,, \qquad \forall i\,. \tag{3.41}$$

Using (3.37), (3.40), (3.41), Lemmas 3.8(i) and 3.9(ii), we get that for $\beta>1$ and $L>-1/2\,\ln(\beta-1)$, for sufficiently large n

$$\begin{aligned} R_N(s,t) &\leq \gamma^{i-j-1}R_N(t,t) \\ &\leq \gamma^{n(t-s)}\cdot\begin{cases} \frac{\beta\sigma}{2\sqrt{n}}+\frac{1}{4n} & \text{for } \; 0\leq t\leq 1-\frac{L\sigma}{\sqrt{n}}-\frac{1}{n}\,, \\ \frac{\sigma}{\sqrt{n}}+\frac{1}{4n} & \text{for } \; 1-\frac{L\sigma}{\sqrt{n}}-\frac{1}{n}\leq t\leq 1\,. \end{cases} \end{aligned}$$

Hence, for large n

$$\left(\mathrm{rad}^{\mathrm{ave}}(\mathrm{Int},N_n,\sigma^2)\right)^2 = 2\int_0^1\int_0^t R(s,t)\,ds\,dt$$

$$\begin{aligned}
&\le\ 2\left(\frac{\beta\sigma}{2\sqrt{n}}+\frac{1}{4n}\right)\int_0^{1-\frac{L\sigma}{\sqrt{n}}-\frac{1}{n}}\int_0^t \gamma^{n(t-s)}\,ds\,dt \\
&\qquad +2\left(\frac{\sigma}{\sqrt{n}}+\frac{1}{4n}\right)\int_{1-\frac{L\sigma}{\sqrt{n}}-\frac{1}{n}}^1\int_0^t \gamma^{n(t-s)}\,ds\,dt \\
&=\ 2\left(\frac{\beta\sigma}{2\sqrt{n}}+\frac{1}{4n}\right)\frac{1}{n\ln 1/\gamma}\left\{1-\frac{L\sigma}{\sqrt{n}}-\frac{1}{n}+\frac{\gamma^{(n-L\sigma\sqrt{n}-1)}-1}{n\ln 1/\gamma}\right\} \\
&\qquad +2\left(\frac{\sigma}{\sqrt{n}}+\frac{1}{4n}\right)\frac{1}{n\ln 1/\gamma}\left\{\frac{L\sigma}{\sqrt{n}}+\frac{1}{n}+\frac{\gamma^n-\gamma^{(n-L\sigma\sqrt{n}-1)}}{n\ln 1/\gamma}\right\} \\
&=\ (*)\,.
\end{aligned}$$

Since

$$\frac{1}{n\ln 1/\gamma}\ =\ \frac{1}{n\ln\left(1+(\sigma\sqrt{n})^{-1}\right)}\ \approx\ \frac{\sigma}{\sqrt{n}}$$

and $\gamma^{n-L\sigma\sqrt{n}-1}\to 0$ as $n\to+\infty$, then

$$(*)\ \approx\ \frac{\beta\sigma^2}{n}+\frac{2\sigma^3 L}{n\sqrt{n}}\ \approx\ \frac{\beta\sigma^2}{n}\,,$$

which gives the upper bound on $\mathrm{rad}^{\mathrm{ave}}(\mathrm{Int},N_n,\sigma^2)$.

To show the lower bound, we use Lemma 3.8(ii) and Lemma 3.9(i). Let $0<\alpha<1$ and $K>\alpha/(1-\alpha^2)$. Then for sufficiently large n we have

$$\frac{t^*_{i-1}-a^*_{i-1}}{t^*_i-a^*_{i-1}}\ \ge\ \delta\ =\ \frac{\alpha\sigma\sqrt{n}}{1+\alpha\sigma\sqrt{n}}\,,\qquad \forall i\ge K\sigma\sqrt{n}+1,$$

and

$$R_N(t,t)\ \ge\ \frac{\alpha\sigma}{2\sqrt{n}}\,,\qquad \forall t\ge\frac{K\sigma}{\sqrt{n}}+\frac{1}{n}\,.$$

From this and (3.40) we obtain that for large n

$$R_N(s,t)\ \ge\ \delta^{n(t-s)}\frac{\alpha\sigma}{2\sqrt{n}}\,,\qquad t\ge s\ge\frac{K\sigma}{\sqrt{n}}+\frac{2}{n}\,,$$

and, as a consequence,

$$\begin{aligned}
&\left(\mathrm{rad}^{\mathrm{ave}}(\mathrm{Int},N_n,\sigma^2)\right)^2\ \ge\ \frac{\sigma}{\sqrt{n}}\int_{\frac{K\sigma}{\sqrt{n}}+\frac{2}{n}}^1\int_0^t \delta^{n(t-s)}\,ds\,dt \\
&=\ \frac{\alpha\sigma}{\sqrt{n}}\,\frac{1}{n\ln 1/\delta}\left\{1-\frac{K\sigma}{\sqrt{n}}-\frac{2}{n}+\frac{1}{n\ln 1/\delta}\left(\delta^n-\delta^{(K\sigma\sqrt{n}+2)}\right)\right\} \\
&\approx\ \frac{\alpha^2\sigma^2}{n}\,.
\end{aligned}$$

This shows the lower bound on $\mathrm{rad}^{\mathrm{ave}}(\mathrm{Int}, N_n, \sigma^2)$ and completes the proof of the theorem.

Lower bounds

Using (3.29), (3.30), and the formulas (3.36) we can easily show that for exact information the actual values of the minimal errors are equal to

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, 0) \;=\; \frac{1}{\sqrt{2(3n+1)}} \;\approx\; \frac{1}{\sqrt{6n}}, \tag{3.42}$$

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, 0) \;=\; \frac{1}{\sqrt{3}(2n+1)} \;\approx\; \frac{1}{2\sqrt{3}n}. \tag{3.43}$$

Furthermore, the optimal sample points are given by $t_i = 3i/(3n+1)$ for function approximation and $t_i = 2i/(2n+1)$ for integration, $1 \le i \le n$. This shows that in the exact information case N_n is nearly optimal. Almost optimality of N_n in the "noisy" case, $\sigma^2 > 0$, follows from the following theorem.

Theorem 3.15 *For any n and σ^2 we have*

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \sigma^2) \;\ge\; \left(\frac{\sigma}{6\sqrt{n}} - \frac{\sigma^2}{3n}\right)^{1/2} \;\approx\; \frac{1}{\sqrt{6}}\left(\frac{\sigma^2}{n}\right)^{1/4}$$

and

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, \sigma^2) \;\ge\; \left(\frac{\sigma^2}{3(n+\sigma^2)}\right)^{1/2} \;\approx\; \frac{1}{\sqrt{3}}\left(\frac{\sigma^2}{n}\right)^{1/2}.$$

To prove the bound on $\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \sigma^2)$, we need the following lemma.

Lemma 3.10 *Let N be an arbitrary information of the form $N(f) = [f(t_1), \ldots, f(t_n)]$. Then for any $0 \le a < t < b \le 1$ we have*

$$R_N(t,t) \;\ge\; \frac{\sigma^2 \psi(t)}{\sigma^2 + s\psi(t)}$$

where

$$\psi(t) \;=\; \frac{(t-a)(b-t)}{b-a}$$

and s is the number of points t_i satisfying $a < t_i < b$.

Proof Let $\{a_i\}$ and $\{b_i\}$ be the sequences defined by (3.29) and (3.30). Observe first that for any k we have

$$a_k \;\leq\; t_k \,-\, \frac{\sigma^2(t_k-a)}{\sigma^2+s_1(t_k-a)}\,, \tag{3.44}$$

where $s_1=s_1(k)$ is the number of points t_i, $i\leq k$, satisfying $a<t_i$. Indeed, (3.44) can be easily shown by induction on s_1. If $s_1=0$ then $t_k\leq a$ and $a_k\leq a$. For $s_1\geq 1$ we have from (3.29) and from the inductive assumption applied for a_{k-1} that

$$a_k \;=\; t_k \,-\, \frac{\sigma^2(t_k-a_{k-1})}{\sigma^2+t_k-a_{k-1}} \;\leq\; t_k \,-\, \frac{\sigma^2(t_k-a)}{\sigma^2+s_1(t_k-a)}\,.$$

In a similar way we can show that for any k

$$b_k \;\geq\; t_k \,+\, \frac{\sigma^2(b-t_k)}{\sigma^2+s_2(b-t_k)} \tag{3.45}$$

where $s_2=s_2(k)$ is the number of points t_i, $i\geq k$, satisfying $t_i<b$.

Now, let $r=\max\{\,i\geq 0:\, t_r\leq t\,\}$. Due to (3.44) we have

$$a_r \;\leq\; t-\frac{\sigma^2(t-a)}{\sigma^2+s_1(t-a)} \;=:\; a_{max} \tag{3.46}$$

where $s_1=s_1(r)$. Hence, if $r=n$ then $s_1=s$ and

$$R_N(t,t) \;\geq\; t\,-\,a_{max} \;\geq\; \frac{\sigma^2\psi(t)}{\sigma^2+s\psi(t)}\,.$$

For $r<n$, from (3.45) we ontain

$$b_{r+1} \;\geq\; t\,-\,\frac{\sigma^2(b-t)}{\sigma^2+s_2(b-t)} \;=:\; b_{min} \tag{3.47}$$

where $s_2=s_2(r+1)$. Since $s_1+s_2=s$, (3.46) and (3.47) yield

$$R_N(t,t) \;\geq\; \frac{(t-a_{max})(b_{min}-t)}{b_{min}-a_{max}} \;=\; \frac{\sigma^2\psi(t)}{\sigma^2+s\psi(t)}\,,$$

as claimed.

Proof of Theorem 3.15 We start with the problem App. Let N be an arbitrary information operator consisting of observations at t_i, $1\leq i\leq n$.

Divide the unit interval on k equal subintervals (u_{i-1}, u_i), $1 \le i \le k$, where $u_i = i/k$. Let s_i be the number of the points t_i belonging to the ith interval, and let $\psi_i(t) = (t - u_{i-1})(u_i - t)/(u_i - u_{i-1})$. Then, for $u_{i-1} < t < u_i$ we have $\psi_i(t) \le 1/4(u_i - u_{i-1}) = 1/(4k)$. This, (3.27), and Lemma 3.10 yield that the radius of N can be estimated as follows:

$$\begin{aligned}\Big(\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, N, \sigma^2)\Big)^2 \;&\ge\; \sum_{i=1}^{k} \int_{u_{i-1}}^{u_i} \frac{\sigma^2 \psi_i(t)}{\sigma^2 + s_i/(4k)} dt \;=\; \frac{2\sigma^2}{3k} \sum_{i=1}^{k} \frac{1}{s_i + 4k\sigma^2} \\ &=: \;\Omega(s_1, \ldots, s_k)\,.\end{aligned}$$

The function Ω, when defined on the set $\{s_1, \ldots, s_k \ge 0 \,|\, \sum_{i=1}^k s_i \le n\}$, has its minimum at $s_i = n/k \;\forall i$. Hence,

$$\Omega(s_1, \ldots, s_k) \;\ge\; \Omega(\underbrace{n/k, \ldots, n/k}_{k}) \;=\; \frac{2\sigma^2 k}{3(n + 4\sigma^2 k^2)}\,.$$

Letting $k_1 = \max\{1, l\}$, where l is the largest integer satisfying $l \le k_{opt} = \sqrt{n}/(2\sigma)$, we obtain

$$\Big(\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, N, \sigma^2)\Big)^2 \;\ge\; \frac{2\sigma^2 k_1}{3(n + 4\sigma^2 k_1^2)} \;\ge\; \frac{2\sigma^2 (k_{opt} - 1)}{3(n + 4\sigma^2 k_{opt}^2)} \;=\; \frac{\sigma}{6\sqrt{n}} - \frac{\sigma^2}{3n}\,,$$

which proves the desired lower bound on $\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \sigma^2)$.

To show the bound on $\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, \sigma^2)$, we use the general results of Section 3.8.1. When applied to the integration problem in the Wiener space, those results read as follows.

Suppose that the class of permissible functionals consists of all L with $\|L\|_w^2 = \int_{\mathbf{C}^0} L^2(f)\, w(df) \le 1$, i.e., $\Lambda = \Lambda^{\mathrm{all}}$. Then the minimal radius corresponding to observations with variance σ^2 equals

$$\Big(\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, \Lambda^{\mathrm{all}} \sigma^2)\Big)^2 \;=\; \frac{\sigma^2}{n + \sigma^2} \int_F \mathrm{Int}^2(f)\, w(df).$$

Since for $L_t(f) = f(t)$ is $\|L_t\|_w^2 = t$, we have $\Lambda^{\mathrm{std}} \subset \Lambda^{\mathrm{all}}$. This, (3.28), and

$$\int_{\mathbf{C}^0} \mathrm{Int}^2(f)\, w(df) \;=\; \int_0^1 \int_0^1 \min\{s, t\}\, ds\, dt \;=\; \frac{1}{3}\,,$$

yield

$$\Big(\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, \Lambda^{\mathrm{std}} \sigma^2)\Big)^2 \;\ge\; \Big(\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, \Lambda^{\mathrm{all}} \sigma^2)\Big)^2 \;=\; \frac{\sigma^2}{3(n + \sigma^2)}\,,$$

as claimed. □

Theorems 3.14, 3.15, and the formulas (3.42), (3.43) yield that information N_n consisting of noisy observations of function values at equidistant points is almost optimal, for both approximation and integration problems. That is, the errors obtained by applying N_n together with the smoothing spline algorithm are at most $\sqrt{3}$ times larger than optimal. We summarize this in the following corollary.

Corollary 3.5 *For any $\sigma^2 \geq 0$ we have*

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \sigma^2) \;\approx\; \frac{1}{\sqrt{6n}} \;+\; p_n \left(\frac{\sigma^2}{4n}\right)^{1/4}$$

and

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, \sigma^2) \;\approx\; \frac{1}{2\sqrt{3n}} \;+\; q_n \left(\frac{\sigma^2}{n}\right)^{1/2}$$

as $n \to +\infty$, where $p_n, q_n \in [1/\sqrt{3}, 1]$. □

It seems interesting to compare these results with those of Section 3.8.1. More precisely, we want to see whether the class Λ^{std} is as powerful as Λ^{all}. Clearly, $\Lambda^{\mathrm{std}} \subset \Lambda^{\mathrm{all}}$.

As we noticed in the proof of Theorem 3.15, for the integration problem we have $\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, \Lambda^{\mathrm{all}}\sigma^2) \asymp \sigma/\sqrt{n}$ $(\sigma^2 \geq 0)$. Hence, for σ^2 the classes Λ^{std} and Λ^{all} give similar minimal errors, while for exact information Λ^{all} is much more powerful than Λ^{std}.

Due to NR 3.20, for approximation the corresponding radius satisfies $\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \Lambda^{\mathrm{all}}, \sigma^2) \asymp 1/\sqrt{n} + \sigma \ln n/\sqrt{n}$. The situation is then quite opposite. We have $\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \Lambda^{\mathrm{all}}, 0) \asymp \mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \Lambda^{\mathrm{std}}, 0)$, while for $\sigma^2 > 0$

$$\frac{\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \Lambda^{\mathrm{all}}, \sigma^2)}{\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \Lambda^{\mathrm{std}}, \sigma^2)} \;\asymp\; \left(\frac{\sigma^2}{n}\right)^{1/4} \ln n.$$

This will not change when we replace Λ^{std} by the class of functionals of the form $\tilde{L}_t(f) = t^{-1/2} f(t)$ for which $\|\tilde{L}_t\|_w = 1$ $\forall t \in (0, 1]$; see E 3.36.

Notes and Remarks

NR 3.18 Most of Section 3.8.1 is based on Plaskota [78] and [81]. Theorem 3.12

is new.

NR 3.19 In the average case setting, we assume that noise of different observations is uncorrelated, e.g., $x \sim \mathcal{N}(0, \sigma^2 I)$. As we already mentioned, in the worst case the uncorrelated noise corresponds to noise bounded in the maximum norm, e.g., $\|x\|_Y = \|x\|_\infty \le \delta$. Lemma 2.15 says that for such noise the worst case radius does not tend to zero with n. This stands in contrast with the average case where the radius can be reduced to an arbitrary small level.

NR 3.20 Consider the problem of approximating multivariate functions $f \in F = \mathbf{C}^{0\ldots 0}_{r_1\ldots r_d}$ in the norm of $G = \mathcal{L}_2((0,1)^d)$, with respect to the Wiener sheet measure $\mu = w_{r_1\ldots r_d}$. That is $S : \mathbf{C}^{0\ldots 0}_{r_1\ldots r_d} \to \mathcal{L}_2((0,1)^d)$, $S(f) = f$. As mentioned in NR 3.10, the abstract Wiener space corresponding to $w_{r_1\ldots r_d}$ is $\{H, F\}$ with $H = W^{0\ldots 0}_{r_1+1\ldots r_d+1}$. Recall that $SC_\mu S^* = S_H S_H^*$. Due to NR 2.30, the eigenvalues of $S_H S_H^*$ are given as

$$\lambda_j \asymp \left(\frac{\ln^{k-1} j}{j} \right)^{2(r+1)} \qquad \text{as} \quad j \to +\infty,$$

where $r = \min\{r_1, \ldots, r_d\}$ and k is the number of such i that $r_i = r$. The results of Section 3.8.1 yield that for $\sigma^2 > 0$ we have

$$\left(\mathrm{r}^{\mathrm{ave}}_n(\sigma^2)\right)^2 \asymp \begin{cases} \sigma^2 \frac{\ln^{2k} n}{n} & r = 0, \\ \sigma^2 \frac{1}{n} & r \ge 1, \end{cases}$$

and $(\mathrm{r}^{\mathrm{ave}}_n(0))^2 \asymp (\ln^{2(k-1)(r+1)} n) n^{-(2r+1)}$.

NR 3.21 There are many papers dealing with integration or approximation in Wiener type spaces, based on exact information. The first papers on this subject are due to Suldin [102] [103] who analyzed integration with respect to the classical Wiener measure on $\mathbf{C}^0$. Other positions include, e.g., Sacks and Ylvisaker [87] [88] [89], Wahba [115], Lee [46], and Lee and Wasilkowski [48]. The multivariate case with exact information was studied, e.g., by Papageorgiou and Wasilkowski [70], Ritter *et al.* [86], Wasilkowski [119], Wasilkowski and Woźniakowski [122], Woźniakowski [127] [128] [129].

The results on noisy information of Section 3.8.2 are based on Plaskota [79].

NR 3.22 We now give a concrete application of the correspondence theorem of Section 3.6.3. We let F to be a Hilbert space,

$$F = W^0 = \{ f : [0,1] \to \mathbb{R} \mid \quad f(0) = 0,\ f\text{–abs. cont.},\ f' \in \mathcal{L}_2(0,1) \},$$

with the inner product $\langle f_1, f_2 \rangle_F = \int_0^1 f_1(t) f_2(t)\, dt$. Consider the problem of approximating the integral $\mathrm{Int}(f)$ in the worst case setting with E being the unit ball of F. Information consists of n function evaluations and the noise is bounded in the Euclidean norm, $\sum_{i=1}^n x_i^2 \le \delta^2$. As we know, $\{W^0, \mathbf{C}^0\}$ is an abstract Wiener space and the classical Wiener measure w is the corresponding to it Gaussian measure on $\mathbf{C}^0$. Hence, we can apply Theorem 3.7 and Corollary 3.5 to get that for this problem the minimal radius is given as

$$\mathrm{r}_n^{\mathrm{wor}}(\mathrm{Int}, \delta) \;\approx\; \frac{1}{2\sqrt{3}n} \,+\, \tilde{q}_n \,\frac{\delta}{\sqrt{n}}$$

where $\tilde{q}_n \in [1/\sqrt{3}, \sqrt{2}]$. These bounds are attained by the 1/2–smoothing spline algorithm using noisy function values at equidistant points.

NR 3.23 We assume that each value $f(t_i)$ is observed with the same variance σ^2. One may consider a model in which $f(t_i)$ is observed with variance σ_i^2 where σ_i's are possibly different. It is easy to verify that in this case Theorem 3.13 remains valid provided that σ^2 is in the formulas (2.5) and (2.6) replaced by σ_i^2. However, formulas for $\mathrm{r}_n^{\mathrm{ave}}(\sigma_1^2, \ldots, \sigma_n^2)$ are unknown.

NR 3.24 The problems App and Int with $F = \mathbf{C}_r^0$ and μ being the r–fold Wiener measure w_r $(r \ge 1)$ were studied in Plaskota [79]. It was shown that if the class Λ consists of function values or derivatives of order at most r, then

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, r, \sigma^2) \;\asymp\; \frac{\sigma}{\sqrt{n}} \,+\, \left(\frac{1}{n}\right)^{r+1/2}$$

and

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, r, \sigma^2) \;\asymp\; \frac{\sigma}{\sqrt{n}} \,+\, \left(\frac{1}{n}\right)^{r+1}.$$

These bounds are attained by information

$$N_n^r(f) \;=\; [\, f^{(r)}(t_1^*), f^{(r)}(t_2^*), \ldots, f^{(r)}(t_n^*)\,] \tag{3.48}$$

where $t_i^* = i/n,\ 1 \le i \le n$; see E 3.38.

One can show, see Plaskota [84], that for integration the same bound can be obtained using only function values. However, this fact does not apply to the function approximation problem, which follows from more general results of Ritter [85]. He considered numerical differentiation, $S(f) = \mathrm{Dif}_k(f) = f^{(k)}$ $(0 \le k \le r)$, with respect to the same r–fold Wiener measure. Assuming that only observations of function values are allowed, he showed that

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Dif}_k, r, \sigma^2) \;\asymp\; \left(\frac{\sigma}{\sqrt{n}}\right)^{\frac{2(r-k)+1}{2r+2}} \,+\, \left(\frac{1}{n}\right)^{r+1/2}.$$

In particular, for approximation from noisy function values $(\sigma^2 > 0)$, the minimal radius has the exponent $(2r+1)/(2r+2)$ which is much worse than 1/2. Hence, for function approximation, noisy information about rth derivatives is much more powerful than information about function values.

Exercises

E 3.30 Let $a_1 \geq a_2 \geq \cdots \geq a_m \geq 0$ and let λ_i', λ_i, $1 \leq i \leq m$, be such that for all $1 \leq r \leq m$, $\sum_{i=1}^r \lambda_i' \leq \sum_{i=1}^r \lambda_i$. Show that then $\sum_{i=1}^m a_i \lambda_i' \leq \sum_{i=1}^m a_i \lambda_i$.

E 3.31 Show that the lower bound in (3.22) is achieved if

$$\sum_{i=s}^n \eta_i^{**} \;\leq\; \sum_{i=s}^n \frac{1}{\sigma_i^2}\,, \qquad 1 \leq s \leq n,$$

where $\eta^{**} = (\eta_1^{**}, \ldots, \eta_n^{**})$ is the solution (3.18) of the problem $(\mathrm{P}(0,\mathrm{n}))$. On the other hand, the upper bound in (3.22) is achieved if for all $0 \leq q < r \leq n$ the solution η^* of $(\mathrm{P}(\mathrm{q},\mathrm{r}))$ satisfies

$$\sum_{j=s}^r \eta_j^* \;\geq\; \sum_{j=s}^r \frac{1}{\sigma_j^2}\,, \qquad q+1 \leq s \leq r.$$

E 3.32 Show that $\mathrm{r}_n^{\mathrm{ave}}(\sigma_1^2, \ldots, \sigma_n^2)$ is a strictly increasing function of each σ_i^2.

E 3.33 Show that the sequence $\mathrm{r}_n^{\mathrm{ave}}(\sigma^2)$ of the minimal radii given by (3.23) is convex, i.e.,

$$\mathrm{r}_n^{\mathrm{ave}}(\sigma^2) \;\leq\; \frac{\mathrm{r}_{n-1}^{\mathrm{ave}}(\sigma^2) + \mathrm{r}_{n+1}^{\mathrm{ave}}(\sigma^2)}{2} \qquad \forall n \geq 1.$$

E 3.34 Consider the pair of problems (P1) and (P2) on page 186 with $\Sigma = I$ and $\delta^2 = \sigma^2$. Suppose that the eigenvalues λ_i of the operator $S_H^* S_H$ are $\lambda_i = i^{-2}$, $i \geq 1$. Show that then

$$\frac{\mathrm{rad}^{\mathrm{wor}}(N_\Sigma, \Delta)}{\mathrm{r}_n^{\mathrm{wor}}(\delta)} \;\asymp\; \ln n \quad \text{and} \quad \frac{\mathrm{rad}^{\mathrm{ave}}(N_\Delta, \Sigma)}{\mathrm{r}_n^{\mathrm{ave}}(\sigma^2)} \;\asymp\; \frac{\sqrt{n}}{\ln n},$$

where N_Σ is the optimal information in the worst case (P1), and N_Δ is the optimal information in the average case (P2). Hence, Theorem 3.12 does not hold if information N^* is replaced by N_Σ or N_Δ.

E 3.35 Suppose that for $f \in C^0$ the values $f(t_i)$ are observed with variances σ_i^2, $1 \leq i \leq n$, where σ_i's are possibly different. Show that then the formula for the conditional distribution given in Theorem 3.13 remains valid provided that σ^2 is in the formulas (3.29) replaced by σ_i^2.

E 3.36 Consider the approximation problem in the Wiener space with the class $\tilde{\Lambda}^{\mathrm{std}}$ consisting of functionals of the form $L(f) = t^{-1/2} f(t)$, $t \in [0,1]$ (or equivalently, assuming that observations of $f(t)$ with the variance $t\sigma^2$ are allowed). Show that then

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \tilde{\Lambda}^{\mathrm{std}}, \sigma^2) \;\asymp\; \mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \Lambda^{\mathrm{std}}, \sigma^2) \;\asymp\; \frac{1}{\sqrt{n}} + \left(\frac{\sigma^2}{n}\right)^{1/4}.$$

Hint: Consider the solution operator $S_a : \mathbf{C}^0 \to \mathcal{L}_2(a,1)$, $(S_a(f))(t) = f(t)$, where $a \in (0,1)$. Observe that for any $\mathbb{N}$ and φ we have $\mathrm{e}^{\mathrm{ave}}(S_a, \mathbb{N}, \varphi) \le \mathrm{e}^{\mathrm{ave}}(\mathrm{App}, \mathbb{N}, \varphi)$. To find a lower bound on $\mathrm{e}^{\mathrm{ave}}(S_a, \mathbb{N}, \varphi)$, use the technique from the proof of Theorem 3.15.

E 3.37 Let w_r be the r–fold Wiener measure on $\mathbf{C}_r^0$, and let $L(f) = f^{(k)}(t)$, $f \in \mathbf{C}_r^0$, with $0 \le k \le r$. Show that

$$\|L\|_{w_r}^2 \;=\; \int_{\mathbf{C}_r^0} L^2(f)\, w_r(df) \;=\; \frac{t^{2(r-k)+1}}{((r-k)!)^2\,(2(r-k)+1)} \;\le\; 1\,.$$

E 3.38 Let $F = \mathbf{C}_r^0$, $\mu = w_r$, and let N_n^r be information defined as in (3.48). Show the inequalities:

$$\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, r+1, N_n^{r+1}) \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathrm{Int}, r, N_n^r) \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, r, N_n^r)\,.$$

Use this and the previous exercise to obtain that for Λ consisting of function values and derivatives of order at most r we have

$$\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, r, \sigma^2) \;\asymp\; \frac{\sigma}{\sqrt{n}} \;\asymp\; \mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, r, \sigma^2)\,,$$

for all $r \ge 1$ and $\sigma^2 > 0$.

3.9 Complexity

In this section we deal with the average problem complexity. Recall that any problem is defined by the solution operator $S : F \to G$, probability measure μ on F, and the class Λ of permissible functionals.

As in the worst case setting, we assume that approximations are obtained by executing a program. The program is defined in Section 2.9. The only difference is in the interpretation of the information statement. Namely,

$$\mathcal{I}(\, d \,|\, L, f, \sigma^2\,)$$

now means that to the real variable d is assigned a value of the real Gaussian random variable whose mean element is $L(f)$ and variance equals σ^2. The cost of executing this statement is $\mathrm{c}(\sigma^2)$ where, as before, c is a nonnegative and nonincreasing cost function which assumes positive values for small $\sigma^2 > 0$.

We recall that the program specifies not only how information is collected, but also which primitive operations are to be performed. The primitive operations are: arithmetic operations and comparisons over $\mathbb{R}$, elementary linear operations over G, and logical operations over the Boolean values.

Let $\mathcal{P}$ be a program which is a realization of an algorithm φ using information operator $\mathbb{N}$. The (average) cost of computing an approximation with the program $\mathcal{P}$ equals

$$\mathrm{cost}^{\mathrm{ave}}(\mathcal{P}) \;=\; \int_Y \mathrm{cost}(\mathcal{P};y)\,\mu_1(dy)$$

where, as before, $\mathrm{cost}(\mathcal{P};y)$ is the cost of computing $\varphi(y)$, and μ_1 is the a priori distribution of noisy information y on Y,

$$\mu_1(B) \;=\; \int_F \pi_f(B)\,\mu(df) \qquad (3.49)$$

(compare with Section 3.2).

The definition of $\mathrm{cost}^{\mathrm{ave}}(\mathcal{P})$ yields the (average) algorithm complexity, $\mathrm{comp}^{\mathrm{ave}}(\mathbb{N},\varphi)$, and the problem complexity, $\mathrm{Comp}^{\mathrm{ave}}(\varepsilon)$. Namely,

$$\mathrm{comp}^{\mathrm{ave}}(\mathbb{N},\varphi) \;=\; \inf\{\,\mathrm{cost}^{\mathrm{ave}}(\mathcal{P}) \mid \quad \mathcal{P} \text{ is a realization of } \varphi \text{ using } \mathbb{N}\,\},$$

and for $\varepsilon \ge 0$,

$$\mathrm{Comp}^{\mathrm{ave}}(\varepsilon) \;=\; \inf\{\,\mathrm{comp}^{\mathrm{ave}}(\mathbb{N},\varphi) \mid \quad \{\mathbb{N},\varphi\} \text{ such that } \mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi) \le \varepsilon\,\}$$

$(\inf\emptyset = +\infty)$.

Our aim now is to obtain general bounds on the average complexity of linear problems with Gaussian measures. We assume that

- S is a continuous linear operator acting between a separable Banach space F and a separable Hilbert space G, and
- μ is a zero mean Gaussian measure on F.

We first show an auxiliary result about relations between nonadaptive and adaptive information.

3.9.1 Adaption versus nonadaption, II

In Section 3.7.2 we compared the radii of adaptive and nonadaptive information. Theorem 3.9 says that for any adaptive information $\mathbb{N}$ there exists $y \in Y$ such that the average radius of the nonadaptive information $\mathbb{N}_y$ is not greater than the average radius of $\mathbb{N}$. We now prove a stronger result.

Let $\mathbb{N} = \{N_y, \Sigma_y\}_{y \in Y}$ be an arbitrary information operator. The average complexity of $\mathbb{N}$ is given as

$$\mathrm{comp}(\mathbb{N}) \;=\; \int_Y \sum_{i=1}^{n(y)} \mathrm{c}(\sigma_i^2(y_1, \ldots, y_{i-1})\,)\, dy.$$

Clearly, if $\mathbb{N}$ is nonadaptive then we simply have $\mathrm{comp}(\mathbb{N}) = \sum_{i=1}^{n} \mathrm{c}(\sigma_i^2)$.

For $a \in \mathbb{R}$ and $y^{(1)}, y^{(2)} \in Y,$, let $\mathbb{N}' = \mathbb{N}'(y^{(1)}, y^{(2)}, a)$ be an information operator defined based on $\mathbb{N}$ in the following way. Denote by n_i the length of $y^{(i)}$ and by y_1 the first component of a vector y. Let

$$Y' \;=\; \{\, y \in \mathbb{R}^{n_1} \mid \quad y_1 \le a \,\} \,\cup\, \{\, y \in \mathbb{R}^{n_2} \mid \quad y_1 > a \,\},$$

and for $y \in Y'$,

$$\{\, N'_y, \Sigma'_y \,\} \;=\; \begin{cases} \{\, N_{y^{(1)}}, \Sigma_{y^{(1)}} \,\} & \text{if} \quad y_1 \le a, \\ \{\, N_{y^{(2)}}, \Sigma_{y^{(2)}} \,\} & \text{if} \quad y_1 > a. \end{cases}$$

Finally, we set $\mathbb{N}' = \{N'_y, \Sigma'_y\}_{y \in Y'}$. Observe that information $\mathbb{N}'$ is almost nonadaptive since it uses only at most two nonadaptive information operators. It turns out that the class of such information operators is as powerful as the class of all adaptive information operators. Namely, we have the following theorem.

Theorem 3.16 *Let $\mathbb{N} = \{N_y, \Sigma_y\}_{y \in Y}$ be an adaptive information operator. Then there exist $y^{(1)}, y^{(2)} \in Y$ and $a \in \mathbb{R}$, such that for the information $\mathbb{N}' = \mathbb{N}'(y^{(1)}, y^{(2)}, a)$ we have*

$$\mathrm{comp}(\mathbb{N}') \;\le\; \mathrm{comp}(\mathbb{N}) \qquad \textit{and} \qquad \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}') \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}).$$

Proof Let ω be the a priori distribution of the variable $y \to \mathrm{comp}(\mathbb{N}_y) = \sum_{i=1}^{n(y)} \mathrm{c}(\sigma_i^2(y_1, \ldots, y_{i-1})\,)$ on $\mathbb{R}$, i.e.,

$$\omega(B) \;=\; \mu_1(\,\{y \in Y \mid \quad \mathrm{comp}(\mathbb{N}_y) \in B\}\,), \qquad \forall\, B\text{–Borel set of } \mathbb{R}$$

where μ_1 is given by (3.49). Clearly,

$$\mathrm{comp}(\mathbb{N}) \;=\; \int_{\mathbb{R}} T\, \omega(dT). \tag{3.50}$$

The measure μ_1 can be decomposed with respect to the mapping $y \to \mathrm{comp}(\mathbb{N}_y)$ as

$$\mu_1(\cdot) \;=\; \int_{\mathbb{R}} \mu_1(\cdot|T)\,\omega(dT),$$

where $\mu_1(\cdot|T)$ is a probability measure on Y which is supported on the set $Y_T \,=\, \{y\,|\; \mathrm{comp}(\mathbb{N}_y) \,=\, T\}$, for all T such that $Y_T \neq \emptyset$. This, (3.1) and Theorem 3.2 yield

$$(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}))^2 \;=\; \int_Y (r(\nu_2(\cdot|y))\,)^2\;\mu_1(dy) \;=\; \int_{\mathrm{R}} \psi(T)\,\omega(dT) \tag{3.51}$$

where

$$\psi(T) \;=\; \begin{cases} \int_Y (r(\nu_2(\cdot|y)\,)^2\,\mu_1(dy|T) & \text{if} \quad Y_T \neq \emptyset, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.52}$$

Here $\nu_2(\cdot|y) = \mu_2(S^{-1}(\cdot)|y)$ is the conditional distribution of $S(f)$ given y, and $r(\cdot)$ is the radius of a measure.

We now show that it is possible to select real numbers $0 \le T_1 \le T_2 < +\infty$ and $0 \le \alpha^* \le 1$ such that

$$\alpha^*\,T_1 \;+\; (1-\alpha^*)\,T_2 \;\le\; \int_{\mathbb{R}} T\,\omega(dT) \tag{3.53}$$

and

$$\alpha^*\,\psi(T_1) \;+\; (1-\alpha^*)\,\psi(T_2) \;\le\; \int_{\mathbb{R}} \psi(T)\,\omega(dT). \tag{3.54}$$

To this end, let $T_0 = \int_{\mathbb{R}} T\,\omega(dT)$ and $\psi_0 = \int_{\mathbb{R}} \psi(T)\,\omega(dT)$. If such numbers did not exist, for any $T > T_0$ the graph of ψ on the interval $[0, T_0]$ would lie above the line passing through the points (T_0, ψ_0) and $(T, \psi(T))$, i.e.,

$$\psi(R) \;>\; \tilde{\psi}_{\beta_T}(R) \;=\; \beta_T\,(R - T_0) \;+\; \psi_0, \quad \forall R \in [0, T_0],$$

where $\beta_T = (\psi(T) - \psi_0)/(T - T_0)$. Let $\beta = \inf_{T > T_0} \beta_T$. Then $\beta > -\infty$ and for all $T \ge 0$ we have $\psi(T) \ge \tilde{\psi}_\beta(T)$. Moreover, the last inequality "$\ge$" can be replaced by "$>$" on the interval $[0, T_0]$ or on $[T_0, +\infty)$. Hence, we obtain

$$\int_{\mathbb{R}} \psi(T)\,\omega(dT) \;>\; \int_{\mathbb{R}} \tilde{\psi}_\beta(T)\,\omega(dT) \;=\; \psi_0 \;=\; \int_{\mathbb{R}} \psi(T)\,\omega(dT),$$

which is a contradiction.

Let T_1, T_2 and α^* satisfy (3.53), (3.54). We now choose two vectors $y^{(j)}$, $j = 1, 2$, in such a way that $\mathrm{comp}(\mathbb{N}_{y^{(j)}}) = T_j$ and

$$\int_F (r(\nu_2(\cdot|y^{(j)})\,))^2\,\mu_2(df|z^{(j)}) \;\le\; \psi(T_j),$$

as well as the number a such that

$$\int_{-\infty}^{a} \exp\left(\frac{-x^2}{2\sigma_*^2}\right)\,dx \;=\; \alpha^*$$

where $\sigma_*^2 = L_1(C_\mu L_1) + \sigma_1^2$ is the variance of the Gaussian random variable y_1. From (3.50) to (3.54) it now follows that for the information $\mathbb{N}' = \mathbb{N}'(y^{(1)}, y^{(2)}, a)$ we have

$$\begin{aligned}\mathrm{comp}(\mathbb{N}') \;&=\; \alpha^*\,\mathrm{comp}(\mathbb{N}_{y^{(1)}}) \,+\, (1-\alpha^*)\mathrm{comp}(\mathbb{N}_{y^{(2)}})\\ &\le\; \int_{\mathbb{R}} T\,\omega(dT) \;=\; \mathrm{comp}(\mathbb{N})\end{aligned}$$

and

$$\begin{aligned}\left(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}')\right)^2 \;&=\; \alpha^*\,\left(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_{y^{(1)}})\right)^2 \,+\, (1-\alpha^*)\,\left(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_{y^{(2)}})\right)^2\\ &\le\; \alpha^*\,\psi(T_1) \,+\, (1-\alpha^*)\psi(T_2) \;\le\; \int_{\mathbb{R}} \psi(T)\,\omega(dT)\\ &=\; \left(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})\right)^2,\end{aligned}$$

as claimed. □

We now make the following observation. Assume without loss of generality that $\mathrm{comp}(\mathbb{N}_{y^{(1)}}) \le \mathrm{comp}(\mathbb{N})$ and $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_{y^{(2)}}) \le \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$ (if this were not true, it would be possible to select $y^{(1)} = y^{(2)}$). Let $0 < p < 1$. Then for $\alpha^* \ge p$ we have

$$\mathrm{comp}(\mathbb{N}_{y^{(1)}}) \;\le\; \mathrm{comp}(\mathbb{N}) \qquad \text{and} \qquad \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_{y^{(1)}}) \;\le\; \frac{1}{\sqrt{p}}\,\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}),$$

while for $\alpha^* < p$

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_{y^{(2)}}) \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \qquad \text{and} \qquad \mathrm{comp}(\mathbb{N}_{y^{(2)}}) \;\le\; \frac{1}{1-p}\,\mathrm{comp}(\mathbb{N}).$$

This yields the following corollary.

Corollary 3.6 *Let $0 < p < 1$. For any adaptive information $\mathbb{N} = \{\mathbb{N}_y\}_{y\in Y}$ there exists $y^* \in Y$ such that*

$$\mathrm{comp}(\mathbb{N}_y) \;\le\; \frac{1}{1-p}\,\mathrm{comp}(\mathbb{N}) \qquad \textit{and} \qquad \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_y) \;\le\; \frac{1}{\sqrt{p}}\,\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})\,. \quad \square$$

In particular, one can take $p = 1/2$ to get

$$\mathrm{comp}(\mathbb{N}_y) \;\le\; 2\cdot\mathrm{comp}(\mathbb{N}) \qquad \text{and} \qquad \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_y) \;\le\; \sqrt{2}\cdot\mathrm{rad}^{\mathrm{ave}}(\mathbb{N})\,.$$

3.9.2 General bounds

We are now ready to present general bounds on the average ε–complexity of linear problems with Gaussian measures. Let

$$\begin{aligned}\mathrm{IComp}(\varepsilon) \;=\; \inf\{\,\mathrm{comp}(\mathbb{N}) \;\mid\; & \mathbb{N}\text{–adaptive, and there exists } \varphi \\ & \text{such that } \mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi)\le\varepsilon\,\}\end{aligned}$$

be the ε–information complexity, and let

$$\begin{aligned}\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \;=\; \inf\{\,\mathrm{comp}(\mathbb{N}) \;\mid\; & \mathbb{N}\text{–nonadaptive, and there exists } \varphi \\ & \text{such that } \mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi)\le\varepsilon\,\}\end{aligned}$$

be the corresponding quantity for nonadaptive information.

We start with the following theorem which corresponds to Theorem 2.19 of the worst case setting.

Theorem 3.17 *(i) For any $0 < p < 1$ we have*

$$\mathrm{Comp}(\varepsilon) \;\ge\; (1-p)\,\mathrm{IComp}^{\mathrm{non}}\left(\frac{\varepsilon}{\sqrt{p}}\right).$$

(ii) Let $\rho \ge 1$. Let $\mathbb{N}_\varepsilon$ be nonadaptive information using $n(\varepsilon)$ observations and such that $\mathrm{comp}(\mathbb{N}) \le \rho\,\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$. Then

$$\mathrm{Comp}(\varepsilon) \;\le\; \rho\cdot\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \;+\; (2\,n(\varepsilon)-1)\,\mathrm{g}.$$

Proof (i) follows immediately from Corollary 3.6 since it yields

$$\mathrm{Comp}(\varepsilon) \;\ge\; \mathrm{IComp}(\varepsilon) \;\ge\; (1-p)\,\mathrm{IComp}^{\mathrm{non}}\left(\frac{\varepsilon}{\sqrt{p}}\right).$$

To show (ii) observe that for the spline algorithm we have $\mathrm{e}^{\mathrm{ave}}(\mathbb{N}_\varepsilon, \varphi_{\mathrm{spl}}) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}_\varepsilon)$. Since φ_{spl} is linear, the complexity of φ_{spl} using $\mathbb{N}_\varepsilon$ equals $\mathrm{comp}(\mathbb{N}_\varepsilon) + (2\,n(\varepsilon) - 1)\mathrm{g}$. This completes the proof. □

Theorem 3.17 immediately yields the following corollary.

Corollary 3.7 *If the assumptions of Theorem 3.17 are fulfilled and, additionally,*

$$\mathrm{IComp}^{\mathrm{non}}(\varepsilon) \;=\; O(\mathrm{IComp}^{\mathrm{non}}(p^{-1/2}\varepsilon)\,) \qquad \textit{and} \qquad n(\varepsilon) \;=\; O(\mathrm{IComp}^{\mathrm{non}}(\varepsilon)\,),$$

then

$$\mathrm{Comp}(\varepsilon) \;\asymp \mathrm{IComp}^{\mathrm{non}}(\varepsilon) \qquad \textit{as} \quad \varepsilon \to 0^+. \quad \Box$$

Recall that the assumption $n(\varepsilon) = O(\mathrm{IComp}^{\mathrm{non}}(\varepsilon))$ is satisfied when the cost function is bounded from below by a positive constant, $\mathrm{c}(\sigma^2) \geq c_0 > 0$. The second assumption, $\mathrm{IComp}^{\mathrm{non}}(\varepsilon) = O(\mathrm{IComp}^{\mathrm{non}}(p^{-1/2}\varepsilon))$, means that $\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$ increases at most polynomially in $1/\varepsilon$ as $\varepsilon \to 0^+$. This condition can often be replaced by semiconvexity of $\mathrm{IComp}^{\mathrm{non}}(\sqrt{\varepsilon})$. Namely, we have the following result.

Lemma 3.11 *Suppose that the function* $\varepsilon \to \mathrm{IComp}^{\mathrm{non}}(\sqrt{\varepsilon})$ *is semiconvex, i.e., there exist* $\varepsilon_0 \geq 0,\ 0 < \alpha \leq \beta,$ *and a convex function* $h : [0,+\infty) \to [0,+\infty]$ *such that*

$$\alpha \cdot h(\varepsilon) \;\leq\; \mathrm{IComp}^{\mathrm{non}}(\sqrt{\varepsilon}) \qquad \forall\, \varepsilon \geq 0,$$

and

$$\mathrm{IComp}^{\mathrm{non}}(\sqrt{\varepsilon}) \;\leq\; \beta \cdot h(\varepsilon) \qquad \forall\, 0 \leq \varepsilon \leq \varepsilon_0.$$

Then

$$\mathrm{IComp}(\varepsilon) \;\geq\; \frac{\alpha}{\beta} \cdot \mathrm{IComp}^{\mathrm{non}}(\varepsilon) \qquad \forall\, 0 \leq \varepsilon \leq \varepsilon_0.$$

Proof Let $\mathbb{N} = \{\mathbb{N}_y\}_{y \in Y}$ be arbitrary information with radius $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \leq \varepsilon \leq \varepsilon_0$. Let

$$\psi(y) \;=\; \left(r(\mu_2(\cdot|y))\,\right)^2.$$

Define the probability measure ω on $\mathbb{R}$ as

$$\omega(B) \;=\; \mu_1(\,\{\, y \in Y \mid \quad \psi(y) \in B\,\}\,), \qquad \forall B\text{–Borel set of } \mathbb{R}.$$

Due to convexity of h and the inequality

$$\mathrm{comp}(\mathbb{N}_y) \;\ge\; \mathrm{IComp}^{\mathrm{non}}\left(\sqrt{\psi(y)}\right) \;\ge\; \alpha\cdot h(\psi(z)\,),$$

we have

$$\begin{aligned}
\mathrm{IComp}(\mathbb{N}) \;&=\; \int_Y \mathrm{comp}(\mathbb{N}_y)\,\mu_1(dy) \;\ge\; \alpha\cdot\int_Y h(\psi(z))\,\mu_1(dz)\\
&=\; \alpha\cdot\int_{\mathbb{R}} h(x)\,\omega(dx) \;\ge\; \alpha\cdot h\left(\int_{\mathbb{R}} x\,\omega(dx)\right)\\
&=\; \alpha\cdot h\left((\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}))^2\right) \;\ge\; \frac{\alpha}{\beta}\cdot\mathrm{IComp}^{\mathrm{non}}(\,\mathrm{e}^{\mathrm{ave}}(\mathbb{N})\,)\\
&\ge\; \frac{\alpha}{\beta}\cdot\mathrm{IComp}^{\mathrm{non}}(\varepsilon).
\end{aligned}$$

Since $\mathbb{N}$ was arbitrary, the lemma follows. □

Similarly to the worst case setting, the main tool for deriving ε–complexity will be the Tth minimal (average) radius which is defined as

$$\mathrm{R}(T) \;=\; \inf\left\{ \mathrm{r}_n^{\mathrm{ave}}(\sigma_1^2,\ldots,\sigma_n^2)\;\middle|\; n\ge 1,\ \sum_{i=1}^n \mathrm{c}(\sigma_i^2)\le T\right\}.$$

Knowing $\mathrm{R}(T)$ we can find its inverse function,

$$\overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon) \;=\; \inf\{\,T\mid\ \ \mathrm{R}(T)\le\varepsilon\,\}$$

which, similarly to Lemma 2.16, satisfies

$$\lim_{\alpha\to 0+}\overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon-\alpha)\;\ge\;\mathrm{IComp}^{\mathrm{non}}(\varepsilon)\;\ge\;\overline{\mathrm{IComp}}^{\mathrm{non}}(\varepsilon).$$

These inequalities allow to evaluate $\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$.

Notes and Remarks

NR 3.25 First results on adaption versus nonadaption in the average case setting were obtained by Wasilkowski [118] who studied exact information, see also Traub *et al.* [108, Sect. 5.6 of Chap. 6]. The results on adaptive information with noise have been taken mainly from Plaskota [80].

NR 3.26 In terms of $\mathrm{IComp}(\varepsilon)$ and $\mathrm{IComp}^{\mathrm{non}}(\varepsilon)$, the results of Theorem 3.16 mean that for any ε and $0 < p < 1$, at least one of the two following inequalities holds:

$$\mathrm{IComp}(\varepsilon) \;\geq\; \mathrm{IComp}^{\mathrm{non}}\left(\frac{\varepsilon}{\sqrt{p}}\right)$$

or

$$\mathrm{IComp}(\varepsilon) \;\geq\; (1-p)\,\mathrm{IComp}^{\mathrm{non}}(\varepsilon).$$

It turns out that this estimate is sharp. More precisely, it was proven in Plaskota [81] that for exact information (i.e. for the cost function $\mathrm{c} \equiv$ const > 0) the following theorem holds.

Let the nonzero solution operator $S : F \to G$ and the Gaussian measure μ with $\dim \operatorname{supp}\mu = +\infty$ be given. Then there exists a class $\Lambda \subset F^*$ of permissible information functionals such that:

(i) For any $\alpha, \beta > 0$ satisfying $\alpha + \beta > 1$, and for any $\varepsilon_0 > 0$, there exists $\varepsilon < \varepsilon_0$ such that

$$\mathrm{IComp}(\varepsilon) \;<\; \mathrm{IComp}^{\mathrm{non}}\left(\frac{\varepsilon}{\sqrt{\alpha}}\right) \qquad \text{and} \qquad \mathrm{IComp}(\varepsilon) \;<\; \beta\cdot\mathrm{IComp}^{\mathrm{non}}(\varepsilon).$$

(ii) For any $\gamma > 0$ and $\varepsilon_0 > 0$ there exists $\varepsilon < \varepsilon_0$ such that

$$\mathrm{IComp}(\varepsilon) \;<\; \mathrm{IComp}^{\mathrm{non}}\left(\frac{\varepsilon}{\gamma}\right).$$

(iii) For any $\gamma > 0$ and $\varepsilon_0 > 0$ there exists $\varepsilon < \varepsilon_0$ such that

$$\mathrm{IComp}(\varepsilon) \;<\; \gamma\cdot\mathrm{IComp}^{\mathrm{non}}(\varepsilon).$$

Exercises

E 3.39 Let

$$\begin{aligned} h(\varepsilon) \;=\; \inf \quad \{ \quad & \alpha\,\mathrm{IComp}(\sqrt{\varepsilon_1}) + (1-\alpha)\,\mathrm{IComp}(\sqrt{\varepsilon_2}) \mid \\ & 0 \leq \varepsilon_1 \leq \varepsilon \leq \varepsilon_2,\; \alpha\varepsilon_1 + (1-\alpha)\varepsilon_2 = \varepsilon \,\}, \end{aligned}$$

and let $c_{\min} = \inf_{x\geq 0} \mathrm{c}(x)$. Show that the function $h(\varepsilon)$ is convex and

$$h(\varepsilon) \;\leq\; \mathrm{IComp}(\sqrt{\varepsilon}) \;\leq\; h(\varepsilon) \,+\, c_{\min} \qquad \forall \varepsilon \geq 0.$$

E 3.40 Suppose that the function $T \to \mathrm{R}^2(T)$ is semiconvex, i.e., there exist $T_0 \geq 0$, $0 < \alpha \leq \beta$, and a convex function $h : [0, +\infty) \to [0, +\infty)$ such that

$$\alpha \cdot h(T) \;\leq\; \mathrm{R}^2(T) \qquad \forall T \geq 0,$$

and

$$\mathrm{R}^2(T) \;\leq\; \beta \cdot h(T) \qquad \forall T \geq T_0.$$

Show that then for any information $\mathbb{N}$ with $\mathrm{comp}(\mathbb{N}) \leq T$ we have

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) \;\geq\; \sqrt{\frac{\alpha}{\beta}}\,\mathrm{R}(T) \qquad \forall T \geq T_0.$$

3.10 Complexity of special problems

In this section we analyze the ε–complexity of the problems considered in Section 3.8.

3.10.1 Linear problems with Gaussian measures

We begin with the problem defined in Section 3.8.1. That is, $S : F \to G$ is an arbitrary continuous linear operator, μ is a zero mean Gaussian measure and the class Λ consists of linear functionals bounded by 1 in the μ–norm. The technique of evaluating $\mathrm{Comp}(\varepsilon)$ will be similar to that used in Section 2.10.1 where the corresponding problem in the worst case setting was studied. Therefore we only sketch some proofs.

For a given cost function c, we let $\tilde{\mathrm{c}}(x) = c(x^{-1})$, $0 < x < +\infty$. We assume that the function $\tilde{c}$ is concave or convex, and $c(0) = +\infty$.

We recall that $\{\xi_i\}_{i=1}^{\dim G}$ is the complete orthonormal system of eigenelements of $SC_\mu S^*$, $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ are the corresponding eigenvalues, and $K_i^* = \lambda_i^{-1/2} S^* \xi_i$. The function Ω is given by (3.13).

Lemma 3.12 *The Tth minimal radius is equal to*

$$\mathrm{R}(T) \;=\; \sqrt{\inf\,\Omega(\eta_1,\ldots,\eta_n)}$$

where the infimum is taken over all n and $\eta_i \ge 0$, $1 \le i \le n$, satisfying
(a1) for $\tilde{\mathrm{c}}$–concave

$$\sum_{i=1}^n \tilde{c}(\eta_i) \le T,$$

(b1) for $\tilde{\mathrm{c}}$–convex

$$n\,\tilde{c}\left(\frac{1}{n}\sum_{i=1}^n \eta_i\right) \le T.$$

Moreover, if the infimum is achieved for some n^* and $\eta^* = (\eta_1^*,\ldots,\eta_{n^*}^*)$, then

$$\mathrm{R}(T) \;=\; \mathrm{rad}^{\mathrm{ave}}(\,\{N_T,\Sigma_T\}\,)$$

where
(a2) for $\tilde{\mathrm{c}}$–concave

$$\Sigma_T \;=\; \left[\,1/\sqrt{\eta_1^*},\ldots,1/\sqrt{\eta_{n^*}^*}\,\right], \qquad N_T \;=\; [\,K_1^*,\ldots,K_{n^*}^*\,],$$

(b2) for $\tilde{\mathrm{c}}$*–convex*

$$\Sigma_T \;=\; \Big[\,\underbrace{1/\sqrt{\eta_0^*},\ldots,1/\sqrt{\eta_0^*}}_{n^*}\,\Big], \qquad N_T \;=\; [\,L_1^*,\ldots,L_{n^*}^*\,],$$

where $\eta_0^* = 1/n^* \sum_{i=1}^{n^*} \eta_i^*$ *and* L_i^**'s are as in Theorem 3.10 with* $\sigma_i^2 = 1/\eta_0^*$ $\forall\, i$.

Proof The proof goes as the proof of Lemma 2.17. If $\tilde{\mathrm{c}}$ is concave then for any n and $\eta_1,\ldots,\eta_n$ satisfying (3.14) we have $\sum_{i=1}^n \tilde{\mathrm{c}}(\eta_i) \le \sum_{i=1}^n \tilde{c}(1/\sigma_i^2)$. This yields

$$\begin{aligned}\mathrm{R}^2(T) &= \inf\left\{ (\mathrm{r}_n^{\mathrm{ave}}(\sigma_1^2,\ldots,\sigma_n^2))^2 \;\Big|\; n\ge 1,\;\; \sum_{i=1}^n c(\sigma_i^2) \le T \right\} \\ &= \inf\left\{ \Omega(\eta_1,\ldots,\eta_n) \;\Big|\; n\ge 1,\;\; \sum_{i=1}^n \tilde{\mathrm{c}}(\eta_i) \le T \right\}.\end{aligned}$$

On the other hand, for convex $\tilde{\mathrm{c}}$ we have $\sum_{i=1}^n \tilde{\mathrm{c}}(\eta_i) \ge n\,\tilde{\mathrm{c}}(\eta_0)$ where $\eta_0 = 1/n\,\sum_{i=1}^n \eta_i$. Hence,

$$\begin{aligned}\mathrm{R}^2(T) &= \inf\left\{ (\mathrm{r}_n^{\mathrm{ave}}(\underbrace{\sigma^2,\ldots,\sigma^2}_{n}))^2 \;\Big|\; n\ge 1,\; n\,c(\sigma^2) \le T \right\} \\ &= \inf\left\{ \Omega(\eta_1,\ldots,\eta_n) \;\Big|\; n\ge 1,\; n\,\tilde{\mathrm{c}}\left(\frac{1}{n}\sum_{i=1}^n \eta_i\right) \le T \right\}.\end{aligned}$$

The rest of the lemma follows from Theorem 3.10. □

Consider the cost function $\mathrm{c} = \mathrm{c}_{\mathrm{lin}}$. That is, $\mathrm{c}_{\mathrm{lin}}(\sigma^2) = \sigma^{-2}$ for $\sigma^2 > 0$, and $\mathrm{c}_{\mathrm{lin}}(0) = +\infty$. This cost function possesses a similar property as in the worst case – the quality of n observations of $L(f)$ with precisions σ_i^2 depends only on the total cost $\sum_{i=1}^n \sigma_i^{-2}$, and not on the number n of them. Due to Lemma 3.12 we have

$$\mathrm{R}^2(\mathrm{c}_{\mathrm{lin}};T) \;=\; \frac{\left(\sum_{i=1}^n \lambda_i^{1/2}\right)^2}{T+n} + \sum_{j=n+1}^{\infty} \lambda_j \tag{3.55}$$

where $n = n(T)$ is the largest integer satisfying

$$\sum_{i=1}^n \lambda_i^{1/2} \;\le\; \lambda_n^{1/2}\,(T+n). \tag{3.56}$$

Observe that $\mathrm{R}(\mathrm{c}_{\mathrm{lin}};T)$ is well defined since for large n the condition (3.56) is not satisfied. We also have that $\psi(T) = \mathrm{R}^2(\mathrm{c}_{\mathrm{lin}};T)$ is a strictly convex function. To see this, for $n \geq 1$ we let

$$T_n \;=\; \sum_{j=1}^{n} \left(\frac{\lambda_j^{1/2}}{\lambda_n^{1/2}} \,-\, 1 \right) \tag{3.57}$$

(if $\lambda_n = 0$ then $T_n = +\infty$). Then $n = n(T)$ iff $T \in [T_n, T_{n+1})$. On each interval (T_n, T_{n+1}) the function $\psi(T)$ is convex. Hence, for the convexity of ψ on $[0,+\infty)$ it suffices that ψ and $d\psi/dT$ are continuous at T_n. Indeed, due to (3.57) we have

$$\begin{aligned} \psi(T_n^+) \;&=\; \lambda_n(n+T_n) \,+\, \sum_{j=n+1}^{\infty} \lambda_j \\ &=\; \lambda_n(n-1+T_n) \,+\, \sum_{j=n}^{\infty} \lambda_j \;=\; \psi(T_n^-) \end{aligned}$$

and

$$\frac{d\psi}{dT}(T_n^+) \;=\; -\lambda_n \;=\; \frac{d\psi}{dT}(T_n^-).$$

Convexity of $\mathrm{R}^2(\mathrm{c}_{\mathrm{lin}};T)$ implies convexity of $\mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{lin}};\sqrt{\varepsilon})$. Hence, due to Theorem 3.11 we have

$$\mathrm{IComp}(\mathrm{c}_{\mathrm{lin}};\varepsilon) \;=\; \mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{lin}};\varepsilon) \;=\; \inf\{\, T \geq 0 \mid \;\; \mathrm{R}(T) \leq \varepsilon \,\}.$$

If the number $n = n(T)$ defined by (3.56) satisfies $n(T) = O(T)$ $(T \to +\infty)$, then $\mathrm{IComp}(\mathrm{c}_{\mathrm{lin}};\varepsilon)$ is attained by information that uses $O(T)$ observations, and the ε-complexity behaves as $\mathrm{IComp}(\mathrm{c}_{\mathrm{lin}};\varepsilon)$.

Observe that the condition $n(T) = O(T)$ means that zero is not an attraction point of the sequence $n^{-1}\sum_{j=1}^{n}\left(\lambda_j^{1/2}/\lambda_n^{1/2} - 1\right)$. When this is the case, we can show that $\mathrm{c}_{\mathrm{lin}}$ is the "worst" cost function – a result corresponding to Lemma 2.18 of the worst case setting.

Lemma 3.13 *Let* c *be an arbitrary cost function. Let σ_0^2 be such that $\mathrm{c}(\sigma_0^2) < +\infty$. If there exists $a > 0$ such that for sufficiently large n*

$$\frac{1}{n} \sum_{j=1}^{n} \left(\frac{\lambda_j^{1/2}}{\lambda_n^{1/2}} - 1 \right) \;\geq\; a\,, \tag{3.58}$$

then for small $\varepsilon > 0$ we have

$$\mathrm{Comp}(\mathrm{c};\varepsilon) \;\leq\; M\cdot\mathrm{Comp}(\mathrm{c}_{\mathrm{lin}};\varepsilon)$$

where $M = M(\mathrm{c},\sigma_0^2) = a^{-1}\lceil 2\,a\,\sigma_0^2\rceil(\mathrm{c}(\sigma_0^2)+2\mathrm{g})$.

Proof Let n_0 be such that (3.58) holds for all $n \geq n_0$. Let ε_0 satisfy $\varepsilon_0 \leq \mathrm{R}(\mathrm{c}_{\mathrm{lin}}; an_0)$ and $\mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{lin}};\varepsilon_0) \geq a$. We shall show that the required inequality holds for all $\varepsilon < \varepsilon_0$. To this end, we proceed similarly to the proof of Lemma 2.18.

We choose information $\mathbb{N}$ for which $\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}) = \varepsilon$ and $\mathrm{comp}(\mathrm{c}_{\mathrm{lin}};\mathbb{N}) = \mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{lin}};\varepsilon)$. Due to the condition (3.58), we can assume that $\mathbb{N}$ uses $n = \lfloor \mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{lin}};\varepsilon)/a\rfloor$ observations with the same variances σ^2, $\sigma^{-2} = \mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{lin}};\varepsilon)/n$. Let $k = \lfloor 2a\sigma_0^2\rfloor$. Then for the information $\tilde{\mathbb{N}}$ which repeats k times the same observations as $\mathbb{N}$ but with variances $\tilde{\sigma}^2$, $\sigma^{-2} = \sigma^{-2}/k$, we have $\mathrm{rad}^{\mathrm{ave}}(\tilde{\mathbb{N}}) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N})$ and

$$\begin{aligned}\mathrm{comp}(\mathrm{c};\tilde{\mathbb{N}}) \;&\leq\; K\,n\,\tilde{\mathrm{c}}\left(\frac{\mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{lin}};\varepsilon)}{k\,n}\right)\\ &\leq\; k\,n\,\tilde{\mathrm{c}}(2a/k) \;\leq\; a^{-1}k\,\mathrm{c}(\sigma_0^2)\,\mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{lin}};\varepsilon).\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{Comp}(\mathrm{c};\varepsilon) \;&\leq\; a^{-1}k\,\mathrm{c}(\sigma_0^2)\,\mathrm{Comp}(\mathrm{c}_{\mathrm{lin}};\varepsilon) \;+\; (2\,k\,n-1)\mathrm{g}\\ &\leq\; a^{-1}k\,(\mathrm{c}(\sigma_0^2)+2\mathrm{g})\,\mathrm{Comp}(\mathrm{c}_{\mathrm{lin}};\varepsilon),\end{aligned}$$

as claimed. □

We note that the condition (3.58) holds for many sequences $\{\lambda_j\}$ of interest. For instance, for $\lambda_j = j^{-p}$ with $p > 1$ we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\lambda_j^{1/2}}{\lambda_n^{1/2}}-1\right) \;=\; \begin{cases} p/(2-p) & 1<p<2,\\ +\infty & p\geq 2.\end{cases}$$

Hence, we can take $a = 1$. This means, in particular, that $\mathrm{Comp}(\mathrm{c}_{\mathrm{lin}};\varepsilon)$ can be achieved by using no more than $\lfloor\mathrm{Comp}(\mathrm{c}_{\mathrm{lin}};\varepsilon)\rfloor$ observations.

There are however sequences $\{\lambda_j\}$ for which (3.58) is not satisfied, and consequently the Tth minimal radius cannot be achieved by information using $O(T)$ observations. An example is given in E.3.42.

Clearly, when the cost function is bounded from below by a positive constsnt, the lower bound (modulo a constant) on the ε–complexity is provided by $\mathrm{Comp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{exa}};\varepsilon)$ where $\mathrm{c}_{\mathrm{exa}} \equiv 1$ is the cost function for exact information. In this case, letting $n = n(\varepsilon) \geq 0$ to be the minimal n for which

$$\sum_{i=n+1}^{\infty} \lambda_i \;\leq\; \varepsilon^2,$$

we have

$$n(\varepsilon) - 1 \;\leq\; \mathrm{IComp}(\mathrm{c}_{\mathrm{exa}};\varepsilon) \;\leq\; n(\varepsilon) \;=\; \mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{fix}};\varepsilon).$$

Note that $\mathrm{IComp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{exa}};\sqrt{\varepsilon})$ is a semiconvex, but *not* a strictly convex function.

Assume now that the cost function is given as

$$\mathrm{c}_q(\sigma^2) \;=\; \begin{cases} (1+\sigma^{-2})^q & \sigma^2 > 0, \\ +\infty & \sigma^2 = 0, \end{cases}$$

where $q \geq 0$.

Note that for $q = 0$ we have exact information. Assuming (3.58), for $q > 1$ we have $\mathrm{Comp}(q;\varepsilon) \asymp \mathrm{Comp}(1;\varepsilon)$. Therefore in the following calculations we restrict ourselves to $0 < q \leq 1$. Using Lemma 3.12 we obtain

$$\mathrm{R}(q;T)^2 \;=\; \left(\frac{1}{T}\right)^{1/q} \left(\sum_{i=1}^{n} \lambda_i^r\right)^{1/r} \;+\; \sum_{j=n+1}^{\infty} \lambda_j \tag{3.59}$$

where $r = q/(1+q)$ and $n = n(T)$ is the largest integer satisfying

$$\left(1 + \sum_{i=1}^{n-1} \left(\frac{\lambda_i}{\lambda_n}\right)^r\right)^{1/r} - \left(\sum_{i=1}^{n-1} \left(\frac{\lambda_i}{\lambda_n}\right)^r\right)^{1/r} \;\leq\; T^{1/q}.$$

Furthermore, $\mathrm{R}(q;T)$ is attained by observing the functionals $K_1^*,\dots,K_n^*$ with variances

$$\sigma_i^2 \;=\; \left(\lambda_i^{1/(1+q)} \left(\frac{T}{\sum_{j=1}^{n} \lambda_j^r}\right)^{1/q} - 1\right)^{-1}, \qquad 1 \leq i \leq n.$$

Consider now a problem for which the eigenvalues

$$\lambda_j \asymp \left(\frac{\ln^s j}{j}\right)^p$$

where $p > 1$ and $s \geq 0$. Recall that such a behavior of the eigenvalues can be observed for the function approximation with respect to the Wiener sheet measure, see NR 3.10. Then we have

$$\mathrm{R}(q,p,s;T) \asymp \begin{cases} \left(\frac{1}{T}\right)^{1/\tilde{q}} & (p-1)\tilde{q} > 1, \\ \left(\frac{1}{T}\right)^{p-1} (\ln T)^{(s+1)p} & (p-1)\tilde{q} = 1, \\ \left(\frac{1}{T}\right)^{p-1} (\ln T)^{sp} & 0 \leq (p-1)\tilde{q} < 1, \end{cases}$$

as $T \longrightarrow +\infty$, where $\tilde{q} = \min\{1, q\}$. We check that $R(q,p,s;T)^2$ is a semiconvex function of T and that the sequence $\{\lambda_j\}$ satisfies (3.58). Hence, $\mathrm{Comp}^{\mathrm{non}}(q,p,s;\sqrt{\varepsilon})$ is also semiconvex and we obtain the following formulas for the ε–complexity.

Theorem 3.18

$$\mathrm{Comp}^{\mathrm{ave}}(q,p,s;\varepsilon) \asymp \begin{cases} \left(\frac{1}{\varepsilon}\right)^{2\tilde{q}} & (p-1)\tilde{q} > 1, \\ \left(\frac{1}{\varepsilon}\right)^{2/(p-1)} \left(\ln \frac{1}{\varepsilon}\right)^{(s+1)p/(p-1)} & (p-1)\tilde{q} = 1, \\ \left(\frac{1}{\varepsilon}\right)^{2/(p-1)} \left(\ln \frac{1}{\varepsilon}\right)^{sp/(p-1)} & 0 \leq (p-1)\tilde{q} < 1, \end{cases}$$

as $\varepsilon \to 0$. □

The situation is then as for the corresponding problem of the worst case setting (see Theorem 2.20). That is, the complexity may behave, roughly speaking, in only two different ways: as for $q = 1$ (i.e. for the "worst" cost function), or as for $q = 0$ (exact information). Indeed, for $(p-1)\tilde{q} > 1$ we have $\mathrm{Comp}(q,p,s;\varepsilon) \asymp \mathrm{Comp}(1,p,s;\varepsilon)$, while for $(p-1)\tilde{q} < 1$ we have $\mathrm{Comp}(q,p,s;\varepsilon) \asymp \mathrm{Comp}(0,p,s;\varepsilon)$. Furthermore, for $p < 2$, i.e., when the eigenvalues tend to zero sufficiently slowly, the behavior of $\mathrm{Comp}(q,p,s;\varepsilon)$ is independent of q.

3.10.2 Approximation and integration on the Wiener space

We pass to the approximation and integration problems of Section 3.8.2. Recall that both problems are defined on the Wiener space of continuous functions and information consists of noisy observations of function values. In that section we proved tight bounds on the minimal errors $\mathrm{r}_n^{\mathrm{ave}}(\mathrm{App}, \sigma^2)$ and $\mathrm{r}_n^{\mathrm{ave}}(\mathrm{Int}, \sigma^2)$ where $\sigma^2 \geq 0$. They allow to find bounds on the complexity in the case of observations with fixed variance σ_0^2 or, in other words, when the cost function is $\mathrm{c}_{\mathrm{fix}}(\sigma^2) = c_0 > 0$ for $\sigma^2 \geq \sigma_0^2$, and $\mathrm{c}_{\mathrm{fix}}(\sigma^2) = +\infty$ for $\sigma^2 < \sigma_0^2$. Namely, we have $\mathrm{R}(\mathrm{c}_{\mathrm{fix}}; T) = \mathrm{r}_n^{\mathrm{ave}}(\sigma_0^2)$ with $n = n(T) = \lfloor T/c_0 \rfloor$, and due to Corollary 3.5,

$$\mathrm{Comp}^{\mathrm{non}}(\mathrm{App}, \mathrm{c}_{\mathrm{fix}}; \varepsilon) \;\approx\; c_0 \left(\frac{1}{6\varepsilon^2} + p_n^4 \frac{\sigma_0^2}{4\varepsilon^4} \right)$$

and

$$\mathrm{Comp}^{\mathrm{non}}(\mathrm{Int}, \mathrm{c}_{\mathrm{fix}}; \varepsilon) \;\approx\; c_0 \left(\frac{1}{2\sqrt{3}\,\varepsilon} + q_n^2 \frac{\sigma_0^2}{\varepsilon^2} \right)$$

where $p_n, q_n \in [1/\sqrt{3}, 1]$. Since for both problems $\mathrm{Comp}^{\mathrm{non}}(\mathrm{c}_{\mathrm{fix}}; \sqrt{\varepsilon})$ is a semiconvex function, we obtain the following theorem.

Theorem 3.19 *For the cost function* $\mathrm{c}_{\mathrm{fix}}$ *with* $\sigma_0^2 \geq 0$ *we have*

$$\mathrm{Comp}^{\mathrm{ave}}(\mathrm{App}, \mathrm{c}_{\mathrm{fix}}; \varepsilon) \;\asymp\; \frac{1}{\varepsilon^2} + \frac{\sigma_0^2}{\varepsilon^4}$$

and

$$\mathrm{Comp}^{\mathrm{ave}}(\mathrm{Int}, \mathrm{c}_{\mathrm{fix}}; \varepsilon) \;\asymp\; \frac{1}{\varepsilon} + \frac{\sigma_0^2}{\varepsilon^2}$$

where the constants in the "$\asymp$" notation do not depend on σ_0^2. □

It turns out that similar bounds can be proven for the cost function $\mathrm{c}_{\mathrm{lin}}(\sigma^2) = \sigma^{-2}$. Indeed, the upper bound on $\mathrm{Comp}(\mathrm{c}_{\mathrm{lin}}; \varepsilon)$ is provided by $\mathrm{Comp}(\mathrm{c}_{\mathrm{fix}}; \varepsilon)$ with $\sigma_0^2 = 1 = c_0$, while the lower bound follows from the following lemma.

Lemma 3.14 *For all* T *we have*

$$\mathrm{R}(\mathrm{App}, \mathrm{c}_{\mathrm{lin}}; T)^2 \;\geq\; \frac{1}{6\sqrt{T}} - \frac{1}{6T} \;\approx\; \frac{1}{6\sqrt{T}}$$

and

$$\mathrm{R}(\mathrm{Int}, \mathrm{c}_{\mathrm{lin}}; T)^2 \;\geq\; \frac{1}{3(1+T)} \;\approx\; \frac{1}{3T}$$

as $T \to +\infty$.

Proof Let $\mathbb{N}$ be an arbitrary nonadaptive information using observations at t_i's with variances σ_i^2, $1 \le i \le n$, and such that $\mathrm{comp}(\mathrm{c_{lin}}; \mathbb{N}) = \sum_{i=1}^n \sigma_i^{-2} \le T$.

Consider first the approximation problem. Proceeding exactly as in the proof of Lemma 3.10 we can show the following generalization of that lemma. Namely, for any $0 \le a < t < b \le 1$, the covariance kernel of the conditional distribution, $R_N(t,t)$, satisfies

$$R_N(t,t) \;\ge\; \frac{\psi(t)}{1 + T_{ab}\psi(t)}, \tag{3.60}$$

where $\psi(t) = (t-a)(b-t)/(b-a)$, $T_{ab} = \sum \sigma_i^{-2}$, and the summation is taken over all i such that $t_i \in (a,b)$.

We now use (3.60) to obtain the lower bound on $\mathrm{R}(\mathrm{App}, \mathrm{c_{lin}}; T)$. To this end, we divide the unit interval on k equal subintervals (u_{i-1}, u_i), $1 \le i \le k$. For $1 \le i \le n$, let $T_i = \sum_{j \in A_i} \sigma_j^2$ where

$$A_i \;=\; \{\, j \mid \quad 1 \le j \le n,\, t_j \in (u_{i-1}, u_i) \,\}.$$

Denoting $\psi_i(t) = (t - u_{i-1})(u_i - t)/(u_i - u_{i-1})$ and applying (3.27) and (3.60) we obtain

$$(\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, \mathbb{N}))^2 \;\ge\; \sum_{i=1}^k \int_{u_{i-1}}^{u_i} \frac{\psi_i(t)}{1 + T_i/(4k)}\, dt \;=\; \frac{2}{3k} \sum_{i=1}^k \frac{1}{T_i + 4k}.$$

The last quantity, as a function of the nonnegative arguments $T_1, \ldots, T_k$, $\sum_{i=1}^k T_i \le T$, is minimized for $T_i = T/k$. Hence, for any k

$$(\mathrm{rad}^{\mathrm{ave}}(\mathrm{App}, \mathbb{N}))^2 \;\ge\; \frac{2k}{3(T + 4k^2)}.$$

Taking $k = \lfloor \sqrt{T/4} \rfloor$ we obtain the desired bound.

For the integration we have

$$(\mathrm{rad}^{\mathrm{ave}}(\mathrm{Int}, \mathbb{N}))^2 \;\ge\; \frac{\lambda_1}{1+T}$$

where $\lambda_1 = \int_F \mathrm{Int}^2(f)\, w(df) = 1/3$. This completes the proof $\square$.

Thus we have proven the following fact.

Corollary 3.8 *For $S \in \{\mathrm{App}, \mathrm{Int}\}$ we have*

$$\mathrm{Comp}^{\mathrm{ave}}(S, \mathrm{c}_{\mathrm{lin}}; \varepsilon) \asymp \mathrm{Comp}^{\mathrm{ave}}(S, \mathrm{c}_{\mathrm{fix}}; \varepsilon) \qquad \text{as} \quad \varepsilon \to 0^+.$$

Notes and Remarks

NR 3.27 Most of Section 3.10.1 is based on Plaskota [80]. Section 3.10.2 is original.

NR 3.28 We can apply Theorem 3.18 to the multivariate approximation with respect to the Wiener sheet measure – the problem formally defined in NR 3.20. We obtain

$$\mathrm{Comp}^{\mathrm{ave}}(\varepsilon) \asymp \begin{cases} \left(\frac{1}{\varepsilon}\right)^{2\tilde{q}} & \tilde{q} > (r+1/2)^{-1}, \\ \left(\frac{1}{\varepsilon}\right)^{1/(r+1/2)} \left(\ln \frac{1}{\varepsilon}\right)^{k(r+1)/(r+1/2)} & \tilde{q} = (r+1/2)^{-1}, \\ \left(\frac{1}{\varepsilon}\right)^{1/(r+1/2)} \left(\ln \frac{1}{\varepsilon}\right)^{(k-1)(r+1)/(r+1/2)} & \tilde{q} < (r+1/2)^{-1}, \end{cases}$$

where k and r are as in NR 3.20, and $\tilde{q}$ is as in Theorem 3.18.

NR 3.29 Some complexity results for the function approximation and integration with respect to the r–fold Wiener measure can be derived from Plaskota [79], see also NR 3.24. Namely, suppose that the class Λ consists of function values and derivatives of order at most r, and that the cost function $\mathrm{c} = \mathrm{c}_{\mathrm{fix}}$, i.e., observations are performed with the same variance $\sigma_0^2 \geq 0$ and with cost c_0. Then

$$\mathrm{Comp}(\mathrm{App}; \varepsilon) \asymp \left(\frac{\sigma_0}{\varepsilon}\right)^2 + \left(\frac{1}{\varepsilon}\right)^{1/(r+1/2)}$$

and

$$\mathrm{Comp}(\mathrm{Int}; \varepsilon) \asymp \left(\frac{\sigma_0}{\varepsilon}\right)^2 + \left(\frac{1}{\varepsilon}\right)^{1/(r+1)}.$$

The ε–complexity in the case when only observations of function values are allowed, or for other cost functions, is not known.

NR 3.30 We recall that for the solution operator S being a functional we have the correspondence Theorem 3.7. It says that for the corresponding problems the worst case and average case radii of the same information are equal, modulo a constant $\sqrt{2}$. We can formulate an analogous correspondence theorem about the worst and average complexities.

Let $\{H, F\}$ be an abstract Wiener space and μ the associated with it Gaussian measure on F. Let the solution operator $S : F \to \mathbb{R}$ be a continuous linear functional. Let the class Λ of permissible functionals be given. Consider the problem of finding the ε–complexity in the two settings:

P1: The worst case setting with respect to $E = \{f \in H \mid \|f\|_H \le 1\}$, noise bounded in the weighted Euclidean norm, $\sum_{i=1}^n (y_i - L_i(f))^2/\delta_i^2 \le 1$, and a cost function $\mathrm{c}_w(\delta)$,

P2: The average case setting with respect to the measure μ, independent noise with $(y_i - L_i(f)) \sim \mathcal{N}(0, \sigma_i^2)$, and a cost function $\mathrm{c}_a(\sigma^2)$.

If for $\delta^2 = \sigma^2$ is $\mathrm{c}_w(\delta) = \mathrm{c}_a(\sigma^2)$ then

$$(\mathrm{IComp}^{\mathrm{non}})^{\mathrm{wor}}(\sqrt{2}\,\varepsilon) \;\le\; (\mathrm{IComp}^{\mathrm{non}})^{\mathrm{ave}}(\varepsilon) \;\le\; (\mathrm{IComp}^{\mathrm{non}})^{\mathrm{wor}}(\varepsilon).$$

If, moreover, $(\mathrm{IComp}^{\mathrm{non}})^{\mathrm{ave}}(\sqrt{\varepsilon})$ is semiconvex and $(\mathrm{IComp}^{\mathrm{non}})^{\mathrm{ave}}(\sqrt{2}\varepsilon)$ behaves as $(\mathrm{IComp}^{\mathrm{non}})^{\mathrm{ave}}(\varepsilon)$, then

$$\mathrm{Comp}^{\mathrm{wor}}(\varepsilon) \;\asymp\; \mathrm{Comp}^{\mathrm{ave}}(\varepsilon) \qquad \text{as} \quad \varepsilon \to 0^+.$$

For instance, the results of Section 3.10.2 can be applied to get complexity results for the corresponding problem in the worst case setting (compare also with NR 3.22).

Exercises

E 3.41 Show that the condition $\sum_{j=1}^\infty \lambda_j^{1/2} < +\infty$ implies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n \left(\frac{\lambda_j^{1/2}}{\lambda_n^{1/2}} - 1 \right) \;=\; +\infty.$$

That is, for such eigenvalues Lemma 3.13 can be applied.

E 3.42 Let $1/2 < p < 1$. Let $a_n = n^{-p}$ and $P_n = n^{-1} \sum_{i=1}^n a_i/a_n$, $n \ge 1$. For $\alpha_1 > \alpha_2 > \cdots \to 0$, let $0 = n_0 < n_1 < \cdots$ be the sequence of integers defined inductively by the condition

$$P_{n_i - 1} \left(\frac{n_{i-1}}{n_i} \right)^{1-p} - \frac{n_{i-1}}{n_i} \;<\; \alpha_i$$

($P_0 = 0$). Finally, for $n \ge 1$ we let $\lambda_n = a_{n_i}^2$, where i is the unique positive integer such that $n_{i-1} < n \le n_i$. Show that for any n satisfying

$$\sum_{i=1}^n \lambda_i^{1/2} \;\ge\; \lambda_n^{1/2}(T_i + n)$$

with $T_i = \alpha_i n_i$, we have $n/T_i \ge 1/\alpha_i \to +\infty$ as $i \to +\infty$.

E 3.43 Let $0 < q \le 1$. Show that $\mathrm{Comp}^{\mathrm{non}}(q, p, s; \sqrt{\varepsilon})$ is *not* a strictly convex function of ε.

E 3.44 Show that for the cost function

$$\mathrm{c}_1(\sigma^2) \;=\; \begin{cases} 1+\sigma^{-2} & \sigma^2 > 0, \\ 0 & \sigma^2 = 0, \end{cases}$$

we have

$$\mathrm{R}(\mathrm{c}_1; T)^2 \;=\; \frac{1}{T} \cdot \left(\sum_{i=1}^{n} \lambda_i^{1/2} \right)^2 \;+\; \sum_{j=n+1}^{\infty} \lambda_j$$

where $n = n(T)$ is the largest integer satisfying

$$\sum_{i=1}^{n} \lambda_i^{1/2} \;\le\; \lambda_n^{1/2} \left(\frac{T+1}{2} \right).$$

Chapter 4

First mixed setting

4.1 Introduction

In the previous two sections, we studied settings in which we have exclusively deterministic assumptions on the problem elements f and information noise x (worst case setting), or exclusively stochastic assumptions (average case setting). In the first case we analyze the worst peformance of algorithms, while in the other we are interested in the average performance. The deterministic and stochastic assumptions can be combined to obtain *mixed settings.*

In this chapter we study the first mixed setting. We want to approximate values $S(f)$ of a solution operator, for elements f belonging to a set $E \subset F$. Information about f is given with random noise. That is, a nonadaptive or adaptive information operator is defined as in the average case setting of Chapter 3. The error of an algorithm φ that uses information $\mathbb{N}$ is given as

$$\mathrm{e}^{\mathrm{w-a}}(\mathbb{N},\varphi) \;=\; \sup_{f\in E} \sqrt{\int_Y \|S(f)-\varphi(y)\|^2 \pi_f(dy)},$$

where Y is the set of all possible values y of noisy information, and $\pi_f = \mathbb{N}(f)$ is the distribution of y for the element f.

This setting has been extensively studied in statistics. Therefore it is often called *statistical estimation* and the problem of minimizing the error over a class of algorithms – the minimax (statistical) problem. As this setting is one of several settings we study in this monograph, in order to keep our terminology consistent we use the name mixed or *worst–average case setting* which is justified by the definition of error.

In the mixed settings, the complexity results are not as rich as in the worst or average settings. The reason for that lies in the technical difficulty. For instance, even for apparently simple one–dimensional problems, optimal algorithms turn out to be nonlinear (nonaffine) and they are actually not known exactly.

This chapter consists of three sections. In Section 4.2, we study approximation of a linear functional over a convex set E. We consider nonadaptive linear information with Gaussian noise. It turns out that, although optimal algorithms are nonaffine, we lose at most $11,1\ldots\%$ by using affine algorithms. Hence, once more affine approximations prove to be (almost) optimal. Optimal affine algorithms are constructed. These results are obtained by using the concept of a hardest one–dimensional subproblem, and by establishing a relation between the worst-average and worst case settings. In particular, it turns out that appropriately celebrating the levels of random noise in one setting and deterministic noise in the other setting, we get the same optimal affine algorithm.

If E is the unit ball in a Hilbert norm, there are also close relations between the worst-average and the corresponding average case settings. This enables us to show almost equivalence of the three settings. In any of them the same smoothing spline algorithm is almost optimal.

The situation becomes much more complicated when the solution operator is not a functional. This case is considered in Section 4.3. We present only some special results about optimal algorithms when, roughly speaking, information is given "coordinatewise". In particular, we show optimality of the least squares when $E = \mathbb{R}^d$. For arbitrary information, optimal algorithms are unknown, even for problems defined on Hilbert spaces.

4.2 Affine algorithms for linear functionals

For approximating a linear functional in the worst and average case settings, optimal algorithms often turn out to be linear or affine. In this section, we investigate whether a similar result holds in the mixed worst–average case setting.

To begin with, we consider a one–dimensional problem. We shall see that even in this simple case the situation is rather complicated.

4.2.1 The one–dimensional problem

Consider the problem of approximating a real parameter $f \in [-\tau, \tau]$ from data $y = f + x$ where x is distributed according to the zero mean one-dimensional Gaussian measure with variance $\sigma^2 \geq 0$. That is, we formally have $S : \mathbb{R} \to \mathbb{R}$, $S(f) = f$, and $\mathbb{N}(f) = \mathcal{N}(f, \sigma^2)$. Observe that the problem of approximating a linear functional $S : F \to \mathbb{R}$ from data $y = S(f) + x$, $x \sim \mathcal{N}(0, \sigma^2)$, and for $f \in E$–a balanced and convex set, reduces to this case. Indeed, then we approximate $g = S(f) \in \mathbb{R}$ from information $y = g+x$ where $|g| \leq \tau = \sup_{f \in E} S(f)$.

To avoid the trivial case, we assume $\tau > 0$. Clearly, for $\sigma^2 = 0$ we have exact information. For any f the algorithm $\varphi(y) = y$ gives exact value of $S(f)$ with probability 1 and its error is zero. For $\sigma^2 > 0$, the error of any algorithm $\varphi : \mathbb{R} \to \mathbb{R}$ is positive and given as

$$\begin{aligned} \mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi) &= \mathrm{e}^{\mathrm{w-a}}(\tau, \sigma^2; \varphi) \\ &= \sup_{\|f\| \leq \tau} \sqrt{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} |f - \varphi(f + x)|^2 \exp\{-x^2/(2\sigma^2)\}\, dx}. \end{aligned}$$

Consider first linear algorithms. That is, assume that φ is of the form $\varphi(y) = c\, y$ for all $y \in \mathbb{R}$. Let

$$r_{\mathrm{lin}}(\tau, \sigma^2) \;=\; \inf \{\, \mathrm{e}^{\mathrm{w-a}}(\tau, \sigma^2; \varphi) \mid \quad \varphi - \text{linear } \}$$

be the minimal error of linear algorithms.

Lemma 4.1 *For any τ and σ^2 we have*

$$r_{\mathrm{lin}}(\tau, \sigma^2) \;=\; \sigma \cdot \sqrt{\frac{\tau^2}{\tau^2 + \sigma^2}}.$$

The optimal coefficient $c_{\mathrm{opt}} = c_{\mathrm{opt}}(\tau, \sigma^2)$ of a linear algorithm is unique and given as

$$c_{\mathrm{opt}}(\tau, \sigma^2) \;=\; \frac{\tau^2}{\tau^2 + \sigma^2}.$$

Proof We have already noticed that the lemma is true for $\sigma^2 = 0$. Let $\sigma^2 > 0$. Then for any linear algorithm $\varphi(y) = cy$ and $f \in \mathbb{R}$ we have

$$\begin{aligned} (\mathrm{e}^{\mathrm{w-a}}(\tau, \sigma^2; \varphi))^2 &= \sup_{|f| \leq \tau} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} |f - \varphi(f + x)|^2\, e^{-x^2/(2\sigma^2)}\, dx \\ &= \sup_{|f| \leq \tau} f^2(1 - c)^2 \,+\, \sigma^2 c^2 \;=\; \tau^2(1 - c)^2 \,+\, \sigma^2 c^2. \end{aligned}$$

The lemma now follows by taking the minimum of the last expression over $c \in \mathbb{R}$. □

Hence, the optimal coefficient $c_{\rm opt}$ is determined uniquely and it is a function of σ/τ, i.e., $c_{\rm opt}(\tau,\sigma^2) = c_{\rm opt}(1,\sigma^2/\tau^2)$. For the minimal error we have

$$r_{\rm lin}(\tau,\sigma^2) \;=\; \tau \cdot r_{\rm lin}(1,\sigma^2/\tau^2). \tag{4.1}$$

Furthermore, for $\sigma^2 \to 0$ we have $r_{\rm lin}(\tau,\sigma^2) \approx \sigma$, and for $\sigma^2 \to \infty$ we have $r_{\rm lin}(\tau,\sigma^2) \to \tau$.

Obviously, the linear algorithm $y \to c_{\rm opt} y$ is optimal also in the class of affine algorithms. However, if we consider arbitrary algorithms, then it is not difficult to see that we can do better.

Example 4.1 Observe that for large $|y|$, $|y| > \tau + \sigma^2/\tau$, we have $c_{\rm opt} y \notin [-\tau,\tau]$, and $\tau\, sgn(y)$ provides better approximation to any f from $[-\tau,\tau]$ than $c_{\rm opt} y$. Hence, for the nonlinear algorithm

$$\varphi_{\rm non}(y) \;=\; \begin{cases} c_{\rm opt}(\tau,\sigma^2)\, y & |y| \le \tau + \sigma^2/\tau, \\ \tau \cdot sgn(y) & |y| > \tau + \sigma^2/\tau, \end{cases}$$

we have ${\rm e}^{\rm w-a}(\tau,\sigma^2;\varphi_{\rm non}) < r_{\rm lin}(\tau,\sigma^2)$. □

The fact that nonlinear algorithms are better than linear ones should be contrasted to the results of worst and average case settings where, for the corresponding problems, linear algorithms are optimal, see E 2.13 and E 3.19.

It turns out, however, that we never gain much. Namely, let

$$r_{\rm arb}(\tau,\sigma^2) \;=\; \inf\{\, {\rm e}^{\rm w-a}(\tau,\sigma^2;\varphi) \mid \quad \varphi - \text{arbitrary} \,\}$$

be the minimal error of arbitrary algorithms.

Theorem 4.1 *We have*

$$\lim_{\sigma^2/\tau^2 \to 0} \frac{r_{\rm lin}(\tau,\sigma^2)}{r_{\rm arb}(\tau,\sigma^2)} \;=\; 1 \;=\; \lim_{\sigma^2/\tau^2 \to \infty} \frac{r_{\rm lin}(\tau,\sigma^2)}{r_{\rm arb}(\tau,\sigma^2)}.$$

Furthermore, there exists an absolute constant κ_1 such that

$$1 \;\le\; \frac{r_{\rm lin}(\tau,\sigma^2)}{r_{\rm arb}(\tau,\sigma^2)} \;\le\; \kappa_1 \qquad \forall\, \tau,\sigma^2.$$

Proof Without loss of generality we can restrict ourselves to the case $\tau = 1$. Indeed, setting $\tilde{f} = f/\tau$, $\tilde{x} = x/\tau$, and for arbitrary φ, $\tilde{\varphi}(y) = \varphi(\tau y)/\tau$, we get $\mathrm{e}^{\mathrm{w-a}}(\tau, \sigma^2; \varphi) = \tau \mathrm{e}^{\mathrm{w-a}}(1, \sigma^2/\tau^2; \tilde{\varphi})$. Hence,

$$r_{\mathrm{arb}}(\tau, \sigma^2) \;=\; \tau \cdot r_{\mathrm{arb}}(1, \sigma^2/\tau^2).$$

This and (4.1) yield

$$\frac{r_{\mathrm{lin}}(\tau, \sigma^2)}{r_{\mathrm{arb}}(\tau, \sigma^2)} \;=\; \frac{r_{\mathrm{lin}}(1, \sigma^2/\tau^2)}{r_{\mathrm{arb}}(1, \sigma^2/\tau^2)}.$$

To obtain the first limit in the theorem it suffices to show $r_{\mathrm{arb}}(1, \sigma^2) \approx \sigma$. To this end, observe that for any φ we have

$$\begin{aligned}(\mathrm{e}^{\mathrm{w-a}}(1, \sigma^2; \varphi))^2 \;&\geq\; \frac{1}{2}\int_{-1}^{1}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}}(f - \varphi(y))^2 e^{-\frac{(y-f)^2}{2\sigma^2}} dy\right\} df \\ &=\; \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}}\left\{\int_{-1}^{1}(f - \varphi(y))^2 e^{-\frac{(y-f)^2}{2\sigma^2}} df\right\} dy. (4.2)\end{aligned}$$

The integral in the last parenthesis is minimized by

$$\varphi_1(y) \;=\; \frac{\int_{-1}^{1} x e^{-\frac{(y-x)^2}{2\sigma^2}} dx}{\int_{-1}^{1} e^{-\frac{(y-x)^2}{2\sigma^2}} dx} \;=\; y \;-\; \sigma\,\frac{\int x e^{-x^2/2} dx}{\int e^{-x^2/2} dx}$$

where the integrals are taken from $(y-1)/\sigma$ to $(y+1)/\sigma$. Put $\varphi = \varphi_1$ and change variables in (4.2), $y = f + \sigma u$. After some calculations we get

$$(\mathrm{e}^{\mathrm{w-a}}(1, \sigma^2; \varphi))^2 \;\geq\; \sigma^2 \cdot \frac{1}{2}\int_{-1}^{1}\left\{\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \psi_1^2(f, u, \sigma^2)\, e^{-u^2/2}\, du\right\} df$$

where

$$\psi_1(f, u, \sigma^2) \;=\; \frac{\int (u - x)\, e^{-x^2/2} dx}{\int e^{-x^2/2} dx},$$

the integrals taken from $u + (f-1)/\sigma$ to $u + (f+1)/\sigma$. Observe now that for $|f| \leq a < 1$ and $|u| < A < +\infty$, the function $\psi_1(f, u, \sigma^2)$ converges uniformly to u as $\sigma^2 \to 0$. Hence,

$$\begin{aligned}\lim_{\sigma^2 \to 0} \quad & \frac{1}{2}\int_{-1}^{1}\left\{\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \psi_1^2(f, u, \sigma^2) e^{-u^2/2} du\right\} df \\ =\; & \frac{1}{2}\int_{-1}^{1}\left\{\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} u^2 e^{-u^2/2} du\right\} df \;=\; 1.\end{aligned}$$

Consequently, $r_{\mathrm{arb}}(1,\sigma^2) \approx \sigma$, as claimed.

On the other hand, the error of φ satisfies

$$\begin{aligned}&(\mathrm{e}^{\mathrm{w-a}}(1,\sigma^2;\varphi))^2 \\ \geq\ & \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}}(f-\varphi(y))^2 e^{-\frac{(y-f)^2}{2\sigma^2}} + (f+\varphi(y))^2 e^{-\frac{(y+f)^2}{2\sigma^2}}\,dy \qquad (4.3)\end{aligned}$$

where f is arbitrary from $[-1,1]$. This is minimized by

$$\varphi_2(y) \ = \ \frac{a_- - a_+}{a_- + a_+}\,f, \qquad a_\pm = e^{-\frac{(y\pm f)^2}{2\sigma^2}}. \qquad (4.4)$$

Putting $\varphi = \varphi_2$ in (4.3) we obtain

$$(\mathrm{e}^{\mathrm{w-a}}(1,\sigma^2;\varphi))^2 \ \geq \ f^2\,\psi(f/\sigma) \qquad (4.5)$$

where

$$\psi(x) \ = \ e^{-x^2/2}\sqrt{\frac{2}{\pi}}\int_0^\infty \frac{e^{-u^2/2}}{\cosh(ux)}\,du.$$

Take $f=1$. Then

$$(\mathrm{e}^{\mathrm{w-a}}(1,\sigma^2;\varphi))^2 \ \geq \ \psi(1/\sigma) \ \geq \ \frac{e^{-\frac{1}{2\sigma^2}}}{\cosh(1/\sqrt{\sigma})}\sqrt{\frac{2}{\pi}}\int_0^{\sqrt{\sigma}} e^{-u^2/2}\,du$$

which tends to 1 as $\sigma^2 \to \infty$. This yields the second limit of the theorem.

It remains to show existence of κ_1. For $\sigma^2 \geq 1$ we have $r_{\mathrm{lin}}(1,\sigma^2) \leq 1$ and $r_{\mathrm{arb}}(1,\sigma^2) \geq \sqrt{\psi(1)}$, where the last inequality follows from (4.5) and monotonicity of ψ. On the other hand, for $\sigma^2 < 1$ we have $r_{\mathrm{lin}}(1,\sigma^2) \leq \sigma$ and $r_{\mathrm{arb}}(1,\sigma^2) \geq \sigma\sqrt{\psi(1)}$, where this time the last inequality follows from (4.5) by taking $f=\sigma$. Thus, for any σ^2

$$\frac{r_{\mathrm{lin}}(1,\sigma^2)}{r_{\mathrm{arb}}(1,\sigma^2)} \ \leq \ \frac{1}{\sqrt{\psi(1)}}$$

and we can take $\kappa_1 = \psi^{-1/2}(1)$. □

Let us now define the constant

$$\kappa_1^* \ = \ \sup_{\tau,\sigma^2}\frac{r_{\mathrm{lin}}(\tau,\sigma^2)}{r_{\mathrm{arb}}(\tau,\sigma^2)}. \qquad (4.6)$$

We showed that $1 < \kappa_1^* \leq \psi^{-1/2}(1) = 1.49\ldots$. Actually, the value of κ_1^* is known much more precisely, see NR 4.2.

4.2.2 Almost optimality of affine algorithms

We pass to the general problem. We assume that the functional S is defined on a linear space F. We want to approximate $S(f)$ for f belonging to a convex set $E \subset F$ based on linear information with Gaussian noise. That is, we have at disposal information $y = N(f) + x \in \mathbb{R}^n$ where $N : F \to Y = \mathbb{R}^n$ is a linear operator and the noise $x \sim \mathcal{N}(0, \sigma^2 \Sigma)$. The symmetric matrix $\Sigma \in \mathbb{R}^{n \times n}$ is assumed to be positive definite. It induces the inner product $\langle \cdot, \cdot \rangle_Y$ in $\mathbb{R}^n$, $\langle y, z \rangle_Y = \langle \Sigma^{-1} y, z \rangle_2$.

We denote by $\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{aff}}(\mathbb{N}; E)$ and $\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}; E)$ the minimal errors of affine and arbitrary algorithms over E,

$$\begin{aligned} \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{aff}}(\mathbb{N}; E) &= \inf \{ \mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi; E) \mid \quad \varphi - \text{affine } \}, \\ \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}; E) &= \inf \{ \mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi; E) \mid \quad \varphi - \text{arbitrary } \}. \end{aligned}$$

Algorithms that attain the first and second infimum will be called optimal affine and optimal, respectively.

We need the following fact.

Lemma 4.2 *Consider the one–dimensional problem of Section 4.2.1 with $f \in [-\tau, \tau]$ and data $y = f + x$, $x \sim \mathcal{N}(0, \sigma^2)$. Suppose we allow algorithms which additionally use some (independent of y) "pure noise" data $t \in T$ where $T = \mathbb{R}^k$ and $t \sim \omega = \mathcal{N}(0, \sigma^2 I)$, say. Then we cannot make use of t and best affine and arbitrary algorithms use y alone.*

Proof This is the consequence of a more general fact. Namely, suppose that for the corresponding class $\mathcal{A}$ of algorithms which use only y, there exists a least favorable probability distribution μ on E for which

$$\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}, \mathcal{A}; \mu) \;=\; \mathrm{rad}^{\mathrm{w-a}}(\mathbb{N}, \mathcal{A}; E). \tag{4.7}$$

Let $\varphi : Y \times T \to G$ be an arbitrary algorithm using also t, such that $\varphi(\cdot, t) \in \mathcal{A}$, $\forall t$. Denote

$$e^2(f, t) \;=\; \int_Y (S(f) - \varphi(N(f) + x, t))^2 \, \pi(dx) \,.$$

Then, using the mean value theorem we obtain

$$\begin{aligned} (\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}; \varphi(\cdot, \cdot)))^2 &= \sup_{f \in E} \int_T e^2(f, t) \, \omega(dt) \;\geq\; \int_E \int_T e^2(f, t) \, \omega(dt) \, \mu(df) \\ &= \int_T \int_E e^2(f, t) \, \mu(df) \, \omega(dt) \;\geq\; \int_E e^2(f, t^*) \, \mu(df) \\ &\geq (\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}, \mathcal{A}; \mu))^2 \;\geq\; (\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N}, \mathcal{A}; E))^2. \end{aligned}$$

For the one–dimensional problem under consideration the least favorable μ satisfying (4.7) exists. For affine algorithms it puts equal mass at $\pm\tau$, $\mu(\{-\tau\}) = \mu(\{\tau\}) = 1/2$, which follows from the fact that the error of any linear algorithm is attained at the end points. For arbitrary algorithms, μ is concentrated on a finite set of points, see NR 4.1. □

Consider now the case where E is an interval. That is, $E = I = \{\,\alpha f_{-1} + (1-\alpha) f_1 \mid 0 \le \alpha \le 1\,\}$ for some $f_{-1}, f_1 \in F$.

Lemma 4.3 *Let E be the one–dimensional set, $E = I(f_{-1}, f_1)$. Let $h = (f_1 - f_{-1})/2$ and $f_0 = (f_1 + f_{-1})/2$.*

(i) If $N(h) = 0$ then

$$\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{aff}}(\mathbb{N}, I) \;=\; \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}; I) \;=\; |S(h)|$$

and the optimal algorithm is $\varphi \equiv S(f_0)$.

(ii) If $N(h) \neq 0$ then

$$\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{aff}}(\mathbb{N}; I) \;=\; |S(h)|\, r_{\mathrm{lin}}\left(1, \frac{\sigma^2}{\|N(h)\|_Y^2}\right),$$

$$\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}; I) \;=\; |S(h)|\, r_{\mathrm{arb}}\left(1, \frac{\sigma^2}{\|N(h)\|_Y^2}\right),$$

and the optimal affine algorithm is given as

$$\varphi_{\mathrm{aff}}(y) \;=\; S(f_0) \,+\, c_{\mathrm{opt}}\left(1, \frac{\sigma^2}{\|N(h)\|_Y^2}\right) \frac{S(h)}{\|N(h)\|_Y} \left\langle y - N(f_0), \frac{N(h)}{\|N(h)\|_Y} \right\rangle_Y$$

where $r_{\mathrm{lin}}(\cdot,\cdot)$, $r_{\mathrm{arb}}(\cdot,\cdot)$ and $c_{\mathrm{opt}}(\cdot,\cdot)$ are as in Lemma 4.1.

Proof (i) Since for any $f \in I$ we have $N(f) = N(f_0)$, information consists of pure noise only. Hence, in view of Lemma 4.2, such information is useless. The optimal algorithm is the center of $S(I)$ and the formula for the minimal error follows.

(ii) For $f \in I$, let $\alpha = \alpha(f)$ be defined by $f = f_0 + \alpha h$. Clearly, $f \in I$ iff $|\alpha| \le 1$. Transform the data $y = N(f) + x$ to

$$z \;=\; \frac{\Sigma^{-1/2}(y - N(f_0))}{\|N(h)\|_Y} \;=\; \alpha\, w \,+\, x',$$

where $w = (\Sigma^{-1/2}N(h))/\|N(h)\|_Y$ and

$$x' \;=\; \frac{\Sigma^{-1/2}x}{\|N(h)\|_Y} \;\sim\; \mathcal{N}\left(0, \frac{\sigma^2}{\|N(h)\|_Y}\, I\right).$$

We now choose Q to be such an orthogonal matrix that $Qw = e_1$ (the first versor). Then the problem of approximating $S(f)$ from data y is equivalent to that of approximating $s(\alpha) = S(f_0) + \alpha\, S(h)$, $-1 \le \alpha \le 1$, from data

$$\tilde{y} \;=\; Q\,z \;=\; [\,\alpha + \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n] \in \mathbb{R}^n$$

where $\tilde{x}_i$ are independent, $\tilde{x}_i \sim \mathcal{N}(0, \sigma^2/\|N(h)\|_Y^2)$, $1 \le i \le n$. We see that only the first component of $\tilde{y}$ is not pure noise. From Lemma 4.2 it follows that we cannot make use of $\tilde{x}_2, \ldots, \tilde{x}_n$ to reduce the error, and we can restrict ourselves to data $\tilde{y}_1 = \alpha + \tilde{x}_1$.

Thus we have reduced the original problem to the one dimensional problem of approximating $s(\alpha)$ from $\tilde{y}_1 = \alpha + \tilde{x}_1$. The formulas for the minimal errors now follow from Lemma 4.1. The optimal affine algorithm is given as

$$\varphi_{\mathrm{aff}}(y) \;=\; S(f_0) \;+\; S(h)\, c_{\mathrm{opt}}\left(1, \frac{\sigma^2}{\|N(h)\|_Y^2}\right)\, \tilde{y}_1.$$

To complete the proof, observe that

$$\begin{aligned} \tilde{y}_1 \;&=\; \langle Qz, e_1\rangle_2 \;=\; \langle z, Q^{-1}e_1\rangle_2 \;=\; \langle z, w\rangle_2 \\ &=\; \left\langle \frac{\Sigma^{-1/2}(y - N(f_0))}{\|N(h)\|_Y}, \frac{\Sigma^{-1/2}N(h)}{\|N(h)\|_Y} \right\rangle_2 \\ &=\; \frac{1}{\|N(h)\|_Y} \left\langle y - N(f_0), \frac{N(h)}{\|N(h)\|_Y} \right\rangle_Y. \quad \square \end{aligned}$$

We now find optimal affine algorithms for arbitrary convex set E. For $\delta \ge 0$, let

$$r(\delta) \;=\; \sup\{\, S(h) \mid \quad h \in \mathrm{bal}(E),\; \|N(h)\|_Y \le \delta\,\}$$

($\mathrm{bal}(E) = (E-E)/2$). Recall that $r(\delta)$ is the worst case radius of information $\mathbb{N}_\delta(f) = \{\, N(f) + x \mid \|x\|_Y \le \delta\}$ with respect to E, $r(\delta) = \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}_\delta; E)$. Let φ_δ be the worst case optimal affine algorithm for information $\mathbb{N}_\delta$. We know from Section 2.4 that it exists and it is optimal among all algorithms, $\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}_\delta; E) = \mathrm{e}^{\mathrm{wor}}(\mathbb{N}_\delta, \varphi_\delta; E)$. Moreover, φ_δ is of the form

$$\varphi_\delta \;=\; g_\delta \;+\; d_\delta\, \langle \cdot, w_\delta\rangle_Y$$

where $g_\delta \in \mathbb{R}$, $w_\delta \in \mathbb{R}^n$ with $\|w_\delta\|_Y = 1$, and $d_\delta \ge 0$ is an arbitrary number such that $r(\gamma) \le r(\delta) + d_\delta(\gamma - \delta)$, $\forall \gamma$. The set of all such d_δ will be denoted by $\partial r(\delta)$. (Compare with Theorem 2.6.)

Observe that taking $\delta^2 = \sigma^2$, we obtain an algorithm which is close to optimal affine in the mixed worst–average case. Indeed, for any affine $\varphi = g + d\,\langle \cdot, w\rangle_Y$ with $\|w\|_Y = 1$, we have

$$\begin{aligned} & |S(f) - \varphi(N(f)+x)|^2 \\ = \; & |S(f) - \varphi(N(f))|^2 \; - \; 2\,dS(f)\,\langle w, \Sigma^{-1}x\rangle_2 \; + \; d^2\langle w, \Sigma^{-1}x\rangle_2^2. \end{aligned}$$

If we integrate this over $x \sim \pi = \mathcal{N}(0, \sigma^2\Sigma)$, the second component will vanish and the third one will become $\sigma^2 d^2$. Hence,

$$\begin{aligned} \mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi; E) \; &= \; \sup_{f \in E} \sqrt{\int_{\mathbb{R}^n} |S(f) - \varphi(N(f)+x)|^2 \, \pi(dx)} \\ &= \; \sqrt{\sup_{f \in E} |S(f) - \varphi(N(f))|^2 \; + \; \sigma^2\, d^2}. \end{aligned} \tag{4.8}$$

Since $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi; E) = \sup_{f \in E} |S(f) - \varphi(N(f))| + \delta\, d$, for $\delta^2 = \sigma^2$ we have

$$\frac{1}{\sqrt{2}}\,\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi; E) \; \le \; \mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi; E) \; \le \; \mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi; E).$$

In particular, this implies

$$\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_\delta; E) \; \le \; \sqrt{2} \cdot \mathrm{rad}^{\mathrm{w-a}}(\mathbb{N}; E) \qquad (\delta^2 = \sigma^2).$$

It turns out that for appropriately chosen δ the algorithm φ_δ is strictly optimal affine.

Theorem 4.2 *Let $\sigma^2 > 0$. Suppose that there exist $\delta = \delta(\sigma) > 0$ and $d_\delta \in \partial r(\delta)$ such that*

$$d_\delta \; = \; \frac{\delta\, r(\delta)}{\sigma^2 \, + \, \delta^2}. \tag{4.9}$$

Then the algorithm φ_δ is optimal affine in the mixed worst–average setting and

$$\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N}; E) \; = \; \mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_\delta; E) \; = \; \frac{\sigma\, r(\delta)}{\sqrt{\sigma^2 + \delta^2}}.$$

Proof For $\varepsilon > 0$, let $h = (f_1 - f_{-1})/2 \in \mathrm{bal}(E)$, $f_1, f_{-1} \in E$, be such that $\|N(h)\|_Y \le \delta$ and $S(h) \ge r(\delta) - \varepsilon$. Let $I = I(f_{-1}, f_1)$. Then

$$\sup_{f\in E} |S(f) - \varphi_\delta(N(f))| \;\le\; \sup_{f\in I} |S(f) - \varphi_\delta(N(f))| \,+\, \varepsilon.$$

This and (4.8) yield

$$\begin{aligned} (\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_\delta; E))^2 \;&=\; \sup_{f\in E} |S(f) - \varphi_\delta(N(f))|^2 \,+\, \sigma^2 d_\delta^2 \\ &\le\; (\sup_{f\in I} |S(f) - \varphi_\delta(N(f))| + \varepsilon\,)^2 \,+\, \sigma^2 d_\delta^2. \end{aligned}$$

Since ε can be arbitrarily small, the last inequality and the formula (4.9) for d_δ give

$$\begin{aligned} (\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_\delta; E))^2 \;&\le\; \sup_{f\in I} |S(f) - N(f)|^2 \,+\, \sigma^2 d_\delta^2 \\ &=\; (r(\delta) - \delta\, d_\delta)^2 \,+\, \sigma^2 d_\delta^2 \;=\; \frac{\sigma^2\, r^2(\delta)}{\sigma^2 + \delta^2}. \end{aligned} \tag{4.10}$$

On the other hand, the error over E is not smaller than the error over the interval I. Using the formula for $\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N}; I)$ given in Lemma 4.3 we obtain

$$\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N}; E) \;\ge\; \mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N}; I) \;\ge\; \frac{\sigma\,(r(\delta) - \varepsilon)}{\sqrt{\sigma^2 + \delta^2}},$$

and since ε is arbitrary,

$$\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N}; E) \;\ge\; \frac{\sigma\, r(\delta)}{\sqrt{\sigma^2 + \delta^2}}. \tag{4.11}$$

The theorem now follows from (4.10) and (4.11). □

Hence, under the assumption (4.9), the optimal algorithm in the mixed setting turns out to be optimal in the worst case with appropriately chosen δ.

Observe that in the proof we also showed that the minimal linear worst–average error over E equals the minimal linear worst–average error over the hardest one–dimensional subset $I \subset E$. We emphasize this important fact in the following corollary.

Corollary 4.1 *If there exist $\delta = \delta(\sigma)$ and $d_\delta \in \partial r(\delta)$ satisfying (4.9) then*

$$\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N};E) \;=\; \sup_{I\subset E} \mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N};I).$$

Furthermore, if the worst case radius $r(\delta)$ is attained at $h^* = (f_1^* - f_{-1}^*)/2 \in \mathrm{bal}(E)$, $f_1^*, f_{-1}^* \in E$, then the interval $I^* = I(f_{-1}^*, f_1^*)$ is hardest possible, i.e., $\mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N};E) \;=\; \mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N};I^*)$. □

We note that if E is not only convex but also balanced, then the optimal affine algorithm φ_δ is linear and the hardest one–dimensional subclass is symmetric about zero.

It is now natural to ask when the crucial assumption (4.9) is satisfied.

Lemma 4.4 *Suppose that for the worst case radius we have $r'(0^+) > 0$. If*

$$\sup_{h\in\mathrm{bal}(E)} \|N(h)\|_Y \;<\; +\infty \tag{4.12}$$

then for any $\sigma^2 > 0$ there exists $\delta = \delta(\sigma) > 0$ and $d \in \partial r(\delta)$ which satisfy (4.9).

Proof Since $r(\gamma)$ is a concave function of δ, the set $\{\,(\gamma,d)\,|\;\gamma \geq 0,\, d \in \partial r(\gamma)\,\}$ forms a continuous curve. The assumption (4.12) implies that for sufficiently large γ the radius $r(\gamma)$ is constant, which means that for large γ we have $\partial r(\gamma) = \{0\}$. We also have $\partial r(0) = r'(0^+)$. On the other hand, the function $\gamma \to \gamma r(\gamma)(\sigma^2+\gamma^2)^{-1}$ is nonnegative, continuous and it takes zero for $\gamma = 0$. Hence, these two curves must have nonempty intersection at some $\gamma > 0$, as claimed. □

Clearly, the condition (4.12) (and consequently also (4.9)) is satisfied if, for instance, the space F can be equipped with a norm with respect to which E is a bounded set and $N : F \to \mathbb{R}^n$ is a continuous operator. Sometimes it may happen that Theorem 4.2 applies although (4.12) does not hold.

Example 4.2 Consider the integration problem over the class E of periodic 1–Lipschitz functions $f : [0,1] \to \mathbb{R}$ as in Example 2.8. The information is given as $y_i = f(i/n) + x_i$, $1 \leq i \leq n$, where x_i are independent and $x_i \sim \mathcal{N}(0,\sigma^2)$.

Recall that then $r(\gamma) = \gamma/\sqrt{n} + 1/(4n)$. The worst case optimal algorithm is independent of γ and equals $\varphi_{\mathrm{lin}}(y) = n^{-1}\sum_{i=1}^n y_i$. We check that (4.9) holds for $\delta = \delta(\sigma) = 4\,\sigma^2\sqrt{n}$. Hence, Theorem 4.2 yields that the algorithm φ_{lin} is optimal linear also in the mixed case for any σ and

$$\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N};E) \;=\; \sqrt{\frac{\sigma^2}{n} + \frac{1}{16\,n^2}}.$$

The hardest one–dimensional subclass is $[-h^*, h^*]$ where

$$h^* \;=\; 4\,\sigma^2 \,+\, \frac{1}{2n} \,-\, \left| t - \frac{2i-1}{2n} \right|, \quad \frac{i-1}{n} \le t \le \frac{i}{n},\; 1 \le i \le n.$$

In this example, $\sup_{h\in\mathrm{bal}(E)} \|N(h)\|_2 = +\infty$ and (4.12) does not hold. □

We now pass to arbitrary algorithms. The just proven relations between the mixed and worst case settings enable us to show the following result.

Theorem 4.3 *If (4.9) holds then*

$$1 \;\le\; \frac{\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N};E)}{\mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N};E)} \;\le\; \kappa_1^* \tag{4.13}$$

where κ_1^* is defined by (4.6). Furthermore, $\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N};E) \approx \mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N};E)$ *as* $\sigma^2 \to 0^+$.

Proof Due to Lemma 4.3, (4.13) holds for E being an interval. This and Corollary 4.1 yield

$$\begin{aligned} \mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N};E) \;&=\; \sup_{I\subset E}\, \mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N};E) \\ \le\; \kappa_1^* \cdot \sup_{I\subset E}\, \mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N};I) \;&\le\; \kappa_1^* \cdot \mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N};E). \end{aligned}$$

We now prove the remaining part of the theorem. We can assume without loss of generality that $r'(\delta) > 0$ since otherwise information is useless. Then, in view of Lemma 4.3 and Theorem 4.1, we have to show that it is possible to select $\delta = \delta(\sigma^2)$ and d_δ in such a way that σ^2/δ^2 converges to 0 or to $+\infty$, as $\sigma^2 \to 0^+$. Indeed, if this were not true, we would have $\delta \to 0^+$. However, as

$$\frac{\sigma^2}{\delta^2} \;=\; \frac{r(\delta)}{\delta\, d_\delta} \,-\, 1,$$

the limit

$$\lim_{\delta\to 0^+}\frac{\sigma^2}{\delta^2} = \begin{cases} 0 & \text{if} \quad r(0)=0, \\ +\infty & \text{if} \quad r(0)>0, \end{cases}$$

as claimed. □

Thus, we have shown that nonaffine algorithms can only be slightly better than affine algorithms. Moreover, optimal affine algorithm is asymptotically optimal among arbitrary algorithms.

We end this section by considering a special case where E is the unit ball in a separable Hilbert space F and the functional S as well as the operator N are continuous. That is, $S = \langle\cdot, f_S\rangle_F$ and $N = [\langle\cdot, f_1\rangle_F, \ldots, \langle\cdot, f_n\rangle_F]$ for some f_S and f_i from F. Obviously, the condition (4.12) is satisfied and all results of this section are valid. However, in this special case we can obtain more specific results. To do this, we will refer to the average case setting rather than to the worst case.

Suppose that $\tilde F \supset F$ is such a separable Banach space that S and N can be extended to a continuous functional $\tilde S$ and continuous operator $\tilde N = [\tilde L_1, \ldots, \tilde L_n]$ defined on $\tilde F$. Suppose also that the pair $\{F, \tilde F\}$ is an abstract Wiener space, and let μ be the corresponding to it zero mean Gaussian measure. As always, we denote by C_μ the correlation operator of μ. Consider the problem of approximating $\tilde S(f)$ in the average case setting with respect to the measure μ, based on information $y = \tilde N(f) + x$, $x \sim \mathcal{N}(0, \sigma^2\Sigma)$.

Lemma 4.5 *For any linear algorithm $\varphi_{\rm lin}$ we have*

$$\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_{\rm lin}; E) \;=\; \mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\rm lin}; \mu).$$

Proof Indeed, denoting $\varphi_{\rm lin} = d\langle\cdot, w\rangle_Y$, $\|w\|_Y = 1$, we obtain

$$\begin{aligned}(\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_{\rm lin}; \mu))^2 \;&=\; \int_{\tilde F}\int_{\mathbb{R}^n} (\tilde S(f) - \varphi_{\rm lin}(\tilde N(f) + x)\,)^2\, \pi(dx)\, \mu(df) \\ &=\; \int_{\tilde F} (\tilde S(f) - \varphi_{\rm lin}(\tilde N(f))\,)^2 \mu(df) \;+\; \sigma^2 d^2.\end{aligned}$$

Recall now that for any continuous functional L defined on $\tilde F$ we have $\int_{\tilde F} L^2(f)\mu(df) = \|f_L\|_F^2$ where $f_L \in F$ is the representer of L in F, see Section 3.3.2. Since $K = \tilde S(\cdot) - \varphi_{\rm lin}(\tilde N(\cdot))$ is a continuous functional in $\tilde F$

and $f_K = f_S - d\sum_{i=1}^n w_i f_i$, we get

$$\begin{aligned} (\mathrm{e}^{\mathrm{ave}}(\mathbb{N},\varphi_{\mathrm{lin}};\mu))^2 &= \|f_S - d\sum_{i=1}^n w_i f_i\|^2 + \sigma^2 d^2 \\ &= \sup_{\|f\|_F\le 1} (S(f) - \varphi_{\mathrm{lin}}(N(f)))^2 + \sigma^2 d^2 \\ &= (\mathrm{e}^{\mathrm{w-a}}(\mathbb{N},\varphi_{\mathrm{lin}};E))^2, \end{aligned}$$

as claimed. □

Observe that the space $\tilde F$ satisfying the desired assumptions always exists. For instance, it can be constructed as follows. Let W be the space spanned by $f_S, f_1, \ldots, f_n$ and let $W^\perp$ be the orthogonal complement of W. Let $\{f_j\}_{j>n}$ be a complete orthonormal basis of $W^\perp$. Define $\tilde F$ as the closure of F with respect to the norm

$$\|f\|_{\tilde F}^2 = \|f_W\|_F^2 + \sum_{j=n+1}^\infty \lambda_j \alpha_j^2, \quad f = f_W + \sum_{j=n+1}^\infty \alpha_j f_j,$$

where $\{\lambda_j\}$ is a positive sequence with $\sum_{j=n+1}^\infty \lambda_j < +\infty$. Then $\{F, \tilde F\}$ is an abstract Wiener space. Furthermore, it is easy to see that $\tilde S(f) = S(f_W)$ and $\tilde N(f) = N(f_W)$.

Existence of $\tilde F$ together with Lemma 4.5 and the formulas for the optimal algorithm in the average case given at the beginning of Section 3.5 yields the following result.

Theorem 4.4 *Let E be the unit ball in a separable Hilbert space F, and let S and N be continuous linear. Then, in the mixed worst–average setting, the optimal affine algorithm is linear, unique, and given as $\varphi_{\mathrm{lin}}(y) = \langle y, w\rangle_2$ where w is the solution of $(\sigma^2\Sigma + G_N)w = N(f_S)$ and the matrix $G_N = \{\langle f_i, f_j\rangle_F\}_{i,j=1}^n$. Furthermore,*

$$\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N};E) = \sqrt{\|f_S\|_F^2 - \langle w, N(f_S)\rangle_2}. \quad \square$$

In the Hilbert case, the hardest one–dimensional subproblem can also be shown explicite. Indeed, we know from Theorem 3.4 that the average case approximation of $S(f)$ with respect to the measure μ is as difficult as the average case approximation with respect to the measure μ_{K^*} whose mean element is zero and correlation operator

$$A_{K^*}(L) \;=\; \frac{\langle L, K^*\rangle_\mu}{\|K^*\|_\mu^2}\, C_\mu K^*, \qquad L \in F^*,$$

$K^* = S - \langle w, N(\cdot)\rangle_2$. Furthermore, in both cases the algorithm φ_{lin} is optimal. Note that μ_{K^*} is concentrated on the one–dimensional subspace $V = \mathrm{span}\{C_\mu K^*\}$. Hence, due to Lemma 4.5, φ_{lin} is also optimal linear in the mixed setting with the set $E_{K^*} = \{\, \alpha C_\mu K^* \in V \mid |\alpha| \|K^*\|_{\mu_{K^*}} \le 1 \,\}$, and $\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_{\mathrm{lin}}; E_{K^*}) = \mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_{\mathrm{lin}}; E)$. Since

$$\|K^*\|_{\mu_{K^*}}^2 \;=\; K^*(A_{K^*}K^*) \;=\; \|K^*\|_\mu^2 \;=\; \|C_\mu K^*\|_F^2,$$

we have $E_{K^*} = [-h^*, h^*]$ where

$$h^* \;=\; \frac{C_\mu K^*}{\|C_\mu K^*\|_F} \;=\; \frac{f_S - \sum_{j=1}^n w_j f_j}{\|f_S - \sum_{j=1}^n w_j f_j\|_F},$$

and $E_{K^*} \subset E$. The interval $[-h^*, h^*]$ is thus the hardest one–dimensional subproblem.

4.2.3 A correspondence theorem

The relations between the mixed worst–average and other settings discovered in Section 4.2.2 yield the following correspondence theorem.

Let S be a linear functional on a linear space F. Let information about f be given as $y = N(f) + x$. Consider the problem of approximating $S(f)$ from data y in the following three settings.

P1: Mixed worst-average setting with a convex set $E \subset F$ and the noise $x \sim \mathcal{N}(0, \sigma^2\Sigma)$.

P2: Worst case setting with a convex set $E \subset F$ and the noise bounded by $\|x\|_Y = \sqrt{\langle \Sigma^{-1}x, x\rangle_2} \le \delta$.

P3: Average case setting with a Gaussian measure μ defined on F and $x \sim \mathcal{N}(0, \sigma^2\Sigma)$.

We denote by φ_σ the optimal affine algorithm in the mixed setting (P1). Recall that $\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_\sigma; E) \le \kappa_1^* \,\mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N}; E)$.

Theorem 4.5 *(i) Suppose the condition (4.9) is satisfied and $\delta^2 = \sigma^2$. Then* $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi_\sigma; E) \le \sqrt{2}\,\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E)$ *and*

$$\frac{1}{\kappa_1^* \sqrt{2}} \cdot \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E) \;\le\; \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}; E) \;\le\; \mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E).$$

(ii) Suppose the measure μ is induced by the abstract Wiener space $\{H, F\}$, and E is the unit ball of H. Then $\mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi_\sigma; \mu) = \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu)$ *and*

$$\frac{1}{\kappa_1^*} \cdot \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu) \;\le\; \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}; E) \;\le\; \mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu). \quad \Box$$

We can say even more. For any $\sigma^2 \in [0, +\infty]$ there is $\delta = \delta(\sigma) \in [0, +\infty]$ such that the algorithm φ_δ is optimal affine in the mixed setting ($P1$) and in the worst case setting ($P2$) (with convention that φ_∞ is a constant). And vice versa. For any $\delta \in [0, +\infty]$ there is $\sigma^2 = \sigma^2(\delta) \in [0, +\infty]$ ($\sigma^2 = \delta(r(\delta)/d_\delta - \delta)$) such that the algorithm φ_σ is optimal affine for ($P1$) and ($P2$). Since φ_σ is also optimal affine in the average case ($P3$), similar relations hold between the worst case ($P2$) and average case ($P3$).

Example 4.3 Consider the abstract Wiener space $\{H, F\}$ where $H = W^0_{r+1}$ and $F = C^0_r$ ($r \ge 0$), and corresponding to it r–fold Wiener measure w_r (see Example 3.4). Suppose we want to approximate a functional $S \in F^*$, e.g., $S(f) = \int_0^1 f(t)\,dt$, from noisy information $y = N(f) + x$, where

$$N(f) \;=\; [\,f(t_1), f(t_2), \ldots, f(t_n)\,].$$

We know that in the average case ($P3$) with $\mu = w_r$, the unique optimal algorithm is the smoothing spline algorithm. It is given as $\varphi_\sigma(y) = S(\mathbf{s}(y))$ where $\mathbf{s}(y)$ is the natural polynomial spline of order r which belongs to W^0_r and minimizes

$$\int_0^1 (f^{(r+1)}(t))^2\,dt \;+\; \frac{1}{\sigma^2} \cdot \sum_{j=1}^{n} (y_i - f(t_i))^2$$

(for $\sigma^2 = 0$, $\mathbf{s}(y)$ interpolates data y_i exactly, $\mathbf{s}(y)(t_i) = y_i$, $\forall i$). Hence, this algorithm is unique optimal affine in the mixed setting ($P1$) with E being the unit ball of H, and close to optimal among arbitrary algorithms in the mixed and worst settings ($P1$) and ($P2$).

Let $\{\varphi_\sigma\}$ be the family of smoothing spline algorithms where σ runs from zero to infinity. These are all optimal affine algorithms in any of the three cases, for different δ or σ^2, respectively.

Notes and Remarks

NR 4.1 The one–dimensional problem of Section 4.2.1 was studied by many authors. Bickel [5], Casella and Strawderman [8], Levit [49] looked for optimal nonlinear algorithms. It is known that the optimal algorithm is the Bayes estimator with respect to the least favorable distribution on $[-\tau,\tau]$. This least favorable distribution is concentrated on a finite number of points. Moreover, for τ/σ sufficiently small, $\tau/\sigma < 1.05$, it assigns mass 1/2 each to $\pm\tau$. Hence, in this case the algorithm φ_2 defined by (4.4) with $f=\tau$ is optimal and

$$(r_{\mathrm{arb}}(\tau,\sigma^2))^2 \;=\; \tau^2\, e^{-\frac{1}{2}(\tau/\sigma)^2}\frac{1}{\sqrt{2\pi}}\int_0^\infty \frac{e^{-u^2/2}}{\cosh(u\tau/\sigma)}\,du.$$

As τ/σ increases, the number of points also increases and the least favorable prior "tends" to uniform distribution.

NR 4.2 The fact that the ratio $r_{\mathrm{lin}}(\tau,\sigma^2)/r_{\mathrm{arb}}(\tau,\sigma^2)$ is bounded from above by a finite constant was pointed out by Ibragimov and Hasminski [25] who studied the case $N=I$ and convex and balanced E. Donoho *et al.* [14] and Brown and Feldman [7] independently precisely calculated the value of κ_1^*. It is 1.11... .

NR 4.3 Li [50] and Speckman [97] showed optimal properties of smoothing splines for approximating functionals defined on Hilbert spaces. The main line of proving results of Section 4.2.2 for arbitrary convex class E follows Donoho [12] who considered also some other error criteria. (However, we did not assume that the worst case radius is always attained.) The result about asymptotic optimality of affine algorithms and special results for the Hilbert case seem to be new. Lemma 4.5 was pointed to me by K. Ritter in a conversation.

Exercises

E 4.1 Suppose we want to approximate $f\in[-\tau,\tau]$ based on n independent observations of f, $y_i=f+x_i$ where $x_i\sim\mathcal{N}(0,\sigma^2)$. Show that then the sample mean, $\varphi_n(y)=n^{-1}\sum_{j=1}^n y_j$, is an asymptotically optimal algorithm,

$$\mathrm{e}^{\mathrm{w-a}}(\varphi_n) \;=\; \frac{\sigma}{\sqrt{n}} \;\approx\; \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(n), \qquad \text{as } n\to+\infty,$$

where $\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(n)$ is the corresponding nth minimal error of arbitrary algorithms.

E 4.2 Consider the problem of E 4.1 with $f\in\mathbb{R}$ ($\tau=+\infty$) and observations with possibly different variances, $x_i\sim\mathcal{N}(0,\sigma_i^2)$, $1\le i\le n$. Show that then the algorithm

$$\varphi(y) \;=\; \frac{\sum_{i=1}^n \sigma_i^{-2} y_i}{\sum_{i=1}^n \sigma_i^{-2}}$$

is optimal among arbitrary algorithms and its error equals $(\sum_{i=1}^n \sigma_i^{-2})^{-1/2}$.

E 4.3 Consider the problem of approximating values of a linear functional S with $F = \mathbb{R}^n$, based on information $y = f + x$, $x \sim \mathcal{N}(0, \sigma^2 I)$. Show that for a convex and balanced class $E \subset \mathbb{R}^n$ we have

$$\mathrm{rad}_{\mathrm{aff}}^{\mathrm{w-a}}(\mathbb{N}; E) \;=\; \sigma \cdot \sup_{f \in E} \sqrt{\frac{S^2(f)}{\sigma^2 + \|f\|_F^2}}.$$

E 4.4 Prove the uniqueness of the optimal affine algorithm $\varphi_{\delta(\sigma^2)}$ of Theorem 4.2 (if it exists).
Hint: Consider first E being an interval.

E 4.5 Show that for the one–dimensional problem of Section 4.2.1 with $E = \mathbb{R}$ the condition (4.9) is not satisfied.

E 4.6 Show that in the Hilbert case the number $\sigma^2(\delta)$ is determined uniquely. Moreover, in the worst case setting, the regularization parameter $\gamma = \gamma(\delta)$ equals $\sigma^2(\delta)$.

E 4.7 Let F be a separable Hilbert space. Consider approximation of a nonzero functional $S = \langle \cdot, s \rangle_F$, from information $y = N(f) + x$ where

$$N \;=\; [\,\langle \cdot, f_1 \rangle_F, \ldots, \langle \cdot, f_n \rangle_F],$$

$\langle f_i, f_j \rangle_F = \delta_{ij}$, and $x \sim \mathcal{N}(0, \sigma^2 I)$.

Denote by s_1 the orthogonal projection of s onto $\mathrm{span}\{f_1, \ldots, f_n\}$ and by s_2 its orthogonal complement. Show that

$$\delta(\sigma^2) \;=\; \frac{\left(\frac{\sigma^2}{1+\sigma^2}\right)\,\|s_1\|_F}{\sqrt{\left(\frac{\sigma^2}{1+\sigma^2}\right)^2 \|s_1\|_F^2 + \|s_2\|_F^2}} \qquad \text{for } 0 \le \sigma^2 \le +\infty,$$

and

$$\sigma^2(\delta) \;=\; \frac{\delta\,\|s_2\|_F}{\sqrt{1-\delta^2}\,\|s_1\|_F - \delta\,\|s_2\|_2} \qquad \text{for } 0 \le \delta < \frac{\|s_1\|_F}{\|s\|_F},$$

$\sigma^2(\delta) = +\infty$ for $\delta \ge \|s_1\|_F / \|s\|_F$. Hence, in particular, the regularization parameter $\gamma(\delta) = \sigma^2(\delta) \approx \delta \|s_2\|_F / \|s_1\|_F$ as $\delta \to 0^+$.

E 4.8 Show that in the general case $\sigma^2(\delta) \to 0$ as $\delta \to 0^+$ and the convergence is at least linear.

E 4.9 Prove Corollary 4.1 (and consequently also Theorem 4.3) in the Hilbert case using only Lemma 4.5 and Theorem 3.4.

4.3 Approximation of operators

In this section, we present some results about approximation of linear operators in the mixed worst-average case setting.

4.3.1 Ellipsoidal problems in $\mathbb{R}^n$

Suppose we want to approximate a vector $f = (f_1, \dots, f_n) \in \mathbb{R}^n$ which is known to belong to a rectangle

$$E \;=\; \mathcal{R}(\tau) \;=\; \{\, f \in \mathbb{R}^n \mid \quad |f_i| \le \tau_i,\ 1 \le i \le n \,\}$$

where $\tau = (\tau_1, \dots, \tau_n) \in \mathbb{R}^n$, $\tau_i \ge 0 \;\; \forall i$. Information y about f is given coordinatewise, i.e., $y_i = f_i + x_i$, $1 \le i \le n$, where x_i are independent and $x_i \sim \mathcal{N}(0, \sigma_i^2)$.

Lemma 4.6 *For the rectangular problem we have*

$$\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{lin}}(\mathbb{N}, \mathcal{R}(\tau)) \;=\; \sqrt{\sum_{i=1}^n r^2_{\mathrm{lin}}(\tau_i, \sigma_i^2)} \;=\; \sqrt{\sum_{i=1}^n \frac{\sigma_i^2 \tau_i^2}{\sigma_i^2 + \tau_i^2}},$$

$$\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}, \mathcal{R}(\tau)) \;=\; \sqrt{\sum_{i=1}^n r^2_{\mathrm{arb}}(\tau_i, \sigma_i^2)},$$

and the (unique) optimal linear algorithm is given as $\varphi_\tau(y) = (c_i y_1, \dots, c_n y_n)$ *where*

$$c_i \;=\; c_{\mathrm{opt}}(\tau_i, \sigma_i^2) \;=\; \frac{\tau_i^2}{\sigma_i^2 + \tau_i^2}, \qquad 1 \le i \le n.$$

Proof Indeed, in this case the error of any algorithm $\varphi = (\varphi_1, \dots, \varphi_n)$, $\varphi_i : \mathbb{R}^n \to \mathbb{R}$, can be written as

$$\begin{aligned}(\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi; \mathcal{R}(\tau))\,)^2 &= \sup_{f \in \mathcal{R}(\tau)} \int_{\mathbb{R}^n} \sum_{i=1}^n (f_i - \varphi_i(f+x))^2\, \pi(dx) \\ &= \sum_{i=1}^n \left(\sup_{|f_i| \le \tau_i} \int_{\mathbb{R}} (f_i - \varphi_i(f+x)\,)^2\, \pi_i(dx) \right)\end{aligned}$$

where $\pi_i = \mathcal{N}(0, \sigma_i^2)$. Hence, the optimal (linear or nonlinear) approximation is coordinatewise. Moreover, as y_i's are independent, optimal φ_i's use y_i only.

The lemma now follows from results about the one–dimensional problem given in Lemma 4.1. □

It turns out that such a "coordinatewise" algorithm is optimal linear over a larger set than the rectangle. Indeed, observe that the squared error of φ_τ at any $f \in \mathbb{R}^n$ is $\sum_{i=1}^n (1-c_i)^2 f_i^2 + \sigma_i^2 c_i^2$. Since for $f \in \mathcal{R}(\tau)$ this error is maximized at $f = \tau$, the algorithm φ_τ is optimal also over the set of f satisfying the inequality

$$\sum_{i=1}^n (1-c_i)^2 f_i^2 \;\le\; \sum_{i=1}^n (1-c_i)^2 \tau_i^2 .$$

Taking into account the formulas for c_i we get that this set is ellipsoidal,

$$\mathcal{E}(\tau) \;=\; \left\{ f \in \mathbb{R}^n \;\Big|\; \sum_{i=1}^n f_i^2/a_i^2 \le 1 \right\}$$

where

$$a_i^2 \;=\; a_i^2(\tau) \;=\; (1+\tau_i^2/\sigma_i^2)^2 \cdot \sum_{j=1}^n \frac{\tau_j^2}{(1+\tau_j^2/\sigma_j^2)^2}, \quad 1 \le i \le n.$$

Moreover, $\mathcal{E}(\tau)$ is the largest set for which φ_τ is optimal linear. Hence, we have the following corollary.

Corollary 4.2 *Let $E \subset \mathbb{R}^n$. Suppose there is $\tau^* \in \mathbb{R}^n$ such that*

$$\mathcal{R}(\tau^*) \;\subset\; \overline{E} \;\subset\; \mathcal{E}(\tau^*). \tag{4.14}$$

Then the algorithm φ_{τ^*} is optimal linear over E and

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N};\mathcal{R}(\tau^*)) \;=\; \mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N};E) \;=\; \mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N};\mathcal{E}(\tau^*)). \quad \square$$

In words, $\mathcal{R}(\tau^*)$ is the hardest rectangular subproblem contained in E. The notion of the hardest rectangular subproblem thus corresponds to the hardest one–dimensional subproblem for approximating functionals.

The condition (4.14) is satisfied by many sets. The most important example are ellipsoids. In this case, the formulas for τ^*, and consequently for the minimal error, can be found explicite. Namely, we have the following theorem.

Theorem 4.6 *Let E be an ellipsoid,*

$$E = \left\{ f \in \mathbb{R}^n \;\Big|\; \sum_{i=1}^{n} f_i^2/b_i^2 \le 1 \right\}$$

where $b_1 \ge b_2 \ge \cdots \ge b_n > b_{n+1} = 0$. Let

$$\tau_i^* = \begin{cases} \sigma_i \sqrt{b_i \left(\frac{1+\sum_{j=1}^k \sigma_j^2/b_j^2}{\sum_{j=1}^k \sigma_j^2/b_j} \right) - 1} & 1 \le i \le k, \\ 0 & k+1 \le i \le n, \end{cases}$$

where k is the smallest positive integer satisfying

$$b_{k+1} \le \frac{\sum_{j=1}^k \sigma_j^2/b_j}{1 + \sum_{j=1}^k \sigma_j^2/b_j^2}.$$

Then the "coordinatewise" algorithm φ_{τ^*} is (unique) optimal linear,

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N}; E) = \mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_{\tau^*}; E) = \sqrt{\sum_{j=1}^{k} \sigma_j^2 - \frac{(\sum_{j=1}^k \sigma_j^2/b_j)^2}{1 + \sum_{j=1}^k \sigma_j^2/b_j^2}}.$$

Furthermore,

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N}; E) \le \kappa_1^* \cdot \mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N}; E)$$

and φ_{τ^*} is asymptotically optimal among arbitrary algorithms, i.e.,

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N}; E) \approx \mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N}; E) \qquad \text{as} \quad \sigma_i^2 \to 0^+, \quad 1 \le i \le n.$$

Proof Observe first that τ^* is well defined. Indeed, the definition of k implies

$$\frac{1 + \sum_{j=1}^k \sigma_j^2/b_j^2}{\sum_{j=1}^k \sigma_j^2/b_j} > \frac{1 + \sum_{j=1}^k \sigma_j^2}{b_k \left(1 + \sum_{j=1}^{k-1} \sigma_j^2/b_j^2\right) + \sigma_k^2/b_k} = \frac{1}{b_k}$$

($\sum_1^0 = 0$), which means that $b_i(1 + \sum_{j=1}^k \sigma_j^2/b_j^2)/(\sum_{j=1}^k \sigma_j^2/b_j) - 1 > 0 \;\; \forall i$.

Using the standard technique we find that $\mathcal{R}(\tau^*)$ is the (unique) hardest rectangular subproblem contained in E. We can also easily check that $E \subset \mathcal{E}(\tau^*)$. Indeed, for $1 \le i \le k$ we have $a_i(\tau^*) = b_i$, while for $k+1 \le i \le n$ we have

$$a_i(\tau^*) = \frac{\sum_{j=1}^k \sigma_j^2/b_j}{1 + \sum_{j=1}^k \sigma_j^2/b_j^2} \ge b_{k+1} \ge b_i.$$

Due to Corollary 4.2, the algorithm φ_{τ^*} is thus optimal linear, and the radius $\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{lin}}(\mathbb{N};E) = \mathrm{e}^{\mathrm{w-a}}_{\mathrm{lin}}(\mathbb{N},\varphi_{\tau^*};E)$ can easily be calculated.

As $r_{\mathrm{lin}}(\tau_i,\sigma_i^2) \le \kappa_1^* r_{\mathrm{arb}}(\tau_i,\sigma_i^2)$, in view of Lemma 4.6 we have

$$\begin{aligned} \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{lin}}(\mathbb{N};E) &= \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{lin}}(\mathbb{N};\mathcal{R}(\tau^*)\,) \\ &\le \kappa_1^*\cdot\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N};\mathcal{R}(\tau^*)\,) \;\le\; \kappa_1^*\cdot\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N};E). \end{aligned}$$

We also have $r_{\mathrm{lin}}(\tau_i,\sigma_i^2) \approx r_{\mathrm{arb}}(\tau_i,\sigma_i^2)$ as $\sigma_i^2/\tau_i^2 \to 0$. Hence, to complete the proof it suffices to observe that $\sigma_i^2/(\tau_i^*)^2 \to 0$ as all σ_i's decrease to zero. □

Another characterization of problems whose difficulty is determined by the difficulty of the hardest rectangular subproblem is given as follows.

We shall say that a set E is orthosymmetric iff $(f_1,\ldots,f_n)\in E$ implies $(s_1f_2,\ldots,s_nf_n)\in E$, for all choices of $s_i\in\{+1,-1\}$. A set E is quadratically convex iff

$$Q(E) \;=\; \{\,(f_1^2,\ldots,f_n^2)\mid \quad f\in E\,\}$$

is convex. Examples of orthosymmetric and quadratically convex sets include rectangles, ellipsoids, and l_p–bodies with $p\ge 2$,

$$E \;=\; \left\{ f\in\mathbb{R}^n \;\Big|\; \sum_{i=1}^n |f_i|^p/|a_i|^p \le 1 \right\}.$$

Lemma 4.7 *Let E be a bounded convex set of $\mathbb{R}^n$. If E is orthosymmetric and quadratically convex then the condition (4.14) holds, i.e.,*

$$\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N};E) \;=\; \sup_{\tau\in E}\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N};\mathcal{R}(\tau)\,).$$

Proof Let τ^* be the maximizer of $\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N};\mathcal{R}(\tau)\,)$ over $\tau\in\overline{E}$. As E is orthosymmetric and convex, $\mathcal{R}(\tau^*)\subset\overline{E}$ and it is the hardest rectangular subproblem contained in $\overline{E}$. We need to show that $E\subset\mathcal{E}(\tau^*)$.

For $x_i\ge 0$ $\forall i$, let

$$\psi(x_1,\ldots,x_n) \;=\; (\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N};\mathcal{R}(\sqrt{x_1},\ldots,\sqrt{x_n}))\,)^2 \;=\; \sum_{i=1}^n \frac{\sigma_i^2x_i}{\sigma_i^2+x_i}.$$

Denoting by ∂A the boundary of a set A, we have that $P=Q(\,\partial\mathcal{E}(\tau^*)\,)$ is a hyperpline which is adjacent to the set

$$B \;=\; \{\,x\mid \quad \psi(x)\ge(\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N};\mathcal{R}(\tau^*))\,)^2\,\}.$$

As $Q(E)$ is convex and the interiors of $Q(E)$ and $Q(B)$ have empty intersection, both sets are separated by P. Hence, $Q(E) \subset Q(\mathcal{E}(\tau^*))$ which implies $E \subset \mathcal{E}(\tau^*))$, as claimed.

4.3.2 The Hilbert case

We now apply the obtained results to get optimal algorithms for some problems defined on Hilbert spaces. We assume that S is a compact operator acting between separable Hilbert spaces F and G. We want to approximate $S(f)$ for f from the unit ball $E \subset F$. Information is linear with Gaussian noise, i.e., $y = N(f)+x$ where $N = [\langle\cdot, f_1\rangle_F, \ldots, \langle\cdot, f_n\rangle_F]$ and $x \sim \mathcal{N}(0, \sigma^2\Sigma)$, $\Sigma > 0$. As always, $\langle\cdot,\cdot\rangle_Y = \langle\Sigma^{-1}(\cdot), \cdot\rangle_2$.

We will also assume that the operators S^*S and N^*N, where N^* is meant with respect to the inner products $\langle\cdot,\cdot\rangle_F$ and $\langle\cdot,\cdot\rangle_Y$, have a common basis of eigenelements. Denote this basis as $\{\xi_i\}_{i\geq 1}$ and the corresponding eigenvalues as λ_i and η_i,

$$S^*S\,\xi_i \;=\; \lambda_i\,\xi_i, \qquad N^*N\,\xi_i \;=\; \eta_i\,\xi_i, \qquad i \geq 1,$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq 0$ and $\lim_{j\to\infty} \lambda_j = 0$.

Our aim is to find the optimal linear algorithm and its error. It is clear that we can restrict our considerations to such φ that $\varphi(\mathbb{R}^n) \subset \overline{S(F)}$ since otherwise we would project φ onto $\overline{S(F)}$ to obtain a better algorithm. We write φ in the form $\varphi(y) = \sum_j \varphi_j(y) S(\xi_j)$ where $\varphi_j : \mathbb{R}^n \to \mathbb{R}$ and the summation is taken over all $j \geq 1$ with $\lambda_j > 0$. As the elements $S(\xi_j)$ are orthogonal and $\|S(\xi_j)\|^2 = \lambda_j$, for such φ we have

$$\begin{aligned}
&(\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi))^2 \\
&= \sup_{\|f\|_F \leq 1} \int_{\mathbb{R}^n} \Big\| \sum_j \left(\langle f, \xi_j\rangle_F \,-\, \varphi_j(y)\,\right) S(\xi_j) \Big\|^2 \pi_f(dy) \\
&= \sup_{\sum_i \langle f,\xi_i\rangle_F^2 \leq 1} \sum_j \int_{\mathbb{R}^n} \lambda_j \left(\,\langle f, \xi_j\rangle_F \,-\, \varphi_j(y)\,\right)^2 \pi_f(dy).
\end{aligned}$$

We now change variables as follows. Let $\mathcal{I} = \{\,i_1, i_2, \ldots, i_m\,\}$ $(i_1 < i_2 < \cdots i_m)$ be the set of all indices $i \geq 1$ such that $\eta_i > 0$. Clearly, $m \leq n$. For $j \in \mathcal{I}$, let $q_j = N\xi_j/\sqrt{\eta_j}$. Then the vectors q_j are orthonormal in $\mathbb{R}^n$ with respect to the inner product $\langle\cdot,\cdot\rangle_Y$, and $\Sigma^{-1/2}q_j$ are orthonormal with respect to $\langle\cdot,\cdot\rangle_2$. Let Q be an othogonal $n \times n$ matrix whose m first columns are $\Sigma^{-1/2}q_{i_j}$, and let

$$D_1 \;=\; \mathrm{diag}\Big\{\eta_{i_1}^{-1/2},\ldots,\eta_{i_m}^{-1/2},\underbrace{1,\ldots,1}_{n-m}\Big\}.$$

Setting $\tilde y \,=\, D_1Q^T\Sigma^{-1/2}y$ we transform the data $y \,=\, N(f)+x$ to $\tilde y \,=\, M(f)+\tilde x$, where

$$M(f) \;=\; \Big[\,\langle f,\xi_{i_1}\rangle_F,\ldots,\langle f,\xi_{i_m}\rangle_F,\underbrace{0,\ldots,0}_{n-m}\,\Big]$$

and $\tilde x_j$ are independent,

$$\tilde x \;\sim\; \tilde\pi \;=\; \mathcal{N}\Big(0,\sigma^2\mathrm{diag}\Big\{\eta_{i_1}^{-1},\ldots,\eta_{i_m}^{-1},\underbrace{1,\ldots,1}_{n-m}\Big\}\Big).$$

(Compare with analogous transformation in Section 3.4.2). Denoting $\tilde\varphi(\tilde y) = \varphi(y)$ and $f_j = \langle f,\xi_j\rangle_F$ we obtain

$$(\mathrm{e}^{\mathrm{w-a}}(\mathbb{N},\varphi))^2 \;=\; \sup_{\sum_i f_i^2\le 1}\,\sum_j\int_{\mathbb{R}^n}\lambda_j(f_j-\tilde\varphi_j(M(f)+\tilde x)\,)^2\;\tilde\pi(d\tilde x).$$

Changing the variables once more to $h_j = \sqrt{\lambda_j}f_j$, $t_j = \sqrt{\lambda_j}\tilde x_j$ for $1\le j\le m$, $t_j = \tilde x_j$ for $m+1\le j\le n$, and denoting

$$\psi_j(y_1,\ldots,y_n) \;=\; \sqrt{\lambda_j}\cdot\tilde\varphi_j\left(y_1/\sqrt{\lambda_{i_1}},\ldots,y_m/\sqrt{\lambda_{i_m}},y_{m+1},\ldots,y_n\right),$$

we finally get that the squared error $(\mathrm{e}^{\mathrm{w-a}}(\mathbb{N},\varphi))^2$ equals

$$\sup_{\sum_i h_i^2/\lambda_i\le 1}\,\sum_j\int_{\mathbb{R}^n}(h_j-\psi_j(h_{i_1}+t_1,\ldots,h_{i_m}+t_m,t_{m+1},\ldots,t_n)\,)\,\omega(dt)$$

where

$$\omega \;=\; \mathcal{N}\Big(\,0,\,\sigma^2\cdot\mathrm{diag}\Big\{\lambda_{i_1}/\eta_{i_1},\ldots,\lambda_{i_m}/\eta_{i_m},\underbrace{1,\ldots,1}_{n-m}\Big\}\Big).$$

Observe that information about $h = (h_1,h_2,\ldots)\in l_2$ is now given coordinatewise. Thus we can apply the whole machinery with rectangular subproblems which are now given as $\mathcal{R}(\tau) = \{\,h\in l_2\mid |h_i|\le\tau_i,\,i\ge 1\,\}$ where $\tau\in l_2$ is in the ellipsoid $\sum_i\tau_i^2/\lambda_i\le 1$.

It is not difficult to see that the hardest τ^* is given as follows. Let

$$s \;=\; \min\{\, i \geq 0 \mid \quad \eta_{i+1} = 0 \text{ or } \lambda_{i+1} = 0\,\}. \tag{4.15}$$

Then $\tau_i^* = 0$ for $i \geq s+2$, and $\tau_1^*, \ldots, \tau_{s+1}^*$ is the maximizer of

$$\sum_{j=1}^{s} \frac{\sigma_j^2 \tau_j^2}{\sigma_j^2 + \tau_j^2} \;+\; \tau_{s+1}^2$$

over the ellipsoid $\sum_{i=1}^{s+1} \tau_i^2/\lambda_i \;\leq\; 1$ (if $\lambda_{s+1} \;=\; 0$ then $\tau_{s+1} \;=\; 0$ and the summation is taken from 1 to s), where the noise levels $\sigma_j^2 \;=\; \sigma^2\lambda_j/\eta_j$. Solving this maximization problem, we obtain the following formulas.

Let k be the smallest integer satisfying $k \in \{1, 2, \ldots, s\}$ and

$$\sqrt{\lambda_{k+1}} \;\leq\; \frac{\sigma^2\,\sum_{j=1}^{k}(\sqrt{\lambda_j}\eta_j)^{-1}}{1 + \sigma^2\,\sum_{j=1}^{k}\eta_j^{-1}}, \tag{4.16}$$

or $k = s+1$ if such a number does not exist. We have two cases:

(i) If $1 \leq k \leq s$ then

$$(\tau_i^*)^2 \;=\; \begin{cases} \sigma^2\,\frac{\lambda_i}{\eta_i}\,\left(\sqrt{\lambda_i}\,\left(\frac{1+\sigma^2\sum_{j=1}^{k}\eta_j^{-1}}{\sigma^2\sum_{j=1}^{k}(\sqrt{\lambda_j}\eta_j)^{-1}}\right) \;-\; 1\right) & 1 \leq i \leq s, \\ 0 & i \geq s+1. \end{cases} \tag{4.17}$$

(ii) If $k = s+1$ then

$$(\tau_i^*)^2 \;=\; \begin{cases} \sigma^2\,\frac{\lambda_i}{\eta_i}\,\left(\frac{\sqrt{\lambda_i}}{\sqrt{\lambda_{s+1}}} \;-\; 1\right) & 1 \leq i \leq s, \\ \lambda_{s+1} \;-\; \sigma^2\,\sqrt{\lambda_{s+1}}\,\sum_{j=1}^{s}\left(\frac{\sqrt{\lambda_j}-\sqrt{\lambda_{s+1}}}{\eta_j}\right) & i = s+1, \\ 0 & i \geq s+2. \end{cases} \tag{4.18}$$

Now we can check that the "coordinatewise" algorithm φ_{τ^*} is optimal not only for the hardest rectangular subproblem $\mathcal{R}(\tau^*)$, but also for the ellipsoid $\sum_{j=1}^{s+1} h_j^2/\lambda_j \leq 1$. The minimal linear error is then equal to the error of φ_{τ^*} and nonlinear algorithms can be only slightly better. We summarize our analysis in the following theorem.

Theorem 4.7 *Suppose the operators S^*S and N^*N have a common orthonormal basis of eigenelements $\{\xi_i\}$ and the corresponding eigenvalues are*

λ_i and η_i, respectively, $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$. Let s and k be defined by (4.15) and (4.16).

(i) If $1 \le k \le s$ then

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N}) \;=\; \sigma \cdot \sqrt{\sum_{j=1}^{k} \frac{\lambda_j}{\eta_j} \;-\; \frac{\sigma^2\,\left(\sum_{j=1}^{k}(\sqrt{\lambda_j}\eta_j)^{-1}\right)^2}{1+\sigma^2\,\sum_{j=1}^{k}\eta_j^{-1}}}\;.$$

(ii) If $k = s+1$ then

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N}) \;=\; \sqrt{\lambda_{s+1} \,+\, \sigma^2 \sum_{j=1}^{s} \frac{(\sqrt{\lambda_j}-\sqrt{\lambda_{s+1}})^2}{\eta_j}}\;.$$

In both cases, the optimal linear algorithm is given as

$$\varphi_{\mathrm{lin}}(y) \;=\; \sum_{j=1}^{s} \frac{\sigma^2(\tau_j^*)^2}{\sigma^2+\eta_j(\tau_j^*)^2}\, z_j\, S(\xi_j)$$

where τ_j^* are given by (4.17) and (4.18), and $z_j = \eta_j^{-1}\langle N(\xi_j),\Sigma^{-1}y\rangle_2$, $1 \le j \le s$.

For nonlinear algorithms we have $\mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N}) \le \kappa_1^*\,\mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N})$. □

These rather complicated formulas take much simpler form when S is the identity operator in $\mathbb{R}^d$. More precisely, suppose we approximate a vector $f \in \mathbb{R}^d$, $\|f\|_2 \le 1$, from information $y_i = \langle f, f_i\rangle_2 + x_i$, $1 \le i \le n$, where $x \sim \mathcal{N}(0,\sigma^2\Sigma)$ and the vectors f_i span the space $\mathbb{R}^d$. (If the last assumption is not satisfied then $\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N}) = 1$.) It is clear that then the orthonormal eigenvectors ξ_i of N^*N are also the eigenvectors of $S^*S = I$ and $\lambda_i = 1$ $\forall i$. From Theorem 4.7 we obtain that the minimal error depends only on $\sum_{i=1}^{d}\eta_i^{-1} = \mathrm{trace}((N^*N)^{-1})$ and equals

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{w-a}}(\mathbb{N}) \;=\; \sigma \cdot \sqrt{\frac{\mathrm{trace}(\,N^*N)^{-1})}{1+\sigma^2\,\mathrm{trace}(\,(N^*N)^{-1})}}.$$

The hardest rectangular subproblem is independent of σ^2 and given as

$$\mathcal{R}(\tau^*) \;=\; \left\{\, f = \sum_{i=1}^{d}\alpha_i\xi_i \in \mathbb{R}^d \;\middle|\; |\alpha_i| \le \tau_i^*,\ 1 \le i \le d \right\}$$

where $\tau_i^* = \eta_i^{-1}/\mathrm{trace}(\,(N^*N)^{-1})$. Hence, the optimal linear algorithm is

$$\varphi_{\mathrm{lin}}(y) \;=\; \frac{1}{1+\sigma^2\,\mathrm{trace}(\,(N^*N)^{-1})}\,\sum_{j=1}^d z_j\xi_j$$

where $z_j = \eta_j^{-1}\langle N(\xi_j), \Sigma^{-1}y\rangle_2$. In particular, if N is the identity and Σ is a diagonal matrix (with elements $\eta_j^{-1}, 1 \le j \le d$), then $z = y$ and $\varphi_{\mathrm{lin}}(y) = (1+\sigma^2\mathrm{trace}(\,(N^*N)^{-1})\,)^{-1}\,y$.

We see that the optimal linear algorithm is in this case *not* a smoothing spline since the latter puts different coefficients, $c_j = (1+\gamma/\eta_j)^{-1}$, for different j. Thus, in the mixed worst–average case the situation changes as compared to worst and average settings where, in the Hilbert case, smoothing splines are optimal algorithms.

The assumption that S^*S and N^*N have a common basis of eigenelements was essential. When this is not satisfied, optimal (or almost optimal) algorithms are known only for some special problems. We now present one of them, for other results see NR 4.5.

Suppose we approximate values of a linear operator S defined on $F = \mathbb{R}^d$, for all elements $f \in \mathbb{R}^d$, i.e., $E = \mathbb{R}^d$. Information $\mathbb{N}$ is assumed to be arbitrary. It turns out that in this case the optimal algorithm are the least squares, even in the class of arbitrary algorithms. Indeed, the (generalized) least squares algorithm is defined as $\varphi_{\mathrm{ls}}(y) \;=\; SN^{-1}P_N(y)$ where P_N is the orthogonal projection (with respect to $\langle\cdot,\cdot\rangle_Y$) in $\mathbb{R}^n$ onto $N(\mathbb{R}^d)$. From the proof of Theorem 3.6 we know that for any f

$$\int_{\mathbb{R}^n} \|S(f)-\varphi_{\mathrm{ls}}(N(f)+x)\,\|^2\,\pi(dx) \;=\; \sigma^2\,\mathrm{trace}(S(N^*N)^{-1}S^*).$$

Hence, $(\mathrm{e}^{\mathrm{w-a}}(\mathbb{N},\varphi_{\mathrm{ls}}))^2 = \sigma^2\mathrm{trace}(S(N^*N)^{-1}S^*)$. On the other hand, a lower bound on $\mathrm{rad}^{\mathrm{w-a}}(\mathbb{N};\mathbb{R}^d)$ can be obtained by calculating the average radius of the same information with respect to the measure $\mu_\lambda = \mathcal{N}(0,\lambda I)$. Using Corollary 3.1 we obtain

$$\begin{aligned}(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N};\mu_\lambda))^2 \;&=\; \lambda\cdot\mathrm{trace}(SS^*)\;-\;\sum_{j=1}^d\frac{\lambda\,\|S(\xi_j)\|^2}{1+\sigma^2/(\eta_j\lambda)}\\ &=\; \sum_{j=1}^d\frac{\sigma^2\lambda}{\sigma^2+\eta_j\lambda}\,\|S(\xi_j)\|^2.\end{aligned}$$

Now, letting $\lambda \to +\infty$ we get

$$\begin{aligned}(\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N};\mathbb{R}^d))^2 &\geq \lim_{\lambda\to\infty}(\mathrm{rad}^{\mathrm{ave}}(\mathbb{N};\mu_\lambda))^2 \\ &= \sigma^2\sum_{j=1}^d \frac{\|S(\xi_j)\|^2}{\eta_j} = \sigma^2\sum_{j=1}^d \|SN^{-1}(N\xi_j/\sqrt{\eta_j})\|^2 \\ &= \sigma^2\,\mathrm{trace}(S(N^*N)^{-1}S^*).\end{aligned}$$

Hence, we have proven the following theorem.

Theorem 4.8 *Let $E=F=\mathbb{R}^d$ and $\dim N(F)=d$. Then the generalized least squares φ_{ls} are optimal among arbitrary algorithms and*

$$\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N};\mathbb{R}^d) = \sigma\sqrt{trace(S(N^*N)^{-1}S^*)}.$$

Notes and Remarks

NR 4.4 Section 4.3.1 is based on the results of Donoho *et al.* [14] where the model with infinitely many observations is studied. Corollary 4.2 is however new. The proof of Lemma 4.7 is also different. Results of Section 4.3.2 seem to be new.

Asymptotic optimality of linear algorithms for ellipsoidal problems was first shown by Pinsker [76].

NR 4.5 The model with "coordinatewise" observations turns out to be the limiting model in curve estimation. This fact together with results of Section 4.3.1 can be used to derive results about optimal algorithms for some other problems. We now give one example.

Suppose we want to approximate a function $f:[0,1]\to\mathbb{R}$ in $\mathcal{L}_2$–norm from the class $E=E_P=\{\,f\in W_r \mid \int_0^1 (f^{(r)}(t))^2\,dt \leq P^2\,\}$, based on noisy values of f at equidistant points, $y_i = f(i/n)+x_i$, $0\leq i\leq n$, and $x\sim\mathcal{N}(0,\sigma^2 I)$ $(\sigma^2>0)$. In the statistical literature, this is called a *nonparametric regression* model and was studied, e.g., in Golubev and Nussbaum [19], Nussbaum [67], Speckman [98], Stone [100] (see also the book of Eubank [15]). It is known that for this problem the minimal error is asymptotically (as $n\to\infty$) achieved by a version of the smoothing spline algorithm, and that this minimal error satisfies

$$\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(n) \asymp n^{-\frac{r}{2r+1}}.$$

The asymptotic constant Γ in the "$\asymp$" notation was evaluated by Nussbaum [67] who showed that

$$\Gamma = \gamma(r)\sigma^{\frac{2r}{2r+1}}P^{\frac{1}{2r+1}}$$

where $\gamma(r) = (2r+1)^{1/(4r+2)}(r/\pi(r+1))^{r/(2r+1)}$ is Pinsker's constant. The main idea of proving this result is as follows. As for large n, the $\mathcal{L}_2$–norm of a function f essentially equals $\|f\|_n = (n^{-1}\sum_{i=1}^n f^2(i/n)\,)^{1/2}$, we can consider the error with respect to the seminorm $\|\cdot\|_n$ instead of the $\mathcal{L}_2$–norm. The set $E_n = \{\,(f(0), f(1/n), \ldots, f((n-1)/n), f(1)\,)\mid f\in E_P\,\}$ is an ellipsoid. Hence, the original problem can be reduced to that of approximating a vector $v\in E_n\subset\mathbb{R}^n$ from information $y = v+x$ where $x\sim\mathcal{N}(0,\sigma^2 I)$. If we find the coordinates of E_n (which is the main difficulty in this problem), results of Section 4.3.1 can be applied.

Golubev and Nussbaum [19] showed that we cannot do much better by performing observations at other than equidistant points.

NR 4.6 Recently, Donoho and Johnstone [13] (see also Donoho *et al.* [11]) developed a new algorithm for approximating functions from their noisy samples at equidistant points. The algorithm is nonlinear. It uses the wavelet transform and relies on translating the empirical wavelet coefficients towards the origin by an amount $\sqrt{2\log(n)}\sigma/\sqrt{n}$. Surprisingly enough, such a simple algorithm turns out to be nearly optimal for estimating many classes of functions, including standard Hölder and Sobolev classes, but also more general Besov and Triebel bodies.

More precisely, suppose that f is in the unit ball of the Besov space $B^s_{p,r}$ or Triebel space $F^s_{p,q}$. Then the minimal errors of arbitrary and linear algorithms using n noisy samples are given as

$$\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(n) \asymp n^{-k} \qquad \text{and} \qquad \mathrm{rad}^{\mathrm{w-a}}_{\mathrm{lin}}(n) \asymp n^{-k'},$$

where

$$k = \frac{s}{s+1/2} \qquad \text{and} \qquad k' = \frac{s+(1/p_- - 1/p)}{s+1/2+(1/p_- - 1/p)},$$

$p_- = \max\{p, 2\}$, correspondingly. Hence, for $p<2$, no linear algorithm can achieve the optimal rate of convergence.

For information on Besov and Triebel spaces see, e.g., Triebel [110].

NR 4.7 The computational complexity in the mixed worst–average case setting is studied very rarely. There are still two main difficulties that yet have to be overcome before finding concrete complexity formulas.

The first difficulty lies in obtaining optimal information. As optimal algorithms are not known exactly even for problems defined in Hilbert spaces, results on optimal information are rather limited. (For some special cases see NR 4.5 and E 4.13 4.14.)

The second difficulty is in the problem of adaptive information. For instance, we do not know sufficient conditions under which adaption does not help. Ibragimov and Hasminski [24] and Golubev [20] proved that in the nonparametric regression model the equidistant design is asymptotically optimal even in the class of adaptive designes. This is however no longer valid if we consider the integration problem. Namely, one can show that adaption can significantly help for multivariate integration over convex and balanced classes of functions, see Plaskota [84].

Exercises

E 4.10 Show that the number k in Theorem 4.6 can equivalently be defined as the largest integer satisfying $1 \le k \le n$ and

$$b_k \;<\; \frac{\sum_{j=1}^k \sigma_j^2/b_j}{1+\sum_{j=1}^k \sigma_j^2/b_j^2}.$$

E 4.11 Let E be the l_p–body with $1 \le p < 1$. Show that then the condition (4.14) is not satisfied and, in particular, E is not quadratically convex.

E 4.12 Suppose the least squares algorithm φ_{ls} is applied for $S = I$ and E being the unit ball of $\mathbb{R}^d$. Show that

$$\mathrm{e}^{\mathrm{w-a}}(\mathbb{N}, \varphi_{\mathrm{ls}}) \;\approx\; \mathrm{rad}_{\mathrm{arb}}^{\mathrm{w-a}}(\mathbb{N}) \qquad \text{as } \; \mathrm{trace}(\,(N^*N)^{-1}) \to 0.$$

E 4.13 Consider the problem of approximating a vector f from the unit ball of $\mathbb{R}^d$. Let $r_n(\sigma_1, \dots, \sigma_n)$ $(0 < \sigma_1^2 \le \cdots \le \sigma_n^2)$ be the minimal error that can be attained by linear algorithms that use n $(n \ge d)$ independent observations $y_i = \langle f, f_i \rangle_2 + x_i$ with Gaussian noise, $x_i \sim \mathcal{N}(0, \sigma_i^2)$, where $\|f_i\|_2 \le 1$, $1 \le i \le n$. Show that

$$r_n(\sigma_1, \dots, \sigma_n) \;=\; \min \sqrt{\frac{\sum_{i=1}^d \eta_i^{-1}}{1+\sum_{i=1}^d \eta_i^{-1}}}$$

where the minimum is taken over all $\eta_i \ge 0$ such that

$$\sum_{j=r}^n \eta_j \;\le\; \sum_{j=r}^n \sigma_j^{-2}, \qquad 1 \le r \le n. \tag{4.19}$$

In particular, for $n = d$ we have

$$r_n(\sigma_1, \dots, \sigma_n) \;=\; \sqrt{\frac{\sum_{i=1}^d \sigma_i^2}{1+\sum_{i=1}^d \sigma_i^2}}.$$

What is the optimal information?
Hint: To obtain the upper bound and optimal information, use Lemma 2.14.

E 4.14 Consider the optimal information problem as in E 4.13, but with $E = \mathbb{R}^d$ and arbitrary solution operator $S : \mathbb{R}^d \to G$. Let $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$ be the eigenvalues of S^*S. Show that

$$r_n(\sigma_1, \dots, \sigma_n) \;=\; \min \sqrt{\sum_{i=1}^d \frac{\lambda_i}{\eta_i}}$$

where the minimum is taken over all $\eta_i \geq 0$ satisfying (4.19). In particular, for equal variances $\sigma_i^2 = \sigma^2$ we have

$$r_n(\sigma) \;=\; \frac{\sigma}{\sqrt{n}} \cdot \sum_{i=1}^{d} \lambda_i^{1/2}.$$

Find the optimal information.

Chapter 5

Second mixed setting

5.1 Introduction

When we vary stochastic and deterministic assumptions on the problem elements f and noise x, the *mixed average-worst case setting* will appear quite naturally as the fourth possible way of treating problems with noisy information. In this setting, we assume some probability distribution μ on the domain F of the solution operator S. The information operator is defined as in the worst case of Chapter 2. That is, $\mathbb{N}(f)$ is a set of finite real sequences. The error of an algorithm φ that uses information $y \in \mathbb{N}(f)$ is given as

$$\mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi) \;=\; \sqrt{\int_F \sup_{y\in\mathbb{N}(f)} \|S(f)-\varphi(y)\|^2 \,\mu(df)}.$$

As the mixed average–worst setting seem to be of less importance than the other settings, it is studied very rarely. Nevertheless, it leads to interesting and nontrivial theoretical problems.

The main results of this chapter concern approximating of functionals and are presented in Section 5.2. It turns out that in this case the mixed average–worst setting can be analyzed similarly to the worst–average setting, although they seem to be completely different. Assuming that the information noise is bounded in a Hilbert norm, we establish a close relation between the average–worst and a corresponding average case settings. Namely, optimal linear algorithms in both settings belong to the same class of smoothing spline algorithms. Moreover, the minimal achievable errors differ only by a small constant and can be (almost) attained by the same algorithm. Using once more the concept of hardest one–dimensional subproblem, we also show that for approximating functionals nonlinear algorithms cannot be much better than linear algorithms.

The relation between the average-worst and average settings together with relations established in the previous sections, enables us to formulate a theorem about (almost) equivalence of all four corresponding settings in the case when the solution operator is a linear functional.

In Section 5.3, we present some results about approximating of operators. In particular, we show that for sufficiently small noise level, the least squares are the optimal linear algorithm.

5.2 Linear algorithms for linear functionals

In this section, we construct almost optimal algorithms for the case when the solution operator S is a linear functional. To do this, we use ideas similar to those of the worst–average setting.

5.2.1 The one–dimensional problem

Suppose we want to approximate a real random variable f which has zero mean normal distribution with variance $\lambda > 0$, $f \sim \mathcal{N}(0,\lambda)$. We assume that instead of f we know only its noisy value $y = f + x$ where $|x| \le \delta$. That is, the information operator $\mathbb{N}(f) = [f-\delta, f+\delta]$. In this case, the error of an algorithm φ is given as

$$\begin{aligned} \mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi) &= \mathrm{e}^{\mathrm{a-w}}(\lambda,\delta;\varphi) \\ &= \sqrt{\frac{1}{\sqrt{2\pi\lambda}}\int_{\mathbb{R}} \sup_{|x|\le\delta} |f-\varphi(f+x)|^2 \exp\{-f^2/(2\lambda)\}\, df}\ . \end{aligned}$$

We note that the general problem with $N = S$ reduces to this case. Indeed, as μ is Gaussian, we have $S(f) \sim \mathcal{N}(0, S(C_\mu S))$. Then we approximate $g = S(f) \in \mathbb{R}$ based on information $y = g + x$ where $|x| \le \delta$.

First, we consider linear algorithms. Let

$$r_{\mathrm{lin}}(\lambda,\delta) \;=\; \inf\{\, \mathrm{e}^{\mathrm{a-w}}(\lambda,\delta;\varphi) \mid \quad \varphi - \text{linear} \;\}.$$

Lemma 5.1 *For any λ and δ we have*

$$r_{\mathrm{lin}}(\lambda,\delta) \;=\; \begin{cases} \delta & \delta^2 \le \frac{2}{\pi}\lambda, \\ \sqrt{\frac{\lambda\delta^2(1-2/\pi)}{\lambda+\delta^2-2\delta(2\lambda/\pi)^{1/2}}} & \frac{2}{\pi}\lambda < \delta^2 < \frac{\pi}{2}\lambda, \\ \sqrt{\lambda} & \frac{\pi}{2}\lambda \le \delta^2\,. \end{cases}$$

The optimal coefficient $c_{\rm opt} = c_{\rm opt}(\lambda,\delta)$ *of linear algorithms is unique and given as*

$$c_{\rm opt}(\lambda,\delta) \;=\; \begin{cases} 1 & \delta^2 \le \frac{2}{\pi}\lambda, \\ \frac{\lambda-\delta\sqrt{2\lambda/\pi}}{\lambda+\delta^2-2\delta\sqrt{2\lambda/\pi}} & \frac{2}{\pi}\lambda < \delta^2 < \frac{\pi}{2}\lambda, \\ 0 & \frac{\pi}{2}\lambda \le \delta^2\,. \end{cases}$$

Proof For a linear algorithm $\varphi(y) \;=\; c\,y$ and $f\in\mathbb{R}$ we have

$$\begin{aligned}\sup_{|x|\le\delta}|f-\varphi(f+x)|^2 \;&=\; \sup_{|x|\le\delta}|(1-c)f-cx|^2 \\ &=\; (1-c)^2f^2 \,+\, c^2\,\delta^2 \,+\, 2\,\delta\,|\,(1-c)\,c\,|\,|f\,|\,.\end{aligned}$$

Taking the integral over f we get

$$\mathrm{e}^{\mathrm{a-w}}(\lambda,\delta;\varphi) \;=\; (1-c)^2\lambda \,+\, c^2\delta^2 \,+\, 2\delta|(1-c)c|\sqrt{\frac{2\lambda}{\pi}}.$$

To obtain the desired result it is now enough to minimize the obtained expression with respect to c. □

Observe that the optimal coefficient $c_{\rm opt}$ is determined uniquely and it is a function of δ^2/λ, i.e., $c_{\rm opt}(\lambda,\delta) \;=\; c_{\rm opt}(1,\delta/\sqrt{\lambda})$. Furthermore,

$$r_{\rm lin}(\lambda,\delta) \;=\; \sqrt{\lambda}\cdot r_{\rm lin}(1,\delta/\sqrt{\lambda})\,. \tag{5.1}$$

It is clear that we can do better by using nonlinear algorithms.

Example 5.1 Let $\lambda>0$ and $\delta^2\ge\lambda\pi/2$. Then the nonlinear algorithm

$$\varphi_{\rm non}(y) \;=\; \begin{cases} y+\delta & y<-\delta, \\ 0 & -\delta\le y\le\delta, \\ y-\delta & \delta<y, \end{cases}$$

has smaller error than the optimal linear one $\varphi_{\rm lin}\equiv 0$. Indeed, it is easy to check that for any f we have

$$\sup_{|x|\le\delta}|f-\varphi_{\rm non}(f+x)|^2 \;=\; \min\{\,|f|^2,\,4\delta^2\,\},$$

while for $\varphi\equiv 0$ the above quantity equals $|f|^2$. Hence, $\mathrm{e}^{\mathrm{a-w}}(\lambda,\delta;\varphi_{\rm non}) < \mathrm{e}^{\mathrm{a-w}}(\lambda,\delta;0)$. □

However, as in the first mixed setting, we never gain much. Indeed, let

$$r_{\rm arb}(\lambda,\delta) \;=\; \inf\{\, {\rm e}^{\rm a-w}(\lambda,\delta;\varphi) \;|\quad \varphi - \text{arbitrary} \;\}.$$

Theorem 5.1 *We have*

$$\lim_{\delta/\sqrt{\lambda}\to 0} \frac{r_{\rm lin}(\lambda,\delta)}{r_{\rm arb}(\lambda,\delta)} \;=\; 1 \;=\; \lim_{\delta/\sqrt{\lambda}\to\infty} \frac{r_{\rm lin}(\lambda,\delta)}{r_{\rm arb}(\lambda,\delta)}\,.$$

Furthermore, there exists an absolute constant κ_2 such that

$$1 \;\le\; \frac{r_{\rm lin}(\lambda,\delta)}{r_{\rm arb}(\lambda,\delta)} \;\le\; \kappa_2 \qquad \forall\,\lambda,\delta.$$

Proof Due to the same argument as in the proof of Theorem 4.1, we can assume without loss of generality that $\lambda = 1$.

To obtain a lower bound on the error of φ, note that

$$\sup_{|x|\le\delta} |f-\varphi(f+x)|^2 \;\ge\; \frac{1}{2}\left(|f-\varphi(f+\delta)|^2 + |f-\varphi(f-\delta)|^2\right).$$

From this we get

$$\begin{aligned}
&({\rm e}^{\rm a-w}(1,\delta;\varphi)\,)^2\\
\ge\;& \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{1}{2}\,(\,|f-\varphi(f+\delta)|^2 \;+\; |f-\varphi(f-\delta)|^2\,)\, e^{-f^2/2}\,df\\
=\;& \frac{1}{\sqrt{8\pi}}\int_{\mathbb{R}} (y-\varphi(y)-\delta)^2\, e^{-(y-\delta)^2/2} \;+\; (y-\varphi(y)+\delta)^2\, e^{-(y+\delta)^2/2}\,df\,.
\end{aligned}$$

For each y, the last integrand is minimized by

$$\varphi^*(y) \;=\; y \;-\; \delta\,\frac{a_- \;-\; a_+}{a_- \;+ a_+}$$

where $a_- \;=\; e^{-(y-\delta)^2/2}$ and $a_+ \;=\; e^{-(y+\delta)^2/2}$. Hence, setting $\varphi = \varphi^*$ and performing some elementary transformations we finally arrive at the following bound:

$$ {\rm e}^{\rm a-w}(1,\delta;\varphi) \;\ge\; \delta^2\,\psi(\delta) \tag{5.2}$$

where

$$\psi(\delta) \;=\; e^{-\delta^2/2}\sqrt{\frac{2}{\pi}}\int_0^{+\infty} \frac{\exp\{-y^2/2\}}{\cosh\{\delta\, y\}}\,dy\,.$$

Observe that for all δ

$$\psi(\delta) \;\geq\; \sqrt{\frac{2}{\pi}}\int_{\delta}^{+\infty} \exp\{-y^2/2\}\,dy\,.$$

This, (5.2) and Lemma 5.1 yield

$$\lim_{\delta\to 0}\frac{r_{\mathrm{lin}}(1,\delta)}{r_{\mathrm{arb}}(1,\delta)} \;=\; 1\,.$$

The second limit of the theorem follows from the fact that for any φ and $|f|\leq\delta$ we have $\sup_{|x|\leq\delta}|f-\varphi(f+x)| \;\geq\; |f-\varphi(0)|$. This yields

$$\begin{aligned} \mathrm{e}^{\mathrm{a-w}}(1,\delta;\varphi) \;&\geq\; \frac{1}{\sqrt{2\pi}}\int_{-\delta}^{\delta}(\,f-\varphi(0)\,)^2\,e^{-f^2/2}\,df \\ &\geq\; \frac{1}{\sqrt{2\pi}}\int_{-\delta}^{\delta} f^2\,e^{-f^2/2}\,df \;\longrightarrow\; 1\,, \end{aligned}$$

as $\delta\to+\infty$.

Since the inequality $r_{\mathrm{arb}}(\lambda,\delta)\leq r_{\mathrm{lin}}(\lambda,\delta)$ is obvious, it remains to show existence of κ_2. To this end, observe that the function ψ is decreasing. Hence, from (5.2) and Lemma 5.1 we get that for $\delta\in[0,1]$

$$\frac{r_{\mathrm{lin}}(1,\delta)}{r_{\mathrm{arb}}(1,\delta)} \;\leq\; \frac{\delta^2}{\delta^2\,\psi(\delta)} \;\leq\; \frac{1}{\psi(1)}\,.$$

On the other hand, for $\delta\in(1,+\infty)$ we have

$$\frac{r_{\mathrm{lin}}(1,\delta)}{r_{\mathrm{arb}}(1,\delta)} \;\leq\; \frac{1}{r_{\mathrm{arb}}(1,1)} \;\leq\; \frac{1}{\psi(1)}\,.$$

Hence, we can take $\kappa_2 \;=\; 1/\psi(1)$. □

As in the first mixed setting, we can define the constant

$$\kappa_2^* \;=\; \sup_{\lambda,\delta}\frac{r_{\mathrm{lin}}(\lambda,\delta)}{r_{\mathrm{arb}}(\lambda,\delta)}\,. \tag{5.3}$$

From the proof of Theorem 5.1 we have that $\kappa_2^* \;\leq\; \psi^{-1/2}(1) \;=\; 1.49\ldots$. The exact value of κ_2^* is however not known.

5.2.2 Almost optimality of linear algorithms

We now consider a general case. That is, we assume that S is an arbitrary continuous linear functional defined on a separable Banach space F, and that μ is a zero mean Gaussian measure on F with correlation operator $C_\mu : F^* \to F$. Information about $f \in F$ is linear with noise bounded uniformly in a Hilbert norm. That is, $y = N(f) + x \in \mathbb{R}^n$ where $N = [L_1, , \ldots, L_n]$ ($L_i \in F^*$) and $\|x\|_Y = \sqrt{\langle \Sigma^{-1}x, x\rangle_2} \le \delta$, $\Sigma = \Sigma^* > 0$. Let

$$\begin{aligned} \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N};\mu) &= \inf\{\, \mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi;\mu) \mid \quad \varphi \text{ - linear}\,\} \\ \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N};\mu) &= \inf\{\, \mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi;\mu) \mid \quad \varphi \text{ - arbitrary}\,\} \end{aligned}$$

be the minimal errors of linear and arbitrary approximations with respect to the measure μ.

Consider first the case where μ is concentrated on a one dimensional subspace.

Lemma 5.2 *Let $h \in F$ and let μ be the zero mean Gaussian measure with correlation operator*

$$C_\mu(L) \;=\; L(h)\,h, \qquad \forall\, L \in F^*.$$

(i) If $N(h) = 0$ then

$$\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N};\mu) \;=\; \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N},\mu) \;=\; |S(h)|$$

and $\varphi \equiv 0$ is the optimal algorithm.
(ii) If $N(h) \neq 0$ then

$$\begin{aligned} \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N};\mu) &= |S(h)|\, r_{\mathrm{lin}}\left(1, \frac{\delta}{\|N(h)\|_Y}\right), \\ \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N};\mu) &= |S(h)|\, r_{\mathrm{arb}}\left(1, \frac{\delta}{\|N(h)\|_Y}\right), \end{aligned}$$

and the optimal linear algorithm is given as

$$\varphi(y) \;=\; c_{\mathrm{opt}}\left(1, \frac{\delta}{\|N(h)\|_Y}\right) \frac{S(h)}{\|N(h)\|_Y} \left\langle y, \frac{N(h)}{\|N(h)\|_Y} \right\rangle_Y$$

where $r_{\mathrm{lin}}(\cdot,\cdot)$, $r_{\mathrm{arb}}(\cdot,\cdot)$ and $c_{\mathrm{opt}}(\cdot,\cdot)$ are as in Lemma 5.1.

Proof (i) Since $N(f)$ vanishes on span$\{h\}$, information consists of pure noise only. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary algorithm. Then, for any $a \in \mathbb{R}^n$ with $\|a\|_Y \leq \delta$, the error of the constant algorithm $\varphi_a \equiv \varphi(a)$ is not larger than the error of φ. Hence, zero provides the best approximation and the minimal error equals $\sqrt{S(C_\mu S)} = |S(h)|$, as claimed.

(ii) Define the random variable $\alpha = \alpha(f)$ by $f = \alpha h$. Then α has standard Gaussian distribution. Similarly as in the proof of Lemma 4.3, letting $z = \Sigma^{-1/2} y / \|N(h)\|_Y$, we can transform the data $y = N(f) + x$ to $z = \alpha\, w + x'$ where $w = \Sigma^{-1/2} N(h)/\|N(h)\|_Y$ and $\|x'\|_2 \leq \delta/\|N(h)\|_Y$. Using an orthogonal matrix Q with $Qw = e_1$ we get that the original problem of approximating $S(f)$ from data y is equivalent to that of approximating $s(\alpha) = \alpha\, S(h)$, $\alpha \sim \mathcal{N}(0,1)$, from data

$$\tilde{y} \;=\; Qz \;=\; [\,\alpha + \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n]$$

where $\|x'\|_2 \leq \delta/\|N(h)\|_Y$.

It is now easily seen that the 'pure noise data' $\tilde{y}_2, \ldots, \tilde{y}_n$ do not count. Indeed, for an arbitrary algorithm $\varphi : \mathbb{R}^n \to \mathbb{R}$ we can define another algorithm $\varphi_0(\tilde{y}) = \varphi(\tilde{y}_1, \underbrace{0, \ldots, 0}_{n-1})$ which uses $\tilde{y}_1$ only. Then for any $\alpha \in \mathbb{R}$ we have ($\delta_h = \delta/\|N(h)\|_Y$)

$$\begin{aligned} \sup_{\|\tilde{x}\|_2 \leq \delta_h} |s(\alpha) - \varphi(\tilde{y})| \;&\geq\; \sup_{|\tilde{x}_1| \leq \delta_h} |s(\alpha) - \varphi(\tilde{y}_1, \underbrace{0, \ldots, 0}_{n-1})| \\ &=\; \sup_{\|\tilde{x}\|_2 \leq \delta_h} |s(\alpha) - \varphi_0(\tilde{y})| \end{aligned}$$

and hence $\mathrm{e}^{\mathrm{a\text{-}w}}(\varphi) \geq \mathrm{e}^{\mathrm{a\text{-}w}}(\varphi_0)$.

Thus, we have reduced the original problem to that of approximating $s(\alpha)$ from $\tilde{y}_1 = \alpha + \tilde{x}_1$ where $\alpha \sim \mathcal{N}(0,1)$ and $|\tilde{x}_1| \leq \delta/\|N(h)\|_Y$. The formulas for the minimal errors and optimal linear algorithm now follow from Lemma 5.1 and the fact that $\tilde{y}_1 = \langle y, N(h)\rangle_Y / \|N(h)\|_Y^2$. □

Assume now that the Gaussian measure μ is arbitrary. For $\sigma^2 \geq 0$, let φ_σ be the optimal algorithm in the average case setting with the measure μ and information $y = N(f) + x$ where the noise $x \sim \mathcal{N}(0, \sigma^2\Sigma)$. Recall that φ_σ is given as $\varphi_\sigma(y) = \langle y, w_\sigma\rangle_2$, where w_σ is the only solution of $(\sigma^2\Sigma + G_N) w_\sigma = N(C_\mu S)$ belonging to $(\sigma^2\Sigma + G_N)(\mathbb{R}^n)$, and the matrix $G_N = \{L_i(C_\mu L_j)\}_{i,j}$; see Section 3.5.

It is easy to see that for $\sigma^2 = \delta^2$ the algorithm φ_σ is almost optimal linear in the mixed average-worst setting. Indeed, let $\varphi = d\langle \cdot, w\rangle_Y$ with $\|w\|_Y = 1$ and $d \geq 0$ be a linear algorithm. Then the worst case error for $f \in F$ equals

$$\sup_{\|x\|_Y \leq \delta} |S(f) - \varphi(N(f)+x)|^2 \;=\; (|S(f) - d\langle N(f), w\rangle_Y| \,+\, \delta\, d\,)^2,$$

while the average case error for f equals

$$\int_{\mathbb{R}^n} |S(f) - \varphi(N(f)+x)|^2 \,\pi(dx) \;=\; |S(f) - d\langle N(f), w\rangle_Y|^2 \,+\, \delta^2 d^2.$$

Hence,

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}, \varphi; \mu) \;\leq\; \mathrm{e}^{\mathrm{a-w}}(\mathbb{N}, \varphi; \mu) \;\leq\; \sqrt{2} \cdot \mathrm{e}^{\mathrm{ave}}(\mathbb{N}, \varphi; \mu),$$

and consequently

$$\mathrm{e}^{\mathrm{a-w}}(\mathbb{N}, \varphi_\sigma; \mu) \;\leq\; \sqrt{2} \cdot \mathrm{rad}^{\mathrm{a-w}}(\mathbb{N}; \mu) \qquad (\sigma^2 = \delta^2).$$

(Compare with the corresponding discussion in the worst-average case of Section 4.2.2).

For an appropriately chosen σ the algorithm φ_σ turns out to be strictly optimal linear. To show this, we first need some preliminary facts.

For $\sigma \geq 0$, let $K_\sigma = S - \varphi_\sigma(N(\cdot)\,)$. Recall that, in the average case setting, the functional K_σ determines a family of one-dimensional subproblems which are as difficult as the original problem, see Theorem 3.4 of Section 3.5. Define the function $\rho : [0, +\infty) \to (0, +\infty]$ as

$$\rho(\sigma) \;=\; \frac{\|K_\sigma\|_\mu}{\|N(C_\mu K_\sigma)\|_Y}$$

for $\sigma^2 > 0$, and $\rho(0) = \lim_{\sigma \to 0^+} \rho(\sigma)$.

Lemma 5.3 *If $N(C_\mu S) \neq 0$ then the function $\rho(\sigma)$ is well defined for all $\sigma \geq 0$.*

Proof We first show that for $\sigma^2 > 0$ is $N(C_\mu K_\sigma) \neq 0$. Indeed, using the formula for φ_σ we obtain

$$\begin{aligned} N(C_\mu K_\sigma) &= N(C_\mu S) - \sum_{j=1}^n w_{\sigma,j} N(C_\mu L_j) \\ &= N(C_\mu S) - G_N w_\sigma \;=\; \sigma^2 \Sigma w_\sigma. \end{aligned}$$

As $N(C_\mu S) \neq 0$, we also have $w_\sigma \neq 0$ and consequently $N(C_\mu K_\sigma) \neq 0$, as claimed.

Now, let us see what happens when $\sigma \to 0^+$. If the average radius of exact information ($\sigma = 0$) is positive then $N(C_\mu K_\sigma) = \sigma^2 \Sigma w_\sigma \to 0$ and $\|K_\sigma\|_\mu \to \|K_0\|_\mu = \|S - \varphi_0 N\|_\mu > 0$, which means that $\lim_{\sigma\to 0^+} \rho(\sigma) = +\infty$. Otherwise we have $\|K_0\|_\mu = 0$ and $S = \sum_{j=1}^n w_{0,j} L_j$ a.e. on F. This yields

$$\begin{aligned}\|K_\sigma\|_\mu^2 &= \sum_{i,j=1}^n (w_0 - w_\sigma)_i (w_0 - w_\sigma)_j L_i(C_\mu L_j) \\ &= \langle G_N(w_0 - w_\sigma), (w_0 - w_\sigma)\rangle_2 \;=\; \sigma^2 \langle \Sigma w_\sigma, w_0 - w_\sigma\rangle_2 .\end{aligned}$$

As $\|N(C_\mu K_\sigma)\|_Y^2 = \sigma^4 \|\Sigma w_\sigma\|_Y^2 = \sigma^4 \langle \Sigma w_\sigma, w_\sigma\rangle_2$, in this case ρ takes the form

$$\rho^2(\sigma) \;=\; \frac{\langle \Sigma w_\sigma, (w_0 - w_\sigma)\rangle_2}{\sigma^2 \langle \Sigma w_\sigma, w_\sigma\rangle_2}.$$

Let P_N be the orthogonal projection in $\mathbb{R}^n$ onto $X = G_N(\mathbb{R}^n)$ with respect to the Euclidean inner product. As $\Sigma w_\sigma \in X$, we have $(\sigma^2 P_N \Sigma + G_N) w_\sigma = N(C_\mu S) = G_N w_0$ which yields $(w_0 - w_\sigma) = \sigma^2 G_N^{-1} P_N \Sigma w_\sigma$. (For $x \in X$, $G_N^{-1} x$ is the only element $y \in X$ such that $G_N y = x$.) Thus, we finally obtain

$$\rho^2(\sigma) \;=\; \frac{\langle \Sigma w_\sigma, G_N^{-1} P_N \Sigma w_\sigma\rangle_2}{\langle \Sigma w_\sigma, w_\sigma\rangle_2} \;\longrightarrow\; \frac{\langle \Sigma w_0, G_N^{-1} P_N \Sigma w_0\rangle_2}{\langle \Sigma w_0, w_0\rangle_2}$$

as $\sigma \to 0^+$. □

We are ready to state the theorem about optimal linear algorithms.

Theorem 5.2 *Let $\delta > 0$.*

(i) If $\delta \|S\|_\mu \geq \sqrt{\pi/2}\, \|N(C_\mu S)\|_Y$ then the zero algorithm is optimal and $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) = \|S\|_\mu$.

(ii) If $\delta \|S\|_\mu < \sqrt{\pi/2}\, \|N(C_\mu S)\|$ then the optimal linear algorithm is φ_σ where $\sigma = \sigma(\delta) \geq 0$ is the (existing) solution of

$$c_{\mathrm{opt}}(\,1,\, \delta\, \rho(\sigma)\,) \;=\; \frac{1}{1 + \sigma^2 \rho^2(\sigma)}. \tag{5.4}$$

Furthermore, for $\sigma(\delta) > 0$ we have

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \;=\; \left(\|K_\sigma\|_\mu \,+\, \rho^{-1}(\sigma) \sqrt{\langle \Sigma w_\sigma, w_\sigma\rangle_2} \right) r_{\mathrm{lin}}(1, \delta\rho(\sigma)\,),$$

while for $\sigma(\delta) = 0$ we have $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) = \delta \sqrt{\langle \Sigma w_0, w_0\rangle_2}$.

Proof (i) We can assume that $N(C_\mu S) \neq 0$ (and consequently $\|S\|_\mu \neq 0$) since otherwise the theorem is obvious. Let $\mu_S(\cdot|g)$ be the conditional measure of μ given $g = P_S(f) = f - S(f)C_\mu S/\|S\|_\mu^2$. Due to Lemma 3.5, $\mu_S(\cdot|g)$ has the mean g and correlation operator

$$A_S(L) \;=\; \frac{L(C_\mu S)}{\|S\|_\mu^2} C_\mu S, \qquad \forall g \text{ a.e.}$$

This, Lemma 5.2 and Lemma 5.1 give

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N}; \mu_S(\cdot|g)\,) \;\geq\; \frac{S(C_\mu S)}{\|S\|_\mu} r_{\mathrm{lin}}\left(1, \frac{\delta\,\|S\|_\mu}{\|N(C_\mu S)\|_Y}\right) \;=\; \|S\|_\mu.$$

Hence,

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \;\geq\; \sqrt{\int_{P_S(F)} \Big(\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu_S(\cdot|g))\Big)^2 \,\mu P_S^{-1}(dg)} \;\geq\; \|S\|_\mu.$$

On the other hand, the error $\|S\|_\mu$ is achieved by $\varphi \equiv 0$. Hence, the zero algorithm is optimal linear.

(ii) We first show that there exists $\sigma = \sigma(\delta)$ satisfying the equality (5.4). Let ψ_l and ψ_r denote the left and right hand side of (5.4), respectively. As w_σ depends continuously on σ, ψ_l and ψ_r are continuous functions of σ on $(0,+\infty)$. If $\rho(0) < +\infty$, we also have continuity at 0. Hence, for existence of $\sigma(\delta)$ it suffices that the function $(\psi_l - \psi_r)(\sigma)$ takes positive and nonpositive values. Indeed, for $\sigma \to +\infty$ we have $w_\sigma \to 0$ and $\rho(\sigma) \to \|S\|_\mu/\|N(C_\mu S)\|_Y$. Hence,

$$\lim_{\sigma\to\infty} c_{\mathrm{opt}}(1, \delta\,\rho(\sigma)\,) \;=\; c_{\mathrm{opt}}\left(1, \frac{\delta\,\|S\|_\mu}{\|N(C_\mu S)\|_Y}\right) \;>\; c_{\mathrm{opt}}\left(1, \sqrt{\pi/2}\right) \;=\; 0.$$

On the other hand, $\lim_{\sigma\to\infty}(1+\sigma^2\rho^2(\sigma)\,)^{-1} = 0$, which means that for large σ is $\psi_l(\sigma) > \psi_r(\sigma)$. If $\rho(0) = +\infty$ then for small positive σ we have $\psi_l(\sigma) = 0 < \psi_r(\sigma)$. If $\rho(0) < +\infty$ then $\psi_l(0) \leq 1 = \psi_r(0)$.

Hence, there always exists $\sigma = \sigma(\delta) \geq 0$ such that $\psi_l(\sigma) = \psi_r(\sigma)$. Note that $\sigma(\delta) = 0$ only if $\rho(0) < +\infty$ and $c_{\mathrm{opt}}(1, \delta\rho(0)) = 1$.

We now prove optimality of φ_σ. Assume first that $\sigma(\delta) > 0$. Then

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \;\geq\; \sqrt{\int_{P_{K_\sigma}(F)} \Big(\mathrm{rad}_{\mathrm{aff}}^{\mathrm{a-w}}(\mathbb{N};\mu_{K_\sigma}(\cdot|g)\,)\Big)^2 \,\mu P_{K_\sigma}^{-1}(dg)}$$

where $P_{K_\sigma}(f) = f - K_\sigma(f) C_\mu K_\sigma / \|K_\sigma\|_\mu^2$. As the measures $\mu_{K_\sigma}(\cdot|g)$ have the same (independent of g) correlation operator

$$A_{K_\sigma}(L) \;=\; \frac{L(C_\mu K_\sigma)}{\|K_\sigma\|_\mu^2}\, C_\mu K_\sigma \qquad \forall g \text{ a.e.},$$

and they differ only by the mean g, the minimal error of affine algorithms over $\mu_{K_\sigma}(\cdot|g)$ is independent of g. That is,

$$\mathrm{rad}_{\mathrm{aff}}^{\mathrm{a-w}}(\mathbb{N}; \mu_{K_\sigma}(\cdot|g)\,) \;=\; \mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N}; \mu_{K_\sigma})$$

where $\mu_{K_\sigma} = \mu_{K_\sigma}(\cdot|0)$. Now we can use Lemma 5.2 to get the affine algorithm φ_g attaining $\mathrm{rad}_{\mathrm{aff}}^{\mathrm{a-w}}(\mathbb{N}; \mu_{K_\sigma}(\cdot|g))$. We obtain

$$\varphi_g(y) \;=\; S(g) \,+\, c_{\mathrm{opt}}\left(1, \frac{\delta}{\|N(h_\sigma)\|_Y}\right) \frac{S(h_\sigma)}{\|N(h_\sigma\|_Y} \left\langle y - N(g), \frac{N(h_\sigma)}{\|N(h_\sigma)\|_Y} \right\rangle_Y$$

where $h_\sigma = C_\mu K_\sigma / \|K_\sigma\|_\mu$. We find that

$$N(h_\sigma) \;=\; \frac{N(C_\mu K_\sigma)}{\|K_\sigma\|_\mu} \;=\; \frac{\sigma^2\,\Sigma\, w_\sigma}{\|K_\sigma\|_\mu},$$

$\|N(h_\sigma)\|_Y = \rho^{-1}(\sigma)$, and

$$S(h_\sigma) \;=\; \frac{S(C_\mu K_\sigma)}{\|K_\sigma\|_\mu} \;=\; \frac{\|K_\sigma\|^2 + \sigma^{-2}\|N(C_\mu K_\sigma)\|_Y^2}{\|K_\mu\|_\mu}$$

(compare with the proof of Theorem 3.4). Hence,

$$\varphi_g(y) \;=\; S(g) \,+\, c_{\mathrm{opt}}(1, \delta\rho(\sigma)\,)\,(1 + \sigma^2\rho^2(\sigma)\,)\,\langle y, w_\sigma\rangle_2.$$

Since σ satisfies (5.4) and for all $g \in P_{K_\sigma}(F)$ is $S(g) - \langle w_\sigma, N(g)\rangle_2 = K_\sigma(g) = 0$, we finally obtain

$$\varphi_g \;=\; \langle \cdot, w_\sigma\rangle_2 \;=\; \varphi_\sigma.$$

Thus the same (linear) algorithm φ_σ minimizes the errors over $\mu_{K_\sigma}(\cdot|g)$, $\forall g$ a.e. Hence, φ_σ is optimal linear.

To find the error of φ_σ, observe that $S(h_\sigma)$ can be written as

$$S(h_\sigma) \;=\; \Big(\|K_\sigma\|_\mu \,+\, \rho^{-1}(\sigma)\Big)\,\sqrt{\langle \Sigma w_\sigma, w_\sigma\rangle_2}.$$

This and Lemma 5.2 give

$$\begin{aligned}
\mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi_\sigma;\mu) &= \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N};\mu) \;=\; \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N};\mu_{K_\sigma}) \\
&= |S(h_\sigma)|\, r_{\mathrm{lin}}(1,\delta\,\rho(\sigma)\,) \\
&= \left(\|K_\sigma\|_\mu \,+\, \rho^{-1}(\sigma)\,\sqrt{\langle \Sigma w_\sigma, w_\sigma\rangle_2} \right) r_{\mathrm{lin}}(1,\delta\rho(\sigma)\,).
\end{aligned}$$

Consider now the case when $\sigma(\delta)=0$. Proceeding as for $\sigma(\delta)>0$ we get that for any $\gamma>0$

$$\begin{aligned}
\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N};\mu) &\geq \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N};\mu_{K_\gamma}) \\
&= \Big(\|K_\gamma\|_\mu \,+\, \rho^{-1}(\gamma) \Big)\, \sqrt{\langle \Sigma w_\gamma, w_\gamma\rangle_2}.
\end{aligned}$$

We have already noticed that in the case $\sigma(\delta)=0$ we have $\rho(0)<+\infty$ and $c_{\mathrm{opt}}(1,\delta\rho(0))=1$. In view of Lemma 5.1, this means that $r_{\mathrm{lin}}(1,\delta\rho(0))=\delta\rho(0)$. We also have $\|K_0\|_\mu=0$ which follows from the prove of Lemma 5.3. Hence, letting $\gamma\to 0^+$ and using continuity arguments we get

$$\begin{aligned}
\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N};\mu) &\geq \left(\|K_0\|_\mu \,+\, \rho^{-1}(0)\sqrt{\langle \Sigma w_0, w_0\rangle_2} \right) r_{\mathrm{lin}}(1,\delta\rho(0)\,) \\
&= \delta\,\sqrt{\langle \Sigma w_0, w_0\rangle_2}. \qquad (5.5)
\end{aligned}$$

On the other hand, in the case $\sigma(\delta)=0$ we have $S(f)=\varphi_0(Nf)$, $\forall f$ a.e. Hence,

$$\sup_{\|x\|_Y\leq\delta} |S(f)-\varphi_0(Nf)| \;=\; \sup_{\|x\|_Y\leq\delta} |\langle w_0,x\rangle_2| \;=\; \delta\,\sqrt{\langle \Sigma w_0, w_0\rangle_2}.$$

This and (5.5) give optimality of φ_0. The proof is complete. □

Similarly to the average case setting, we can introduce a concept of a family of one-dimensional subproblems. Any such a family is determined by a functional $K\in F^*$ and indexed by $g\in P_K(F)$ where $P_K(f)=f-K(f)C_\mu K/\|K\|_\mu^2$. For given g, the subproblem relies on minimizing the average-worst case error of linear algorithms with respect to the conditional measure $\mu(\cdot|g)$ whose mean is g and correlation operator $A_K(L)=L(C_\mu K)C_\mu(K)/\|K\|_\mu^2$. (Equivalently, the subproblem relies on minimizing the error with respect to μ using additional information that $P_K(f)=g$.) From the proof of Theorem 5.2 it follows that the subproblems determined by the functional K_σ are as difficult as the original problem. Denoting as before $\mu_K=\mu(\cdot|0)$, we have the following corollary.

Corollary 5.1 *Let $\sigma = \sigma(\delta)$ be defined by the equation (5.4). Then*

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \;=\; \sup_{K\in F^*} \mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu_K) \;=\; \mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu_{K_\sigma}). \quad \square$$

For arbitrary algorithms, we can show a result corresponding to Theorem 4.3 of the first mixed setting.

Theorem 5.3 *We have*

$$1 \;\le\; \frac{\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu)}{\mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N};\mu)} \;\le\; \kappa_2^*$$

where κ_2^* is defined by (5.3). Furthermore, $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \approx \mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N};\mu)$ *as $\delta \to 0^+$.*

Proof Take $\sigma = \sigma(\delta)$ such that the algorithm φ_σ is optimal linear. In view of Lemma 5.2 we have

$$\frac{\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu_{K_\sigma}(\cdot|g))}{\mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N};\mu_{K_\sigma}(\cdot|g))} \;\le\; \kappa_2^* \qquad \forall\, g \text{ a.e.}$$

This and Theorem 5.2 yield that for an arbitrary algorithm φ

$$\begin{aligned}
\mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi;\mu) \;&=\; \sqrt{\int_{P_{K_\sigma}(F)} (\mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi;\mu_{K_\sigma}(\cdot|g)\,)\,)^2\, \mu P_{K_\sigma}^{-1}(dg)} \\
&\ge\; \frac{1}{\kappa_2^*}\sqrt{\int_{P_{K_\sigma}(F)} (\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu_{K_\sigma}(\cdot|g)\,)\,)^2\, \mu P_{K_\sigma}^{-1}(dg)} \\
&=\; \frac{1}{\kappa_2^*}\,\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu)\,,
\end{aligned}$$

which proves the first part of the theorem.

Let r_0 be the error of exact information ($\delta = 0$). To show $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \approx \mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N};\mu)$, it suffices to consider $r_0 = 0$ since otherwise $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \to r_0$ and $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \to r_0$ as $\delta \to 0$. However, $r_0 = 0$ implies $\rho(0) < +\infty$ and consequently $\delta\rho(\sigma(\delta)\,) \to 0$ as $\delta \to 0^+$. Using again Lemma 5.2 and Theorem 5.1 we obtain

$$\frac{\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu)}{\mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N};\mu)} \;\le\; \frac{r_{\mathrm{lin}}(1,\delta\rho(\sigma(\delta))\,)}{r_{\mathrm{arb}}(1,\delta\rho(\sigma(\delta))\,)} \;\to\; 1 \qquad \text{as } \delta \to 0^+.$$

This completes the proof. $\square$

Thus nonlinear algorithms can be only slightly better than linear algorithms.

5.2.3 A correspondence theorem

In Section 4.2.3 we established close relations between optimal approximation of functionals in the mixed worst-average setting and in the other settings. In this section we show similar relations for the mixed average-worst setting. They follow from the results of Section 5.2.2.

Let S be a continuous linear functional defined on a separable Banach space F. Let μ be a zero mean Gaussian measure on F and $\Sigma = \Sigma^* > 0$. Let $H \subset F$ be the associated with μ Hilbert space, so that $\{H, F_1\}$, $F_1 = \mathrm{supp}\ \mu$, is an abstract Wiener space. We consider the problem of approximating $S(f)$ from noisy information $y = N(f) + x$ in the following four settings.

P1: Mixed average-worst setting with the measure μ on F and the noise $\|x\|_Y = \sqrt{\langle \Sigma^{-1}x, x\rangle_2} \le \delta$.

P2: Worst case setting with E being the unit ball in H and $\|x\|_Y = \sqrt{\langle \Sigma^{-1}x, x\rangle_2} \le \delta$.

P3: Average case setting with the measure μ and the noise $x \sim \mathcal{N}(0, \sigma^2\Sigma)$.

P4: Mixed worst-average setting with E being the unit ball in H and $x \sim \mathcal{N}(0, \sigma^2\Sigma)$.

As always, we denote by φ_σ the optimal (linear) algorithm in the average case setting. Recall that φ_σ can be interpreted as the smoothing spline algorithm, $\varphi_\sigma(y) = S(\mathbf{s}(y)\,)$ where $\mathbf{s}(y)$ is the minimizer of $\|f\|_H^2 + \sigma^{-2}\|y - N(f)\|_Y^2$ in H.

Theorem 5.4 *Let $\sigma^2 = \delta^2$. Then we have*

$$\mathrm{e}^{\mathrm{a-w}}(\mathbb{N}, \varphi_\sigma; \mu) \;\le\; \sqrt{2}\,\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N}; \mu) \;\le\; \kappa_2^*\sqrt{2}\,\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N}; \mu)$$

and

$$\begin{aligned}
\frac{1}{\kappa_2^*\sqrt{2}}\,\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E) \;&\le\; \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N}; \mu) \;\le\; \sqrt{2}\,\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}; E),\\
\frac{1}{\kappa_2^*}\,\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu) \;&\le\; \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N}; \mu) \;\le\; \sqrt{2}\,\mathrm{rad}^{\mathrm{ave}}(\mathbb{N}; \mu),\\
\frac{1}{\kappa_2^*}\,\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}; E) \;&\le\; \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N}; \mu) \;\le\; \sqrt{2}\,\mathrm{rad}^{\mathrm{w-a}}_{\mathrm{arb}}(\mathbb{N}; E). \quad \Box
\end{aligned}$$

We also showed that for any $\delta \in [0, +\infty]$ we can find $\sigma = \sigma(\delta) \in [0 + \infty]$ such that the (smoothing spline) algorithm φ_σ (with convention $\varphi_\infty \equiv 0$) is optimal linear in the mixed setting (P1) and in the average case setting (P3). Clearly, the inverse relation is also true. For any $\sigma^2 \in [0 + \infty]$ there is $\delta = \delta(\sigma^2) \in [0. + \infty]$ such that φ_σ is optimal in both settings.

Taking together Theorems 5.4, 4.5, wnd 3.7, we obtain an almost equivalence of *all* four settings for approximating linear functionals. Namely, if only the set E is the unit ball induced by the measure μ, and $\delta^2 = \sigma^2$, then in all four settings:

- the minimal achievable errors are almost the same,
- the same algorithm φ_σ is almost optimal, and
- for varying δ and σ^2, the optimal linear algorithms belong to the common class of smoothing spline algorithms.

Notes and Remarks

NR 5.1 The main results of this section come from Plaskota [82]. Theorem 5.4 is new.

Exercises

E 5.1 Suppose we want to approximate a real parameter $f \sim \mathcal{N}(0, \lambda)$ based on n observations $y_i = f + x_i$ with noise satisfying $\|x\|_2 = \sqrt{\sum_{j=1}^n x_i^2} \le \delta$. Show that the sample mean, $\varphi_n(y) = n^{-1} \sum_{j=1}^n y_j$, is an asymptotically optimal algorithm,

$$\mathrm{e}^{\mathrm{a-w}}(\varphi_n) \;=\; \frac{\delta}{\sqrt{n}} \;\approx\; \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(n) \qquad \text{as } n \to +\infty,$$

where $\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(n)$ is the corresponding nth minimal error of arbitrary algorithms.

E 5.2 Suppose that the noise in the problem of E 5.1 satisfies $\sum_{j=1}^n x_j^2/\delta_j^2 \le 1$. Show that for the algorithm

$$\varphi(y) \;=\; \frac{\sum_{i=1}^n \delta_i^{-2} y_i}{\sum_{i=1}^n \delta_i^{-2}}$$

we have

$$\mathrm{e}^{\mathrm{a-w}}(\varphi) \;=\; \sqrt{\frac{1}{\sum_{i=1}^n \delta_i^{-2}}} \;\approx\; \mathrm{rad}^{\mathrm{a-w}}(\mathbb{N})$$

as $\lambda \to +\infty$.

E 5.3 Prove the uniqueness of the optimal linear algorithm of Theorem 5.2.
Hint: Consider first the case when μ is concentrated on a one-dimensional subspace.

E 5.4 Show that the solution $\sigma = \sigma(\delta)$ of (5.4) not only exists for any $\delta > 0$, but it is also determined uniquely.

E 5.5 Let μ be the standard Gaussian distribution on $F = \mathbb{R}^n$, $\mu = \mathcal{N}(0, I)$. Consider approximation of a functional S from information $y = f + x \in \mathbb{R}^n$ where $\|x\|_2 \le \delta$. Show that for $\delta \ge \sqrt{\pi/2}$ the optimal algorithm is $\varphi_\infty \equiv 0$, while for $\delta < \sqrt{\pi/2}$ it is given as $\varphi_\sigma(y) = (1+\sigma^2)^{-1} S(y)$ where $\sigma = \sigma(\delta) = \sqrt{1/c_{\mathrm{opt}}(1,\delta) - 1}$. Furthermore, the error of φ_σ equals $\|S\|_2 \, r_{\mathrm{lin}}(1,\delta)$.

E 5.6 Let $\delta > 0$. Show that the necessary and sufficient condition for the algorithm φ_0 to be optimal is that $K_0 = S - \varphi_0 N = 0$ a.e. on F, and

$$\delta^2 \, \frac{\pi}{2} \, \frac{\langle \Sigma w_0, G_N^{-1} P_N \Sigma w_0 \rangle_2}{\langle \Sigma w_0, w_0 \rangle_2} \;\le\; 1,$$

where w_0, G_N and P_N are as in the proof of Lemma 5.3.

5.3 Approximation of operators

As we already noticed, we know very little about approximation of operators which are not functionals. Here we present some very special results.

Suppose we approximate a vector $f = (f_1, \ldots, f_n) \in \mathbb{R}^n$ whose coordinates f_i have independent normal distributions, $f_i \sim \mathcal{N}(0, \lambda_i)$ where $\lambda_i > 0$, $1 \le i \le n$. That is, the joint probability distribution μ on $\mathbb{R}^n$ is zero mean Gaussian and its correlation matrix is diagonal. Information about f is given coordinatewise, $y_i = f_i + x_i$ where $|x_i| \le \delta_i$, $1 \le i \le n$.

Lemma 5.4 *We have*

$$\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N}) \;=\; \sqrt{\sum_{i=1}^n r^2_{\mathrm{lin}}(\lambda_i, \delta_i)}\,,$$

$$\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N}) \;=\; \sqrt{\sum_{i=1}^n r^2_{\mathrm{arb}}(\lambda_i, \delta_i)}\,,$$

and the (unique) optimal linear algorithm is given as $\varphi(c_1 y_1, \ldots, c_n y_n)$ where $c_i = c_{\mathrm{opt}}(\lambda_i, \delta_i)$, $1 \le i \le n$.

Proof Let $\mu_i = \mathcal{N}(0, \lambda_i)$. Due to independence of f_i's and x_i's, the error of any algorithm $\varphi = (\varphi_1, \ldots, \varphi_n)$ equals

$$\left(\mathrm{e}^{\mathrm{a-w}}(\mathbb{N}, \varphi)\right)^2 \;=\; \sum_{i=1}^{n} \left(\int_{\mathbb{R}} \sup_{|x_i| \le \delta_i} \; (f_i - \varphi_i(f_i + x_i)\,)^2 \, \mu_i(df_i) \right).$$

Hence, the lemma follows from Lemma 5.1 about optimal algorithms for the one-dimensional problem. □

Recall that for $\delta_i \le \sqrt{2\lambda_i/\pi}$ we have $c_i = 1$ and $r_{\mathrm{lin}}(\lambda_i, \delta_i) = \delta_i$. Hence, for sufficiently small noise (or for sufficiently large λ_i's), the algorithm $\varphi(y) = y$ is optimal linear and its error equals $(\sum_{i=1}^n \delta_i^2)^{1/2}$. This observation can be generalized as follows.

Consider the same problem but with noise x belonging to a set $B \subset \mathbb{R}^n$, i.e., $\mathbb{N}(f) = \{\, f + x \mid x \in B \,\}$. Denote

$$\rho(B) \;=\; \sup_{x \in B} \|x\|_2.$$

Lemma 5.5 *Suppose the set B is convex and orthosymmetric. If there exists $\overline{x} \in \overline{B}$ such that $\|\overline{x}\|_2 = \rho(B)$ and $|\overline{x}_i| \le \sqrt{2\lambda_i/\pi}$, $1 \le i \le n$, then the algorithm $\varphi(y) = y$ is optimal linear and* $\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N}) = \rho(B)$. *We also have* $\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N}) \approx \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N})$ *as $\rho(B) \to 0$.*

Proof From convexity and orthosymmetry it follows that $\overline{B}$ includes the rectangle $\mathcal{R} = \{\, x \in \mathbb{R}^n \mid |x_i| \le |\overline{x}_i|,\ 1 \le i \le n \,\}$. Hence, from Lemma 5.4 we obtain

$$\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N}) \;\ge\; \sqrt{\sum_{i=1}^n |\overline{x}_i|^2} \;=\; \rho(B).$$

On the other hand, for the identity algorithm we have $\mathrm{e}^{\mathrm{a-w}}(\mathbb{N}, \varphi; E) = \rho(B)$.

To show the remaining part of the lemma, observe that for $\rho(B) \to 0^+$

$$\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N}) \;\ge\; \sqrt{\sum_{i=1}^n r^2_{\mathrm{arb}}(\lambda_i, |\overline{x}_i|)} \;\approx\; \sqrt{\sum_{i=1}^n r^2_{\mathrm{lin}}(\lambda_i, |\overline{x}_i|)} \;=\; \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N}).$$

Hence, $\mathrm{rad}^{\mathrm{a-w}}_{\mathrm{arb}}(\mathbb{N}) \approx \mathrm{rad}^{\mathrm{a-w}}_{\mathrm{lin}}(\mathbb{N})$, as claimed. □

Thus the identity algorithm is optimal linear if the noise belongs to a convex and orthosymmetric set B whose radius is sufficiently small. If E is an ellipsoid, $\sum_{i=1}^n x_i^2/\delta_i^2$ with $\delta_1 \geq \cdots \geq \delta_n > 0$, then $\overline{x} = (\delta_1, \underbrace{0, \ldots, 0}_{n-1})$. In this case, the sufficient condition for the identity algorithm to be optimal linear is that $\delta_1 \leq \sqrt{2\lambda_1/\pi}$.

In the end, let us consider the (generalized) least squares algorithm $\varphi_{\mathrm{ls}}(y)$ for linear problems S defined on $\mathbb{R}^d$. We assume that the information operator $\mathbb{N}$ is linear with $N : \mathbb{R}^d \to Y = \mathbb{R}^n$, $\dim N(\mathbb{R}^d) = d$, and noise x is bounded in a Hilbert norm, $\|x\|_Y = \sqrt{\langle x, x\rangle_Y} \leq \delta$. The Gaussian measure μ on $\mathbb{R}^d$ has the mean zero and its correlation operator C_μ is positive definite.

Recall that $\varphi_{\mathrm{ls}} = SN^{-1}P_N$ (where P_N is the orthogonal projection onto $N(\mathbb{R}^d)$ with respect to $\langle \cdot, \cdot \rangle_Y$). For small noise level δ, in all three previously analyzed settings φ_{ls} is either optimal or close to optimal, see Theorems 2.12, 3.6, 4.8. The following theorem shows optimality properties of φ_{ls} in the average-worst setting. It can be viewed as a generalization of Lemma 5.5 for ellipsoidal B.

Theorem 5.5 *Let $\overline{g} \in G$ be such that $\|\overline{g}\| = 1$ and*

$$\|S(N^*N)^{-1}S^*\overline{g}\| \;=\; \|S(N^*N)^{-1}S^*\|.$$

Then for sufficiently small δ,

$$\delta^2 \cdot \left\langle S(N^*N)^{-1}C_\mu^{-1}(N^*N)^{-1}S^*\overline{g},\, \overline{g} \right\rangle \;\leq\; \frac{2}{\pi}\,\|S(N^*N)^{-1}S^*\|, \qquad (5.6)$$

the generalized least squares φ_{ls} is an optimal linear algorithm and

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N}) \;=\; \delta \cdot \sqrt{\|S(N^*N)^{-1}S^*\|}\,.$$

Furthermore, $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N}) \approx \mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N})$ *as $\delta \to 0^+$.*

Proof We shall use once more the concept of the one dimensional subproblem. Namely, suppose that we have additional information that f is in the subspace $\mathrm{span}\,\overline{h}$ where $\overline{h} = (N^*N)^{-1}S^*\overline{g}$. Due to Lemma 3.5, this corresponds to changing the measure μ to $\tilde{\mu}$ which is zero mean Gaussian and its correlation operator equals

$$A \;=\; \frac{\langle \cdot, \overline{h}\rangle_2\, \overline{h}}{\|C_\mu^{-1}\overline{h}\|_\mu^2} \;=\; \langle \cdot, h\rangle_2\, h,$$

$h = \overline{h}/\|C_\mu^{-1}\overline{h}\|_\mu$. We obviously have $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \ge \mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\tilde{\mu})$, and using Lemma 5.2,

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\tilde{\mu}) \;=\; \|S(h)\|\, r_{\mathrm{lin}}\left(1, \frac{\delta}{\|N(h)\|_Y}\right).$$

As

$$\|N(\overline{h})\|_Y^2 \;=\; \langle\, S(N^*N)^{-1}S^*\overline{g},\overline{g}\,\rangle \;=\; \|S(N^*N)^{-1}S^*\|$$

and

$$\|C_\mu^{-1}\overline{h}\|_\mu^2 \;=\; \langle \overline{h}, C_\mu^{-1}\overline{h}\rangle_2 \;=\; \langle\, S(N^*N)^{-1}C_\mu^{-1}(N^*N)^{-1}S^*\overline{g},\overline{g}\rangle,$$

the condition (5.6) is equivalent to $\delta/\|N(h)\|_Y \le \sqrt{2/\pi}$. This means that

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\tilde{\mu}) \;=\; \|S(h)\|\,\frac{\delta}{\|N(h)\|_Y} \;=\; \delta\cdot\|S(N^*N)^{-1}S^*\|^{1/2}.$$

On the other hand, we know that for the least squares we have

$$\sup_{\|x\|_Y\le\delta} \|S(f)-\varphi_{\mathrm{ls}}(N(f)+x)\,\|^2 \;=\; \delta^2\cdot\|S(N^*N)^{-1}S^*\|$$

(compare this with the corresponding part of the proof of Theorem 2.12). Since the last expression is independent of f, we obtain $\mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi_{\mathrm{ls}};\mu) = \delta\,\|S(N^*N)^{-1}S^*\|$, and consequently $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) = \mathrm{e}^{\mathrm{a-w}}(\mathbb{N},\varphi_{\mathrm{ls}};\mu)$.

To complete the proof, observe that for $\delta\to 0^+$ we have

$$\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \;=\; \mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\tilde{\mu}) \;\approx\; \mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N};\tilde{\mu}) \;\le\; \mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N};\mu).$$

Thus $\mathrm{rad}_{\mathrm{lin}}^{\mathrm{a-w}}(\mathbb{N};\mu) \approx \mathrm{rad}_{\mathrm{arb}}^{\mathrm{a-w}}(\mathbb{N};\mu)$, as claimed.

Notes and Remarks

NR 5.2 All results of Section 5.3 are original.

Exercises

E 5.7 Suppose we approximate an operator $S: F\to G$ from information $y = [N(f)+x,t]$ where the noise $(x,t)\in E$. Prove that if E satisfies

$$(x,t)\in E \qquad\Longrightarrow\qquad (x,0)\in E, \tag{5.7}$$

then the "pure noise" data t do not count. That is, the optimal algorithm uses $y^{(1)} = N(f)+x$ only. Give an example showing that the condition (5.7) is essential.

E 5.8 Consider approximation of $f \in \mathbb{R}^n$ where f_i are independent and $f_i \sim \mathcal{N}(0, \lambda_i)$, based on information $y = f + x$, $x \in E$. Show that if the set E is sufficiently large,

$$E \supset \{ x \in \mathbb{R}^n \mid \quad |x_i|^2 \le \lambda_i \, \pi/2, \, 1 \le i \le n \},$$

then zero is the best linear algorithm. In particular, if E is an ellipsoidal set, $E = \{ x \in \mathbb{R}^n \mid \sum_{j=1}^n x_j^2/\delta_j^2 \le 1\}$, then the sufficient condition for the zero algorithm to be optimal linear is $\sum_{j=1}^n \lambda_j/\delta_j^2 \le 2/\pi$.

E 5.9 Consider the problem of approximating a vector $f \in \mathbb{R}^d$ whose distribution is zero mean Gaussian with full support. Let $r_n(\delta_1, \ldots, \delta_n)$ $(0 < \delta_1 \le \cdots \le \delta_n)$ be the minimal error that can be attained by linear algorithms using n $(n \ge d)$ observations $y_i = \langle f, f_i \rangle_2 + x_i$ with noise $\sum_{j=1}^n x_j^2/\delta_j^2 \le 1$, where $\|f_i\|_2 \le 1$, $1 \le i \le n$. Denote by $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$ the eigenvalues of S^*S. Show that for sufficiently small δ_i's we have

$$r_n(\delta_1, \ldots, \delta_n) \;=\; \min \max_{1 \le i \le d} \sqrt{\frac{\lambda_i}{\eta_i}},$$

where the minimum is taken over all $\eta_i \ge 0$ satisfying

$$\sum_{j=r}^n \eta_j \;\le\; \sum_{j=r}^n \delta_j^{-2}, \qquad 1 \le r \le n.$$

In particular, for fixed noise levels, $\delta_i = \delta$ $\forall i$, and large n we have

$$r_n(\delta) \;=\; \frac{\delta}{\sqrt{n}} \cdot \sqrt{\sum_{j=1}^d \lambda_j} \, .$$

Find the optimal information.

Hint: To get the optimal information, use Lemma 2.14.

Chapter 6

Asymptotic setting

6.1 Introduction

In Chapters 2 to 5 we were interested in finding a *single* information and algorithm which minimize an error or cost of approximation. In this chapter we study asymptotic behavior of algorithms. The aim is to construct a *sequence* of algorithms, such that for any problem element f the error of successive approximations vanishes as fast as possible, as the number of observations increases to infinity.

A motivation for analyzing the asymptotic setting comes from real–life computations. It suffices to mention only the Romberg algorithm for computing integrals, or finite element methods (FEM) for solving partial differential equations. When dealing with these and other numerical algorithms, we are usually interested in how fast they converge to a solution. Another motivation comes from some negative results in the worst case setting.

Example 6.1 Consider a problem with one-to-one compact solution operator S acting between separable Hilbert spaces F and G. Assume, for simplicity, that information is exact and given as

$$N^n \;=\; [\,\langle\cdot,\xi_1\rangle_F, \langle\cdot,\xi_2\rangle_F, \ldots, \langle\cdot,\xi_n\rangle_F\,]\,.$$

If we want to study the worst case and do not have any a priori knowledge about $\|f\|_F$, then we have to assume that the worst case error is taken over the whole space F. In this case, error of any algorithm is infinite.

On the other hand, for the corresponding to N^n spline algorithm φ^n_{spl} we have

$$\|S(f) - \varphi^n_{\mathrm{spl}}(y^n)\| \;\le\; \|f\|_F \cdot \sup\{\, \|S(h)\| \;|\; \ \|h\|_F \le 1,\ h \in \ker N^n \,\}\,,$$

for all $f \in F$ and $y^n = N^n(f)$. Hence, If only the elements ξ_j are selected in such a way that $F = \overline{\text{span}\,\{\xi_1, \xi_2, \ldots\}}$, the succesive approximations $\varphi^n_{\text{spl}}(y^n)$ converge to $S(f)$ with $n \to \infty$, and this convergence is independent of $\|f\|_F$. Hence, although the worst case error is infinite, we can construct an algorithm which converges to the solution. □

We present two kinds of results dependent on wheather we have deterministic or stochastic assumptions on the problem elements and information noise. In both cases, we assume that the solution operator as well as information is linear. We focus attention on relations between the asymptotic and worst or average case settings, correspondingly.

This chapter consists of two sections. In Section 6.2, we study relations between the asymptotic and worst case settings in the case when information noise is bounded in a norm. We show that an upper bound on the rate of convergence of algorithms is provided by the worst case radii $\text{rad}^{\text{wor}}(\mathbb{N}^n)$ taken over the unit ball of F. It turns out that if F is a Banach space, this convergence cannot be essentially beaten by any algorithm. More precisely, in any ball of F we can find an element f such the for some information $y^n \in \mathbb{N}^n(f)$, $n \geq 1$, the error $\|S(f) - \varphi^n(y^n)\|$ essentially behaves as $\text{rad}^{\text{wor}}(\mathbb{N}^n)$. Hence, algorithms optimal in the worst case are also optimal in the asymptotic setting. The assumption that F is a Banach space is crucial. We also consider the problem of optimal information.

In Section 6.3, we assume that information noise is Gaussian and that we have some Gaussian measure on F. In this case, we show relations between the asymptotic and average case settings. Namely, we first prove that the spline algorithm (which is optimal in the average case) gives the best possible convergence. Any other algorithm can converge better only on a set of measure zero.

Then we investigate the rate of convergence of the spline algorithm. We show that this convergence can be characterized by the sequence of average radii $\text{rad}^{\text{ave}}(\mathbb{N}^n)$. Finally, we give results on optimal information.

6.2 Asymptotic and worst case settings

We start with the formal description of the asymptotic setting with deterministic information noise.

6.2.1 Information, algorithm and error

In the asymptotic setting, we are interested in the behavior of algorithms as the number of observations increases to infinity. Therefore it is convenient to define information and algorithm as infinite sequences. Namely, a nonadaptive information operator $\mathbb{N}$ is a pair, $\mathbb{N} = \{N, \Delta\}$ where $N : F \to \mathbb{R}^\infty$ is an exact information operator,

$$N \;=\; [\,L_1, L_2, L_3, \dots\,],$$

and $\Delta \in \mathbb{R}^\infty$ is a precision sequence,

$$\Delta \;=\; [\,\delta_1, \delta_2, \delta_3, \dots\,].$$

For given $\mathbb{N}$, by N^n and Δ^n we denote the first n components of N and Δ. In particular,

$$N^n(f) \;=\; [\,L_1(f), L_2(f), \dots, L_n(f)\,].$$

We say that an infinite sequence $y = [y_1, y_2, y_3, \dots] \in \mathbb{R}^\infty$ is (noisy) information about $f \in F$ and write $y \in \mathbb{N}(f)$ iff for all $n \geq 1$ the vector $x^n = y^n - N^n(f)$ is in the given set $B(\Delta^n, N^n(f)) \subset \mathbb{R}^n$ of all possible values of the nth information noise corresponding to exact information $N^n(f)$. Here $y^n = [y_1, \dots, y_n]$ and $B(\Delta^n, N^n(f))$ satisfy conditions of Section 2.7.1. That is, 1. $B(0, z) = \{0\}$, 2. If $\Delta^n \leq \bar{\Delta}^n$ then $B(\Delta^n, z) \subset B(\bar{\Delta}^n), z)$, and 3. $B(\Delta^n, z^n) = \{\, x \in \mathbb{R}^n \mid \exists a \in \mathbb{R},\, [x, a] \in B(\Delta^{n+1}, z^{n+1})\,\}$.

Defining the nth information operator $\mathbb{N}^n = \{N^n, \Delta^n\}$ as

$$\mathbb{N}^n(f) \;=\; \{\, y^n \in \mathbb{R}^n \mid \quad y^n - N^n(f) \in B(\Delta^n, N^n(f))\,\},$$

we can equivalently say that $y \in \mathbb{R}^n$ is noisy information about f iff for all $n \geq 1$ the vector y^n is the nth noisy information about f, $y^n \in \mathbb{N}^n(f)$.

We will consider only the case when $B(\Delta^n, z^n) = B(\Delta^n)$ are unit balls in some extended norms $\|\cdot\|_{\Delta^n}$ of $\mathbb{R}^n$. We recall that in this case the conditions 1.-3. imply

$$\|x\|_{\Delta^n} \;=\; \min_{t \in \mathbb{R}} \|[x, t]\|_{\Delta^{n+1}}, \qquad n \geq 1,\; x \in \mathbb{R}^n \tag{6.1}$$

(see E 2.45).

We now pass to adaptive information. An adaptive information operator is a family $\mathbb{N} = \{\mathbb{N}_y\}_{y \in \mathbb{R}^\infty}$ where $\mathbb{N}_y = \{N_y, \Delta_y\}$ is nonadaptive information with

$$N_y \;=\; [\,L_1, L_2(\cdot; y_1), \dots, L_n(\cdot; y_1, \dots, y_{n-1}), \dots\,]$$

and

$$\Delta_y \;=\; [\,\delta_1, \delta_2(y_1), \ldots, \delta_n(y_1, \ldots, y_{n-1}), \ldots\,].$$

For adaptive $\mathbb{N}$, a sequence y is called (noisy) information about f iff $y^n \in \mathbb{N}_y^n(f)$, $\forall n \geq 1$.

For a given solution operator $S : F \to G$ where F and G are normed spaces, an approximation to $S(f)$ is provided by an algorithm. By an algorithm we mean a sequence of transformations, $\varphi = \{\varphi^n\}_{n \geq 0}$, where $\varphi^n : \mathbb{R}^n \to G$. ($\varphi^0$ is a fixed element of G.) The nth error of approximating $S(f)$ based on (adaptive or nonadaptive) information $y \in \mathbb{N}(f)$ is defined by the difference $\|S(f) - \varphi^n(y^n)\|$.

6.2.2 Optimal algorithms

Our first goal is to characterize the best possible behavior of the error $\|S(f) - \varphi^n(y^n)\|$, $f \in F$, $y \in \mathbb{N}(f)$, for a fixed (adaptive) information operator $\mathbb{N}$. A crucial role in our analysis will play the nth (worst case) radii of nonadaptive information $\mathbb{N}_y$. They are given as the usual worst case radii of $\mathbb{N}_y^n$ with respect to the unit ball of F,

$$\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y) \;=\; \inf_{\varphi^n} \sup_{\|f\|_F \leq 1} \sup_{z \in \mathbb{N}_y(f)} \|S(f) - \varphi^n(z^n)\|.$$

We assume that the solution operator S is linear. Recall that in this case the following formula is valid:

$$\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y) \;=\; \alpha \cdot \sup\{\, \|S(h)\| \;|\; \|h\|_F \leq 1,\, \|N_y^n(h)\|_{\Delta_y^n} \leq 1\,\} \qquad (6.2)$$

where $\alpha \in [1, 2]$ (comp. with Theorem 2.2).

Given $\mathbb{N}$, it is not difficult to construct an algorithm for which the error converges to zero at least as fast as the sequence $\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y)$, for all $f \in F$ and $y \in \mathbb{N}(f)$. Namely, it suffices to consider the (ordinary) spline algorithm $\varphi_o = \{\varphi_o^n\}$. It was first defined in Section 2.5.1 for nonadaptive information. A natural generalization of this algorithm for adaptive $\mathbb{N}$ is as follows:

$$\varphi_o^n(y^n) \;=\; S(\mathbf{s}_o^n(y^n)), \qquad y^n \in \mathbb{N}^n(F),$$

where $\mathbf{s}_o^n(y^n)$ is the ordinary spline, i.e.,

1. $y^n \in \mathbb{N}^n(\mathbf{s}_o^n(y^n))$,
2. $\|\mathbf{s}_o^n(y^n)\|_F \;\leq\; \rho \cdot \inf\{\, \|f\|_F \;|\; y^n \in \mathbb{N}^n(f)\,\}$

$(\rho > 1)$. Proceeding as in the proof of Theorem 2.7 we show that

$$\|S(f) - \varphi_o^n(y^n)\| \;\le\; c(f) \cdot \mathrm{diam}(\mathbb{N}_y^n) \;\le\; 2\,c(f) \cdot \mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y) \qquad (6.3)$$

where $c(f) = \max\{1, \frac{1+\rho}{2}\|f\|_F\}$.

Corollary 6.1 *For the algorithm φ_o, the error $\|S(f) - \varphi_o^n(y^n)\|$ converges to zero at least as fast as the nth worst case radii $\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y)$, for all $f \in F$ and $y \in \mathbb{N}(f)$.* □

As we know, the ordinary spline algorithm is usually nonlinear. It would be nice to have a linear algorithm for which Corollary 6.1 also holds. More precisely, we shall say that an algorithm $\varphi = \{\varphi^n\}$ is linear iff the mappings $\varphi^n : \mathbb{R}^n \to G$ are linear for all $n \ge 1$.

Lemma 6.1 *Let information $\mathbb{N}$ be nonadaptive. Suppose that there exists a linear algorithm φ_{lin} and $M \ge 1$ such that for all n*

$$\mathrm{e}^{\mathrm{wor}}(\mathbb{N}^n, \varphi_{\mathrm{lin}}^n) \;\le\; M \cdot \mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}).$$

Then for all $f \in F$ and $y \in \mathbb{N}(f)$ we have

$$\|S(f) - \varphi_{\mathrm{lin}}^n(y^n)\| \;\le\; M \cdot \max\{\,1,\; \|f\|_F\,\} \cdot \mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}^n).$$

Proof The lemma is obviously true for $\|f\|_F \le 1$. For $\|f\|_F > 1$, we have that $y' = y/\|f\|_F$ is noisy information about $f' = f/\|f\|_F$, and $\|f'\|_F = 1$. Hence,

$$\|S(f) - \varphi_{\mathrm{lin}}^n(y^n)\| \;=\; \|f\|_F\,\|S(f') - \varphi_{\mathrm{lin}}^n((y')^n)\| \;\le\; M\,\|f\|_F\,\mathrm{rad}^{\mathrm{wor}}(\mathbb{N}^n),$$

as claimed. □

Lemma 6.1 can be applied, for instance, in the case when F is a Hilbert space and the noise is always bounded uniformly in a Hilbert norm. Due to Lemma 2.7, in this case the α–smoothing spline algorithm φ_α (with any $\alpha \in (0,1)$) is almost optimal, and $M = \max\{\alpha^{-1/2}, (1-\alpha)^{-1/2}\}$.

Are there algorithms for which convergence is better than $\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y)$? We now give two examples showing that the answer depends on a particular problem.

Example 6.2 Let $\dim F \geq 1$, S be a continuous embedding, $S(f) = f$, and let $\mathbb{N}$ be the zero information, $N = [0, 0, \ldots]$. Then $\mathrm{rad}_n^{\mathrm{wor}} = \|S\|_F > 0$, $\forall n$. Since we always have $y = 0$, any algorithm φ is just a sequence $\{\varphi^n\}$ of elements in G. Hence, the error can converge to zero for at most one element of F.

Example 6.3 Let $F = G$ be the space of Lipschitz functions $f : [0, 1] \to \mathbb{R}$ with the supremum norm. Let S be the identity. Suppose the nth approximation to f is based on data $y^n = [y_1, \ldots, y_n]$ where $y_i = f(t_i) + x_i$ and $|x_i| \leq \delta_i$. It is easy to see that for any choice of the points t_n, precisions δ_n and mappings φ^n, the nth worst case error, $\mathrm{e}^{\mathrm{wor}}(\mathbb{N}^n, \varphi^n)$, is at least 1.

Now, let t_n and δ_n be given as

$$t_n \;=\; (2i+1)/2^{k+1}, \qquad \delta_n \;=\; 2^{-(k+1)}, \qquad n \geq 1,$$

where $n = 2^k + i$, $0 \leq i \leq 2^k - 1$. Let the nth approximation $\varphi^n(y^n)$ be given by the linear spline interpolating data y^n. Then for any $f \in F$ and $y^n \in \mathbb{N}^n(f)$ we have

$$\|f - \varphi^n(y^n)\|_\infty \;\leq\; M/n,$$

where $M = M(f)$ depends only on the Lipschitz constant for f. Hence, we have at least linear convergence of the successive approximations to f, while the radii $\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N})$ do not converge at all. □

In the last example, F is *not* a Banach space. The completeness of the space F and continuity of information turn out to be crucial assumptions. That is, assume additionally that

- S is a Banach space, and
- the linear functionals $L_i(\cdot; y_1, \ldots, y_{i-1})$ forming information are continuous for all $i \geq 1$ and $y \in \mathbb{R}^\infty$.

Then $\{\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y)\}$ establishes a lower bound on the speed of convergence of the error $\|S(f) - \varphi^n(y^n)\|$. Namely, we have the following theorem.

Theorem 6.1 *Let $\mathbb{N}$ and φ be an arbitrary information operator and algorithm. Let $\tau(y)$, $y \in \mathbb{R}^\infty$, be the family of infinite positive sequences such that*

$$\tau(y) \;=\; [\tau_1, \tau_2(y^1), \ldots, \tau_n(y^{n-1}), \ldots]$$

and $\lim_{n\to\infty} \tau_n(y) \,=\, 0$*. Then the set*

$$A \;=\; \left\{ f \in F \;\middle|\; \quad \forall y \in \mathbb{N}(f), \quad \limsup_{n\to\infty} \frac{\|S(f) - \varphi^n(y^n)\|}{\tau_n(y)\,\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y)} < +\infty \right\}$$

is boundary, i.e., it does not contain any ball in F*. (Here* $0/0 = \infty$*.)*

Proof Suppose to the contrary that A contains a closed ball B with radius r, $0 < r \le 1$. We shall show that then it is possible to find an element $f^* \in B$ and information y^* about f^* such that

$$\limsup_{n\to\infty} \frac{\|S(f^*) - \varphi((y^*)^n)\|}{\tau_n(y^*)\,\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_{y^*})} \;=\; +\infty. \tag{6.4}$$

1. We first construct by induction a sequence $\{f_k\}_{k\ge1} \subset B$, a sequence of integers $0 = n_0 < n_1 < \cdots$, and $y_k \in \mathbb{N}_{y_k}(f_k)$, $k \ge 1$, which satisfy the following conditions:

$$\begin{aligned} y_{k+1}^{n_k} &= y_k^{n_k}, \\ \|y_k^{n_k} - N_{y_k}^{n_k}(f_{k+1})\|_{\Delta_{y_k}^{n_k}} &\le \sum_{j=1}^{k} 2^{-j}, \\ \|f_{k+1} - f_k\|_F &\le (r/2)^k, \end{aligned}$$

for all $k \ge 1$.

Let f_1 be the center of B. Suppose that for some $k \ge 1$ we have constructed $f_1, \ldots, f_k$, $n_0 < \cdots < n_{k-1}$, and $y_1, \ldots, y_{k-1}$. We select $y_k = [y_{k,1}, y_{k,2}, \ldots]$ in such a way that $y_{k,i} = y_{k-1,i}$ for $1 \le i \le n_{k-1}$, and

$$\|y_k^{i+1} - N_{y_k}^{i+1}(f_k)\|_{\Delta_{y_k}^{i+1}} \;=\; \|y_k^i - N_{y_k}^i(f_k)\|_{\Delta_{y_k}^i}, \quad i \ge n_{k-1} + 1.$$

Note that in view of (6.1) this selection is possible. (For $k = 1$ we set $y_{1,i} = L_i(f_k; y_1^{i-1})$ so that $y_1 = N_{y_1}(f_1)$.) Clearly, y_k is noisy information about f_k.

Now, we choose $r_k > 0$ such that for $\|f - f_k\| \le r_k$ is $\|S(f) - S(f_k)\| \le 1/3\,\|S(f_k) - S(f_{k-1})\|$. (For $k = 1$ we choose r_1 to be such that $\|S(f) - S(f_1)\| \le 1/3$ for $\|f - f_1\| \le r_1$.) Since $f_k \in B$, there is an integer $n_k > n_{k-1}$ for which

$$\sqrt{\tau_{n_k}(y_k)} \;\le\; \min\{r_k,\, (r/2)^k\}$$

and

$$\frac{\|S(f_k) - \varphi^{n_k}(y_k^{n_k})\|}{\tau_{n_k}(y_k)\,\mathrm{rad}^{\mathrm{wor}}_{n_k}(\mathbb{N}_{y_k})} \;\le\; \frac{1}{10\sqrt{\tau_{n_k}(y_k)}}.$$

Due to (6.2) there exists $h_k \in F$ such that

(i) $\|N_{y_k}^{n_k}(h_k)\|_{\Delta_{y_k}^{n_k}} \;\le\; \sqrt{\tau_{n_k}(y_k)}$,

(ii) $\|h_k\|_F \;\le\; \sqrt{\tau_{n_k}(y_k)}$, and

(iii) $\|S(h_k)\| \;\ge\; 1/4\sqrt{\tau_{n_k}(y_k)}\,\mathrm{rad}^{\mathrm{wor}}_{n_k}(\mathbb{N}_{y_k})$.

We now set $f_{k+1} = f_k + h_k$. Observe that for $k = 1$ we have

$$\|y_1^{n_1} - N_{y_1}^{n_1}(f_2)\|_{\Delta_{y_1}^{n_1}} \;=\; \|N_{y_1}^{n_1}(h_1)\|_{\Delta_{y_1}^{n_1}} \;\le\; \sqrt{\tau_{n_1}(y_1)} \;\le 1/2,$$

while for $k \ge 2$ we have

$$\begin{aligned}\|y_k^{n_k} - N_{y_k}^{n_k}(f_{k+1})\|_{\Delta_{y_k}^{n_k}} &\le \|y_k^{n_k} - N_{y_k}^{n_k}(f_k)\|_{\Delta_{y_k}^{n_k}} + \|N_{y_k}^{n_k}(h_k)\|_{\Delta_{y_k}^{n_k}} \\ &\le \|y_{k-1}^{n_{k-1}} - N_{y_{k-1}}^{n_{k-1}}(f_k)\|_{\Delta_{y_{k-1}}^{n_{k-1}}} + \sqrt{\tau_{n_k}(y_{n_k})} \\ &\le \sum_{j=1}^{k} 2^{-j}.\end{aligned}$$

Furthermore, $\|f_{k+1} - f_1\|_F \le \sum_{j=1}^k \|f_{j+1} - f_j\|_F \le r$, so that $f_{k+1} \in B$. This completes the construction.

2. The sequence $\{f_k\}$ satisfies the Couchy condition. Indeed, for any $m > k$ we have

$$\|f_m - f_k\|_F \;\le\; \sum_{j=k}^{m-1} \|f_{j+1} - f_j\|_F \;\le\; \sum_{j=1}^{m-1} (r/2)^j \;\le\; 2\,(r/2)^k.$$

Hence, there exists the limit $f^* = \lim_{k\to\infty} f_k \in B$.

We now show a property of f^*. Since $\|f_{k+1} - f_k\|_F \le r_k$, we have $\|S(f_{k+1}) - S(f_k)\| \le 1/3 \cdot \|S(f_k) - S(f_{k-1})\|$, $k \ge 2$. This gives for $m > k$

$$\begin{aligned}\|S(f_m) - S(f_k)\| &\ge \|S(f_{k+1}) - S(f_k)\| - \sum_{j=k+1}^{m-1} \|S(f_{j+1}) - S(f_j)\| \\ &\ge \left(1 - \sum_{j=k+1}^{m-1} (1/3)^{j-k}\right) \|S(f_{k+1}) - S(f_k)\| \\ &\ge 1/2 \cdot \|S(f_{k+1}) - S(f_k)\|.\end{aligned}$$

By letting $m \to +\infty$ we get

$$\|S(f^*) - S(f_k)\| \;\geq\; 1/2 \cdot \|S(f_{k+1}) - S(f_k)\|, \qquad k \geq 1.$$

3. Define now the sequence $y^* \in \mathbb{R}^\infty$ as

$$y^* \;=\; [y_{1,1}, \dots, y_{1,n_1}, \dots, y_{k,n_{k-1}+1}, \dots, y_{k,n_k}, \dots].$$

That is, $(y^*)^{n_k} = y_k^{n_k}$, $k \geq 1$, where $y_k \in \mathbb{R}^\infty$ are constructed in **1**. We shall show that y^* is noisy information about f^*. Indeed, for $m > k$ we have

$$\begin{aligned}
& \|(y^*)^{n_k} - N_{y^*}^{n_k}(f_m)\|_{\Delta_{y^*}^{n_k}} \\
\leq\;\; & \|(y^*)^{n_k} - N_{y^*}^{n_k}(f_{k+1})\|_{\Delta_{y^*}^{n_k}} \;+\; \sum_{j=k+1}^{m-1} \|N_{y^*}^{n_k}(f_{j+1} - f_j)\|_{\Delta_{y^*}^{n_k}} \\
\leq\;\; & \sum_{j=1}^{k} 2^{-j} \;+\; \sum_{j=k+1}^{m-1} \|N_{y^*}^{n_j}(h_j)\|_{\Delta_{y^*}^{n_j}} \;\leq\; \sum_{j=1}^{m-1} 2^{-j} \;\leq\; 1.
\end{aligned}$$

Letting $m \to +\infty$ and using continuity of $N_{y^*}^{n_k}$, we find that $\|(y^*)^{n_k} - N_{y^*}^{n_k}(f^*)\|_{\Delta_{y^*}^{n_k}} \leq 1$. This in turn yields $y^* \in \mathbb{N}_{y^*}^{n}(f^*)$ for all $n \geq 1$, i.e., y^* is noisy information about f^*.

Finally, for $k \geq 1$ we obtain

$$\begin{aligned}
& \|S(f^*) - \varphi^{n_k}((y^*)^{n_k}\| \\
\geq\;\; & \|S(f^*) - S(f_k)\| \;-\; \|S(f_k) - \varphi^{n_k}((y^*)^{n_k})\| \\
\geq\;\; & 1/2 \cdot \|S(f_{k+1}) - S(f_k)\| \;-\; 1/10 \cdot \sqrt{\tau_{n_k}(y^*)}\, \mathrm{rad}_{n_k}^{\mathrm{wor}}(\mathbb{N}_{y^*}) \\
\geq\;\; & 1/40 \cdot \sqrt{\tau_{n_k}(y^*)}\, \mathrm{rad}_{n_k}^{\mathrm{wor}}(\mathbb{N}_{y^*}),
\end{aligned}$$

which implies (6.4) and contradicts $f^* \in B$. The proof is complete. □

Observe that the sequences $\tau(y)$, $y \in \mathbb{R}^\infty$, can be selected in such a way that they converge to zero arbitrarily slowly. This means that the speed of convergence of the radii $\{\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_y)\}$ cannot be essentially beaten by any algorithm φ. In this sense, the optimal convergence rate is given by that of the radii, and the (ordinary) spline algorithm φ_o is optimal.

We now give an additional example which should explain the role of $\tau(y)$.

Example 6.4 Let $F = G$ be a separable, infinite dimensional Hilbert space with the orthonormal basis $\{\xi_i\}_{i\geq 1}$. Let S be given as $S\xi_i = \lambda_i\xi_i$, $i \geq 1$, where $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$. Finally, let information $\mathbb{N}$ be exact with $N = [\langle\cdot,\xi_1\rangle_F, \langle\cdot,\xi_2\rangle_F, \ldots]$. In this case we have $\mathrm{rad}_n^{\mathrm{wor}}(N) = |\lambda_{n+1}|$. On the other hand, for the algorithm $\varphi = \{\varphi^n\}$ where $\varphi^n(y^n) = \sum_{i=1}^n y_i\lambda_i\xi_i$, we have

$$\|S(f) - \varphi^n(y^n)\| \;=\; \sqrt{\sum_{i=n+1}^{\infty} \langle f,\xi_i\rangle_F^2 \lambda_i^2} \;\leq\; |\lambda_{n+1}| \sqrt{\sum_{i=n+1}^{\infty} \langle f,\xi_i\rangle_F^2}$$

which converges to zero faster than $|\lambda_{n+1}|$. However, due to Theorem 6.1, for a dense set of f the ratio $\|S(f)-\varphi^n(y^n)\|/|\lambda_{n+1}|$ tends to zero arbitrarily slowly.

6.2.3 Optimal information

Let $\Lambda \subset F^*$ be a given class of continuous functionals. Let $\mathcal{N}$ be the class of nonadaptive (exact) information operators $N = [L_1, L_2, \ldots]$ with the functionals $L_i \in \Lambda$, $i \geq 1$. Suppose the precision sequence is fixed, $\Delta = [\delta_1, \delta_2, \ldots]$, and we want to select information $N \in \mathcal{N}$ in such a way as to maximize the speed of convergence of the error $\|S(f) - \varphi_o^n(y^n)\|$ where $y \in \mathbb{N}(f)$ and $\mathbb{N} = \{N, \Delta\}$.

Due to Theorem 6.1, the error cannot tend to zero essentially faster than the sequence of the nth minimal worst case radii $\{\mathrm{r}_n^{\mathrm{wor}}(\Delta)\}$ where

$$\mathrm{r}_n^{\mathrm{wor}}(\Delta) \;=\; \inf_{N\in\mathcal{N}} \mathrm{rad}_n^{\mathrm{wor}}(N^n, \Delta^n), \qquad n \geq 1$$

(compare with the definition in Section 2.8). (Actually, this kind of convergence connot be beaten even when the information functionals are selected adaptively.) We shall show that it is often possible to construct information $N \in \mathcal{N}$ for which that convergence is achieved.

We assume that the extended norms $\|\cdot\|_{\Delta^n}$ satisfy the following condition. Let $n \geq 1$, $[\delta_1, \ldots, \delta_n] \in \mathbb{R}^n$ be a precision vector, and let $\{p_i\}_{i=1}^n$ be a permutation of $\{1, 2, \ldots, n\}$. Then

$$\|[x_1, \ldots, x_n]\|_{[\delta_1,\ldots,\delta_n]} \;=\; \|[x_{p_1}, \ldots, x_{p_n}]\|_{[\delta_{p_1},\ldots,\delta_{p_n}]}, \quad \forall [x_1, \ldots, x_n] \in \mathbb{R}^n. \tag{6.5}$$

This condition expresses the property that the power of information does not depend on the order of peforming n nonadaptive observations. Indeed, for two information operators $\mathbb{N}_1 = \{[L_1,\ldots,L_n],[\delta_1,\ldots,\delta_n]\}$ and $\mathbb{N}_2 = \{[L_{p_1},\ldots,L_{p_n}],[\delta_{p_1},\ldots,\delta_{p_n}]\}$ we have $[y_1,\ldots,y_n] \in \mathbb{N}_1(f)$ if and only if $[y_{p_1},\ldots,y_{p_n}] \in \mathbb{N}_2(f)$. Clearly, (6.5) holds, for instance, for the weighted sup or Euclidean norms.

Let $\eta > 1$. For any $n \geq 1$, let information $N_n \in \mathcal{N}$ be such that

$$\mathrm{rad}_n^{\mathrm{wor}}(N_n,\Delta) \;\leq\; \eta \cdot \mathrm{r}_n^{\mathrm{wor}}(\Delta).$$

Define

$$N_\Delta \;=\; [\,N_1^1, N_2^2, N_4^4, \ldots, N_{2^k}^{2^k}, \ldots\,]$$

where, as always, N_n^n denotes the first n functionals of N_n. The following theorem yields in many cases optimality of N_Δ.

Theorem 6.2 *Suppose the precision sequence $\Delta = [\delta_1,\delta_2,\ldots]$ satisfies*

$$\delta_1 \geq \delta_2 \geq \delta_3 \geq \cdots \geq 0. \tag{6.6}$$

Then for information $\mathbb{N}_\Delta = \{N_\Delta,\Delta\}$ and the ordinary spline algorithm φ_o we have

$$\|S(f) - \varphi_o^n(y^n)\| \;\leq\; K(f)\cdot \mathrm{r}^{\mathrm{wor}}_{\lceil\frac{n+1}{4}\rceil}(\Delta), \qquad f \in F,\; y \in \mathbb{N}_\Delta(f),$$

where $K(f) = \eta \cdot \max\{2, (1+\rho)\,\|f\|_F\}$.

Proof For $n \geq 1$, let $k = k(n)$ be the largest integer satisfying $n \geq \sum_{i=0}^k 2^i = 2^{k+1}-1$. Then all the functionals of $N_{2^k}^{2^k}$ are contained in N_Δ^n and, due to (6.6), these functionals are observed more precisely using information $\mathbb{N}_\Delta$ than $\{N_{2^k}^{2^k}, \Delta^{2^k}\}$. This, (6.1) and (6.5) yield

$$\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_\Delta) \;\leq\; \mathrm{rad}_{2^k}^{\mathrm{wor}}(\{N_{2^k}^{2^k}, \Delta^{2^k}\}) \;\leq\; \eta\cdot \mathrm{r}_{2^k}^{\mathrm{wor}}(\Delta).$$

Using (6.3), for any $f \in F$ and $y \in \mathbb{N}_\Delta(f)$ we obtain

$$\begin{aligned}\|S(f) - \varphi_o^n(y^n)\| \;&\leq\; \max\{2,(1+\rho)\,\|f\|_F\}\,\mathrm{rad}_n^{\mathrm{wor}}(\mathbb{N}_\Delta)\\ &\leq\; \eta\,\max\{2,(1+\rho)\,\{f\|_F\}\,\mathrm{r}_{2^k}^{\mathrm{wor}}(\Delta).\end{aligned}$$

The theorem now follows from the fact that $2^k \geq \lceil (n+1)/4 \rceil$. □

We have already convinced ourselves that for many problems the nth minimal radius $\mathrm{r}_n^{\mathrm{wor}}(\Delta)$ behaves polynomially in $1/n$. In such cases the error

$\|S(f) - \varphi_o^n(y^n)\|$ achieves the optimal convergence rate which is $\mathrm{r}_n^{\mathrm{wor}}(\Delta)$, and information N_Δ can be called optimal.

Theorem 6.2 is of general character. It is clear that for some special problems it is possible to construct the optimal information more effectively, even when $\mathrm{r}^{\mathrm{wor}}_{\lceil\frac{n+1}{4}\rceil}(\Delta) \asymp \mathrm{r}_n^{\mathrm{wor}}(\Delta)$ does non hold. An example is given in E 6.4.

Notes and Remarks

NR 6.1 The first results which revealed relations between the asymptotic and worst case settings were obtained by Trojan [111] who analyzed the linear case with exact information. (See also Traub *et al.* [108, pp. 383–295].) His results were then generalized by Kacewicz [27] to the nonlinear case with exact information. Particular nonlinear problems of evaluating the global maximum and zero finding were studied in Plaskota [77] and Sikorski and Trojan [94], correspondingly.

The results on the asymptotic setting with noisy information were obtained by Kacewicz and Plaskota [33] [34] [35]. This section is based mainly on the last three papers.

NR 6.2 In this section we analyzed behavior of algorithms as the number of observations increases to infinity. It is also possible to study behavior of the cost of computing an ε–approximation, as $\varepsilon \to 0$. Clearly, we want this cost to grow as slowly as possible for all $f \in F$ and information y about f. The corresponding computational model would be as follows.

The approximations are obtained by executing a program $\mathcal{P}$. This time, however, the result of computations is a sequence $g_0, g_1, g_2, \ldots$ of approximations rather than a single approximation. That is, the execution consists (at least theoretically) of infinitely many steps. At each nth step a noisy value y_n of a functional $L_n(f; y_1, \ldots, y_{n-1})$ is observed and then the nth approximation $g_n = \varphi^n(y_1, \ldots, y_n)$ is computed. Obviously, such an infinite process usually requires infinitely many constants and variables. However, we assume that for any n, the nth approximation is obtained using a finite number of constants and variables, as well as a finite number of primitive operations. In other words, a program reduced to only n first steps is a program in the sense of the worst case setting of Section 2.9.1.

Let $\mathcal{P}$ be a program that realizes an algorithm φ using information $\mathbb{N}$. For $\varepsilon \geq 0$, let

$$m(\mathcal{P}; f, y)(\varepsilon) \;=\; \min\{\, k \geq 0 \mid \quad \forall i \geq k \quad \|S(f) - \varphi^i(y^i)\| \leq \varepsilon \,\} \qquad (6.7)$$

be the minimal number of steps for which all elements $g_m, g_{m+1}, g_{m+2}, \ldots$ are ε–approximations to $S(f)$. (If such a k does not exist then $m(\mathcal{P}; f, y) = +\infty$.) Then the cost of obtaining an ε–approximation using the program $\mathcal{P}$ is given as

$$\mathrm{cost}(\mathcal{P}; f, y)(\varepsilon) \;=\; \mathrm{cost}_m(\mathcal{P}, y), \quad f \in F,\ y \in \mathbb{N}(f),$$

where $m = m(\mathcal{P}; f, y)(\varepsilon)$ is defined by (6.7), and $\mathrm{cost}_m(\mathcal{P}; y)$ is the cost of performing m steps using the program $\mathcal{P}$ with information y. (If $m = +\infty$ then $\mathrm{cost}(\mathcal{P}; f, y)(\varepsilon) = +\infty$.)

A similar model of the (asymptotic) cost was studied by Kacewicz and Plaskota [34] [35]. They showed that, under some additional assumptions, the best behavior of $\mathrm{cost}(\mathcal{P}; f, y)(\varepsilon)$ is essentially determined by the worst case complexity $\mathrm{Comp}^{\mathrm{wor}}(\varepsilon)$. Hence, there are close relations between the asymptotic and worst case settings not only with respect to the error but also with respect the the cost of approximation.

NR 6.3 In the prvious remark we assumed that the computational process is infinite. It is clear that in practice the computations must be somewhere interrupted. The choice of an adequate rule for terminating calculations is an important practical problem. Obviously, (6.7) cannot serve as a computable stopping rule since $m(\mathcal{P}; f, y)(\varepsilon)$ explicitly depends on f which is unknown.

Suppose that we want to compute approximations using a program $\mathcal{P}$ which realizes a linear algorithm φ using nonadaptive information $\mathbb{N}$. Suppose also that we know some bound on the norm of f, say $\|f\|_F \le K$. Then, to obtain an ε–approximation it is enough to stop calculations after

$$m \;=\; \min\{\, i \mid \quad \mathrm{e}^{\mathrm{wor}}(\mathbb{N}^i, \varphi^i) \le \varepsilon \, \min\{1, 1/K\}\,\}$$

steps. In view of the results of Kacewicz and Plaskota [34] [35], this is best we can do. On the other hand, if we do not have any additional information about $\|f\|_F$, then any computable termination rule fails; see E 6.5.

Exercises

E 6.1 Let $\Delta \in \mathbb{R}^\infty$ be a given precision sequence. Show that

$$\|x\|_\Delta \;=\; \lim_{n\to\infty} \|x^n\|_{\Delta^n}, \qquad x \in \mathbb{R}^\infty,$$

is a well defined extended norm in $\mathbb{R}^\infty$, and that for all $n \ge 1$ we have

$$\|x^n\|_{\Delta^n} \;=\; \min_{z \in \mathbb{R}^\infty} \|[x^n, z]\|_\Delta, \qquad x^n \in \mathbb{R}^n.$$

E 6.2 Show that Corollary 6.1 holds also for the smoothing spline algorithm φ_∞ defined in Section 2.5.1.

E 6.3 Let $\mathbb{N}$ and φ be an arbitrary information and algorithm. Show that for the spline algorithm φ_o and $\tau_n(y)$ as in Theorem 6.1, the set

$$\left\{ f \in F \;\middle|\; \forall y \in \mathbb{N}(f), \quad \limsup_{n\to\infty} \frac{\|S(f) - \varphi^n(y^n)\|}{\tau_n(y) \cdot \|S(f) - \varphi_o^n(y^n)\|} < +\infty \right\}$$

does not contain any ball.

E 6.4 Suppose that the solution operator S is compact and it acts between separable Hilbert spaces, and that observations of all functionals with norm bouded by 1 are allowed. Let

$$N_0 \;=\; [\,\langle\cdot,\xi_1\rangle_F,\langle\cdot,\xi_2\rangle_F,\ldots],$$

where $\{\xi_i\}$ is the complete orthonormal basis of eigenelements of S^*S and the corresponding eigenelements satisfy $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Assuming exact observations, $\Delta = [0,0,0,\ldots]$, show that for the spline algorithm φ_{spl} we have

$$\|S(f)-\varphi^n_{\mathrm{spl}}(y^n)\| \;\leq\; \|f\|\cdot \mathrm{r}_n^{\mathrm{wor}}(0), \qquad f\in F,\ y=N_0(f).$$

That is, N_0 is the optimal information independently of the behavior of $\mathrm{r}_n^{\mathrm{wor}}(0) = \sqrt{\lambda_{n+1}}$.

E 6.5 (Kacewicz and Plaskota) Let $\mathcal{P}$ be a program realizing an algorithm φ using nonadaptive information $\mathbb{N} = \{N,\Delta\}$ such that $\mathrm{r}_n^{\mathrm{wor}}(\{N,0\}) > 0\ \forall n \geq 1$. Let $t_n : \mathbb{R}^\infty \to \{0,1\}$ be arbitrary termination functions. That is, for $f \in F$ and $y \in \mathbb{N}(f)$ calculations are terminated after

$$m(y) \;=\; \min\{\, i\geq 0 \mid \quad t_i(y_1,\ldots,t_i)=1\,\}$$

steps. Show that for any $\varepsilon > 0$ and $y \in \mathbb{N}(F)$, there exists $f \in F$ such that $y \in \mathbb{N}(f)$ and $\|S(f)-\varphi^{m(y)}(y)\| \;>\; \varepsilon$.

6.3 Asymptotic and average case settings

In this section, we assume that F is a separable Banach space equipped with a zero mean Gaussian measure μ. The solution operator S is continuous linear and it acts between F and a separable Hilbert space G. The information noise has random character.

More specifically, an (in general adaptive) information operator is given as a family $\mathbb{N} = \{N_y,\Sigma_y\}_{y\in\mathbb{R}^\infty}$ where

$$N_y \;=\; [\,L_1(\cdot), L_2(\cdot;y_1),\ldots,L_n(\cdot;y_1,\ldots,y_{n-1}),\ldots]$$

and

$$\Sigma_y \;=\; [\,\sigma_1^2,\sigma_2^2(y_1),\ldots,\sigma_n^2(y_1,\ldots,y_{n-1}),\ldots]$$

are infinite sequences of continuous linear functionals $L_i = L_i(\cdot;y_1,\ldots,y_{i-1})$ and nonnegative reals $\sigma_i^2 = \sigma_i^2(y_1,\ldots,y_{i-1})$, $i \geq 1$, respectively. For $f \in F$, $\pi_f = \mathbb{N}(f)$ is a probability measure representing the probability of occuring sequences $y = [y_1,y_2,\ldots] \in \mathbb{R}^\infty$ when gaining information about f. These measures are defined on the σ–field generated by the cylindrical sets of the form $B = A \times \mathbb{R}^\infty$ where A is a Borel set of $\mathbb{R}^n$ and $n \geq 1$. For any such B we have $\pi_f(B) = \pi_f^n(A)$ where π_f^n is the distribution of $[y_1, \ldots, y_n] \in \mathbb{R}^n$ corresponding to the first n observations. That is, π_f^n is defined as in Section 3.7.1 for the information operator $\mathbb{N}^n$ with

$$N_y^n \;=\; [\,L_1(\cdot), L_2(\cdot; y_1), \ldots, L_n(\cdot; y_1, \ldots, y_{n-1})\,]$$

and $\Delta_y^n = [\delta_1, \delta_2(y_1), \ldots, \delta_n(y_1, \ldots, y_{n-1})]$. Hence, for any Borel set $B \subset \mathbb{R}^\infty$ we have

$$\pi_f(B) \;=\; \lim_{n\to\infty} \pi_f^n(B^n) \tag{6.8}$$

where $B^n = \{\, y^n \in \mathbb{R}^n \mid y \in B\,\}$ is the projection of B onto $\mathbb{R}^n$.

Note that the measure π_f possesses the following property. For a given vector $(y_1, \ldots, y_{m-1}) \in \mathbb{R}^n$, the distribution of y_m is Gaussian with mean $L_m(f; y_1, \ldots, y_{m-1})$ and variance $\sigma_m^2(y_1, \ldots, y_{m-1})$.

Noisy information about f is any realization $y \in \mathbb{R}^\infty$ of the random variable distributed according to π_f.

6.3.1 Optimal algorithms

We now deal with the problem of optimal algorithm. Recall that in the average case setting the optimal algorithms φ_{opt} are obtained by applying S on the mean of the conditional distribution given information about f. Also, φ_{opt} can be interpreted as a smoothing spline algorithm, $\varphi_{\text{opt}} = \varphi_{\text{spl}}$. We now show that the same kind of algorithms can be successfully used in the asymptotic setting.

More specifically, for $y \in \mathbb{R}^\infty$, let the algorithm $\varphi_{\text{spl}} = \{\varphi_{\text{spl}}^n\}$ be given as $\varphi_{\text{spl}}^n(y^n) = S(m(y^n)\,)$ where

$$m(y^n) \;=\; \sum_{j=1}^n z_j^n\,(C_\mu(L_j(\cdot; y^{j-1})\,)\,),$$

z^n is the solution of

$$\left(\Sigma_y^n + G_{N_y}^n\right)\, z^n \;=\; y^n\,,$$

$\Sigma_y^n =$ diag $\{\sigma_1^2, \sigma_2^2(y^1), \ldots, \sigma_n^2(y^{n-1})\,\}$, and $G_{N_y}^n$ is the Gram matrix, $G_{N_y}^n = \{L_i(\cdot; y^{i-1}), L_j(\cdot; y^{j-1})\}_{i,j=1}^n$. Denoting by H the Hilbert space associated with μ, $m(y^n)$ can be equivalently defined as the minimizer in H of the functional

$$\|f\|_H^2 \;+\; \sum_{j=1}^n \frac{1}{\sigma_j^2(y^{j-1})}\,(y_j - L_j(f; y^{j-1})\,)^2$$

(compare with Sections 3.6 and 3.7.2).

In what follows, we shall use the joint distribution $\tilde{\mu}$ on the space $F \times \mathbb{R}^\infty$. It represents the probability of occurring $f \in F$ and information y about f, and is generated by the measure μ and the distributions π_f. Namely, for measurable sets $A \subset F$ and $B \subset \mathbb{R}^\infty$, we have

$$\tilde{\mu}(A \times B) \;=\; \int_A \pi_f(B)\,\mu(df).$$

Observe that in view of (6.8) we can also write

$$\tilde{\mu}(A \times B) \;=\; \lim_{n\to\infty} \tilde{\mu}^n(A \times B^n)$$

where $\tilde{\mu}^n(A \times B^n) = \int_A \pi_f^n(B^n)\,\mu(df)$ is the joint probability on $F \times \mathbb{R}^n$ or, in other words, it is the projection of $\tilde{\mu}$ onto $F \times \mathbb{R}^n$. Obviously, $m(y^n)$ is the mean element of the conditional distribution $\mu_2(\cdot|y^n)$ on F. Hence, φ_{spl}^n minimizes the average error over $\tilde{\mu}^n$.

The algorithm φ_{spl} is optimal in the following sense.

Theorem 6.3 *For any algorithm $\varphi = \{\varphi^n\}$, its error almost nowhere tends to zero faster than the error of φ_{spl}. That is, the set*

$$A \;=\; \left(\left\{\, (f,y) \in F \times \mathbb{R}^\infty \;\middle|\; \lim_{n\to\infty} \frac{\|S(f) - \varphi^n(y^n)\|}{\|S(f) - \varphi_{\mathrm{spl}}^n(y^n)\|} \;=\; 0 \right\}\right)$$

is of $\tilde{\mu}$–measure zero. (By convention, $0/0 = 1$.)

The proof of this theorem is based on the following fact.

Lemma 6.2 *Let ω be a Gaussian measure on G with mean m_ω. Then for any $g_0 \in G$ and $q \in (0,1)$ we have*

$$\omega\left(\{\, g \in G \mid \quad \|g - g_0\| < q\,\|g - m_\omega\|\,\}\right) \;\le\; \beta\,\frac{q}{1-q} \tag{6.9}$$

where $\beta = \sqrt{2/(\pi\, e)}$.

Proof Suppose first that $m_\omega = 0$. Let $g = c\,g_0 + g_1$ where $g_1 \perp g_0$. Then $\|g - g_0\| < q\,\|g\|$ is equivalent to

$$(1-c)^2 c^2 \|g_0\|^2 \;+\; \|g_1\|^2 \;<\; q^2\,(\,c^2\|g_0\|^2 + \|g_1\|^2)$$

which, in particular, implies $(1-c)^2 < c^2q^2$ and $c \in ((1+q)^{-1},(1-q)^{-1})$. This in turn means that

$$\frac{\|g_0\|^2}{1+q} \;<\; \langle\, g, g_0 \,\rangle \;<\; \frac{\|g_0\|^2}{1-q}. \tag{6.10}$$

Let B be the set of all g satisfying (6.10). As $g \to \langle g, g_0\rangle$ is the zero mean Gaussian random variable with variance $\lambda = \langle C_\omega g_0, g_0\rangle$ (where $C_\omega : G \to G$ is the correlation operator of μ), for $\lambda = 0$ we have $\omega(B) = 0$, while for $\lambda > 0$

$$\omega(B) \;=\; \frac{1}{\sqrt{2\pi\lambda}} \int_{\|g_0\|^2/(1+q)}^{\|g_0\|^2/(1-q)} e^{-x^2/(2\lambda)}\,dx \;=\; \frac{1}{\sqrt{2\pi}} \int_{a/(1+q)}^{a/(1-q)} e^{-x^2/2}\,dx,$$

where $a = \|g_0\|^2\lambda^{-1/2}$. Hence,

$$\begin{aligned} \omega(B) \;&\le\; \frac{a}{\sqrt{2\pi}}\left(\frac{1}{1-q} - \frac{1}{1+q}\right) \exp\left\{-\frac{1}{2}\left(\frac{a}{1+q}\right)^2\right\} \\ &=\; \sqrt{\frac{2}{\pi}}\,\frac{q}{1-q}\left(\frac{a}{1+q}\right) \exp\left\{-\frac{1}{2}\left(\frac{a}{1+q}\right)^2\right\}. \end{aligned}$$

To get (6.9), it suffices to observe that the maximal value of $x\,e^{-x^2/2}$ is $e^{-1/2}$.

If $m_\omega \neq 0$ then we let $\omega(\cdot) = \omega_1(\cdot - m_\omega)$ and $\bar g_0 = g_0 - m_\omega$. Then zero is the mean element of ω and $\omega(A) = \omega_1(\{g \mid \|g - \bar g_0\| < q\|g\|\,\})$.

Proof of Theorem 6.3 Choose $q \in (0,1)$. For $n \ge 1$, define the sets

$$A_n \;=\; \Big\{\,(f,y) \in F \times \mathbb{R}^\infty \mid \quad \|S(f) - \varphi^n(y^n)\| < q \cdot \|S(f) - \varphi^n_{\mathrm{spl}}(y^n)\|\,\Big\}.$$

Observe that if $(f,y) \in A$ then for all sufficiently large n we have $(f,y) \in A_n$, and consequently

$$A \;\subset\; \bigcup_{j=1}^{\infty}\,\bigcap_{n=j}^{\infty} A_n.$$

Hence,

$$\tilde\mu(A) \;\le\; \lim_{j\to\infty}\,\tilde\mu\left(\bigcap_{n=j}^{\infty} A_n\right) \;\le\; \limsup_{n\to\infty}\,\tilde\mu(A_n). \tag{6.11}$$

We now estimate the measure $\tilde\mu$ of A_n. Let

$$A_n^n \;=\; \{\,(f,y^n) \in F \times \mathbb{R}^n \mid \quad (f,y) \in A_n\,\}.$$

Then $\tilde{\mu}(A_n) = \tilde{\mu}^n(A_n^n)$. Using decomposition of $\tilde{\mu}^n$ with respect to the nth information y^n we get

$$\tilde{\mu}^n(A_n^n) \;=\; \int_{\mathbb{R}^n} \mu_2(A_n^n|y^n\,)\,\mu_1(dy^n),$$

where μ_1 is the a priori distribution of y^n, $\mu_2(\cdot|y^n)$ is the conditional distribution on F given y^n, and

$$A_n^n(y^n) \;=\; \{\, f \in F \mid \quad (f, y^n) \in A_n^n\,\}.$$

Due to Theorem 3.8, the measures $\mu_2(\cdot|y^n)$ are Gaussian. Denote $\nu_2(\cdot|y^n) = \mu_2(S^{-1}(\cdot)|y^n)$ and

$$B_n^n(y^n) \;=\; S(A_n^n(y^n)\,) \;=\; \{\, g \in G \mid \quad \|g - \varphi^n(y^n)\| < q\,\|g - \varphi^n_{\mathrm{spl}}(y^n)\|\,\}.$$

Since $\nu_2(\cdot|y^n)$ is also Gaussian and its mean element equals $\varphi^n_{\mathrm{spl}}(y^n)$, Lemma 6.2 gives $\nu_2(B_n^n(y^n)) \le \beta\, q/(1-q)$, and consequently

$$\tilde{\mu}^n(A_n^n) \;=\; \int_{\mathbb{R}^n} \nu_2(B_n^n(y^n))\,\mu_1(dy^n) \;\le\; \beta\,\frac{q}{1-q}.$$

Thus the set A_n has the $\tilde{\mu}$ measure at most $\beta\, q/(1-q)$. In view of (6.11), this means that also $\tilde{\mu}(A) \le \beta\, q/(1-q)$. Since q can be arbitrarily close to 0, we finally obtain $\tilde{\mu}(A) = 0$, as claimed. □

The algorithm minimizing the nth average errors turned out to be optimal also in the asymptotic setting. There is no algorithm φ such that the successive approximations $\varphi^n(y^n)$ converge to the solution $S(f)$ faster than $\varphi^n_{\mathrm{spl}}(y^n)$.

6.3.2 Convergence rate

Theorem 6.3 does not say anything about the rate of convergence of $\varphi^n_{\mathrm{spl}}(y^n)$ to $S(f)$. For deterministic noise, the best behavior of the error can essentially be compared to that of the worst case radii. It turns out that for random noise a similar role plays the sequence of nth average radii. The nth average radius of nonadaptive information $\mathbb{N}_y$ ($y \in \mathbb{R}^\infty$) is given as

$$\mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y) \;=\; \sqrt{\int_{\mathbb{R}^n} r^2(\mu_2(\cdot|y^n)\,)}\,,$$

where $\mu_2(\cdot|y^n)$ is the conditional distribution on F given y^n, and $r^2(\cdot)$ is the squared radius of a measure, see Section 3.2.

Before we state theorems about the rate of convergence of φ_{spl}, we first need some estimations for Gaussian measures of balls in G. In what follows, we denote by $B_r(a)$ the ball of radius r and centered at a.

Lemma 6.3 *Let ω be a zero mean Gaussian measure on G. Then for any $r \geq 0$ and $a \in G$ we have*

$$\omega(B_r(0)\,) \;\geq\; \omega(B_r(a)\,).$$

Proof Assume first that $G = \mathbb{R}^n$ and the coordinates g_i, $1 \leq i \leq n$, are independent random variables. For $n = 1$ the lemma is obvious. Let $n \geq 2$. Let ω^{n-1} be the joint distribution of $g^{n-1} = (g_1, \ldots, g_{n-1})$ and ω_n be the distribution of g_n. Then

$$\begin{aligned} & \omega(B_r(a)\,) \\ = \;& \int_{\mathbb{R}^n} \omega_n \left(\left\{ g_n \;\Big|\; |g_n - a_n| \leq \sqrt{r^2 - \|g^{n-1} - a^{n-1}\|^2} \right\} \right) \omega^{n-1}(dg^{n-1}) \\ \leq \;& \int_{\mathbb{R}^n} \omega_n \left(\left\{ g_n \;\Big|\; |g_n| \leq \sqrt{r^2 - \|g^{n-1} - a^{n-1}\|^2} \right\} \right) \omega^{n-1}(dg^{n-1}) \\ = \;& \omega(B_r(a^{n-1}, 0)\,)\,. \end{aligned}$$

Proceeding in this way with successive coordinates we obtain

$$\omega(B_r(a^{n-1},0)\,) \;\leq\; \omega(B_r(a^{n-2},0,0)\,) \;\leq\; \cdots \;\leq\; \omega(B_r(\underbrace{0,\ldots,0}_{n})\,),$$

and consequently $\omega(B_r(a)\,) \leq \omega(B_r(0)\,)$.

Consider now the general case. Let $\{\xi_j\}_{j\geq 1}$ be the complete orthonormal system of eigenelements of C_ω. Then $g_j = \langle g, \xi_j \rangle$ are independent zero mean Gaussian random variables and $B_r(a) = \{\, g \in G \,|\, \sum_j (g_j - a_j)^2 \leq r^2 \,\}$. Denoting by ω^n the joint distribution of $(g_1, \ldots, g_n)$ and by $B_r^n(a^n)$ the ball in $\mathbb{R}^n$ with center $a^n = (a_1, \ldots, a_n)$ and radius r, we have

$$\omega(B_r(a)\,) \;=\; \lim_{n\to\infty} \omega^n(B_r^n(a^n)\,) \;\leq\; \lim_{n\to\infty} \omega^n(B_r^n(0)\,) \;=\; \omega(B_r(0)\,)\,,$$

as claimed.

Lemma 6.4 *Let ω be a Gaussian measure on G and let C_ω be its correlation operator. Then*

$$\nu(B_r(a)) \;\le\; \frac{4}{3}\,\psi\left(\frac{2\,r}{\sqrt{\mathrm{trace}(C_\omega)}}\right)$$

where $\psi(x) = \sqrt{2/\pi}\,\int_0^x e^{-t^2/2}dt$.

Proof We can assume without loss of generality that the mean element of ω is zero since we always can shift the measure towards the origin. In view of Lemma 6.3, we can also assume that the ball is centered at zero. In this case we write, for brevity, B_r instead of $B_r(0)$.

Let $d = \dim G \le +\infty$. Let $\{\xi_j\}$ be the complete orthonormal system of eigenelements of C_ω, $C_\omega \xi_j = \lambda_j \xi_j$. Then the random variables $g_j = \langle g, \xi_j\rangle$ are independent and $g_j \sim \mathcal{N}(0, \lambda_j)$.

Let t_j be independent random variables which take -1 and $+1$ each with probability $1/2$, and let $t = (t_j)_{j=1}^d$. Denote by p the joint probability on $T = \{-1, +1\}^d$, and by $\tilde\omega$ the joint probability on $T \times G$. Then

$$\begin{aligned}
&\tilde\omega\left(\left\{ (t,g) \in T \times G \;\Big|\; \Big|\sum_{j=1}^d t_j\, g_j\Big| \;\le\; 2\,r \right\}\right) \qquad (6.12)\\
\ge\;& \int_{B_r} p\left(\left\{ t \in T \;\Big|\; \Big|\sum_{j=1}^d t_j\, g_j\Big| \le 2\,r \right\}\right) \omega(dg)\\
\ge\;& \gamma \cdot \omega(B_r)
\end{aligned}$$

where

$$\gamma \;=\; \inf p\left(\left\{ t \in T \;\Big|\; \Big|\sum_{j=1}^d t_j\, c_j\Big| \;\le\; 2 \right\}\right),$$

the infimum taken over all c_j with $\sum_{j=1}^d c_j^2 \le 1$.

On the other hand, $\{t_j g_j\}$ are independent random variables and $t_j g_j \sim \mathcal{N}(0, \lambda_j)$, which implies that $\sum_{j=1}^d t_j g_j \sim \mathcal{N}(0, \lambda)$ where $\lambda = \mathrm{trace}(C_\omega)$. Hence, (6.12) equals

$$\frac{1}{\sqrt{2\pi\lambda}} \int_{-2r}^{2r} e^{-t^2/(2\lambda)}\, dt \;=\; \psi\left(\frac{2r}{\sqrt{\lambda}}\right),$$

and consequently

$$\omega(B_r) \;\le\; \frac{1}{\gamma}\,\psi\left(\frac{2r}{\sqrt{\mathrm{trace}(C_\omega)}}\right).$$

We now estimate γ^{-1}. Since for any $c=(c_1,c_2,\ldots)$ the random variable $\sum_{j=1}^d t_j c_j$ has mean zero and variance $\sum_{j=1}^d c_j^2$, we can use the well known Chebyshev's inequality to get

$$p\left(\left\{ t\in T \;\Big|\; \Big|\sum_{j=1}^d t_j c_j\Big| > 2 \right\}\right) \;\le\; \frac{1}{4}\cdot\sum_{j=1}^d c_j^2 \;\le\; \frac{1}{4}\,.$$

Hence, $\gamma\ge 1-1/4$ and $\gamma^{-1}\le 4/3$. The proof is complete. □

We are now ready to show that the error of any algorithm (and particularly the error of φ_{spl}) cannot converge to zero faster than $\{\mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y)\}$.

Theorem 6.4 *For any algorithm φ the set*

$$A_1 \;=\; \left\{ (f,y)\in F\times\mathbb{R}^\infty \;\Big|\; \lim_{n\to\infty}\frac{\|S(f)-\varphi_{\mathrm{spl}}^n(y^n)\|}{\mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y)} \;=\; 0 \right\}$$

has the $\tilde\mu$–measure zero.

Proof We choose $q\in(0,1)$ and define

$$A_{1,n} \;=\; \{\,(f,y)\in F\times\mathbb{R}^\infty \mid \|S(f)-\varphi_{\mathrm{spl}}^n(y^n)\| < q\cdot \mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y)\,\}$$

and $B_{1,n}=\{\,(f,y^n)\in F\times\mathbb{R}^n \mid (f,y)\in A_{1,n}\,\}$. Similarly to the proof of Theorem 6.3, we have $A_1\subset\cup_{i=1}^\infty\cap_{n=i}^\infty A_{1,n}$ and $\tilde\mu(A_1)\le\limsup_{n\to\infty}\tilde\mu^n(B_{1,n})$. It now suffices to show that the last limit tends to zero as $q\to 0^+$.

Indeed, using the conditional distribution of $\tilde\mu^n$ we get

$$\begin{aligned}\tilde\mu^n(B_{1,n}) \;&=\; \int_{\mathbb{R}^n}\mu_2(\{\,f\in F\mid\; (f,y^n)\in B_{1,n}\,\}\,|\,y^n\,)\,\mu_1(dy^n)\\ &=\; \int_{\mathbb{R}^n}\nu_2(\{\,g\in G\mid\; \|g-g_{y^n}\|<q\cdot\mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y)\,\}\,|\,y^n\,)\,\mu_1(dy^n)\,.\end{aligned}$$

Since $\mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y)=\mathrm{trace}(C_{\nu_{2,y^n}})$ where $C_{\nu_{2,y^n}}$ is the correlation operator of the Gaussian measure $\nu_2(\cdot|y^n)$, we can use Lemma 6.4 to get that

$$\nu_2\left(\{\,g\in G\mid\; \|g-g_{g^n}\|<q\cdot\mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y)\,\}\;|\,y^n\,\right)\;\le\;\frac{4}{3}\,\psi(2q)\,.$$

Thus $\tilde\mu^n(B_{1,n})\le 4/3\,\psi(2q)$. Since this tends to zero with $q\to 0^+$, $\tilde\mu(A_1)=0$. □

We now show that in some sense the sequence $\{\mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y)\}$ provides also an upper bound on the convergence rate.

Theorem 6.5 *For the algorithm $\varphi_{\rm spl}$ the set*

$$A_2 \;=\; \left\{ (f,y)\in F\times\mathbb{R}^\infty \;\Big|\; \lim_{n\to\infty} \frac{\mathrm{rad}_n^{\rm ave}(\mathbb{N}_y)}{\|S(f)-\varphi_{\rm spl}^n(y^n)\|} \;=\; 0 \right\}.$$

has the $\tilde\mu$–measure zero.

Proof Choose $q\in(0,1)$ and define

$$A_{2,n} \;=\; \{\,(f,y)\in F\times\mathbb{R}^\infty \mid \;\; \|S(f)-\varphi_{\rm spl}^n(y^n)\| \ge 1/q\cdot \mathrm{rad}_n^{\rm ave}(\mathbb{N}_y)\,\}$$

and $B_{2,n}=\{\,(f,y^n)\in F\times\mathbb{R}^\infty \mid (f,y)\in A_{2,n}\,\}$. Then $A_2\subset \cup_{i=1}^\infty \cap_{n=i}^\infty A_{2,n}$ and $\tilde\mu(A_2)\le \limsup_{n\to\infty}\tilde\mu^n(B_{2,n})$. Using decomposition of $\tilde\mu$ we have

$$\begin{aligned} &\tilde\mu^n(B_{2,n}) \qquad\qquad (6.13)\\ &=\; \int_{\mathbb{R}^n} \nu_2\left(\left\{ g\in G \mid \;\; \|g-g_{y^n}\| \ge 1/q\,\sqrt{\mathrm{trace}(C_{\nu,y^n})} \right\} \Big| y^n \right)\,\mu_1(dy^n)\,. \end{aligned}$$

We now use a slight generalization of the Chebyshev's inequality to estimate the Gaussian measure of the set of all g which are not in the ball centered at the mean element. Namely, if ω is a Gaussian measure on G then for any $r>0$

$$\begin{aligned} \mathrm{trace}(C_\omega) \;&=\; \int_G \|g-m_\omega\|^2\,\omega(dg)\\ &\ge\; \int_{\|g-m_\omega\|>r} \|g-m_\omega\|^2\,\omega(dg) \;\ge\; r^2\,\omega(\,G\setminus B_r(m_\omega)\,)\,, \end{aligned}$$

and consequently

$$\omega\,(\{\,g\in G \mid \;\; \|g-m_\omega\|\ge r\,\}) \;\le\; \frac{\mathrm{trace}(C_\omega)}{r^2}.$$

For $r=1/q\,\sqrt{\mathrm{trace}(C_\omega)}$, the right hand side of the last inequality is just q^2. Hence, (6.13) is bounded from above by q^2.

Using the same argument as in the proof of Theorem 6.4 we conclude that $\tilde\mu(A_2)=0$. □

Theorem 6.5 says that for all (f,y) a.e. some subsequence $\|S(f)-\varphi_{\rm spl}^{n_k}(y^{n_k})\|$ converges to zero at least as fast as $\mathrm{rad}_{n_k}^{\rm ave}(\mathbb{N}_y)$, as $k\to\infty$. Unfortunately, the word "subsequence" above is necessary. That is, we cannot claim in general that with probability one the sequence $\|S(f)-\varphi_{\rm spl}^n(y^n)\|$ behaves at least as well as $\mathrm{rad}_n^{\rm ave}(\mathbb{N}_y)$. Actually, this probability can be even zero, as illustrated in the following example.

Example 6.5 Let $F = G$ be the space of infinite sequences with $\|f\|_F^2 = \sum_{j=1}^{\infty} f_j^2 < +\infty$. We equip F with the zero mean Gaussian measure μ such that $C_\mu e_i = \lambda_i e_i$ where $\lambda_i = a^j$ and $0 < a < 1$. Consider approximation of $f \in F$ from exact information about coordinates of f, i.e., $N(f) = [f_1, f_2, f_3 \ldots]$ and $\Sigma = [0, 0, 0, \ldots]$. We shall see that then the set

$$A_3 \;=\; \left\{ (f, y) \in F \times \mathbb{R}^\infty \;\Big|\; \limsup_{n\to\infty} \|S(f) - \varphi_{\text{spl}}^n(y^n)\| / \text{rad}_n^{\text{ave}}(\mathbb{N}) < +\infty \right\}$$

has the $\tilde{\mu}$–measure zero.

Indeed, as noise does not exist, the measure $\tilde{\mu}$ is concentrated on the subspace $\{(f, N(f)) \mid f \in F\}$ and $\tilde{\mu}(A_3)$ equals the μ–measure of the set $B = \{\, f \in F \mid (f, N(f)\,) \in A_3 \,\}$. Moreover, since in this case $\varphi_{\text{spl}}^n(y^n) = [y_1, \ldots, y_n, 0, 0, 0, \ldots]$ and $\text{rad}_n^{\text{ave}}(\mathbb{N}) = \sqrt{\sum_{i=n+1}^{\infty} \lambda_i}$, we have $B = \bigcup_{k=1}^{\infty} B_k$ where

$$B_k \;=\; \left\{ f \in F \;\Big|\; \sum_{i=n}^{\infty} f_i^2 \le k^2 \sum_{i=n}^{\infty} \lambda_i, \quad \forall n \ge 1 \right\}.$$

Observe now that the condition $\sum_{i=n}^{\infty} f_i^2 \le k^2 \sum_{i=n}^{\infty} \lambda_i$ implies

$$|f_n| \;\le\; k \sqrt{\sum_{i=n}^{\infty} \lambda_i} \;=\; k \sqrt{\frac{a^n}{1-a}} \;=\; \frac{k}{\sqrt{1-a}} \sqrt{\lambda_n}.$$

Hence,

$$\begin{aligned} \mu(B_k) \;&\le\; \prod_{n=1}^{\infty} \mu\left(\left\{ f \in F \;\Big|\; |f_n| \le \frac{k}{\sqrt{1-a}} \sqrt{\lambda_n} \right\} \right) \\ &=\; \prod_{n=1}^{\infty} \left(\sqrt{\frac{2}{\pi}} \int_0^{\frac{k}{\sqrt{1-a}}} e^{-x^2/2}\, dx \right) \;=\; 0, \end{aligned}$$

and consequently $\tilde{\mu}(A_3) = \mu(B) = \lim_{k\to\infty} \mu(B_k) = 0$.

6.3.3 Optimal information

In the end, let us consider the problem of optimal information. That is, we fix the precision sequence $\Sigma = [\sigma_1^2, \sigma_2^2, \ldots]$ and want to select the infinite sequence of functionals $N = [L_1, L_2, \ldots]$ in such a way that the errors $\|S(f) - \varphi_{\text{spl}}^n(y^n)\|$ converge to zero as fast as possible. We assume that N belongs to the class $\mathcal{N}$ of all information for which the functionals L_i are in a given class $\Lambda \subset F^*$.

Theorems 6.4 and 6.5 say that for given information $\mathbb{N} = \{N, \Sigma\}$ the behavior of errors can be essentially characterized by that of the nth radii of $\mathbb{N}$. Hence, it seems natural to call optimal this information for which the sequence $\mathrm{rad}_n^{\mathrm{ave}}(N, \Sigma)$ vanishes with fastest rate.

For $n \geq 1$, let

$$\mathrm{r}_n^{\mathrm{ave}}(\Sigma) \;=\; \inf_{N \in \mathcal{N}} \mathrm{rad}_n^{\mathrm{ave}}(N^n, \Sigma^n)$$

be the minimal average error that can be achieved using first n nonadaptive observations. It is clear that for any information $N \in \mathcal{N}$ we have $\mathrm{rad}_n^{\mathrm{ave}}(N, \Sigma) \geq \mathrm{r}_n^{\mathrm{ave}}(\Sigma)$, i.e., the radii $\mathrm{rad}_n^{\mathrm{ave}}(N, \Sigma)$ do not converge faster than $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$. Consequently, Theorem 6.4 yields that for arbitrary information $N \in \mathcal{N}$ the $\tilde{\mu}$–measure of the set

$$\left\{ (f, y) \in F \times \mathbb{R}^\infty \;\middle|\; \lim_{n \to \infty} \frac{\|S(f) - \varphi_{\mathrm{spl}}^n(y^n)\|}{\mathrm{r}_n^{\mathrm{ave}}(\Sigma)} = 0 \right\}$$

is zero. We now give information N_Σ whose radii behave in many cases as $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$. To this end, we use construction of Section 6.2.3. That is, we let $\eta > 1$ and for $n \geq 1$ choose $N_n \in \mathcal{N}$ in such a way that

$$\mathrm{rad}_n^{\mathrm{ave}}(N_n, \Sigma) \;\leq\; \eta \cdot \mathrm{r}_n^{\mathrm{ave}}(\Sigma).$$

Then

$$N_\Sigma \;=\; [N_1^1, N_2^2, N_4^4, \ldots, N_{2^k}^{2^k}, \ldots].$$

Theorem 6.6 *Suppose the precision sequence $\Sigma = [\sigma_1^2, \sigma_2^2, \ldots]$ satisfies*

$$\sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2 \geq \cdots \geq 0.$$

Then for the information N_Σ and algorithm φ_{spl} the set

$$\left\{ (f, y) \in F \times \mathbb{R}^\infty \;\middle|\; \lim_{n \to \infty} \frac{\mathrm{r}_{\lceil \frac{n+1}{4} \rceil}^{\mathrm{ave}}(\Sigma)}{\|S(f) - \varphi_{\mathrm{spl}}^n(y^n)\|} = 0 \right\}$$

has the $\tilde{\mu}$–measure zero.

Proof Proceeding as in the proof of Theorem 6.2 we can show that

$$\mathrm{rad}_n^{\mathrm{ave}}(N_\Sigma, \Sigma) \;\leq\; \eta \cdot \mathrm{r}_{\lceil \frac{n+1}{4} \rceil}^{\mathrm{ave}}(\Sigma) \tag{6.14}$$

Hence, the theorem is a consequence of (6.14) and Theorem 6.5. □.

If $\mathrm{r}_n^{\mathrm{ave}}(\Sigma)$ behaves polynomially in $1/n$ (which holds for problems analyzed in Chapter 3), then $\mathrm{r}^{\mathrm{ave}}_{\lceil\frac{n+1}{4}\rceil}(\Sigma) \asymp \mathrm{r}_n^{\mathrm{ave}}(\Sigma)$. In such cases information N_Σ is optimal.

Notes and Remarks

NR 6.4 This section is original. However, the technique of proving Theorems 6.3, 6.4 and 6.5 is adopted from Wasilkowski and Woźniakowski [121] where the exact information case is studied and relations between the asymptotic and average case settings were established for the first time. Lemma 6.4 is due to Kwapień and also comes from the cited paper.

Exercises

E 6.6 Consider the problem of approximating a parameter $f \in \mathbb{R}$ from information $y = [y_1, y_2, y_3, \ldots] \in \mathbb{R}^\infty$ where $y_i = f + x_i$ and x_i's are independent, $x_i \sim \mathcal{N}(0, \sigma^2)$, $i \geq 1$. Show that then for any f

$$\pi_f \left(\left\{ y \in \mathbb{R}^\infty \;\middle|\; \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n y_j = f \right\} \right) \;=\; 1\,.$$

That is, the algorithm $\varphi^n(y^n) = 1/n \sum_{j=1}^n y_j$ converges to the "true" solution f with probability 1.

E 6.7 Consider the one dimensional problem of E 6.6. For y belonging to the set

$$C \;=\; \left\{ y \in \mathbb{R}^\infty \;\middle|\; \text{the limit} \quad m(y) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n y_j \quad \text{exists and is finite} \right\},$$

let ω_y be the Dirac measure on $\mathbb{R}$ centered at $m(y)$. Let μ_1 be the prior distribution of information $y \in \mathbb{R}^\infty$,

$$\mu_1(\cdot) \;=\; \int_F \pi_f(\cdot)\, \mu(df).$$

Show that $\mu_1(C) = 1$, and that for any measurable sets $A \subset F$ and $B \subset \mathbb{R}^\infty$ we have

$$\tilde{\mu}(A \times B) \;=\; \int_B \omega_y(A)\, \mu_1(dy).$$

That is, $\{\omega_y\}$ is the family of regular conditional distribution on $\mathbb{R}$ with respect to information $y \in \mathbb{R}^\infty$, $\omega_y = \mu_2(\cdot|y)$ $\forall y$ a.e.

E 6.8 Give an example where

$$\tilde{\mu}\, \left(\left\{\, (f, y) \in F \times \mathbb{R}^\infty \mid \quad \|S(f) - \varphi^n_{\mathrm{spl}}(y^n)\| \asymp \mathrm{rad}_n^{\mathrm{ave}}(\mathbb{N}_y) \,\right\} \right) \;=\; 1\,.$$

E 6.9 Suppose the class Λ consists of functionals whose μ–norm is bounded by 1. Let

$$N_0 \;=\; [\,\langle\cdot,\xi_1\rangle,\langle\cdot,\xi_2\rangle,\ldots\,]$$

where $\{\xi_i\}$ is the complete orthonormal basis of eigenelements of $SC_\mu S^*$ and the corresponding eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Assuming exact observations, $\Sigma = [0,0,0,\ldots]$, show that information N_0 is optimal independently of the behavior of $\mathrm{r}_n^{\mathrm{ave}}(0) = \sqrt{\sum_{j\geq n+1}\lambda_j}$.

Bibliography

[1] B.B. Arestov. Best recovery of operators and original problems. *Tr. MIAN, Nauka*, 189:2–20, 1989.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[3] K.I. Babenko. *Theoretical Background and Constructing Computational Algorithms for Mathematical–Physical Problems.* Nauka, Moscow, 1979. (In Russian).

[4] N.S. Bakhvalov. On the optimality of linear methods for operator approximation in convex classes. *Comput. Math. Math. Phys.*, 11:244–249, 1971.

[5] P.J. Bickel. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Annals of Statistics*, 9:1301–1309, 1981.

[6] L. Blum, M. Shub, and S. Smale. On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines. *Bull. of the AMS (new series)*, 21:1–46, 1989.

[7] L.D. Brown and I. Feldman. Manuscript. 1990.

[8] G. Casella and W.E. Strawderman. Estimating bounded normal mean. *Annals of Statistics*, 9:870–878, 1981.

[9] Z. Ciesielski. On Lévy's Brownian motion with several–dimensional time. volume 472 of *Lecture Notes in Mathematics*, pages 29–56. Springer, New York/London, 1975.

[10] N.J. Cutland. *Computability.* Cambridge Univ. Press, Cambridge, 1980.

[11] D.J. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? Technical Report, 1993.

[12] D.L. Donoho. Statistical estimation and optimal recovery. *Annals of Statistics*, 22:238–270, 1994.

[13] D.L. Donoho and I.M. Johnstone. Minimax estimation via wavelet shrinkage. Technical Report, 1992.

[14] D.L. Donoho, R.C. Liu, and K.B. MacGibbon. Minimax risk over hyperrectangles, and implications. *Annals of Statistics*, 18:1416–1437, 1990.

[15] R.L. Eubank. *Spline smoothing and nonparametric regression.* Dekker, New York, 1988.

[16] I.I. Gikhman and A.V. Skorohod. *Introduction to the theory of random processes.* Nauka, Moscow, 1965. (In Russian).

[17] M. Golomb and H.F. Weinberger. Optimal approximation and error bounds. In R.E. Lager, editor, *On Numerical Approximation*, pages 117–190. Univ. of Wisconsin Press, Madison, 1959.

[18] G.H. Golub, M.T. Heath, and G. Wahba. Validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

[19] G.K. Golubev and M. Nussbaum. A risk bound in Sobolev class regression. *Annals of Statistics*, 18:758–778, 1990.

[20] K.G. Golubev. On sequential experimental designs for nonparametric estimation of smooth regression functions. *Problems Inform. Transmission*, 28:76–79, 1992. (In Russian).

[21] T.N.E. Greville. Introduction to spline functions. In Greville, editor, *Theory and applications of spline functions*, pages 1–35. Academic Press, 1969.

[22] P.C. Hansen. Analysis of discrete ill–posed problems by means of the L–curve. *SIAM Review*, 34:561–580, 1992.

[23] S. Heinrich and J.D. Kern. Parallel information–based complexity. *J.Complexity*, 7:339–370, 1991.

[24] I.A. Ibragimov and R.Z. Hasminski. Bounds for the risk of nonparametric regression estimates. *Theory Probab. Appl.*, 28:81–94, 1982. (In Russian).

[25] I.A. Ibragimov and R.Z. Hasminski. On the nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.*, 29:19–32, 1984. (In Russian).

[26] K. Jensen and N. Wirth. *Pascal. User Manual and Report.* Springer Verlag, Berlin/Heidelberg/New York, 1975.

[27] B.Z. Kacewicz. Asymptotic error of algorithms for solving nonlinear problems. *J. Complexity*, 3:41–56, 1987.

[28] B.Z. Kacewicz. On sequential and parallel solution of initial value problems. *J. Complexity*, 6:136–148, 1990.

[29] B.Z. Kacewicz and M.A. Kowalski. Approximating linear functionals on unitary spaces in the presence of bounded data errors with applications to signal recovery. *Int. J. of Adaptive Control and Signal Processing*, pages ???–???, 1993.

[30] B.Z. Kacewicz and M.A. Kowalski. Recovering linear operators from inaccurate data. *J. Complexity*, 11:???–???, 1995.

[31] B.Z. Kacewicz, M. Milanese, R. Tempo, and A. Vicino. Optimality of central and projection algorithms for bounded uncertainty. *Systems Control Lett.*, 8:161–171, 1986.

[32] B.Z. Kacewicz and L. Plaskota. On the minimal cost of approximating linear problems based on information with deterministic noise. *Numer. Funct. Anal. and Optimiz.*, 11:511–525, 1990.

[33] B.Z. Kacewicz and L. Plaskota. Noisy information for linear problems in the asymptotic setting. *J. Complexity*, 7:35–57, 1991.

[34] B.Z. Kacewicz and L. Plaskota. Termination conditions for approximating linear problems with noisy information. *Math. of Comp.*, 59:503–513, 1992.

[35] B.Z. Kacewicz and L. Plaskota. The minimal cost of approximating linear operators using perturbed information – the asymptotic setting. *J. Complexity*, 9:113–134, 1993.

[36] J.B. Kadane, G.W. Wasilkowski, and H. Woźniakowski. On adaption with noisy information. *J. Complexity*, 4:257–276, 1988.

[37] G.S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495–502, 1970.

[38] Ker-I Ko. Applying techniques of discrete complexity theory to numerical computation. In R.V. Book, editor, *Studies in Complexity Theory*, pages 1–62. Pitman, London, 1986.

[39] M.A. Kon and E. Novak. On the adaptive and continuous information problems. *J. Complexity*, 5:345–362, 1989.

[40] M.A. Kon and E. Novak. The adaption problem for approximating linear operators. *Bull. Amer. Math. Soc.*, 23:159–165, 1990.

[41] N.P. Korneichuk. Optimization of active algorithms for recovery of monotonic functions from Holder's class. *J. Complexity*, pages 265–269, 1994.

[42] M.A. Kowalski. On approximation of band–limited signals. *J. Complexity*, 5:283–302, 1989.

[43] M.A. Kowalski, K. Sikorski, and F. Stenger. *Selected Topics in Approximation and Computation.* Oxford University Press, 1995. To appear.

[44] H.H. Kuo. *Gaussian Measures in Banach Spaces*, volume 463 of *Lecture Notes in Math.* Springer–Verlag, Berlin, 1975.

[45] C.L. Lawson and R.J. Hanson. *Solving Least Squares Problems.* Prentice–Hall, Inglewood Cliffs, N.J., 1974.

[46] D. Lee. Approximation of linear operators on a Wiener space. *Rocky Mount. J. Math.*, 16:641–659, 1986.

[47] D. Lee, T. Pavlidis, and G.W. Wasilkowski. A note on the trade-off between sampling and quantization in signal processing. *J. Complexity*, 3:359–371, 1987.

[48] D. Lee and G.W. Wasilkowski. Approximation of linear functionals on a Banach space with a Gaussian measure. *J. Complexity*, 2:12–43, 1986.

[49] B.Y. Levit. On asymptotic minimax estimates of the second order. *Theory Probab. Appl.*, 25:552–568, 1980.

[50] K-C. Li. Minimaxity of the method of regularization on stochastic processes. *Annals of Statistics*, 10:937–942, 1982.

[51] G.G. Magaril-Il'yaev. Average widths of Sobolev classes on R^n. *J. Approx. Th.*, 76:65–76, 1994.

[52] G.G. Magaril-Il'yaev and K.Yu. Osipenko. On optimal recovery of functionals from inaccurate data. *Matem. Zametki*, 50(6):85–93, 1991. (In Russian).

[53] V. Maiorov. Average n–widths of the Wiener space in the L_∞–norm. *J. Complexity*, 9:222–230, 1993.

[54] V. Maiorow. Linear widths of function spaces uquipped with the Gaussian measure. *J. Approx. Th.*, 77:74–88, 1994.

[55] A.G. Marchuk and K.Y. Osipenko. Best approximation of functions specified with an error at a finite number of points. *Math. Notes*, 17:207–212, 1975.

[56] P. Mathé. s–numbers in information–based complexity. *J. Complexity*, 6:41–66, 1990.

[57] A.A. Melkman and C.A. Micchelli. Optimal estimation of linear operators in Hilbert spaces from inaccurate data. *SIAM, J. Numer. Anal.*, 16:87–105, 1979.

[58] C.A. Micchelli. Optimal estimation of linear operators from inaccurate data: a second look. *Numerical Algorithms*, 5:375–390, 1993.

[59] C.A. Micchelli and T.J. Rivlin. A survey of optimal recovery. In *Estimation in Approximation Theory*, pages 1–54. Plenum, New York, 1977.

[60] V.A. Morozow. *Methods for Solving Incorrectly Posed Problems.* Springer Verlag, New York, 1984.

[61] A. Nemirowski. On parallel complexity of nonsmooth convex optimization. *J. Complexity*, 10:451–463, 1994.

[62] E. Novak. Optimal recovery and n-widths for convex classes of functions. To appear.

[63] E. Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349 of *Lecture Notes in Math.* Springer–Verlag, Berlin, 1988.

[64] E. Novak. Quadrature formulas for convex classes of functions. In H. Braß and G. Hämmerlin, editors, *Numerical Integration IV*. Birkhäuser Verlag, Basel, 1993.

[65] E. Novak. An adaption problem for nonsymmetric convex sets. To appear, 1994.

[66] E. Novak. The real number model in numerical analysis. *J. Complexity*, 11:???–???, 1995.

[67] M. Nussbaum. Spline smoothing in regression model and asymptotic efficiency in l_2. *Annals of Statistics*, 13:984–997, 1985.

[68] K.Yu. Osipenko. Optimal recovery of periodic functions from Fourier coefficients given with an error. Manuscript, 1994.

[69] E.W. Packel. Linear problems (with extended range) have linear optimal algorithms. *Aequationes Math.*, 30:18–25, 1986.

[70] A. Papageorgiou and G.W. Wasilkowski. Average complexity of multivariate problems. *J. Complexity*, 5:1–23, 1990.

[71] K.R. Parthasarathy. *Probability Measures on Metric Spaces.* Academic Press, 1967.

[72] E. Parzen. An approach to time series analysis. *Ann. Math. Statist.*, 32:951–989, 1962.

[73] E. Parzen. Probability density functionals and reproducing kernel Hilbert spaces. In M. Rosenblatt, editor, *Proc. Symposium on Time Series Analysis*, pages 155–169, New York, 1963. Wiley.

[74] S.H. Paskov. Average case complexity of multivariate integration for smooth functions. *J. Complexity*, 9:291–312, 1993.

[75] A. Pinkus. *n–Widths in Approximation Theory.* Springer-Verlag, Berlin, 1985.

[76] M.S. Pinsker. Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission*, 16:52–68, 1980. (In Russian).

[77] L. Plaskota. Asymptotic error for the global maximum of functions in s dimensions. *J. Complexity*, 5:369–378, 1989.

[78] L. Plaskota. On average case complexity of linear problems with noisy information. *J. Complexity*, 6:199–230, 1990.

[79] L. Plaskota. Function approximation and integration on the Wiener space with noisy data. *J. Complexity*, 8:301–323, 1992.

[80] L. Plaskota. A note on varying cardinality in the average case setting. *J. Complexity*, 9:458–470, 1993.

[81] L. Plaskota. Optimal approximation of linear operators based on noisy data on functionals. *J. Approx. Th.*, 73:93–105, 1993.

[82] L. Plaskota. Average case approximation of linear functionals based on information with deterministic noise. *J. of Computing and Information*, 4:21–39, 1994.

[83] L. Plaskota. Average complexity for linear problems in a model with varying noise of information. *J. Complexity*, 11:???–???, 1995.

[84] L. Plaskota. Complexity of multivariate integration with random noise. In progress, 1995.

[85] K. Ritter. Almost optimal differentiation using noisy data. Manuscript, 1994.

[86] K. Ritter, G.W. Wasilkowski, and H. Woźniakowski. Multivariate integration and approximation for random fields satisfying Sacks–Ylvisaker conditions. To appear, 1993.

[87] J. Sacks and D. Ylvisaker. Designs for regression problems with correlated errors. *Ann. Math. Stat.*, 37:66–89, 1966.

[88] J. Sacks and D. Ylvisaker. Designs for regression problems with correlated errors; many parameters. *Ann. Math. Stat.*, 39:49–69, 1968.

[89] J. Sacks and D. Ylvisaker. Designs for regression problems with correlated errors III. *Ann. Math. Stat.*, 41:2057–2074, 1970.

[90] I.J. Schoenberg. On interpolation by spline functions and its minimum properties. *Internat. Ser. Numer. Anal.*, 5:109–129, 1964.

[91] I.J. Schoenberg. Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. U.S.A.*, 52:947–949, 1964.

[92] I.J. Schoenberg and T.N.E. Greville. Smoothing by generalized spline functions. *SIAM Rev.*, 7:617, 1965.

[93] A. Schönhage. Equation solving in terms of computational complexity. In *Proc. Intern. Congress Math.*, Berkeley, 1986.

[94] K. Sikorski and G.M. Trojan. Asymptotic near optimality of the bisection method. *Numer. Math.*, 57:421–433, 1990.

[95] A.V. Skorohod. *Integration in Hilbert Spaces*. Springer-Verlag, New York, 1074.

[96] S.A. Smolyak. *On optimal recovery of functions and functionals of them*. PhD thesis, Moscow State Univ., 1965.

[97] P. Speckman. Minimax estimates of linear functionals in a Hilbert space. Manuscript, 1979.

[98] P. Speckman. Spline smoothing and optimal rates of convergence in nonparametric regression models. *Annals of Statistics*, 13:970–983, 1985.

[99] S.B. Steckin and Yu.N. Subbotin. *Splines in numerical mathematics*. Nauka, Moscow, 1976. (in Russian).

[100] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053, 1982.

[101] A.G. Sukharev. On the existence of optimal affine methods for approximating linear functionals. *J. Complexity*, 2:317–322, 1986.

[102] A.V. Suldin. Wiener measure and its applications to approximation methods, I. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 13:145–158, 1959. (In Russian).

[103] A.V. Suldin. Wiener measure and its applications to approximation methods, II. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 18:165–179, 1960. (In Russian).

[104] Y. Sun and C. Wang. μ-average n-widths on the Wiener space. *J. Complexity*, 10:428–436, 1994.

[105] A.N. Tikhonov. On regularization of ill–posed problems. *Dokl. Akad. Nauk USSR*, 153:49–52, 1963.

[106] A.N. Tikhonov and V.Ja. Arsenin. *Methods for solving ill–posed problems.* Wiley, New York, 1979.

[107] J.F. Traub, G.W. Wasilkowski, and H. Woźniakowski. *Information, Uncertainty, Complexity.* Addison–Wesley, Mass., 1983.

[108] J.F. Traub, G.W. Wasilkowski, and H. Woźniakowski. *Information–based Complexity.* Academic Press, New York, 1988.

[109] J.F. Traub and H. Woźniakowski. *A General Theory of Optimal Algorithms.* Academic Press, New York, 1980.

[110] H. Triebel. *Theory of Function Spaces.* Birkhäuser Verlag, Basel, 1983.

[111] G.M. Trojan. Asymptotic setting for linear problems. Manuscript, 1983.

[112] N.N. Vakhania. *Probability Distributions on Linear Spaces.* North-Holland, New York, 1981.

[113] N.N. Vakhania, V.I. Tarieladze, and S.A. Chobanyan. *Probability Distributions on Banach Spaces.* Reidel, Dordrecht, 1987.

[114] V.S. Varadarajan. Measures on topological spaces. *Mat. Sbornik*, 55:35–100, 1961. (In Russian).

[115] G. Wahba. On the regression design problem of Sacks and Ylvisaker. *Ann. Math. Stat.*, pages 1035–1043, 1971.

[116] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS–NSF Series in Appl. Math.* SIAM, 1990.

[117] G.W. Wasilkowski. Local average error. Columbia University Comp. Sc. Report, 1983.

[118] G.W. Wasilkowski. Information of varying cardinality. *J. Complexity*, 2:204–228, 1986.

[119] G.W. Wasilkowski. Integration and approximation of multivariate functions: average case complexity with isotropic Wiener measure. *J. Approx. Th.*, 77:212–227, 1994.

[120] G.W. Wasilkowski and H.Woźniakowski. There exists a linear problem with infinite combinatory cost. *J. Complexity*, 7:326–337, 1993.

[121] G.W. Wasilkowski and H. Woźniakowski. On optimal algorithms in an asymptotic model with Gaussian measure. *SIAM, J. Math. Anal.*, 3:632–647, 1987.

[122] G.W. Wasilkowski and H. Woźniakowski. Explice cost bounds of algorithms for solving multivariate problems. *J. Complexity*, 11:???–???, 1995.

[123] A.G. Werschulz. An information–based approach to ill–posed problems. *J. Complexity*, 3, 1987.

[124] A.G. Werschulz. *The Computational Complexity of Differential and Integral Equations.* Oxford University Press, Oxford, 1992.

[125] A.G. Werschulz and H. Woźniakowski. Are linear algorithms always good for linear problems? *Aequationes Math.*, 30:202–212, 1986.

[126] A. Wilansky. *Modern Methods in Topological Vector Spaces.* McGraw–Hill, New York, 1978.

[127] H. Woźniakowski. Average case complexity of multivariate integration. *Bull. AMS*, 24:185–194, 1991.

[128] H. Woźniakowski. Average case complexity of multivariate linear problems I, II. *J. Complexity*, 8:337–392, 1992.

[129] H. Woźniakowski. Tractability of linear multivariate problems. *J. Complexity*, 10:96–128, 1994.