INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Noisy Information and Computational Complexity: A Short Survey

Leszek Plaskota[*]

TR-95-055

September 1995

Abstract

In the modern world, the importance of information can be hardly overestimated. Information also plays a prominent role in scientific computations. A branch of computational complexity which deals with problems for which information is partial, noisy, and priced is called *information–based complexity.*

In most of the work on information–based complexity, the emphasis was on partial and exact information. We concentrate our attention on *noisy* information. We consider deterministic and random noise. The analysis of noisy information leads to a variety of new algorithms and complexity results.

This short survey has a reach extension in the form of a monograph 'Noisy Information and Computational Complexity', to be published in Cambridge University Press.

[*] Department of Mathematics, Informatics, and Mechanics, University of Warsaw, Poland.

1 Introduction

In the process of doing scientific computations we always rely on some *information.* In practice, this information is typically *noisy*, i.e., contaminated by error. Sources of noise include: previous computations, inexact measurements, transmission errors, arithmetic limitations, an adversary's lies.

Problems with noisy information have always attracted attention from researchers in many different scientific fields. There is also a vast literature, especially in statistics, where noisy information is analyzed from different perspectives.

In this short survey, noisy information is studied in the context of the computational complexity of solving mathematical problems.

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*, and *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 error. 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 to find 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. Those solving the problem with minimal cost are of special importance and are 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 use a finite set of numbers. Usually, these numbers cannot be entered exactly into the computer's memory. Hence, problems described by infinitely many parameters can be 'solved' only using 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 unified approach to such 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 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 *Problem Complexity* *and Method Efficiency in Optimization* by A.S. Nemirovski and D.B. Yudin, Wiley and Sons, 1983, *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 its analysis. 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 to mean random error that contaminates 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, either in the model or 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.

The general IBC model covers a large variety of problems. We are mainly interested in *linear problems*, i.e., problems which can be described in terms of approximating a linear operator from noisy information about values of some linear functionals. Examples include function approximation and integration, where information is given by noisy function values.

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. We study: *worst case setting*, *average case setting*, *mixed settings*, and *asymptotic setting.* In the worst case setting, the error and cost are defined by their worst performance. In the average case setting, we consider the average error and cost. The mixed settings are obtained by combining the worst and average cases. In the asymptotic setting, we are interested in the asymptotic behavior of algorithms. Other settings such as probabilistic or randomized settings are also important and will be involved in the topics of future research.

2 Worst case setting

Let

$$S : F \to G$$

be a linear operator acting between a linear space F and a normed space G (both over $\mathbb{R}$). Our aim is to approximate elements $S(f)$ for f belonging to a subset $E \subset F$. An approximation is constructed based on some *information* about f. This information is given as a vector of the form

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

where $N : F \to Y = \mathbb{R}^n$ is a linear operator. Hence, y is a perturbed value of exact information $N(f)$. The vector x represents *noise*. We assume that x is bounded in a norm,

$$\|x\|_Y \le \delta.$$

Knowing information y about f, the approximation is given as $\varphi(y)$, where the transformation

$$\varphi : Y \to G$$

is called an *algorithm*. In the worst case setting, the error of an algorithm is defined by its worst behavior with respect to f and x, i.e.,

$$e^{\mathrm{wor}}(N, \delta, \varphi) = \sup_{f \in E} \sup_{\|x\|_Y \le \delta} \|S(f) - \varphi(N(f) + x)\|,$$

where $\|\cdot\|$ is the norm in G.

Our first goal is to find *optimal algorithms*, that minimize the error for given information,

$$r^{\mathrm{wor}}(N, \delta) = \inf_{\varphi} e^{\mathrm{wor}}(N, \delta, \varphi).$$

We are especially interested in existence of optimal algorithms which have a simple form, e.g. which are linear or affine. This holds, for instance, if S is a functional.

Theorem 1 *(Magaril-Il'yaev and Osipenko, 1991) Let S be a linear functional. If the set E is convex then there exists an affine algorithm which is optimal.*

In this case we also have the following nice formula for the optimal error:

$$r^{\mathrm{wor}}(N, \delta) = \sup\{S(h) : \ h \in (E - E)/2, \|N(h)\|_Y \le \delta\}.$$

If S is arbitrary and E is a ball,

$$E = \{f \in F : \ \|f\|_F \le 1\},$$

then (almost) optimal approximations are provided by *spline algorithms*. These are algorithms of the form $\varphi_{\mathrm{spl}}(y) = S(s(y))$, where $s(y)$ is a *spline element*, which can be defined in different ways. For instance,

$$s(y) = \arg\min\{\|f\|_F : \ \|y - N(f)\|_Y \le \delta\}$$

(ordinary spline), or

$$s(y) = \arg\min_{f \in F} \lambda \|f\|_F^2 + \|y - N(f)\|_Y^2$$

(smoothing spline). In the latter case, λ is a nonnegative parameter. (For simplicity, we assume that the minima above are attained.)

Theorem 2 *(Kacewicz and Plaskota, 1991) For the ordinary spline algorithm we have*

$$e^{\text{wor}}(N, \delta, \varphi_{\text{spl}}) \leq 2\, r^{\text{wor}}(N, \delta).$$

A similar theorem (but with a different constant) can be shown for smoothing splines. It turns out that in some cases smoothing splines are (strictly) optimal.

Theorem 3 *(Melkman and Micchelli, 1979) If $\|\cdot\|_F$, $\|\cdot\|$, and $\|\cdot\|_Y$ are Hilbert norms, then there exists $\lambda \in [0, +\infty]$ such that the smoothing spline algorithm with parameter λ is optimal. Moreover,*

$$r^{\text{wor}}(N, \delta) = \sup\{\|S(h)\| : \ \|h\|_F \leq 1, \|N(h)\|_Y \leq \delta\}.$$

Note that in this case φ_{spl} is a linear algorithm. The following lemma says about the optimal choice of λ.

Lemma 1 *(Plaskota, 1995b) Let S be a compact operator. Let $\{\xi_j\}$ be a complete orthonormal basis of N^*N, and let η_j be the corresponding eigenvalues. Then the optimal λ^* minimizes*

$$\psi(\lambda) = \sup_{\|g\|=1} \sum_{j\geq 1} \frac{\langle S(\xi_j), g\rangle^2}{\lambda + \eta_j}, \qquad \lambda \geq 0,$$

and $r^{\text{wor}}(N, \delta) = \sqrt{\psi(\lambda^*)}$. If S is a functional then

$$\psi(\lambda) = \sum_{j\geq 1} \frac{S^2(\xi_j)}{\lambda + \eta_j}.$$

So far we have assumed that information is fixed. Now, we assume that not only algorithms, but also information can vary. More precisely, information is collected by noisy computations (or observations) of functionals belonging to some class Λ, i.e., $y = [y_1, y_2, \ldots, y_n]$, where

$$y_i = L_i(f) + x_i,$$

$L_i \in \Lambda$. In the model with varying information, it is natural to ask for *optimal information.* For given n and δ, we want to minimize the error $r^{\text{wor}}(N, \delta)$ with respect to $N = [L_1, \ldots, L_n]$, where $L_i \in \Lambda$. This problem has been solved in several cases. We give two examples.

Assume first that S is a compact operator. Denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ the dominating eigenvalues of the operator S^*S. Let

$$\Lambda = \{L : \ \|L\|_F \leq 1\},$$

and let $\|\cdot\|_Y = \|\cdot\|_2$ be the Euclidean norm in $\mathbb{R}^n$, i.e., the information noise satisfies $\sum_{j=1}^n x_j^2 \leq \delta^2 \leq 1$.

Theorem 4 *(Plaskota, 1995b) For approximation of a compact operator the minimal error is given as*

$$r^{\mathrm{wor}}(n,\delta) = \sqrt{\lambda_{n+1} + \frac{\delta^2}{n}\sum_{i=1}^{n}(\lambda_j - \lambda_{n+1})}.$$

Actually, this problem has been solved more generally, assuming that $\|x\|_Y$ is a weighted Euclidean norm, $\|x\|_Y^2 = \sum_{j=1}^n \delta_j^{-2} x_j^2$. Optimal information is also known. It relies on observations of some particular functionals belonging to the subspace spanned by the eigenelements of S^*S corresponding to the first n eigenvalues $\lambda_1, \dots, \lambda_n$. In particular, it turns out that neither the optimal information nor optimal algorithm depends on the noise level δ, see Theorems 2.8.1 and 2.8.2 in Plaskota (1995b).

Let us now consider the optimal information in a function space. We assume that E is the class of 1–Lipschitz functions $f : [0,1] \to \mathbb{R}$. Information is given by noisy function values at n points t_i, and the noise

$$|x_i| = |y_i - f(t_i)| \le \delta_i, \qquad 1 \le i \le n.$$

Without loss of generality w assume that $0 \le \delta_1 \le \cdots \le \delta_n$. Denote $\Delta_n = [\delta_1, \dots, \delta_n]$. We consider the approximation problem, $S = \mathrm{App} : E \to C([0,1])$, $\mathrm{App}(f) = f$, and integration, $S(f) = \mathrm{Int}(f) = \int_0^1 f(u)du$.

Theorem 5 *(Plaskota, 1995b) For approximation and integration of Lipschitz functions we have*

$$\begin{aligned} r^{\mathrm{wor}}(\mathrm{App}, \Delta_n) &= \frac{1}{k}\Big(\frac{1}{2} + \sum_{j=1}^{k}\delta_j\Big), \\ r^{\mathrm{wor}}(\mathrm{Int}, \Delta_n) &= \frac{1}{k}\Big(\frac{1}{2} + \sum_{j=1}^{k}\delta_j\Big)^2 - \sum_{j=1}^{k}\delta_j^2, \end{aligned}$$

where k is the largest integer satisfying $1 \le k \le n$ and $\delta_k \le \frac{1}{k}(\frac{1}{2} + \sum_{j=1}^k \delta_j)$. The optimal information consists of k observations at

$$t_i^* = \frac{2i-1}{k}\Big(\frac{1}{2} + \sum_{j-1}^{k}\delta_j\Big) - 2\Big(\sum_{j=1}^{i-1}\delta_j\Big) - \delta_i, \qquad 1 \le i \le k.$$

From the point of view of practical computations it is important that the algorithms not only give approximate solutions with small error, but also that their cost is small. Our *model of computation* is based, roughly speaking, on the following three postulates:

- the arithmetic operations and comparisons over the reals are allowed and any such an operation costs unity,

- linear operations (addition, multiplication by a scalar) in the space G are allowed and cost $\mathbf{g} \geq 1$, and
- obtaining noisy value of a functional with accuracy δ costs $\mathbf{c}(\delta)$, where $\mathbf{c}$ is a *cost function.*

Thus, in our general model we assume that we can select not only the functionals, but also accuracies with which we observe them. Moreover, the functionals L_i, the accuracies δ_i, and the total number n of observations can be chosen adaptively, depending on results of previous observations. Formally, this means that $L_i(\cdot) = L_i(\cdot; y_1, \ldots, y_{i-1}) \in \Lambda$, $\delta_i = \delta_i(y_1, \ldots, y_{i-1})$, and the space Y of information values may consist of vectors with different number of components. Such information is called *adaptive.* If the functionals, accuracies, and their number are given a priori, information is *nonadaptive.*

Note that different cost functions correspond to different practical situations. For instance, the function $\mathbf{c}(\delta) = c_0$ for $\delta \geq \delta_0$, and $\mathbf{c}(\delta) = +\infty$ for $0 \leq \delta < \delta_0$, corresponds to the case when all observations are performed with the same precision δ_0. The function $\mathbf{c}(\delta) = \log 1/\delta$ corresponds, roughly speaking, to the cost of storing information with absolute accuracy δ, in the fixed point arithmetic, etc.

For a given operator N, accuracy Δ, and algorithm φ, let

$$\mathrm{cost}^{\mathrm{wor}}(N, \Delta, \varphi) = \sup_y \mathrm{cost}(N, \Delta, \varphi, y)$$

be the cost of obtaining the approximation for the 'hardest' information y. (It consists of the information cost of obtaining y and combinatory cost of computing $\varphi(y)$.) The quantity

$$\mathrm{comp}^{\mathrm{wor}}(\varepsilon) = \inf\{\mathrm{cost}^{\mathrm{wor}}(N, \Delta, \varphi) : \ N, \Delta, \varphi \text{ such that } e^{\mathrm{wor}}(N, \Delta, \varphi) \leq \varepsilon\}$$

is called the *ε–complexity* of the problem.

It turns out that if observations are performed with a fixed accuracy, $|y_i - L_i(f)| \leq \delta_0$, then the (worst case) complexity is often infinite.

Theorem 6 *(Plaskota, 1995b) Let E be the unit ball in a norm $\|\cdot\|_F$. Suppose that only observations with fixed accuracy δ_0 are allowed, and there exists an element $h^* \in F$ such that $h^* \notin$ ker S and $|L(h^*)| \leq 1$, $\forall L \in \Lambda$. Then, for $\varepsilon < \min\{\delta_0, \|h^*\|_F^{-1}\}\|S(h^*)\|$ we have*

$$\mathrm{comp}^{\mathrm{wor}}(\varepsilon) = +\infty.$$

If arbitrary accuracies are possible, complexity is usually finite, though it may be large. We now give two examples. Consider first the approximation and integration problems, as in Theorem 5.

Theorem 7 *(Plaskota, 1995b) Let the cost function be given as* $\mathbf{c}(\delta) = \delta^{-q}$*, where* $q \geq 0$*. Then*

$$\mathrm{comp}^{\mathrm{wor}}(\mathrm{App}, \varepsilon) \asymp \mathrm{comp}^{\mathrm{wor}}(\mathrm{Int}, \varepsilon) \asymp \left(\frac{1}{\varepsilon}\right)^{q+1}, \qquad \varepsilon \to 0.$$

In the second example we have multivariate approximation. Let E be the unit cube in the Banach space F of functions $f : D = [0,1]^d \to \mathrm{I\!R}$, which are r times continuously differentiable, with the norm

$$\|f\|_F = \max_{0 \leq k_1 + \cdots + k_d = i} \sup_{t \in D} \left| \frac{\partial^k f(t)}{(\partial x^1)^{k_1} \cdots (\partial x^d)^{k_d}} \right|,$$

where $t = (t^1, \ldots, t^d)$. The problem is to approximate f in the sup–norm, based on its noisy values at some points.

Theorem 8 *(Kacewicz and Plaskota, 1990) If* $\mathbf{c}(1/x)$ *tends to* $+\infty$ *polynomially, as* $x \to +\infty$*, then for multivariate approximation we have*

$$\mathrm{comp}^{\mathrm{wor}}(\varepsilon) \asymp \mathbf{c}(\varepsilon)\varepsilon^{-d/r}, \qquad \varepsilon \to 0.$$

Optimal information uses $n \asymp \varepsilon^{d/r}$ *equispaced observations with accuracy* $\delta \asymp \varepsilon$*, and piecewise polynomial approximation is optimal.*

Hence, in this case we have the *curse of dimensionality*, since the complexity depends exponentially on d.

3 Average case setting

In the average case setting, we assume that the elements f as well as the information noise x are Gaussian random variables, and we are interested in minimization of the expected error of algorithms. More precisely, F is a separable Banach space equipped with a zero mean Gaussian measure μ whose correlation operator C_μ is positive definite. Information is given by (1), however, the noise x is now the n–dimensional random variable with normal distribution,

$$x \sim \pi = \mathcal{N}(0, \sigma^2 \Sigma),$$

where the covariance matrix $\Sigma = \Sigma^T > 0$. The parameter σ represents the noise level and hence corresponds do δ from the worst case.

The problem is to approximate $S(f)$, where $S : F \to G$ is a linear continuous operator and G is a separable Hilbert space. The error of an algorithm φ is defined as

$$e^{\mathrm{avg}}(N, \sigma, \varphi) = \left(\int_F \int_Y \|S(f) - \varphi(N(f) + x)\|^2 \pi(dx) \mu(df) \right)^{1/2}.$$

In the average case, optimal algorithms are given as $\varphi(y) = S(m(y))$, where $m(y)$ is the mean element of the conditional distribution on F, given information y. Moreover, these algorithms can be interpreted as smoothing spline algorithms, which was well known is some special cases, see, e.g., Wahba (1990). Here we cite a general result. Let $H \subset F$ be the Hilbert space such that the pair $\{H, F\}$ is an abstract Wiener space for the measure μ. Recall that then for two functionals $L_1, L_2 \in F^*$ we have

$$\int_F L_1(f)L_2(f)\mu(df) = \langle h_1, h_2 \rangle_H,$$

where h_i is the representer of L_i in H, i.e., $L_i(f) = \langle f, h_i \rangle_H$, $\forall f \in H$, $i = 1, 2$. Let $\langle \cdot, \cdot \rangle_Y = \langle \Sigma^{-1}(\cdot), \cdot \rangle_2$. Define a spline algorithm as $\varphi_{\rm spl}(y) = S(s(y))$, where

$$s(y) = \arg \min_{f \in H} \sigma^2 \|f\|_H^2 + \|y - N(f)\|_Y^2. \tag{2}$$

Note that $\varphi_{\rm spl}$ corresponds to the smoothing spline algorithm from the worst case with parameter $\lambda = \sigma^2$.

Theorem 9 *(Plaskota, 1995b) In the average case setting, the spline algorithm $\varphi_{\rm spl}$ is optimal.*

This fact allows us to establish a correspondence between the worst and average case settings. To do this, consider the following two problems.

1. Approximate $S(f)$ for $f \in E \subset F$ from information $y = N(f) + x$, where $\|x\|_Y \le \delta$.
2. Approximate $S(f)$, where $f \in F$ has Gaussian distribution μ, from information $y = N(f) + x$, $x \sim \mathcal{N}(0, \sigma^2 \Sigma)$.

Theorem 10 *(Plaskota) Let $\{H, F\}$ be an abstract Wiener space and E the unit ball in H. If $\delta = \sigma$ then the spline algorithm $\varphi_{\rm spl}$ is optimal for the problem (2) in the average case setting, $e^{\rm avg}(N, \sigma, \varphi_{\rm spl}) = r^{\rm avg}(N, \sigma)$, and almost optimal for the problem (1) in the worst case setting, $e^{\rm wor}(N, \delta, \varphi_{\rm spl}) \le \sqrt{2}\, r^{\rm wor}(N, \delta)$. Furthermore,*

$$r^{\rm wor}(N, \delta) \le \sqrt{2}\, r^{\rm avg}(N, \sigma).$$

If S is a functional then

$$r^{\rm avg}(N, \sigma) \le r^{\rm wor}(N, \delta) \le \sqrt{2}\, r^{\rm avg}(N, \sigma).$$

Let us now consider the problem of optimal information. First, assume that Σ is the identity matrix, and information relies on n noisy observations of functionals from the class

$$\Lambda = \{L \in F^* : \ \|L\|_H = \sup_{\|f\|_H = 1} |L(f)| \le 1\},$$

where H is the Hilbert space for the measure μ. Let $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ be the dominating eigenvalues of the operator $SC_\mu S^* : G \to G$. (Recall that we have $\sum_{i \ge 1} \lambda_i < +\infty$.)

Theorem 11 *(Plaskota, 1990) For approximation of continuous operators we have*

$$r^{\mathrm{avg}}(n,\sigma) = \left(\sigma^2 \frac{(\sum_{i=1}^{k} \lambda_i^{1/2})^2}{n+\sigma^2 k} + \sum_{j=k+1}^{\infty} \lambda_j \right)^{1/2},$$

where k is the largest integer satisfying $1 \le k \le n$ and

$$\sigma^2 \frac{\sum_{j=1}^{k} \lambda_j^{1/2}}{n+\sigma^2 k} \le \lambda_k^{1/2}.$$

This theorem has a generalization to the case when Σ is a diagonal matrix, i.e., when observations are performed with different variances. We also know formulas for the optimal information, see Plaskota (1993a).

As the second example, consider the function approximation and integration in the classical Wiener space. More precisely, we assume that F is the space of continuous functions $f:[0,1] \to \mathbb{R}$ such that $f(0)=0$, and $\mu = w$ is the classical Wiener measure on F. The problem App relies on approximating a function in $\mathcal{L}_2$–norm, while Int on computing $\int_0^1 f(u)du$. Information is given as noisy function values, where the noise of different observations is independent and its variance equals σ^2.

Theorem 12 *(Plaskota, 1992) For the approximation and integration on the Wiener space we have*

$$\begin{aligned} r^{\mathrm{avg}}(\mathrm{App},\sigma,n) &\approx \frac{1}{\sqrt{6n}} + p_n \left(\frac{\sigma^2}{4n}\right)^{1/4}, \\ r^{\mathrm{avg}}(\mathrm{Int},\sigma,n) &\approx \frac{1}{2\sqrt{3n}} + q_n \left(\frac{\sigma^2}{n}\right)^{1/2}, \end{aligned}$$

where $p_n, q_n \in [\sqrt{3}/3, 1]$.

We now give formulas for the ε–complexity of the two problems above. We assume the model of computation as in the worst case setting. The only difference is that observations are contaminated by random noise, and $\mathbf{c}(\sigma)$ denotes the cost of one observation with variance σ^2. Information (which can be adaptive or nonadaptive) is identified with the pair N, Σ, where N denotes the selection of functionals and Σ the selection of variances. The cost of obtaining approximations is given by the average cost,

$$\mathrm{cost}^{\mathrm{avg}}(N,\Sigma,\varphi) = \int_Y \mathrm{cost}(N,\Sigma,\varphi,y)\mu_1(dy),$$

where μ_1 is the a priori distribution of y (induced by the distributions of f and x). The ε-complexity is defined as

$$\mathrm{comp}^{\mathrm{avg}}(\varepsilon) = \inf\{\mathrm{cost}^{\mathrm{avg}}(N,\Sigma,\varphi) : \ N,\Sigma,\varphi \text{ such that } e^{\mathrm{avg}}(N,\Sigma,\varphi) \le \varepsilon\}.$$

While adaptive information is not better than nonadaptive information in the worst case setting (at least for convex and symmetric sets E, see, e.g., Corollary 2.7.1 in Plaskota (1995b)), the situation is quite different in the average case setting. It turns out that the use of varying cardinality can sometimes significantly reduce the cost of obtaining the ε–approximation, see Plaskota (1993b). However, provided some additional assumptions are satisfied, adaption does not help in the average case setting, either. Namely, let $\mathrm{ic}(\varepsilon)$ be the minimal cost of obtaining nonadaptive information from which it is possible to construct an approximation with error at most ε.

Theorem 13 *(Plaskota, 1995a) Let* $\varepsilon_0 = \int_F \|S(f)\|^2\mu(df)$. *Suppose that* $\mathrm{ic}(\sqrt{\varepsilon})$ *is a semi–convex function of* ε *on the interval* $[0,\varepsilon_0]$*, i.e., there exists* $0 < \alpha \leq \beta$ *and a convex function* $h : [0,\varepsilon] \to [0,+\infty]$ *such that*

$$\alpha\, h(\varepsilon) \leq \mathrm{ic}(\sqrt{\varepsilon}) \leq \beta\, h(\varepsilon), \qquad \forall 0 \leq \varepsilon \leq \varepsilon_0.$$

Then

$$\mathrm{comp}^{\mathrm{avg}}(\varepsilon) \geq \frac{\alpha}{\beta}\,\mathrm{ic}(\varepsilon), \qquad \forall 0 \leq \varepsilon \leq \varepsilon_0.$$

The assumptions of this theorem hold for many problems. In particular, they are satisfied by the two problems considered below. We start with approximation of continuous operators from Theorem 11.

Theorem 14 *(Plaskota, 1995a) Suppose that the eigenvalues of* $SC_\mu S^*$ *satisfy*

$$\lambda_j \asymp \left(\frac{ln^s j}{j}\right)^p, \qquad j \to \infty,$$

where $p \geq 1$ *and* $s \geq 0$. *Let the cost function* $\mathbf{c}(\sigma) = (1+\sigma^{-2})^q$, $q \geq 0$. *Then*

$$\mathrm{comp}^{\mathrm{avg}}(\varepsilon) \asymp \begin{cases} (1/\varepsilon)^{2\tilde{q}} & (p-1)\tilde{q} > 1, \\ (1/\varepsilon)^{2/(p-1)}(\ln 1/\varepsilon)^{(s+1)p/(p-1)} & (p-1)\tilde{q} = 1, \\ (1/\varepsilon)^{2/(p-1)}(\ln 1/\varepsilon)^{sp/(p-1)} & 0 \leq (p-1)\tilde{q} < 1, \end{cases}$$

$\varepsilon \to 0$.

This theorem applies, for instance, to $\mathcal{L}_2$–approximation of multivariate functions with respect to a Wiener sheet measure.

Let us now consider the function approximation and integration in the classical Wiener space, as in Theorem 12. Recall that in this case information is restricted to noisy function values.

Theorem 15 *(Plaskota, 1995b) Suppose that the cost function* $\mathbf{c}(\sigma) = \sigma^{-2}$*, or that only observations with fixed variance* $\sigma_0^2 > 0$ *are allowed. Then, for the function approximation and integration problems, we have*

$$\mathrm{comp}^{\mathrm{avg}}(\mathrm{App}, \varepsilon) \asymp \varepsilon^{-4} \quad \textit{and} \quad \mathrm{comp}^{\mathrm{avg}}(\mathrm{Int}, \varepsilon) \asymp \varepsilon^{-2},$$

$\varepsilon \to 0$.

Note that, unlike in the worst case, in the average case setting we can reduce the error to an arbitrary level using observations with fixed variance.

4 Mixed settings

Mixed settings are obtained by combining the deterministic assumptions of the worst case setting with stochastic assumptions of the average case setting.

In the first mixed setting (worst–average case setting), we assume that the problem elements f have deterministic character, $f \in E \subset F$ (as in the worst case), while the information noise has random character, i.e., $x \sim \pi = \mathcal{N}(0, \sigma^2\Sigma)$ (as in the average case). The error of an algorithm is defined as

$$e^{\mathrm{w-a}}(N, \sigma, \varphi) = \sup_{f \in E} \left(\int_Y \|S(f) - \varphi(N(f) + x)\|^2 \pi(dx) \right)^{1/2}.$$

In this setting, the main results are on optimal algorithms for approximating linear functionals S. It turns out that even if the set E is convex, optimal algorithms are in this case nonlinear (comp. with Theorem 1). However, affine algorithms are not much worse. This important fact was shown by Donoho (1994) who used some additional assumptions. The following theorem is a generalization of that result. Let $r_{\mathrm{aff}}^{\mathrm{w-a}}(N, \sigma)$ and $r_{\mathrm{non}}^{\mathrm{w-a}}(N, \sigma)$ be the optimal errors of affine and arbitrary (even nonlinear) algorithms, respectively.

Theorem 16 *(Plaskota, 1995b) If* S *is a linear functional and* E *is a convex set, then*

$$1 \le \frac{r_{\mathrm{aff}}^{\mathrm{w-a}}(N, \sigma)}{r_{\mathrm{non}}^{\mathrm{w-a}}(N, \sigma)} \le 1,11\ldots$$

In the proof of this theorem, one shows a relation between optimal algorithms in the mixed and worst case settings, in which the noise $\|x\|_Y^2 = \langle \Sigma^{-1}x, x \rangle_2 \le \delta^2$. This relation reads as follows.

Theorem 17 *(Plaskota, 1995b) For any* σ *there exists* $\delta = \delta(\sigma)$ *such that the same affine algorithm is optimal in the mixed and worst case settings. And vice versa. For any* δ *there exists* $\sigma = \sigma(\delta)$ *such that the same affine algorithm is optimal in both settings.*

The situation becomes much more complicated if S is a linear operator and not a functional. (Almost) optimal algorithms are known only for some special cases, see e.g., Nussbaum (1985), Donoho et al. (1995).

In the second mixed setting (average–worst case setting) we assume that F is a separable Banach space with a Gaussian measure μ (as in the average case setting), and that the noise satisfies $\|x\|_Y \leq \delta$ (as in the worst case setting). We will assume that $\|\cdot\|_Y$ is the Hilbert norm induced by a matrix $\Sigma = \Sigma^T > 0$. The error of algorithms is given by the formula

$$e^{\mathrm{a-w}}(N, \delta, \varphi) = \left(\int_F \sup_{\|x\|_Y \leq \delta} \|S(f) - \varphi(N(f) + x)\|^2 \mu(df) \right)^{1/2}.$$

Similarly to the first mixed setting, the main results are on approximating linear continuous functionals. As we can easily convince ourselves, the optimal algorithms are again nonlinear. However, linear algorithms (which are now as powerful as affine algorithms) are not much worse. Let $r^{\mathrm{a-w}}_{\mathrm{lin}}(N, \delta)$ and $r^{\mathrm{a-w}}_{\mathrm{non}}(N, \delta)$ be the minimal errors of linear and nonlinear algorithms, respectively.

Theorem 18 *(Plaskota, 1994) For approximating functionals with respect to a Gaussian measure we have*

$$1 \leq \frac{r^{\mathrm{a-w}}_{\mathrm{lin}}(N, \delta)}{r^{\mathrm{a-w}}_{\mathrm{non}}(N, \delta)} \leq 1,49...$$

Again, the proof uses a correspondence between the optimal linear algorithms in the mixed setting and in the average case setting, where $x \sim \mathcal{N}(0, \sigma^2 \Sigma)$.

Theorem 19 *(Plaskota, 1994) For any δ there exists $\sigma = \sigma(\delta)$ such that the same linear algorithm is optimal in the mixed setting and in the average case setting. And vice versa. For any σ there exists $\delta = \delta(\sigma)$ such that the same linear algorithm is optimal in both settings.*

This and earlier cited relations show some kind of equivalence of the four different settings: worst case, average case, and mixed, for approximating linear functionals.

5 Asymptotic setting

The asymptotic setting has somewhat different character. While earlier we wanted to minimize the error of algorithms for a given set E (or for a given measure μ), now we fix the problem element $f \in F$ and want to construct not a single, but a sequence of approximations that converge to the solution $S(f)$ as fast as possible. One may hope that with such an approach it is possible to find a sequence of information N^n and algorithms φ^n such that the error $\|S(f) - \varphi^n(N^n(f) + x^n)\|$ tends to zero much faster than the sequence of corresponding minimal (worst case or average case) errors. We shall show that this is not always true.

In the asymptotic setting it is convenient to assume that information is given as an infinite sequence $y = [y_1, y_2, \ldots] \in \mathbb{R}^\infty$, where

$$y_i = L_i(f; y_1, \ldots, y_{i-1}) + x_i, \qquad i \geq 1,$$

$|x_i| \leq \delta_i(y_1, \ldots, y_{i-1})$. Note that the functionals L_i as well as accuracies δ_i may be chosen adaptively. For information y, a sequence of successive approximations is given as $\{\varphi^n(y^n)\}$, where $y^n = [y_1, \ldots, y_n]$ and $\varphi^n : \mathbb{R}^n \to G$ is the n-th algorithm.

If F is a normed space, we denote by $r_n^{\mathrm{wor}}(N_z, \Delta_z)$ the optimal worst case error with respect to the unit ball in F and nonadaptive observations of the first n functionals $L_i(\cdot; z_1, \ldots, z_{i-1})$ with accuracies $\delta_i(z_1, \ldots, z_{i-1})$.

Theorem 20 *(Kacewicz and Plaskota, 1993) Let F be a Banach space, $S : F \to G$ a linear operator and $L_i(\cdot; y_1, \ldots, y_{i-1})$ continuous linear functionals. Then, for any positive sequence $\{\tau_n\}$ converging to zero, the set*

$$\left\{ f \in F : \limsup_{n\to\infty} \frac{\|S(f) - \varphi^n(y^n)\|}{\tau_n r_n^{\mathrm{wor}}(N_y, \Delta_y)} < +\infty, \ \forall y \textit{ information about } f \right\}$$

has empty interior, i.e., it does not contain any ball in F.

This theorem establishes an equivalence of the asymptotic setting with deterministic information noise, and the worst case setting. We note that the convergence $r^{\mathrm{wor}}(N_y, \Delta_y)$ is attained by spline algorithms, and the optimal convergence (in the case of varying information) by some nonadaptive information.

Consider now the case when the information noise is random. More precisely, assume that noise of successive observations is independent and $x_i \sim \mathcal{N}(0, \sigma_i^2)$, where $\sigma_i = \sigma_i(y_1, \ldots, y_{i-1})$. Also, assume that F is a separable Banach space equipped with a zero mean Gaussian measure, S is linear continuous, and G is a separable Hilbert space.

Let φ_{spl}^n be the spline algorithm (2) corresponding to the first n observations. Then we have the following theorems showing equivalence of the asymptotic setting with random information noise and the average case setting. (The theorems below are generalizations of the results from Wasilkowski and Woźniakowski (1987), where exact information is studied.)

Theorem 21 *(Plaskota, 1995b) For arbitrary sequence of algorithms $\{\varphi^n\}$,*

$$\mathrm{Prob}\left(\lim_{n\to\infty} \frac{\|S(f) - \varphi^n(y^n)\|}{\|S(f) - \varphi_{\mathrm{spl}}^n(y^n)\|} = 0 \right) = 0.$$

(Here, the probability is taken with respect to the a priori measure on $F \times \mathbb{R}^\infty$ which is induced by μ and the distribution of noise.)

Denote by $r_n^{\rm avg}(N_z, \Sigma_z)$ the optimal average error that can be achieved using the first n observations of the functionals $L_i(\cdot; z_1, \ldots, z_{i-1})$ with variances $\sigma_i^2(z_1, \ldots, z_{i-1})$.

Theorem 22 *(Plaskota, 1995b) For any sequence $\{\varphi^n\}$ we have*

$$\mathrm{Prob}\left(\lim_{n\to\infty} \frac{\|S(f) - \varphi^n(y^n)\|}{r_n^{\rm avg}(N_y, \Sigma_y)} = 0 \ \mathit{or}\ + \infty\right) = 0.$$

References

Donoho, D.L. (1994) Statistical estimation and optimal recovery. *Annals of Statistics*, **22**, 238–270.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia? J. Roy. Stat. Soc., **57**, 301–369.

Kacewicz, B.Z. and Plaskota, L. (1990) On the minimal cost of approximating linear problems based on information with deterministic noise. *Numer. Funct. Anal. and Optimiz.*, **11**, 511–525.

Kacewicz, B.Z. and Plaskota, L. (1991) Noisy information for linear problems in the asymptotic setting. *J. of Complexity*, **7**, 35–57.

Kacewicz, B.Z. and Plaskota, L. (1993) The minimal cost of approximating linear operators using perturbed information–the asymptotic setting. *J. of Complexity*, **9**, 113–134.

Magaril–Il'yaev, G.G. and Osipenko, K.Yu. (1991) On optimal recovery of functionals from inaccurate data. *Math. Notes*, **50**, 85–93.

Melkman, A.A. and Micchelli, C.A. (1979) Optimal estimation of linear operators in Hilbert spaces from inaccurate data. *SIAM J. Numer. Anal.*, **16**, 87–105.

Nemirovski, A.S. and Yudin, D.B. (1983) *Problem Complexity and Method Efficiency in Optimization.* Wiley and Sons, New York.

Novak, E. (1988) *Deterministic and Stochastic Error Bounds in Numerical Analysis.* Vol. 1349 of Lecture Notes in Math. Springer, Berlin.

Nussbaum, M. (1985) Spline smoothing in regression model and asymptotic efficiency. *Annals of Statistics*, **13**, 984–997.

Plaskota, L. (1990) On average case complexity of linear problems with noisy information. *J. of Complexity*, **6**, 199–230.

Plaskota, L. (1992) Function approximation and integration on the Wiener space with noisy data. *J. of Complexity*, **8**, 301–323,

Plaskota, L. (1993a) Optimal approximation of linear operators based on noisy data on functionals. *J. of Approx. Theory*, **73**, 93–105.

Plaskota, L. (1993b) A note on varying cardinality in the average case setting. *J. of Complexity*, **9**, 458–470.

Plaskota, L. (1994) Average case approximation of linear functionals based on information with deterministic noise. *J. of Computing and Information*, **4**, 21–39.

Plaskota, L. (1995a) Average complexity for linear problems in a model with varying information noise. *J. of Complexity*, **11**.

Plaskota, L. (1995b) *Noisy Information and Computational Complexity.* To appear in Cambridge University Press, Cambridge.

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

TRAUB, J.F., WASILKOWSKI, G.W. AND WOŹNIAKOWSKI, H. (1983) *Information, Uncertainty, Complexity.* Addison Wesley, Mass.

TRAUB, J.F., WASILKOWSKI, G.W. AND WOŹNIAKOWSKI, H. (1988) *Information–based Complexity.* Academic Press, New York.

WAHBA, G. (1990) *Spline Models for Observational Data.* Vol. 59 of CBMS–NSF Series in Appl. Math., SIAM.

WASILKOWSKI, G.W. AND WOŹNIAKOWSKI, H. (1987) On optimal algorithms in an asymptotic model with Gaussian measure. *SIAM J. of Math. Anal.*, **3**, 632–647.

WERSCHULZ, A.G. (1991) *The Computational Complexity of Differential and Integral Equations.* Oxford University Press, Oxford.