WHEN IS NONADAPTIVE INFORMATION AS POWERFUL AS ADAPTIVE INFORMATION?

J. F. Traub
G. W. Wasilkowski
H. Woźniakowski

CUCS-149-84

Proceedings of 23rd Conference
on Decision and Control
Las Vegas, NV, December 1984

WHEN IS NONADAPTIVE INFORMATION AS POWERFUL AS ADAPTIVE INFORMATION?*

J.F. Traub**, G.W. Wasilkowski*** and H. Woźniakowski***

Abstract

Information based complexity is a unified treatment of problems where only partial or approximate information is available. In this approach one states how well a problem should be solved and indicates the type of information available. The theory then tells one optimal information and optimal algorithm and yields bounds on the problem complexity. In this paper we survey some recent results addressing one of the problems studied in information based complexity. The problem deals with nonadaptive and adaptive information both for the worst case and average case settings.

1. Introduction

The purpose of this paper is to survey recent work in information based complexity on the effectiveness of adaptive versus nonadaptive information.

To explain what we mean by adaptive and nonadaptive information we use a simplified version of the prediction problem. Suppose that for a function f from a given class F one seeks an approximation x to $f(t^*)$. The approximation $x = x(f)$ is constructed depending on some partial information about f which is available at the present time. Typically the information consists of n function values, $N(f) = [f(t_1),..., f(t_n)]$, with some restrictions on sample points t_i, say $t_i < t^*$. Due to the finiteness of the information, $N(f)$ does not identify $f(t^*)$ uniquely and, in general, there exists infinitely many functions from the given class F which share the same information and have different values $f(t^*)$. This means that the information $N(f)$ causes an intrinsic uncertainty which cannot be reduced no matter how one approximates $f(t^*)$. Of course, we are interested in information with the intrinsic uncertainty as small as possible. That is, we are interested in an optimal choice of sampling points to reduce uncertainty. There are two different ways of selecting the sampling points $t_1, t_2, \ldots, t_n$. The first one is by selecting them a priori. In this case, infomation $N = N^{non}$ is called nonadaptive. The second way is by selecting them adaptively, i.e., the choice of the point t_2 depends on the value $f(t_1)$, t_3 depends on $f(t_1)$, $f(t_2)$, and so on. In this case $N = N^{a}$ is called adaptive. Since the structure of adaptive infomation is far richer than the structure of nonadaptive information, one might hope that for adaptively chosen sample points the uncertainty is much smaller than for nonadaptively chosen ones. Is this really the case? Or equivalently,

Is adaption more powerful than nonadaption?

Before answering this question we want to stress that this problem is not merely of theoretical interest. Adaptive information has several undesirable properties:

- it has more complicated structure than nonadaptive information,
- it is ill-suited for parallel or distributed computation whereas nonadaption can be computed very efficiently in parallel,
- the idea of precomputation can not be used when dealing with adaptive information.
- because of its complicated structure, it is far harder to find optimal adaptive information.

Due to these undesirable properties of adaptive infomation one should use adaptive information only if it causes significantly smaller uncertainty than nonadaptive information.

There is a number of papers addressing this question for specific problems. We believe, Kiefer in 1957 [5] was the first one to show that adaption does not help for approximation of the integral of a function f from a certain class F. In 1971, Bakhovalov [1] proved that adaption does not help for the approximation of linear functionals assuming that the given class F is balanced and convex. (Balanced means that $f \in F$ implies $-f \in F$.) This result was generalized by Gal and Micchelli [2] and Traub and Woźniakowski [12] in 1980 for the approximation of linear operators also assuming that F is balanced and convex. Further generalizations can be found in [10].

There is a number of papers addressing the problem of adaptive information for the approximation of nonlinear operators, see [3, 8, 9, 15]. For some nonlinear problems adaption helps, for some it doesn't. For instance, in 1982 Sikorski [7] proved that adaption is exponentially better than nonadaption for the zero-finding problem for the class F of scalar regular functions with different signs at the endpoints.

In all papers cited above the uncertainty was measured by the worst performance, i.e., by the error caused by the hardest element f. It is also known, see [11, 16, 17], that adaption does not help on the average for the approximation of linear operators in a Hilbert space. Here the uncertainty is measured by the average error with respect to some probability measure. The same result holds when the information has stochastic error, see [4].

Based on these results we may conclude that nonadaptive information is as powerful as adaptive infomation for the approximation of linear operators defined on balanced and convex classes. On the other hand, if one approximates a nonlinear operator or a linear operator defined on an unbalanced and/or nonconvex class F, then adaption may help significantly.

We summarize the contents of this paper. In Section 2 we precisely define what we mean by a problem and by information. We discuss measuring uncertainty

*This work was supported in part by the National Science Foundation under Grant MCS-82-14322 and by the Advanced Research Projects Agency under contract N00039-82-C-0427.

**Department of Computer Science, Columbia University, New York, NY,

***Department of Computer Science, Columbia University, New York, NY and Department of Mathematics, Computer Science and Mechanics, University of Warsaw, Poland.

Reprinted from THE 23RD IEEE CONFERENCE ON DECISION AND CONTROL, December 12-14, 1984

by the worst and average performances of algorithms. In Section 3 we formally define the concept of nonadaptive and adaptive information. In Sections 4 and 5 we survey some results which state when adaption does not help in the worst and average case settings respectively. We also give a new example of a linear operator considered on an unbalanced and nonconvex set for which adaption is exponentially better than nonadaption. In Section 6 we briefly discuss the lack of power of adaptive information in the asymptotic setting.

2. Radius of Information

We define "problem" and "information". The prediction problem mentioned in the Introduction is an example of such a problem. We define a fundamental quantity, the radius of information, which measures the intrinsic uncertainty in solving a problem, due to the available information.

For given normd linear spaces F_1 and F_2 let

$$S : F_1 \to F_2$$

be an operator (in general nonlinear). We call S a solution operator. We wish to construct an element x, $x = x(f) \in F_2$, which approximates $S(f)$ as close as possible. In information based complexity we assume that the element f is unknown. Instead we assume that the knowledge about f is provided by $N(f)$ where N, called information, is an operator

$$N : F_1 \to F_3,$$

for some space F_3. In most cases N is many-to-one. We call such information partial. Thus the knowledge of $N(f)$ does not identify f uniquely. Knowing $N(f)$ we construct an approximation $x = x(f)$ to $S(f)$ by an algorithm ϕ,

$$x = \phi(N(f)).$$

Here by an (idealized) algorithm that uses N we mean any mapping

$$\phi: N(F_1) \subseteq F_3 \to F_2.$$

Since in general $N(f)$ does not identify $S(f)$ uniquely, $\phi(N(f))$ has to approximate $S(\tilde{f})$ for all elements $\tilde{f}$ which share the same information as f, $N(\tilde{f}) = N(f)$. This means that partial information N causes intrinsic uncertainty. Here we discuss measuring the uncertainty in two different settings: worst case and average case settings.

We begin with the

(i) Worst Case Setting

Assume we want to approximate $S(f)$ for f from a given subclass F of the space F_1. Typically, F is defined by restricting f. For example $F = \{f \in F_1 : \|f^{(r)}\| \le 1\}$ where F_1 is a space of regular functions. In the worst case model the error of an algorithm is determined by its worst performance. That is, the error of ϕ is defined by

$$e^{w}(\phi,N) = \sup_{f\in F} \|S(f) - \phi(N(f))\|.$$

By the (worst case) radius of information N we mean

$$r^{w}(N) = \inf_{\phi} e^{w}(\phi,N).$$

The radius of information measures the intrinsic uncertainty caused by information N, and no algorithm that uses N can have a smaller error than the radius $r^{w}(N)$. Hence, if one wants to find an algorithm ϕ which approximates $S(f)$ within a given accuracy ε, i.e.,

$$\|S(f) - \phi(N(f))\| < \varepsilon, \quad \forall f \in F,$$

this can be done if and only if $r^{w}(N) < \varepsilon$.

We want to add that the radius of information can be defined independently of the notion of algorithm. It only depends on the solution operator S, the class F and information N. The fact that the radius of N is a sharp lower bound on the error of any algorithm ϕ, is a conclusion. See [12, 13].

One might say that the error of an algorithm ϕ in the worst case setting is defined too pessimistically. The algorithm may perform quite well for "most" elements f. If ϕ performs badly for one element f^* then its error in the worst case model is determined by its bad behavior for f^* and does not reflect the good behavior of ϕ for most elements f. Therefore it is natural to define the error of ϕ by its "average" performance. This leads us to the

(ii) Average Case Setting

Assume that we are given a probability measure μ on the class F. The error of an algorithm is determined by its average performance,

$$e^{avg}(\phi,N) = \sqrt{\int_F \|S(f) - \phi(N(f))\|^2 \mu(df)}.$$

In the average case model the uncertainty caused by information N is measured by the average radius of N defined by

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

Hence, the average radius of information N is the sharp lower bound on the average error of any algorithm ϕ. We can approximate $S(f)$ within ε on the average if and only if $r^{avg}(N) < \varepsilon$.

3. Adaptive and Nonadaptive Information

In this section we define the concepts of nonadaptive and adaptive information. Let $L_i: F_1 \to \mathbb{R}$, $i = 1,2,\ldots$, be a linear functional. We say N is nonadaptive information of cardinality n if

$$N(f) = [L_1(f),\ldots,L_n(f)], \quad \forall f \in F_1.$$

The information N is called nonadaptive since the linear functionals $L_1, L_2, \ldots, L_n$ are given simultaneously. In parallel computation, one can evaluate $N(f)$ by evaluating $L_i(f)$ on different processors. For the prediction problem, $L_i(f) = f(x_i)$ for some x_i. Then nonadaption means that the points $x_1, x_2, \ldots, x_n$ are determined a priori.

We now turn to adaptive information. The essence of adaption is that the functionals L_i are not chosen a priori, and the ith functional depends on the previously computed information. More precisely, N is called adaptive information of cardinality n if

$$N(f) = [L_1(f), L_2(f,y_1), \ldots, L_n(f,y_1,\ldots,y_{n-1})], \quad \forall f \in F_1,$$

where $y_1 = y_1(f) = L_1(f)$, $y_i = y_i(f) = L_i(f,y_1,\ldots,y_{i-1})$. We assume that $L_i(\cdot,y_1,\ldots,y_{i-1})$ are linear functionals for every $y = [y_1,\ldots,y_n] \in \mathbb{R}^n$. For the prediction problem we have $L_i(f,y_1,\ldots,y_{i-1}) = f(x_i(y_1,\ldots,y_{i-1}))$, i.e., the point at which f is evaluated depends on the previously computed values of f.

Observe that nonadaptive information N is a linear operator from F_1 into $\mathbb{R}^n$, whereas adaptive

information N is in general a nonlinear operator.

4. When Does Adaption Not Help for the Worst Case Setting?

In the Introduction we give a brief history of the work on adaptive information for the approximation of linear operators. We now state a theorem which was proven by Gal and Micchelli [2] and Traub and Woźniakowski [12].

Theorem 4.1

Let S be linear and let F be balanced and convex. Then for every adaptive information N there exists nonadaptive information N^{non} of the same cardinality as N, for which

$$r^w(N^{non}) \le 2r^w(N).$$

This theorem states that for linear operators considered on balanced and convex sets, nonadaptive information is (within a constant two) as powerful as adaptive information. We want to add that for many cases we have a stronger result, $r^w(N^{non}) \le r^w(N)$. This holds, for instance, when F is a ball with respect to some semi-innerproduct or when the norm in F_2 is induced by some innerproduct. The assumptions that F is balanced and convex are crucial. We now show an example of approximating a linear operator S on F which is neither balanced nor convex, for which adaption is exponentially more powerful as nonadaption. Since this example is new we provide a sketch of the proof.

Example 4.1

Let F be the set of functions which takes only two values $\{0,1\}$ and which have exactly one discontinuity point. More precisely,

$$F = \{f: [0,1] \to \mathbb{R}: \exists x_f \in [0,1],\ f(x)=0 \text{ for } x\in[0,x_f) \text{ and } f(x)=1 \text{ for } x\in(x_f,1]\}.$$

Let F_1 be any linear normed space containing F as a subset. Let $F_2 = L_2[0,1]$. Define the solution operator $S: F_1 \to F_2$ by $Sf = f$. Note that the solution operator S is linear whereas the class F is neither balanced nor convex. Therefore the assumptions of Theorem 4.1 are not satisfied.

Take arbitrary nonadaptive information N, $N(f) = [f(x_1),\ldots,f(x_n)]$ with $0 = x_0 \le x_1 < x_2 < \ldots < x_n \le x_{n+1} = 1$. For given $y = [y_1,\ldots,y_n] \in N(F)$, let $i = i(y)\in[0,n]$ be the maximal index such that $y_i = 0$. Knowing $y = N(f)$ we conclude that $f(x)=0$ for $x \le x_i$ and $f(x)=1$ for $x > x_{i+1}$. Therefore the set $V(N,y)$ of all functions from F which share the same information y has the following form

$$V(N,y) = \{f\in F: \text{the discontinuity point } x_f \text{ of } f \text{ belongs to } (x_i,x_{i+1}]\}.$$

Note that $f_1,f_2 \in V(N,y)$ implies $\|f_1-f_2\| = \sqrt{|x_{f_1} - x_{f_2}|} \le \sqrt{x_{i+1} - x_i}$ and that this bound is sharp. This yields that the radius of informatin N is given by

$$r^w(N) = \frac{1}{2} \max_{0\le i\le n} \sqrt{x_{i+1}-x_i} \ge \frac{1}{2}\sqrt{\frac{1}{n+1}}.$$

This means that for every nonadaptive information N, $N(f) = [f(x_1),\ldots,f(x_n)]$,

$$r^w(N) \ge \frac{1}{2}\sqrt{\frac{1}{n+1}}.$$

The equality is achieved for nonadaptive information $N_n^{non}(f) = [f(\frac{1}{n+1}),\ldots,f(\frac{n}{n+1})]$,

$$r^w(N_n^{non}) = \frac{1}{2}\sqrt{\frac{1}{n+1}}.$$

This shows that equidistant points $x_i = i/(n+1)$ are optimal for nonadaptive evaluations of f.

We show that using adaptive information one can significantly decrease the uncertainty. Suppose $x_1 = \frac{1}{2}$ and we compute $f(\frac{1}{2})$ for $f \in F$. If $f(\frac{1}{2}) = 1$ then we conclude that $f(x) = 1$ for $x \in [\frac{1}{2},1]$. If $f(\frac{1}{2}) = 0$ then we conclude that $f(x) = 0$ for $x \in [0,\frac{1}{2}]$. In either case we know the function f exactly on a subinterval of length $\frac{1}{2}$. Then we choose the next point x_2 of evaluation of f as the midpoint of the subinterval on which f is unknown, i.e.,

$$x_2 = \begin{cases} \frac{3}{4} & \text{if } f(x_1) = 0, \\ \frac{1}{4} & \text{if } f(x_1) = 1. \end{cases}$$

Note that x_2 depends on $y_1 = f(x_1)$. This means that we use adaption. Knowing $f(x_1)$ and $f(x_2)$ we conclude the behavior of f on the whole interval $[0,1]$ except a subinterval of length $\frac{1}{4}$. The next point x_3 of evaluation will be chosen as a midpoint of this subinterval. Let $N_n^a(f) = [f(x_1), f(x_2(y_1)),\ldots,f(x_n(y_1,\ldots,y_{n-1}))]$ be the adaptive information described as above. That is $x_1 = \frac{1}{2}$, $x_2 = x_2(y_1)$ is given by the formula above and so on. The adaptive information N_n^a is called bisection information since the point x_i bisects the uncertainty interval of f. Note that knowing $y=N_n^a(f)$, the uncertainty interval of f is of the length 2^{-n}. Hence

$$r^w(N_n^a) = \frac{1}{2}\sqrt{2^{-n}}.$$

In fact, one can prove that any adaptive information consisting of n function evaluations at adaptively chosen points has the radius no less than $r^w(N_n^a)$. This means that

the bisection information points are optimal for adaptive evaluations of f.

Compare now the radii $r^w(N_n^{non})$ and $r^w(N_n^a)$. We see that adaption reduces uncertainty exponentially better than nonadaption.

We conclude this section by adding that for some nonlinear problems, i.e., for nonlinear solution operators S, adaption does not help, see [3, 8, 9, 15]. For some other nonlinear problem, adaption does help, see [7, 12].

5. When Does Adaption Not Help for the Average Case Setting?

In this section we report some results addressing the problem of adaptive information on the average. For simplicity we assume that the class F is equal to the whole space F_1, and the spaces F_1 and F_2 are separable Hilbert spaces. Let μ be a Gaussian measure on F_1 with mean element zero and covariance operator S_μ. This means that

$$\int_F e^{i(f,x)}\,\mu(df) = e^{-\frac{1}{2}(S_\mu x,x)},$$

$$\forall\, x \in F_1,\quad i = \sqrt{-1}\ ,$$

see [6]. Assume also that S is a continuous linear operator. Then from [17] we have

Theorem 5.1

For every adaptive information N there exists nonadaptive information N^{non} of the same cardinality as N, for which

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

We want to stress that Theorem 5.1 holds for more general probability measures and for any separable Banach space F_1. If F is a finite dimensional space, Theorem 5.1 was proven in [11]. Theorem 5.1 remains valid assuming that $N(f)$ is computed with stochastic error, see [4]. Adaption also does not help for a more general definition of the average error, see [16].

We conclude this section by continuing Example 4.1. We show that adaption also helps on the average.

Example 5.1:

Let F and S be as in Example 4.1. Let μ be defined on F by

$$\mu(A) = \ell(\{x \in [0,1]:\ x \text{ is the discontinuity point of some } f \text{ from } A\}),$$

where ℓ stands for the Lebesgue measure on $[0,1]$.

For an arbitrary nonadaptive information N, $N(f) = [f(x_1),\ldots,f(x_n)]$ with $0 = x_0 \le x_1 < \ldots < x_n \le x_{n+1} = 1$, N takes only $n+1$ different values: $y^0 = [1,\ldots,1]$, $y^1 = [0,1,\ldots,1]$, $\ldots$, $y^n = [0,\ldots,0]$. Since the sets $V(N,y^i)$ are disjoint and $F = \bigcup_{i=0}^{n} V(N,y^i)$, we have

$$\begin{aligned} r^{avg}(N)^2 &= \inf_{\phi} \int_F \|f-\phi(N(f))\|^2 \mu(df) \\ &= \sum_{i=0}^{n} \inf_{\phi} \int_{V(N,y^i)} \|f-\phi(y^i)\|^2 \mu(df) \\ &= \sum_{i=0}^{n} \inf_{g \in F_2} \int_u^{x_{i+1}} \Big\{ \int_{x_i}^{x_{i+1}} g^2(x)\,dx \\ &\quad + \int_u^{x_{i+1}} (1-g(x))^2\,dx \Big\} du \\ &= \frac{1}{6} \sum_{i=0}^{n} (x_{i+1}-x_i)^2 \ge \frac{1}{6(n+1)}\ . \end{aligned}$$

This means that for arbitrary nonadaptive information N of cardinality n,

$$r^{avg}(N) \ge \sqrt{\frac{1}{6(n+1)}}\ .$$

As in Example 4.1 the equality is achieved for $N_n^{non}(f) = [f(\frac{1}{n+1}),\ldots,f(\frac{n}{n+1})]$,

$$r^{avg}(N_n^{non}) = \sqrt{\frac{1}{6(n+1)}}\ .$$

This shows that equidistant points are optimal for nonadaptive evaluations of f on the average. Similarily, one can find the average radius of the bisection information,

$$r^{avg}(N_n^a) = \sqrt{\frac{1}{6\cdot 2^n}}\ .$$

This shows that adaption is much more powerful than nonadaption also on the average.

6. Asymptotic Setting

One might think that the lack of power of adaption for linear problems is due to the fact that the cardinality of information is fixed for all elements f. One might hope that when information operators of increasing cardinality are used for fixed f, adaption became more powerful. This problem is analyzed in an asymptotic case setting. It turns out that for the approximation of linear operators adaption also does not help in this setting. See [14, 18].

References

1. N.S. Bakhovalov, "On the optimality of linear methods for operator approximation in convex classes of functions", U.S.S.R. Computational Math. and Math. Phys., vol. 11, pp. 244-249, 1971.

2. S. Gal and C.A. Micchelli, "Optimal sequential and nonsequential procedures for evaluating a functional", Appl. Anal., vol. 10, pp. 105-120, 1980.

3. B. Kacewicz, "On the optimal error of algorithms for solving a scalar autonomous ODE", BIT, vol. 22, pp. 503-518, 1982.

4. J.B. Kadane, G.W. Wasilkowski and H. Woźniakowski, "Can adaption help on the average for stochastic information?", in progress.

5. J. Kiefer, "Optimum sequential search and approximation methods under minimum regularity assumptions", J. Soc. Ind. Appl. Math., vol. 5, pp. 105-136.

6. H. Kuo, Gaussian measures in Banach spaces, Lecture Notes in Mathematics 463, Springer-Verlag, 1975.

7. K. Sikorski, "Bisection is optimal", Numer. Math., vol. 40, pp. 111-117, 1982.

8. K. Sikorski and H. Woźniakowski, "For which error criteria can we solve nonlinear equations?", Dept. of Comp. Sci., Report, Columbia University, 1983.

9. A.G. Sukharev, "Optimal strategies of the search for an extremum", U.S.S.R. Computational Math. and Math. Phys., vol. 11, pp. 119-137, 1971.

10. J.F. Traub, G.W. Wasilkowski and H. Woźniakowski, Information, Uncertainty, Complexity, Addison-Wesley, Reading, Mass., 1983.

11. J.F. Traub, G.W. Wasilkowski and H. Woźniakowski, "Average case optimality for linear problems", Th. Comp. Sci., vol. 29, pp. 1-25, 1984.

12. J.F. Traub and H. Woźniakowski, A general theory of optimal algorithms, Acad. Press, New York, 1980.

13. J.F. Traub and H. Woźniakowski, "Information and computation", in Advances in computers, vol. 23, ed. M.C. Yovits, Acad. Press, New York, 1984.

14. G.M. Trojan, "Asymptotic setting for linear problems", in progress.

15. G.W. Wasilkowski, "Some nonlinear problems are as easy as the approximation problem", to appear in Comput. Math. Appl.

16. G.W. Wasilkowski, "Optimal Algorithms for linear problems with Gaussian measure", Dept. of Comp. Sci., Report, Columbia University, 1984.

17. G.W. Wasilkowski and H. Woźniakowski, "Can adaption help on the average?", to appear in Numer. Math.

18. G.W. Wasilkowski and H. Woźniakowski, "Asymptotic setting for linear problems in Hilbert spaces with Gaussian measure", in progress.